SIAM J. Applied Dynamical Systems
Vol. 8, No. 1, pp. 1–20
© 2009 Society for Industrial and Applied Mathematics

Stabilizability via Time-Delayed Feedback: An Eigenvalue Optimization Approach*

Henri Huijberts†, Wim Michiels‡, and Henk Nijmeijer§

Abstract. In this paper we consider the stabilizability of dynamical systems via time-delayed state and output feedback control. Based on an eigenvalue optimization approach in combination with a continuation argument and a result characterizing properties of the optimum eigenvalue configurations, we obtain explicit expressions for stabilizability boundaries for two-dimensional systems. As time-delayed feedback can be used to approximate derivative feedback, we also give a comparison of the stabilizability via both types of feedback.

Key words. stabilizability, time-delayed feedback control, eigenvalue optimization, nonsmooth optimization

AMS subject classifications. 39B82, 39D15

DOI. 10.1137/070708767

1. Introduction.

In recent years a growing interest in the stability of time-delay systems is emerging in the control literature. The reasons for this are manifold, and the interest is motivated by possible applications in networked control systems, chemistry, and biology. For a review on recent results and techniques in stability analysis of time-delay systems, we refer to, e.g., [8], [13], [20].

In the applications indicated above, one mainly deals with either the control of systems whose dynamics involve some inherent time-delay, or the control of systems in the presence of inherent time-delay in the feedback loop. On the other hand, one might ask whether it might be beneficial to impose time-delay in a controller. One of the reasons for this might be to achieve noninvasive stabilization. This leads to the introduction of so-called time-delayed feedback, and in this paper we will study the problem of stabilization of equilibria by means of this time-delayed feedback control. The idea of using time-delayed feedback was originally proposed by Pyragas in [18] in the context of stabilization of periodic solutions of chaotic systems; see also, e.g., [11], [16]. Later, in the physics literature the use of time-delayed feedback was extended to the problem of stabilization of equilibrium points; see, e.g., [19], [10], [22].

*Received by the editors November 20, 2007; accepted for publication (in revised form) by B. Krauskopf June 13, 2008; published electronically January 9, 2009.
http://www.siam.org/journals/siads/8-1/70876.html

†School of Engineering and Materials Science, Queen Mary, University of London, Mile End Road, London E1 4NS, United Kingdom (H.J.C.Huijberts@qmul.ac.uk).

‡Department of Computer Science, Katholieke Universiteit Leuven, Celestijnenlaan 200A, B-3001 Heverlee-Leuven, Belgium (Wim.Michiels@cs.kuleuven.be). The research of this author has been supported by the Belgian Programme on Interuniversity Poles of Attraction, initiated by the Belgian State, Prime Minister's Office for Science, Technology and Culture, and by the Optimization in Engineering Centre OPTEC of the K.U.Leuven. This author is a postdoctoral fellow of the Research Foundation Flanders (Belgium).

§Department of Mechanical Engineering, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands (h.nijmeijer@tue.nl).

Besides this motivation from the physics point of view, there is also envisaged to be a practical motivation from the control engineering point of view. Namely, there is an increasing interest in using derivative control. Typical applications are in vibration control of mechanical systems where the state variables are positions and velocities, while the accelerations, which are the sensed variables, are directly used for feedback; see [5], [17]. In some applications, including vibration suppression, the fact that a derivative feedback keeps the steady state solutions of the uncontrolled system invariant, i.e., the control law is noninvasive, is considered a positive feature. However, there are problems in the application of "pure" derivative feedbacks in that it is often difficult to obtain reliable measurements of derivatives, especially in a noisy environment. In the literature, this problem has been addressed in, e.g., [9], where the authors used so-called washout filters that also possess the noninvasiveness property. Alternatively, it appears that in this respect time-delayed feedback as an approximation of derivative feedback, or even time-delayed feedback per se, seems to be a promising alternative that also has the noninvasiveness property.

In the area of time-delayed stabilization of periodic orbits, it has long been contended that periodic orbits with an odd number of real Floquet multipliers greater than unity cannot be stabilized by time-delayed feedback control; see, e.g., [16], [11]. Recently, it was shown in [7] (see also [12]) that this so-called odd-number limitation does not hold true for the stabilization of periodic orbits. However, it does still hold true for the stabilization of equilibrium points by means of time-delayed feedback control.

Recently, [10], [22] considered the characterization of stabilizability of unstable equilibria of (mainly two-dimensional) dynamical systems by means of time-delayed feedback. Some interesting analytical results regarding this problem, supplemented with numerical simulations, were given in these references. It is to be noted, however, that both references mainly considered diagonal state feedbacks.

The present paper takes inspiration from [10], [22], and our goal is to provide a complete classification of all possible linear time-delayed state and output feedbacks that may be used to stabilize an unstable equilibrium of a two-dimensional dynamical system. Even though we will be dealing with linear systems, the stabilizability problem gives rise to the analysis of the solutions of a parametrized nonlinear eigenvalue problem. For this analysis we employ methods and tools from bifurcation analysis. In particular, our approach is based on the construction of determining systems and the continuation of their solutions.

We consider a two-dimensional single-input-single-output nonlinear control system of the form

$$\left\{ \begin{array}{rcl} \dot{z} & = & f(z,u), \\ y & = & h(z), \end{array} \right. \tag{1.1}$$

with state $z \in \mathbb{R}^2$, input $u \in \mathbb{R}$, and output $y \in \mathbb{R}$. We assume that the origin is a steady state (i.e., $f(0,0) = 0$), which is hyperbolic and unstable. As in, e.g., [10], [22], we then consider the linearization of (1.1) about $(x,u) = (0,0)$, which is given by

$$\left\{ \begin{array}{rcl} \dot{z} & = & Az + Bu, \\ y & = & \begin{pmatrix} c_2 & c_1 \end{pmatrix} z, \end{array} \right. \tag{1.2}$$

where

$$A = D_z f(0,0), \quad B = D_u f(0,0), \quad \begin{pmatrix} c_2 & c_1 \end{pmatrix} = D_z h(0). \tag{1.3}$$

Assuming that the linearized system is controllable, it may then be assumed without loss of generality that A and B have the following forms (see, e.g., [4]):

$$A = \begin{pmatrix} 0 & 1 \\ -a_2 & -a_1 \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{1.4}$$

In this paper we will study the question of under which conditions on the system parameters a_1, a_2, c_1, c_2 there exists either a *time-delayed state feedback* of the form

$$u(t) = -K(z(t) - z(t-\tau)), \quad K = \begin{pmatrix} K_2 & K_1 \end{pmatrix}, \tag{1.5}$$

or a *time-delayed output feedback* of the form

$$u(t) = -k(y(t) - y(t-\tau)) \tag{1.6}$$

such that the origin is an exponentially stable equilibrium point of the closed loop system (1.2), (1.5) or (1.2), (1.6), respectively.

Note that due to the odd number-limitation discussed above, the equilibrium of the system (1.5) can never be stabilized by means of a time-delayed feedback of the form (1.5) or (1.6) if the matrix A has an odd number of positive real eigenvalues. Therefore, we can throughout restrict ourselves to the case that $a_2 \geq 0$.

The aim of this paper is to give a *complete* answer to the question of whether and how the linearized system (1.2) can be stabilized by means of a time-delayed state feedback (1.5) or a time-delayed output feedback (1.6). Our approach uses a methodology from [14], which exploits an eigenvalue optimization approach in combination with a continuation argument and a result characterizing properties of the optimum eigenvalue configurations. The latter argument allows one to "guess" analytical expressions for the boundaries of stabilizability regions that can then be verified by means of numerical computations on a very coarse grid.

The paper is organized as follows. In section 2 we first study the problem of state feedback stabilization. After that, the problem of output feedback stabilization will be discussed in section 3, where we will restrict ourselves to the case that $c_1 c_2 \leq 0$. In both sections 2 and 3 we will first treat the case of a fixed value of the time-delay τ. After this, the conclusions drawn for this case will be used to derive stabilizability conditions in case the time-delay can also be used as a controller parameter. Since time-delayed feedback could be used to approximate derivative feedback, we also include a discussion of the relationships between stabilizability via both types of feedback at the end of section 3. Finally, in section 4 conclusions will be drawn.

2. State feedback stabilization.

In subsection 2.1 we first outline an eigenvalue optimization-based stabilization approach that is instrumental to the solution of our problem. Next, in subsection 2.2 we completely characterize the stabilizable systems in the (a_1, a_2, τ) parameter space.

2.1. Stabilization through eigenvalue optimization. The characteristic equation of the closed loop system (1.2), (1.5) is given by

$$p(s;\, a_1, a_2, \tau, K_1, K_2) = 0,$$

where

$$\begin{aligned} p(s;\, a_1, a_2, \tau, K_1, K_2) &= \det \begin{pmatrix} s & -1 \\ a_2 + K_2(1 - \mathrm{e}^{-s\tau}) & s + a_1 + K_1(1 - \mathrm{e}^{-s\tau}) \end{pmatrix} \\ &= s^2 + \left(a_1 + K_1(1 - \mathrm{e}^{-s\tau})\right) s + \left(a_2 + K_2(1 - \mathrm{e}^{-s\tau})\right). \end{aligned} \tag{2.1}$$

Given a_1, a_2, and τ, we define the *spectral abscissa function* $F : \mathbb{R}^2 \to \mathbb{R}$ by

$$F(K;\, a_1, a_2, \tau) := \sup\{\Re(s) \mid p(s;\, a_1, a_2, \tau, K_1, K_2) = 0\}. \tag{2.2}$$

The zeroes of (2.1) whose real parts are equal to (2.2) are called the *active* eigenvalues. We further define $c(a_1, a_2, \tau)$ as

$$c(a_1, a_2, \tau) := \min_K F(K;\, a_1, a_2, \tau). \tag{2.3}$$

This then gives that the stabilization problem is solvable for given (a_1, a_2, τ) if and only if $c(a_1, a_2, \tau) < 0$, and the boundary of the stabilizability region is determined by values of a_1, a_2, and τ for which $c(a_1, a_2, \tau) = 0$.

From the above we see that the stabilization problem under consideration can be viewed as an optimization problem. Furthermore, assertions about stabilizability can be made by minimizing the spectral abscissa function (2.2).

As shown in [21], the spectral abscissa function is not everywhere differentiable, even not everywhere Lipschitz continuous. Moreover, discontinuities of its derivatives typically occur in the minima, which prohibits the use of most standard optimization methods. However, as the spectral abscissa function is continuous and differentiable almost everywhere, so-called bundle gradient methods like the *gradient sampling algorithms* of [3] are applicable. These methods rely only on the evaluation of the objective function and its gradient in points where the objective function is smooth. In [2], [3] the gradient sampling algorithm, which directly generalizes the steepest descent method to piecewise-smooth objective functions, was successfully applied to the design of stabilizing fixed-order controllers, and recently it was included in HIFOO [1], a software package for fixed-order controller design and $\mathcal{H}_\infty$-optimization. In [21] the algorithm was applied to the stabilization of time-delay systems, which required a combination with the routines of the package DDE-BIFTOOL [6] for the computation of the eigenvalues determining the stability of these systems.

An application of the gradient sampling algorithm, implemented as described in [21], for minimizing the objective function (2.2) yields the results displayed in Figure 1 for $(a_1, a_2) = (-5, 1.5)$ and $\tau = 1$. In Figure 1(a) the evolution of the objective function as a function of the number of iterations of the algorithm is displayed, when initialized with $(K_1, K_2) = (0, 0)$. As the method is gradient-based, the objective function is monotonically decreasing until a local minimum is reached, where $F = -0.04501$. Other simulations run with different, randomly

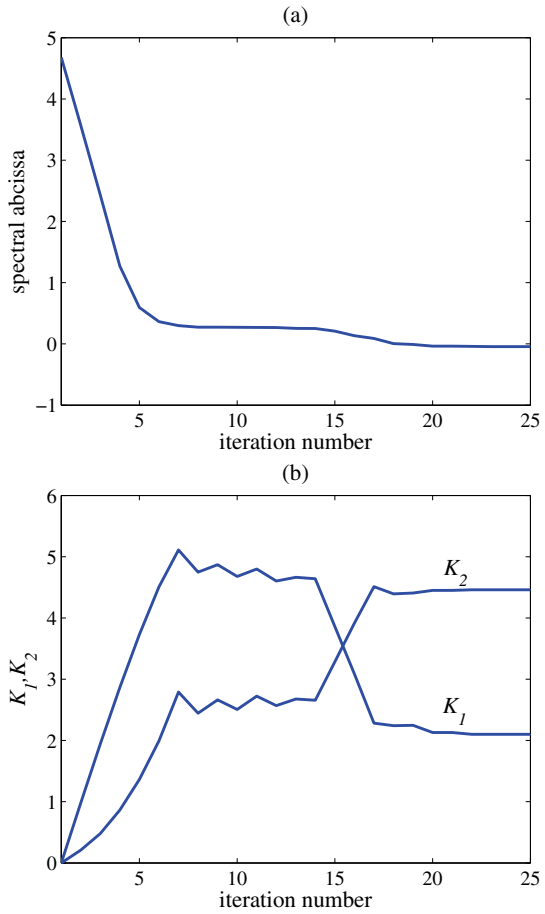


Figure 1. *Iterations of the gradient sampling algorithm applied to the minimization of* (2.2) *for* $(a_1, a_2, \tau) = (-5, 1.5, 1)$. (a) *Evolution of the objective function as a function of the number of iterations of the algorithm,* (b) *corresponding gain values.*

chosen, initial conditions reveal that this minimum corresponds to the global minimum. In Figure 1(b) the evolution of the gain values K_1 and K_2 as a function of the iteration number is shown. As the spectral abscissa is negative in the global minimum the system (1.2), (1.5) is stabilizable for $\tau = 1$ and the aforementioned values of a_1 and a_2. In Figure 2, the active eigenvalues corresponding to the minimum are displayed on two different scales. The minimum is characterized by a pair of complex conjugate active eigenvalues with multiplicity two.

2.2. Characterization of stabilizable systems. It is sufficient to compute the stabilizability region in the (a_1, a_2) plane for one *arbitrary* nonzero delay value, because of the following scaling property:

$$p(s;\, a_1, a_2, \tau, K_1, K_2) = 0 \iff p(\hat{s};\, a_1\tau, a_2\tau^2, 1, K_1\tau, K_2\tau^2) = 0, \quad \hat{s} = s\tau, \tag{2.4}$$

which implies the relation

$$c(a_1, a_2, \tau) = 0 \iff c(a_1\tau, a_2\tau^2, 1) = 0.$$

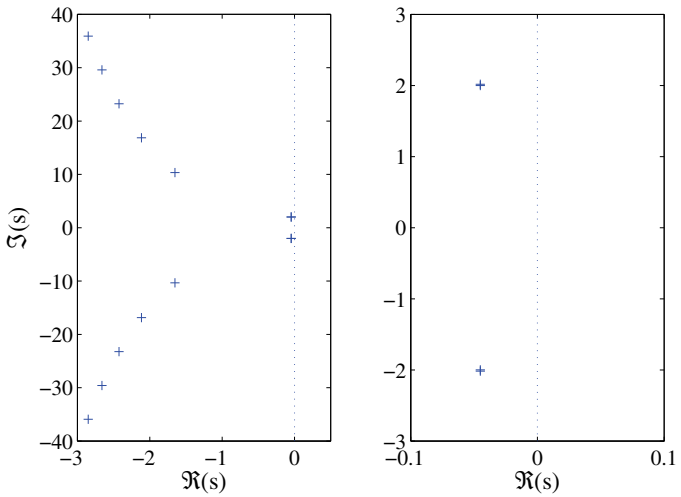


Figure 2. *Rightmost eigenvalues corresponding to the global minimum of* (2.2) *for* $(a_1, a_2, \tau) = (-5, 1.5, 1)$.

For a given value of τ, an exhaustive approach to compute the stabilizability region in the (a_1, a_2) plane would consist of applying the gradient sampling algorithm, illustrated in subsection 2.1, for a large number of values of (a_1, a_2) on a fine grid, to check whether the system is stabilizable. The stabilizability boundary then separates the regions where the system is stabilizable and where it is not. However, a more efficient calculation is possible by taking into account specific properties of the optimization problem of (2.2), as we now explain.

Recall that in the numerical example of subsection 2.1 the minimum is characterized by a pair of active eigenvalues with multiplicity two. This can be generalized due to the following property.

Proposition 2.1. *If* $a_2 \neq 0$ *and* $F(K; a_1, a_2, \tau)$ *in* (2.2) *is minimal, then there are at least three active eigenvalues (counting multiplicities).*

Proof. Because $a_2 \neq 0$, we have $p(0; a_1, a_2, \tau, K_1, K_2) \neq 0$ for all values of K.

When s_1 and $s_2 \in \mathbb{C}\backslash\{0\}$, $s_1 \neq s_2$, are required to be zeroes of p, this can always be achieved by solving K_1, K_2 from

$$\begin{pmatrix} s_1\left(1-\mathrm{e}^{-s_1\tau}\right) & \left(1-\mathrm{e}^{-s_1\tau}\right) \\ s_2\left(1-\mathrm{e}^{-s_2\tau}\right) & \left(1-\mathrm{e}^{-s_2\tau}\right) \end{pmatrix} \begin{pmatrix} K_1 \\ K_2 \end{pmatrix} = -\begin{pmatrix} s_1^2 + a_1 s_1 + a_2 \\ s_2^2 + a_1 s_2 + a_2 \end{pmatrix}. \tag{2.5}$$

Further, if we require $s_0 \in \mathbb{C}\backslash\{0\}$ to be a zero of p with multiplicity larger than one, i.e., $p(s_0) = p'(s_0) = 0$, we can achieve this by solving K_1, K_2 from

$$\begin{pmatrix} s_0\left(1-\mathrm{e}^{-s_0\tau}\right) & \left(1-\mathrm{e}^{-s_0\tau}\right) \\ 1-\mathrm{e}^{-s_0\tau}+s_0\tau\mathrm{e}^{-s_0\tau} & \tau\mathrm{e}^{-s_0\tau} \end{pmatrix} \begin{pmatrix} K_1 \\ K_2 \end{pmatrix} = -\begin{pmatrix} s_0^2 + a_1 s_0 + a_2 \\ 2s_0 + a_1 \end{pmatrix}. \tag{2.6}$$

Assume that K is such that $F(K)$ is minimal, and assume that there are less than three active eigenvalues. Let s_1, s_2 with $\Re(s_1) \leq \Re(s_2)$ be the two rightmost zeroes of p. Using

Table 1

According to Proposition 2.1, two configurations of the active eigenvalues are possible in the global minimum of (2.2). In the second column the mathematical relations, which characterize each configuration, are displayed.

Type	Active eigenvalues (multiplicity)	Determining system	Unknowns
I	$c \pm j\omega$ (2)	$\begin{cases} \Re\left(p(c+j\sigma)\right)=0 \\ \Im\left(p(c+j\sigma)\right)=0 \\ \Re\left(p'(c+j\sigma)\right)=0 \\ \Im\left(p'(c+j\sigma)\right)=0 \end{cases}$	c, σ, K_1, K_2
II	$c \pm j\sigma_1$ (1) $c \pm j\sigma_2$ (1)	$\begin{cases} \Re\left(p(c+j\sigma_1)\right)=0 \\ \Im\left(p(c+j\sigma_1)\right)=0 \\ \Re\left(p(c+j\sigma_2)\right)=0 \\ \Im\left(p(c+j\sigma_2)\right)=0 \end{cases}$	$c, \sigma_1, \sigma_2, K_1, K_2$

(2.5) or (2.6) it is then straightforwardly shown that for every $\epsilon > 0$ there exists a $K_i(\epsilon)$, $i = 1, 2$, such that $s_1 - \epsilon$ and $s_2 - \epsilon$ are roots of $p(s;\, a_1, a_2, \tau, K_1(\epsilon), K_2(\epsilon))$. Moreover, due to continuity we have that $\lim_{\epsilon \to 0} K_i(\epsilon) = K_i$, $i = 1, 2$, and that for ϵ small enough $s_2 - \epsilon$ is the rightmost zero of $p(s;\, a_1, a_2, \tau, K_1(\epsilon), K_2(\epsilon))$. This contradicts the assumption of a minimum of $F(K_1, K_2)$. ■

By Proposition 2.1 the possible configurations of the active eigenvalues in the global minimum of (2.2) can be reduced to the two types shown in Table 1, where we also show the mathematical relations characterizing the configuration of the active eigenvalues. In our numerical experiments only Type I occurred. Theoretically Type II is possible, but it is less generic than Type I. Indeed, for Type I the mathematical relations allow a direct computation of the minimal value $c(a_1, a_2, \tau)$ of (2.2) and the corresponding control parameters K_1 and K_2, provided that good starting values are available. For Type II, however, there is an extra parameter, and therefore the mathematical relations in Table 1 define curves in their unknowns, through the point corresponding to the optimal parameter values. In such a situation, a suitable small parameter change generically reduces the value of (2.2), meaning that the situation does not correspond to the global minimum unless an extra condition is satisfied (e.g., a turning point on the curve, or an additional active eigenvalue). Our numerical experiments indicate that this extra condition is characterized by $\sigma_2 = \sigma_1$ such that Type II reduces to Type I.

In order to compute the stabilizability region in the (a_1, a_2) plane for a particular delay value ($\tau = 1$ was chosen in our simulations), we used the information from Table 1 as follows. First, we freed the parameters (a_1, a_2) and imposed that $c(a_1, a_2, \tau) = 0$, corresponding to a minimum of F of Type I with active eigenvalues of multiplicity 2 at $\pm j\sigma$. In this way we arrived at the four relations

$$\begin{cases} \Re\left(p(j\sigma;\, a_1, a_2, \tau, K_1, K_2)\right) = 0, \\ \Im\left(p(j\sigma;\, a_1, a_2, \tau, K_1, K_2)\right) = 0, \\ \Re\left(p'(j\sigma;\, a_1, a_2, \tau, K_1, K_2)\right) = 0, \\ \Im\left(p'(j\sigma;\, a_1, a_2, \tau, K_1, K_2)\right) = 0 \end{cases} \tag{2.7}$$

in the five unknowns $(\sigma, K_1, K_2, a_1, a_2)$, which define a curve whose projection on the (a_1, a_2) plane forms the stabilizability boundary in regions where Type I minima occur. Next, we

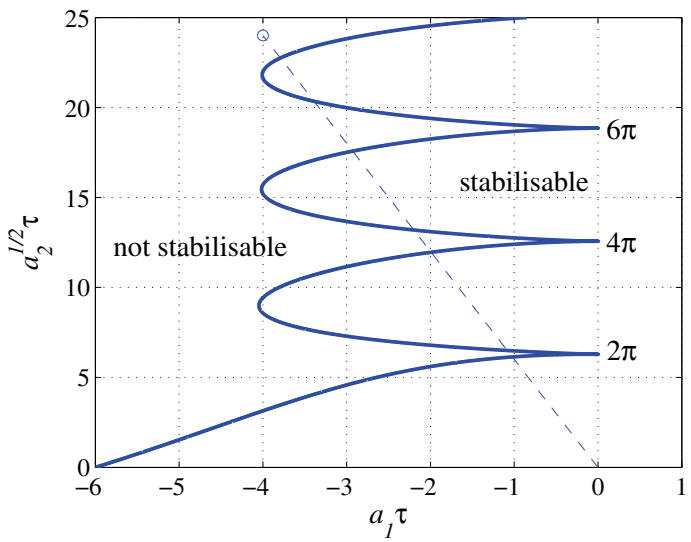


Figure 3. *Stabilizability region for time-delayed state feedback.*

performed a direct minimization of (2.2) for values of (a_1, a_2) chosen on a *coarse grid*, in order to exclude Type II minima. Finally, as Type II minima were not observed, the following result was concluded from (2.7).

Proposition 2.2. $c(a_1, a_2, \tau) = 0$ if and only if

$$\exists \omega \in \mathbb{R}^+ : \quad \begin{cases} a_1\tau &= \frac{2\omega(\cos\omega - 1)}{\omega - \sin\omega}, \\ a_2\tau^2 &= \frac{\omega^2(\omega - \sin\omega)}{\omega + \sin\omega}. \end{cases} \tag{2.8}$$

Moreover, the associated values of K_1 and K_2 are given by

$$\begin{cases} K_1\tau &= \frac{-2\omega(\omega\cos\omega - \sin\omega)}{\omega^2 - \sin^2\omega}, \\ K_2\tau^2 &= \frac{2\omega^3\sin\omega}{\omega^2 - \sin^2\omega}. \end{cases} \tag{2.9}$$

Proof. See Appendix A. ∎

According to Proposition 2.2, the stabilizability region in the (a_1, a_2) plane is depicted in Figure 3.

We conclude the section with some observations. The curve (2.9) asymptotically converges to a scaled cycloid as ω tends to infinity. For $a_1 = 0$ and $a_2 > 0$, which corresponds to an undamped oscillator, stabilization is not possible if

$$a_2\tau^2 = (2\pi\ell)^2, \quad \ell \in \mathbb{N}. \tag{2.10}$$

Physically, this can be interpreted as follows. Condition (2.10) implies that the delay is a multiple of the period of the solutions of the uncontrolled system, around which the control law (1.5) vanishes, and hence time-delayed feedback control cannot stabilize the system.

From Figure 3 we can deduce stabilizability information for a system with fixed parameters (a_1, a_2) as a function of the delay τ. When the delay changes, the normalized plant parameters $(a_1\tau, \tau\sqrt{a_2})$ move on a (half) straight line, as indicated in the figure. Note that several disjunct delay intervals may exist for which the system is stabilizable. If the delay can also be used as a controller parameter, then the system is stabilizable if and only if $a_2 > 0$; i.e., it does not satisfy the odd number condition. Furthermore, stability can always be achieved for sufficiently small values of the delay.

3. Stabilization by means of time-delayed output feedback. In this section we consider the system (1.2) and consider the question of whether or not it can be stabilized with a time-delayed output feedback of the form (1.6). We will first consider the stabilizability for a fixed value of the delay parameter, after which we consider the case where both the gain and delay are considered as controller parameters. We end with a comparison between the time-delayed output feedback approach and the derivative output feedback approach.

Note that now the characteristic equation of the closed loop system is given by

$$0 = p(s;\, a_1, a_2, c_1, c_2, \tau, k) =: s^2 + \left(a_1 + c_1 k(1 - \mathrm{e}^{-s\tau})\right) s + \left(a_2 + c_2 k(1 - \mathrm{e}^{-s\tau})\right). \tag{3.1}$$

Analogously to (2.4) we now have the following scaling property:

$$p(s; a_1, a_2, c_1, c_2, \tau, k) = 0 \Leftrightarrow p\left(\hat{s}; a_1\tau, a_2\tau^2, 1, \frac{c_2}{c_1}\tau, 1, c_1 k\tau\right) = 0, \quad \hat{s} = s\tau. \tag{3.2}$$

As a consequence, it is again sufficient to compute the stabilizability region for one arbitrary nonzero delay value.

3.1. Stabilizability for a fixed delay value. In this section we consider the case that $\tau > 0$ is fixed. We will restrict ourselves to the cases where $c_2 = 0$, $c_1 = -1$ and where $c_1 c_2 < 0$. The cases $c_1 c_2 > 0$ and $c_1 = 0$, $c_2 \neq 0$ are omitted because they can be treated in a completely analogous way, and unlike the other cases, the results are not essential for the analysis of section 3.2.

As there now is only one free parameter (viz. the control parameter k), it may be shown in a similar way as in Proposition 2.1 that if $F(k;\, a_1, a_2, \tau)$ is minimal, there are at least *two* active eigenvalues.

3.1.1. The case that $c_2 = 0$, $c_1 = -1$. In this case, the stabilizability region is the region to the right of the thick black (dashed) and red (solid) curves in Figure 4. In this figure, the red (solid) curve is characterized by the existence of $\sigma \in \mathbb{R}^+$, $\hat{k} \in \mathbb{R}$ such that

$$p(j\sigma;\, a_1, a_2, \hat{k}) = 0, \qquad \frac{d\Re(\lambda_k)}{dk}(\hat{k}) = 0, \tag{3.3}$$

where λ_k satisfies $p(\lambda_k;\, a_1, a_2, k) = 0$, $\lambda_{\hat{k}} = j\sigma$. This represents the case that an unstable eigenvalue in the right-half plane becomes critically stable but moves back into the right-half plane when k is changed. As shown in Appendix B, this gives rise to the following equalities

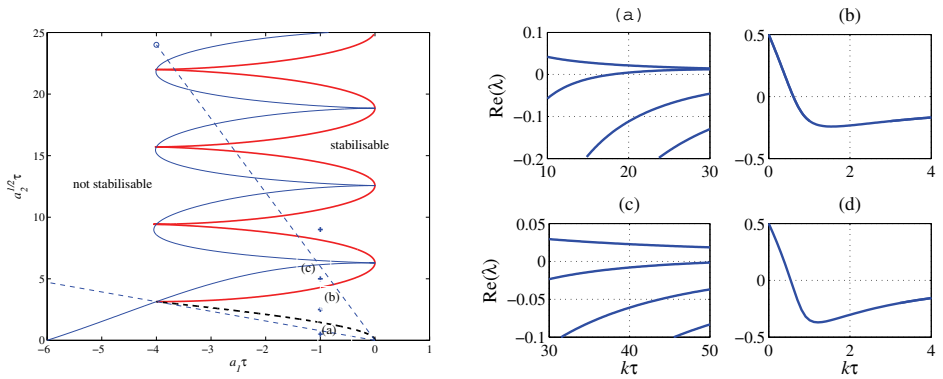


Figure 4. *Left-hand panel: stabilizability boundary for time-delayed output feedback with $c_2 = 0$, $c_1 = -1$. Thick red (solid) and black (dashed) curves: stabilizability boundary. Thin blue (solid) curve: stabilizability boundary for state feedback. Crosses: points at which stabilizability has been checked in right-hand panel. Right-hand panel: Real parts of rightmost eigenvalues as function of $k\tau$.* (a) $(a_1\tau, \tau\sqrt{a_2}) = (-1, 0.5)$, *not stabilizable;* (b) $(a_1\tau, \tau\sqrt{a_2}) = (-1, 2.5)$, *stabilizable;* (c) $(a_1\tau, \tau\sqrt{a_2}) = (-1, 5)$, *not stabilizable;* (d) $(a_1\tau, \tau\sqrt{a_2}) = (-1, 9)$, *stabilizable.*

describing the red curve:

$$\left\{\begin{array}{rcl} a_1\tau &=& \frac{2\omega(\cos\omega-1)}{\omega-\sin\omega}, \\ a_2\tau^2 &=& \frac{\omega^2(\omega+\sin\omega)}{\omega-\sin\omega}, \\ k\tau &=& \frac{2\omega}{(\sin\omega-\omega)}, \\ \omega &=& \sigma\tau. \end{array}\right. \tag{3.4}$$

The black (dashed) curves are characterized by the existence of $\sigma_1, \sigma_2 \in \mathbb{R}^+$ and k such that one pair of roots of $p_k(s)$ crosses to the right-half plane through $\pm j\sigma_1$, while another pair of roots of $p_k(s)$ crosses to the left-half plane through $\pm j\sigma_2$ (or vice versa). As shown in Appendix B, this gives rise to the following equalities:

$$\left\{\begin{array}{l} a_1\tau = k\tau(1-\cos\omega_1), \\ a_2\tau^2 = \omega_1^2 - k\tau\omega_1\sin\omega_1, \\ k\tau = \frac{\omega_2^2-\omega_1^2}{\omega_2\sin\omega_2-\omega_1\sin\omega_1}, \\ \omega_i = \sigma_i\tau \ (i=1,2), \end{array}\right. \tag{3.5}$$

where

$$\cos\omega_1 = \cos\omega_2. \tag{3.6}$$

Given $\omega_1 > 0$, it is obviously the case that there exist multiple $\omega_2 > 0$ such that (3.6) holds. However, from numerical experiments it turned out that the curves described by (3.5), (3.6) are active only for $\omega_1 = \omega \in [0, \pi]$, $\omega_2 = 2\pi - \omega$. This gives that in fact the black (dashed)

curve in Figure 4 is given by

$$\left\{\begin{array}{rcll} a_1\tau &=& \frac{2(\pi-\omega)(\cos\omega-1)}{\sin\omega}, & \\ a_2\tau^2 &=& \omega(2\pi-\omega). & (0\le\omega\le\pi) \end{array}\right. \tag{3.7}$$

This is illustrated by the numerical experiments given in Figure 4.

As before, we can deduce stability information for a system with fixed parameters (a_1, a_2) as a function of the delay τ. As illustrated in Figure 4 by means of the upper thin dashed line, the normalized plant parameters $(a_1\tau, \tau\sqrt{a_2})$ will again move on a (half) straight line. We see from the figure that again several disjunct delay intervals may exist for which the system is stabilizable. Furthermore, as it may be shown that the black (dashed) curve has an infinite slope at the origin, we have, unlike the static state feedback case, that the system is never stabilizable for small enough values of τ. Finally, the lower thin dashed line in Figure 4 illustrates the construction of the stabilizability boundary if k as well as τ are control parameters. Namely, we see that in this case (a_1, a_2) is on the stabilizability boundary if and only if there exists a $\tau > 0$ such that $(a_1\tau, \tau\sqrt{a_2}) = (-4, \pi)$. This then gives that there exist $k \in \mathbb{R}$, $\tau > 0$ such that the closed loop system is stable if and only if

$$\sqrt{a_2} > -\frac{\pi a_1}{4}. \tag{3.8}$$

3.1.2. The case that $c_1c_2 < 0$. Recall that in in this case the characteristic equation of the closed loop system is given by (3.1). Define

$$\widetilde{c} = \frac{c_2}{c_1}\tau. \tag{3.9}$$

As for the case treated in the previous subsection, it will turn out that the stabilizability boundary is formed by points characterized by one of the following two properties:

1. There exist $\sigma \in \mathbb{R}^+$, $\hat{k} \in \mathbb{R}$ such that (3.3) holds. Similarly to the derivation of (3.4), it may be shown that this gives rise to the following equalities describing this curve:

$$\left\{\begin{array}{rcl} a_1\tau &=& \frac{2\omega(-\widetilde{c}^2\cos\omega+\omega^2\cos\omega-\omega^2-2\widetilde{c}\omega\sin\omega-\widetilde{c}^2)}{-2\widetilde{c}\omega\cos\omega+\widetilde{c}^2\omega+\omega^3-\omega^2\sin\omega+\widetilde{c}^2\sin\omega-2\widetilde{c}\omega}, \\ a_2\tau^2 &=& \frac{\omega^2(2\widetilde{c}\omega\cos\omega+\omega^3+\widetilde{c}^2\omega+\omega^2\sin\omega-\widetilde{c}^2\sin\omega-2\widetilde{c}\omega)}{-2\widetilde{c}\omega\cos\omega+\widetilde{c}^2\omega+\omega^3-\omega^2\sin\omega+\widetilde{c}^2\sin\omega-2\widetilde{c}\omega}, \\ c_1k\tau &=& \frac{2\omega^2(\omega\cos\omega-\omega-\widetilde{c}\sin\omega)}{(\cos\omega-1)(-2\widetilde{c}\omega\cos\omega+\widetilde{c}^2\omega+\omega^3-\omega^2\sin\omega+\widetilde{c}\tau^2\sin\omega-2\widetilde{c}\omega)}, \\ \omega &=& \sigma\tau. \end{array}\right. \tag{3.10}$$

2. There exist $\sigma_1, \sigma_2 \in \mathbb{R}^+$, $k \in \mathbb{R}$ such that one pair of roots of $p_k(s)$ crosses to the right-half plane through $j\omega_1$, while another pair of roots of $p_k(s)$ crosses to the left-half plane through $j\omega_2$ (or vice versa). Similarly to the derivation of (3.5), (3.6), it may be shown that this gives rise to the following equalities:

$$\left\{\begin{array}{rcl} a_1\tau &=& -\frac{k\tau}{\omega_1}(\omega_1-\omega_1\cos\omega_1+\widetilde{c}\sin\omega_1), \\ a_2\tau^2 &=& \omega_1(\omega_1+k\sin\omega_1)-\widetilde{c}k\tau(1-\cos\omega_1), \\ c_1k\tau &=& \frac{\omega_2^2-\omega_1^2}{\omega_1\sin\omega_1-\omega_2\sin\omega_2+\widetilde{c}(\cos\omega_1-\cos\omega_2)}, \\ \omega_i &=& \sigma_i\tau \quad (i=1,2), \end{array}\right. \tag{3.11}$$

where

$$-\cos\omega_1 + \frac{\widetilde{c}}{\omega_1}\sin\omega_1 = -\cos\omega_2 + \frac{\widetilde{c}}{\omega_2}\sin\omega_2. \tag{3.12}$$

As in the previous subsection, in principle for each $\omega_1 > 0$ there exist multiple $\omega_2 > 0$ such that (3.12) is satisfied. Again, it turns out that it suffices to consider only the smallest $\omega_2 > \omega_1$ satisfying (3.12).

As for intersections of the two types of curves in the (a_1, a_2) plane, numerical experiments have indicated that at these intersections the values of k associated with both curves are identical. An intersection can then occur if first, for ω associated with point (a_1, a_2) on the first curve, we have that (a_1, a_2) is on the second curve with $\omega_1 = \omega_2 = \omega$. It may be shown that this occurs when both curves cross the static state feedback stability boundary given by (2.8). Second, an intersection may occur when we have the situation where cases 1 and 2 hold at the same time with $\omega_1 \neq \omega_2$ and $\omega_1, \omega_2 \neq \omega$.

In Figure 5 a comparison is given of stabilizability regions for different values of the parameter $\frac{c_2}{c_1}\tau$. In terms of the location of the black (dashed) curves representing case 2, the first four plots are qualitatively different from the last two plots. This difference is further illustrated by means of Figures 6 and 7 for $\frac{c_2}{c_1}\tau = -0.01$ and $\frac{c_2}{c_1}\tau = -10$, respectively.

In Figure 8 a comparison is given between the stabilizability boundaries when $(c_1, c_2\tau) = (1, 0)$ and $\frac{c_2}{c_1}\tau = -0.01$. This clearly illustrates that as $c_2\tau \to 0$, the blue (dotted) and black (dashed) curves approach each other pointwise, but not uniformly.

3.2. Stabilizability using the delay and gain parameter. We will now consider the stabilization problem when the delay and gain parameter can be varied simultaneously for the case that $c_2 \neq 0$. Note that the case that $c_2 = 0$ has already been considered in subsection 3.1.1.

We will first consider the case that $c_1 c_2 < 0$. Recall that for a controller of the form (1.6), the closed-loop characteristic equation is given by (3.1). By defining

$$\sigma = s\tau, \quad \kappa = c_1 k\tau, \quad \gamma = \left|\frac{c_2}{c_1}\right|\tau, \quad \alpha_1 = a_1\gamma\left|\frac{c_1}{c_2}\right|, \quad \alpha_2 = a_2\gamma^2\left(\frac{c_1}{c_2}\right)^2, \tag{3.13}$$

it is straightforwardly seen that (3.1) may be rewritten as

$$0 = \pi(\sigma;\, \alpha_1, \alpha_2, \gamma, \kappa) := \sigma^2 + \left(\alpha_1\gamma + \kappa(1 - \mathrm{e}^{-\sigma})\right)\sigma + \left(\alpha_2\gamma^2 - \gamma\kappa(1 - \mathrm{e}^{-\sigma})\right). \tag{3.14}$$

As a consequence, the stabilizability region in the $\left(a_1\left|\frac{c_1}{c_2}\right|, a_2\left(\frac{c_1}{c_2}\right)^2\right)$ plane can be obtained by using the previous section to obtain the stabilizability boundary in the (α_1, α_2) plane for a system with unit delay and output $z_1 - \gamma z_2$, then plotting this boundary in the $\left(\frac{\alpha_1}{\gamma}, \frac{\alpha_2}{\gamma^2}\right) = \left(a_1\left|\frac{c_1}{c_2}\right|, a_2\left(\frac{c_1}{c_2}\right)^2\right)$ plane, and then taking a union of all stabilizability regions for $\gamma < 0$. In Figure 9 this process is illustrated. From Figure 9(a), one sees that for $a_1\left|\frac{c_1}{c_2}\right| < -1$ the stabilizability boundary is given by the thin blue (solid) curve, which consists of points where the stabilizability boundaries for output feedback cross the stabilizability boundaries for state

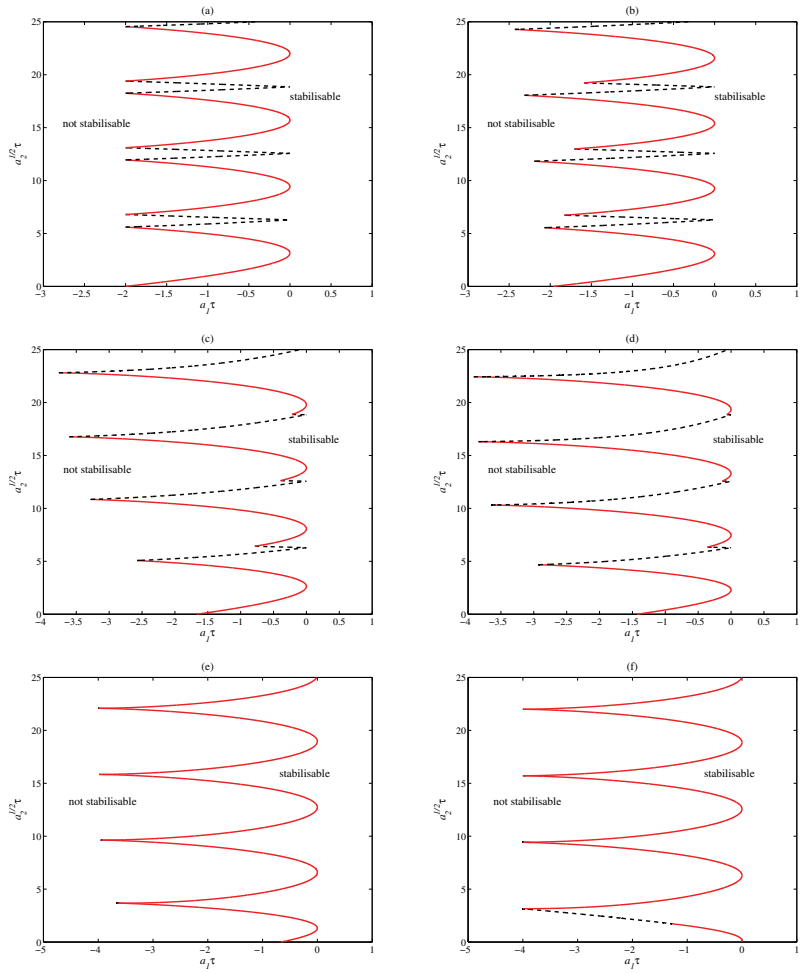


Figure 5. *Stabilizability regions for time-delayed output feedback. Red (solid) curves: case* 1*; black (dashed) curves: case* 2*.* (a) $c_1c_2 \to -\infty$ *(note: this is equivalent to* $c_1 = 0$*,* $c_2 \neq 0$*);* (b) $\frac{c_2}{c_1}\tau = -100$*;* (c) $\frac{c_2}{c_1}\tau = -10$*;* (d) $\frac{c_2}{c_1}\tau = -5$*;* (e) $\frac{c_2}{c_1}\tau = -1$*;* (f) $\frac{c_2}{c_1}\tau = -0.01$.

feedback. It may be shown that this curve is given by

$$\text{(3.15)} \qquad \begin{cases} a_1 \left|\frac{c_1}{c_2}\right| &= 2\frac{(\cos\omega - 1)(\omega\cos\omega - \sin\omega)}{\omega(\omega - \sin\omega)\sin\omega}, \\ a_2 \left(\frac{c_1}{c_2}\right)^2 &= \frac{(\omega - \sin\omega)(\omega\cos\omega - \sin\omega)^2}{\omega^2(\omega + \sin\omega)\sin^2\omega}, \end{cases} \qquad \omega \in (\pi, \zeta),$$

where ζ is the first strictly positive solution of $\tan x = x$. From Figure 9(b), one sees that the whole strip $a_1\left|\frac{c_1}{c_2}\right| > -1$ belongs to the stabilizability region. Thus, one obtains the stabilizability region illustrated in Figure 10.

For the cases that $c_1c_2 > 0$ or $c_1 = 0$, $c_2 \neq 0$, an application of Propositions 2.7 and 3.1 from [15] leads to the following result.

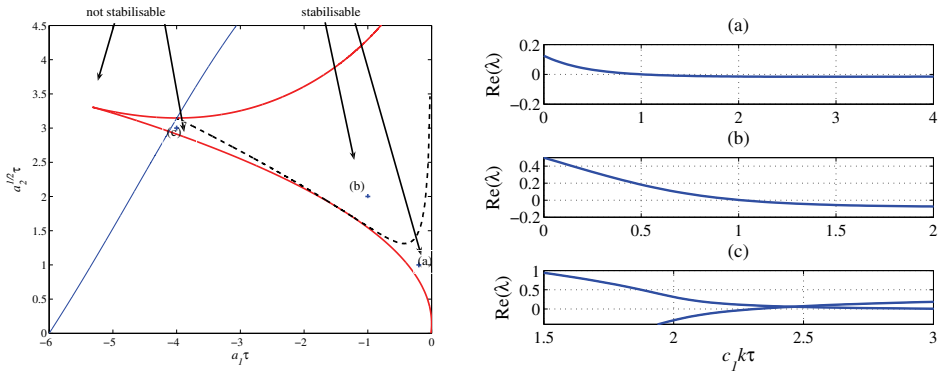


Figure 6. *Left-hand panel: stabilizability boundary for time-delayed output feedback with* $\frac{c_2}{c_1}\tau = -0.01$. *Thin blue (solid) curve: stabilizability boundary for time-delayed state feedback; thick red (solid) curve: case* 1*; thick black (dashed) curve: case* 2. *Crosses: points at which stabilizability has been checked in right-hand panel. Right-hand panel: Real parts of rightmost eigenvalues as functions of* $c_1 k\tau$. (a) $(a_1\tau, \tau\sqrt{a_2}) = (-0.25, 1)$, *stabilizable;* (b) $(a_1\tau, \tau\sqrt{a_2}) = (-1, 2)$, *stabilizable;* (c) $(a_1\tau, \tau\sqrt{a_2}) = (-4, 3)$, *not stabilizable.*

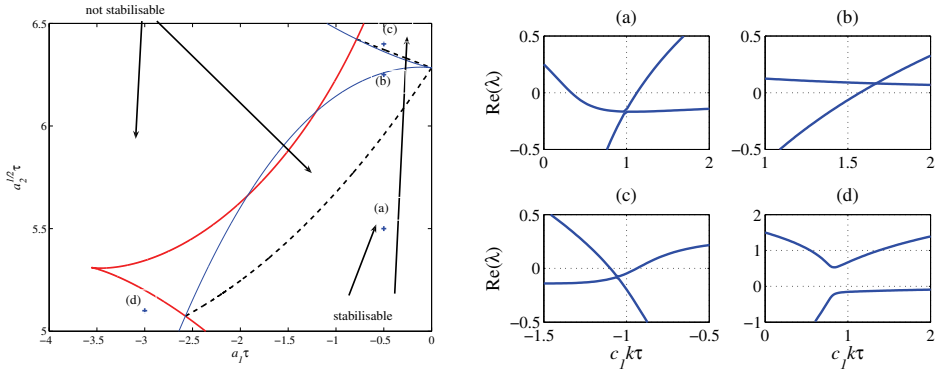


Figure 7. *Left-hand panel: stabilizability boundary for time-delayed output feedback with* $\frac{c_2}{c_1}\tau = -10$. *Thin blue (solid) curve: stabilizability boundary for time-delayed state feedback; thick red (solid) curve: case* 1*; thick black (dashed) curve: case* 2. *Crosses: points at which stabilizability has been checked in right-hand panel. Right-hand panel: Real parts of rightmost eigenvalues as functions of* $c_1 k\tau$. (a) $(a_1\tau, \tau\sqrt{a_2}) = (-0.5, 5.5)$, *stabilizable;* (b) $(a_1\tau, \tau\sqrt{a_2}) = (-0.5, 6.25)$, *not stabilizable;* (c) $(a_1\tau, \tau\sqrt{a_2}) = (-0.5, 6.4)$, *stabilizable;* (d) $(a_1\tau, \tau\sqrt{a_2}) = (-3, 5.1)$, *not stabilizable.*

Lemma 3.1. *If the system* (1.2) *with* $a_2 > 0$ *can be stabilized with a derivative output feedback*

$$u(t) = -k\dot{y}(t), \tag{3.16}$$

where $y(t)$ *is defined in* (1.2) *and* $1 + kc_1 > 0$, *then it can be stabilized with a time-delayed output feedback* (1.6).

Using this result, we then obtain the following.

Theorem 3.2. *If* $c_1c_2 > 0$ *or* $c_1 = 0$, $c_2 \neq 0$, *there exists a stabilizing time-delayed output feedback* (1.6) *for* (1.2) *if and only if* $a_2 > 0$.

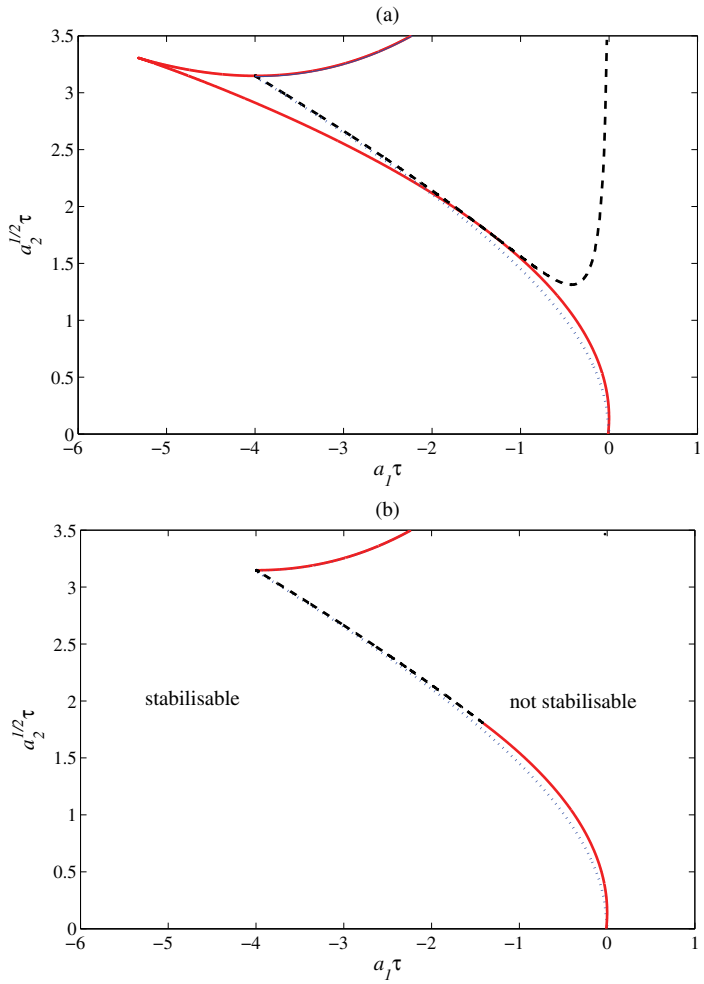


Figure 8. *Comparison of stabilizability regions. Black (dashed) and red (solid) curves define the stabilizability region for* $\frac{c_2}{c_1}\tau = -0.01$ *with resulting active stabilizability boundary illustrated in* (b)*; blue (dotted) curve defines the stabilizability region for* $c_1 = -1$, $c_2 = 0$.

Proof. The necessity follows from the odd-number limitation. For sufficiency, we will show that for $a_2 > 0$ there always exists a stabilizing derivative output feedback (3.16) with $1 + kc_1 > 0$, which by Lemma 3.1 then implies the existence of a stabilizing time-delayed output feedback.

Applying the derivative output feedback (3.16) to (1.2) gives

$$\left\{ \begin{array}{rcl} \dot{z}_1 & = & z_2, \\ \dot{z}_2 & = & -a_2 z_1 - a_1 z_2 - k(c_2 \dot{z}_1 - k c_1 \dot{z}_2), \end{array} \right. \tag{3.17}$$

which for the generic case that $1 + kc_1 \neq 0$ can be rewritten as

$$\begin{pmatrix} \dot{z}_1 \\ \dot{z}_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -\frac{a_2}{1+kc_1} & -\frac{a_1+kc_2}{1+kc_1} \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}. \tag{3.18}$$

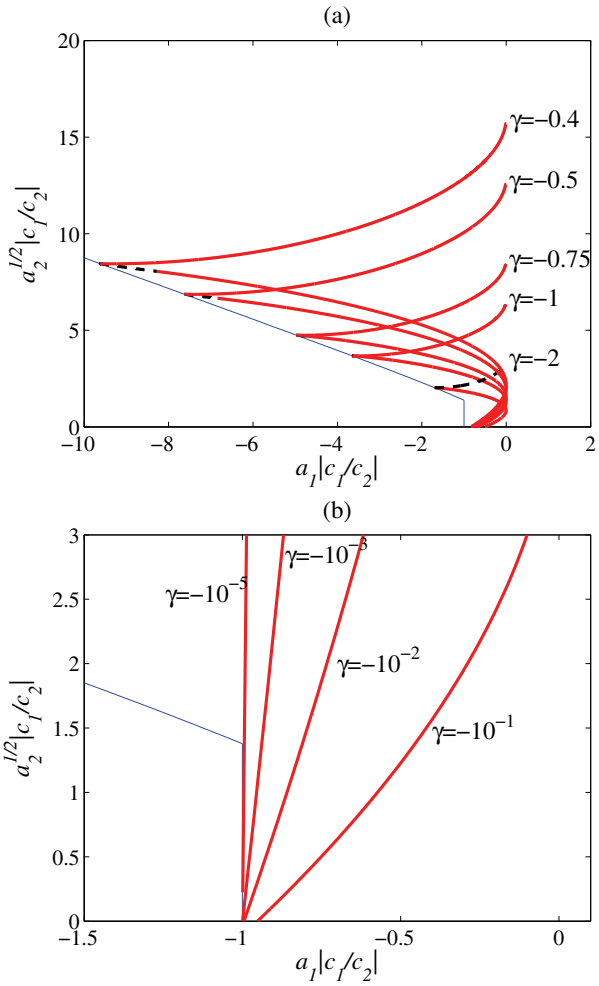


Figure 9. *Construction of stabilizability region using the gain and delay parameter for* $c_1c_2 < 0$.

It is now straightforward to check that if $a_2 > 0$ and $c_1c_2 > 0$ or $c_1 = 0$, $c_2 \neq 0$, there always exists a $k \in \mathbb{R}$ such that

$$\frac{a_2}{1+kc_1}, \frac{a_1+kc_2}{1+kc_1} > 0, \tag{3.19}$$

which implies that there always exists a $k \in \mathbb{R}$ such that the closed loop system (3.17) is stable. Moreover, note that when (3.19) holds, we necessarily have that $1 + kc_1 > 0$. This then establishes our claim. ■

3.3. Comparison with derivative output feedback. In [15] the stabilizability with control laws involving state derivative feedback is addressed, motivated by potential applications in vibration control, where the output of accelerometers is used directly for feedback. Given the fact that for small delays time-delayed feedback can be interpreted as an approximation of

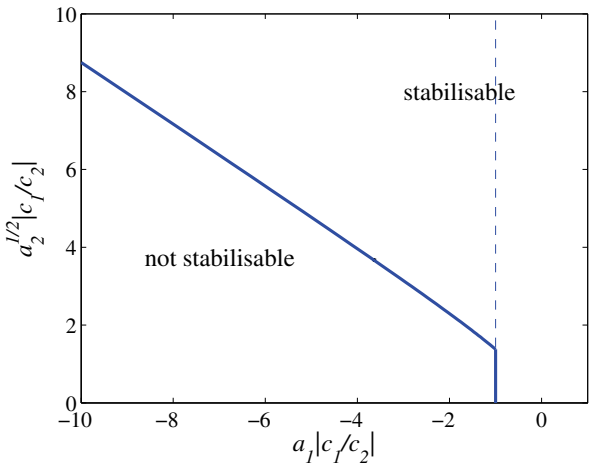


Figure 10. *Stabilizability region using the gain and delay parameter for* $c_1c_2 < 0$.

derivative feedback, following from

$$\dot{x}(t) \approx \frac{x(t) - x(t-\tau)}{\tau},$$

it seems appropriate to make a comparison between the stabilizability regions for time-delayed output feedback (with k and τ as parameters) and the stabilizability regions using the control law (3.16). We will do this here for the different cases also encountered in this section.

- $c_1c_2 > 0$.
  In this case, stabilization by means of derivative feedback is always possible, while due to the odd-number limitation stabilization by means of time-delayed feedback is possible if and only if $a_2 > 0$. At first sight the conclusion regarding the case $a_2 \leq 0$ might look counterintuitive, because it could be argued that derivative feedback is just time-delayed feedback with a very small time-delay. However, it may be shown that the approximation by time-delayed feedback introduces an unstable half plane pole whose real part moves off to infinity as the delay goes to zero. In fact, for $a_2 \leq 0$ the stabilization by means of derivative feedback is *fragile* (see [15]), in that small implementation or modeling errors will lead to an unstable closed loop, even though the feedback is stabilizing for the nominal system.
- $c_1c_2 < 0$.
  In this case, the system is stabilizable by means of derivative output feedback if and only if $\left(a_1 \left|\frac{c_1}{c_2}\right| + 1\right) a_2 \geq 0$. As above, for the part of this region where $a_2 \leq 0$, the stabilizability is fragile. For time-delayed feedback, stabilization is possible if $a_2 > 0$ and stabilization by means of derivative feedback is possible, but there is also still a region where $\left(a_1 \left|\frac{c_1}{c_2}\right| + 1\right) < 0$ and where stabilization is possible; see subsection 3.2.
- $c_1 = 0$, $c_2 \neq 0$.
  In this case stabilization by means of derivative output feedback as well as time-delayed output feedback is possible if and only if $a_2 > 0$.

- $c_1 \neq 0$, $c_2 = 0$.
  Here stabilization by means of derivative output feedback is possible if and only if $a_1 a_2 > 0$, where for $a_2 < 0$ the stabilizability is fragile again. For time-delayed output feedback stabilization is possible if $a_1, a_2 > 0$ and if $a_1 < 0$, $a_2 > \frac{\pi^2 a_1^2}{16}$; see subsection 3.1.1.

4. Conclusions. In this paper we have explored the limits of the stabilization approach via linear time-delayed state and output feedback. We have made a parametrization of stabilizable systems. This has yielded many insights into the mechanism of time-delayed output feedback which could not have been determined by means of conservative analysis methods that yield sufficient but not necessary stability or stabilizability conditions. However, as has become apparent from the shapes of the stabilizability regions, the parametrization problem is a very difficult problem even for the second-order systems that have been discussed in this paper. We have solved the problem by means of a combination of analytical results (see, e.g., Proposition 2.1) and recently developed numerical tools like rightmost eigenvalue computation and nonsmooth optimization. The advantage of our method over a purely numerical approach is that *explicit* expressions of stabilizability boundaries have been obtained based on numerics performed on a *coarse* grid.

It should be noted that even though our paper was inspired by the references [10], [22], we have not explicitly solved the control problems stated in these references. The main reason for this is that from a purely practical control point of view the problem formulations in our paper are more relevant. The results presented in this paper have direct application in controlling certain systems using delayed output feedback, which can be interpreted as either an approximation of derivative feedback or as a feedback that weighs the difference of an output signal over a time-interval of length τ. However, the methods applied in the present paper can, mutatis mutandis, be straightforwardly applied to the control problems in [10], [22].

Appendix A. Proof of Proposition 2.2. Recall from (2.1) that

$$p(s;\, a_1, a_2, \tau, K_1, K_2) = s^2 + \left(a_1 + K_1(1 - \mathrm{e}^{-s\tau})\right) s + \left(a_2 + K_2(1 - \mathrm{e}^{-s\tau})\right), \tag{A.1}$$

which gives that

$$p'(s;\, a_1, a_2, \tau, K_1, K_2) = \left(2 + K_1 \tau s \mathrm{e}^{-s\tau}\right) s + \left(a_1 + K_1(1 - \mathrm{e}^{-s\tau}) + K_2 \tau \mathrm{e}^{-s\tau}\right). \tag{A.2}$$

Defining $\omega = \sigma\tau$, it then follows from (2.7), (A.1), (A.2) that

$$\begin{cases} a_2\tau^2 - \omega^2 - K_1\tau\omega\sin\omega + K_2\tau^2(1 - \cos\omega) = 0, \\ a_1\tau\omega + K_1\tau\omega(1 - \cos\omega) + K - 2\tau^2\sin\omega = 0, \\ a_1\tau + K_1\tau(\omega\sin\omega + 1 - \cos\omega) + K_2\tau^2\cos\omega = 0, \\ 2\omega + K_1\tau(\omega\cos\omega + \sin\omega) - K_2\tau^2\sin\omega = 0. \end{cases} \tag{A.3}$$

Solving (A.3) for $a_1\tau, a_2\tau^2, K_1\tau, K_2\tau^2$ then establishes our claim.

Appendix B. Derivation of (3.4) and (3.5). It follows from (3.1) that for $c_2 = 0$, $c_1 = -1$ we have that

$$p(s;\, a_1, a_2, \tau, k) = s^2 + (a_1 - k(1 - \mathrm{e}^{-s\tau}))s + a_2. \tag{B.1}$$

Let $\lambda_k = x_k + jy_k$ be such that

$$\begin{aligned} 0 = F(x_k, y_k) =: p(\lambda_k;\, a_1, a_2, \tau) = &\left[x_k^2 - y_k^2 + a_1 x_k - k x_k + k\mathrm{e}^{-\tau x_k}(y_k \sin \tau y_k + x_k \cos \tau y_k)\right] \\ &+ j\left[2x_k y_k + a_1 y_k - k y_k - k\mathrm{e}^{-\tau x_k}(x_k \sin \tau y_k - y_k \cos \tau y_k)\right]. \end{aligned} \tag{B.2}$$

For $k = \hat{k}$, $x_{\hat{k}} = 0$, (B.2) then gives the equalities

$$\begin{cases} -y_{\hat{k}}^2 + a_2 + \hat{k} y_{\hat{k}} \sin \tau y_{\hat{k}} = 0, \\ a_1 y_{\hat{k}} - \hat{k} y_{\hat{k}} + \hat{k} y_{\hat{k}} \cos \tau y_{\hat{k}} = 0. \end{cases} \tag{B.3}$$

Next, denoting $x'_k := \frac{dx_k}{dk}$, $y'_k := \frac{dy_k}{dk}$; implicitly differentiating (B.2) with respect to k; and substituting $k = \hat{k}$, $x_{\hat{k}} = 0$, $x'_{\hat{k}} = 0$ gives

$$\begin{cases} y_{\hat{k}} \sin \tau y_{\hat{k}} + (2y_{\hat{k}} - \hat{k} \sin \tau y_{\hat{k}} - \hat{k}\tau y_{\hat{k}} \cos \tau y_{\hat{k}}) y'_{\hat{k}} = 0, \\ y_{\hat{k}} - y_{\hat{k}} \cos \tau y_{\hat{k}} + (a1 - \hat{k} + \hat{k} \cos \tau y_{\hat{k}} - \hat{k}\tau y_{\hat{k}} \sin \tau y_{\hat{k}}) y'_{\hat{k}} = 0. \end{cases} \tag{B.4}$$

Elimination of $y'_{\hat{k}}$ from (B.4) then gives

$$-a_1 \sin \tau y_{\hat{k}} + \hat{k}\tau y_{\hat{k}} + 2y_{\hat{k}} - 2y_{\hat{k}} \cos \tau y_{\hat{k}} - \hat{k}\tau y_{\hat{k}} \cos \tau y_{\hat{k}} = 0. \tag{B.5}$$

Defining $\omega = \tau y_{\hat{k}}$, (3.4) is then straightforwardly derived from (B.3) and (B.5).

For the derivation of (3.5), it follows that we need to have that

$$\begin{aligned} 0 &= p(j\sigma_i;\, a_1, a_2, \tau) = -\sigma_i^2 + (a_1 - k(1 - \mathrm{e}^{-j\sigma_i\tau}))j\sigma_i + a_2 \\ &= (-\sigma_i^2 + k\sigma_i \sin \sigma_i\tau + a_2) + j(\sigma_i a_1 - k\sigma_i + \sigma_i k \cos \sigma_i\tau) \quad (i = 1, 2). \end{aligned} \tag{B.6}$$

Defining $\omega_i = \sigma_i \tau$, it then follows from (B.6) that we have that

$$\begin{cases} a_1\tau &= k\tau(1 - \cos \omega_1), \\ a_2\tau^2 &= \omega_1^2 - k\tau\omega_1 \sin \omega_1, \\ 0 &= (\omega_1^2 - k\tau\omega_1 \sin \omega_1) - (\omega_2^2 - k\tau\omega_2 \sin \omega_2), \\ 0 &= k\tau(1 - \cos \omega_1) - k\tau(1 - \cos \omega_2), \\ \omega_i &= \sigma_i\tau \quad (i = 1, 2). \end{cases} \tag{B.7}$$

The first three equalities then establish (3.5), while the third equality establishes (3.6).

REFERENCES

[1] J. V. Burke, D. Henrion, A. S. Lewis, and M. L. Overton, *HIFOO—A* Matlab *package for fixed-order controller design and H-infinity optimization*, in Proceedings of ROCOND 2006, Toulouse, France, 2006 (CD).

[2] J. V. Burke, D. Henrion, A. S. Lewis, and M. L. Overton, *Stabilization via nonsmooth, nonconvex optimization*, IEEE Trans. Automat. Control, 51 (2006), pp. 1760–1769.

[3] J. V. Burke, A. S. Lewis, and M. L. Overton, *A robust gradient sampling algorithm for nonsmooth, nonconvex optimization*, SIAM J. Optim., 15 (2005), pp. 751–779.

[4] C.-T. Chen, *Linear System Theory and Design*, Oxford University Press, New York, 1999.

[5] S. J. Dyke, B. F. Spencer, P. Quast, M. K. Sain, D. C. Kaspari, and T. T. Soong, *Acceleration feedback control of MDOF structures*, J. Engineering Mech., 122 (1996), pp. 907–917.

[6] K. Engelborghs, T. Luzyanina, and G. Samaey, *DDE-BIFTOOL v.* 2.00: *A* matlab *Package for Bifurcation Analysis of Delay Differential Equations*, TW Report 330, Department of Computer Science, Katholieke Universiteit Leuven, Belgium, 2001.

[7] B. Fiedler, V. Flunkert, M. Georgi, P. Hövel, and E. Schöll, *Refuting the odd-number limitation of time-delayed feedback control*, Phys. Rev. Lett., 98 (2007), paper 114101.

[8] J. K. Hale and S. M. Verduyn Lunel, *Introduction to Functional Differential Equations*, Springer-Verlag, New York, 1993.

[9] M. A. Hassouneh, H.-C. Lee, and E. H. Abed, *Washout filters in feedback control: Benefits, limitations and extensions*, in Proceedings of the 2004 American Control Conference, Boston, MA, 2004, pp. 3950–3955.

[10] P. Hövel and E. Schöll, *Control of unstable steady states by time-delayed feedback methods*, Phys. Rev. E, 72 (2005), paper 046203.

[11] W. Just, T. Bernard, M. Ostheimer, E. Reibold, and H. Benner, *Mechanism of time-delayed feedback control*, Phys. Rev. Lett., 78 (1997), pp. 203–206.

[12] W. Just, B. Fiedler, M. Georgi, V. Flunkert, P. Hövel, and E. Schöll, *Beyond the odd number limitation: A bifurcation analysis of time-delayed feedback control*, Phys. Rev. E, 76 (2007), paper 026210.

[13] W. Michiels and S.-I. Niculescu, *Stability and Stabilization of Time-Delay Dystems: An Eigenvalue-Based Approach*, Adv. Des. Control 12, SIAM, Philadelphia, 2007.

[14] W. Michiels and D. Roose, *Limitations of delayed state feedback: A numerical study*, Int. J. Bifur. Chaos, 12 (2002), pp. 1309–1320.

[15] W. Michiels, T. Vyhlídal, H. Huijberts, and H. Nijmeijer, *Stabilizability and stability robustness of state derivative feedback controllers*, SIAM J. Control Optim., to appear; preprint available as Internal Report TW520, Department of Computer Science, Katholieke Universiteit Leuven, Belgium, online at http://www.cs.kuleuven.be/publicaties/rapporten/tw/TW520.abs.html (2008).

[16] H. Nakajima, *On analytical properties of delayed feedback control of chaos*, Phys. Lett. A, 232 (1997), pp. 207–210.

[17] N. Olgac, H. Elmali, M. Hosek, and M. Renzulli, *Active vibration control of distributed systems using delayed resonator with acceleration feedback*, Trans. ASME J. Dyn. Syst., 119 (1997), pp. 380–389.

[18] K. Pyragas, *Continuous control of chaos by self-controlling feedback*, Phys. Lett. A, 170 (1992), pp. 421–428.

[19] K. Pyragas, *Control of chaos via extended delay feedback*, Phys. Lett. A, 206 (1995), pp. 323–330.

[20] E. Schöll and H. G. Schuster, eds., *Handbook of Chaos Control—Second, Completely Revised and Enlarged Edition*, Wiley-VCH, Weinheim, Germany, 2007.

[21] J. Vanbiervliet, K. Verheyden, W. Michiels, and S. Vandewalle, *A nonsmooth optimization approach for the stabilization of linear time-delay systems*, ESAIM Control Optim. Calc. Var., 14 (2008), pp. 478–493.

[22] S. Yanchuk, M. Wolfrum, P. Hövel, and E. Schöll, *Control of unstable steady states by long delay feedback*, Phys. Rev. E, 74 (2006), paper 026201.

SIAM J. Applied Dynamical Systems
Vol. 8, No. 1, pp. 21–39
© 2009 Society for Industrial and Applied Mathematics

Stabilization of Long-Period Periodic Orbits Using Time-Delayed Feedback Control*

Claire M. Postlethwaite†

Abstract. The Pyragas method of feedback control has attracted much interest as a method of stabilizing unstable periodic orbits in a number of situations. We show that a time-delayed feedback control similar to the Pyragas method can be used to stabilize periodic orbits with arbitrarily large period, specifically those resulting from a resonant bifurcation of a heteroclinic cycle. Our analysis reduces the infinite-dimensional delay-equation governing the system with feedback to a three-dimensional map, by making certain assumptions about the form of the solutions. The stability of a fixed point in this map corresponds to the stability of the periodic orbit in the flow and can be computed analytically. We compare the analytic results to a numerical example and find very good agreement.

Key words. feedback control, heteroclinic cycle, delay equation

AMS subject classifications. 37C27, 37C29

DOI. 10.1137/080721510

1. Introduction.

The stabilization of unstable periodic orbits (UPOs) using feedback control has attracted the attention of many authors over a number of years. The time-delayed feedback method of Pyragas [1] has been of particular interest. Here, the feedback F is proportional to the difference between the current state and a past state of the system. Specifically, $F = K(x(t-\tau) - x(t))$, where $x(t)$ is some state vector, τ is the period of the targeted UPO, and K is a feedback gain matrix. Advantages of this method include the following. First, since the feedback vanishes on any orbit with period τ, the targeted UPO is still a solution of the system with feedback. Control is therefore achieved in a noninvasive manner. Second, the only information required a priori is the period τ of the target UPO, rather than a detailed knowledge of the profile of the orbit, or even any knowledge of the form of the original ODEs, which may be useful in experimental setups. The method has been implemented successfully in a variety of laboratory situations [2, 3, 4, 5, 6, 7, 8], as well as analytically and numerically in spatially extended pattern-forming systems [9, 10, 11, 12]; more examples can be found in a recent review by Pyragas [13].

Until now, there has been little or no study on whether there are limitations to Pyragas feedback control as the period of the targeted orbit, and hence the delay time, becomes large. In this paper, we investigate the use of Pyragas feedback on unstable periodic orbits with arbitrarily large period.

*Received by the editors April 16, 2008; accepted for publication (in revised form) by J. Guckenheimer August 29, 2008; published electronically January 9, 2009. This work was supported in part by grant NSF-DMS-0709232.
http://www.siam.org/journals/siads/8-1/72151.html
†Department of Mathematics, University of Auckland, Private Bag 92019, Auckland, New Zealand, and Department of Engineering Science and Applied Mathematics, Northwestern University, Evanston, IL 60208 (c.postlethwaite@math.auckland.ac.nz).

One mechanism for the generation of long-period periodic orbits is at bifurcations from homoclinic orbits or heteroclinic cycles. In this paper we focus on a subcritical bifurcation from a symmetric heteroclinic cycle, specifically the heteroclinic cycle of Guckenheimer and Holmes [14]. The bifurcation produces a branch of unstable long-period periodic orbits, and we investigate using a time-delayed feedback control similar to the Pyragas feedback as a stabilization mechanism.

The addition of Pyragas feedback to the ODEs considered by Guckenheimer and Holmes results in an infinite-dimensional delay equation. In order to analyze trajectories near the periodic orbit of interest, we make a number of assumptions about the form of solutions to the delay-differential equation and reduce the flow to a three-dimensional map. This method, after the assumptions have been made, is a modified version of the standard "small box and Poincaré map" analysis used by many authors to study the dynamics of trajectories close to heteroclinic cycles. This reduction of an infinite-dimensional delay equation to a finite-dimensional map has not appeared before in the literature. Although our assumptions are not fully rigorously justified, we test the validity of our arguments by comparing our results with a numerical example. We find excellent agreement between the analytical and numerical results. A surprising result of the analysis for the particular example we use is that, as the period of the orbit increases, the amplitude of the gain parameter required to stabilize the unstable orbits decreases.

This paper is organized as follows. In section 2 we give a review of heteroclinic cycles and their bifurcations. We describe the Guckenheimer–Holmes heteroclinic cycle and summarize the standard approach to analyzing trajectories close to heteroclinic cycles. In section 3 we describe how we choose the feedback control terms which are added to the equations. We then perform the reduction of the equations described above, which gives us a method of computing the stability of the periodic orbits. Section 4 contains numerical examples, and section 5 concludes.

2. Review of heteroclinic cycles.

A heteroclinic cycle is a topological circle of connecting orbits between at least two saddle-type equilibria. In generic (nonsymmetric) dynamical systems, heteroclinic cycles are of high codimension, and their existence for open sets of parameter values is unexpected. If a dynamical system contains flow-invariant subspaces, the connecting orbits can be contained within these subspaces, and then the heteroclinic cycle is robust to perturbations of the system that preserve the invariance of these subspaces. Flow-invariant subspaces can arise due to symmetry or due to other restrictions on the flow (such as extinctions in population dynamics models [15]). The review of Krupa [16] contains many examples of robust heteroclinic cycles. In this paper we consider robust heteroclinic cycles in symmetric systems.

2.1. Preliminary definitions.

Consider a continuous-time dynamical system defined by an ODE,

$$\dot{x} = f(x), \qquad x \in \mathbb{R}^n, \tag{2.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}^n$ is a Λ-equivariant vector field, that is,

$$\gamma f(x) = f(\gamma x) \qquad \forall\ \gamma \in \Lambda, \tag{2.2}$$

and $\Lambda \subset \mathbf{O}(n)$ is a finite Lie group. An equilibrium $\xi \in \mathbb{R}^n$ of (2.1) satisfies $f(\xi) = 0$. We consider only hyperbolic equilibria and assume that f is smoothly linearizable about each equilibrium.

Definition 1. *$\phi_j(t)$ is a* heteroclinic connection *between two equilibria ξ_j and ξ_{j+1} of (2.1) if $\phi_j(t)$ is a solution of (2.1) which is backward asymptotic to ξ_j and forward asymptotic to ξ_{j+1}.*

A heteroclinic cycle is an invariant set $X \subset \mathbb{R}^n$ consisting of the union of a set of equilibria $\{\xi_1, \dots, \xi_m\}$ and orbits $\{\phi_1, \dots, \phi_m\}$, where ϕ_j is a heteroclinic connection between ξ_j and ξ_{j+1}, and $\xi_{m+1} \equiv \xi_1$. We require that $m \geq 2$. If $m = 1$, then ϕ_1 is a homoclinic orbit. A heteroclinic cycle is a *homoclinic cycle* if there exists $\gamma \in \Lambda$ such that $\gamma\xi_j = \xi_{j+1}$ for all j.

For $x \in \mathbb{R}^n$ we define the *isotropy subgroup* Σ_x,

$$\Sigma_x = \{\sigma \in \Lambda : \sigma x = x\}. \tag{2.3}$$

For Σ an isotropy subgroup of Λ, we define the *fixed-point subspace*

$$\operatorname{Fix}\Sigma = \{x \in \mathbb{R}^n : \sigma x = x \ \forall \sigma \in \Sigma\}. \tag{2.4}$$

Definition 2. *A heteroclinic cycle X is* robust *if for each j, $1 \leq j \leq m$, there exists a fixed-point subspace, $P_j = \operatorname{Fix}\Sigma_j$, where $\Sigma_j \subset \Lambda$ and*

1. *ξ_j is a saddle and ξ_{j+1} is a sink for the flow restricted to P_j,*
2. *there is a heteroclinic connection from ξ_j to ξ_{j+1} contained in P_j.*

Robust heteroclinic cycles occur as codimension-zero phenomena in systems with symmetry. That is, they can exist for open sets of parameter values. Bifurcations of heteroclinic cycles therefore occur as codimension-one phenomena. We now consider the computation of the stability of heteroclinic cycles and the associated bifurcations.

2.2. Resonant bifurcations. The stability of a heteroclinic cycle is usually computed by constructing Poincaré maps on a Poincaré section of the flow. The flow near the cycle is divided into two parts: the "local" part, near the equilibria, where the flow can be well approximated by the linearized flow about the equilibria, and the "global" part of the flow, where the trajectory is away from the equilibria. The global part of the flow occurs on a much faster timescale than the local part and can be approximated by a linearization of the flow around the heteroclinic connections. The construction of such Poincaré maps is a standard procedure; details can be found in, for example, [17, 18].

Heteroclinic cycles generically lose stability in two ways: resonant bifurcations and transverse bifurcations. Transverse bifurcations occur when one of the eigenvalues at an equilibrium passes through zero; the equilibrium undergoes a local bifurcation. We do not consider transverse bifurcations here; see [19] for details. Throughout this paper, when we refer to "the eigenvalues at an equilibrium," we of course mean the eigenvalues of the Jacobian matrix of the flow linearized about that equilibrium.

At a resonant bifurcation the eigenvalues at the equilibria are generically nonzero but satisfy an algebraic condition that determines a global change in the stability properties of the cycle. Resonant bifurcations were first studied in the nonsymmetric case by Chow, Deng, and Fielder [20] in the context of a bifurcation from a homoclinic orbit. A more recent study [21] considers a codimension-two resonant bifurcation from a robust heteroclinic cycle with complex eigenvalues.

Resonant bifurcations are generically accompanied by the birth or death of a long-period periodic orbit. If ν is the bifurcation parameter controlling the resonant bifurcation (that is, $\nu = 0$ at the bifurcation point), then the period T of the bifurcating periodic orbit generically scales as

$$T \sim \frac{1}{\nu}.$$

Resonant bifurcations can occur in a supercritical or subcritical manner. We consider the subcritical case, when the branching periodic orbits are unstable, and in the following show that Pyragas-type time-delayed feedback can stabilize the periodic orbits. Our analysis focuses on the Guckenheimer–Holmes cycle in $\mathbb{R}^3$.

2.3. The Guckenheimer–Holmes cycle. The Guckenheimer–Holmes (GH) cycle [14] is a prototypical example of a robust heteroclinic cycle. We use this cycle as an example on which to base our analysis. First we review the original case with no feedback.

The equations considered by Guckenheimer and Holmes can be written as

$$\begin{aligned} \dot{x}_1 &= x_1(1 - \mathbf{X} - \mu x_2^2 + \lambda x_3^2), \\ \dot{x}_2 &= x_2(1 - \mathbf{X} - \mu x_3^2 + \lambda x_1^2), \\ \dot{x}_3 &= x_3(1 - \mathbf{X} - \mu x_1^2 + \lambda x_2^2), \end{aligned} \tag{2.5}$$

where $\mathbf{X} = \sum_{i=1}^{3} x_i^2$, and μ and λ are real parameters. The equations are equivariant under the symmetry group $\Lambda = \mathbb{Z}_3 \ltimes \mathbb{Z}_2^3$, generated by a reflection κ_1 and a rotation γ:

$$\begin{aligned} \kappa(x_1, x_2, x_3) &= (-x_1, x_2, x_3), \\ \gamma(x_1, x_2, x_3) &= (x_3, x_1, x_2). \end{aligned}$$

We label the equilibrium on the positive x_j-axis as ξ_j. Here, and throughout the remainder of the paper, subscripts on equilibria, coordinates, and similar objects should be taken mod 3. Each two-dimensional coordinate plane is a fixed-point subspace. If $\mu\lambda > 0$, then the only equilibria in each coordinate plane are those lying on the coordinate axes. We consider the case $\mu, \lambda > 0$, and then it can be shown that in the plane $x_3 = 0$, ξ_1 is a saddle and ξ_2 is a sink. It can additionally be shown that in forward time trajectories are bounded away from infinity, and therefore by the Poincaré–Bendixson theorem there exists a heteroclinic connection from ξ_1 to ξ_2. Similarly, connections also exist from ξ_2 to ξ_3 and ξ_3 to ξ_1. These connections lie in two-dimensional fixed-point subspaces (the two-dimensional coordinate planes), so the cycle is robust. The resulting heteroclinic cycle is shown schematically in Figure 1. Also note that $\gamma\xi_j = \xi_{j+1}$, so the cycle is homoclinic.

The stability of the cycle can be calculated using the methods described above. It is a standard procedure, but we outline the method here, as we use similar ideas later when considering the stability of periodic orbits in the system with added time-delayed feedback. Consider a trajectory which passes close to the equilibrium ξ_1. The linearized flow near ξ_1 is

$$\dot{x}_1 = -2x_1, \tag{2.6}$$

$$\dot{x}_2 = \lambda x_2, \tag{2.7}$$

$$\dot{x}_3 = -\mu x_3. \tag{2.8}$$

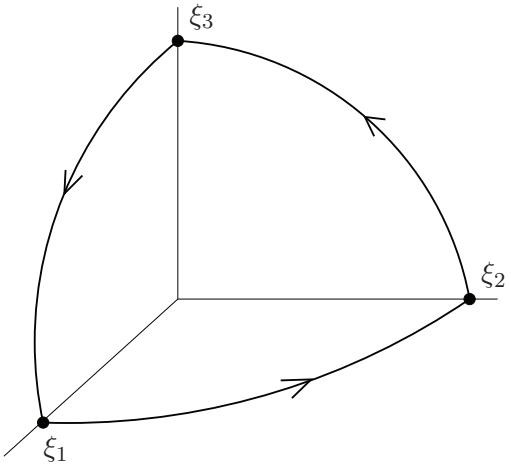


Figure 1. *The GH cycle in* $\mathbb{R}^3$.

The x_1 direction is the "radial" direction, and as shown in [22], for heteroclinic cycles of this type the radial direction does not affect the stability of the cycle. All trajectories move away from the origin, and also away from infinity, and in this case are attracted to an "invariant sphere" [23] which contains the heteroclinic cycle. Therefore, for simplicity, we henceforth ignore this component. We define Poincaré sections close to ξ_1,

$$H_1^{\text{in}} = \{(x_1, x_2, x_3) \mid 0 < x_2 < h,\ x_3 = h\}, \tag{2.9}$$

$$H_1^{\text{out}} = \{(x_1, x_2, x_3) \mid x_2 = h,\ 0 < x_3 < h\}, \tag{2.10}$$

where $0 < h \ll 1$, and construct a Poincaré return map on H_1^{in}. Consider a trajectory which passes through H_1^{in} at time $t = 0$ with $x_2(0) = x_2^i$. The trajectory will hit H_1^{out} at

$$t = T_i \approx (-1/\lambda) \log(x_2^i/h)$$

with

$$x_3(T_i) \equiv x_3^o \approx h^{1+\mu/\lambda} {x_2^i}^{\mu/\lambda}.$$

We thus write down a local map $\phi_{\text{loc}} : H_1^{\text{in}} \mapsto H_1^{\text{out}}$, which describes the flow near the equilibria:

$$x_3^o = \phi_{\text{loc}}(x_2^i) = h^{1+\mu/\lambda} {x_2^i}^{\mu/\lambda}. \tag{2.11}$$

The flow near the heteroclinic connection from H_1^{out} to a plane $H_2^{\text{in}} = \gamma H_1^{\text{in}}$ near ξ_2 is approximated by the global map ϕ_{far}:

$$x_3^i = \phi_{\text{far}}(x_3^o) = A x_3^o + O({x_3^o}^2), \tag{2.12}$$

where x_3^i is the x_3 coordinate of the trajectory when it next hits H_2^{in} and A is a positive constant. Note that the constant term in this expansion of ϕ_{far} is zero because the plane $x_3 = 0$ is invariant. We write $\tilde{\phi} = \phi_{\text{far}} \circ \phi_{\text{loc}}$. Since $H_1^{\text{in}} = \gamma^{-1} H_2^{\text{in}}$, the map $\gamma^{-1}\tilde{\phi}(x_2^i)$ is a return map on H_1^{in}. Write $\phi = \gamma^{-1}\tilde{\phi}$, and then the return map is

$$\phi(x) = A x^{\delta},$$

where $\delta = \mu/\lambda$.

The map $\phi(x)$ has fixed points at $x = 0$ and at $x = x_p = A^{1/(1-\delta)}$. The fixed point at $x = 0$ corresponds to the heteroclinic cycle in the flow and is stable if $\delta > 1$. The heteroclinic cycle loses stability in a resonant bifurcation at $\delta = 1$. The second fixed point at $x = x_p$ corresponds to a branch of periodic orbits, as long as x_p is small and positive. The symmetry γ acts as a spatiotemporal symmetry on the periodic orbits. That is, if we write the periodic solution as a trajectory $x^\star(t; \mu, \lambda)$, with minimal period T, then

$$\gamma^{-1} x^\star(t) = x^\star(t - T/3).$$

The stability of the orbits can be determined by finding the stability of the fixed point x_p in the map ϕ. It is simple to see that if $A > 1$, x_p is small and positive (and hence corresponds to a periodic orbit in the flow) when $\delta > 1$, so the resonant bifurcation is subcritical. We find that

$$\frac{\mathrm{d}\phi}{\mathrm{d}x}_{x=x_p} = \delta,$$

and so x_p is unstable. Conversely, if $A < 1$, then x_p corresponds to a branch of stable periodic orbits if $\delta < 1$, and the bifurcation is supercritical. The period T of the orbit is approximately

$$T = 3\tau \approx \frac{-3 \log A}{\lambda - \mu},$$

where $\tau = -\log(x_p)/\lambda$ and is the time spent by the trajectory each time it passes close to an equilibrium. We are ignoring the time spent away from the equilibria (that is, close to the heteroclinic connections in the invariant planes) since it is much less than T when we are close to the resonant bifurcation, that is, $|1 - \delta| \ll 1$.

In equations (2.5), the resonant heteroclinic bifurcation at $\mu = \lambda$ is degenerate. That is, the branch of periodic orbits exists only at $\mu = \lambda$. This corresponds to the case $A = 1$ in the map ϕ. We add additional higher order terms to break this degeneracy; specifically we consider

$$\begin{aligned} \dot{x}_1 &= x_1(1 - \mathbf{X} - \mu x_2^2 + \lambda x_3^2 + \alpha x_2^2 x_3^2), \\ \dot{x}_2 &= x_2(1 - \mathbf{X} - \mu x_3^2 + \lambda x_1^2 + \alpha x_3^2 x_1^2), \\ \dot{x}_3 &= x_3(1 - \mathbf{X} - \mu x_1^2 + \lambda x_2^2 + \alpha x_1^2 x_2^2). \end{aligned} \tag{2.13}$$

The additional terms preserve the equilibria and the symmetries of the system and also the invariant planes and the heteroclinic cycle. The heteroclinic cycle still loses stability in a resonant bifurcation at $\mu = \lambda$, but now a branch of periodic orbits is created in either $\mu > \lambda$ or $\mu < \lambda$. The sign of α determines the branching direction and whether, in the map ϕ, A is greater or less than 1. If $\alpha > 0$, we see a branch of unstable periodic orbits in $\mu > \lambda$ (and the resonant bifurcation is subcritical). If $\alpha < 0$, we see a branch of stable periodic orbits in $\mu < \lambda$ (and the bifurcation is supercritical). A complete study of the effect of fifth order terms on the dynamics near the GH cycle has not been performed. However, the above assertion can be seen by considering the effect of the new term on the x_j component when the trajectory is close to the $x_j = 0$ plane but away from either coordinate axis.

In the following, we consider the subcritical case, where the periodic orbits are unstable, and add noninvasive time-delayed feedback to stabilize the orbits near the heteroclinic cycle.

3. Addition of feedback terms.

3.1. Choice of coordinates. To ease analysis and improve the accuracy in the numerical computations in section 4, we introduce new coordinates $Y_j = \log(x_j^2)$. Along with a change in timescale, this transforms equations (2.13) into

$$\begin{aligned} \dot{Y}_1 &= (1 - \mathbf{Y} - \mu e^{Y_2} + \lambda e^{Y_3} + \alpha e^{Y_2+Y_3}), \\ \dot{Y}_2 &= (1 - \mathbf{Y} - \mu e^{Y_3} + \lambda e^{Y_1} + \alpha e^{Y_3+Y_1}), \\ \dot{Y}_3 &= (1 - \mathbf{Y} - \mu e^{Y_1} + \lambda e^{Y_2} + \alpha e^{Y_1+Y_2}), \end{aligned} \tag{3.1}$$

where $\mathbf{Y} = \sum_{j=1}^{3} e^{Y_j}$. Note that in these coordinates, the equilibria are at, e.g., $Y_1 = 0$, $Y_2 = Y_3 = -\infty$. The invariant planes in the x_j coordinates are transformed to $Y_j = -\infty$. However, we are not interested in trajectories which lie in the coordinate planes, only those which are close to them.

3.2. Addition of feedback. Pyragas feedback is additive, and it exhibits the form $F = \Gamma(\mathbf{x}(t-T) - \mathbf{x}(T))$, where Γ is a (real) gain matrix and T is the period of the targeted periodic orbit. Our choice of coordinates suggests the following slightly altered functional form for the feedback:

$$F = \Gamma \begin{pmatrix} e^{Y_1(t-T)-Y_1(t)} - 1 \\ e^{Y_2(t-T)-Y_2(t)} - 1 \\ e^{Y_3(t-T)-Y_3(t)} - 1 \end{pmatrix}. \tag{3.2}$$

For trajectories close to the periodic orbit, $Y_j(t-T)-Y_j(t) \ll 1$, and so the feedback terms are approximately of Pyragas form. For this choice of feedback, the equilibria and the invariance of the coordinate planes (in the original x_j coordinates) are preserved. However, we additionally choose to use the symmetries of the system to make a further change in the form of the feedback which simplifies the subsequent analysis. The feedback we use is

$$F = \Gamma \begin{pmatrix} e^{Y_3(t-\tau)-Y_1(t)} - 1 \\ e^{Y_1(t-\tau)-Y_2(t)} - 1 \\ e^{Y_2(t-\tau)-Y_3(t)} - 1 \end{pmatrix}, \tag{3.3}$$

where $\tau = T/3$ is one-third of the period of the orbit. Due to the spatiotemporal symmetry of the periodic orbit under the action of γ, the feedback vanishes at the periodic orbit, and so the periodic orbit is still a solution of the system. However, this feedback does not preserve the equilibria or invariant planes (in the original x_j coordinates).

We choose the matrix Γ in a similar manner to that in [24], as follows. We write

$$\Gamma = EGE^{-1},$$

where

$$G = \begin{pmatrix} 0 & 0 & 0 \\ 0 & b_0 \cos\beta & -b_0 \sin\beta \\ 0 & b_0 \sin\beta & b_0 \cos\beta \end{pmatrix}, \quad E = \begin{pmatrix} 1 & -1/2 & \sqrt{3}/2 \\ 1 & -1/2 & -\sqrt{3}/2 \\ 1 & 1 & 0 \end{pmatrix}. \tag{3.4}$$

The matrix G has the form of the feedback matrix used by Fiedler et al. [25] in a two-dimensional example: stabilizing periodic orbits emanating from a subcritical Hopf bifurcation. Recall that the orbit has two unstable directions and one stable direction—the radial direction. The matrix E is chosen so that the feedback is rotated to align with the unstable directions, and there is no feedback in the stable direction.

The resulting equations with feedback are

$$\begin{pmatrix} \dot{Y}_1 \\ \dot{Y}_2 \\ \dot{Y}_3 \end{pmatrix} = I + M \begin{pmatrix} \mathrm{e}^{Y_1} \\ \mathrm{e}^{Y_2} \\ \mathrm{e}^{Y_3} \end{pmatrix} + \alpha \begin{pmatrix} \mathrm{e}^{Y_2+Y_3} \\ \mathrm{e}^{Y_3+Y_1} \\ \mathrm{e}^{Y_1+Y_2} \end{pmatrix} + \Gamma \begin{pmatrix} \mathrm{e}^{Y_{3\tau}-Y_1} - 1 \\ e^{Y_{1\tau}-Y_2} - 1 \\ e^{Y_{2\tau}-Y_3} - 1 \end{pmatrix}, \tag{3.5}$$

where

$$M = \begin{pmatrix} -1 & -\mu - 1 & \lambda - 1 \\ \lambda - 1 & -1 & -\mu - 1 \\ -\mu - 1 & \lambda - 1 & -1 \end{pmatrix} \qquad \text{and} \qquad Y_{j_\tau} = Y_j(t - \tau).$$

3.3. Stability analysis. We analyze the stability of the periodic orbits close to the heteroclinic cycle in a similar manner to the methods used without feedback. We assume we are close to the resonant bifurcation, that is, $|1-\delta| \ll 1$, so that the periodic orbit lies close to the heteroclinic cycle, and consider the flow close to the periodic orbit. The linearized equations close to the equilibrium ξ_1 are given by

$$\dot{Y}_2 = \lambda + \Gamma_{22}(Y_{1\tau} - Y_2) + \Gamma_{23}(Y_{2\tau} - Y_3), \tag{3.6}$$

$$\dot{Y}_3 = -\mu + \Gamma_{32}(Y_{1\tau} - Y_2) + \Gamma_{33}(Y_{2\tau} - Y_3), \tag{3.7}$$

where the Γ_{jk} are the components of the feedback gain matrix Γ. As before, we neglect the Y_1 equation—since the feedback only acts in directions tangent to the plane containing the periodic orbit, we assume that when trajectories are close enough to the periodic orbit the dynamics in the radial direction are unaffected. That is, near ξ_1, the Y_1 direction will be contracting and so not affect the stability of the orbit. In section 4.1 we show numerical results which support this assumption.

Recall that the periodic orbits we are attempting to stabilize are spatiotemporally symmetric under the action of γ. We make use of this in the following. At each equilibrium, we define a contracting direction and an expanding direction. At ξ_j, the contracting direction Y_c is the Y_{j-1} direction, and the expanding direction Y_e is the Y_{j+1} direction.

Unlike in the case without feedback, we cannot solve the linear equations explicitly, and so we make the following approximations. Let $Y^\star(t;\mu,\lambda)$ be the periodic orbit for the original system (in the logarithmic coordinates). Then $Y^\star$ is still a solution of the system with feedback. Consider solving the delay-differential equation for the system with feedback for a trajectory $Y(t)$ which starts close to $Y^\star$. That is, for $-\tau < t < 0$, $Y(t)$ is close to $Y^\star(t)$. Then for $0 < t < \tau$ the feedback terms in the delay-differential equation will be small; that is, the equations will only be a small perturbation from the original system. By continuity, the solution $Y(t)$ for $0 < t < \tau$ will also be close to $Y^\star$.

Set $t = 0$ as the trajectory intersects the plane H_j^{in} ($Y_{j-1} = H$, for $H = \log h$, $0 < h \ll 1$) on the ith time the trajectory passes close to an equilibrium, ξ_j. With no feedback, the

local part of the trajectory can be written down exactly. The expanding and contracting components, $Y_e(t)$ and $Y_c(t)$, satisfy

$$Y_e(t) = Y_e^i + \lambda t, \tag{3.8}$$

$$Y_c(t) = H - \mu t, \tag{3.9}$$

for $t \in [0, T_i)$, where T_i is the length of time spent near the equilibrium (i.e., in the small box) and Y_e^i is the expanding coordinate of the trajectory as it intersects the plane $Y_c = H$.

For the system with feedback, we cannot explicitly solve the linearized equations. Given the argument above, we assume that we start sufficiently close to the periodic orbit that solutions are only a small perturbation away from those for the case with no feedback. That is, we write, for $t \in [0, T_i)$,

$$Y_e(t) = Y_e^i + \lambda t + f_i(t), \tag{3.10}$$

$$Y_c(t) = H - \mu t + g_i(t), \tag{3.11}$$

where $f_i(t)$ and $g_i(t)$ are functions which satisfy

$$f_i(t), g_i(t) \ll 1, \qquad f_i(0) = g_i(0) = 0,$$

and if the trajectory is exactly the periodic orbit, $f_i(t), g_i(t) \equiv 0$. We will use this assumed form of the local flow together with (3.6) and (3.7) and the global flow as before to derive a new return map. This gives recurrence relations for Y_e^i and the functions $f_i(t)$ and $g_i(t)$. Figure 2 shows a schematic of the local flow past an equilibrium ξ_j, in the original x_j coordinates.

We again ignore the time the trajectory spends near the heteroclinic connections but away from the equilibria, so for $t \in [-T_{i-1}, 0)$ the flow is given by

$$Y_e(t) = Y_e^{i-1} + \lambda t + f_{i-1}(t), \tag{3.12}$$

$$Y_c(t) = H - \mu t + g_{i-1}(t). \tag{3.13}$$

Using the symmetry γ, we can rewrite the linear equations (3.6) and (3.7) as

$$\dot{Y}_e = \lambda + \Gamma_{22}(Y_{e\tau} - Y_e) + \Gamma_{23}(Y_{c\tau} - Y_c), \tag{3.14}$$

$$\dot{Y}_c = -\mu + \Gamma_{32}(Y_{e\tau} - Y_e) + \Gamma_{33}(Y_{c\tau} - Y_c), \tag{3.15}$$

where the delayed terms are the corresponding coordinates near the previous equilibrium, that is,

$$Y_{e\tau} \equiv Y_e(t - \tau) = Y_e(-\tau) + \lambda t + f_{i-1}(t), \tag{3.16}$$

$$Y_{c\tau} \equiv Y_c(t - \tau) = Y_c(-\tau) - \mu t + g_{i-1}(t). \tag{3.17}$$

The time of flight of the trajectory between the planes H_j^{in} and H_j^{out}, T_{i-1}, will not be equal to the delay time τ except when the trajectory is exactly on the periodic orbit (see Figure 2). In order to find the coordinates at $t = -\tau$, we assume that the flow given by (3.12)

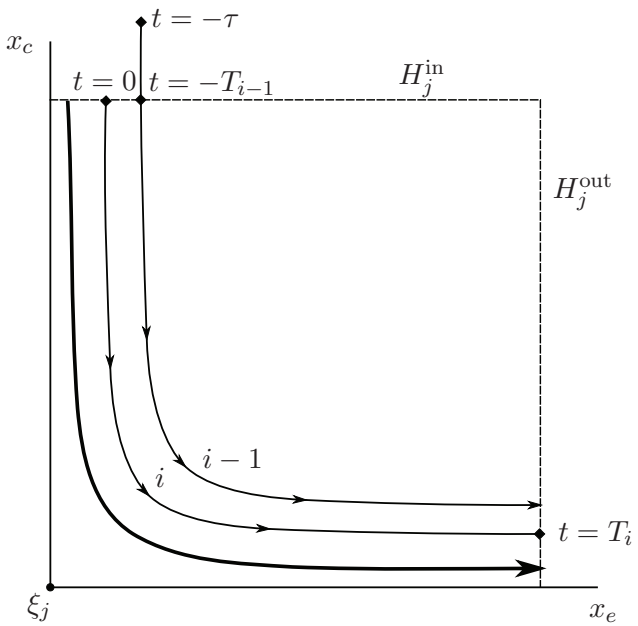


Figure 2. *The figure shows a schematic of the local flow as the trajectory passes an equilibrium ξ_j for the $(i-1)$th and ith time. The bold line indicates the periodic orbit. We set $t = 0$ as the trajectory intersects H_j^{in} on the ith pass. The trajectory intersects H_j^{in} at $t = -T_{i-1}$ on the $(i-1)$th pass, which does not coincide with $t = -\tau$. Note that here we show the $(i-1)$th trajectory passing the same equilibrium as the ith trajectory—in actuality the $(i-1)$th pass is of the previous equilibrium, but because of the symmetries in the system this is equivalent and schematically simpler to show in the figure. The diamonds (♦) indicate the points of the trajectory at which times are indicated on the figure. The dashed lines indicate the surfaces H_j^{in} and H_j^{out}.*

and (3.13) is also valid for $t \in (-\tau, -T_{i-1}]$. The coordinates of the trajectory at $t = -\tau$ are therefore

$$Y_e(-\tau) = Y_e^{i-1} + \lambda(T_{i-1} - \tau) + f_{i-1}(T_{i-1} - \tau), \tag{3.18}$$

$$Y_c(-\tau) = H - \mu(T_{i-1} - \tau) + g_{i-1}(T_{i-1} - \tau). \tag{3.19}$$

Writing $T_i = \tau + \Delta_i$, where $\Delta_i/\tau \ll 1$ (since we are close to the periodic orbit), and expanding f_{i-1} and g_{i-1} about zero gives

$$f_{i-1}(\Delta_{i-1}) = f_{i-1}(0) + \Delta_{i-1} f'_{i-1}(0) + \cdots \approx \Delta_{i-1} f'_{i-1}(0), \tag{3.20}$$

$$g_{i-1}(\Delta_{i-1}) = g_{i-1}(0) + \Delta_{i-1} g'_{i-1}(0) + \cdots \approx \Delta_{i-1} g'_{i-1}(0), \tag{3.21}$$

since $f_i(0) = g_i(0) = 0$. Substituting this into (3.18) and (3.19) gives

$$Y_e(-\tau) = Y_e^{i-1} + \Delta_{i-1}(\lambda + f'_{i-1}(0)), \tag{3.22}$$

$$Y_c(-\tau) = H + \Delta_{i-1}(-\mu + g'_{i-1}(0)). \tag{3.23}$$

We also have that for $t \in [0, T_i)$,

$$\dot{Y}_e(t) = \lambda + f'_i(t), \quad \dot{Y}_c(t) = -\mu + g'_i(t).$$

Substituting (3.10), (3.11), (3.22), and (3.23) into (3.14) and (3.15), we find

$$f_i'(t) = \Gamma_{22}[Y_e^{i-1} - Y_e^i + \Delta_{i-1}(\lambda + f_{i-1}'(0)) + f_{i-1}(t) - f_i(t)] + \Gamma_{23}[\Delta_{i-1}(-\mu + g_{i-1}'(0)) + g_{i-1}(t) - g_i(t)], \tag{3.24}$$

$$g_i'(t) = \Gamma_{32}[Y_e^{i-1} - Y_e^i + \Delta_{i-1}(\lambda + f_{i-1}'(0)) + f_{i-1}(t) - f_i(t)] + \Gamma_{33}[\Delta_{i-1}(-\mu + g_{i-1}'(0)) + g_{i-1}(t) - g_i(t)]. \tag{3.25}$$

These expressions are true for all $t \in [0, T_i)$, so we set $t = 0$ to simplify and find

$$f_i'(0) = \Gamma_{22}[Y_e^{i-1} - Y_e^i + \Delta_{i-1}(\lambda + f_{i-1}'(0))] + \Gamma_{23}[\Delta_{i-1}(-\mu + g_{i-1}'(0))], \tag{3.26}$$

$$g_i'(0) = \Gamma_{32}[Y_e^{i-1} - Y_e^i + \Delta_{i-1}(\lambda + f_{i-1}'(0))] + \Gamma_{33}[\Delta_{i-1}(-\mu + g_{i-1}'(0))], \tag{3.27}$$

that is, a recurrence relation for $f_i'(0)$ and $g_i'(0)$ if the Δ_i and Y_e^i are known. We write $X_i = Y_e^i - H$, $\lambda_i = f_i'(0) + \lambda$, and $\mu_i = g_i'(0) - \mu$ to further simplify:

$$\lambda_i = \lambda + \Gamma_{22}(X_{i-1} - X_i + \Delta_{i-1}\lambda_{i-1}) + \Gamma_{23}\Delta_{i-1}\mu_{i-1}, \tag{3.28}$$

$$\mu_i = -\mu + \Gamma_{32}(X_{i-1} - X_i + \Delta_{i-1}\lambda_{i-1}) + \Gamma_{33}\Delta_{i-1}\mu_{i-1}. \tag{3.29}$$

We next find an expression for Δ_i which we use to find a recurrence relation for the X_i. Recall that $Y_e(T_i) = H$, so from (3.10) we have

$$H = Y_e(T_i) = Y_e^i + \lambda T_i + f_i(T_i), \tag{3.30}$$

$$X_i = -\lambda T_i - f_i(T_i). \tag{3.31}$$

In order to be able to get tractable results in what follows, we need to invert the above equation for T_i. Motivated by numerical results, which we give in section 4.1, we make the following assumption:

$$f_i(T_i) \approx T_i f_i'(0),$$

that is, that f_i is approximately a linear function of t. Using this gives us

$$T_i \approx -\frac{X_i}{\lambda_i}$$

so

$$\Delta_i = T_i - \tau \approx -\frac{X_i}{\lambda_i} + \frac{X^\star}{\lambda}, \tag{3.32}$$

where $X^\star = -\lambda\tau = \log A/(1-\delta)$.

We make a similar assumption on the g_i, that is, $g_i(T_i) \approx T_i g_i'(0)$, and then use (3.11) to find $Y_c(T_i)$,

$$Y_c^o \equiv Y_c(T_i) \approx H + (-\mu + g_i'(\tau))T_i = H - \frac{\mu_i}{\lambda_i}X_i,$$

which is an expression for the local map $\phi_{\text{loc}} : Y_e^i \to Y_c^o$. We assume that the global map ϕ_{far} is of the same form as the case without feedback (2.12) when we are close enough to the periodic orbit, and hence find a return map for the X_i:

$$X_{i+1} = \log A - \frac{\mu_i}{\lambda_i}X_i.$$

Substituting (3.32) into (3.28) and (3.29) results in a third order recurrence system:

$$
\begin{aligned}
X_i &= \log A - \frac{\mu_{i-1}}{\lambda_{i-1}} X_{i-1}, \\
\lambda_i &= \lambda + \Gamma_{22}\left(\frac{\lambda_{i-1}}{\lambda}X^\star + \frac{\mu_{i-1}}{\lambda_{i-1}}X_{i-1}\right) + \Gamma_{23}\mu_{i-1}\left(\frac{X^\star}{\lambda} - \frac{X_{i-1}}{\lambda_{i-1}}\right) - \Gamma_{22}\log A, \\
\mu_i &= -\mu + \Gamma_{32}\left(\frac{\lambda_{i-1}}{\lambda}X^\star + \frac{\mu_{i-1}}{\lambda_{i-1}}X_{i-1}\right) + \Gamma_{33}\mu_{i-1}\left(\frac{X^\star}{\lambda} - \frac{X_{i-1}}{\lambda_{i-1}}\right) - \Gamma_{32}\log A.
\end{aligned} \tag{3.33}
$$

Note that when $\Gamma_{jk} \equiv 0$, the recurrence relation reduces to that for the system with no feedback, as expected. This system of three recurrence relations has a fixed point at

$$\lambda_i = \lambda, \quad \mu_i = -\mu, \quad X_i = X^\star,$$

which corresponds to the periodic orbit in the flow. The stability of the fixed point in the recurrence relation will correspond to the stability of the periodic orbit in the flow.

The Jacobian matrix J of (3.33) at this fixed point is

$$J = \begin{pmatrix} \delta & -\delta\hat{X} & -\hat{X} \\ -\delta(\Gamma_{22}-\Gamma_{23}) & \hat{X}[\Gamma_{22}+\delta(\Gamma_{22}-\Gamma_{23})] & \Gamma_{22}\hat{X} \\ -\delta(\Gamma_{32}-\Gamma_{33}) & \hat{X}[\Gamma_{32}+\delta(\Gamma_{32}-\Gamma_{33})] & \Gamma_{32}\hat{X} \end{pmatrix},$$

where $\hat{X} = X^\star/\lambda$. The characteristic equation of J is

$$m^3 + am^2 + bm + c = 0, \tag{3.34}$$

where

$$a = \hat{X}\delta(\Gamma_{23} - \Gamma_{22}) - \hat{X}(\Gamma_{22} + \Gamma_{32}) - \delta, \tag{3.35}$$

$$b = \hat{X}^2\delta(\Gamma_{22}\Gamma_{33} - \Gamma_{23}\Gamma_{32}) + \hat{X}\delta(\Gamma_{33} + \Gamma_{22}), \tag{3.36}$$

$$c = -\hat{X}^2\delta(\Gamma_{22}\Gamma_{33} - \Gamma_{23}\Gamma_{32}). \tag{3.37}$$

The fixed point will be unstable if (3.34) has any solutions with $|m| > 1$, so curves with $|m| = 1$ define stability boundaries of the periodic orbit. Recall that $\Gamma = EGE^{-1}$ and is a function of just two parameters, b_0 and β. We consider the stability of the periodic orbit as the parameters δ and b_0 are varied.

3.4. Determination of stability boundaries. We split our investigation of the stability boundaries into three cases. We introduce the bifurcation parameter $\nu = \delta - 1$. Without feedback, the heteroclinic cycle is stable in $\nu > 0$, and the periodic orbits exist and are unstable in $\nu > 0$. We consider analytically the limits of the stability boundary curves as $\nu, b_0 \to 0$. The boundaries can actually be computed exactly (although the algebra is rather nasty), since the eigenvalues are the roots of a cubic. We plot the boundaries for specific parameter values in Figure 3. In section 4 we compute the stability of the periodic orbit in the original system (3.5) numerically using the continuation package DDE-BIFTOOL.

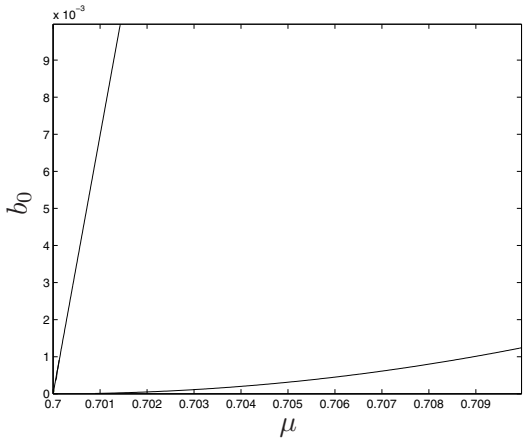


Figure 3. *The figure shows the stability boundaries of the periodic orbit, that is, curves of solutions $|m| = 1$ to (3.34), as μ and b_0 are varied. The lower boundary is a steady-state bifurcation ($m = 1$), and the left-hand boundary is a period-doubling bifurcation ($m = -1$). The orbit is stable in the wedge between these lines. Remaining parameters are $\lambda = 0.7$, $\beta = \pi/4$, $\alpha = 0.01$, so $\nu = 0$ at $\mu = 0.7$ (left-hand side of figure).*

Case 1: $m = 1$. A stability boundary with $m = 1$ corresponds to a steady-state bifurcation of the periodic orbit. This occurs when $1 + a + b + c = 0$, that is,

$$1 - \delta + \hat{X}\delta(\Gamma_{23} + \Gamma_{33}) - \hat{X}(\Gamma_{22} + \Gamma_{32}) = 0.$$

It can easily be computed that

$$\Gamma_{23} + \Gamma_{33} = \frac{b_0}{3}(\cos\beta - \sqrt{3}\sin\beta) \tag{3.38}$$

and

$$\Gamma_{22} + \Gamma_{32} = \frac{b_0}{3}(\cos\beta + \sqrt{3}\sin\beta). \tag{3.39}$$

In the limit $\nu, b_0 \to 0$, using $\hat{X} = -\frac{\log A}{\lambda\nu}$, we find

$$\nu^2 = p_1 b_0, \qquad p_1 = \frac{2\log A \sin\beta}{\sqrt{3}\lambda}. \tag{3.40}$$

It is also simple to calculate that in the limit $\nu \to 0$ the eigenvalue which goes through $m = 1$ as this curve is crossed is greater than 1 if $\nu^2 > p_1 b_0$ and less than 1 if $\nu^2 < p_1 b_0$.

Case 2: $m = -1$. A stability boundary with $m = -1$ will correspond to a period-doubling bifurcation of the periodic orbit. These curves will have $-1 + a - b + c = 0$, that is,

$$-1 - \delta + \hat{X}\delta(\Gamma_{23} - 2\Gamma_{22} - \Gamma_{33}) - \hat{X}(\Gamma_{22} + \Gamma_{32}) - 2\hat{X}^2\delta(\Gamma_{22}\Gamma_{33} - \Gamma_{23}\Gamma_{32}) = 0.$$

Again, we can compute the coefficients

$$\Gamma_{22}\Gamma_{33} - \Gamma_{23}\Gamma_{32} = \frac{b_0^2}{3}, \tag{3.41}$$

$$\Gamma_{23} - 2\Gamma_{22} - \Gamma_{33} = -\frac{b_0}{3}(7\cos\beta + \sqrt{3}\sin\beta), \tag{3.42}$$

and in the same limit as above, we find

$$1 - \frac{\log A}{3\lambda}(4\cos\beta + \sqrt{3}\sin\beta)\frac{b_0}{\nu} + \frac{(\log A)^2}{3\lambda^2}\frac{b_0^2}{\nu^2} = 0, \tag{3.43}$$

so there are two solutions $b_0 = c_\pm \nu$ for some $c_\pm$ function of A, λ, and β. The direction of the bifurcation as these lines are crossed in this case depends on β.

Case 3: $m = e^{i\theta}$, $\theta \neq n\pi$. For $m = e^{i\theta}$, $\theta \neq n\pi$, it can easily be computed that we must have $b + c^2 - ac = 1$. The computations in this case are messier, so we omit them and give the resulting curve in the limit $b_0, \nu \to 0$,

$$\left(\frac{b_0}{\nu}\right)^4 \frac{(\log A)^4}{3\lambda^4} - \left(\frac{b_0}{\nu}\right)^3 \frac{2(\log A)^3}{3\lambda^3}(2\cos\beta + \sqrt{3}\sin\beta) - \frac{b_0^2}{\nu}\frac{(\log A)^2}{\lambda^2} - \frac{b_0}{\nu}\frac{4\log A}{\lambda}\cos\beta = 3. \tag{3.44}$$

The direction of the bifurcation again will depend on β.

The curves (3.40), (3.43), and (3.44) describe the limiting cases of the stability boundaries of the periodic orbit as the point $\nu = b_0 = 0$ is approached. Since the characteristic polynomial (3.34) is cubic, it can be solved for any values of ν and b_0. In Figure 3 we plot the solutions $|m| = 1$ of (3.34) for a specific set of parameter values. In this case only the curves corresponding to $m = 1$ and $m = -1$ are stability boundaries. The lower boundary is the quadratic curve for $m = 1$, and the left-hand boundary is a straight line corresponding to $m = -1$. The remaining curves with $|m| = 1$ do not form stability boundaries in this case because the periodic orbit is already unstable in the regions in which they exist.

We can see that for these parameter values the periodic orbit is stable for a wide range of parameters and, specifically, can be stabilized arbitrarily close to the heteroclinic cycle, that is, for arbitrarily large period. That is, for any $\nu > 0$ we can find a b_0 for which the periodic orbit is stable. In fact, for this particular case, we see that as ν gets smaller, in order for the orbit to be stable, we have to choose b_0, the gain parameter, to be increasingly small. This seems a rather surprising result—that as the period of the targeted orbit increases, the amplitude of the gain parameter tends towards zero.

3.5. Stability of the heteroclinic cycle. We note that the recurrence relations (3.33) have a second solution, $\lambda_i = \lambda$, $\mu_i = -\mu$, $X_i \to -\infty$, which corresponds to the heteroclinic cycle. We can consider the stability of this solution by considering the solutions to (3.34) in the limit $\hat{X} \to -\infty$. In this limit, $a \sim O(\hat{X})$, $b \sim O(\hat{X}^2)$, and $c \sim O(\hat{X}^2)$. The cubic equation (3.34) therefore has one solution with $m \sim O(1)$ and two solutions with $m \sim O(\hat{X})$. Therefore this fixed point is always unstable in the recurrence relation, and so the heteroclinic cycle is always unstable in the flow.

4. Numerical results. We use the MATLAB package DDE-BIFTOOL [26] to numerically analyze the stability of periodic orbits in the system (3.5). The delay time τ was set equal to the period of the bifurcating periodic orbits (and so is a function of μ) and was calculated numerically from the system with no feedback. Parameters used were the same as those used to produce Figure 3.

Figure 4 shows a contour plot of the amplitude of the largest Floquet multiplier as the parameters b_0 and μ are varied. The periodic orbit is stable when all Floquet multipliers have

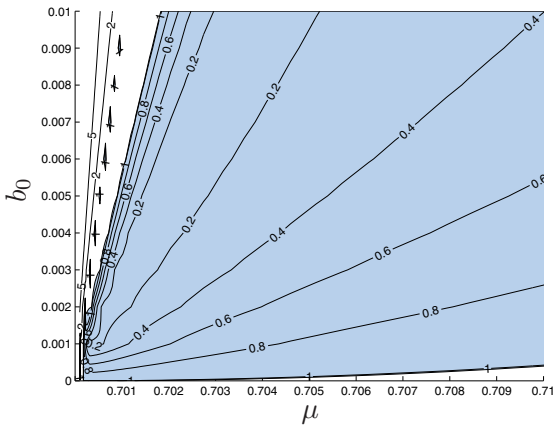


Figure 4. *The figure shows a contour plot of the magnitude of the largest Floquet multipliers of the periodic orbit, as the parameters b_0 and μ are varied. The orbit is stable when the largest Floquet multiplier has magnitude less than one, which is indicated by the shading. Remaining parameter values are $\lambda = 0.7$, $\beta = \pi/4$, $\alpha = 0.01$, so $\nu = 0$ at $\mu = 0.7$ (left-hand side of figure). This figure was produced using* DDE-BIFTOOL.

amplitude less than 1, and this region is indicated by the shading in Figure 4. Comparison with Figure 3, showing the stability as calculated analytically, shows a very good agreement between the location of the stability boundaries. The shapes of the boundaries also agrees; that is, the left-hand boundary is a straight line, whereas the lower boundary is part of a parabola. The nature of the bifurcations that occur as the boundaries are crossed also agrees with the analytical result. That is, the left-hand boundary is a period-doubling bifurcation, with a critical Floquet multiplier equal to -1, and the lower boundary is a steady-state bifurcation with a critical Floquet multiplier of $+1$.

Forward integration of the equations (3.5) also confirms the stability results. In Figure 5 we show results from such an integration. We also show the derivative of the coordinates and the feedback terms. It can be seen that as the periodic orbit is approached, the derivative of the expanding coordinate tends to λ (in this case, $\lambda = 0.7$), and the feedback terms tend toward zero.

4.1. Justification of assumptions. In section 3.3 we make a number of assumptions regarding the form of solutions to the delay differential equations. First, we assume that the "radial" direction does not affect the stability of the periodic orbits, and so we neglect this coordinate in our construction of a Poincaré map. Second, that trajectories starting near the periodic orbits will be only small perturbations from the form of solutions to the original equations without feedback. Third, we make the assumption that $f_i(T_i) \approx T_i f_i'(0)$ and $g_i(T_i) \approx T_i g_i'(0)$.

Here, we address each assumption in turn and show that our numerical results support these assumptions.

Figure 5(c) shows the feedback terms in a forward integration of equations (3.5) as the periodic orbit is approached. It can be seen here that the feedback terms corresponding to the radial direction are much smaller than the other feedback terms—on this scale they cannot be

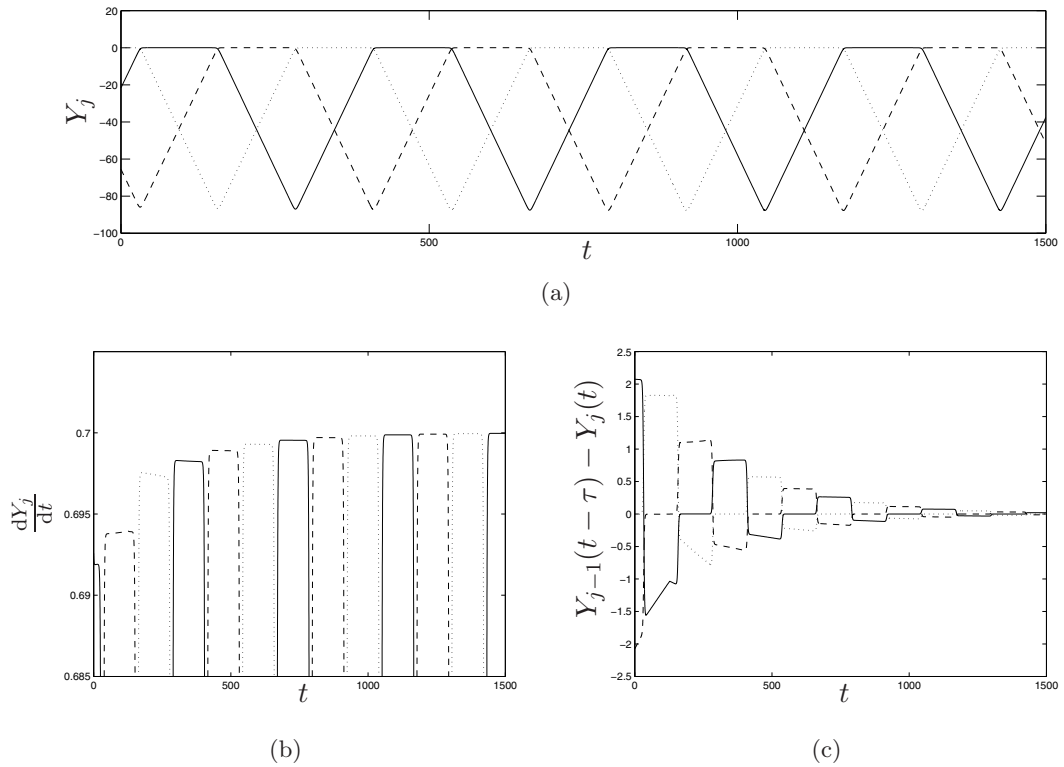


Figure 5. *Figure* (a) *shows forward integration of* (3.5)*, using* MATLAB *routine* `dde23`*. Figure* (b) *shows the derivative of the curves shown in* (a)*: it can be seen that the derivative of the expanding direction at each equilibrium tends to* 0.7 *as the periodic orbit is approached. Figure* (c) *shows the feedback terms,* $Y_{j-1}(t-\tau) - Y_j(t)$*, which clearly tends to zero as the orbit is approached. We can also see that the feedback term corresponding to the coordinates* $Y_{j-1}(t-\tau) - Y_j(t)$ *when we are near equilibria* ξ_i *is much smaller than the other feedback terms. Parameters are* $\mu = 0.7012$*,* $\lambda = 0.7$*,* $b_0 = 0.0015$*,* $\beta = \pi/4$*,* $\alpha = 0.01$*.*

distinguished from zero. Hence the effect of the feedback on the radial direction is negligible, and this assumption is justified.

Regarding the second assumption, it can be clearly seen in Figure 5(c) that the feedback terms decay to zero as the periodic orbit is approached. However, this is to be expected in the case that the periodic orbit is stable. In Figure 6 we show the results of an integration in which the periodic orbit is unstable. It can be seen from the time series in 6(a) that the trajectories still remain approximately of the form of the periodic orbit even though the trajectory is moving away. Figure 6(b) shows the feedback terms, a measure of how close the trajectory is to the periodic orbit. Although they are increasing in magnitude, they do so in the same manner one would expect for an unstable periodic orbit in ODEs. That is, by starting trajectories close enough to the periodic orbit, the feedback magnitude can be bounded above for arbitrarily long time.

The third assumption is that that $f_i(T_i) \approx T_i f_i'(0)$ and $g_i(T_i) \approx T_i g_i'(0)$. Note that in the recurrence relations (3.33), the terms in $f_i'(0)$ appear only in the combination $\lambda + f_i'(0)$ (similarly with $g_i'(0)$ in the combination $-\mu + g_i'(0)$). Therefore, we only need to show that the

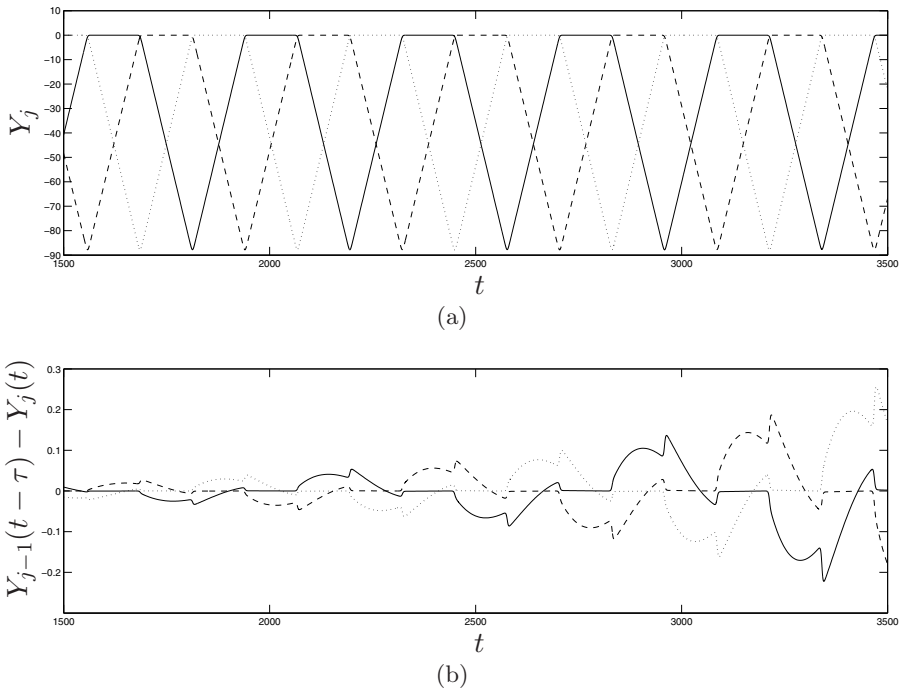


Figure 6. *The figure shows forward integration of* (3.5)*, using* MATLAB *routine* `dde23`*, for a parameter set in which the periodic orbit is unstable. Figure* (a) *shows a time series of the trajectory, and figure* (b) *shows the feedback terms,* $Y_{j-1}(t-\tau) - Y_j(t)$*. Parameters are* $\mu = 0.7012$*,* $\lambda = 0.7$*,* $b_0 = 0.008$*,* $\beta = \pi/4$*,* $\alpha = 0.01$*.*

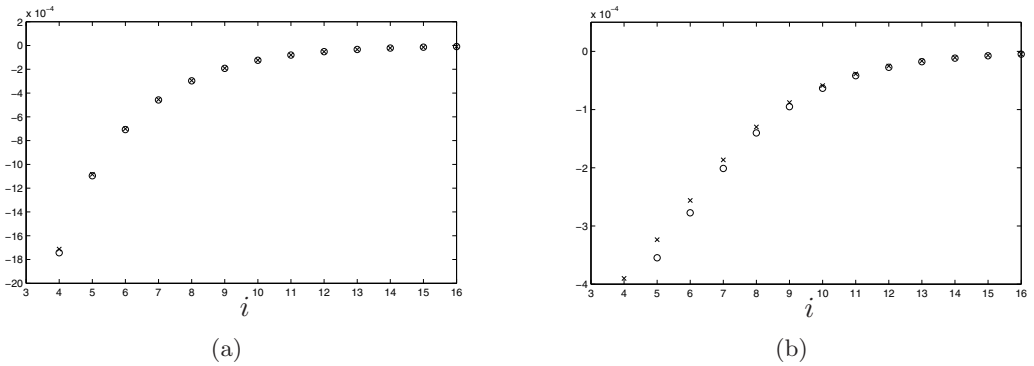


Figure 7. *The left figure shows the values of* $f_i'(0)$ *(*$\times$*'s) and* $f_i(T_i)/T_i$ *(open circles), and the right figure shows the values of* $g_i'(0)$ *(*$\times$*'s) and* $g_i(T_i)/T_i$ *(open circles) on each pass the trajectory makes past an equilibrium (i) for the integration shown in Figure* 5(a)*.*

difference between $f_i'(0)$ and $f_i(T_i)/T_i$ is much smaller than λ to justify our assumption (and similarly for the g_i). For the integration we perform for Figure 5(a) we compute the values of $f_i'(0)$, T_i, and $f_i(T_i)$ (and the corresponding values for g_i) on each pass the trajectory makes past an equilibrium. We plot these values in Figure 7. It can be seen that the difference between $f_i'(0)/T_i$ and $f_i(T_i)$ (and between $g_i'(0)/T_i$ and $g_i(T_i)$) is clearly very small and is

much less than $\lambda = 0.7$ ($\mu = 0.7012$) for this example.

5. Discussion. We have shown that a time-delayed feedback control mechanism similar to that first introduced by Pyragas can be used to stabilize periodic orbits of arbitrarily large period, specifically those resulting from a resonant bifurcation from a heteroclinic cycle. Our analytical results are based on an analysis of the stabilization of orbits near the GH cycle. These results are asymptotic; that is, they are correct in the limit of the periodic orbit being close to the heteroclinic cycle. However, in comparison with numerical results (which, conversely, are much harder to obtain when the orbit is close to the cycle due to the long period of the orbit), the results actually agree for some large(ish) range of parameters away from the bifurcation point.

It should also be possible to extend this analysis so that it applies to resonant bifurcations from higher-dimensional heteroclinic cycles. However, care may need to be taken with the transverse eigenvalues.

As the resonant bifurcation is approached, the period of the bifurcating periodic orbit grows like $1/\nu$, where ν is the bifurcation parameter. This is in contrast to the homoclinic bifurcation, in which case the bifurcating periodic orbit has a period which grows like $-\log \nu$. This difference in scaling between the growth rate of the periods of the orbits indicates that the results of adding similar time-delayed feedback near a subcritical homoclinic bifurcation may be quite different from the results given here. Work on this problem is ongoing.

Acknowledgments. The author would like to thank Mary Silber for many useful discussions regarding this work, and David Barton for assistance using DDE-BIFTOOL. Two anonymous referees also provided some helpful comments.

REFERENCES

[1] K. Pyragas, *Continuous control of chaos by self-controlling feedback*, Phys. Lett. A, 170 (1992), pp. 421–428.

[2] K. Pyragas and A. Tamaševičius, *Experimental control of chaos by delayed self-controlling feedback*, Phys. Lett. A, 180 (1993), pp. 99–102.

[3] D. J. Gauthier, D. W. Sukow, H. M. Concannon, and J. E. S. Socolar, *Stabilizing unstable periodic orbits in a fast diode resonator using continuous time-delay autosynchronization*, Phys. Rev. E, 50 (1994), pp. 2343–2346.

[4] S. Bielawski, D. Derozier, and P. Glorieux, *Controlling unstable periodic orbits by a delayed continuous feedback*, Phys. Rev. E, 49 (1994), pp. R971–R974.

[5] Th. Pierre, G. Bonhomme, and A. Atipo, *Controlling the chaotic regime of nonlinear ionization waves using the time-delay autosynchronization method*, Phys. Rev. Lett., 76 (1996), pp. 2290–2293.

[6] T. Fukuyama, H. Shirahama, and Y. Kawai, *Dynamical control of the chaotic state of the current-driven ion acoustic instability in a laboratory plasma using delayed feedback*, Physics of Plasmas, 9 (2002), pp. 4525–4529.

[7] F. W. Schneider, R. Blittersdorf, A. Förster, T. Hauck, D. Lebender, and J. Müller, *Continuous control of chemical chaos by time delayed feedback*, J. Phys. Chem., 97 (1993), pp. 12244–12248.

[8] A. Lekebusch, A. Förster, and F. W. Schneider, *Chaos control in an enzymatic reaction*, J. Phys. Chem., 99 (1995), pp. 681–686.

[9] M. E. Bleich and J. E. S. Socolar, *Controlling spatiotemporal dynamics with time-delay feedback*, Phys. Rev. E, 54 (1996), pp. R17–R20.

[10] K. Montgomery and M. Silber, *Feedback control of traveling wave solutions of the complex Ginzburg Landau equation*, Nonlinearity, 17 (2004), pp. 2225–2248.

[11] W. Lu, D. Yu, and R. G. Harrison, *Control of patterns in spatiotemporal chaos in optics*, Phys. Rev. Lett., 76 (1996), pp. 3316–3319.
[12] C. M. Postlethwaite and M. Silber, *Spatial and temporal feedback control of traveling wave solutions of the two-dimensional complex Ginzburg–Landau equation*, Phys. D, 236 (2007), pp. 65–74.
[13] K. Pyragas, *Delayed feedback control of chaos*, Phil. Trans. R. Soc. A, 364 (2006), pp. 2309–2334.
[14] J. Guckenheimer and P. Holmes, *Structurally stable heteroclinic cycles*, Math. Proc. Cambridge Philos. Soc., 103 (1988), pp. 189–192.
[15] J. Hofbauer and K. Sigmund, *Evolutionary Games and Population Dynamics*, Cambridge University Press, Cambridge, UK, 1998.
[16] M. Krupa, *Robust heteroclinic cycles*, J. Nonlinear Sci., 7 (1997), pp. 129–176.
[17] V. Kirk and M. Silber, *A competition between heteroclinic cycles*, Nonlinearity, 7 (1994), pp. 1605–1621.
[18] M. Krupa and I. Melbourne, *Asymptotic stability of heteroclinic cycles in systems with symmetry.* II, Proc. Roy. Soc. Ed. A, 134 (2004), pp. 1177–1197.
[19] P. Chossat, M. Krupa, I. Melbourne, and A. Scheel, *Transverse bifurcations of homoclinic cycles*, Phys. D, 100 (1997), pp. 85–100.
[20] S.-N. Chow, B. Deng, and B. Fielder, *Homoclinic bifurcation at resonant eigenvalues*, J. Dynam. Differential Equations, 2 (1990), pp. 177–244.
[21] C. M. Postlethwaite and J. H. P. Dawes, *A codimension-two resonant bifurcation from a heteroclinic cycle with complex eigenvalues*, Dyn. Syst., 21 (2006), pp. 313–336.
[22] M. Krupa and I. Melbourne, *Asymptotic stability of heteroclinic cycles in systems with symmetry*, Ergodic Theory Dynam. Systems, 15 (1995), pp. 121–147.
[23] M. J. Field, *Lectures on Bifurcations, Dynamics and Symmetry*, Pitman Res. Notes Math. 356, Pitman, Boston, 1996.
[24] C. M. Postlethwaite and M. Silber, *Stabilizing unstable periodic orbits in the Lorenz equations using time-delayed feedback control*, Phys. Rev. E, 76 (2007), paper 056214
[25] B. Fiedler, V. Flunkert, M. Georgi, P. Hovel, and E. Scholl, *Refuting the odd number limitation of time-delayed feedback*, Phys. Rev. Lett., 98 (2007), paper 114101.
[26] K. Engelborghs, T. Luzyanina, and G. Samaey, DDE-BIFTOOL *v.* 2.00: *A* Matlab *Package for Bifurcation Analysis of Delay Differential Equations*, Technical Report TW-330, Department of Computer Science, Katholik University Leuven, Leuven, Belgium, 2001.

SIAM J. Applied Dynamical Systems
Vol. 8, No. 1, pp. 40–75
© 2009 Society for Industrial and Applied Mathematics

Bifurcations from Synchrony in Homogeneous Networks: Linear Theory*

Martin Golubitsky† and Reiner Lauterbach‡

Abstract. A *regular* network is a network with one kind of node and one kind of coupling. We show that a codimension one bifurcation from a synchronous equilibrium in a regular network is at linear level isomorphic to a generalized eigenspace of the adjacency matrix of the network, at least when the dimension of the internal dynamics of each node is greater than 1. We also introduce the notion of a product network—a network where the nodes of one network are replaced by copies of another network. We show that generically the center subspace of a bifurcation in product networks is the tensor product of eigenspaces of the adjacency matrices of the two networks.

Key words. coupled systems, bifurcation theory, networks

AMS subject classifications. 37G10, 37G40, 34C23

DOI. 10.1137/070704873

1. Introduction.

The fundamental result in this paper, Theorem 2.7, states that the center subspace of a codimension one synchrony-breaking bifurcation in a regular network is isomorphic to a generalized eigenspace of the adjacency matrix of that network, at least when the dimension of the network nodes is greater than 1. Moreover, the action of the Jacobian matrix on that center subspace is conjugate to the action of the adjacency matrix on the associated generalized eigenspace. We now discuss the various terms in these statements.

Stewart and coworkers [13, 21, 16] developed a theory for the dynamics of a network of coupled systems based on network architecture. In this theory a network consists of a finite number of nodes or cells $C = \{c_1, \dots, c_r\}$ and a finite number of arrows. The cells represent systems of differential equations, and the arrows represent coupling between those systems. More precisely, a *cell* consists of a state space, usually taken to be a Euclidean space $\mathbf{R}^k$, and a system of differential equations associated with that cell. The system associated with cell j has the form

$$\dot{x}_j = f_j(x_j, x_{i_1}, \dots, x_{i_m}),$$

where the first argument in f_j represents the internal dynamics of the cell and the remaining variables represent coupling. Note that x_p appears in f_j only if cell p is coupled to cell j. The general theory allows for *self-coupling* (some i_q equals j) and *multiple arrows* (some subsets of indices i_q are equal). The *network* of a coupled cell system is a graph that shows which cells have the same equations and which cells are coupled to which.

*Received by the editors October 9, 2007; accepted for publication (in revised form) by T. Kaper September 12, 2008; published electronically January 9, 2009. This research was supported in part by NSF grant DMS-0604429 and ARP grant 003652-0009-2006.
http://www.siam.org/journals/siads/8-1/70487.html
†Department of Mathematics, University of Houston, Houston, TX 77204-3008. Current address: Mathematical Biosciences Institute, Ohio State University, 364 Jennings Hall, 1735 Neil Avenue, Columbus, OH 43210 (mg@mbi.osu.edu).
‡Department Mathematik, Universität Hamburg, 20146 Hamburg, Germany (lauterbach@math.uni-hamburg.de).

Examples of such networks have been considered previously by a number of authors; we mention, in particular, Pecora and Carroll [19]. These papers discuss the stability of fully synchronous solutions. In this context the tensor structure of the linearization given in (2.2) has been observed. The master stability function, which recovers the eigenvalues of the Jacobian matrix in terms of the eigenvalues of the adjacency matrix, is related to our Lemma 2.3. Pecora and Carroll use this structure to discuss stability, that is, to determine the maximum real part of the eigenvalues or, in a more general context, the Liapunov exponents. In this paper we use the tensor structure to discuss the surprisingly nontrivial Jordan structure of the Jacobian restricted to center subspaces that is sometimes forced by network architecture. This issue is relevant for bifurcations from synchrony. For example, it has been shown in [10, 8] that the length of Jordan blocks determines the amplitude of certain bifurcating solutions obtained by Hopf bifurcation in feed-forward chains.

It is well known in the pattern formation literature that symmetry-breaking bifurcations from a stable fully symmetric equilibrium in a single reaction-diffusion equation are not possible. Such bifurcations can occur only in systems consisting of several equations. For example, Ermentrout and Lewis [9] show the need for three species to obtain a Turing symmetry-breaking Hopf bifurcation. A similar issue occurs in bifurcation from synchrony in coupled cell systems, where the role of the number of equations in reaction-diffusion systems is replaced by the dimensionality k of the internal cell dynamics. In section 7 we show that when the eigenvalues of the adjacency matrix are real, then all such steady-state synchrony-breaking bifurcations are possible when $k \geq 2$ (Theorem 7.1) and all such Hopf bifurcations are possible when $k \geq 3$ (Theorem 7.2). The last theorem (proved by Amit Vutha) generalizes, in the context of coupled cell systems, the observation in [9].

In this paper we consider only networks in which all cells are *identical*; that is, the state spaces are all identical ($\mathbf{R}^{k_j} = \mathbf{R}^k$ for all j), and the systems of differential equations are all equal ($f_j = f$ for all j); that is,

$$\dot{x}_j = f(x_j, x_{i_1}, \ldots, x_{i_m}). \tag{1.1}$$

Definition 1.1. *A* homogeneous *coupled cell network is a network in which all cells are identical.*

The diagonal $\Delta_0 = \{x : x_1 = \cdots = x_r\}$ consists of points where the coordinates in each cell are identical and is always flow-invariant in homogeneous cell systems. Solutions in Δ_0 can be found by solving the single system of differential equations

$$\dot{y} = f(y; y, \ldots, y). \tag{1.2}$$

Synchronous equilibria, zeros of f in (1.2), occur naturally in homogeneous cell systems. In particular, it is reasonable to assume that there is an equilibrium of an admissible vector field in Δ_0 which we can assume, without loss of generality, is at the origin.

A homogeneous network can be depicted by a graph, which consists of nodes and arrows. This graph has identical node symbols but can have different types of arrows. However, for the network to be homogeneous each node must receive the same number of inputs of each arrow type. Examples of homogeneous three-cell rings are given in Figure 1. Observe that in the third figure there are two types of coupling indicated by solid and dashed arrows.

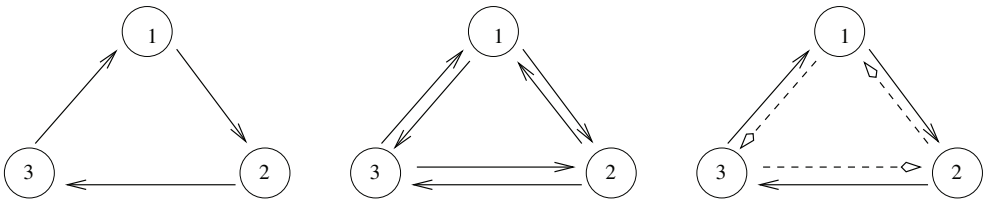


Figure 1. *Three-cell homogeneous networks: (left) unidirectional ring; (center) bidirectional ring; (right) directionally coupled ring.*

Each network architecture has associated with it a class of admissible vector fields, as described above. The classes of admissible vector fields associated with each of the networks in Figure 1 are given by

$$\begin{array}{rcl} \dot{x}_1 &=& f(x_1,x_3),\\ \dot{x}_2 &=& f(x_2,x_1),\\ \dot{x}_3 &=& f(x_3,x_2), \end{array} \qquad \begin{array}{rcl} \dot{x}_1 &=& g(x_1,\overline{x_2,x_3}),\\ \dot{x}_2 &=& g(x_2,\overline{x_3,x_1}),\\ \dot{x}_3 &=& g(x_3,\overline{x_1,x_2}), \end{array} \qquad \begin{array}{rcl} \dot{x}_1 &=& h(x_1,x_2,x_3),\\ \dot{x}_2 &=& h(x_2,x_3,x_1),\\ \dot{x}_3 &=& h(x_3,x_1,x_2). \end{array}$$

The overbar in g indicates that the two couplings to a given cell are identical and, because of this, the coupling arguments can be interchanged; that is, $g(x,y,z) = g(x,z,y)$. The theory in [21, 16] explains how the assignment of admissible vector fields to network architecture can be done in a functorial way (the admissible vector fields are just those that commute with the symmetry groupoid of the graph).

Definition 1.2. *A homogeneous network is* regular *if all couplings are of the same type. The* valency *of a regular network is the number of arrows that input to each cell.*

Note that the first two networks in Figure 1 are regular, whereas the third is not.

This paper considers the structure of synchrony-breaking bifurcations from a fully synchronous equilibrium in a homogeneous cell system. More specifically, suppose that

$$\dot{x} = F(x)$$

is an admissible system for a homogeneous network with $F(0) = 0$. By a *synchrony-breaking* bifurcation we mean that at the origin there is a critical eigenvector in the center subspace E^c of the Jacobian $J = (dF)_0$ that is not in Δ_0. One goal of this paper is to describe the Jordan normal form that $J|E^c$ takes at a generic (*codimension one*) synchrony-breaking bifurcation from a synchronous equilibrium. We will show that for regular networks (section 2) synchrony-breaking bifurcations can be identified with the Jordan structure of the adjacency matrix associated with the network architecture.

Definition 1.3. *The* adjacency matrix *of a regular network is the matrix* $A = [a_{ij}]$*, where* a_{ij} *is the number of arrows from cell* j *to cell* i*.*

Note that the each row sum of the adjacency matrix equals the network valency.

We prove (Theorem 2.7) that when the dimension of the internal dynamics of each cell k is at least 2, then the center subspace at bifurcation is isomorphic to a generalized eigenspace of the adjacency matrix. This theorem generalizes observations by Leite and Golubitsky [18], who studied bifurcations in three-cell systems. There is a striking analogy between eigenspaces

of the adjacency matrix in synchrony-breaking bifurcations in regular cell networks and irreducible representations in symmetry-breaking bifurcations for equivariant dynamical systems (see Remark 2.9).

Leite and Golubitsky [18] and Elmhirst and Golubitsky [8] analyzed the codimension one steady-state and Hopf synchrony-breaking bifurcations in valency one and two regular three-cell networks. It was shown in [18] that there are 34 different such networks and that all manner of center subspaces are possible in codimension one bifurcations. For simplicity assume that the dimension k of the internal dynamics phase space is 1. Then [18] shows that there are networks where the center subspace E^c at a synchrony-breaking bifurcation is one-dimensional, or is two-dimensional with two independent eigenvectors, or is two-dimensional with only one independent eigenvector (the *nilpotent* case). It is even possible for a codimension one bifurcation to have two simultaneously critical eigenvectors, with one in the subspace Δ_0 and one not in that subspace (see network 12 in [18]). In particular, the Jordan structure of the Jacobian at a synchronous equilibrium in a regular system can be quite complicated in codimension one bifurcations.

After discussing the relationship between synchrony-breaking bifurcations in regular networks and symmetry-breaking bifurcations in equivariant bifurcation theory, we focus on networks that are networks of networks. The product of two regular networks is defined in section 3, and such networks are homogeneous. Product networks have two types of coupling (one from each regular network) and two adjacency matrices. Theorem 5.5 states that generically the center subspaces of codimension one synchrony-breaking bifurcations for product networks are isomorphic to tensor products of eigenspaces of adjacency matrices, one from each of the regular networks.

The remainder of the paper is devoted to discussions of partial results for the nonlinear bifurcations whose linear parts have been described in the beginning sections. In particular, sections 6 and 8 discuss examples of (nonlinear) synchrony-breaking bifurcations for regular networks and for product networks. A discussion of a general strategy for analyzing synchrony-breaking bifurcations is given in section 9.

2. Center subspaces in regular networks. In this section we discuss the generic structure of critical eigenspaces of regular networks. We fix a regular network with r cells and, as before, assume that an admissible vector field F has a synchronous equilibrium at the origin; that is, $F(0) = 0$. Suppose that the internal dynamics of each cell has dimension $k = 1$. Then

$$(dF)_0 = \alpha I + \beta A,$$

where A is the adjacency matrix. Every eigenvalue of $J = (dF)_0$ has the form $\alpha + \beta\mu$, where μ is an eigenvalue of A. It also follows that generalized eigenspaces of J are generalized eigenspaces of A. In this section, we will prove that generically center subspaces at synchrony-breaking bifurcations are isomorphic to the real parts of generalized eigenspaces of A, but this conclusion is valid in general only when $k > 1$.

Remark 2.1. Suppose $k = 1$. Then the real parts of the eigenvalues of J are just $\alpha + \mathrm{Re}(\mu_j)\beta$. So, if two of the A eigenvalues have the same real part, say $\mathrm{Re}(\mu_1) = \mathrm{Re}(\mu_2)$, then the real parts of $\alpha + \mu_1\beta$ and $\alpha + \mu_2\beta$ will be equal for all α and β. As an example consider

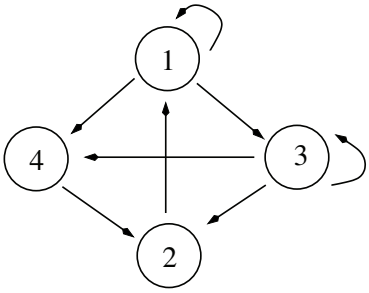


Figure 2. *Four-cell network whose adjacency matrix has eigenvalues* $2, 0, \pm i$.

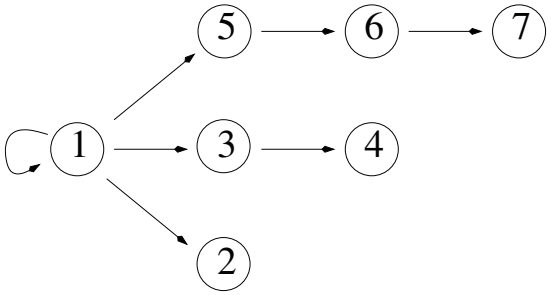


Figure 3. *Branching feed-forward network that can exhibit the degenerate Jordan normal form* J *in* (2.1) *in a codimension one bifurcation from a synchronous equilibrium.*

the four-cell network shown in Figure 2 whose adjacency matrix is

$$A = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix}.$$

The eigenvalues of A are $2, 0, \pm i$. At a codimension one steady-state synchrony-breaking bifurcation, the center subspace is three-dimensional—the sum of the eigenspaces corresponding to the eigenvalues 0 and $\pm i$.

Remark 2.2. Every Jordan structure J associated to a 0 eigenvalue can be the Jordan structure of a real eigenvalue of an adjacency matrix of a regular network. For example, suppose that

$$J = \left[\begin{array}{c|cc|cc|c} 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right]. \tag{2.1}$$

Then an example of a network whose adjacency matrix has the the Jordan structure J (corresponding to the zero eigenvalue) is given in Figure 3. We see that in the general case we can always choose that network to be a branching network of feed-forward subnets.

Next assume that the internal dynamics of each cell is k-dimensional, where $k > 1$. At a fully synchronous state, the Jacobian is determined by two $k \times k$ matrices: the linearized internal dynamics α and the linearized coupling β. These matrices are found by differentiating F, as defined in (1.1), with respect to the first two variables. Using the discussion of linearizations given in [18, section 3], we have

$$(dF)_0 = \alpha \otimes \mathbf{I}_r + \beta \otimes A. \tag{2.2}$$

To see this, first consider the linearization when the coupling matrix β is zero. Then the linearization has the form $\alpha \otimes \mathbf{I}_r$ as in [18]. Assume next that the linearized internal dynamics α is zero to obtain the second term. Again this proceeds as in [18].

Let $\mu_1, \ldots, \mu_s$ be the distinct eigenvalues of A. Moreover, let

$$M_{\mu_i} = \alpha + \mu_i \beta.$$

Lemma 2.3. *The kr eigenvalues of the Jacobian J are the union of the eigenvalues of the $k \times k$ matrices M_{μ_j} for $1 \le j \le s$. Specifically, suppose $v \in \mathbf{C}^r$ is an eigenvector of A. Then*

$$J(u \otimes v) = (M_\mu u) \otimes v. \tag{2.3}$$

Therefore, if $u \in \mathbf{C}^k$ is an eigenvector of M_μ, then $u \otimes v$ is an eigenvector of J.

Proof. Suppose that $\mu \in \mathbf{C}$ is an eigenvalue of A with eigenvector $v \in \mathbf{C}^r$. Let

$$Y_v = \{u \otimes v : u \in \mathbf{C}^k\} \subset \mathbf{C}^k \otimes \mathbf{C}^r.$$

We claim that the subspace Y_v is J-invariant and that $J|Y_v = M_\mu$. It then follows that the k eigenvalues of $J|Y_v$ are eigenvalues of J. To verify the claim, calculate

$$\begin{aligned} J(u \otimes v) &= (\alpha u) \otimes v + (\beta u) \otimes Av \\ &= (\alpha u) \otimes v + \mu(\beta u) \otimes v \\ &= ([\alpha + \mu\beta]u) \otimes v \\ &= (M_\mu u) \otimes v, \end{aligned}$$

thus verifying (2.3). Specifically, $J|Y_v$ is the matrix M_μ. The fact that $u \otimes v$ is an eigenvector of J whenever u is an eigenvector of M_μ follows directly from (2.3). ∎

Our primary goal is to classify the center subspaces that occur at codimension one bifurcations in regular networks. Let

$$\rho_1, \ldots, \rho_p \quad \text{and} \quad \eta_1, \ldots, \eta_q \tag{2.4}$$

be distinct eigenvalues of A. The ρ_j are the distinct real eigenvalues, and the η_i are the distinct complex eigenvalues of A whose imaginary parts are positive.

Definition 2.4. *A pair of real $k \times k$ matrices α, β is a* codimension one pair *if the real parts of the eigenvalues of the matrices $M_{\rho_1}, \ldots, M_{\rho_p}, M_{\eta_1}, \ldots, M_{\eta_q}$ are distinct (except for complex conjugate pairs of eigenvalues of the M_{ρ_j}). The set of all codimension one pairs is denoted by $\mathcal{M}_2(k)$.*

It is clear that $\mathcal{M}_2(k)$ is an open set of $L(k)^2$, where $L(k)$ is the space of real $k \times k$ matrices. In Proposition 2.5 we prove that when $k > 1$ the set $\mathcal{M}_2(k)$ is also a dense subset. Therefore, the generic codimension one bifurcation will occur with a Jacobian J associated to a codimension one pair, and J will have a center subspace that corresponds to one critical eigenvalue.

Proposition 2.5. *For $k > 1$, $\mathcal{M}_2(k) \subset L(k)^2$ is open and dense. For $k = 1$, $\mathcal{M}_2(k)$ is open and dense when the eigenvalues of A in (2.4) have distinct real parts.*

The second main result in this section states that over the complex numbers generically the generalized eigenspace of the Jacobian matrix J is isomorphic to a generalized eigenspace of the adjacency matrix A. We begin with a straightforward lemma.

Lemma 2.6. *Let $\mu \in \mathbf{C}$ be an eigenvalue of A, and let $\mathcal{G}_A(\mu) \subset \mathbf{C}^r$ denote the associated complex generalized eigenspace. Then $\mathbf{C}^k \otimes \mathcal{G}_A(\mu)$ is invariant under J.*

Proof. Let $A_\mu = A - \mu\mathbf{I}$ and write

$$J = \alpha \otimes \mathbf{I}_r + \beta \otimes A = M_\mu \otimes \mathbf{I}_r + \beta \otimes A_\mu.$$

Clearly both summands leave the space $\mathbf{C}^k \otimes \mathcal{G}_A(\mu)$ invariant, and hence so does J. ∎

Let $\sigma \in \mathbf{C}$ be an eigenvalue of J, and let $\mathcal{G}_J(\sigma)$ denote the generalized eigenspace of $J_\sigma = J - \sigma\mathbf{I}$ restricted to the invariant subspace $\mathbf{C}^k \otimes \mathcal{G}_A(\mu)$.

Proposition 2.5 allows us to assume that all eigenvalues of M_μ are simple and have distinct real parts. It follows that for each μ there is a basis of eigenvectors $u_1, \ldots, u_k$ of M_μ in $\mathbf{C}^k$. In order to state the next theorem we must impose further genericity assumptions on the coupling matrix β as follows. For each eigenvalue σ of M_μ with associated eigenvector u write

$$\beta u = \sum_j \zeta_j u_j. \tag{2.5}$$

Assume that all of the

$$\zeta_j \neq 0 \quad \text{for } j = 1, \ldots, k. \tag{2.6}$$

Observe that the set of β that satisfy (2.5), (2.6) for a fixed σ is open and dense. Since there are a finite number of σ, the set of β that satisfy (2.5), (2.6) for all σ is also open and dense.

Theorem 2.7. *Assume (2.5), (2.6). Let $\mu \in \mathbf{C}$ be an eigenvalue of A, and let $\sigma \in \mathbf{C}$ be a simple eigenvalue of M_μ. Then there exists an isomorphism*

$$\eta : \mathcal{G}_A(\mu) \to \mathcal{G}_J(\sigma)$$

such that on $\mathcal{G}_A(\mu)$

$$J_\sigma \circ \eta = \eta \circ A_\mu. \tag{2.7}$$

Remark 2.8. A similar situation occurs if we study synchrony-breaking bifurcations from a fully synchronous limit cycle. Observe that each time dependent matrix in the variational equation has a tensor structure as in (2.2). It follows that the autonomous Floquet system keeps this tensor product structure, which leads to invariant subspaces as in Lemma 2.6.

Since the critical Floquet multiplier corresponds to the trivial eigenvalue of the adjacency matrix, this tensor product structure is inherited by the Poincaré map of the limit cycle. If the linearization of the Poincaré map satisfies the genericity conditions (2.5) and (2.6), then the result of Theorem 2.7 holds in this context as well. Although we do not have a formal proof for this last statement, we believe it is true.

Elmhirst and Golubitsky [8] proved that the amplitude growth of periodic solutions obtained from codimension one synchrony-breaking Hopf bifurcations in feed-forward networks can be unusual. Recently, Comanici [4] showed that a similar effect can be found in codimension one period-doubling bifurcations in maps. These results strongly suggest that the codimension one bifurcations from synchronous limit cycles in feed-forward networks will lead to solutions with interesting amplitude growth. However, here we have not attempted to work out the nonlinear theory that is needed to establish such a result.

Remark 2.9. Proposition 2.5 and Theorem 2.7 provide an analogy between synchrony-breaking bifurcations in regular networks and symmetry-breaking bifurcations in equivariant theory; however, to see this analogy clearly, we must discuss center subspaces of the real Jacobian J, not just eigenspaces of the complexification of J. In this analogy, center subspaces of J associated with real eigenvalues of A correspond to absolutely irreducible representations, and center subspaces of J associated with complex eigenvalues of A correspond to non–absolutely irreducible representations.

Proposition 2.5 implies that generically, at a synchrony-breaking bifurcation, the center subspace of J is determined by the eigenspace corresponding to one eigenvalue μ of A and one simple critical eigenvalue σ of M_μ. It then follows, as we explain next, that the center subspaces of J, and hence the codimension one synchrony-breaking bifurcations, divide into three types, which are described best by considering separately the cases when μ is real and when μ is not real.

(a) μ is real. In this case the matrix M_μ is an arbitrary real $k \times k$ matrix. As J varies with a parameter, so does M_μ, and a critical eigenvalue σ of M_μ is either 0 or purely imaginary.
Suppose that $\sigma = 0$. In this case all of the constructions in Theorem 2.7 are real, and the real subspaces $\mathcal{G}_J(0)$ and $\mathcal{G}_A(\mu)$ are isomorphic. Hence $E^c(J) \cong \mathcal{G}_A(\mu)$.
Suppose that σ is purely imaginary. Theorem 2.7 states that the complex vector space $\mathcal{G}_J(\sigma)$ and the complexification of $\mathcal{G}_A(\mu)$ are isomorphic. However, the fact that M_μ is a real matrix implies that $\overline{\sigma}$ is also a critical eigenvalue of M_μ. Since the center subspace $E^c(J)$ is the real part of $\mathcal{G}_J(\sigma) \oplus \mathcal{G}_J(\overline{\sigma})$, it follows that $E^c(J) \cong \mathcal{G}_A(\mu) \oplus \mathcal{G}_A(\mu)$.

(b) μ is not real. In this case the matrix M_μ is an arbitrary complex $k \times k$ matrix. As J varies with a parameter, so does M_μ, and a critical eigenvalue σ of M_μ is generically purely imaginary. Note, however, that generically $\overline{\sigma}$ is not an eigenvalue of M_μ, though it is always an eigenvalue of $M_{\overline{\mu}}$. Hence $E^c(J) \cong \mathrm{Re}(\mathcal{G}_A(\mu) \oplus \mathcal{G}_A(\overline{\mu}))$.

In Γ-equivariant bifurcation theory, steady-state bifurcations correspond to absolutely irreducible representations, and Hopf bifurcations correspond to Γ-simple representations (see [15]). Recall that Γ-simple representations are either the direct sum of two copies of an absolutely irreducible representation or a single copy of a nonabsolutely irreducible representation. As we have seen, generically there can be synchrony-breaking steady-state and Hopf bifurcations in regular networks that correspond to real eigenvalues of A and Hopf bifurcations

that correspond to complex eigenvalues of A.

As an immediate consequence of Remark 2.9 we have the following characterization of center subspaces for codimension one bifurcations in regular networks.

Corollary 2.10. *Generically in α, β, the center subspace $E^c(J)$ at a codimension one steady-state bifurcation in a regular network is isomorphic to $\mathcal{G}_A(\mu)$ for some real eigenvalue μ of A. At a Hopf bifurcation, $E^c(J)$ is isomorphic to $\mathcal{G}_A(\mu) \oplus \mathcal{G}_A(\mu)$ if μ is real or to $\mathrm{Re}(\mathcal{G}_A(\mu) \oplus \mathcal{G}_A(\overline{\mu}))$ if μ is not real.*

Proof of Proposition 2.5. To prove the proposition we must establish that a certain finite number of sets in the pairs α, β are open and dense. The intersection of these sets will be open and dense and contain the desired pairs of matrices.

To begin, take two eigenvalues ν_a, ν_b of A from the set (2.4) and consider the two sets of matrix pairs α, β:

$$\begin{aligned} D_a &= \{(\alpha,\beta) \in L(k)^2 : M_{\nu_a} \text{ has eigenvalues with distinct real parts}\}, \\ E_{a,b} &= \{(\alpha,\beta) \in D_a \cap D_b : M_{\nu_a}, M_{\nu_b} \text{ have eigenvalues with different real parts}\}. \end{aligned} \tag{2.8}$$

Note that the sets D_a and $E_{a,b}$ are open and that $\mathcal{M}_2(k)$ consists of matrices in the intersection of the sets $E_{a,b}$. Moreover, if the sets D_a are dense in $L(k)^2$, then proving that $E_{a,b}$ is dense in $D_a \cap D_b$ suffices to prove that $E_{a,b}$ is dense in $L(k)^2$.

From the point of view of this proposition the distinct eigenvalues of the adjacency matrix A in (2.4) just form a finite set of real and complex numbers. This set can be shifted by any real number c, since

$$M_\mu = (\alpha + c\beta) + (\mu - c)\beta.$$

It follows that the sets in (2.8) are open and dense for ν_a, ν_b if and only if they are open and dense for the set $\nu_a - c$, $\nu_b - c$. Similarly, we can scale the set of eigenvalues, since

$$M_\mu = \alpha + (c\mu)\left(\frac{1}{c}\beta\right)$$

for any real constant $c \neq 0$.

Thus, when we consider the sets D_a we may assume, without loss of generality, that $\nu_a = 0$ or $\nu_a = i$. In the first case, $M_{\nu_a} = \alpha$. Using real Jordan normal forms, it is easy to show that the small perturbations of α have eigenvalues all of whose real parts are unequal. In the second case, $M_{\nu_a} = \alpha + i\beta$, which is an arbitrary complex matrix. Now using the basis whose existence is guaranteed by complex Jordan normal form, we can see that small perturbations of α and β will move, arbitrarily, the real parts of the eigenvalues of M_{ν_a}. Hence, in all cases, D_a is dense.

To prove the density of the sets $E_{a,b}$ we need to consider pairs of eigenvalues ν_a and ν_b in (2.4). These eigenvalues either have real parts unequal or real parts equal. Therefore, we can shift and scale these eigenvalues so that they are equal to one of the pairs:

$$\begin{aligned} (\nu_a, \nu_b) &\qquad \mathrm{Re}(\nu_a) = 0;\ \mathrm{Re}(\nu_b) \neq 0, \\ (i, i\omega) &\qquad 0 \leq \omega < 1. \end{aligned} \tag{2.9}$$

The sets $E_{a,b}$ can be shown to be dense by explicit constructions in each case. Since α and β are in $D_a \cap D_b$, the matrices M_{ν_a} and M_{ν_b} have distinct eigenvalues and a basis of eigenvectors.

Case $\mathrm{Re}(\nu_a) = 0$*;* $\mathrm{Re}(\nu_b) \neq 0$. In this case we keep α fixed and perturb β to $\beta + \varepsilon I_k$. This perturbation shifts the eigenvalues of M_{ν_a} by $\varepsilon\nu_a$, which is purely imaginary; hence the perturbation keeps the real parts of the eigenvalues of this matrix fixed. This perturbation also shifts the eigenvalues of M_{ν_b} by $\varepsilon\nu_b$. Since $\mathrm{Re}(\nu_b) \neq 0$, this perturbation shifts the real parts of the eigenvalues by $\varepsilon\,\mathrm{Re}(\nu_b)$. Thus, the sets E_{i,ν_b} are dense.

Case $\nu_a = i$*;* $\nu_b = i\omega$*;* $0 \leq \omega < 1$. When $k = 1$ the real parts of the eigenvalues of M_{ν_a} and M_{ν_b} are always equal, and separation of the real parts of the eigenvalues is impossible. So we assume $k \geq 2$.

The matrix $M_{\nu_a} = \alpha + i\beta$ is an arbitrary complex matrix. We can assume, as above, that the eigenvalues of M_{ν_a} are distinct and, by an arbitrarily small perturbation, that these eigenvalues are all complex. Fix an eigenvalue λ_a of M_{ν_a} with eigenvector $y = y_R + iy_I$. We can also assume, after an arbitrarily small perturbation, that y_R and y_I are linearly independent so that the subspace $Y = \mathbf{R}\{y_R, y_I\}$ is two-dimensional. Let $p + iq$ be a perturbation matrix for which

$$(p + iq)y = 0. \tag{2.10}$$

Note that $(M_{\nu_a} + (p+iq))y = \lambda_a y$. So the real part of the perturbed eigenvalue of $M_{\nu_a} + (p+iq)$ is the same as the real part of the eigenvalue λ_a of M_{ν_a}.

Next assume that $M_{\nu_b} = \alpha + i\omega\beta$ (where $\omega \neq 1$) has an eigenvalue λ_b with eigenvector $z = z_R + iz_I$ and $\mathrm{Re}(\lambda_a) = \mathrm{Re}(\lambda_b)$. Assume that p, q also satisfy

$$(p + i\omega q)z = \varepsilon z \tag{2.11}$$

for some nonzero $\varepsilon \in \mathbf{R}$. Equation (2.11) implies that the real part of the eigenvalue of $M_{\nu_b} + (p + i\omega q)$ associated with the eigenvector z is $\lambda_b + \varepsilon$. Thus the sets $E_{i,\omega i}$ are dense if the perturbation constraints (2.10) and (2.11) can hold simultaneously.

Assume that p is invertible, and let $B = p^{-1}q$. Then (2.10) can be written as

$$(I + iB)y = 0$$

or as $By = iy$. Since y_R and y_I are linearly independent, we can choose B to satisfy $By_R = -y_I$ and $By_I = y_R$. Note that the eigenvalues of B on Y are $\pm i$. It follows that we can choose B so that $I + \omega iB$ is invertible (both on and off Y), since $\omega \neq \pm 1$. Next observe that (2.11) reduces to

$$p^{-1}z = \frac{1}{\varepsilon}(I + i\omega B)z.$$

Since $I + i\omega B$ is invertible, we can choose an invertible p to satisfy this equation. Then we set $q = pB$.

Note that the only place where we needed to assume $k > 1$ was in the second case, the case where two eigenvalues of the adjacency matrix A have equal real parts. ■

Proof of Theorem 2.7. Proposition 2.5 allows us to assume that all eigenvalues of M_μ are simple and have distinct real parts. Then for each μ there is a basis of eigenvectors $u_1, \ldots, u_k$ of M_μ in $\mathbf{C}^k$. By assumptions (2.5) and (2.6) we assume that all of the ζ_j are nonzero.

For the remainder of the proof we fix the eigenvalue μ. Let σ be an eigenvalue of M_μ and hence an eigenvalue of J. Let $\mathcal{G}_J(\sigma) \subset \mathbf{C}^k \otimes \mathcal{G}_A(\mu)$ denote the generalized eigenspace corresponding to the eigenvalue σ of J. We prove the theorem by constructing the isomorphism η.

In fact, we need only show that η is an injection. If it is, then $\dim \mathcal{G}_J(\sigma) \geq \dim \mathcal{G}_A(\mu)$. Indeed, if some η were not a surjection, then it would follow that

$$\sum_{i=1}^{k} \dim \mathcal{G}_J(\sigma_i) > k \dim \mathcal{G}_A(\mu),$$

where $\sigma_1, \ldots, \sigma_k$ are the eigenvalues of M_μ. However, since

$$\mathcal{G}_J(\sigma_1) \oplus \cdots \oplus \mathcal{G}_J(\sigma_k) = \mathbf{C}^k \otimes \mathcal{G}_A(\mu),$$

it follows that

$$\sum_{i=1}^{k} \dim \mathcal{G}_J(\sigma_i) = k \dim \mathcal{G}_A(\mu).$$

Hence, η must be surjective as well as injective. Observe that as a consequence of the dimension count, we have also proved that all of the eigenvalues of J are the ones that came from the M_μ's in Lemma 2.3. Since generically the real parts of all of the eigenvalues σ are distinct (Proposition 2.5), it follows that $E^c = \mathrm{Re}(\mathcal{G}_J(\sigma))$.

Next we choose a basis in $\mathcal{G}_A(\mu)$ for which A is in Jordan normal form. Let

$$\mathcal{G}_A^j(\mu) = \ker(A_\mu^j)$$

for $j = 1, \ldots, s$, where s is the smallest integer for which $\mathcal{G}_A^s(\mu) = \mathcal{G}_A(\mu)$. Let $P^1 = \mathcal{G}_A^1(\mu)$ consist of the eigenvectors of A with eigenvalue μ, and choose subspaces P^j for $j = 2, \ldots, s$ so that

$$\mathcal{G}_A^j(\mu) = P^j \oplus \mathcal{G}_A^{j-1}.$$

Let $d_j = \dim P^j$ and note that $d_s \leq \cdots \leq d_1$. Moreover, we can choose bases $\{v_{j,1}, \ldots, v_{j,d_j}\}$ of P^j so that

$$A_\mu(v_{j,m}) = v_{j-1,m}.$$

We define η on Jordan blocks of A restricted to $\mathcal{G}_A(\mu)$. More precisely, each string of generalized eigenvectors $v_{1,m}, v_{2,m}, \ldots, v_{\ell,m}$ (where $d_{\ell+1} < m$) is a basis for a Jordan block of A. The construction of η proceeds inductively on each string.

To simplify the notation denote the string with single subscripts by $V_1, \ldots, V_\ell$. We claim that there exist vectors $W_1, \ldots, W_\ell$ in $\mathbf{C}^k \otimes \mathcal{G}_A(\mu)$ such that $W_1 \neq 0$, $J_\sigma(W_1) = 0$, and

$$J_\sigma(W_j) = W_{j-1}$$

for $2 \leq j \leq \ell \leq s$. We then define $\eta(V_j) = W_j$ and extend by linearity. By construction η satisfies (2.7). Moreover, the set of W_j so constructed is linearly independent. Note that none of the W_j equals zero. Next suppose that

$$W = \rho_1 W_1 + \cdots + \rho_s W_s = 0,$$

where $\rho_j \in \mathbf{C}$. It follows that $J_\sigma^{s-1}(W) = \rho_s W_1 = 0$. Hence $\rho_s = 0$ since $W_1 \neq 0$. Since $\rho_s = 0$, it follows that $J_\sigma^{s-2}(W) = \rho_{s-1} W_1 = 0$. Hence, $\rho_{s-1} = 0$. Proceed inductively.

It is also straightforward to show that the union of vectors W_j that come from all the different strings are also linearly independent. Thus the map $\eta : \mathcal{G}_A(\mu) \to \mathcal{G}_J(\sigma)$ is an injection. We now continue with the construction of the W_j.

Since the eigenvalues of M_μ are simple, we can write

$$\mathbf{C}^k = \mathbf{C}\{U\} \oplus Q, \tag{2.12}$$

where $M_\mu U_1 = \sigma U_1$ is an eigenvector and Q is the sum of the eigenspaces corresponding to eigenvalues other than σ. Let $M_{\mu,\sigma} = M_\mu - \sigma \mathbf{I}_k$. Observe that $M_{\mu,\sigma} : \mathbf{C}^k \to Q$ and $M_{\mu,\sigma}$ restricted to Q is invertible.

Let U be an eigenvector of M_μ with eigenvalue σ. We claim that we can choose $U_1^j = \gamma_j U$ where $\gamma_j \in \mathbf{C}$ is nonzero and $U_2^j, \ldots, U_j^j \in Q$ so that W_j has the form

$$W_j = U_1^j \otimes V_j + \cdots + U_j^j \otimes V_1. \tag{2.13}$$

To begin, define $W_1 = U_1^1 \otimes V_1$, where $U_1^1 = U$. Since the vector V_1 is an eigenvector of A with eigenvalue μ, Lemma 2.3 states that $W_1 \neq 0$ is an eigenvector of J with eigenvalue σ; that is, $J_\sigma(W_1) = 0$.

Next we find $\gamma_2 \in \mathbf{C}$, $U_2^2 \in Q$, and W_2. Compute

$$\begin{aligned} J_\sigma(W_2) &= J_\sigma(U_1^1 \otimes V_2 + U_2^2 \otimes V_1) \\ &= \gamma_2(M_{\mu,\sigma}U) \otimes V_2 + \gamma_2(\beta U) \otimes (A_\mu V_2) + (M_{\mu,\sigma}U_2^2) \otimes V_1 + (\beta U_2^2) \otimes (A_\mu V_1) \\ &= \gamma_2(\beta U) \otimes V_1 + (M_{\mu,\sigma}U_2^2) \otimes V_1 \end{aligned}$$

since $M_{\mu,\sigma}U = 0$, $A_\mu V_1 = 0$, and $A_\mu V_2 = V_1$. It follows that to solve $J_\sigma(W_2) = W_1 = U \otimes V_1$ we must solve the equation

$$\gamma_2 \beta U + M_{\mu,\sigma}U_2^2 - U = 0 \tag{2.14}$$

for $U_2^2 \in \mathbf{C}^k$ and $\gamma_2 \in \mathbf{C}$. Using the genericity condition (2.5), we can set $\gamma_2 = \zeta_j^{-1}$, where u_j is the eigenvector in (2.5) that corresponds to the eigenvalue σ. It follows that (2.14) can now be written in the form

$$M_{\mu,\sigma}U_2^2 = q,$$

where $q \in Q$. Since $M_{\mu,\sigma}$ is invertible on Q, this equation can be solved for $U_2^2 \in Q$.

Next, we assume that W_{j-1} has been defined in $\mathbf{C}^k \otimes \mathcal{G}_A(\mu)$ so that $J_\sigma(W_{j-1}) = W_{j-2}$. To complete the induction, we must find $W_j \in \mathbf{C}^k \otimes \mathcal{G}_A(\mu)$ and $\gamma_j \in \mathbf{C}$ so that $J_\sigma(W_j) = W_{j-1}$. Let W_j have the form in (2.13). Compute

$$\begin{aligned} J_\sigma(W_j) &= J_\sigma(U_1^\ell \otimes V_j + \cdots + U_j^j \otimes V_1) \\ &= J_\sigma(U_1^j \otimes V_j) + \cdots + J_\sigma(U_j^j \otimes V_1) \\ &= \gamma_j(\beta U) \otimes V_{j-1} + \textstyle\sum_{p=2}^{j-1}((M_{\mu,\sigma}U_p^j) \otimes V_{j+1-p} + (\beta U_p^j) \otimes V_{j-p}) + (M_{\mu,\sigma}U_j^j) \otimes V_1 \\ &= \textstyle\sum_{p=1}^{j-1}(\beta U_p^j + M_{\mu,\sigma}U_{p+1}^j) \otimes V_{j-p}. \end{aligned}$$

Recall that W_{j-1} has the form

$$W_{j-1} = U_1^{j-1} \otimes V_{j-1} + \cdots + U_{j-1}^{j-1} \otimes V_1,$$

where $U_1^{j-1} = \gamma_{j-1}U$ for some $\gamma_{j-1} \in \mathbf{C}$. Thus to solve for W_j such that $J_\sigma(W_j) = W_{j-1}$, we need to solve the system of equations

$$\begin{array}{rcl} (\gamma_j\beta - \gamma_{j-1})U + M_{\mu,\sigma}U_2^j &=& 0, \\ \beta U_2^j - U_2^{j-1} + M_{\mu,\sigma}U_3^j &=& 0, \\ \vdots &=& \vdots \\ \beta U_{j-1}^j - U_{j-1}^{j-1} + M_{\mu,\sigma}U_j^j &=& 0, \end{array}$$

for $U_2^j, \ldots, U_j^j \in Q$ and $\gamma_j \in \mathbf{C}$. As in the case $j = 2$ we can solve the first equation by choosing γ_j so that $(\gamma_j\beta - \gamma_{j-1})U \in Q$. Then one can invert $M_{\mu,\sigma}$ to solve for $U_2^j \in Q$. The remaining equations have the form

$$\begin{array}{rcl} M_{\mu,\sigma}U_3^j &=& q_3, \\ \vdots &=& \vdots \\ M_{\mu,\sigma}U_j^j &=& q_j, \end{array}$$

where $q_3, \ldots, q_j \in Q$, and can also be solved for $U_3^j, \ldots, U_j^j \in Q$ by inverting $M_{\mu,\sigma}$. ■

3. Product networks. Suppose that N_1 and N_2 are two regular networks of sizes r_1 and r_2 with cells $c_1, \ldots, c_{r_1}$ and $d_1, \ldots, d_{r_2}$, respectively. Form the *product* network $N = N_1 \boxtimes N_2$ of size $r = r_1r_2$, where each node c_i in network N_1 is replaced by a copy of network N_2. Let p_{ij} be the jth cell in the copy of network N_2 that replaces cell c_i. So p_{ij} is a copy of cell d_j. Examples of product networks from the point of view of symmetry were considered by Dangelmayr, Güttinger, and Wegelin [6] and Dionne, Golubitsky, and Stewart [7]. The stability of synchronous dynamics in products of bidirectional graphs (without self-coupling or multiple arrows) was also considered by Atay and Biyikoğlu [2]. In particular, these authors calculated the eigenvalues of the Laplacian matrix which, in the kinds of graphs they considered, is directly related to the eigenvalues of the adjacency matrix.

We assume that there is an arrow from cell p_{ij} to cell $p_{\ell j}$ if and only if there is an arrow from cell c_i to cell c_ℓ in network N_1. We also assume that there is an arrow from cell p_{ij} to cell $p_{i\ell}$ if and only if there is an arrow from cell d_j to cell d_ℓ in network N_2. Finally, we assume that these two types of arrows are different. So there are two arrow types in the homogeneous network N—those arrows that connect cells within a given copy of N_2 and those arrows that connect cells between copies of N_2. In this sense, the product networks we consider differ from product networks in the graph theory literature.

As an example, consider the three-cell feed-forward network N_1 pictured in Figure 4(left). Let N_2 be the unidirectional ring pictured in Figure 1(left). Then $N = N_1 \boxtimes N_2$ is the 9-cell network in Figure 4(right). Note that N is homogeneous with two types of coupling. The locomotor central pattern generators modeling the gaits of n legged animals introduced in [14] provide a second example. Those networks are the product of a unidirectional ring of n cells with a ring of two cells.

It is easy to write the adjacency matrices corresponding to the two types of coupling in terms of tensor products. Let A_1, A_2 be the adjacency matrices for the networks N_1, N_2. Formally, write $p_{ij} = c_i \otimes d_j$. Then the adjacency matrix of couplings in network N corresponding

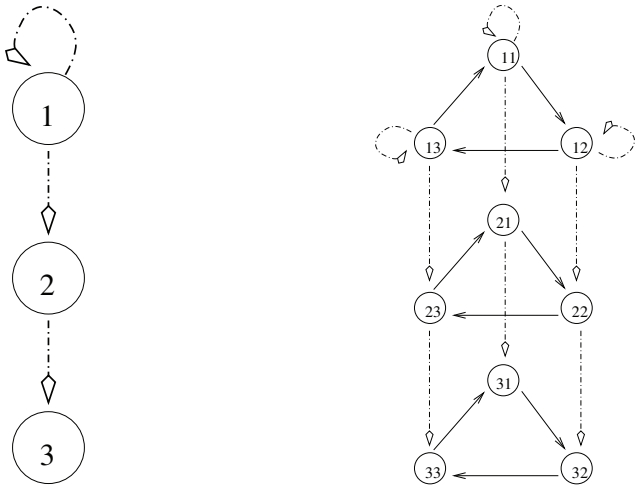


Figure 4. *(Left) three-cell feed-forward network; (right) nine-cell product of three-cell feed-forward network with three-cell unidirectional ring.*

to the couplings in network N_1 is $A_1 \otimes \mathbf{I}_{r_2}$, and the adjacency matrix corresponding to the couplings in network N_2 is $\mathbf{I}_{r_1} \otimes A_2$.

Similarly, if we assume that the internal dynamics of cells in N is one-dimensional, it is convenient to write the state space of N as

$$\mathbf{R}^r = \mathbf{R}^{r_1} \otimes \mathbf{R}^{r_2}.$$

Let $u_1, \ldots, u_{r_1}$ be the standard basis for $\mathbf{R}^{r_1}$ and $v_1, \ldots, v_{r_2}$ be the standard basis for $\mathbf{R}^{r_2}$. Then $u_i \otimes v_j$ is the standard basis for $\mathbf{R}^r$.

Next suppose that $\dot{z} = F(z)$ is an admissible vector field for the product network with $F(0) = 0$. Then the Jacobian has the form

$$(dF)_0 = \alpha \mathbf{I}_{r_1} \otimes \mathbf{I}_{r_2} + \beta_1 A_1 \otimes \mathbf{I}_{r_2} + \beta_2 \mathbf{I}_{r_1} \otimes A_2, \tag{3.1}$$

where α is the linearized internal dynamics and β_1, β_2 are the linearized coupling strengths for the couplings in networks N_1, N_2. It follows that if $v \in \mathbf{R}^{r_1}$ is an eigenvector of A_1 with eigenvalue μ and $w \in \mathbf{R}^{r_2}$ is an eigenvector of A_2 with eigenvalue ν, then $v \otimes w$ is an eigenvector of $(dF)_0$ with eigenvalue $\alpha + \beta_1 \mu + \beta_2 \nu$.

The assumption that the product network internal dynamics is one-dimensional is convenient because we can then identify the phase space P_{ij} for cell i, j in the product network with $P_i^1 \otimes P_j^2$, where P_i^1 is the phase space for cell i in network N_1 and P_j^2 is the phase space for cell j in network N_2. This identification, which is possible because $\mathbf{R} = \mathbf{R} \otimes \mathbf{R}$, allows us to write the linearization of a vector field at a synchronous equilibrium in the simple form given in (3.1). However, the assumption of one-dimensional internal dynamics is restrictive, because many generic codimension one Hopf bifurcations cannot occur unless the dimension of the internal dynamics is at least 2.

In the theory of coupled cell systems developed in [21, 16] the dimensions of cell phase spaces are fixed but arbitrary. In homogeneous networks, these phase spaces all have the same

dimension. However, in a product network, where the dimension of P_{ij} is $k > 1$, there is no natural way to think of this space as a tensor product. Said more strongly, there is no natural way to think of the admissible vector fields on a product network as tensor products of vector fields on the original networks.

We now assume that $P_{ij} = \mathbf{R}^k$, where $k \geq 1$. The full phase space of a coupled system corresponding to the product network consists of $r_1 r_2$ copies of $\mathbf{R}^k$. Thus, we can write the total phase as

$$P = \mathbf{R}^k \otimes \mathbf{R}^{r_1} \otimes \mathbf{R}^{r_2},$$

and we can use this tensor product decomposition of phase space to compute the Jacobian of a general admissible vector field at a synchronous equilibrium. To begin, recall from [16] applied to a product network that an admissible vector field has the form

$$\dot{x}_{ij} = f(x_{ij}, x_{I_{ij}}),$$

where $x_{ij} \in \mathbf{R}^k$ are the coordinates associated to cell i, j; I_{ij} is an index set consisting of cells connected to cell i, j; and the function f is independent of i, j since the product of homogeneous networks is homogeneous. However, there are two types of coupling in product networks. We can denote the indices of those cells that are coupled to cell i, j using the couplings in network N_1 by I_i, j, where I_i consists of those cells in network N_1 that are coupled to cell i. Similarly, those that are coupled to cell i, j using the couplings in network N_2 can be denoted by i, J_j. Thus, a product network admissible vector field has the form

$$\dot{x}_{ij} = f(x_{ij}, \overline{x_{I_i,j}}, \overline{x_{i,J_j}}), \tag{3.2}$$

where the function f is invariant under permutation of the coordinates under an overbar.

At a fully synchronous state, the Jacobian is determined by three $k \times k$ matrices: the linearized internal dynamics α, the linearized coupling β_1 due to the couplings in network N_1, and the linearized coupling β_2 due to the couplings in network N_2. These matrices are found by differentiating f in (3.2) in the first, second, and third types of variables, respectively. Using the discussion of linearizations given in [18, section 3], we can now generalize the formula for the Jacobian in (3.1) to the case $k > 1$. Specifically,

$$(dF)_0 = \alpha \otimes \mathbf{I}_{r_1} \otimes \mathbf{I}_{r_2} + \beta_1 \otimes A_1 \otimes \mathbf{I}_{r_2} + \beta_2 \otimes \mathbf{I}_{r_1} \otimes A_2. \tag{3.3}$$

To see this first consider the linearization when the coupling matrices β_j are zero. Then the linearization has the form $\alpha \otimes \mathbf{I}_{r_1} \otimes \mathbf{I}_{r_2}$ as in [18]. Assume next that the linearized internal dynamics α and one of the coupling matrices are zero to obtain the second and third terms. Again this proceeds as in [18].

It is possible to rewrite (3.3) in a more convenient form for later calculations. Let μ and ν be complex numbers and let

$$M_{\mu,\nu} = \alpha + \mu\beta_1 + \nu\beta_2, \qquad A_{1,\mu} = A_1 - \mu\mathbf{I}_{r_1}, \qquad A_{2,\nu} = A_2 - \nu\mathbf{I}_{r_2}. \tag{3.4}$$

Let $J = (dF)_0$. Then

$$J = M_{\mu,\nu} \otimes \mathbf{I}_{r_1} \otimes \mathbf{I}_{r_2} + \beta_1 \otimes A_{1,\mu} \otimes \mathbf{I}_{r_2} + \beta_2 \otimes \mathbf{I}_{r_1} \otimes A_{2,\nu}. \tag{3.5}$$

To verify (3.5) compute

$$\begin{aligned} J &= \alpha \otimes \mathbf{I}_{r_1} \otimes \mathbf{I}_{r_2} + \beta_1 \otimes A_1 \otimes \mathbf{I}_{r_2} + \beta_2 \otimes \mathbf{I}_{r_1} \otimes A_2 \\ &= \alpha \otimes \mathbf{I}_{r_1} \otimes \mathbf{I}_{r_2} + \beta_1 \otimes (A_{1,\mu} + \mu \mathbf{I}_{r_1}) \otimes \mathbf{I}_{r_2} + \beta_2 \otimes \mathbf{I}_{r_1} \otimes (A_{2,\nu} + \nu \mathbf{I}_{r_2}) \\ &= (\alpha + \mu\beta_1 + \nu\beta_2) \otimes \mathbf{I}_{r_1} \otimes \mathbf{I}_{r_2} + \beta_1 \otimes A_{1,\mu} \otimes \mathbf{I}_{r_2} + \beta_2 \otimes \mathbf{I}_{r_1} \otimes A_{2,\nu} \\ &= M_{\mu,\nu} \otimes \mathbf{I}_{r_1} \otimes \mathbf{I}_{r_2} + \beta_1 \otimes A_{1,\mu} \otimes \mathbf{I}_{r_2} + \beta_2 \otimes \mathbf{I}_{r_1} \otimes A_{2,\nu}. \end{aligned}$$

Proposition 3.1. *Let $\mathcal{G}_1(\mu)$ be the complex generalized eigenspace of A_1 corresponding to the eigenvalue μ, and let $\mathcal{G}_2(\nu)$ be the complex generalized eigenspace of A_2 corresponding to the eigenvalue ν. Then the following hold:*

(a) *the complex subspace*

$$V_{\mu,\nu} = \mathbf{C}^k \otimes \mathcal{G}_1(\mu) \otimes \mathcal{G}_2(\nu) \subset \mathbf{C}^k \otimes \mathbf{C}^{r_1} \otimes \mathbf{C}^{r_2} \tag{3.6}$$

is invariant under J, where the Jacobian matrix J is defined in (3.5).

(b) *Let $A_1 v = \mu v$ and $A_2 w = \nu w$. Then*

$$J(u \otimes v \otimes w) = (M_{\mu,\nu} u) \otimes v \otimes w. \tag{3.7}$$

Hence every eigenvalue of $M_{\mu,\nu}$ is an eigenvalue of J.

Proof. (a) Let $u \otimes v \otimes w$ be in $V_{\mu,\nu}$, and use (3.5) to compute

$$J(u \otimes v \otimes w) = (M_{\mu,\nu} u) \otimes v \otimes w + (\beta_1 u) \otimes (A_{1,\mu} v) \otimes w + (\beta_2 u) \otimes v \otimes (A_{2,\nu} w).$$

Since $A_{1,\mu} v \in \mathcal{G}_1(\mu)$ and $A_{2,\nu} w \in \mathcal{G}_2(\nu)$, invariance of $V_{\mu,\nu}$ is established.

(b) By assumption, $A_{1,\mu} v = A_{2,\nu} w = 0$. Therefore

$$J(u \otimes v \otimes w) = (M_{\mu,\nu} u) \otimes v \otimes w.$$

So every eigenvalue of $M_{\mu,\nu}$ is an eigenvalue of J. ∎

Remark 3.2. Suppose that μ and ν are both real. Then $M_{\mu,\nu}$ is a real matrix. Let $\mathcal{G}_1^R(\mu)$ and $\mathcal{G}_2^R(\nu)$ denote the real generalized eigenspaces of A_1 and A_2 corresponding to the eigenvalues μ and ν, respectively. Then the real subspace

$$V_{\mu,\nu}^R = \mathbf{R}^k \otimes \mathcal{G}_1^R(\mu) \otimes \mathcal{G}_2^R(\nu) \subset \mathbf{R}^k \otimes \mathbf{R}^{r_1} \otimes \mathbf{R}^{r_2} \tag{3.8}$$

is invariant under J.

Let $\mu_1, \ldots, \mu_{s_1}$ be the distinct eigenvalues of the adjacency matrix A_1, and let $\nu_1, \ldots, \nu_{s_2}$ be the distinct eigenvalues of the adjacency matrix A_2.

Definition 3.3. *A triple of real $k \times k$ matrices α, β_1, β_2 is a* codimension one triple *if, except for complex conjugate eigenvalues, the real parts of the $s_1 s_2 k$ eigenvalues of the matrices M_{μ_j,ν_i} are all distinct. We denote the set of all codimension one triples by $\mathcal{M}_3(k)$.*

It is clear that $\mathcal{M}_3(k)$ is an open set of $L(k)^3$, where $L(k)$ is the space of $k \times k$ matrices. In Proposition 3.4 we prove that when $k > 1$, the set $\mathcal{M}_3(k)$ is also a dense subset. Therefore, the generic codimension one bifurcation will occur with a Jacobian J associated with a codimension one triple, and J will have a center subspace that corresponds to one critical eigenvalue.

Proposition 3.4. *For $k > 1$, $\mathcal{M}_3(k) \subset L(k)^3$ is open and dense. If $k = 1$, $\mathcal{M}_3(k)$ is open and dense unless either A_1 or A_2 have eigenvalues with equal real parts.*

Proof. Let μ_1, μ_2 be eigenvalues of A_1 and ν_1, ν_2 be eigenvalues of A_2. Assume that $k = 1$. Then the corresponding eigenvalues are

$$M_{\mu_1,\nu_1} = \alpha + \mu_1\beta_1 + \nu_1\beta_2 \quad \text{and} \quad M_{\mu_2,\nu_2} = \alpha + \mu_2\beta_1 + \nu_2\beta_2,$$

and they have equal real part if and only if

$$\beta_1(\mathrm{Re}(\mu_1) - \mathrm{Re}(\mu_2)) + \beta_2(\mathrm{Re}(\nu_1) - \mathrm{Re}(\nu_2)) = 0. \tag{3.9}$$

Identity (3.9) can hold for an open set of β_1, β_2 if and only if both

$$\mathrm{Re}(\mu_1) = \mathrm{Re}(\mu_2) \quad \text{and} \quad \mathrm{Re}(\nu_1) = \mathrm{Re}(\nu_2). \tag{3.10}$$

Suppose that μ_1 and μ_2 have equal real parts and $\nu_1 = \nu_2$. Then the real parts of the eigenvalues M_{μ_1,ν_1} and M_{μ_2,ν_1} will be the same independent of the choice α, β_1, β_2. A similar conclusion holds if the ν's have equal real parts and the μ's are equal (or, indeed, if both the μ's and ν's have equal real parts).

We can now assume that $k \geq 2$. Suppose now that two eigenvalues σ_1 from the matrix M_{μ_1,ν_1} and σ_2 from the matrix M_{μ_2,ν_2} have equal real part. We can perturb β_1 to $\beta_1 + \varepsilon_1 I_k$ and β_2 to $\beta_2 + \varepsilon_2 I_k$. The perturbed eigenvalues $\sigma_j + \varepsilon_1\mu_j + \varepsilon_2\nu_j$ can have equal real parts for an open set of ε_j if and only if (3.10) is valid. Thus we have reduced the proof to two cases:

(a) $\mathrm{Re}(\mu_1) = \mathrm{Re}(\mu_2)$ and $\nu_1 = \nu_2$, where $\mu_1 \neq \mu_2$, or the reverse in the μ's and ν's.

(b) $\mathrm{Re}(\mu_1) = \mathrm{Re}(\mu_2)$ and $\mathrm{Re}(\nu_1) = \mathrm{Re}(\nu_2)$, where $\mu_1 \neq \mu_2$ and $\nu_1 \neq \nu_2$.

Case (a). As in the proof of Proposition 2.5, we can shift and scale the eigenvalues ν so that $\nu_1 = \nu_2 = 0$ or $\nu_1 = \nu_2 = i$. In the first case, the matrices are $M_{\mu_1,\nu_1} = \alpha + \mu_1\beta_1$ and $M_{\mu_2,\nu_2} = \alpha + \mu_2\beta_1$. For these matrices we can apply Proposition 2.5 directly to perturb the matrices α, β_1 to separate the real parts of the eigenvalues of $M_{\mu_1,\nu_1}, M_{\mu_2,\nu_2}$. So we can assume

$$M_{\mu_1,\nu_1} = \alpha + \mu_1\beta_1 + i\beta_2 \quad \text{and} \quad M_{\mu_2,\nu_2} = \alpha + \mu_2\beta_1 + i\beta_2.$$

Choose vectors v_1, v_2 so that $M_{\mu_1,\nu_1}v_1 = \sigma_1 v_1$ and $M_{\mu_2,\nu_2}v_2 = \sigma_2 v_2$. Moreover, we can perturb β_2 so that the vectors $\beta_2 v_1$ and $\beta_2 v_2$ are linearly independent over $\mathbf{C}$. Note that $C = \alpha + i\beta_2$ is an arbitrary complex matrix and that

$$Cv_1 = -(\sigma_1 + \beta_2)v_1 \quad \text{and} \quad Cv_2 = -(\sigma_2 + \beta_2)v_2.$$

It follows from linear independence that we can perturb C so that Cv_2 is unchanged whereas $Cv_1 = -(\sigma_1 + \varepsilon + \beta_2)v_1$. It follows that the real parts of the eigenvalues σ_1, σ_2 are separated.

Case (b). Write

$$M_{\mu\nu} = \alpha + \mu\beta_1 + \nu\beta_2$$

and use the scaling argument as before to write

$$M_{\mu,\nu} = (\alpha + c_1\beta_1 + c_2\beta_2) + (\mu - c_1)\beta_1 + (\nu - c_2)\beta_2.$$

This way we can achieve that the coefficients of β_j are purely imaginary. By the second scaling,

$$\mu\beta_1 \mapsto (c\mu)\frac{1}{c}\beta_1,$$

and a similar argument for the second coefficient we have

$$\mu_1 = 0 \qquad \text{or} \qquad \mu_1 = i$$

and similarly

$$\nu_1 = 0 \qquad \text{or} \qquad \nu_1 = i.$$

Applying the same scalings to μ_2 and ν_2 gives us

$$\mu_2 = si \qquad \text{and} \qquad \nu_2 = ti$$

for some real numbers $s, t \neq 0, 1$. We have three cases to distinguish:

1. $\mu_1 = \nu_1 = 0$, $\mu_2 = \nu_2 = i$.
2. $\mu_1 = 0$, $\nu_1 = i$ and $\mu_2 = i$, $\nu_2 = ti$, $0 \neq t \neq 1$.
3. $\mu_1 = i$, $\nu_1 = i$ and $\mu_2 = si$, $\nu_2 = ti$, $0 \neq s, t \neq 1$.

In the first case we consider

$$M_{\mu_1,\nu_1} = \alpha \quad \text{and} \quad M_{\mu_2,\nu_2} = \alpha + i(\beta_1 + \beta_2).$$

In this case our perturbation argument is the same as in part (a).

In the second case we look at the matrices

$$\alpha + i\beta_2 \quad \text{and} \quad \alpha + is\beta_1 + it\beta_2,$$

where s, t are as before. Again the first matrix is an arbitrary complex matrix. We can apply the argument from $\nu_a = i$; $\nu_b = i\omega$ given in the second part of the proof of Proposition 2.5.

In the third case, we look at matrices

$$M_{\mu_1,\nu_1} = \alpha + i(\beta_1 + \beta_2) \quad \text{and} \quad M_{\mu_2,\nu_2} = \alpha + is\beta_1 + it\beta_2$$

with $s, t \neq 0, 1$. Again we can use the argument given in the second part of the proof of Proposition 2.5 to show that a perturbation exists which splits the real parts of the eigenvalues. ■

4. Center subspaces for product networks. As we have seen, the eigenvalues σ of the Jacobian J of an admissible system of differential equations associated with a product network are the union of the eigenvalues of $M_{\mu,\nu}$, where μ, ν vary over the sets of eigenvalues of the adjacency matrices A_1 and A_2, respectively. Recall that the matrix $M_{\mu,\nu}$ is defined in (3.7). We assume that the $k \times k$ matrices α, β_1, β_2 that define J form a codimension one triple. Note that

$$\overline{M_{\mu,\nu}} = M_{\overline{\mu},\overline{\nu}},$$

so that there may be some redundancy in the enumeration of eigenvalues of J. To eliminate possible redundancies, let $\mathcal{P}$ be a set of eigenvalue pairs that includes precisely one of the

eigenvalue pairs μ, ν and $\overline{\mu}, \overline{\nu}$. It follows from Proposition 3.4 that the eigenvalues σ corresponding to $M_{\mu,\nu}$ with the pair μ, ν in $\mathcal{P}$ all have distinct real parts, at least when $k > 1$, which we now assume. It is also the case that generically σ is a simple eigenvalue of $M_{\mu,\nu}$ and generically σ is not real if either μ or ν is not real.

Given these genericity assumptions, the center subspace of J is the real part of the generalized eigenspace of J corresponding to an eigenvalue σ with 0 real part. To make the content of this observation more transparent we distinguish two cases:

(a) $\sigma = 0$,

(b) $\sigma = \omega i$, where $\omega \neq 0$.

In case (a) μ and ν are real, and the matrix $M_{\mu,\nu}$ is also real. Proposition 3.4 implies that generically there are no other pairs μ', ν' such that 0 is an eigenvalue of $M_{\mu',\nu'}$ and that no other eigenvalue of J has zero real part. Remark 3.2 shows that the real subspace $V^R_{\mu,\nu}$ defined in (3.8) is J-invariant. Let $U^R \subset \mathbf{R}^k$ be the real one-dimensional kernel of $M_{\mu,\nu}$; then the center subspace of J is the generalized eigenspace corresponding to the eigenvalue 0 and is given by $U^R \otimes \mathcal{G}_1^R(\mu) \otimes \mathcal{G}_2^R(\nu) \cong \mathcal{G}_1^R(\mu) \otimes \mathcal{G}_2^R(\nu)$. The last isomorphism is valid because U^R is real one-dimensional.

Case (b) itself divides into two parts: either μ and ν are both real or at least one of μ, ν is not real. In the first part $M_{\mu,\nu}$ is a real matrix and $-\omega i$ is also an eigenvalue of $M_{\mu,\nu}$. So the pair $\pm\omega i$ are both eigenvalues of J, and there are no other eigenvalues of J with 0 real part. Then the real subspace $V^R_{\mu,\nu}$ is invariant under J. Let U^R be the two-dimensional invariant subspace of $\mathbf{R}^k$ corresponding to the simple eigenvalues $\pm\omega i$ of $M_{\mu,\nu}$; then the center subspace of J is the invariant subspace of J corresponding to these eigenvalues and is given by $U^R \otimes \mathcal{G}_1^R(\mu) \otimes \mathcal{G}_2^R(\nu) \cong (\mathcal{G}_1^R(\mu) \otimes \mathcal{G}_2^R(\nu))^2$. The last isomorphism is valid because U^R is real two-dimensional.

In the second part either μ or ν or both are not real. Then $-\omega i$ is an eigenvalue of $\overline{M_{\mu,\nu}} = M_{\bar{\mu},\bar{\nu}}$, and the center subspace of J corresponds to the generalized eigenspaces of the eigenvalues $\pm\omega i$ of J, since no other eigenvalue of J has 0 real part. More precisely, let $U \subset \mathbf{C}^k$ be the complex one-dimensional subspace of $\mathbf{C}^k$ that consists of eigenvectors of $M_{\mu\nu}$ corresponding to the eigenvalue ωi, and let $\overline{U}$ be the subspace of eigenvectors of $M_{\bar{\mu},\bar{\nu}}$ corresponding to $-\omega i$. Then the space

$$(U \otimes \mathcal{G}_1(\mu) \otimes \mathcal{G}_2(\nu)) \oplus (\overline{U} \otimes \mathcal{G}_1(\bar{\mu}) \otimes \mathcal{G}_2(\bar{\nu}))$$

is invariant under J, and its real part

$$\{w + \bar{w} : w \in U \otimes \mathcal{G}_1(\mu) \otimes \mathcal{G}_2(\nu)\} = \mathrm{Re}(U \otimes \mathcal{G}_1(\mu) \otimes \mathcal{G}_2(\nu)) \cong \mathrm{Re}(\mathcal{G}_1(\mu) \otimes \mathcal{G}_2(\nu)) \tag{4.1}$$

is the generalized eigenspace of J corresponding to the eigenvalues $\pm\omega i$. The last isomorphism is valid because U is complex one-dimensional.

5. Jordan structure on center subspaces. Before we prove a result similar to Theorem 2.7 for product networks, we need to develop a deeper understanding of tensor products of Jordan blocks. Specifically, we consider

$$\mathcal{A} : \mathcal{G}_1(\mu) \otimes \mathcal{G}_2(\nu) \to \mathcal{G}_1(\mu) \otimes \mathcal{G}_2(\nu),$$

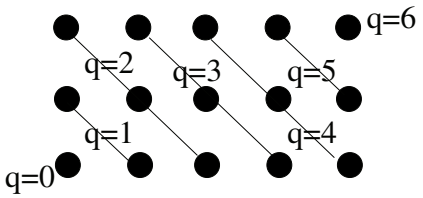


Figure 5. *Rectangle of lattice sites in* $\mathbf{R}^2$ *with* $\ell = 5$ *and* $m = 3$.

where

$$\mathcal{A} = A_{1,\mu} \otimes \mathbf{I}_{r_2} + \mathbf{I}_{r_1} \otimes A_{2,\nu}.$$

We can write

$$\mathcal{G}_1(\mu) = \bigoplus_{s=1}^{q_1} \mathsf{J}_s^1(\mu) \quad \text{and} \quad \mathcal{G}_2(\nu) = \bigoplus_{s=1}^{q_2} \mathsf{J}_s^2(\nu),$$

where J_s^j stands for a Jordan block in $\mathcal{G}_j$. Each block has dimension $d_s(\mu)$ or $d_s(\nu)$, and we see easily that the tensor products

$$\mathsf{J}_s^1(\mu) \otimes \mathsf{J}_t^2(\nu)$$

are invariant under $\mathcal{A}$. Moreover,

$$\bigoplus_{s,t} \mathsf{J}_s^1(\mu) \otimes \mathsf{J}_t^2(\nu) = \mathcal{G}_1(\mu) \otimes \mathcal{G}_2(\nu).$$

It is important to observe that the tensor products $\mathsf{J}_s^1(\mu) \otimes \mathsf{J}_t^2(\nu)$ are not the Jordan blocks corresponding to $\mathcal{A}$.

Fix Jordan blocks $\mathsf{J}^1(\mu)$ for A_1 and $\mathsf{J}^2(\nu)$ for A_2. Let v_0 be an eigenvector of A_1 in $\mathsf{J}^1(\mu)$, and let ℓ be the smallest positive integer for which $A_{1,\mu}^{\ell}$ is identically zero on $\mathsf{J}^1(\mu)$. Similarly, let w_0 be an eigenvector for A_2 in $\mathsf{J}^2(\nu)$, and let m be the smallest positive integer for which $A_{2,\nu}^{m}$ is zero on $\mathsf{J}^2(\nu)$. A basis for the space $\mathcal{G}_1(\mu) \otimes \mathcal{G}_2(\nu)$ is given by the tensors $v_j \otimes w_r$, where $0 \le j \le \ell - 1$ and $0 \le r \le m - 1$. For each $q = 0, \ldots, \ell + m - 2$, let Z_q be the subspace spanned by the tensors $v_s \otimes w_t$ with $s + t = q$. Then for $q \ge 1$

$$\mathcal{A} : Z_q \to Z_{q-1}.$$

We say that a vector in Z_q has *order* q.

A good way to think about the subspaces Z_q is to look at the rectangle of lattice sites in $\mathbf{R}^2$ (see Figure 5). The pairs (j, r) for which $j + r = q$, which lie on lines with slope -1 in that figure, form a basis for the subspace Z_q. Without loss of generality assume $\ell \ge m$. Then we have a sequence of maps

$$Z_{\ell+m-2} \to Z_{\ell-m-3} \to \cdots \to Z_{\ell-1} \to \cdots \to Z_{m-1} \to \cdots \to Z_0. \tag{5.1}$$

Writing $\mathcal{A}$ in coordinates with respect to the basis, we see that $\mathcal{A}$ has maximal rank as a mapping of $Z_q \to Z_{q-1}$ for all q.

Looking at the geometric picture in Figure 5, we see that

$$1 = \dim Z_{\ell+m-2} < \dim Z_{\ell+m-3} = 2 < \cdots < \dim Z_{\ell-1} = \dim Z_{\ell-2} = \cdots = \dim Z_{m-1}$$

and

$$1 = \dim Z_0 < \cdots < \dim Z_{m-1}.$$

Therefore, $\mathcal{A} : Z_q \to Z_{q-1}$ is

- surjective for $q \le m$,
- an isomorphism for $m+1 \le q \le \ell$, and
- injective for $\ell + 1 \le q$.

So it remains to describe precisely the Jordan blocks. For each $0 \le q \le m-1$ we have to give a vector in the kernel of $\mathcal{A}$, and in a similar way we have to give vectors in the complement to the range of $\mathcal{A}$, or the *cokernel* of $\mathcal{A}$, for $q > \ell$.

Proposition 5.1. *The kernel of $\mathcal{A}$ is spanned by the vectors*

$$z_q = \sum_{j=0}^{q} (-1)^j v_j \otimes w_{q-j},$$

where $1 \le q \le m-1$, and a basis for the cokernel of $\mathcal{A}$ is

$$z^q = \sum_{j=0}^{q} (-1)^j \alpha_j v_{\ell-j-1} \otimes w_{m-q+j-1},$$

with appropriately chosen α_j. Moreover, we have

$$\mathcal{A}^{\ell+m-2(1+q)-1} z^q = z_q. \tag{5.2}$$

Remark 5.2. Observe that the vectors for the complement of the ranges are not uniquely determined. We could set all $\alpha_j = 1$; however, for (5.2) to hold, we need the freedom to choose the α_j.

Proof. Observe that $A_{1,\mu}, A_{2,\nu}$ are nilpotent, allowing us to use representation theory for $\mathbf{sl}(2, \mathbf{R})$. We decompose the representation, on the tensor product into representations, and we obtain the length of the Jordan blocks from this Clebsch–Gordan decomposition; compare [20, 5]. ∎

In order to describe the way that this decomposition carries over to the decomposition into Jordan blocks for J_σ we need some auxiliary results.

Lemma 5.3.

$$U \otimes \mathsf{J}^1(\mu) \otimes \mathsf{J}^2(\nu)$$

is invariant under J_σ.

Proof. Let us write

$$M_{\mu,\nu}(\sigma) = M_{\mu,\nu} - \sigma \mathbf{I}.$$

Then, for $u \in U$, $v \in \mathsf{J}^1(\mu)$, $w \in \mathsf{J}^2(\nu)$ we have

$$J_\sigma(u \otimes v \otimes w) = (M_{\mu,\nu}(\sigma)u) \otimes v \otimes w + \beta_1 u \otimes A_{1,\mu} v \otimes w + \beta_2 u \otimes v \otimes A_{2,\nu} w.$$

All three summands are in $U \otimes \mathsf{J}(\mu) \otimes \mathsf{J}(\nu)$, thus proving the invariance. ■

So we can reduce our attention to tensor products of Jordan blocks. We need one result about being able to solve certain equations up to terms of lower order, which we will state first. Assuming that all eigenvalues of $M_{\mu,\nu}$ are simple (cf. Proposition 3.4), we can write each $u \in U$ uniquely as

$$u = \sum_{i=1}^{k} \xi_i u_i,$$

where each u_i is an eigenvector of $M_{\mu,\nu}$. Fix an eigenvalue σ of $M_{\mu,\nu}$ and denote the corresponding eigenvector by u_0.

Lemma 5.4. *Given $p+1$ vectors $x_j \in U$ and p real numbers $\zeta_1, \dots, \zeta_p$, where for each vector x_j the u_0 component is zero. Then for*

$$z = \sum_{j=0}^{p} (-1)^j x_j \otimes v_j \otimes w_{p-j}$$

there exist vectors

$$y_0, \dots, y_p \in U$$

such that

$$Y = \sum_{j=0}^{p} y_j \otimes v_j \otimes w_{p-j}$$

solves the equation

$$J_\sigma Y = z$$

up to terms of lower order. That is,

$$z - J_\sigma Y = \sum_{j=0}^{p-1} h_j \otimes v_j \otimes w_{p-1-j},$$

with the additional property that the u_0-component of each h_j is $\zeta_j u_0$.

Proof. We begin with the case $p = 1$; then we have two vectors x_1, x_2, neither having a component in the direction of u_0. We look at

$$z = x_0 \otimes v_1 \otimes w_0 - x_1 \otimes v_0 \otimes w_1.$$

Make the ansatz

$$Y = y_0 \otimes v_1 \otimes w_0 - y_1 \otimes v_0 \otimes w_1$$

and compute $J_\sigma Y - z$:

$$J_\sigma(y_0 \otimes v_1 \otimes w_0 - y_1 \otimes v_0 \otimes w_1) - x_0 \otimes v_1 \otimes w_0 - x_1 \otimes v_0 \otimes w_1.$$

We find (up to terms of lower order)

$$J_\sigma(y_0 v_1 \otimes w_0 - y_1 \otimes v_0 \otimes w_1) - z = (M_{\mu,\nu}(\sigma) y_0 - x_0) \otimes v_1 \otimes w_0 + (M_{\mu,\nu}(\sigma) y_1 - x_1).$$

Solving this equation requires the partial invertibility of $M_{\mu,\nu}(\sigma)$. Using the definition of Q as in (2.12), we find that $M_{\mu,\nu}$ restricted to Q is invertible. We write $M_{\mu,\nu}(\sigma)^{-1}$ for this partial inverse if we obtain solutions

$$y_a = M_{\mu,\nu}(\sigma)^{-1} x_a$$

for $a = 0, 1$. Observe for any $t_a \in \mathbf{R}$

$$M_{\mu,\nu}(\sigma)(y_a + t_a u_0) = x_a.$$

Now consider the application of the remaining part of J_σ to

$$(y_0 + t_0 u_0) \otimes v_1 \otimes w_0 + (y_1 + t_1 u_0) \otimes v_0 \otimes w_1.$$

We obtain for the terms of order $p - 1$

$$(\beta_1 y_0 + t_0 \beta_1 u_0) \otimes v_0 \otimes w_0 + (\beta_2 y_1 + t_1 \beta_2 u_0) \otimes v_0 \otimes w_0.$$

Now, if we make the nondegeneracy assumption that $\beta_j u_0$ has a component in the direction of u_0, we choose t_0, t_1 so that the u_0-component in this expression is $\zeta_j u_0$.

Repeating this computation for general q gives the induction, and we have proved the claim. ■

Theorem 5.5 needs an assumption on β_1, β_2 that is similar to the genericity assumptions (2.5) and (2.6). Its precise form will be given in the proof.

Theorem 5.5. *Let μ, ν be eigenvalues of A_1, A_2, and let σ be a simple eigenvalue of $M_{\mu,\nu}$. Then generically in α, β_1, β_2 the corresponding eigenspace of J is isomorphic to $\mathrm{Re}(\mathcal{G}_1(\mu) \otimes \mathcal{G}_2(\nu))$. The isomorphism sends each eigenvector z_q of $\mathcal{A}$ to an eigenvector $\tilde{z}_q$ of J_σ, where the top part of $\tilde{z}_q$ is given by $u_0 \otimes z_q$.*

As in the case of regular networks we can immediately describe the generic center subspaces for codimension one bifurcations in product networks.

Corollary 5.6. *Assume $k > 1$. Let E^c be the center subspace corresponding to a codimension one bifurcation from a fully synchronous equilibrium in a product network. Then, generically in α, β_1, β_2, E^c is as a vector space isomorphic to $\mathrm{Re}(\mathcal{G}_1(\mu) \otimes \mathcal{G}_2(\nu))$ or to $\mathrm{Re}(\mathcal{G}_1(\mu) \otimes \mathcal{G}_2(\nu)) \oplus \mathrm{Re}(\mathcal{G}_1(\mu) \otimes \mathcal{G}_2(\nu))$.*

Example 5.7. We consider the case of a product of the three-cell feed-forward network with a unidirectional ring with three cells and a two-dimensional internal dynamics leading to Hopf bifurcation. The eigenvalues of A_1 are $0, 0, 1$ with a geometrically simple eigenvalue 0. The eigenvalues of A_2 are $1, \zeta, \zeta^2$ with $\zeta^3 = 1$. We consider $\mu_1 = 0$ and $\nu_1 = 1$ (any other choice here gives similar results). Then we can find real 2×2-matrices α, β_1, β_2 such that the corresponding eigenvalue σ is purely imaginary and equal to i. Then the critical eigenvalues are $\pm i$ which are algebraically double and geometrically simple.

Proof. The proof of the theorem is more involved than that of Theorem 2.7. We concentrate on one tensor block. In a first step, we show that for each element in the kernel of $\mathcal{A}$ there is an eigenvector of J corresponding to the eigenvalue σ.

The spaces Z_q give a grading of the space $\mathsf{J}^1(\mu) \otimes \mathsf{J}^2(\nu)$. We will use this structure and Lemma 5.4 to solve equations recursively on the various levels.

Now let us consider the kernel vectors of $\mathcal{A}$ within Z_q, for $q > 1$. For the eigenvector we make the ansatz

$$\tilde{z}_q = \sum_{j=0}^{q} \alpha_j u_0 \otimes v_j \otimes w_{q-j} + \sum_{p<q} \sum_{j=0}^{p} x_j^p \otimes v_j \otimes w_{p-j}, \quad x_j \in U.$$

We have to show that we can solve the equation $J_\sigma \tilde{z}_q = 0$.

In a first step we consider the first term (of order q), and we describe how to adjust the other terms $p < q$ in a second step. Applying $M_{\mu,\nu}(\sigma)$ to the first term gives 0 since u_0 is an eigenvector of $M_{\mu,\nu}$ for the eigenvalue σ. Therefore we just have to apply the other two operators. This gives a term of the form

$$\sum_{j=0}^{p-1} (\alpha_j \beta_2 u_0 - \alpha_{j+1} \beta_1 u_0) \otimes v_j \otimes w_{p-1-j}.$$

Observe that we can choose the α_j to adjust the u_0 contribution in each term to be 0. Set

$$z = \sum_{j=0}^{p-1} (\alpha_j \beta_2 u_0 - \alpha_{j+1} \beta_1 u_0) \otimes v_j \otimes w_{p-1-j}.$$

By the induction step we can choose vectors y_j such that $J_\sigma Y - z$ is zero on level $q-1$, and the terms in level $q-2$ have no u_0 contribution. We can iterate this procedure to solve the equation $J_\sigma \tilde{z}_q = 0$. This proves the result on eigenvectors. ■

Lemma 5.8. *Let $Q(\lambda)$ be a smooth family of nilpotent matrices, where* $\dim \ker Q(\alpha)$ *is constant. Then the kernel varies smoothly with λ. If the length of all chains of generalized eigenvectors are different, then the generalized eigenvectors vary smoothly in λ.*

Proof. First, we solve the equation $Q(\lambda)x = 0$ using a Liapunov–Schmidt decomposition. Without loss of generality we consider the problem near $\lambda = 0$. For this, let K be the kernel of $Q(0)$, let M be a complement to K, let R be the range of $Q(0)$, and let N be a complement to N. Let $E : \mathbf{R}^n \to R$ and $(I - E) : \mathbf{R}^n \to N$ be projections. Write u for the variable in the kernel, w for the variable in M. Then $Q(\lambda)x = 0$ is equivalent to

$$\begin{array}{rcl} f(u, w, \lambda) & \equiv & (I - E)Q(\lambda)(u + w) = 0, \\ g(u, w, \lambda) & \equiv & EQ(\lambda)(u + w) = 0. \end{array}$$

Then $g_w : M \to R$ is an isomorphism which leads to a smooth solution of

$$g(u, w, \lambda) = 0$$

as $w = w(u, \lambda)$, where $w(u, 0) = 0$ and w is linear in u. Then we have the equivalent equation

$$(I - E)Q(\lambda)(u + w(u, \lambda)) = 0.$$

From the hypotheses we know that the solution set of this equation has the same dimension as the kernel of $Q(0)$. Therefore, the kernel of $Q(\lambda)$ is given by

$$\ker Q(\lambda) = \{u + w(u, \lambda) \ : \ u \in K\}.$$

This is a smooth deformation of K. Especially we find a smooth family of projection operators $P(\lambda)$ mapping $\mathbf{R}^n$ onto $\ker Q(\lambda)$.

Secondly we look at the generalized eigenvector g for Q which has the longest chain. There is a positive integer q with

$$Q^q g = 0$$

and

$$Q^i \neq 0$$

for $1 \leq i < q$. Now we want to solve an equation of the form

$$Q^q(\lambda)(w) = 0.$$

We solve the equivalent equation

$$P(\lambda)\, Q^{q-1}(\lambda) x\,(\lambda) = 0$$

for some $x(\lambda)$ near g. Q^{q-1} is an isomorphism from a complement of the kernel to its range. Write w for the coordinate in the complement of the kernel. Then the implicit function theorem tells us that we can solve the equation

$$P(\lambda) Q^{q-1}(\lambda)(g + w) = 0$$

uniquely by some $w(\lambda)$. Since $w(0) = 0$, we have that $Q^i(\lambda)g + w(\lambda)$ is uniformly near the points $Q^i g$. This means that the element $g + w(\lambda)$ defines a chain of the required length.

By restriction of the problem to a complement of the longest chain, we can iteratively prove the result. ∎

Corollary 5.9. *Generically, the generalized eigenspaces of J_σ are in* 1:1 *correspondence with those of $\mathcal{A}$.*

Proof. We note that if $\beta_1 = \beta_2$, the result follows from Theorem 2.7, because there the matrix $\mathcal{A}$ is the adjacency matrix for the regular network obtained from $N_1 \boxtimes N_2$ by setting the couplings equal. Next we perturb β_1, β_2 near the diagonal $\beta_1 = \beta_2$. Theorem 5.5 guarantees that the kernel contains at least the given set of vectors. Generic perturbations will not increase the dimension of the kernel, and therefore the dimension stays locally constant. Since the Jordan blocks for J_σ in the case of $\beta_1 = \beta_2$ have the same structure as those for $\mathcal{A}$, which is given by the Clebsch–Gordan theorem, all the length of the chains are different. Therefore the hypotheses of the previous Lemma 5.8 are satisfied, and we obtain that the Jordan structure is the same for $\|\beta_1 - \beta_2\|$ small. ∎

Remark 5.10. The Jordan structure within the center subspace $E^Q(\lambda_1 + \lambda_2)$ can be complicated. Proposition 3.1 gives a bound on the maximal length of Jordan blocks within this space; however, it does not enumerate the number of Jordan blocks. Consider the matrices

$$A_1 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad A_2 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

Since $A_1^3 = 0$ and $A_2^2 = 0$, Proposition 3.1 states that $Q^4 = 0$. In this case the lengths of the Jordan blocks are $\ell_1 = 3$ and $\ell_2 = 2$. Observe that the Jordan block structure of

Q contains two blocks—one of length 4 and one of length 2. In general, when A_1 and A_2 are Jordan blocks of lengths $\ell_1 \geq \ell_2$, the tensor product will consist of ℓ_2 blocks of length $\ell_1 - \ell_2 + 1, \ell_1 - \ell_2 + 3, \ldots, \ell_1 + \ell_2 - 1$.

6. Branching of solutions. In this section we prove an analogue of the equivariant branching lemma [15] in the context of homogeneous cell systems.

Definition 6.1. *Let y in phase space be given. The isotropy subspace containing y is the smallest synchrony subspace Δ_y that contains y.*

The subspace $\Delta = \{(y, \ldots, y)\}$ consists of fully synchronous points. Suppose that $\dot{x} = F(x, \lambda)$ is a coupled cell system and $x_0 \in \Delta$ is a fully synchronous equilibrium at λ_0; that is, $F(x_0, \lambda_0) = 0$. We call x_0 a *synchrony-breaking* bifurcation point if

$$K = \ker(dF)_{x_0,\lambda_0} \neq 0 \quad \text{and} \quad K \cap \Delta = \{0\}.$$

Without loss of generality, we can assume that $\lambda_0 = 0$. At a synchrony-breaking bifurcation $(dF)_{x_0,0}|\Delta$ is nonsingular. Therefore, the implicit function theorem guarantees that there is a parameterized family of fully synchronous equilibria $x(\lambda)$, where $x(0) = x_0$. Again, we can assume, without loss of generality, that $x(\lambda) = 0$. It follows that $x_0 = 0$ and $F(0, \lambda) = 0$.

Definition 6.2. *An isotropy subspace Δ_y is* axial *if* $\dim(\Delta_y \cap K) = 1$.

Theorem 6.3. *Let Δ_y be an axial polydiagonal, where $\Delta_y \cap K = \mathbf{R}\{v\}$. Then generically there exists a unique branch of zeros of F in Δ_y.*

Proof. Let A be the $r \times r$ adjacency matrix of an network with r nodes. Let $\bowtie$ be the equivalence relation corresponding to input equivalence. Considering A as a linear mapping on $\mathbf{R}^r$ (or $\mathbf{C}^r$), we have an invariant linear subspace S corresponding to $\bowtie$. We do not distinguish between the real and complex versions of S. Let μ be an eigenvalue of A, $\mathcal{G}_A(\mu)$ be the corresponding generalized eigenspace, and $E_A(\mu)$ the kernel of $A - \mu\mathbf{I}$.

Next we consider the case of a k-dimensional internal dynamics for the network. The state space is isomorphic to $\mathbf{R}^k \otimes \mathbf{R}^r$. Let

$$\overline{S} = \mathbf{R}^k \otimes S,$$

which is invariant under the map $\mathbf{I} \otimes A$. We have

$$M_\mu = \alpha + \mu\beta$$

for real $k \times k$ matrices α, β. Let σ be a simple eigenvalue of A. Let

$$J = \alpha \otimes \mathbf{I}_r + \beta \otimes A.$$

Then Theorem 2.7 implies that there exists an isomorphism η,

$$\eta : \mathcal{G}_A(\mu) \to \mathcal{G}_J(\sigma).$$

We can also consider the spaces

$$\mathcal{G}_A^S(\mu) = S \cap \mathcal{G}_A(\mu)$$

and

$$\mathcal{G}_J^{\overline{S}}(\sigma) = \overline{S} \cap \mathcal{G}_J(\sigma).$$

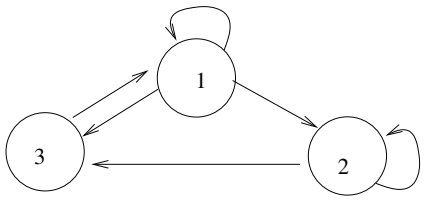


Figure 6. *Three-cell network with multiple eigenvalues in the adjacency matrix.*

If we apply this same construction, we conclude that in the case that σ is a simple eigenvalue of M_μ, then

$$\eta^S : \mathcal{G}_A^S(\mu) \to \mathcal{G}_J^{\overline{S}}(\sigma)$$

is an isomorphism. Especially we have that

$$\dim E_A^S(\mu) = \dim E_J^{\overline{S}}(\sigma).$$

Let $\hat{F} = F|\Delta_y$. Observe that $\hat{F}(0,\lambda) = 0$ and that $(d\hat{F})_{0,0}$ has a one-dimensional kernel equal to $\Delta_y \cap K = \mathbf{R}\{v\}$. Let $G : \mathbf{R} \times \mathbf{R} \to \mathbf{R}$ be the reduced equation obtained by Liapunov–Schmidt reduction of $\hat{F}$ to $\mathbf{R}\{v\}$. Observe that $G(0,\lambda) = 0$. Therefore, by Taylor's theorem,

$$G(t,\lambda) = tH(t,\lambda).$$

Generically, $H(0,0)$ is nonzero, and the implicit function theorem guarantees the existence of a branch of solutions to $H(t,\lambda) = 0$ parametrized by t. ■

Example 6.4. Network 6 in [18], shown in Figure 6, provides an example where Theorem 6.3 can be applied when 0 is a multiple eigenvalue of the adjacency matrix and therefore of dF. The adjacency matrix is

$$A_6 = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix}$$

with a double zero eigenvalue and a single nullvector $v = (1,-1,-1)$.

Observe that setting cell 2 equal to cell 3 leads to a balanced coloring and that $\Delta_v \cap K = \mathbf{R}\{v\}$ is one-dimensional. Therefore, Theorem 6.3 implies the existence of a branch of equilibria with $x_2 = x_3$. We note that [18, Theorem 4.11] also proves the existence of an asynchronous branch of solutions emanating from this bifurcation.

Suppose that the coupled cell system has a symmetry group Γ, and suppose that Σ is an axial subgroup for the action of Γ on K. Then $\mathrm{Fix}_K(\Sigma) = \mathrm{Fix}(\Sigma) \cap K$ is one-dimensional, and Theorem 6.3 reduces to the equivariant branching lemma.

Example 6.5. The four-cell network pictured in Figure 7 shows that in principle Theorem 6.3 gives more refined information than does the equivariant branching lemma.

To see this, observe that the adjacency matrix of this network,

$$A = \begin{bmatrix} 0 & 2 & 0 & 0 \\ 2 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix},$$

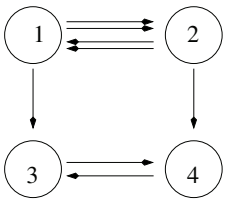


Figure 7. *A homogeneous four-cell network with* $\mathbf{Z}_2$ *symmetry.*

has eigenvectors

$$(1,1,1,1)^t, \qquad (1,-1,-1,1)^t, \qquad (0,0,1,-1)^t, \qquad (0,0,1,1)^t,$$

corresponding to the eigenvalues $2,-2,-1,1$. Bifurcations associated with the corresponding center subspace, which are covered by Theorem 6.3, lead to synchronous solutions. The first eigenvector leads to a fully synchronous solution, and the remaining three eigenvectors lead to solutions in the synchrony subspace corresponding to the balanced colorings in Figure 8. Note that the equivariant branching lemma discusses only the pattern of synchrony corresponding to the eigenvector $(0,0,1,1)^t$, which is associated with $\mathrm{Fix}(\mathbf{Z}_2)$.

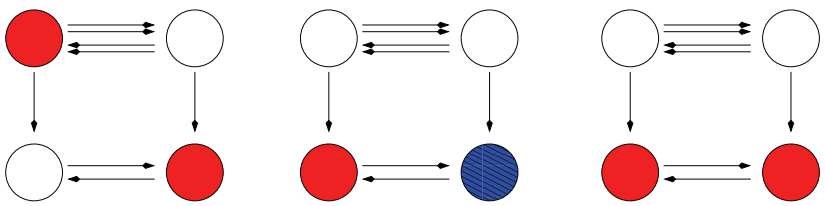


Figure 8. *Three nontrivial balanced colorings of the four-cell network with* $\mathbf{Z}_2$ *symmetry in Figure* 7.

If the internal dynamics is one-dimensional, the center subspace corresponding to the eigenvalues $-1,1$ cannot occur as the first bifurcation. However, if two-dimensional internal dynamics is assumed, then a bifurcation corresponding to each eigenvalue of A can be the first one to undergo bifurcation. Observe that the eigenvalues of $(dF)_0$ that correspond to the eigenvalues $2,-2,-1,1$ of A are the eigenvalues of the matrices $\alpha+2\beta$, $\alpha-2\beta$, $\alpha-\beta$, and $\alpha+\beta$, respectively. So to find a bifurcation associated with the eigenvalue 1 of A, we must find α and β so that the six eigenvalues of the three 2×2 matrices $\alpha+2\beta$, $\alpha-2\beta$, $\alpha-\beta$ all have negative real part, and one eigenvalue of $\alpha+\beta$ is 0, whereas the other one is negative. The matrices

$$\alpha = \begin{bmatrix} -1 & -2 \\ -2.5 & -4.5 \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

give such an example.

Theorem 6.6. *Suppose that bifurcations with axial polydiagonals in* E^{A_1} *and* E^{A_2} *lead to solutions with balanced colorings* $\bowtie_1$ *and* $\bowtie_2$*. Then bifurcations in the product network corresponding to the axial polydiagonal* $E^{A_1}\otimes E^{A_2}$ *will lead to solutions with balanced coloring* $\bowtie = \bowtie_1 \boxtimes \bowtie_2$.

A general theorem for networks that is analogous to the equivariant Hopf theorem [15] (EHT) is difficult to state. The issue concerns spatio-temporal symmetries. The EHT uses

the fact that Hopf bifurcation itself induces a natural $\mathbf{S}^1$ action corresponding to phase shift and the interaction of this $\mathbf{S}^1$ symmetry with the spatial symmetries of the system. In the general network context there is mounting evidence that periodic solutions with (robust) phase shift synchrony can occur only if an associated quotient network has nontrivial symmetry. See Stewart and Parker [22, 23, 24, 25]. However, it is not clear how to state a worthwhile analogue of EHT using this conjecture. Even in the special context of interior symmetries [11], where the analogue of EHT can easily be stated, the proof remained elusive and was given only recently by Antoneli, Dias, and Paiva [1].

Two examples illustrate some of the simpler difficulties: the three-cell unidirectional ring (Figure 1(left)) and the three-cell feed-forward network (Figure 4(left)).

Hopf bifurcation from a synchronous equilibrium in the unidirectional ring can occur with one-dimensional internal dynamics and leads, because of the $\mathbf{Z}_3$ symmetry, to solutions where any two cells oscillate a third of a period out of phase. It is a curious fact that bifurcation to synchronous oscillation can occur only if the internal dynamics is at least two-dimensional. These facts follow trivially from the center subspaces associated with the adjacency matrix,

$$A_{\text{udr}} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}. \tag{6.1}$$

The eigenvalues of A_{udr} are $\mu_1 = 1$, $\mu_2 = (-1+i\sqrt{3})/2$, and $\overline{\mu_2}$ with associated real eigenspaces

$$V_1 = \mathbf{R}\{(1,1,1)^t\} \quad \text{and} \quad V_2 = \{(x_1,x_2,x_3) : x_1+x_2+x_3 = 0\}. \tag{6.2}$$

It was shown in [10, 8] that synchrony-breaking Hopf bifurcation in the feed-forward network leads to two branches of periodic solutions. The amplitude of the solutions on one branch grows at the expected rate of $\lambda^{\frac{1}{2}}$ (where λ is the bifurcation parameter), and the amplitude of solutions along the other branch grows at the unexpected rate of $\lambda^{\frac{1}{6}}$. Neither the existence of multiple solutions nor the rapid growth rate is due to symmetry; rather they are due in part to multiple eigenvalues in the adjacency matrix caused by network architecture. The adjacency matrix of the feed-forward network is

$$A_{\text{ff}} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}. \tag{6.3}$$

The eigenvalues of A_{ff} are $\nu_1 = 1$ and $\nu_2 = 0$ with algebraic multiplicity 2 and geometric multiplicity 1. The associated real (generalized) eigenspaces are

$$W_1 = \mathbf{R}\{(1,1,1)^t\} \quad \text{and} \quad W_2 = \mathbf{R}\{(0,1,0)^t, (0,0,1)^t\}. \tag{6.4}$$

The nontrivial Jordan structure of A_{ff} is responsible for the existence of multiple solution branches in synchrony-breaking Hopf bifurcations in this network, but the complete picture is even more complicated. Elmhirst and Golubitsky [8] show that there are other regular networks that have Jordan structures in their adjacency matrices that are identical to that in the feed-forward network, but whose nonlinear bifurcations generate solutions with different multiplicities and different growth rates.

Nevertheless, once these solution structures are understood, they seem to lead naturally to solution structures in the product network, as we discuss in the next section.

7. Bifurcations from stable synchronous solutions. Until now we have classified the center subspaces of possible codimension one bifurcations from a fully synchronous equilibrium that can occur in regular cell systems, but we have not discussed whether these bifurcations can lead to stable solutions. More precisely, suppose that $\mu_1, \ldots, \mu_r$ are the distinct eigenvalues of the adjacency matrix. A synchrony-breaking bifurcation associated with the eigenvalue μ_j can lead to stable solutions only if there exist $k \times k$ matrices α, β such that all eigenvalues of the matrices M_{μ_i} for $i \neq j$ have negative real part and all noncritical eigenvalues of M_{μ_j} have negative real parts. If this necessary condition is satisfied at a codimension one bifurcation, we call that bifurcation a *first bifurcation*.

We conjecture that if k is large enough, there can be a codimension one synchrony-breaking first bifurcation associated with any eigenvalue μ_j of the adjacency matrix. However, we can prove this conjecture only for those regular networks whose adjacency matrices have real eigenvalues.

Theorem 7.1. *Assume that all eigenvalues of the adjacency matrix A of a regular network are real. Let $\mu_1 < \cdots < \mu_r$ be the distinct eigenvalues of A. Suppose that $k = 2$. Then, for any $1 \leq j \leq r$, there exist 2×2 matrices α and β such that the eigenvalues of M_{μ_i} have negative real part for all $i \neq j$ and the eigenvalues of M_{μ_j} are 0 and negative.*

Proof. Without loss of generality we can translate the numbers μ_i by the same constant c. More precisely,

$$\alpha + \mu\beta = (\alpha + c\beta) + (\mu - c)\beta.$$

So, without loss of generality, we may assume that $\mu_j = 0$.

The theorem is easily seen to be valid for $j = 1$, even when $k = 1$. Just take $\alpha = 0$ and $\beta = -1$. Similarly, when $j = r$, set $\alpha = 0$ and $\beta = 1$. So we can assume $1 < j < r$. In this case (when $k = 2$) we must find matrices α, β so that $\operatorname{tr}(M_{\mu_i}) < 0$ for all i and $\det(M_{\mu_i}) > 0$ for all $i \neq j$. We claim that this can be done by setting

$$\alpha = \begin{bmatrix} 0 & 0 \\ 0 & a \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} -1 & b \\ 1 & 0 \end{bmatrix},$$

where

$$a < \mu_1 \tag{7.1}$$

and

$$b < -\frac{a}{\mu_{j-1}}. \tag{7.2}$$

Observe that

$$\operatorname{tr}(\alpha + \mu\beta) = a - \mu \quad \text{and} \quad \det(\alpha + \mu\beta) = -a\mu - b\mu^2.$$

So assumption (7.1) implies that

$$\operatorname{tr}(M_{\mu_i}) = a - \mu_i < a - \mu_1 < 0$$

for all i, since $\mu_1 \leq \mu_i$. We claim $\det(M_{\mu_i}) > 0$ if $i \neq j$. To verify this point, note that for $i \neq j$, $\det(M_{\mu_i}) > 0$ if and only if

$$b < (-a)\frac{1}{\mu_i},$$

which is equivalent to

$$-\frac{a}{\mu_i} - b > 0$$

for all $i \neq j$. Since $-a > 0$, we need to choose b sufficiently negative, namely, less than $-a$ times the largest negative value of $\frac{1}{\mu_i}$. Because of the ordering of the μ_i, that value is $\frac{1}{\mu_{j-1}}$, which we have assumed in (7.2). ■

Suppose that the eigenvalues of the adjacency matrix of a regular network are real. When the dimension of the internal dynamics is $k \geq 2$, Theorem 7.1 implies that there is a codimension one steady-state first bifurcation associated with any eigenvalue μ. This statement is not true for Hopf bifurcations, at least when $k = 2$.

Suppose that a regular network has an adjacency matrix A with a negative eigenvalue μ_1, a zero eigenvalue, and a positive eigenvalue μ_r. Suppose also that $k = 2$. Then a Hopf bifurcation associated with the 0 eigenvalue of A cannot be a first bifurcation.

To verify this point, note that if the 2×2 matrix $M_0 = \alpha$ has purely imaginary eigenvalues, it follows that $\operatorname{tr}(\alpha) = 0$ and hence that

$$\operatorname{tr}(M_1) = \mu_1 \operatorname{tr}(\beta) \quad \text{and} \quad \operatorname{tr}(M_r) = \mu_r \operatorname{tr}(\beta).$$

If $\operatorname{tr}(\beta) \neq 0$, then the traces of the 2×2 matrices M_1 and M_r have opposite sign, and one of them must have an eigenvalue with positive real part. So the eigenvalue 0 of the adjacency matrix cannot correspond to a first (Hopf) bifurcation. If $\operatorname{tr}(\beta) = 0$, then $\operatorname{tr}(M_1) = 0$ and the bifurcation is either not codimension one (if $\det(M_1) \geq 0$) or not a first bifurcation (if $\det(M_1) < 0$). In short, to have a first Hopf bifurcation associated to a general adjacency matrix eigenvalue μ, one must assume that $k \geq 3$. Observations such as these have been made previously in the pattern formation literature (see, for example, [9]). On the other hand, if $k \geq 3$, then all types of Hopf bifurcation can occur as first bifurcations. The following theorem was proved by Amit Vutha [26].

Theorem 7.2. *Assume that all eigenvalues of the adjacency matrix A of a regular network are real. Let $\mu_1 < \cdots < \mu_r$ be the distinct eigenvalues of A. Suppose that $k = 3$. Then, for any $1 \leq j \leq r$, there exist 3×3 matrices α and β such that the eigenvalues of M_{μ_i} have negative real part for all $i \neq j$ and the eigenvalues of M_{μ_j} are purely imaginary and negative.*

Proof. We begin by recalling the Routh–Hurwitz criterion. Given a polynomial of degree 3,

$$\lambda^3 + a_1\lambda^2 + a_2\lambda + a_3 = 0.$$

All roots of this polynomial have negative real part if and only if

$$a_1 > 0, \qquad a_3 > 0, \qquad a_1a_2 > a_3.$$

The polynomial has a pair of purely imaginary roots if and only if

$$a_1 > 0, \qquad a_3 > 0, \qquad a_1a_2 = a_3.$$

We prove that the eigenvalue μ_j can have a first Hopf bifurcation when $k = 3$. As in the proof of Theorem 7.1, we may translate all of the eigenvalues μ_i so that $\mu_j = 0$. We wish to find 3×3 matrices α and β such that

(i) $\alpha + \mu_i\beta$ has all eigenvalues with negative real parts for $i \neq j$,
(ii) α has a pair of purely imaginary eigenvalues and a real negative eigenvalue.

We proceed by first choosing $a > 0$ satisfying $a > |\mu_i|$ for all i. Then let

$$\alpha = \begin{bmatrix} -a & a & -a \\ 0 & 0 & -a \\ 1 & 0 & 0 \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}.$$

Then

$$\alpha + \mu_i\beta = \begin{bmatrix} -a & a+\mu_i & -a \\ 0 & 0 & -a+\mu_i \\ 1 & 0 & 0 \end{bmatrix},$$

whose characteristic polynomial is

$$\lambda^3 + a\lambda^2 + a\lambda + (a^2 - \mu_i^2) = 0. \tag{7.3}$$

For $i = j$ we have $\mu_i = 0$, and (7.3) has roots $-a$ and $\pm i\sqrt{a}$, whereas for $i \neq j$, (7.3) has all roots with negative real part by the Routh–Hurwitz criterion. ■

8. An example of bifurcation in a product network. In this section we discuss the codimension one synchrony-breaking bifurcations in the nine-cell product network shown in Figure 4(right). For simplicity we assume that the internal dynamics is one-dimensional, even though this restricts some of the bifurcation types that are possible in the general case.

The adjacency matrices for this network are $A_{\text{udr}} \otimes I_3$ and $I_3 \otimes A_{\text{ff}}$ defined in (6.1) and (6.3). Suppose that a system of differential equations associated with this network is denoted by $\dot{X} = F(x)$ and that $J = (dF)_{X_0}$ is the Jacobian of this system at a synchronous equilibrium, which we can take to be $X_0 = 0$. From (3.1) it follows that

$$J = \alpha \mathbf{I}_3 \otimes \mathbf{I}_3 + \beta_1 A_{\text{udr}} \otimes \mathbf{I}_3 + \beta_2 \mathbf{I}_3 \otimes A_{\text{ff}}, \tag{8.1}$$

where α is the linearized internal dynamics and β_1, β_2 are the linearized coupling strengths for the couplings in the unidirectional ring and the feed-forward networks, respectively. It follows from the discussion in section 3 that the eigenvalues of J are $\alpha + \beta_1\mu + \beta_2\nu$, where μ is an eigenvalue of A_{udr} and ν is an eigenvalue of A_{ff}. As noted in section 6, the eigenvalues of A_{udr} are $\mu_1 = 1$, $\mu_2 = (-1 + i\sqrt{3})/2$, and $\overline{\mu_2}$, and the eigenvalues of A_{ff} are $\nu_1 = 1$ and $\nu_2 = 0$ with algebraic multiplicity 2 and geometric multiplicity 1.

The real parts of the eigenvalues of the 9×9 matrix J are

$$\begin{aligned} \sigma_{11} &= \alpha + \beta_1 + \beta_2, \\ \sigma_{21} &= \alpha - \tfrac{1}{2}\beta_1 + \beta_2, \\ \sigma_{12} &= \alpha + \beta_1, \\ \sigma_{22} &= \alpha - \tfrac{1}{2}\beta_1, \end{aligned}$$

with multiplicities 1, 2, 2, and 4. It is possible to choose α, β_1, β_2 so that any one of the $\sigma_{ij} = 0$ and the other three are negative. Sample values are given in the following table:

(8.2)

α	β_1	β_2	σ_{11}	σ_{21}	σ_{12}	σ_{22}
−4	2	2	0	−2	−1	−3
−3	−2	2	−3	0	−5	−2
−2	2	−2	−2	−5	0	−1
−1	−2	−2	−5	−2	−3	0

It follows that each of the four types of bifurcations can lead, in principle, to branches of stable solutions.

If σ_{11} is the critical eigenvalue, then the bifurcation occurs in the fully synchronous subspace with all coordinates equal. Generically we expect a saddle node bifurcation and no synchrony-breaking.

If σ_{21} is the critical eigenvalue, then the bifurcation occurs in the synchrony subspace where all cell coordinates in each ring are equal. In this case the quotient network is the feed-forward network, and the generic (steady-state) bifurcation was analyzed in [18]. Indeed, [18, Theorem 4.6] proves that two nontrivial branches of solutions bifurcate from the trivial branch, one transcritical and one a pitchfork. Moreover, when the internal dynamics is one-dimensional, generically, the bifurcating solutions are unstable. Note that if the product network has two-dimensional internal dynamics, then Hopf bifurcation associated with this center subspace is possible and periodic solutions with the $\lambda^{\frac{1}{6}}$ growth rate may occur.

If σ_{12} is the critical eigenvalue, then the bifurcation occurs in the synchrony subspace where all cell coordinates in each feed-forward network are equal. In this case the quotient network is the unidirectional ring, and we have purely imaginary eigenvalues at criticality. Hopf bifurcation to discrete rotating waves occurs.

If σ_{22} is the critical eigenvalue, then bifurcation occurs with four purely imaginary eigenvalues, and the center subspace is a four-dimensional real subspace. This is an example of (4.1) where U is complex one-dimensional, $\mathcal{G}_1(\mu)$ is complex two-dimensional, and $\mathcal{G}_2(\nu)$ is complex one-dimensional. The tensor product in (4.1) is complex two-dimensional, and the center subspace is real four-dimensional.

The Jordan form at bifurcation is identical to the Hopf bifurcation in the feed-forward network. Of course, the nonlinear bifurcation need not be identical—but it is. In particular, there are two solution branches. The first is found by setting the coordinates in the first two rings to 0. Then there is a standard Hopf bifurcation to a branch of discrete rotating waves in the third ring (because of the $\mathbf{Z}_3$ symmetry), whose amplitude grows at rate $\lambda^{\frac{1}{2}}$. Second, setting the coordinates in the upper ring to 0 implies that the middle ring equations undergo a standard Hopf bifurcation to a branch of discrete rotating waves (again because of the $\mathbf{Z}_3$ symmetry), whose amplitude also grows at rate $\lambda^{\frac{1}{2}}$. Finally, the equations in the third ring are forced by the periodic solutions in the middle ring, and just as in the three-cell feed-forward network there is a unique branch of periodic solutions in the bottom ring. These solutions also form a discrete rotating wave (because of symmetry and uniqueness) with amplitude growth $\lambda^{\frac{1}{6}}$. This may be checked using [8] or by a numerical experiment (we did the latter). Simulations using the cell vector field

$$f(y_1, y_2, y_3) = -0.3y_1 - y_2 - y_3 - y_1^3,$$

that is, $\beta_1 = \beta_2 = -1$, $\alpha = -0.5$, and $\lambda = 0.2$, are shown in Figure 9.

9. Conclusions. The number of different regular networks grows superexponentially in the number of cells r (even for fixed valency). Therefore, the bifurcation analysis for three-cell systems given in [18, 8] cannot be repeated for r-cell systems if the bifurcation analysis is truly different for each network. The results of this paper and those in [8] give some hope

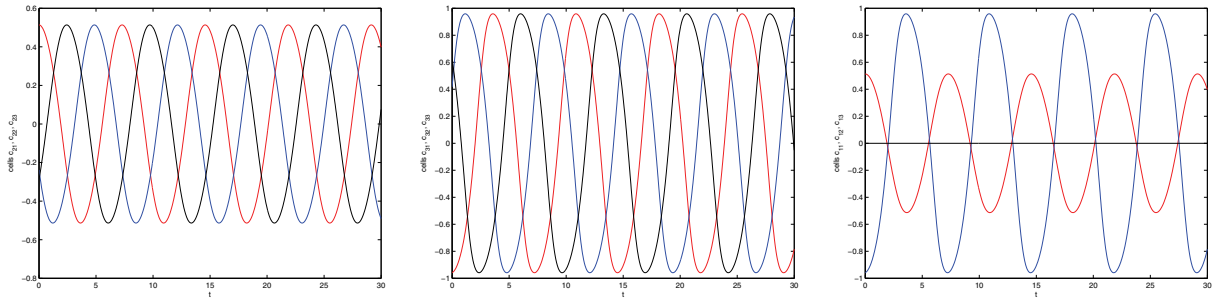


Figure 9. *Simulations from product network. (Left, center) Discrete rotating waves are seen in second and third rings. (Right) Time series from cells in the first feed-forward network; note the increase in amplitude from the second to the third cell in that chain.*

that bifurcations in general cell systems can be understood in a way analogous to equivariant bifurcation theory.

In equivariant bifurcation theory [15, 3, 12] it is known that each irreducible representation of a given group Γ leads to a type of Hopf bifurcation and that each absolutely irreducible representation of Γ leads to a type of steady-state bifurcation theory. Each of these bifurcations is analyzed using a set of techniques that have now become standard (normal form theory, singularity theory, equivariant branching lemma, the EHT, etc.). The important point is that the number of different bifurcations that need to be analyzed is indexed by the number of distinct irreducible representations of Γ and not by the number of representations of Γ.

Here we suggest that the number of bifurcations in regular networks that need to be analyzed is indexed by the number of Jordan structures of adjacency matrices and not by the number of networks. But the situation is not so clear as in the equivariant case. Elmhirst and Golubitsky [8] study *nilpotent Hopf bifurcations*, that is, those Hopf bifurcations where the algebraic multiplicities of the critical eigenvalues are each two but the geometric multiplicities are one. They show that there are (at least) three or four different (nonlinear) bifurcation scenarios in regular networks that can occur with such a fixed center subspace structure. We conjecture that each type of center subspace leads to just a small finite number of different bifurcation scenarios (unlike in equivariant bifurcations, where each irreducible representation leads to one bifurcation scenario), and we explain our reasoning using nilpotent Hopf bifurcations.

The codimension one nilpotent Hopf bifurcation scenarios for regular networks analyzed in [8] can be described by the number of bifurcating branches of periodic solutions and the growth rate of the amplitudes of solutions along these branches. The four possibilities are

(a) two branches with growth rates $\lambda^{\frac{1}{2}}$ and $\lambda^{\frac{1}{6}}$,
(b) two branches each with growth rate λ^{1},
(c) two or four branches each with growth rate $\lambda^{\frac{1}{2}}$, and
(d) two branches with growth rates $\lambda^{\frac{1}{2}}$ and $\lambda^{\frac{1}{4}}$,

where λ is the bifurcation parameter. The three-cell feed-forward network in Figure 4(left) is an example of (a). Other examples are given in [8]. The five-cell network in Figure 10(left) is an example where scenario (b) occurs. The three-cell network in Figure 10(right) is an example of scenario (c); other examples are given in [8]. As of now no examples of a network

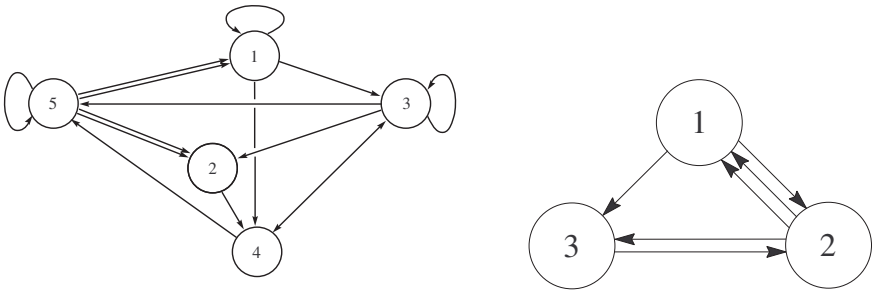


Figure 10. *Regular networks that exhibit nilpotent Hopf bifurcations in codimension one.*

that yields scenario (d) are known.

Krupa [17] observed that if one analyzes the normal form of the center manifold reduction to the four-dimensional center manifold of a nilpotent Hopf bifurcation, then all scenarios can be seen in the third order truncation. This suggests more strongly that generically, for any network, codimension one nilpotent Hopf bifurcations occur in at most four types in any regular network, but we have no proof of this fact. Moreover, the stability of these solutions is sometimes determined at order greater than 3. Nevertheless, these results together give some credence to the conjecture we have made in this section.

Acknowledgments. We thank Vern Paulsen for his help in proving Proposition 2.5, Martin Krupa for many helpful conversations, and Fatima Drubi for her careful reading of the original manuscript. We thank a referee for pointing out that our theory should also apply to bifurcation from synchronous limit cycles; see Remark 2.8. RL thanks the University of Houston for its hospitality.

REFERENCES

[1] F. Antoneli, A. P. S. Dias, and R. C. Paiva, *Hopf bifurcation in coupled cell networks with interior symmetries*, SIAM J. Appl. Dyn. Syst., 7 (2008), pp. 220–248.

[2] F. M. Atay and T. Biyikoğlu, *Graph operations and synchronization of complex networks*, Phys. Rev. E, 72 (2005), paper 016217.

[3] P. Chossat and R. Lauterbach, *Methods in Equivariant Bifurcations and Dynamical Systems*, Adv. Ser. Nonlinear Dynam. 15, World Scientific, Singapore, 2000.

[4] A. N. Comanici, *Synchrony-breaking period-doubling bifurcations for three-cell homogeneous coupled maps*, J. Bifur. & Chaos, to appear.

[5] R. Cushman and J. Sanders, *Nilpotent normal forms and representation theory of* $\mathbf{sl}(2,\mathbf{R})$, in Multiparameter Bifurcation Theory, M. Golubitsky and J. Guckenheimer, eds., AMS, Providence, RI, 1986, pp. 31–51.

[6] G. Dangelmayr, W. Güttinger, and M. Wegelin, *Hopf bifurcation with* $\mathbf{D}_3 \times \mathbf{D}_3$*-symmetry*, Z. Angew. Math. Phys., 44 (1993), pp. 595–638.

[7] B. Dionne, M. Golubitsky, and I. Stewart, *Coupled cells with internal symmetry.* II: *Direct products*, Nonlinearity, 9 (1996), pp. 575–599.

[8] T. Elmhirst and M. Golubitsky, *Nilpotent Hopf bifurcations in coupled cell systems*, SIAM J. Appl. Dyn. Syst., 5 (2006), pp. 205–251.

[9] B. Ermentrout and M. Lewis, *Pattern formation in systems with one spatially distributed species*, Bull. Math. Biol., 59 (1997), pp. 533–549.

[10] M. Golubitsky, M. Nicol, and I. Stewart, *Some curious phenomena in coupled cell networks*, J. Nonlinear Sci., 14 (2004), pp. 207–236.
[11] M. Golubitsky, M. Pivato, and I. Stewart, *Interior symmetry and local bifurcation in coupled cell networks*, Dyn. Syst., 19 (2004), pp. 389–407.
[12] M. Golubitsky and I. Stewart, *The Symmetry Perspective: From Equilibrium to Chaos in Phase Space and Physical Space*, Birkhäuser Boston, Cambridge, MA, 2002.
[13] M. Golubitsky and I. Stewart, *Nonlinear dynamics of networks: The groupoid formalism*, Bull. Amer. Math. Soc., 43 (2006), pp. 305–364.
[14] M. Golubitsky, I. Stewart, P. L. Buono, and J. J. Collins, *A modular network for legged locomotion*, Phys. D, 115 (1998), pp. 56–72.
[15] M. Golubitsky, I. Stewart, and D. G. Schaeffer, *Singularities and Groups in Bifurcation Theory* II, Appl. Math. Sci. 69, Springer-Verlag, New York, 1988.
[16] M. Golubitsky, I. Stewart, and A. Török, *Patterns of synchrony in coupled cell networks with multiple arrows*, SIAM J. Appl. Dyn. Syst., 4 (2005), pp. 78–100.
[17] M. Krupa, *private communication*.
[18] M. C. A. Leite and M. Golubitsky, *Homogeneous three-cell networks*, Nonlinearity, 19 (2006), pp. 2313–2363.
[19] L. M. Pecora and T. L. Carroll, *Master stability functions for synchronized coupled systems*, Phys. Rev. Lett., 80 (1998), pp. 2109–2112.
[20] D. H. Sattinger, *Group Theoretic Methods in Bifurcation Theory*, Lecture Notes in Math. 762, Springer-Verlag, Berlin, 1979.
[21] I. Stewart, M. Golubitsky, and M. Pivato, *Symmetry groupoids and patterns of synchrony in coupled cell networks*, SIAM J. Appl. Dyn. Syst., 2 (2003), pp. 609–646.
[22] I. Stewart and M. Parker, *Period dynamics of coupled cell networks* I: *Rigid patterns of synchrony and phase relations*, Dyn. Syst., 22 (2007), pp. 389–450.
[23] I. Stewart and M. Parker, *Period dynamics of coupled cell networks* II: *Cyclic symmetry*, Dyn. Syst., 23 (2008), pp. 17–41.
[24] I. Stewart and M. Parker, *Period dynamics of coupled cell networks* III: *Rigid phase patterns*, Dyn. Syst., to appear.
[25] I. Stewart and M. Parker, *Period Dynamics of Coupled Cell Networks* IV: *Cyclic Symmetry in Non-Vertex Trivial Networks*, preprint.
[26] A. Vutha, *private communication*, Department of Mathematics, The Ohio State University, Columbus, OH.

SIAM J. Applied Dynamical Systems
Vol. 8, No. 1, pp. 76–128
© 2009 Society for Industrial and Applied Mathematics

Quadratic Volume-Preserving Maps: Invariant Circles and Bifurcations*

H. R. Dullin† and J. D. Meiss‡

Abstract. We study the dynamics of the five-parameter quadratic family of volume-preserving diffeomorphisms of $\mathbb{R}^3$. This family is the unfolded normal form for a bifurcation of a fixed point with a triple-one multiplier and is also the general form of a quadratic three-dimensional map with a quadratic inverse. Much of the nontrivial dynamics of this map occurs when its two fixed points are saddle-foci with intersecting two-dimensional stable and unstable manifolds that bound a spherical "vortex-bubble." We show that this occurs near a saddle-center-Neimark–Sacker (SCNS) bifurcation that also creates, at least in its normal form, an elliptic invariant circle. We develop a simple algorithm to accurately compute these elliptic invariant circles and their longitudinal and transverse rotation numbers and use it to study their bifurcations, classifying them by the resonances between the rotation numbers. In particular, rational values of the longitudinal rotation number are shown to give rise to a *string of pearls* that creates multiple copies of the original spherical structure for an iterate of the map.

Key words. volume-preserving map, incompressible flow, bifurcation, saddle-center, Hopf, Neimark–Sacker

AMS subject classifications. 37J20, 37G05, 34C28

DOI. 10.1137/080728160

1. Introduction.

The area-preserving Hénon map [Hén69] is the universal form for quadratic diffeomorphisms of the plane and provides perhaps the prototype for conservative systems with chaotic dynamics. As such it has deservedly been the subject of much study. We believe that the natural generalization of this map to three dimensions is the volume and orientation-preserving diffeomorphism

$$f(x,y,z) = \begin{pmatrix} x+y \\ y+z-\varepsilon+\mu y+P(x,y) \\ z-\varepsilon+\mu y+P(x,y) \end{pmatrix}, \tag{1}$$

where P is the quadratic form

$$P(x,y) = ax^2+bxy+cy^2. \tag{2}$$

Just as the Hénon map arises as the normal form for a saddle-center bifurcation, the map f is the normal form for the equivalent bifurcation in $\mathbb{R}^3$: a triple-one multiplier [DM08]. More

*Received by the editors June 23, 2008; accepted for publication (in revised form) by T. Kaper September 16, 2008; published electronically January 9, 2009.
http://www.siam.org/journals/siads/8-1/72816.html
†School of Mathematics and Statistics, The University of Sydney, Sydney, NSW 2006, Australia (hdullin@usyd.edu.au). This author was supported in part by a Leverhulme Research Fellowship, and would like to thank the Department of Applied Mathematics in Boulder for its hospitality.
‡Department of Applied Mathematics, University of Colorado, Boulder, CO 80309-0526 (James.Meiss@colorado.edu). This author was supported in part by NSF grant DMS-0707659 and by the Mathematical Sciences Research Institute in Berkeley.

generally, this normal form is given by (1) to any order, when cubic and higher degree terms are added to P. Also like the Hénon map on $\mathbb{R}^2$, the map f on $\mathbb{R}^3$ is the universal form for the family of quadratic diffeomorphisms that have quadratic inverses [LM98]. Indeed, for any polynomial P, f is a diffeomorphism whose inverse has the same degree as P. Finally, both the Hénon map and f arise as normal forms for certain homoclinic bifurcations [GMO06]; see section 2.

In this paper we will study some of the dynamics of the map (1). Numerical evidence presented in section 4 will show that this map has a nonzero measure of bounded orbits primarily near the simultaneous saddle-center and Neimark–Sacker (SCNS) bifurcation that occurs when $\varepsilon = 0$ and $-4 < \mu < 0$. We will study the normal form for this bifurcation in section 5. The bounded orbits of f that appear in such a regime are built around the skeleton formed from its two saddle-focus fixed points, their stable and unstable manifolds, and the elliptic invariant circle that is created when this bifurcation is supercritical.

As we recall in section 5.1, the structure of this "saddle-center-Hopf" bifurcation is well known for the case of a volume-preserving flow [Bro81, Hol84]. For the supercritical case, the bifurcation creates a vortex-bubble structure analogous to the Hill vortex of fluid mechanics [Lam45] or the spheromak configuration of plasma physics [GICH80]. There are two saddle-focus equilibria whose two-dimensional stable and unstable manifolds coincide, forming a sphere. The interior of the sphere is foliated by a family of two-tori enclosing an invariant circle that is normally elliptic. For the map, the normal form of the SCNS bifurcation is no longer integrable, but it still has two saddle-foci, a Cantor family of tori, and an elliptic invariant circle; see section 5.2.

The quadratic map (1) is approximately described by this normal form for small ε and away from low-order resonances. We observe numerically that many of the invariant two-tori and the central invariant circle appear to persist for moderate values of ε. In this paper we will concentrate on the persistence and bifurcations of the elliptic invariant circle. We develop an algorithm to accurately compute this circle in section 6. When it is elliptic, the circle has two rotation numbers $\omega = (\omega_L, \omega_T)$, longitudinal and transverse, respectively. We compute these and compare the numerical results with the normal form calculations.

Resonances of the form $m \cdot \omega = n$ lead to bifurcations that may give rise to new elliptic invariant circles and may result in the destruction or change of stability of the original circle. Interestingly, there are two types of doubling (or m-tupling) bifurcations, one in which the new invariant circle is a single circle that winds multiple times around the original circle (like in a flow), the other one in which k new invariant circles appear that are mapped to each other. The latter case appears when m_1 and m_2 are not coprime.

Perhaps the most interesting of these bifurcations we call a *string of pearls*. It occurs when ω_L becomes rational and typically results in a new SCNS bifurcation for some power of f; see section 7. When this bifurcation is supercritical, the invariant circle is replaced by a set of small vortex-bubbles or *pearls* linked by nearly coincident one-dimensional invariant manifolds of a pair of saddle-focus periodic orbits, the *string*. Recently a similar bifurcation has been observed for dissipative three-dimensional maps [BSV08].

2. Contexts.

The quadratic map (1) arises naturally in at least three contexts.

Suppose that $f : \mathbb{R}^3 \to \mathbb{R}^3$ is a smooth volume- and orientation-preserving map, $\det Df(\xi) = 1$, where $\xi = (x, y, z)$. If ξ^* is a point on a period-n orbit, $f^n(\xi^*) = \xi^*$, then the multipliers

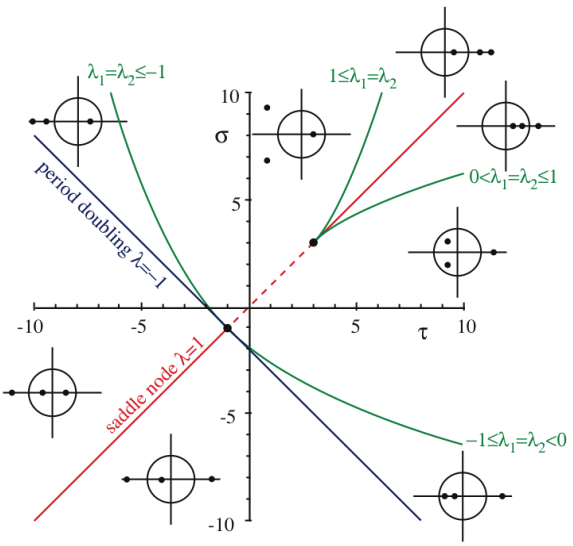


Figure 1. *Classification of the eigenvalues for a three-dimensional; volume-preserving map as a function of the trace τ and second trace σ. The eight insets are the complex planes showing the multiplier configurations relative to the unit circle.*

of this orbit must satisfy $\lambda_1\lambda_2\lambda_3 = 1$. Consequently, the characteristic polynomial of the linearization about the periodic orbit,

$$\det(\lambda I - Df^n(\xi^*)) = \lambda^3 - \tau\lambda^2 + \sigma\lambda - 1, \tag{3}$$

contains two parameters: the trace τ and second trace σ. There are eight stability regimes in the (τ, σ) plane; see Figure 1. The boundaries between these regimes contain two codimension-two points; one corresponds to a triple multiplier $\{1, 1, 1\}$ at $(\tau, \sigma) = (3, 3)$ and the second to multipliers $\{-1, -1, 1\}$ at $(\tau, \sigma) = (-1, -1)$. These two cases form organizing centers for the dynamics in all eight regimes. Notice that in the interior of each of the eight regions the fixed point is hyperbolic and unstable.

If the Jacobian, $Df^n(\xi^*)$, has a multiplier $\lambda = 1$ with algebraic multiplicity three, then λ generically has a one-dimensional eigenspace (i.e., geometric multiplicity one); consequently, the Jacobian is similar to the Jordan block

$$J = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}. \tag{4}$$

We showed in [DM08] that near ξ^*, f^n is formally conjugate to the normal form

$$\begin{aligned} (x', y', z')^T &= J\,(x, y, z + p(x, y, \varepsilon, \mu))^T, \\ p(x, y, \varepsilon, \mu) &= -\varepsilon + \mu_1 x + \mu_2 y + a(\varepsilon, \mu)x^2 + b(\varepsilon, \mu)xy + c(\varepsilon, \mu)y^2 + \cdots, \end{aligned} \tag{5}$$

where the $\cdots$ indicates higher-order terms in x and y only. The normal form is "formal" in the sense that if f is expanded in a power series to any finite degree d in the variables ξ and a set of sufficiently general parameters, then this degree-d map is conjugate to (5). Generically,

one of the two parameters (μ_1, μ_2) in (5) can be eliminated (see Appendix A), leaving two unfolding parameters; for example, if $a \neq 0$, then μ_1 can be set to zero, leaving the two parameters $(\varepsilon, \mu = \mu_2)$.

The normal form provides a remarkable simplification of the full map since all of the nonlinearity is contained in the single polynomial p that depends upon only two of the variables. To lowest nonlinear degree, the map (5) is quadratic, and if we view (a, b, c) as independent parameters, it becomes (1).

The quadratic map (1) also arises in the study of polynomial diffeomorphisms: it was shown in [LM98] that any quadratic diffeomorphism of $\mathbb{R}^3$ that has a quadratic inverse and nontrivial dynamics is affinely conjugate to the *shift-like* [BP98] map

$$(x, y, z) \mapsto (-\varepsilon + \tau x + \sigma y + z + Q(x, y), x, y)\,,$$

where $Q(x, y)$ is a quadratic form. This map is linearly conjugate to the quadratic case of (5) under the orientation reversing transformation $\xi \mapsto U\xi$, where

$$U = \begin{pmatrix} 0 & 1 & 0 \\ 1 & -1 & 0 \\ 1 & -2 & 1 \end{pmatrix},$$

providing that $Q(x, y) = p(y, x - y)$, $\tau = \mu_2 + 3$, and $\sigma = \mu_1 - \mu_2 - 3$. Thus the map (5) is also a normal form in the sense of polynomial automorphisms. However, it is not known how to generalize this result to cubic or higher degree—unlike the planar case that was treated by Friedland and Milnor [FM89]. Quadratic diffeomorphisms may have inverses of degree up to four; these were classified for $\mathbb{C}^3$ in [FW98, Mae01]. Our map (5) does have an inverse of the same degree for any polynomial P.

The map (5) also appears in a third context as the normal form for certain homoclinic bifurcations [GMO06]. In particular, consider a three-dimensional map with a saddle-focus fixed point such that the saddle value—the product of the multipliers of the fixed point—is 1; this, of course, is true for the volume-preserving case. Suppose that the one-dimensional invariant manifold of the saddle-focus has a quadratic tangency with its two-dimensional (spiral) manifold. Unfolding this singularity leads to a map of the form (5) that describes the dynamics of the return map in a neighborhood of the homoclinic point. There are two distinct cases: when the saddle-focus is of type $(2, 1)$ (or *type-A* [Cow73]), with a two-dimensional (spiral) stable manifold and a one-dimensional unstable manifold, then the homoclinic normal form has $P(x, y) = a(x + y)^2$, so that the quadratic form (2) is a perfect square. When the saddle-focus is of type $(1, 2)$ (or *type-B*), then $P(x, y) = ax^2$.

An additional parameter can be introduced in (1) to make $\det Df = b \neq 1$. Maps similar to this often have strange attractors and have been studied in [ACS83, AA84, GOST05, DLM06].

These considerations support our assertion that the normal form (5) is an appropriate *generalized Hénon* map for three-dimensional dynamics.

3. Fixed points. We begin the study of the dynamics of (1) looking at its fixed points; more details are given in Appendix C.

When $\varepsilon a < 0$ the quadratic map (1) has no fixed points; when $\varepsilon a > 0$ it has two at

$$\xi_\pm = (x_\pm, 0, 0), \quad x_\pm = \pm\sqrt{\frac{\varepsilon}{a}}. \tag{6}$$

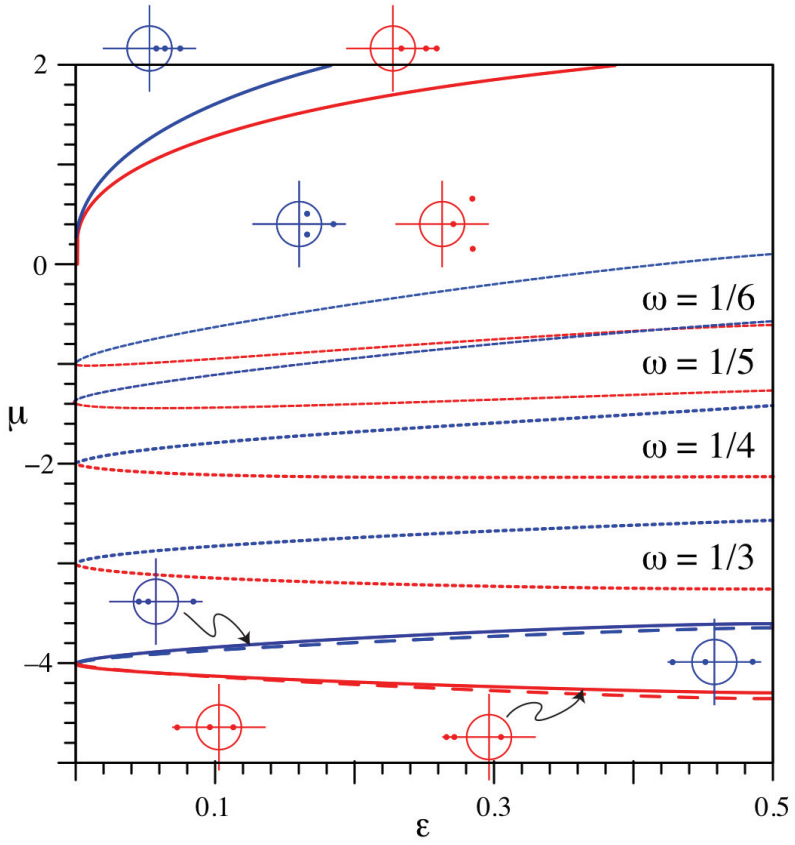


Figure 2. *Stability diagram for the map (1) with $a = 1$, $b = 0.5$, and c arbitrary. The solid curves correspond to a double eigenvalue, the dashed curves to period doubling, and the dotted curves to a complex conjugate pair at $re^{\pm 2\pi i\omega}$ with ω as indicated. Bifurcation curves for the fixed point x_+ are blue, and those for x_- are red.*

Thus a pair of fixed points is created upon crossing the *saddle-center* line $\varepsilon = 0$. Stability of these fixed points is determined from the general stability diagram, Figure 1, by computing (τ, σ) for the fixed points; see Appendix C. For fixed $(a \neq 0, b, c)$, the stability of each fixed point depends on the two parameters (ε, μ). When $a > 0$ the stability diagram of x_+ is a diffeomorphic copy of the half-plane $\tau > \sigma$; the corresponding diagram for x_- covers the half-plane $\tau < \sigma$; see Figure 2. In this figure, the blue curves correspond to the fixed point x_+ and the red curves to x_-. Note that over most of the parameter range, when one of the fixed points is type $(2, 1)$ (two-dimensional stable manifold), the other is type $(1, 2)$ (two-dimensional unstable manifold); see Appendix C.

When $\varepsilon = 0$ and $-4 < \mu < 0$, the fixed points are created with a pair of multipliers $\lambda_\pm = e^{\pm 2\pi i\omega_0}$ on the unit circle with rotation number

$$\omega_0(\mu) = \frac{1}{\pi} \arcsin \sqrt{-\frac{\mu}{4}}. \tag{7}$$

As we will see next, this range of μ seems also to correspond to the range for which there are other nontrivial bounded orbits for the map.

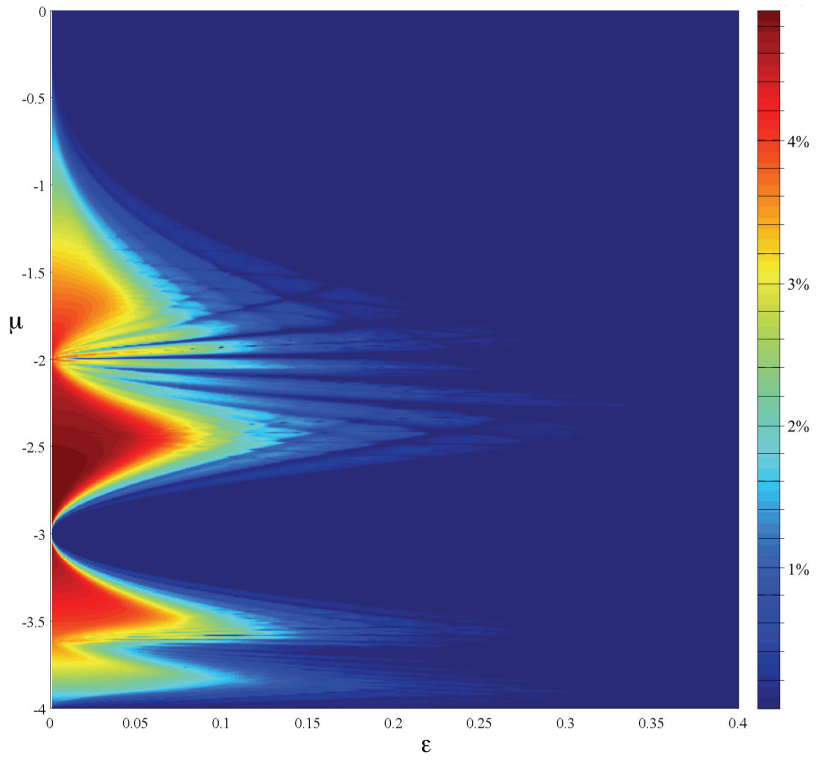


Figure 3. *Fraction of bounded orbits in $\mathcal{B}$ for the map (1) with nonlinearity (2) and $a = 1.0$ and $b = c = \frac{1}{2}$. Orbits are "bounded" if they stay inside the ball of radius 10 for $N = 10^4$ iterates. For this calculation, the map was iterated a total of 1.69×10^{15} times over $500^3 \times 500^2$ initial conditions. The maximum number of trapped orbits was 6139040, corresponding to 4.91% of the initial conditions in $\mathcal{B}$ when $(\varepsilon, \mu) = (0.00078, -2.73)$. When $\varepsilon = 0.4$ there are no bounded orbits except near $\mu = -2.26$ and -3.95, where less than 2×10^{-5} of the initial conditions are bounded.*

4. Bounded orbits. When the quadratic form P of (2) is positive definite, it is not hard to show that all of the bounded orbits of (1) are contained in a cube; see Appendix B. To find regions of parameter space that have nontrivial bounded orbits, we iterated a grid of initial conditions, declaring an orbit to be "bounded" if it remained in a ball of radius 10 for N iterations. The resulting fraction of bounded orbits as a function of the parameters (ε, μ) is shown in Figure 3. For this figure, as for most of the numerical studies reported below, we fixed the parameters of P, choosing $a = 1$, $b = c = \frac{1}{2}$.

Our numerical studies indicate that the domain containing bounded orbits shrinks to zero as $\sqrt{\varepsilon}$. To reflect this, we considered initial conditions on a three-dimensional grid in the box

$$\mathcal{B} \equiv [-2x_+, 2x_+] \times [-4x_+, 4x_+] \times [-8x_+, 8x_+], \tag{8}$$

where x_+ is the position of the fixed point (6). In Figure 3 we plot the fraction of bounded orbits with $N = 10^4$, choosing initial conditions on a cubical grid of 500^3 points that covers $\mathcal{B}$. It appears that there are bounded orbits only when $\varepsilon \geq 0$ and $-4.1 \lesssim \mu < 0$, though we

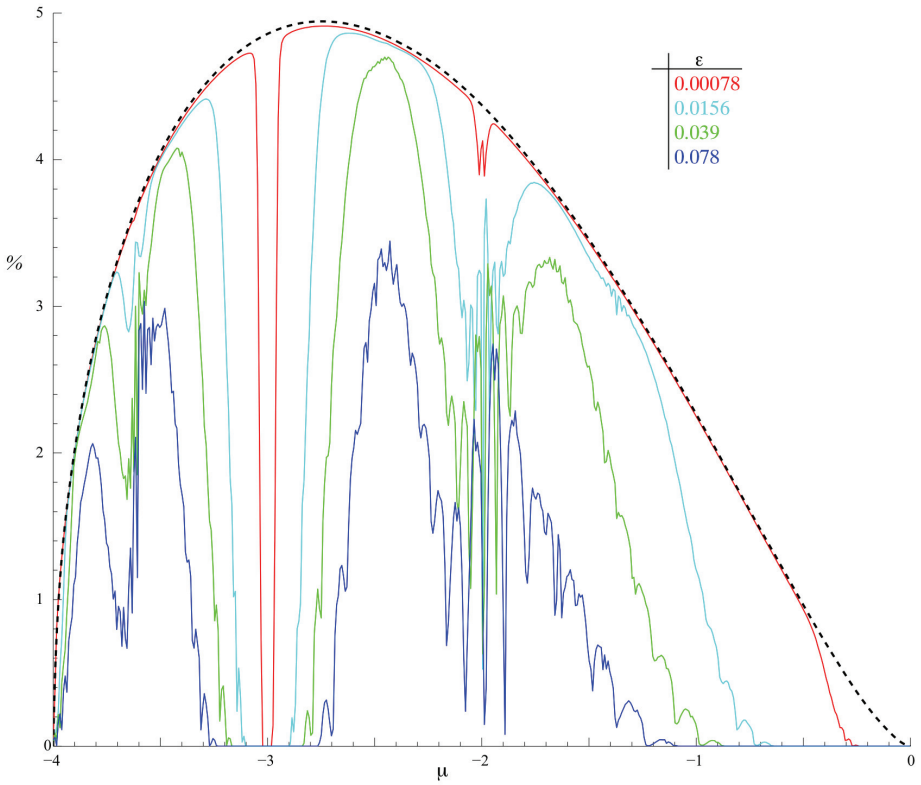


Figure 4. *Percentage of the box $\mathcal{B}$ that is occupied by bounded orbits as a function of μ for several values of ε. The dashed curve is the theoretical result (23) that is valid as $\varepsilon \to 0^+$, away from the major resonances.*

have not exhaustively searched outside this parameter domain. We also observe that, for the parameters of Figure 3, there appear to be essentially no bounded orbits when $\varepsilon \gtrsim 0.5$.

In the figure, the dark blue regions correspond to "no" bounded orbits, and the darkest red to the maximal fraction, 4.9%. There is considerable "resonance tongue" structure in the figure. Some of this is similar to Figure 2 for the multipliers of the fixed point. In particular the bounded fraction appears to be nearly zero at the doubling and tripling points $(\varepsilon, \mu) = (0, -4)$ and $(0, -3)$, respectively, and is small near the quadrupling point $(\varepsilon, \mu) = (0, -2)$. Indeed, if we increase N to 10^6 and double the grid resolution, then we find no bounded orbits in $\mathcal{B}$ whenever $\varepsilon \geq 0.001$ for $\mu = -3$ and at most 0.043% when $\mu = -4$.

Several constant-ε slices through the full data set are shown in Figure 4. Note that as $\varepsilon \to 0^+$, the bounded fraction appears to converge to a smooth curve, shown as the dashed line, except for the resonances near $\mu = 0$, -2, and -3. This curve will be derived in section 5.3.

Much of the structure of Figures 3 and 4 can be attributed to the orbits trapped by the two-dimensional manifolds of the fixed points, as we discuss below.

5. SCNS bifurcation. In most of the regions with bounded orbits in Figure 3, the fixed points (6) are spiral foci with two-dimensional stable and unstable manifolds, respectively. These manifolds appear to intersect transversely for many parameter regimes and, roughly speaking, enclose a ball that appears to contain all of the bounded orbits [LM98]. The

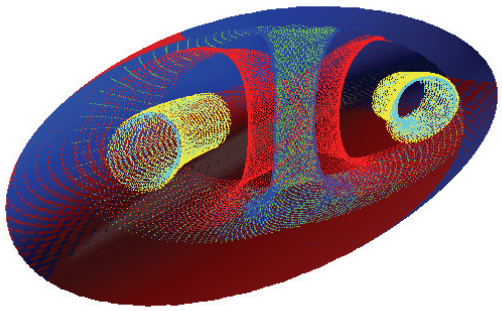


Figure 5. *Cutaway view of several orbits of the map* (1) *for* $a = 1.0$, $b = c = 0.5$, *and* $(\varepsilon, \mu) = (0.01, -2.4)$. *Also shown are the two-dimensional stable (blue) and unstable (red) manifolds of the fixed points.*

structure is analogous to a vortex bubble in an incompressible fluid [Lam45, Mac94]. Such a flow is generic for a volume-preserving vector field near a saddle-center-Hopf bifurcation, as we review in section 5.1. The topology of the intersections changes as the parameters vary, but it most often seems to include infinite spiral curves that connect the fixed points. This is necessarily true for a flow since every intersection point lies on a homoclinic orbit; however, for a map a number of different topological types of intersections can also occur [LM00].

The most prominent bounded orbits correspond to a family of invariant two-tori enclosing an invariant circle that lies approximately in the plane $x = 0$; see Figure 5. If the rotation number (7) of the fixed points is not close to a low-order rational number, then as ε tends to zero this structure limits on a ball that shrinks to zero as $\sqrt{\varepsilon}$. The stable and unstable manifolds in this limit appear to nearly coincide, and the family of invariant tori nearly completely fills the interior of the ball: the structure appears to be nearly "integrable."

Since the bounded orbits of the map (1) occur primarily along the saddle-center-Hopf line, we discuss in section 5.2 the normal form for a map near such a bifurcation. We will show that both supercritical and subcritical bifurcations can occur. In the supercritical case, the normal form creates a pair of fixed points whose two-dimensional stable and unstable manifolds intersect and enclose a sphere that generically contains a Cantor-family of invariant two-tori. In section 5.3, we will transform (1) into normal form, thereby obtaining the relationship between its parameters and those of the normal form. This will demonstrate that this bifurcation occurs on the saddle-center-Hopf segment of Figure 2.

Before proceeding to the discussion of the map, we review the standard results for volume-preserving flow in section 5.1. The time $\sqrt{\varepsilon}$ flow of this vector field will be shown to approximate the map in section 5.2.

5.1. Saddle-center-Hopf bifurcation. For a system of differential equations, the saddle-center-Hopf or Gavrilov–Guckenheimer bifurcation is the codimension-two bifurcation that occurs when an equilibrium has simultaneously one zero and one pair of imaginary eigenvalues. The three-dimensional, center manifold reduction to normal form is discussed in depth in [DI96, DIK01, GH02, Kuz04]. The key simplification is that the "formal" normal form has cylindrical symmetry to all orders in the power series expansion: it exhibits the symmetry of the linearized system. This symmetry is typically only formal and is broken by terms

"beyond-all-orders." For a system of divergence-free differential equations this bifurcation is codimension-one, and its unfolding is considerably simpler [Bro81].

In the neighborhood of an equilibrium point with eigenvalues $(2\pi i\omega, -2\pi i\omega, 0)$, the normal form can be most easily obtained in the complex coordinates, $(u, v = \bar{u}, z)$, that diagonalize the linearization. The flow of the linear system then commutes with the rotation $u \mapsto u\exp(2\pi i\varphi)$. Consequently, in symplectic cylindrical coordinates (r, θ, z) with

$$u = \sqrt{2r}e^{2\pi i\theta}, \quad v = \sqrt{2r}e^{-2\pi i\theta}, \tag{9}$$

the linearized vector field has the form

$$V = \omega\partial_\theta.$$

The normal form has the same symmetry, and since the rotation axis $r = 0$ is a fixed set of the symmetry, it is invariant under the dynamics. The normal form vector field therefore has the form

$$V = rF(r, z)\partial_r + \Omega(r, z)\partial_\theta + Z(r, z)\partial_z,$$

which is divergence-free when $\partial_z Z + \partial_r(rF) = 0$.[1] The divergence-free condition implies that the two-dimensional projection of a vector field onto (r, z) is Hamiltonian with

$$H(r, z) = rG(r, z)$$

such that $F = -G_z$ and $Z = G + rG_r$.[2] Thus the normal form is a skew-product system of the form

$$\begin{aligned} \dot{r} &= -\frac{\partial H}{\partial z}, \\ \dot{\theta} &= \Omega(r, z), \\ \dot{z} &= \frac{\partial H}{\partial r}. \end{aligned} \tag{10}$$

Now consider the dynamics of (10) near the origin. The first few terms in a power series expansion of the Hamiltonian are

$$H(r, z) = r(A_{0,0} + A_{1,0}r + A_{0,1}z + A_{0,2}z^2 + \cdots).$$

When $A_{0,2} \neq 0$, the implicit function theorem implies that the coefficient $A_{0,1}$ can be eliminated by an affine shift in z. In the new coordinates, we replace $A_{0,0}$ by $-\delta$; this represents the unfolding parameter. The shape of the contours of H near the origin for small δ is determined by the privileged scaling $r = \mathcal{O}(\delta)$ and $z = \mathcal{O}(\sqrt{\delta})$. In this case, as $\delta \to 0$, H is equivalent to

$$H(r, z) = r(-\delta + \beta r + \alpha z^2) + \mathcal{O}(\delta^{5/2}). \tag{11}$$

[1] The divergence in the new coordinates looks Euclidean because the transformation to symplectic cylindrical coordinates is volume-preserving.

[2] Indeed, any three-dimensional, volume-preserving flow with a Lie symmetry has an invariant that gives rise to a Hamiltonian structure on the projection of the manifold by the group orbits [HM98]. Specifically, suppose V has the Lie symmetry W, i.e., $[V, W] = 0$. Then if both V and W preserve the volume form Ω, we have $dH = -i_V i_W \Omega$, and the symplectic form is $\omega = -i_W \Omega$.

When $\alpha\beta\delta \neq 0$, the shape of the contours of H depends only upon the two signs

$$s_1 = \operatorname{sgn}(\delta\alpha), \quad s_2 = \operatorname{sgn}(\alpha\beta). \tag{12}$$

Indeed, the transformation

$$\rho = 2\left|\frac{\beta}{\delta}\right| r, \quad \zeta = \left|\frac{\alpha}{\delta}\right|^{\frac{1}{2}} z, \quad \tau = ht, \quad h \equiv \sqrt{|\delta\alpha|}\,\operatorname{sgn}\alpha \tag{13}$$

leads to the scaled Hamiltonian

$$\tilde{H}(\rho,\zeta) = -s_1\rho + \frac{s_2}{2}\rho^2 + \rho\zeta^2 + \mathcal{O}(\sqrt{\delta}), \tag{14}$$

which has no continuous parameters. The implication is that there are two types of saddle-center-Hopf bifurcation; see Figure 6. When $s_2 = 1$, three equilibria are created as $\delta\alpha$ changes from negative to positive. Two, at $(\rho,\zeta) = (0,\pm 1)$, are saddle-foci of the three-dimensional flow when Ω is nonzero. The third equilibrium is a center at $(1,0)$; this corresponds to an elliptic invariant circle in $\mathbb{R}^3$.

By contrast, when $s_2 = -1$, the bifurcation is subcritical: the saddle-foci exist when $s_1 < 0$; they annihilate at $\delta = 0$; and a hyperbolic invariant circle is created for $s_1 > 0$.

The separatrix of stable and unstable manifolds of the saddles is the contour $H = 0$; it is the ζ-axis together with the parabola $\rho = 2s_2\left(s_1 - \zeta^2\right)$. When $s_1 = s_2 = 1$ this separatrix encloses a region of closed trajectories that become, for the three-dimensional flow, a family of two-dimensional invariant tori that surround the invariant circle. In the original unscaled variables, the volume of the vortex-bubble as delimited by the invariant manifolds of the fixed points on the symmetry axis can be easily computed to be

$$V_H = \int_{H<0} dr \wedge d\theta \wedge dz = \frac{8\pi\delta}{3\beta}\sqrt{\frac{\delta}{\alpha}}. \tag{15}$$

However, if $s_1 = s_2 = -1$, then the stable and unstable manifolds of the saddles are unbounded and there is no heteroclinic connection apart from the ζ-axis.

On any invariant torus, the dynamics is conjugate to a rigid rotation. Since the flow of H is integrable, there exist angle-action coordinates (ϕ, I) that are valid inside the separatrix. In these coordinates the normal form vector field would become

$$V = \nu(I)\partial_\phi + \Omega(\phi, I)\partial_\theta,$$

where $\nu(I) = \partial H/\partial I$ is the Hamiltonian frequency. Thus the dynamics of the cylindrical angle can be trivially solved to obtain

$$\theta(t) = \theta(0) + \frac{1}{\nu(I)}\int_{\phi(0)}^{\phi(t)} \Omega(\phi, I)d\phi.$$

Consequently, the winding number on the torus is

$$\frac{1}{\nu(I)}\left\langle \Omega(\phi, I)\right\rangle_\phi.$$

As we will see below, the flow is a good model of the map for $\delta \ll 1$. Consequently, the structure that we observed in Figure 5 corresponds to the creation of a bubble of bounded orbits in a supercritical saddle-center-Hopf bifurcation.

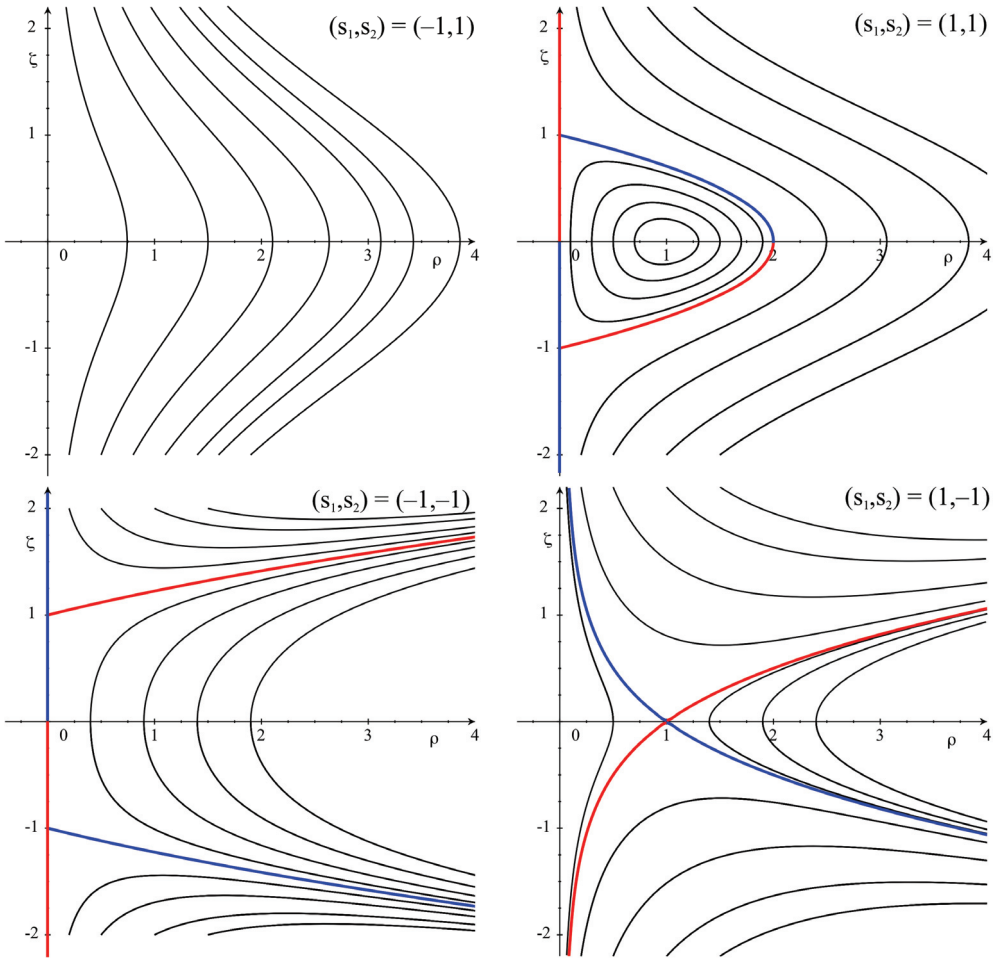


Figure 6. *Contours of the Hamiltonian* (14) *for the four choices of the signs* $s_1 = \operatorname{sgn}\delta\alpha$ *and* $s_2 = \operatorname{sgn}\alpha\beta$. *The upper two panes illustrate the supercritical case,* $s_2 = 1$, *and the lower two the subcritical case,* $s_2 = -1$. *Stable and unstable manifolds of the saddle equilibria are shown in blue and red, respectively.*

5.2. SCNS bifurcation. Consider a three-dimensional volume-preserving map with a fixed point whose multipliers are

$$\lambda_1 = \bar{\lambda}_2 = \lambda \equiv e^{2\pi i\omega}, \quad \lambda_3 = 1, \tag{16}$$

where, without loss of generality, $\omega \in (0, \frac{1}{2})$. As for the flow case, it is convenient to use complex coordinates to diagonalize the linearization: let $\zeta = (u, v, z)$, where u is the complex eigen-coordinate for λ_1, $v = \bar{u}$ for λ_2, and z is the real eigen-coordinate for λ_3.

The dynamics in the neighborhood of this fixed point can be studied by a standard normal form analysis [Mur03, BC93]. We summarize the aspects that are new for the volume-preserving case in Appendix D.

As for the flow case, the normal form, (43), commutes to all orders with the phase shift $u \to ue^{i\psi}$. Consequently, the dynamics of $|u|$ and of z are independent of those of the

argument of u. To make this explicit, we use the symplectic, cylindrical coordinates (9) so that if $u = x + iy$, the volume form obeys

$$dx \wedge dy \wedge dz = \frac{i}{2} du \wedge d\bar{u} \wedge dz = dr \wedge d\theta \wedge dz. \tag{17}$$

To cubic degree, the volume-preserving normal form obtained in Appendix D, (43), is

$$\begin{aligned} r' &= r\left(1 - 2\alpha z - (\gamma + 2\alpha\beta)r + (4\alpha^2 - 3\kappa)z^2\right), \\ \theta' &= \theta + \Omega(r, z), \\ z' &= -\delta + z + \alpha z^2 + 2\beta r + 2\gamma r z + \kappa z^3, \end{aligned} \tag{18}$$

where

$$2\pi\Omega(r, z) = 2\pi\omega + A_i z + 2C_i r + (\alpha A_i + B_i)z^2 + (2\alpha C_i - 2A_i C_r)rz.$$

Here $4C_r \equiv -\gamma - 2\alpha\beta$, but the remaining eight parameters $(\alpha, \beta, \gamma, \delta, \kappa, A_i, B_i, C_i)$ are independent. We summarize this result as the following theorem.

Theorem 1. *Suppose that f_p is a family of C^3 volume-preserving maps of $\mathbb{R}^3$ and that f_0 has a fixed point x^* for which*

- *$Df_0(x^*)$ has eigenvalues $(e^{2\pi i\omega}, e^{-2\pi i\omega}, 1)$;*
- *$\omega \notin \{0, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}\}$;*
- *the vector $\frac{\partial f_p(x^*)}{\partial p}|_{p=0}$ is not in the range of $Df_0(x^*) - I$.*

Then for p and $x - x^*$ sufficiently small there is a coordinate change so that f_p reduces to (18) through cubic order.

The (r, z) components of the map (18) are independent of θ; consequently, the (r, z) map is area-preserving to $\mathcal{O}(3)$. A typical phase portrait for this two-dimensional map is shown in Figure 7.

As $\delta \to 0$, the (r, z) projection of (18) is approximately equivalent to the time h map of the flow of the Hamiltonian (14). To see this, introduce the scaled coordinates (13) to the map to obtain

$$\begin{aligned} \rho' &= \rho - 2h\rho\zeta + \mathcal{O}(h^2), \\ \zeta' &= \zeta + h(-s_1 + \zeta^2 + s_2\rho) + \mathcal{O}(h^2). \end{aligned}$$

This, to $\mathcal{O}(h)$, is the flow of (14); consequently, the analysis of section 5.1 implies that the phase portrait of the map will asymptotically approach the pictures shown in Figure 6 as $\delta \to 0$.

To obtain the portrait in more detail, we study the map directly. Since we have expanded in a power series, the valid fixed points of (18) emerge from the origin when $\delta = 0$. Assuming that $\alpha\beta \neq 0$, there are three such fixed points:

$$\begin{aligned} (0, z_\pm) &= \left(0, \pm\sqrt{\frac{\delta}{\alpha}}\right) + \mathcal{O}(\delta^{3/2}), \\ (r_c, z_c) &= \left(\frac{\delta}{2\beta}, \frac{C_r\delta}{\alpha\beta}\right) + \mathcal{O}(\delta^2), \end{aligned} \tag{19}$$

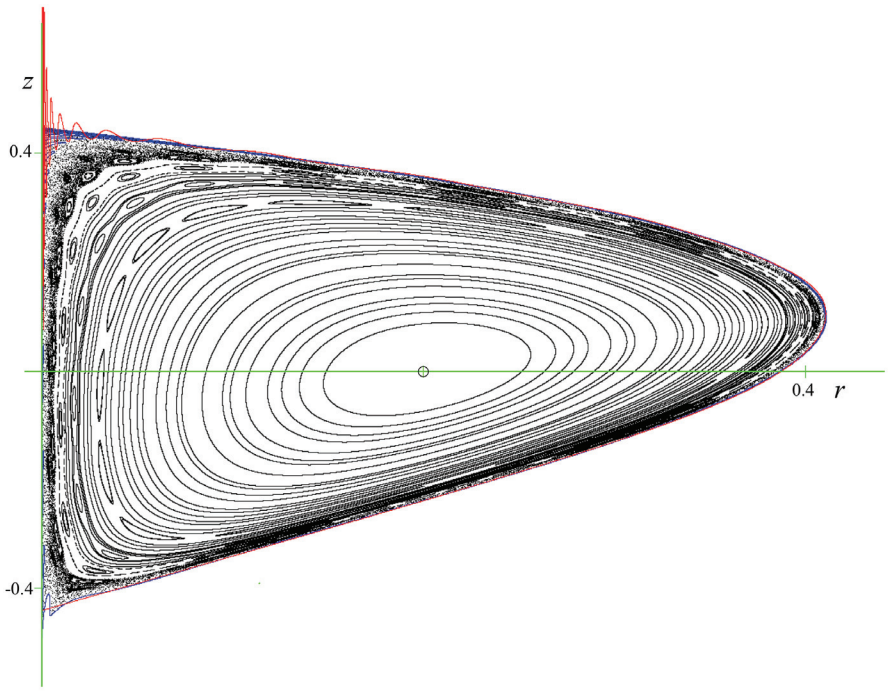


Figure 7. *Phase portrait of the map (18) for $\delta = 0.2$, $\alpha = \beta = 1$, and $C_r = \gamma = \kappa = 0$ in the (r, z) plane. The unstable manifold of the lower saddle is shown in red, and the stable manifold of the upper saddle in blue. Since C_r has been set to zero, the invariant circle is artificially at $z = 0$.*

where C_r is given by (44).

The stability of the fixed points can be classified by computing the "residue"

$$R \equiv \frac{1}{4}(2 - \operatorname{tr} Df) = \alpha\beta r - \alpha^2 z^2 + \mathcal{O}(3).$$

Recall that fixed points are hyperbolic saddles when $R < 0$, elliptic when $0 < R < 1$, and hyperbolic reflection-saddles when $R > 1$.

The fixed points on the z-axis exist when $\delta\alpha > 0$ and have residue

$$R_\pm = -\delta\alpha + \mathcal{O}(\delta^{3/2}).$$

Thus these points are always hyperbolic saddles when they exist. The z-axis corresponds to the one-dimensional stable/unstable manifolds of these two fixed points with the corresponding multiplier $\lambda_\pm = 1 + 2\alpha z_\pm + \mathcal{O}(\delta^{3/2})$. Thus when $\alpha > 0$ the z-axis is the unstable manifold of the upper fixed point and the stable manifold of the lower fixed point. For the three-dimensional map, the other two multipliers form a complex conjugate pair,

$$\lambda_\pm = (1 - 2\alpha z_\pm) e^{2\pi i \omega_\pm} + \mathcal{O}(\delta^{3/2}),$$

provided that $\omega_\pm = \Omega(0, z_\pm) \neq 0$. The graph of two-dimensional invariant manifolds for these

fixed points has the form[3]

$$z = W(r) = z_\pm - \frac{\beta}{2\alpha z_\pm} r - \frac{\beta^2}{8\alpha^2 z_\pm^3} r^2 + \mathcal{O}(r^3).$$

Note that the manifold of the upper (lower) fixed point has negative (positive) slope when $\alpha\beta > 0$; these manifolds are thus inclined towards each other and, as we observed in Figure 7, they generally intersect, enclosing, roughly speaking, a topological sphere. When $\alpha\beta < 0$ the manifolds are inclined away from each other and, as $\delta \to 0$, there is no local intersection of the manifolds.

The third fixed point corresponds to an invariant circle $\mathcal{C}$ of the three-dimensional map (18). Since $r \geq 0$, the circle exists only when $\delta\beta > 0$. The dynamics restricted to the circle is a rigid rotation with rotation number

$$\omega_L = \Omega(r_c, z_c) = \omega + \frac{1}{2\pi\alpha\beta} \left(\alpha C_i + C_r A_i\right) \delta + \mathcal{O}(\delta^2). \tag{20}$$

We call this the *longitudinal rotation number* of $\mathcal{C}$.

The invariant circle corresponds to a fixed point in the (r, z) plane with residue

$$R_c = \frac{\alpha\delta}{2} + \mathcal{O}(\delta^2).$$

Consequently if $\alpha\beta < 0$, the invariant circle is transversely hyperbolic. Alternatively, if $\alpha\beta > 0$, the invariant circle is elliptic when $\alpha\delta < 2$. Nearby circles in the (r, z) plane have a rotation number defined through $R = \sin^2(\pi\omega_T)$,

$$\omega_T = \frac{\sqrt{2\alpha\delta}}{2\pi} + \mathcal{O}(\delta^{3/2}). \tag{21}$$

This is the *transverse rotation number* of the invariant circle $\mathcal{C}$.

5.3. Quadratic map near the saddle-center-Hopf line. When ε is small and $-4 < \mu < 0$, the normal form (1) can be transformed to the saddle-center-Hopf form, (18). As this transformation is somewhat tedious, we use computer algebra to perform the manipulations.

The fixed point at the origin for $\varepsilon = 0$ and $-4 < \mu < 0$ has the linearization

$$Df = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1+\mu & 1 \\ 0 & \mu & 1 \end{pmatrix}.$$

To apply the results of section 5.2 this matrix is transformed to the diagonal form M, (38), using

$$T = \begin{pmatrix} \frac{1}{\mu} & \frac{1}{\mu} & 1 \\ \frac{\lambda}{\lambda-1} & -\frac{1}{\lambda-1} & 0 \\ 1 & 1 & 0 \end{pmatrix}.$$

[3]To make sense of this expansion, we should use the scaled variables (13). This scales the bubble to fixed size as $\delta \to 0$. In this case each of the three terms in the series is $\mathcal{O}(\delta^0)$.

Applying this transformation to (1) gives the new map $T^{-1}\circ f\circ T$. Note that the Jacobian of this transformation is $\det T = i\cot\pi\omega$, so the volume in the transformed coordinate system must be scaled by this factor. The transformation ψ to the new coordinates is constructed by Lie series and is obtained first to $\mathcal{O}(\varepsilon^0)$ through second order in ξ and then to first order in ε. As discussed in Appendix D, this gives a map of the form (47). A final affine transformation on z restores the magnitude of the multipliers to 1, giving a map of the form (18) with parameters

(22)
$$\alpha = -\frac{a}{\mu},\quad \beta = -\frac{2a+(b-2c)\mu}{\mu^3},\quad \gamma = \frac{2(a-b)(3+\mu)}{\mu(4+\mu)}\beta,\quad \delta = -\frac{\varepsilon}{\mu},$$
$$A_i = \frac{a-b}{\sqrt{-\mu(4+\mu)}},$$
$$B_i = -\frac{\big(4(\mu+2)(\mu+3)a^2+\mu(4(\mu+4)c-(5\mu+16)b)a+b^2\mu(\mu+2)\big)}{2\mu(-\mu(\mu+4))^{3/2}},$$
$$C_i = \frac{-2(\mu+1)a^2+\mu(2c(5\mu+13)-b(3\mu+5))a-\mu\left(4\mu(2\mu+5)c^2-2b\mu(3\mu+7)c+b^2(\mu+1)^2\right)}{2\mu^4(\mu+3)\sqrt{-\mu(\mu+4)}},$$
$$\omega = \omega_0(\mu)+\varepsilon\frac{b^2-4ac}{4\pi a\mu\sqrt{-\mu(4+\mu)}}+\mathcal{O}(\varepsilon^2),$$

where $\omega_0(\mu)$ is given in (7). Note that these coefficients have singularities when $\mu\in\{-4,-3,0\}$ corresponding to the resonant frequencies $\omega\in\{\frac12,\frac13,0\}$, respectively.[4]

Thus the dynamics of (1) near the saddle-center-Hopf line are approximated by the normal form (18) away from the low-order resonances. The signs (12) that determine the character of the bifurcation now become

$$s_1 = \operatorname{sgn}(\delta\alpha) = \operatorname{sgn}(\varepsilon a),$$
$$s_2 = \operatorname{sgn}(\alpha\beta) = \operatorname{sgn}(a)\operatorname{sgn}(2a+(b-2c)\mu).$$

Recall from section 5.1 that the bifurcation is super(sub)critical for $s_2 = 1$ $(=-1)$; see Figure 6. For the parameter values $(a,b,c)=(1,\frac12,\frac12)$ that we have primarily studied, $s_2=1$ for all $\mu<0$; thus the bifurcation is always supercritical. Consequently as ε increases through zero, an elliptic invariant circle and a pair of saddle-focus fixed points whose two-dimensional manifolds intersect are created. This is confirmed by numerical computations; for example, in Figure 5 where $\varepsilon=0.01$, the map is nearly integrable and the ball enclosed by the two-dimensional manifolds is predominantly filled with invariant tori. As ε increases, the intersection angle between the two-dimensional stable and unstable manifolds of the saddle-foci grows, many of the tori are destroyed, and orbits near these manifolds become unbounded; see Figure 8. Along the way there are many torus bifurcations that create new families helically wound around the original invariant circle. We postpone a discussion of these until section 7.

[4]The value $(b-2c)\mu+2a=0$ is also special since at this point both β and γ vanish. This corresponds to the boundary between supercritical and subcritical saddle-center-Hopf bifurcations. To study the dynamics at this point would require computing higher-order terms.

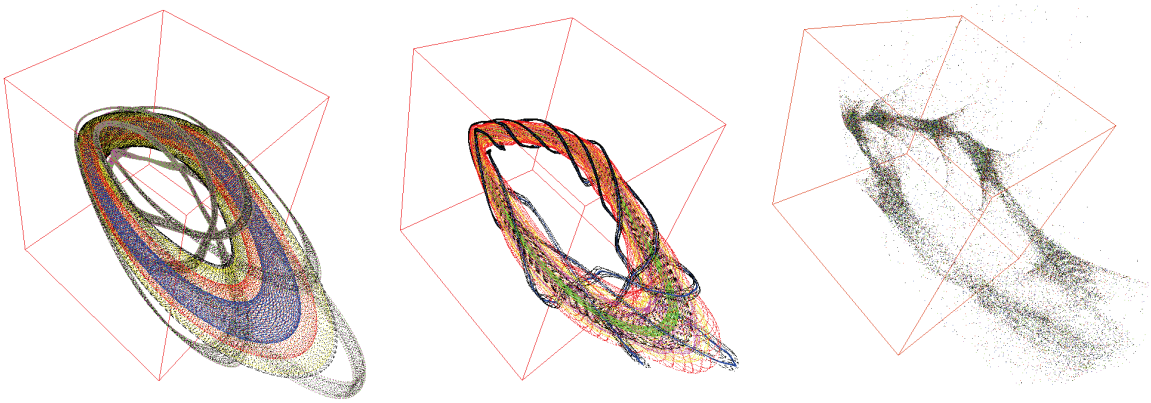


Figure 8. *Orbits of the map (1) for $\mu = -2.4$, $a = 1$, $b = c = \frac{1}{2}$, and $\varepsilon = 0.1$ (left), 0.2 (center), and 0.32 (right). In the last case there appear to be no bounded orbits; however, there is probably a pair of period-7 saddles. The orbits that are shown lie near the stable and unstable manifolds of these orbits. The red cubes are centered on the origin and have sides with length $\frac{10}{3}\sqrt{\varepsilon}$.*

Meanwhile, these results permit comparison with Figure 4 near $\varepsilon = 0$. Taking into account the scaling of the volume due to the transformation T and that due to the complex transformation (17), the fraction of bounded orbits in the box $\mathcal{B}$, (8), is

$$V_{\text{rel}} = 2i\frac{V_H}{\text{vol}(\mathcal{B})\det T} = \frac{\pi a}{96\beta}\sqrt{\frac{4+\mu}{-\mu^3}}, \tag{23}$$

where V_H is the volume (15). This formula gives the dashed curve shown in Figure 4. Since $\beta = O(\mu^{-3})$, (23) implies that the bounded fraction is $O(\mu^{3/2})$ when $\mu \to 0^-$. The actual results in Figure 4 deviate from this result when $|\mu| < 0.05$ due to the $\omega = 0$ resonance. Moreover, the resonances near $\mu = -2$ and -3 are also not captured by this lowest-order formula. However, (23) agrees well with the numerical results near the period-doubling point $\mu = -4$.

From (22) we can compute the position of the fixed points and invariant circle (19). These expressions are rather complicated and not especially useful. However, the rotation numbers of the invariant circle will be used extensively in the following sections. For the values $(a, b, c) = (1, \frac{1}{2}, \frac{1}{2})$ these become

$$\begin{aligned}
\omega_L &= \omega_0(\mu) + \frac{5\mu^3 + 19\mu^2 - 28\mu - 64}{16\pi\mu(\mu+3)(\mu-4)(-\mu(\mu+4))^{3/2}}\epsilon + \mathcal{O}(\varepsilon^2), \\
\omega_T &= -\frac{1}{\pi\mu}\sqrt{\frac{\varepsilon}{2}}\left(1 - g(\mu)\varepsilon + \mathcal{O}(\varepsilon^2)\right), \\
g(\mu) &\equiv \frac{58\mu^7 + 648\mu^6 - 377\mu^5 - 23526\mu^4 - 59027\mu^3 + 139032\mu^2 + 689616\mu + 695808}{96\mu^3(\mu-4)^2(\mu+3)^2(\mu+4)^2}.
\end{aligned} \tag{24}$$

As a first step in understanding the invariant tori, we will study the evolution of the invariant circle. As we will see, resonances between the rotation numbers (ω_L, ω_T) will be key to this discussion.

6. Invariant circles. In a volume-preserving flow, the supercritical saddle-center-Hopf bifurcation creates a pair of saddle-focus fixed points, one of type $(2,1)$ and one of type $(1,2)$ separated by $\mathcal{O}(\sqrt{\delta})$; recall section 5.1. The two-dimensional unstable and stable manifolds of these fixed points coincide in the formal normal form and bound a ball that is foliated by a family of two-dimensional invariant tori. These tori enclose an invariant circle that has a diameter of size $\mathcal{O}(\sqrt{\delta})$.

As we showed in section 5.2, the analogous bifurcation for a volume-preserving map has a normal form (18) that is approximated by this flow when $\delta \ll 1$. However, the cylindrical symmetry of both normal forms is only formal: even when δ is very small, the invariant circles presumably exist only for a Cantor set of parameter values when their rotation numbers satisfy appropriate Diophantine conditions [CS90a, Xia96].

Our goal in this section is to study numerically the invariant circles for the quadratic map (1) that are created in the SCNS bifurcation. In order to do this, we must first develop an algorithm to accurately compute elliptic invariant circles.

6.1. Ellipsoid algorithm. A number of algorithms have been proposed for finding invariant circles and tori; some of the history is discussed in [SOV05, HDLL06b]. Many of the techniques assume that the dynamics on the invariant circle is conjugate to a rigid rotation with a given rotation number ω_L. In this case, the conjugacy can be expanded in a Fourier series and a Newton method employed to compute the Fourier coefficients. Castello and Jorba have used this idea for Hamiltonian systems that have a Cantor set of circles whose frequencies vary [CJ00]. If the gaps in the Cantor set are "small," they can fix a rotation number and search for the corresponding circle. Haro and de la Llave [HDLL06a, HDLL06b, HDLL07] also use spectral methods to find tori in quasi-periodically forced maps. The important simplification for the forced case is that the rotation vector and the conjugacy to rigid rotation are automatically known. A very different method was developed by Edoh and Lorenz [EL03] to find nonsmooth but attracting invariant circles; here the asymptotic stability of the invariant circles is key. A related method is to define a variational method, using the distance between the circle and its image, and obtain the circle by following the gradient flow [LCC06].

An iteration-based method was developed by Simó [Sim98]. Letting $T = \frac{1}{\omega_L}$ be the "irrational period," Simó defined the Tth iterate of the map by interpolation between the points $f^{\lfloor T \rfloor}$ and $f^{\lceil T \rceil}$.[5] An invariant circle corresponds, roughly speaking, to a fixed point of f^T. An advantage of this method is that it does not depend upon the existence of a conjugacy. It can also be used to determine ω_L by fixing a section and looking for approximate returns to define T [CJ00].

Our method is also iteration-based but, in contrast to Simó's idea, uses a slice to localize iterates and a number of points to locate the invariant circle. Suppose that $\mathcal{C}$ is an elliptic invariant circle with a local cross section Σ; for (1), the section

$$\Sigma = \{(x, y, 0) : y > 0\}$$

works in most cases. Let $\Sigma_\Delta = \{|z| < \Delta, y > 0\}$ be a thin slice enclosing Σ. Assume that there is a neighborhood of $\mathcal{C}$ in which the local dynamics predominantly lies on two-dimensional

[5] Note that these points may be far apart, and so the interpolation may not be accurate.

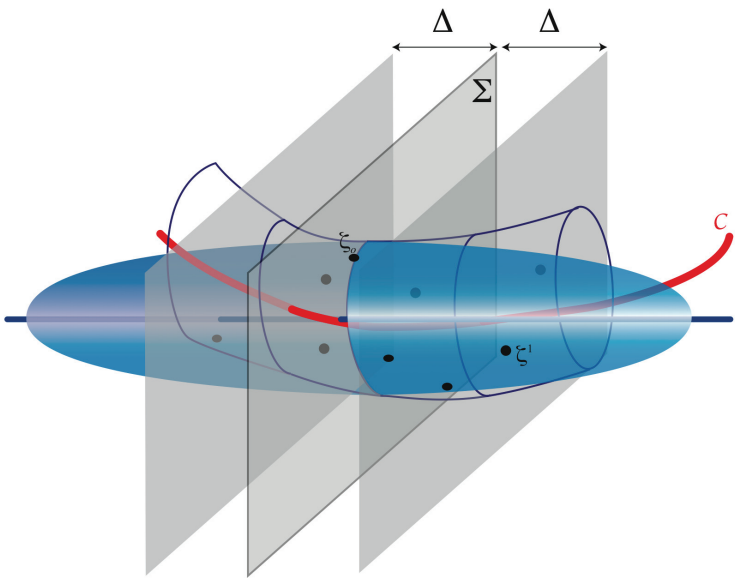


Figure 9. *Finding an invariant circle by fitting a section of a nearby torus to an ellipsoid.*

invariant tori that enclose $\mathcal{C}$. Thus the intersection of a typical orbit near $\mathcal{C}$ with Σ_Δ is a slightly curved, elliptic cylinder; see Figure 9.

Begin with a point $\zeta_0 \in \Sigma$ in the neighborhood of $\mathcal{C}$ and find a set of N returns of its orbit to Σ_Δ:

$$\zeta^i = f^{t_i}(\zeta_0) \in \Sigma_\Delta, \quad 0 < t_1 < t_2 < \cdots < t_N, \quad i = 1, \ldots, N.$$

By assumption, these points lie on an invariant torus; the goal is to find the "axis" of this torus. Instead of fitting the points to an elliptic cylinder, it is easier to fit them to a general ellipsoid, defined as level set of a quadratic form

$$e(x, y, z) = Ax^2 + By^2 + Cz^2 + Dxy + Exz + Fyz + Gx + Hy + Jz.$$

Since the returning points are, presumably, all close to the circle and since they lie on the ellipsoid by hypothesis, consider the N differences $\delta\zeta^i = \zeta^i - \zeta_0$. In this coordinate system the ellipsoid goes through the origin, so $e(\delta\zeta^i) = 0$. Since all the $\delta\zeta^i$ approach zero when the method converges, we normalize by $l = \frac{1}{N}\sum_{i=1}^{N} |\delta\zeta^i|$ and introduce $\delta\xi^i = \delta\zeta^i/l$. This removes the singular behavior from the algorithm that originates from the fact that the fitted ellipsoid becomes thinner, and it also helps avoid round-off errors.

Since the equations $e(\delta\xi^i) = 0$ are homogeneous in the coefficients, an additional equation is needed to fix the solution; we chose to normalize the sum of the squared coefficients of e, giving $N + 1$ linear equations in the coefficients:

$$\begin{aligned} e(\delta\xi^i) &= 0, \quad i = 1, \ldots, N, \\ A + B + C &= 1. \end{aligned}$$

We must choose at least $N = 8$ to fix the 9 coefficients of e. It seems most convenient to use

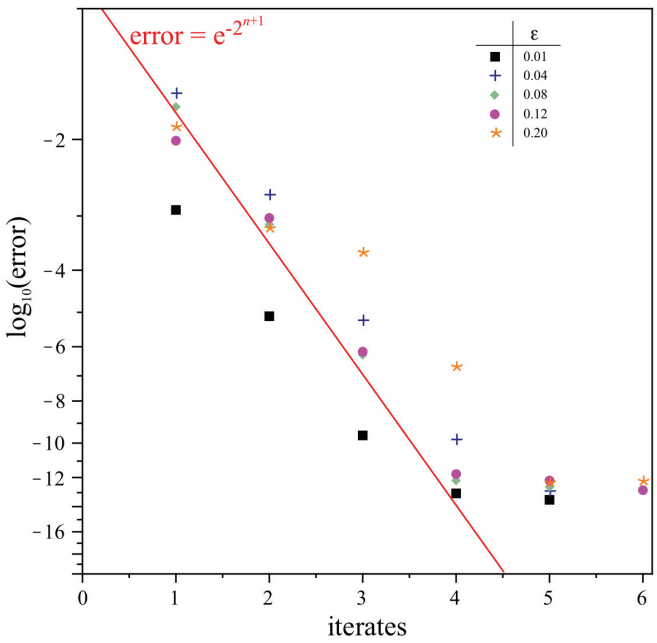


Figure 10. *Error on a log(log) scale for successive iterates of the invariant circle position for* $\mu = -2.4$*, with* $\Delta = 10^{-7}$ *and the values of* ε *indicated. The line represents an error of* $e^{-2^{n+1}}$*, which would mean quadratic convergence. With this value of* Δ*, the minimal error seems to be about* 10^{-13}*, and is achieved in* 4–6 *iterations.*

exactly 8 returns, since the number of iterates to find a return can be quite large, especially when the slice thickness, Δ, is small.[6]

Once the ellipsoid is determined, the center of the ellipse in the section Σ is used as the next guess for a point on $\mathcal{C}$; for the section $\Sigma = \{z = 0\}$, this gives the iterative step

$$\zeta_0 \mapsto \zeta_0 + l\frac{1}{4AC - B^2}(BH - 2CG, BG - 2AH, 0).$$

The whole process is then repeated, finding a new set of returns to the slice Σ_Δ and fitting a new ellipsoid. The error is estimated as the change in distance between the new point and the previous one. The iteration stops either when the error decreases below an error tolerance (we typically chose 10^{-10}) or after a fixed number of iterations (we chose 10). If the given tolerance is not obtained for a fixed Δ, we decrease Δ by a factor of 100 and then restart the algorithm with the best previous value as the initial point ζ_0 (the results typically converged when $\Delta = 10^{-6}$).

This process appears to be quadratically convergent; see Figure 10. Indeed, it is not hard to show that this is the case for a two-dimensional version of this algorithm for finding an elliptic fixed point: since the linearized map has invariant ellipses, the linearization of our iteration at a solution is superstable.

[6]This could be a problem if the points do not lie in a general position; to get around this one could use more points and solve the system using least squares.

Table 1

Position $(x, y, 0)$ on Σ and rotation numbers of the elliptic invariant circle of (1) for $\mu = -2.4$ and $(a, b, c) = (1, \frac{1}{2}, \frac{1}{2})$ as functions of ε. The error is the distance on Σ between the last and the penultimate iterates. We attempted to find a circle with an error $< 10^{-10}$ using a slice thickness $\Delta \leq 10^{-6}$.

ε	x	y	error	ω_L	ω_T
0.01	-0.000077100956	0.110334966059	1.6e-11	0.282171317669	0.00939482
0.02	-0.000218667148	0.156587932592	2.7e-11	0.282294227518	0.01331187
0.03	-0.000401253015	0.192348589289	7.8e-12	0.282415802529	0.01633514
0.04	-0.000606226180	0.222722556427	1.5e-11	0.282536006132	0.01889745
0.05	-0.000263469053	0.250886913814	1.1e-11	0.282654815793	0.02117252
0.06	-0.001281607227	0.273786477131	1.4e-13	0.282772154554	0.02324300
0.07	-0.001578623424	0.296421783954	8.4e-11	0.282888017858	0.02516009
0.08	-0.001928385257	0.317549729412	9.6e-12	0.283002357933	0.02695743
0.09	-0.002313763896	0.337472296148	3.3e-12	0.283115135201	0.02865947
0.10	-0.002732253247	0.356389819268	7.2e-13	0.283226310896	0.03028362
0.11	-0.003191438702	0.374443365374	1.1e-11	0.283335847312	0.03184166
0.12	-0.003632379081	0.391797224416	8.9e-11	0.283443708609	0.03309808
0.13	-0.004150372053	0.408459677586	3.4e-12	0.283549859693	0.03480205
0.14	-0.004689339128	0.424551322589	2.5e-12	0.283654271356	0.03622014
0.15	-0.005256682085	0.440125047284	2.9e-11	0.283756917316	0.03760506
0.16	-0.005618175619	0.455286683823	5.7e-11	0.283857779176	0.03896231
0.17	-0.006487124526	0.469903756471	5.7e-12	0.283956842148	0.04029660
0.18	-0.007146262564	0.484186136060	4.7e-12	0.284054109448	0.04161036
0.19	-0.007838687257	0.498105595397	6.6e-12	0.284149593962	0.04291542
0.20	-0.008559583750	0.511685650160	4.9e-11	0.284243328335	0.04420893
0.21	-0.009315347590	0.524951884960	6.6e-11	0.284335369737	0.04549702
0.22	-0.010104271994	0.537923761008	2.9e-12	0.284425807465	0.04678551
0.23	-0.010926038044	0.550616855400	5.2e-11	0.284514773462	0.04807963
0.24	-0.011781835115	0.563047279372	1.4e-11	0.284602457095	0.04938499
0.25	-0.012671957974	0.575230413835	3.6e-12	0.284689126532	0.05070952
0.26	-0.013597339065	0.587176221467	9.2e-11	0.284775161033	0.05206063
0.27	-0.014558693626	0.598897558338	2.4e-11	0.284861102396	0.05345099
0.28	-0.015555864549	0.610403755970	2.8e-13	0.284947742217	0.05489500
0.29	-0.016590212710	0.621704492198	5.2e-12	0.285036283041	0.05642285
0.30	-0.017662158192	0.632808106481	3.0e-11	0.285128671076	0.05807352
0.31	-0.018772488310	0.643722521310	8.1e-12	0.285228402490	0.06007135

We used our algorithm to compute invariant circles for the normal form (1) for a grid of parameters near the saddle-center-Hopf line with the standard values $a = 1$, $b = c = \frac{1}{2}$. Typically we started at a fixed value of μ with $\varepsilon = 0.01$. If we are able to find the invariant circle, ε is incremented using the previous solution as an initial guess for the new parameters.

An example of the output of this algorithm with $\mu = -2.4$ is given in Table 1. For this value of μ the algorithm finds an invariant circle when $\varepsilon < 0.3129$. Beyond this value it no longer converges; moreover, when $\varepsilon > 0.326$ the computation of section 4 (with a 500^3 resolution and 10^4 iterates) also gives no bounded orbits. Consequently we believe that the circle is either destroyed or becomes unstable in this range of ε. Several of the computed invariant circles are overlaid in Figure 11. This figure also shows a chaotic orbit for $\varepsilon = 0.315$ near where the invariant circle used to be.

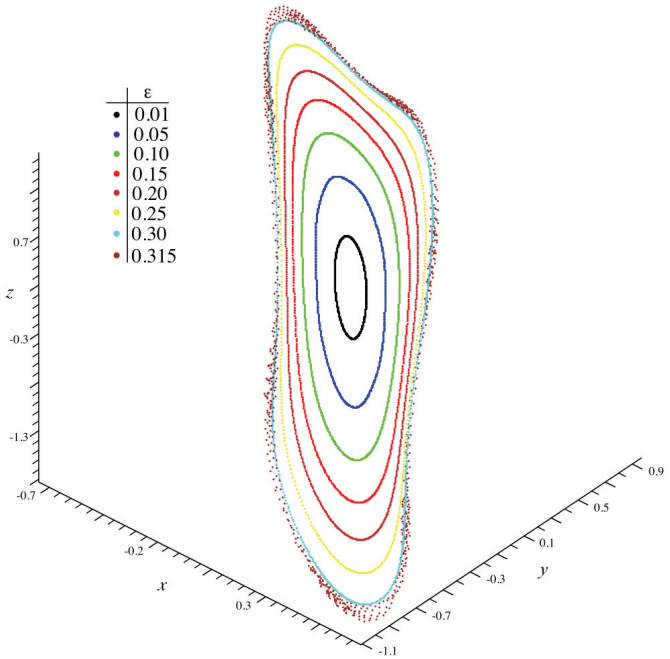


Figure 11. *Several of the invariant circles for $\mu = -2.4$ and ε values as shown. Each circle is displayed with 1000 iterates. For the outermost set of points, at $\varepsilon = 0.315$, there is apparently no invariant circle.*

The ellipsoid algorithm can fail in two ways. On the one hand, if the invariant circle has a very small stable neighborhood, the initial guess ζ_0 may lie on an unbounded trajectory even when there is an elliptic circle. This typically happens as the parameters approach a bifurcation where $\mathcal{C}$ becomes hyperbolic. To approach the stability border, we can simply take smaller steps in ε. This is apparently what happens for $\varepsilon > 0.31$ when $\mu = -2.4$. A second failure mode occurs when the orbits remain bounded, but the given error tolerance cannot be achieved; this indicates that there is no elliptic invariant circle. For example, when $\mu = -2.4$, the algorithm does not converge for $0.3018 \le \varepsilon \le 0.3044$; in this range, the circle appears to be unstable and to have undergone a doubling bifurcation; see section 7.

In our computations for other values of μ, we observe that the invariant circle often undergoes a complex sequence of bifurcations as ε changes, sometimes losing stability or simply vanishing for one ε and then regaining stability or reforming for a larger value of ε; more details are discussed in section 7 below. In every case there appears to be a maximal value of ε beyond which the circle ceases to exist or never regains stability. To make more sense of these bifurcations, we compute the longitudinal and transverse rotation numbers.

6.2. Longitudinal rotation number. If there is an invariant circle, then the restriction of the map to the circle, $f|_{\mathcal{C}}$, is a homeomorphism. Recall that the rotation number of a circle homeomorphism always exists and is independent of the initial point. This is the longitudinal rotation number of the circle, ω_L.

The longitudinal number can easily be computed under the assumption that the invariant

circle projects to the plane $x = 0$ as a Jordan curve that encloses the the origin,[7] and that the angle advance between iterates "avoids an angle" [DSM00]. The avoided angle is chosen as the location for the branch cut of the arctan function. Typically the avoided angle can be chosen as the angle diametrically opposite to the average rotation angle. Then we simply sum the polar angular increments between images, defining

$$\Theta(N) = \frac{1}{2\pi} \sum_{t=1}^{N} \operatorname{atan2}(y_t y_{t-1} + z_t z_{t-1}, y_t z_{t-1} - z_t y_{t-1}),$$

where $\operatorname{atan2}(y, x) = \arg(x + iy)$. An approximate rotation number is then simply $\Theta(N)/N + \mathcal{O}(N^{-1})$.

However, as was first suggested by Hénon (see [EV01]), a more accurate value for ω_L is easily obtained by carefully choosing N. To do this, compute the continued fraction expansion of ω_L by constructing the sequence, $\frac{p_j}{q_j}$, of its continued fraction convergents. Each convergent is defined to be a rational number closer to ω_L than any others with denominators $q < q_j$. Since each orbit on $\mathcal{C}$ must be ordered as a rigid rotation with rotation number ω_L, we can obtain the convergents by finding the closest returns to the initial point. Let q_j be the time of the next closest return to the initial point, $d_j = |\zeta_{q_j} - \zeta_0| < d_{j-1}$, and $p_j = \lfloor \Theta(q_j) + \frac{1}{2} \rfloor$ be the nearest integral number of rotations. To start the process we arbitrarily chose $d_{j_0} = 0.01$; this means that the first computed convergent will not be the leading convergent of the continued fraction for ω_L; however, it is easy to find the earlier convergents from the continued fraction expansion for p_1/q_1. For example, with $(\varepsilon, \mu) = (0.2, -2.4)$, the computed convergents are

$$q_j = \{971, 3205, 7381, 150825, 158206, 309031, 776268\},$$
$$p_j = \{276, 911, 2098, 42871, 44969, 87840, 220649\},$$

which implies that the continued fraction for ω_L is

$$\omega_L \approx \frac{220649}{776268} = [0, 3, 1, 1, 13, 3, 3, 3, 2, 20, 1, 1, 2].$$

Note that this rotation number is relatively close to the rational $\frac{2}{7} = [0, 3, 1, 1]$—see section 7. Recall that the rotation number is close to its convergents:

$$\left| \omega_L - \frac{p_j}{q_j} \right| < \frac{1}{q_j^2}.$$

Indeed, this is the error observed in practice: Figure 12 shows the difference between the value at the jth step and the final value. As discussed in [EV01], a value correct to $\mathcal{O}(q_j^{-4})$ can be obtained by adding to ω_L the average angular deviation from exact q_j-periodicity over the next q_j iterations; however, the estimate above is sufficient for our purposes.

The values in Table 1 were computed using the sequence of closest returns for $t \leq 10^6$, giving a result accurate to $\mathcal{O}(10^{-12})$. We can also use this computation to check that the orbit is properly ordered on the circle. Indeed, the sequence q_j must obey the recursion

$$q_{j+1} = a_j q_j + q_{j-1},$$

[7] If this is not the case, we could also use the *self-rotation* number [DSM00].

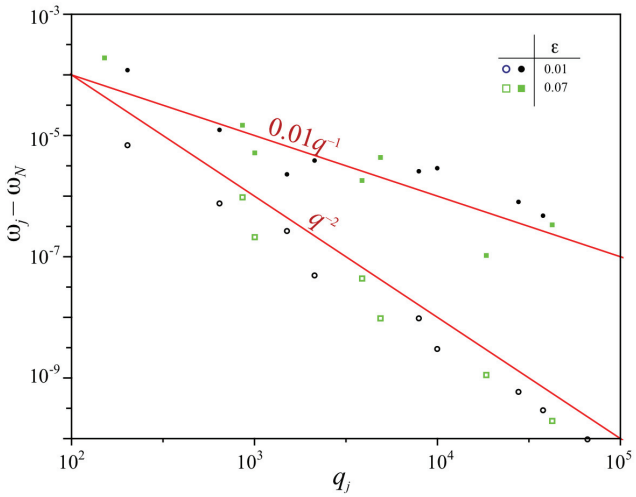


Figure 12. *Error in the computation of the rotation numbers of an invariant circle at $\mu = -2.4$, for two different ε values as shown. The open points correspond to ω_L and the solid to ω_T. The error in the longitudinal rotation number is $\mathcal{O}(q^{-2})$, but that in the transverse rotation number is only $\mathcal{O}(q^{-1})$.*

where $q_{-1} = 0$ and $q_0 = 1$, and $a_j \in \mathbb{N}$ are the continued fraction elements. Thus a necessary condition for a valid sequence of convergents is that

$$(q_{j+1} - q_{j-1}) \mod q_j = 0.$$

Usually this criterion fails only when the error bound in the circle algorithm is also not achieved.

The computations of ω_L are compared with the normal form results (24) in Figure 13. The dominant behavior is the zeroth-order rotation number $\omega_0(\mu)$ in (7). For the figure we subtracted this value from the computed results and then compared with the theoretical $\mathcal{O}(\varepsilon)$ term. The curves in the figure show the results for fixed values of ε as a function of μ. The agreement between the numerical results and the theory is nearly perfect for $\varepsilon < 0.15$ away from the resonances where the normal form is not valid. The computations show that an elliptical invariant circle does not even exist near the main resonances $\mu = -4$, -3, -2, and 0; see section 7. The numerical results indicate an additional singularity in ω_L near $\mu = -2$ that is not present in the normal form to $\mathcal{O}(\varepsilon)$. To find this singularity in the normal form we would have to keep quartic terms.

6.3. Transverse rotation number. Assume that f does have an invariant circle $\mathcal{C} : \{\zeta(\theta_L) : \theta_L \in \mathbb{S}^1\}$ that is C^1 and on which the dynamics is conjugate to rigid rotation with irrational rotation number ω_L,

$$\zeta(\theta_L + \omega_L) = f(\zeta(\theta_L)).$$

The linearization of f then gives rise to a quasi-periodic skew-product on $\mathbb{R}^3 \times \mathbb{S}^1$,

$$\begin{aligned} \xi' &= A(\theta_L)\xi, \\ \theta_L' &= \theta_L + \omega_L \mod 1, \end{aligned} \tag{25}$$

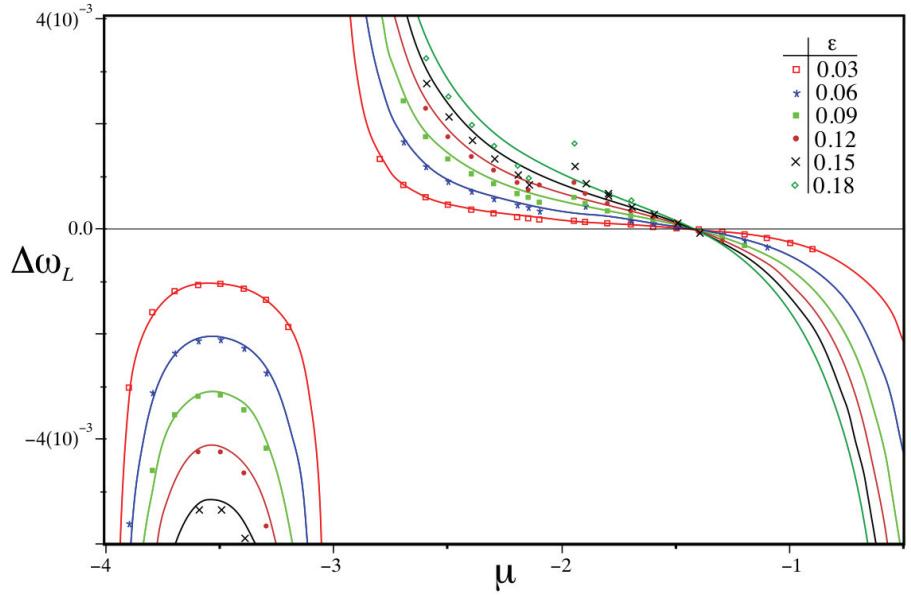


Figure 13. *Comparison of the computed* $\Delta\omega_L = \omega_L - \omega_0(\mu)$ *(dots) with the normal form* (24) *(curves) as a function of* μ *for* $(a, b, c) = (1, \frac{1}{2}, \frac{1}{2})$ *and the values of* ε *indicated.*

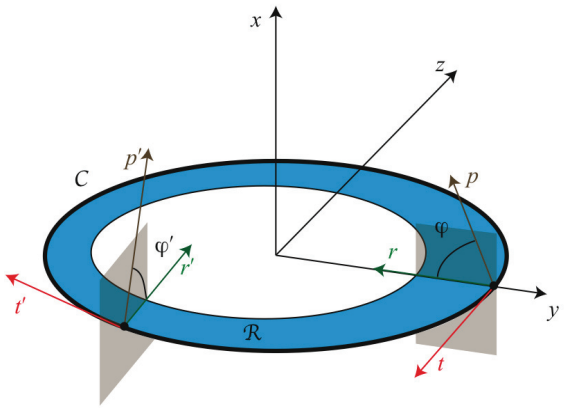


Figure 14. *Computing the transverse rotation number for an invariant circle* $\mathcal{C}$ *relative to a ribbon* $\mathcal{R}$.

where $A(\theta_L) = Df(\zeta(\theta_L))$ is a periodic matrix.

The transverse rotation number of $\mathcal{C}$ is the average rotation rate of a transverse vector ξ "around" $\mathcal{C}$, if that average exists. Indeed, Herman proved that the "fiberwise" rotation number for a quasi-periodically forced circle map always exists [Her83]. The map (25) reduces to this case if the ξ dynamics are projected onto a family of circles transverse to $\mathcal{C}$. To do this, define the transverse angle φ relative to a ribbon $\mathcal{R}$ attached to the invariant circle; see Figure 14. The transverse dynamics then induce a circle map $\varphi \mapsto g(\varphi, \theta_L)$. Consequently, under the assumption that $\mathcal{C}$ exists and its dynamics are conjugate to a rigid rotation, the

rotation number of g exists and is independent of the initial θ and φ.

Let $\hat{t}(\theta_L)$ be the unit tangent vector to $\mathcal{C}$ at $\zeta(\theta_L)$; it can be approximated using the closest approaches from the computation of ω_L. Since the invariant circle of (1) appears to be almost always everywhere transverse to the x-axis,[8] define the ribbon by the vector $r(\theta_L) = \hat{t}(\theta_L) \times \hat{e}_1$. Beginning with an arbitrary vector v_0 attached to the point $\zeta(0) \in \mathcal{C}$, we iterate to obtain the sequence $v_{j+1} = A(j\omega_L)v_j$. These vectors are then projected onto a plane orthogonal to the local tangent vector $\hat{t}(j\omega_L)$; for convenience, we also rescale, defining

$$p = |r| \left(v - (\hat{t} \cdot v)\hat{t}\right).$$

The projected vector is effectively two-dimensional with components $p = (p^{\parallel}, p^{\perp})$ defined relative to the ribbon direction

$$\begin{aligned} p^{\parallel} &= p \cdot \hat{r} = \hat{e}_1 \cdot \hat{t} \times v, \\ p^{\perp} &= \hat{t} \cdot p \times \hat{r} = \hat{e}_1 \cdot (v - (\hat{t} \cdot v)\hat{t}). \end{aligned}$$

Thus the angle of p is $\varphi = \arctan(p^{\perp}/p^{\parallel})$. As before, we compute the change in rotation angle at the jth iterate with the two-argument arc-tangent:

$$\Delta\varphi_j = \operatorname{atan2}(p^{\parallel}_{j-1}p^{\parallel}_j + p^{\perp}_{j-1}p^{\perp}_j, p^{\parallel}_{j-1}p^{\perp}_j - p^{\perp}_{j-1}p^{\parallel}_j).$$

The transverse rotation number is the average change along the orbit

$$\omega_T \approx \frac{1}{2\pi q} \sum_{j=1}^{q} \Delta\varphi_t.$$

It is not clear how to optimize the error in this computation as we did for the longitudinal rotation number. We simply use the time of closest approach that we computed for the longitudinal rotation number; however, as can be seen in Figure 12, the accuracy for this computation is only $\mathcal{O}(q^{-1})$.

We compare the computations of ω_T to the normal form results (24) in Figure 15. Again, the agreement between the two results is extremely good for small ε and away from the resonant values. The $\mathcal{O}(\varepsilon^{1/2})$ normal form result is finite at the resonances at $\mu = -4$, -3, and -2, but, as before, the numerical computations indicate that the actual rotation number diverges there. Such divergences are found in the normal form at $\mathcal{O}(\varepsilon^{3/2})$; however, since these corrections to the rotation number are normally very small, we do not show them in the figure.

6.4. Frequency map. If μ and (a, b, c) are held fixed, the rotation numbers (ω_L, ω_T) of the circle vary along a curve as ε changes; an example is shown in Figure 16 for $\mu = -2.4$. For small ε this curve lies close to the parabola (24) defined by the normal form results. Though we expect the curve to be defined only for a Cantor set of parameter values, it appears to be

[8] An exception is shown in Figure 17 for $(\varepsilon, \mu) = (0.068, -1.9)$, where the invariant circle develops "curlicues." Here the algorithm computes ω_T incorrectly because the projection of the circle on the $x = 0$ plane is not one-to-one.

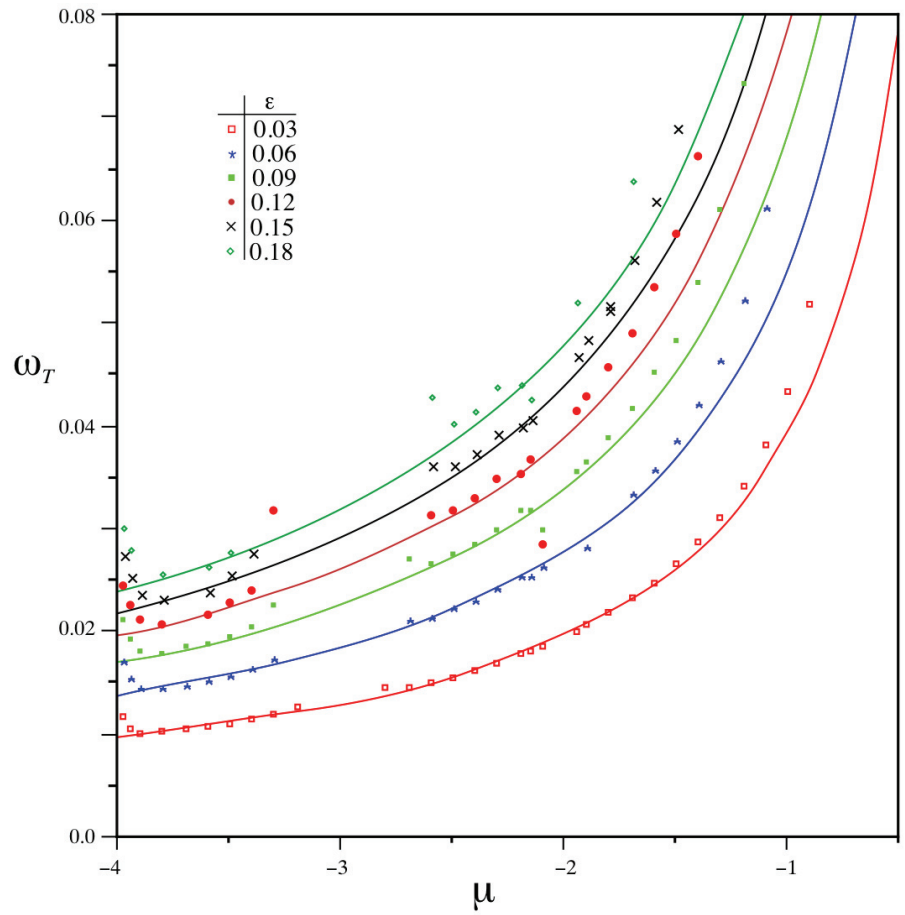


Figure 15. *Comparison of the the measured transverse rotation number (dots) with the theory (24) as a function of μ for $(a, b, c) = (1, \frac{1}{2}, \frac{1}{2})$ and the values of ε indicated.*

continuous for most values of ε; there are, however, several small intervals visible in which the ellipsoid algorithm does not converge. The algorithm can find an invariant circle up to $\varepsilon = 0.312$, at which point the circle is apparently destroyed by a resonant bifurcation; see section 7. The invariant circles shown in Figure 16 are also discussed in more detail in section 7.

For some values of μ, the frequency map exhibits singularities as ε varies; two examples are shown in Figure 17. Note that the normal form curve (brown) still fits the results for small ε. The singularities are in ω_L; those shown occur near $(\varepsilon, \mu) = (0.1, -1.5)$ and $(\varepsilon, \mu) = (0.07, -1.9)$. Near the singularities, the original circle (blue) becomes highly distorted, and as ε increases it can no longer be found. The singularities appear to be associated with a pair of circle-saddle-center bifurcations: the old circle is destroyed in a saddle-center bifurcation below the resonance line, while a new elliptic invariant circle is born above (red); see section 7.3. Remarkably, the new circle persists to a much larger value of ε, and its rotation numbers approximately follow those of the normal form. In each case the old circle eventually appears

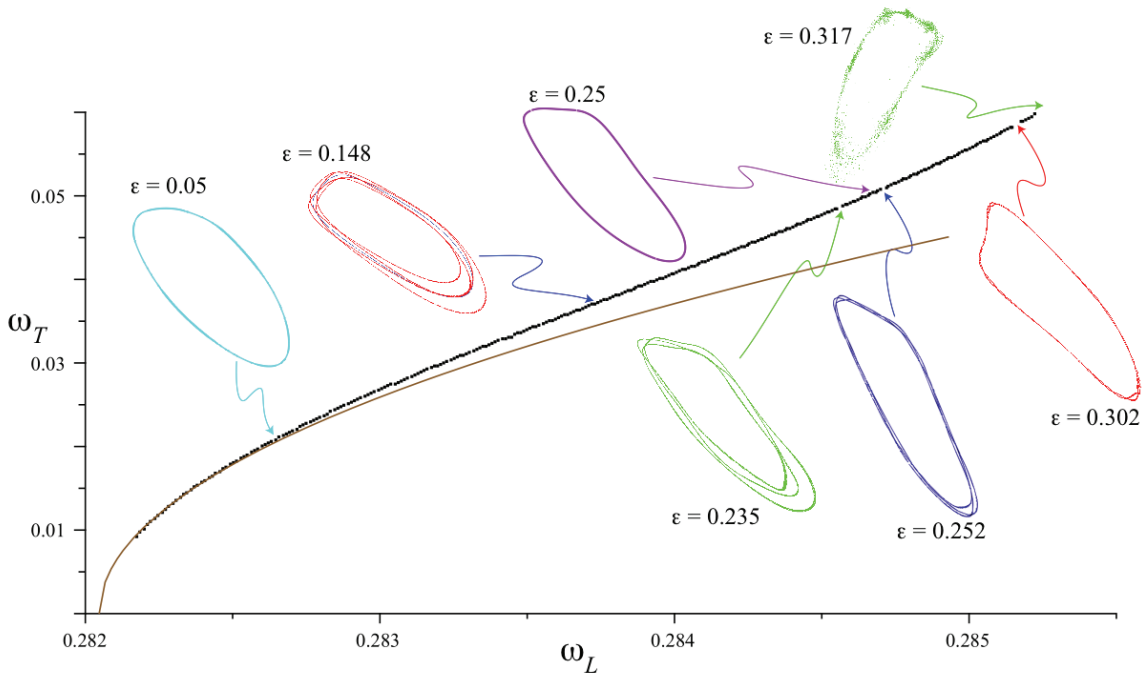


Figure 16. *Numerically computed frequency map for the invariant circle with* $\mu = -2.4$ *and the standard values* $(a, b, c) = (1, \frac{1}{2}, \frac{1}{2})$ *as* ε *varies from* 0.01 *to* 0.312 *(black dots). The curve (brown) represents the normal form of* (24)*. Insets show several phase portraits of the circle and/or nearby orbits. The illustrated resonances in the order of increasing* ϵ *are* $(7, 1, 2)$, $(3, 4, 1)$, $(3, 3, 1)$, $(10, 3, 3)$, $(46, -2, 13)$, $(7, 0, 2)$.

to lose smoothness and is apparently destroyed.

The image of the frequency map for $\mu = -3.9$ is shown in Figure 18. For this μ the longitudinal rotation number decreases with ε as predicted by (24). The computed curve is fairly smooth, but there are bifurcations along the way, e.g., for $\varepsilon = 0.0767$, but the resolution is too rough to resolve them. The gap between $0.0173 < \varepsilon < 0.0177$ is caused by a period doubling, and the doubled circle is shown in green.

A summary of our computations is given by the two-parameter frequency map

$$\Omega : (\varepsilon, \mu) \mapsto (\omega_L, \omega_T),$$

shown in Figure 19. Here we computed the invariant circle for $\mu \in [-4, 0]$ in steps of 0.1 and for $\varepsilon \geq 0.01$. We were able to find invariant circles for some range of ε for each μ except for the values $\{0, -2, -3, -4\}$ where there are strong resonances. When $-0.6 < \mu < 0$, an invariant circle is stable only for $\varepsilon < 0.01$, so these points do not appear in the figure.

There are many cases in which the algorithm fails to find a circle at one value of ε but then succeeds for a slightly larger value. These gaps are typically very small but can be seen in several of the fixed μ curves; in most cases they correspond to resonant bifurcations that cause the circle either to lose transverse stability or to be destroyed. It is to the study of these bifurcations that we turn next.

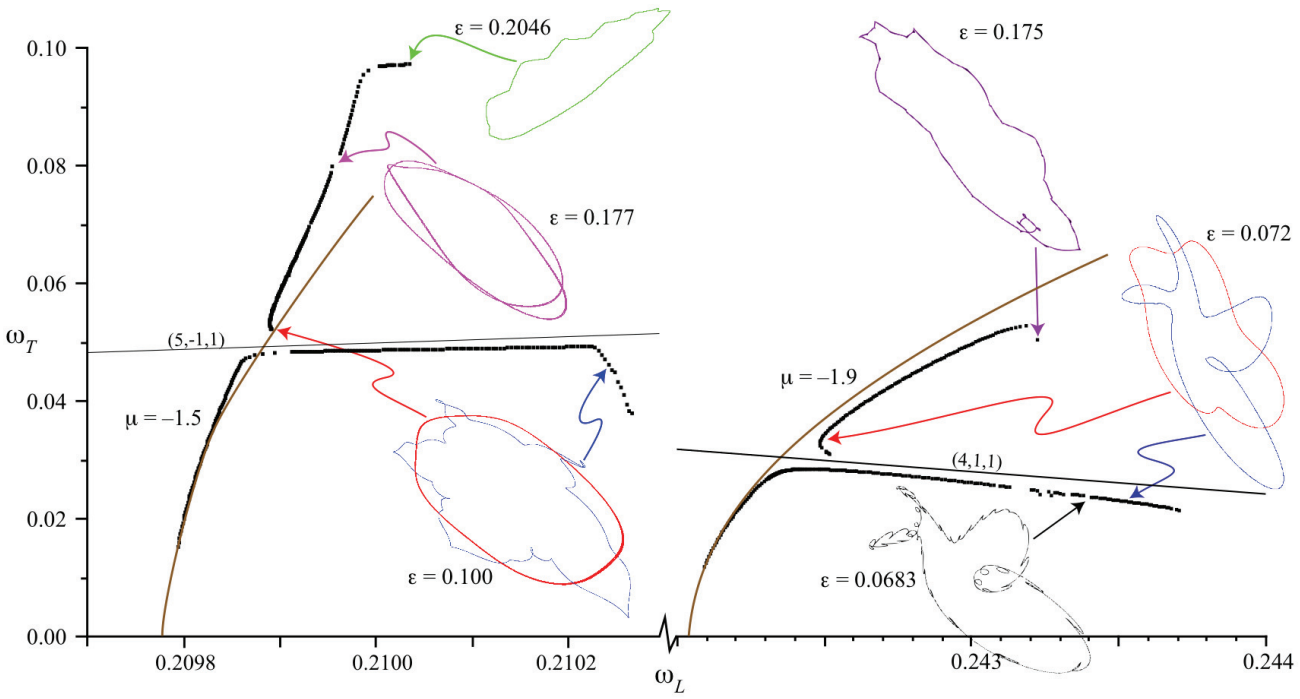


Figure 17. *Numerically computed frequency map for $\mu = -1.5$ and -1.9 (dots) and the corresponding normal form results (brown curves). Also shown are the computed invariant circles for several values of ε. The singularities in the frequency maps are due to $(5,-1,1)$ and $(4,1,1)$ resonances that result in pairs of circle-saddle-center bifurcations; see section 7.3.*

7. Resonant bifurcations of invariant circles. The rotation vector $\omega = (\omega_L, \omega_T)$ is resonant if there are integers (m_1, m_2, n), not all zero, such that

$$m \cdot \omega = n. \tag{26}$$

If ω does not satisfy such a relation, then it is "nonresonant."

Since (26) is homogeneous, if (m, n) is a solution, then so is (lm, ln) for $l \in \mathbb{Z}$; consequently, the set of integers that satisfy (26) forms a sublattice of $\mathbb{Z}^3$,

$$\mathcal{L}(\omega) = \{(m, n) \in \mathbb{Z}^3 : m \cdot \omega = n\}, \tag{27}$$

called the *resonance module*. The dimension of $\mathcal{L}$ is the number of independent resonance conditions. Using homogeneity, we will assume that n is nonnegative and $\gcd(m_1, m_2, n) = 1$.

In ω-space, (26) defines a line for each (m, n); a few such lines are shown in Figure 19. Using the normal form results (24), the resonance curves for the invariant circle can also be obtained in parameter space. The fine structure in the bounded volume Figure 3 is caused by these resonances.

For example, resonances with $|m_2| = 2$ are often observed to result in the destruction or loss of stability of the invariant circle $\mathcal{C}$, as well as a dramatic drop in the volume of bounded orbits. Figure 19 shows that $\mathcal{C}$ appears to lose stability at the $(1, -2, 0)$ resonance

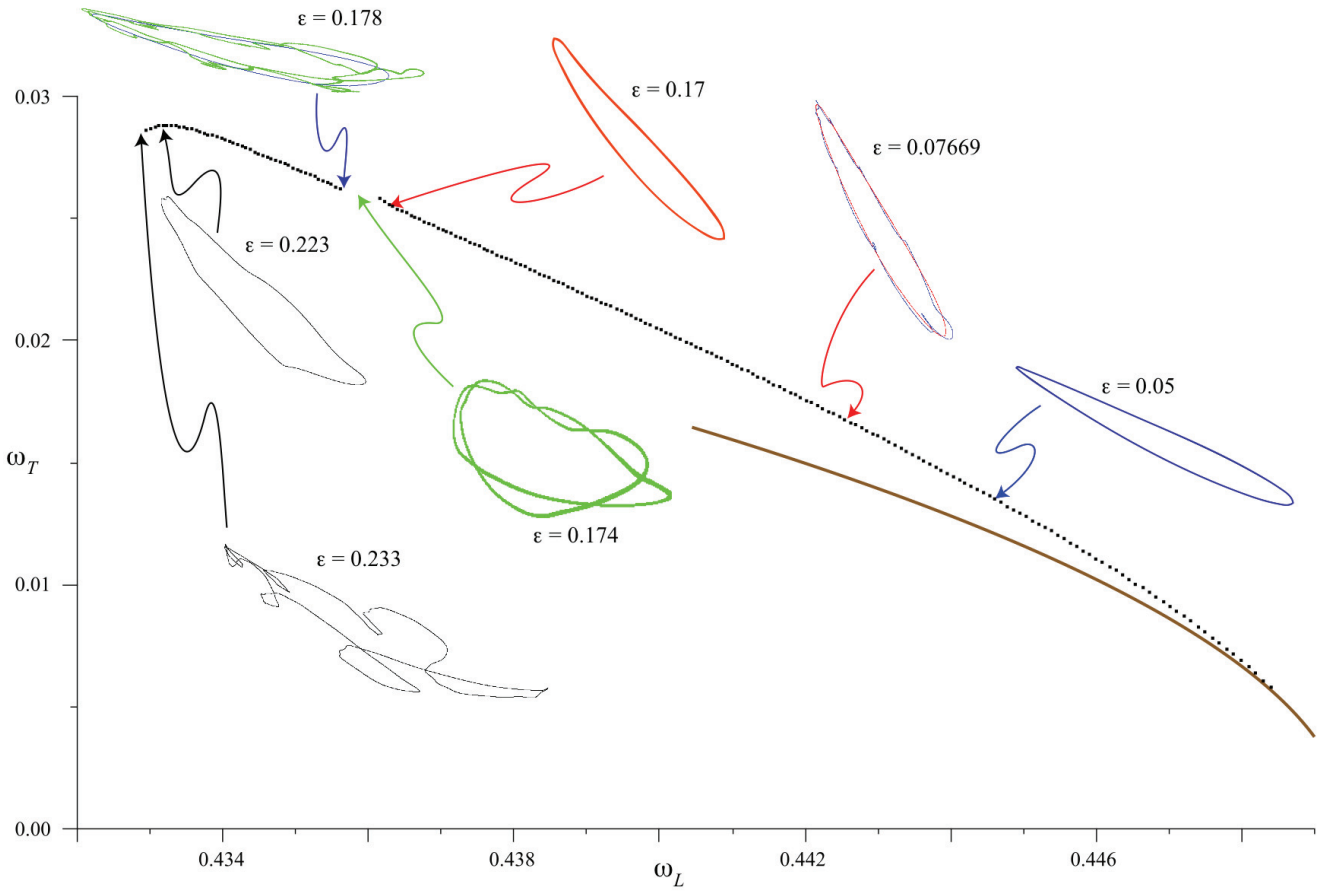


Figure 18. *Frequency map for $\mu = -3.9$ as ε ranges from 0.01 to 0.233. Insets show invariant circles for seven values of ε. The doubled circle in the gap near $\varepsilon = 0.174$ is caused by a $(7, -2, 3)$ resonance.*

for $-1.4 < \mu < 0$ and at the $(3, \pm 2, 1)$ resonances for $-3.3 < \mu < -2.5$. In each case, a new elliptic doubled circle is created when the original circle crosses the resonance; phase space portraits are shown in the top and bottom panels of Figure 20. Since $\omega_T/\omega_L = \frac{1}{2}$ for the $(1, -2, 0)$ resonance, tori near $\mathcal{C}$ just below the bifurcation approximate a Möbius strip with a half-twist, and the new circle undergoes one-half turn about the original circle for each longitudinal period. For the second case, the $(3, \pm 2, 1)$ resonance, the new circle rotates by three turns transversally in two longitudinal rotations. In both cases the new elliptic invariant circle is a double covering of the original one, as is familiar for period doubling of periodic orbits of flows.

Interestingly, there is also another type of period doubling bifurcation that cannot occur in flows. This happens, for example, at the $(4, 2, 1)$ resonance; a phase space portrait is shown in the middle panels of Figure 20. Instead of one invariant circle that is a double cover of the original, here there are two disjoint invariant circles, each covering the original circle once. Though the new invariant circles are geometrically disjoint, they are mapped onto one another by the dynamics. The reason for the different behavior is that $\gcd(m_1, m_2) \neq 1$ in this case.

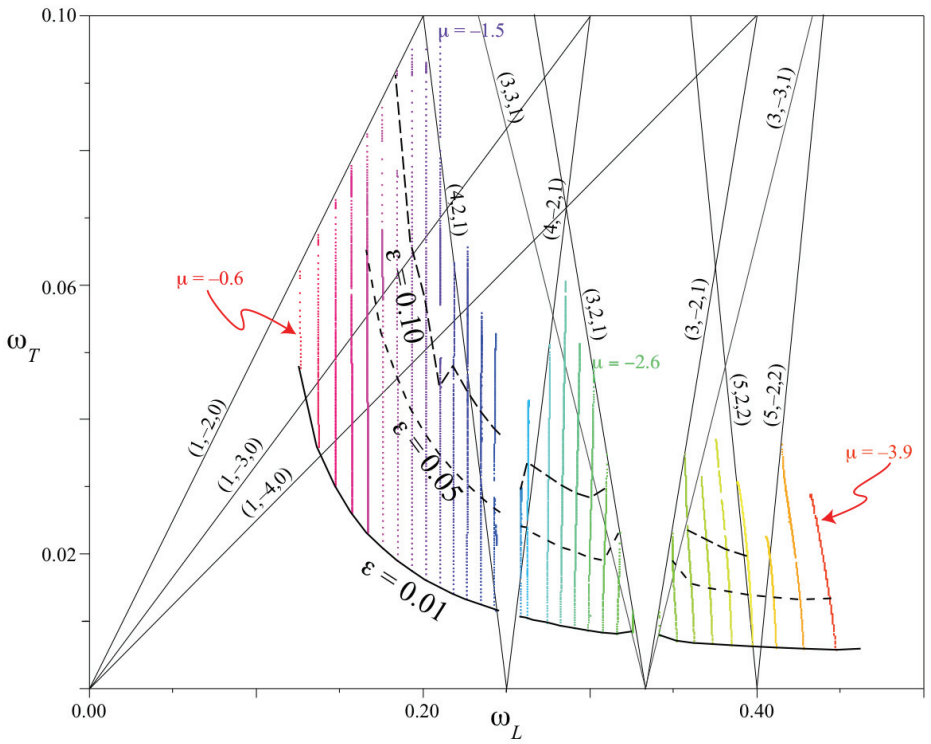


Figure 19. *Frequency map* (ω_L, ω_T) *for* $\mu \in [-3.9, -0.6]$ *with step* 0.1 *and a grid of* ε *values from* 0.01 *up to the largest* ε *for which the ellipsoid method converged to an invariant circle. Also shown are several of the important resonance lines defined by* (26).

This is the simplest example of a general scheme that is easily understood by studying an integrable map of the two-torus with resonant rotation vector ω.

Lemma 2. *The orbits of the torus map*

$$x' = x + \omega \mod 1, \tag{28}$$

for $x, \omega \in \mathbb{T}^2$, *are either*

1. *dense on* $\mathbb{T}^2$ *if* ω *is nonresonant* ($\dim \mathcal{L}(\omega) = 0$);
2. *dense on* k *circles if* ω *satisfies one resonant relation and* $\gcd(m_1, m_2) = k$ ($\dim \mathcal{L}(\omega) = 1$); *or*
3. *periodic if* ω *satisfies more than one resonance relation* ($\dim \mathcal{L}(\omega) = 2$).

Proof. If ω is nonresonant, then the result is standard [CFS82]. Suppose now that there is a single resonance relation ($\dim \mathcal{L} = 1$), let $k = \gcd(m_1, m_2)$ be the common divisor, and define

$$(m_1, m_2) = k(p, q), \tag{29}$$

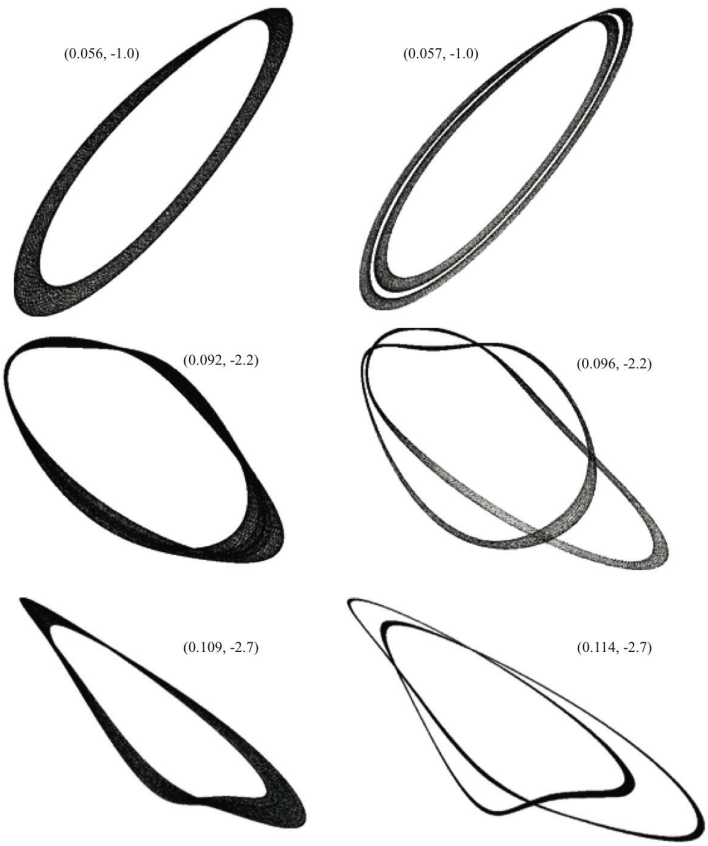


Figure 20. *Circle doubling bifurcations. Shown are some of the orbits in the neighborhood of an elliptical invariant circle (left panels) just before the doubling bifurcations, and orbits in the neighborhood of the doubled circle (right panels) just after. The corresponding resonances are* $(1,-2,0)$*,* $(4,2,1)$*, and* $(3,-2,1)$ *from top to bottom. The values of* (ε,μ) *are shown.*

where p, q are coprime.[9] As is shown in [HW79], there is a solution $(\hat{p},\hat{q}) \in \mathbb{Z}^2$ to the equation

$$p\hat{q} - q\hat{p} = 1$$

if and only if p, q are coprime. Taking any such solution, define the unimodular transformation

$$y = \begin{pmatrix} p & q \\ \hat{p} & \hat{q} \end{pmatrix} x.$$

The variables y are simply a new set of coordinates on $\mathbb{T}^2$. The map in these new coordinates takes the form

$$\begin{aligned} y_1' &= y_1 + \frac{n}{k}, \\ y_2' &= y_2 + \hat{p}\omega_1 + \hat{q}\omega_2. \end{aligned} \tag{30}$$

[9]Note that if one frequency is rational, then $m_1 m_2 = 0$ and hence $pq = 0$. This case can be formally included here with $k = m_1 + m_2$ and p or q equal to 1.

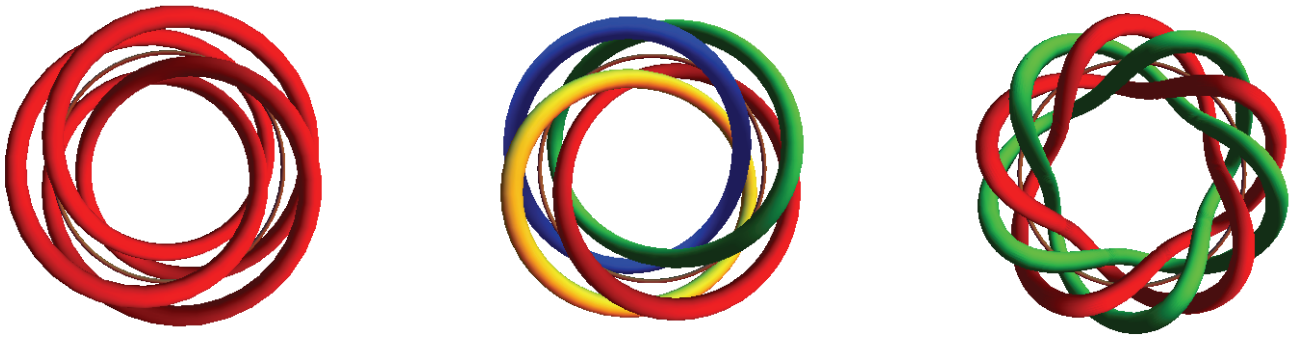


Figure 21. *Schematic pictures of torus links of type* $(m_1, m_2) = (3, 4)$, $(4, 4)$, *and* $(10, 4)$. *Different colors distinguish the* $k = \gcd(m_1, m_2)$ *disjoint circles of the link. The central invariant circle (not part of the torus link) is also shown.*

By assumption, k and n are coprime integers, so the orbit of y_1 is periodic with period k (in particular, if $n = 0$, then $k = 1$). Moreover, since ω satisfies exactly one resonance relation, the combination $(\hat{p}, \hat{q}) \cdot \omega$ is irrational. Thus the orbit of y_2 is dense in $\mathbb{T}^1$. Hence the combined orbit on $\mathbb{T}^2$ is dense on k circles. This still holds true in the original coordinate system, but the circles will wrap around the torus in both directions.

Finally, if ω satisfies two independent resonance relations ($\dim \mathcal{L} = 2$), $m \cdot \omega = n$ and $\hat{m} \cdot \omega = \hat{n}$, then the components of ω must each be rational because then

$$\begin{pmatrix} m^T \\ \hat{m}^T \end{pmatrix} \omega = \begin{pmatrix} n \\ \hat{n} \end{pmatrix},$$

and the matrix on the left is nonsingular since the vectors m and $\hat{m}$ are not parallel. Thus both frequencies are rational, and all orbits are periodic. ■

Structures similar to the second case of Lemma 2 are also found in quasi-periodically forced circle maps [JK06, JS06] where they have been called (k, q)-invariant graphs.

We now embed these dynamics into three dimensions as invariant sets of a family of "integrable" volume-preserving maps f_0 on the solid torus, $\mathbb{T}^2 \times \mathbb{R}^+$,

$$f_0(\theta, r) = (\theta + \omega(r, \mu), r). \tag{31}$$

Here r represents the radius, so that $r = 0$ corresponds to the toroidal or longitudinal axis, θ_L is the toroidal angle, and θ_T is the poloidal or meridional angle. We assume there is a parameter μ to unfold the frequency map (just as for our model (1)). The phase space of (31) is foliated into invariant tori $r = r_0$, and on each such torus f_0 becomes (28).

Now suppose that at (r_0, μ_0) the rotation vector of (31) satisfies a single resonance condition, that is, $\omega(r_0, \mu_0) \cdot m = n$, where m satisfies (29). Each orbit on this resonant torus densely covers a circle ($k = 1$) or set of circles ($k > 1$) on this torus. These curve(s) correspond to a torus knot (torus link) [Ada04]; several examples are shown in Figure 21.[10] A torus knot

[10] Famous knots and links include the trefoil knot $(3, 2)$, the Hopf link $(2, 2)$, and Solomon's knot (which is a link) $(4, 2)$. Note that Borromean rings are not included in the class of torus links.

of type (p, q) coprime corresponds to the closed loop $\{(qt, -pt, r_0) : t \in \mathbb{R}\}$ that has p poloidal wraps for each q toroidal circuits. A torus link of type (m_1, m_2) is the collection of k loops given by $\{(qt, -pt + i/m_1, r_0) : i = 0, \ldots, k-1,\ t \in \mathbb{R}\}$.

For example, the $(3, 4, 1)$ resonance leads to a single $(3, 4)$ torus knot, and the $(4, 4, 1)$ resonance leads to four $(1, 1)$ torus knots (each of which is not knotted), which together give the $(4, 4)$ torus link. Similarly, the $(10, 4, 3)$ resonance leads to two $(5, 2)$ torus knots, together forming a torus link of type $(10, 4)$.

Note that the number n does not enter into the geometric specification of the knot or link. However, by (30), n mod k does determine the dynamics on the torus link.

Here we are specifically interested in local bifurcations at the elliptic invariant circle $\mathcal{C} = \{r = 0\}$ of (31). Some of these are analogous to the generic bifurcations of an elliptic fixed point of an area-preserving map in which periodic orbits are created or destroyed when the multiplier of the linearization passes through a root of unity. For the invariant circle, this would correspond to the transverse rotation number ω_T passing through a rational, n/m_2: an $(0, m_2, n)$ resonance. More generally, suppose that $\mathcal{C}$ is resonant, i.e., that $\omega(0, 0)$ satisfies a single resonance condition (26), and that $\partial_\mu \omega(0, 0) \neq 0$. Then as μ passes through zero, a resonant torus will be created in the neighborhood of $\mathcal{C}$. As above, this resonant torus is foliated into a one-parameter family of torus knots or links. If we now perturb (31), $f_\varepsilon = f_0 + O(\varepsilon)$, then by analogy with the Poincaré–Birkhoff theorem for area-preserving maps we expect that all but a finite number of these circles will be destroyed. Indeed, a generalization of this theorem to the volume-preserving map case has been obtained by Cheng and Sun [CS90b]. Thus the bifurcation should create a finite number of invariant circles in the neighborhood of $\mathcal{C}$.

Though the torus map (28) does not distinguish between its two angle variables, the embedding of (31) into $\mathbb{R}^3$ assigns different roles to the longitudinal (θ_L) and transverse (θ_T) angles. The two-dimensional analogy then implies that for the volume-preserving case, m_2 may play a more important role than m_1. In particular, recall that an elliptic fixed point of an area-preserving map is "strongly" resonant if the denominator of ω_T is small, i.e., $m_2 \leq 4$. An analogous classification will pertain to the volume-preserving case.

From these qualitative considerations and our numerical observations, we propose the following conjecture.

Conjecture 1. *An elliptic invariant circle of a volume-preserving map with frequencies $\omega = (\omega_L, \omega_T)$ that satisfy the single resonance condition $\omega_L m_1 + \omega_T m_2 = n$ ($\gcd(m_1, m_2, n) = 1$, $\gcd(m_1, m_2) = k$) generically undergoes one of the following bifurcations:*

- *$m_2 = 0$ (string of pearls bifurcation): The circle is destroyed, and a pair of saddle period-m_1 orbits are born in an SCNS bifurcation; recall section 5.2. The one-dimensional invariant manifolds of the saddles nearly coincide along the location of the destroyed circle. If the bifurcation is supercritical, it also creates a family of m_1 almost-invariant balls (pearls) bounded by the manifolds of neighboring points on the orbits and containing elliptic circles.*
- *$|m_2| = 1$ (saddle-center bifurcation): The transverse multiplier of the invariant circle becomes 1, and the circle is destroyed in a saddle-center bifurcation as the parameters cross the resonance line.*
- *$|m_2| \geq 2$ (torus-link bifurcation): The invariant circle persists but may lose stability.*

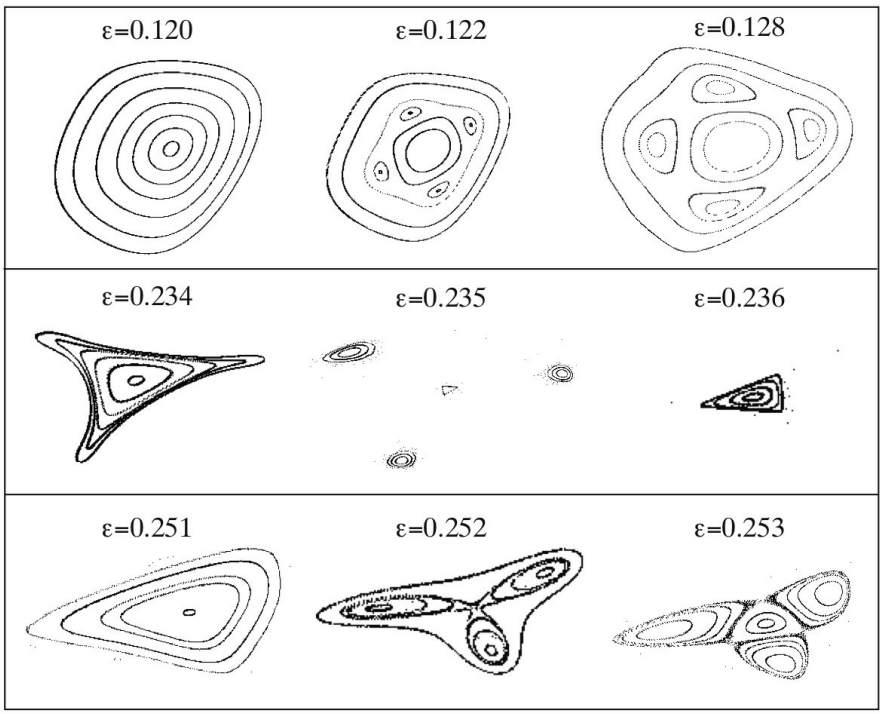


Figure 22. *Poincaré slices Σ_Δ for $\mu = -2.4$ and the ε values shown. The upper sequence crosses the $(4, -4, 1)$ resonance, the middle the $(3, 3, 1)$, and the lower the $(10, 3, 3)$ resonances.*

In its neighborhood, k invariant circles are born that form an (m_1, m_2) torus link.

We believe that this classification of different types of m_2-tupling bifurcations of invariant circles is new. In the following subsections, we will separately treat each of the three cases of the conjecture.

7.1. Torus-link bifurcations, $|m_2| \geq 2$. In the beginning of this section, we discussed several examples of period doubling with $|m_2| = 2$, showing that a torus knot (link) is created if m_1 is odd (even); recall Figure 20. We will now present more examples with larger $|m_2|$, and then give an explanation of the observed bifurcations.

Several of such bifurcations for $\mu = -2.4$ are illustrated using Poincaré *slices* in Figure 22. A slice is analogous to a Poincaré section of a three-dimensional flow; however, since an orbit of a three-dimensional map would almost never intersect a two-dimensional section, we must instead consider a slice of some nonzero thickness. In the figure, we use the same slice, $\Sigma_\Delta = \{(x, y, z) : y > 0, |z| < \Delta\}$, that was used to find the elliptic circles in section 6.1.

For example, at $(\varepsilon, \mu) \approx (0.121, -2.4)$, the elliptic circle crosses the $(4, -4, 1)$ resonance. Since (m_1, m_2) have a common factor, this bifurcation creates an elliptic torus link of type $(4, -4)$ which consists of four torus knots of type $(1, -1)$; recall the schematic Figure 21. The new circles move away from the central circle as ε grows. In the Poincaré slice—the top row of Figure 22—this bifurcation looks just like a generic quadrupling bifurcation of an

elliptic fixed point in an area-preserving map. The slice reveals that both an elliptic and a hyperbolic $(4,-4)$ torus link are created, and that the central circle appears to persist and remain stable through the bifurcation. We observe qualitatively similar Poincaré slices for the $(3,4,1)$ resonance (at $(\varepsilon,\mu)\approx(0.147,-2.4)$; recall Figure 16), and $(10,4,3)$ (at $(0.169,-2.4)$); however, the corresponding torus-links are quite different; recall Figure 21.

It is easy to observe many such resonances; a list of the lower-order resonances encountered for $\mu=-2.4$ is given in Table 3 in Appendix E. It is interesting to note that many resonances appear in sequences: $(3,4,1)$ and $(10,4,3)$ are the two first elements of the sequence $(3+7n,4,1+2n)$. These resonance lines have one conjugate point, $(\omega_L,\omega_T)=(\frac{2}{7},\frac{1}{28})$, in common. Since this limit point is close to the image of the frequency map for $\mu=-2.4$, many of these resonances are encountered in the family. However, since the slope of these resonance lines is proportional to n and the last observable, stable invariant circle has $\omega_L\approx 0.2853=[0,3,1,1,44,\ldots]$, which is close to the rational $\frac{2}{7}$, the frequency curve will eventually miss the resonance lines. Any low-order doubly rational point in the image of the frequency map can similarly serve as a limit point for many resonance sequences. In the present case we observe, for example, the sequences $(7n,1,2n)$, $(7n,2,2n)$, $(3+7n,3,1+2n)$, and $(4+7n,-3,1+2n)$. As a consequence, the invariant circle will repeatedly undergo bifurcations that have the same m_2 but that correspond to different associated torus links.

Resonances with larger m_2 have similar structure: for example, a $(1,-5,0)$ resonance at $(\varepsilon,\mu)\approx(0.026,-1.1)$ creates a new elliptic and a new hyperbolic circle that, as ε grows, give rise to a five-island structure when viewed in a Poincaré slice.

Tripling bifurcations ($|m_2|=3$) may result in a brief loss of stability of the invariant circle. For example, there is a gap in the $\mu=-2.4$ frequency map of Figure 16 near $\varepsilon=0.235$ where the circle crosses the $(3,3,1)$ resonance. This bifurcation again resembles the two-dimensional case. Indeed, the new hyperbolic and elliptic $(3,3)$ torus links are created in a circle-saddle-center bifurcation before the "tripling," near $\varepsilon=0.234$. Moreover, the central circle appears to lose stability in a collision with the hyperbolic period-three family near $\varepsilon=0.235$; cross sections are shown in Figure 22. This and the similar $(3,-3,1)$ resonance occur for a range of μ, as can be seen in Figure 19. Another tripling for $\mu=-2.4$ occurs in a gap around $\varepsilon=0.252$; this is a $(10,3,3)$ resonance. For this case (m_1,m_2) are coprime and the bifurcation results in the formation of a single elliptic torus-knot with three longitudinal wraps and ten meridional wraps (the dark blue circle in Figure 16). Similar tripling bifurcations occur whenever $\mu>-1.6$ as the invariant circle crosses the $(1,-3,0)$ resonance. However, for these cases the tripled circle is a $(1,-3)$ torus knot.

From the examples discussed so far it seems as if the bifurcations of invariant circles are simply those of fixed points of area-preserving maps multiplied by a (twisted) circle. However, this appears to be false. We found many portraits in the Poincaré slice that are not generic in area-preserving maps. To discuss this, we first recall the generic bifurcations of a fixed point of an area-preserving map. When the multiplier of the fixed point is $\exp(\pm 2\pi i n/m)$, n,m coprime, then there are five cases (assuming that the generic higher-order transversality conditions are satisfied) [MH92]:

- $m=1$ (saddle-center): collision of saddle and center fixed points (double multiplier 1, not semisimple).
- $m=2$ (doubling): a periodic orbit of period two is created, and the fixed point changes

stability (double multiplier -1, not semisimple).

- $m = 3$ (tripling): a saddle-center bifurcation creates a pair of period-three orbits away from the fixed point, and the saddle passes through the fixed point, causing a momentary loss of stability.
- $m = 4$ (quadrupling): this bifurcation can behave either like $m = 3$ (the hyperbolic orbit colliding with the fixed point with loss of stability) or like $m > 4$ (elliptic and hyperbolic pair being created without loss of stability), depending on the relative strengths of higher-order terms.
- $m > 4$ (m-tupling): a pair of elliptic and hyperbolic periodic orbits of period m is created at the origin without loss of stability. These orbits form an island chain with m islands.

As a result, resonances are usually grouped into "low-order" or "weak" ($m \leq 4$) and "high-order" or "strong" ($m \geq 4$) cases. High-order resonances do not cause loss of stability of the fixed point at the bifurcation.

For $m \geq 3$ these resonances are described by an area-preserving map of the form

$$z' = e^{2\pi i n/m} z + O(2).$$

The resonant terms correspond to the monomials $z^{k_1}\bar{z}^{k_2}$ with $k_1, k_2 \in \mathbb{N}$ and $|k_1 - k_2|$ divisible by m. Upon setting $z = \sqrt{2I}\exp i\phi$, for angle-action variables (ϕ, I), these bifurcations can be described by an approximate Hamiltonian

$$H(\phi, I) = \mu\alpha I + \beta I^2 + \cdots + (\gamma + \delta I + \cdots) I^{m/2} \cos m\phi, \tag{32}$$

where the bifurcation parameter μ detunes the resonance. When $m = 3$ the resonant term γ is the leading-order nonlinear term, and thus the behavior is different than for $m > 4$. When $m = 4$ the leading-order resonant and nonresonant terms have the same order, and the behavior depends on whether $|\beta| \gtrless |\gamma|$. For $m \geq 5$ the resonant term is not of leading order and the usual island chain is born.

Resonant bifurcations of invariant circles of volume-preserving maps follow a similar scheme, but there are more possibilities. This is for two reasons. First, instead of the single period m there is now a vector (m_1, m_2). Roughly speaking, an area-preserving map describes a slice of the dynamics transverse to the invariant circle with "period" m_2. In the simplest situation, these two-dimensional dynamics rotate as the slice moves along the invariant circle; this is described by m_1. Second and less obvious, the distinction between low-order and high-order resonances is not so clear-cut: indeed all the "low-order" resonances can also appear without loss of stability. Our observations can be summarized in three cases:

- $m_2 = 2$: the central circle loses stability, and an elliptic doubled circle appears. Depending on $k = \gcd(m_1, m_2)$ either there is one circle that is a double cover of the original one ($k = 1$), or there are two circles each covering the original circle once that are mapped into each other ($k = 2$). These two circles form an (m_1, m_2) torus link. However, unlike generic bifurcations in Hamiltonian systems this bifurcation may be "weak," creating a pair of elliptic and hyperbolic circles. This happens, e.g., for the $(7, 2, 2)$ resonance at $(0.0167, -2.4)$.

- $m_2 = 3$: the tripling may lead to a momentary loss of stability of the central circle. When m_1 is divisible by 3, then three circles are created, and $n \bmod 3$ determines how they are mapped to each other. Unlike the generic area-preserving case, the tripling bifurcation can be weak. This occurs, e.g., for the $(11, -3, 3)$ resonance at $(0.1791, -2.4)$.
- $m_2 \geq 4$: here a pair of elliptic/hyperbolic torus links of type (m_1, m_2) is created. If m_1 and m_2 are coprime, there is only a single circle, while if $\gcd(m_1, m_2) = k$, then there are k disjoint circles. This difference is invisible in a slice, and these bifurcations do look like generic bifurcations of area-preserving maps. The order in which the k circles are mapped into each other is determined by $n \bmod k$.

It is possible that the additional cases when $m_2 < 4$ could be described as nongeneric bifurcations of area-preserving maps with $m = m_2$ but in which the lowest-order resonant term vanishes. The approximate Hamiltonian would still be given by (32), but now $\gamma = 0$. For $m = 2$ there are now two cases, depending on whether $|\beta| \gtrless |\delta|$. Thus $m = 2$ becomes similar in some sense to $m = 4$ for the area-preserving case. When $|\beta| > |\delta|$, an elliptic/hyperbolic pair is born like the generic case for $m \geq 5$. However, when $|\beta| < |\delta|$, the bifurcation behaves like the generic $m = 3$ case; i.e., a hyperbolic orbit passes through the center with loss of stability. We have not yet observed this case, but we expect that it is also generic for bifurcations of invariant circles in volume-preserving maps. If $\gamma = 0$ for $m = 3$, an elliptic/hyperbolic pair of orbits are born: this case behaves like the generic case $m \geq 5$. The mechanism by which the lowest-order resonant terms vanishes in these cases presumably is discrete symmetry. We hope to return to this question in a future work.

Besides the local bifurcations discussed so far there are also some interesting sequences of bifurcations associated with period doubling. We have already discussed several doubling bifurcations ($|m_2| = 2$) that result in the permanent loss of stability of the invariant circle, and thus indicate the loss of most of the bounded orbits. There are also doublings for which $\mathcal{C}$ is unstable only for a small range of ε. For example, the frequency map curves shown in Figure 19 sometimes pass through the $(4, \pm 2, 1)$ and $(5, \pm 2, 1)$ resonance lines—with a short interruption—indicating that the original circle regains stability after creating the doubled circle. Even when $|m_1|$ is large, a doubling bifurcation can result in the temporary loss of stability of $\mathcal{C}$. For example, the gap in the frequency map of Figure 16 near $\varepsilon = 0.302$ corresponds to a $(46, -2, 13)$ resonance. A pair of stable circles is created that each wrap around the longitudinal direction once, while encircling the original circle in the transverse direction 23 times; an enlargement is shown in Figure 23.

The class of torus knot (or torus link) bifurcations of invariant circles of volume-preserving maps of $\mathbb{R}^3$ shows a rich bifurcation structure. While there are similarities to area-preserving maps, there is also nontrivial three-dimensional structure, and slices reveal some bifurcations that would be considered nongeneric in the area-preserving category. A more detailed investigation of these bifurcations would be very interesting.

7.2. String of pearls bifurcations. Resonances of an invariant circle with $m_2 = 0$ correspond to rational values of the longitudinal rotation number. We will argue below that this bifurcation corresponds to an SCNS bifurcation for the map f^{m_1}, i.e., the bifurcation discussed in section 5.2. In the supercritical case the invariant circle typically degenerates into a pair of period-m_1 saddle-foci whose two-dimensional stable and unstable manifolds enclose a periodic

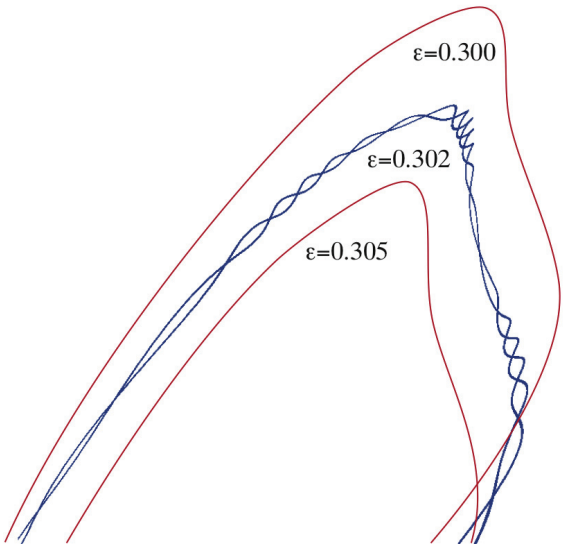


Figure 23. *Portions of the elliptic invariant circles near the* $(46, -2, 13)$ *resonance when* $\mu = -2.4$. *In the gap* $0.3017 < \varepsilon < 0.3045$ *the central circle is unstable. The circles are shifted vertically for clarity.*

family of elliptic invariant circles. The result is a structure that looks like a *string of pearls* on a necklace. This is a common phenomenon in our quadratic family, and we expect this to be a generic bifurcation of invariant circles of volume-preserving maps of $\mathbb{R}^3$. Moreover, a similar bifurcation occurs in dissipative three-dimensional maps [BSV08]; they studied in detail an example with $\omega_L = \frac{1}{5}$.

For example, when $\varepsilon = 0$ and $\mu = -2.0$ for the map (1), $\omega_L = \frac{1}{4}$. As ε is increased from zero, there is a period-four string of pearls created as shown in Figure 24. In the figure the new period-four family of invariant two-tori and several orbits that spiral around the one-dimensional manifolds of the period-four saddle-foci are clearly visible.

This structure can be expected generically. Suppose that as $\omega_L \to \frac{n}{m_1}$ the invariant circle $\mathcal{C}$ remains elliptic and that f can be approximately described by the SCNS normal form (18). When ω_L is rational every point on the invariant circle of the normal form becomes m_1-periodic with multipliers $(1, \lambda^{m_1}, \lambda^{-m_1})$, where $\lambda = e^{2\pi i \omega_T}$. In the resonant case, there are additional terms that must be kept in the normal form, but if $|m_1| > 3$ and ω_T is irrational, then the quadratic terms in f do not change. Consequently f^{m_1} corresponds to the critical case on the saddle-center-Hopf line of Figure 1 and, when any parameter is varied, it will generically undergo an SCNS bifurcation. If this bifurcation is supercritical, the singular fixed point of f^{m_1} will be replaced by a pair of saddle-focus fixed points, and a new elliptic invariant circle will be created. For the original map this corresponds to the creation of a pair of period-m_1 orbits and a period-m_1 family of invariant circles linked with the original circle. Recall from Appendix D that when the unfolding of this bifurcation is constructed, the coordinate transformations include a shift along the direction of the unit eigenvector: this shift selects the particular points along $\mathcal{C}$ that will correspond to the fixed points.

Another example of this bifurcation is shown in Figure 25 for $\omega_L = \frac{1}{5}$. When $\varepsilon = 0$, this longitudinal rotation number occurs at $\mu \approx -1.3819$ according to (7). Fixing $\mu = -1.383$

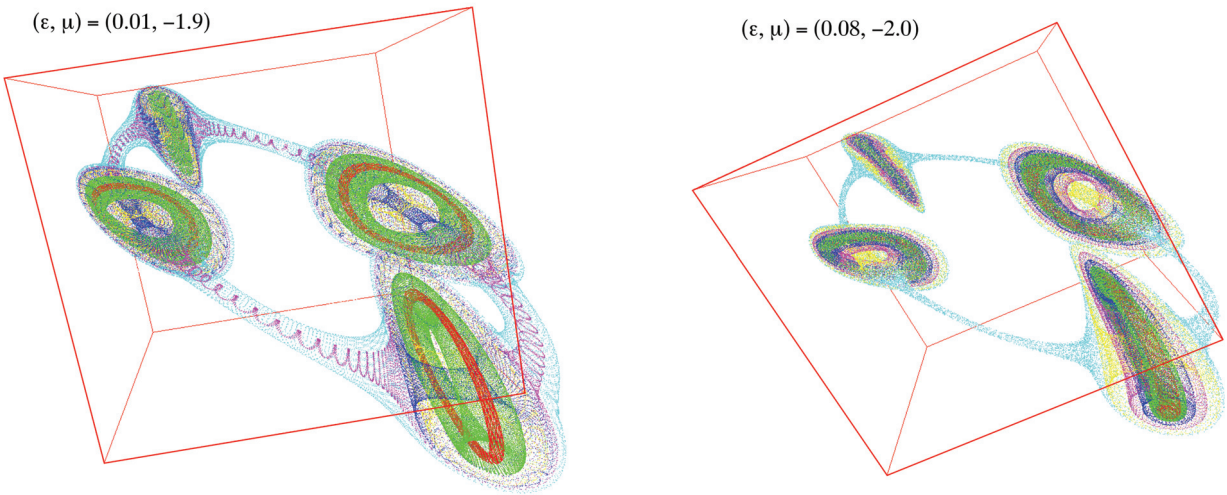


Figure 24. *Phase space near the $\omega_L = \frac{1}{4}$ string of pearls bifurcation. The red cube is centered at the origin and has sides of length $\frac{10}{3}\sqrt{\varepsilon}$.*

and increasing ε from zero, the $(5, 0, 1)$ resonance occurs at $\varepsilon \approx 0.069$ where the "circle" develops five small bulges formed from the nearly coincident two-dimensional manifolds of two new period-five saddle-foci. As ε increases, the points on the new periodic orbits move apart and the corresponding pearls grow in size. The one-dimensional manifolds of the saddle-foci also nearly coincide, forming an approximate circle up to $\varepsilon \approx 0.09$. The string of pearls is still enclosed by a family of two-tori; one of these is shown in the uppermost portrait at $\varepsilon = 0.11$. These tori and the new period-five tori apparently vanish near $\varepsilon = 0.12$, and when ε reaches 0.18 the computation of section 4 reports no bounded orbits. A similar scenario occurs at $(\varepsilon, \mu) = (0.076, -2.6)$, where $\omega_L = \frac{3}{10}$; see Figure 26. In this case the pearls exist only for a very small window in ε, and the circle reforms as ε is increased. Note that the ellipse algorithm is typically able to continue the circle through the string of pearls bifurcation when the slice is not too close to the periodic orbits.

In some cases an $(m_1, 0, n)$ resonance is responsible for the final destruction of the invariant circle. This happens for $\mu = -3.5$ where the circle undergoes an $(8, 0, 3)$ bifurcation at $\varepsilon = 0.266$; see Figure 27. A sequence of three portraits for this case is also shown in Figure 26. As for the $\frac{1}{5}$ and $\frac{3}{10}$ cases, an approximate, frequency-locked invariant circle appears to persist for a small parameter interval beyond the bifurcation (at least up to $\varepsilon = 0.2686$); however, in this case the circle does not reconstitute as ε increases. Destruction of the invariant circle also occurs for $\mu = -2.4$ when $\omega_L = \frac{2}{7}$ at $\varepsilon = 0.32$. This bifurcation was shown in the frequency map Figure 16. The $(7, 0, 2)$ bifurcation appears to be subcritical as we can find no stable orbits beyond $\varepsilon = 0.32$. This is suggested by the rightmost pane of Figure 8 where the first iterates of seven unbounded orbits are plotted. These orbits appear to be near the two-dimensional unstable manifolds of a period-seven saddle, and these manifolds do not appear to intersect as would be predicted by the subcritical case in Figure 6.

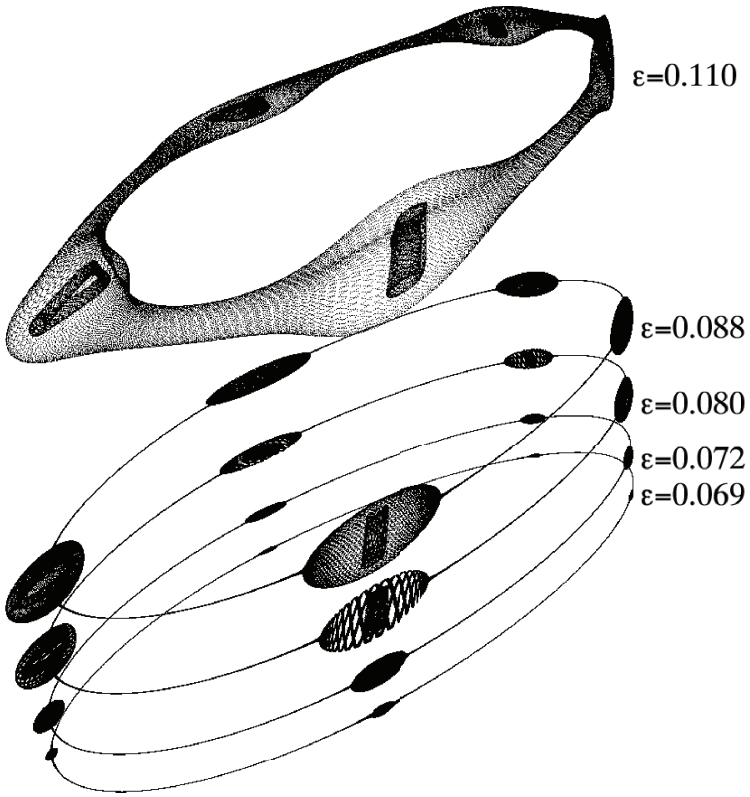


Figure 25. *String of pearls bifurcation for $\mu = -1.383$ and the values of ε shown. The circles are shifted vertically and scaled by $\varepsilon^{-\frac{1}{2}}$ so that their size appears to be roughly fixed.*

7.3. Pairs of saddle-center bifurcations of invariant circles. As we have seen, resonances on an elliptic invariant circle can lead to a local bifurcation that creates new invariant circles linked with the original. Invariant circles can also be created by resonances on two-tori and more generally by saddle-center bifurcations that are not associated with any regular structure. Sometimes saddle-center bifurcations are related to other resonances. For example, near the tripling resonance with $m_2 = \pm 3$ (if $\gamma \neq 0$) a pair of new invariant circles is born in a saddle-center bifurcation, as seen in Figure 22. In the following we discuss a new saddle-center bifurcation that is caused by a resonance with $m_2 = \pm 1$.

We observe that these circle-saddle-center bifurcations come in pairs, as can be seen in the frequency maps of Figure 17, for $m_2 = \pm 1$. In these cases, the circle $\mathcal{C}$ created at $\varepsilon = 0$ approaches a resonance line of the form $(m_1, \pm 1, n)$, and its frequency map curves dramatically to "avoid" crossing the resonance. As this happens, an elliptic/hyperbolic pair of circles are born in a circle-saddle-center bifurcation far from the central circle. This may happen inside a chaotic region or region of unbounded motion or within the family of elliptic two-tori surrounding the invariant circle. In the latter case a torus bifurcates into a figure eight crossed with a circle. According to our observations, the rotation numbers of the elliptic member of the new pair lie on the opposite side of the resonance line. When $\mu = -1.5$, this is the $(5, -1, 1)$ resonance line shown in Figure 17. A sequence of Poincaré sections for this

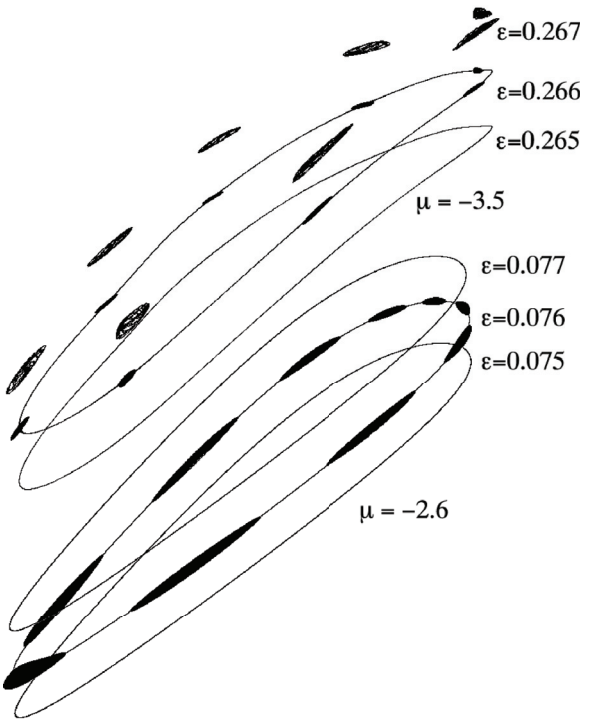


Figure 26. *String of pearls bifurcations for $\mu = -2.6$ (lower three portraits) and $\mu = -3.5$ (upper three portraits), where $\omega_L = \frac{3}{10}$ and $\frac{3}{8}$, respectively.*

case is shown in Figure 28. The new circles appear near $\varepsilon = 0.097$; they form a $(5, -1)$ torus knot. Note, however, that the new elliptic circle (red in the figures) is relatively smooth, and it is more appropriate to think of the highly deformed original circle (blue) as a $(5, -1)$ torus knot wrapping about the new one. As ε grows, the stable region around the original circle shrinks and that around the newly created elliptic circle grows until it takes over as the primary elliptic circle and persists up to $\varepsilon = 0.175$. The old orbit (blue) eventually dies in a circle-saddle-center bifurcation. So overall there is a pair of saddle-center bifurcations of invariant circles, first the creation of a new elliptic circle, and later the destruction of the old elliptic circle. This bifurcation also occurs at $(\varepsilon, \mu) = (0.049, -1.1)$ corresponding to the $(6, -1, 1)$ resonance. When $\mu = -1.9$ a similar bifurcation appears to occur near $\varepsilon = 0.07$ for the resonance $(4, 1, 1)$; however, there are also many other resonances near these parameter values that complicate this case.

8. Conclusion. The three-dimensional analogue of the two-dimensional, area-preserving Hénon map can be found by normal form expansion and unfolding near a triple-one multiplier [DM08]; we believe this quadratic map, (1), should serve as the prototype for volume-preserving dynamics in $\mathbb{R}^3$.

We computed the volume of bounded orbits for (1) as a function of its two essential parameters, (ϵ, μ), obtaining an intricate structure reminiscent of Arnold tongues for circle maps. In the Hénon map bounded motion appears near the elliptic fixed point. By contrast, in the qua-

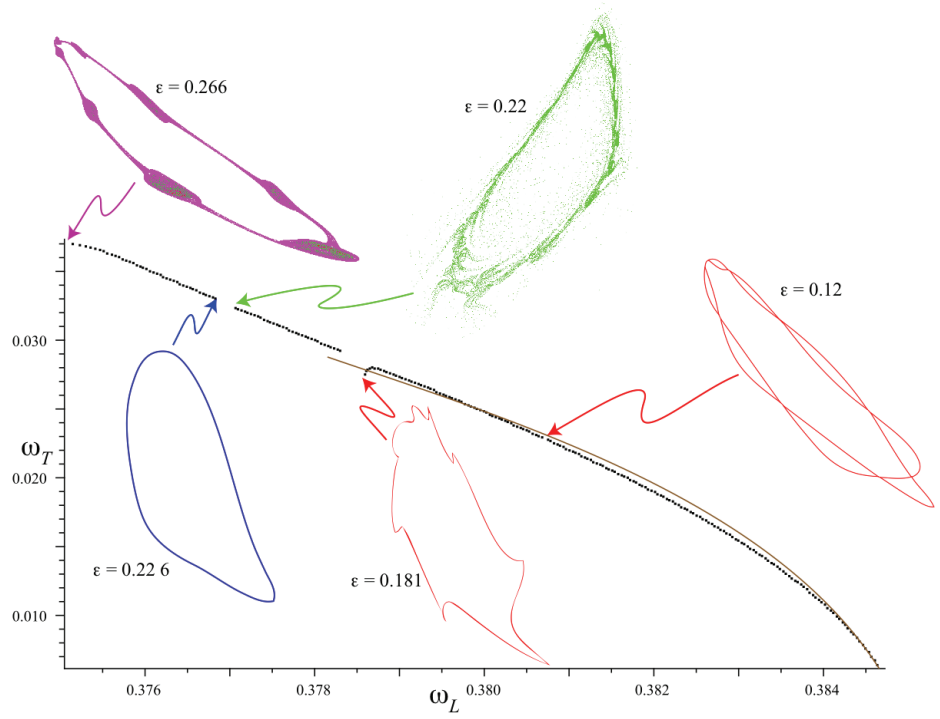


Figure 27. *Frequency map for $\mu = -3.5$. Inserts show the $(8, -2, 3)$ resonance at $\varepsilon = 0.12$ resulting in a doubled circle, $(7, 12, 3)$ at $\varepsilon = 0.181$, $(8, 2, 3)$ at $\varepsilon = 0.22$, and the final disintegration of the invariant circle in the $(8, 0, 3)$ resonance at $\varepsilon = 0.266$ leading to the formation of a string of pearls.*

dratic volume-preserving map bounded motion appears near elliptic invariant circles. These invariant circles in turn are created in SCNS bifurcations. These occur on a codimension-one line in parameter space that emanates from the triple-one-multiplier codimension-two point. The SCNS bifurcation also creates two fixed points that are at the poles of the vortex bubble containing the bounded orbits.

Near the SCNS bifurcation, the map can be decoupled in cylindrical coordinates into a semidirect product of an area-preserving map independent of the angle and a map of the angle depending on all coordinates, (18). A consequence is that, close to the bifurcation, the two fixed points have one-dimensional manifolds that nearly coincide along the bubble's polar axis and two-dimensional manifolds that form the bubble's outer boundary. The curves of transversal intersections of the latter manifolds undergo bifurcations that were described in [LM98]; these give rise to bounded chaotic motion near the boundary of the vortex bubble.

When the SCNC bifurcation is supercritical, it creates an invariant circle $\mathcal{C}$ in the vortex bubble that corresponds to the third, elliptic, fixed point of the area-preserving, decoupled map. We computed the location and frequencies of $\mathcal{C}$ for the normal form and obtained analytical expansions for the longitudinal, ω_L, and transverse, ω_T, frequencies of the invariant circle (or, in fact, for any torus nearby) for the map (1).

The frequency map Ω is an essential object of Kolmogorov–Arnold–Moser (KAM) theory. In a Hamiltonian setting Ω maps n actions into n frequencies for tori of maximal dimension.

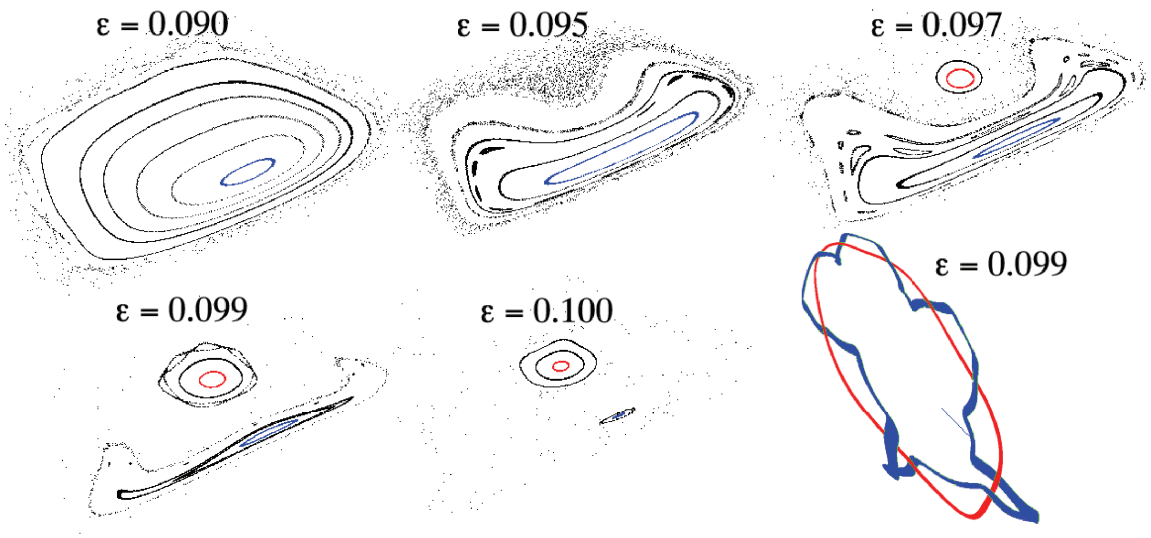


Figure 28. *Circle-saddle-center bifurcation for $\mu = -1.5$. Five Poincaré slices show the creation of a new elliptic circle with $(5,-1)$ helicity near $\varepsilon = 0.096$. The stable region around the new circle grows while that around the original circle shrinks, until near $\varepsilon = 0.101$ the original circle loses stability. Tori near the new original circle are shown in blue and near the new circle are shown in red at $\varepsilon = 0.099$ in the last inset.*

In the volume-preserving case the dimension of the action and angle spaces are typically not equal. In $\mathbb{R}^3$, the tori in the neighborhood of the circle $\mathcal{C}$ have effectively one action, and Ω defines a curve in the two-frequency space that depends upon that single action. As is done in the study of lower-dimensional tori in Hamiltonian systems, the frequency map can be extended to a local diffeomorphism by including parameters in the domain. In this way we study the frequency map $\Omega : (\varepsilon, \mu) \to (\omega_L, \omega_T)$ of the invariant circle of the quadratic map. This mapping organizes the bifurcations of invariant circles. Our analytic approximation for Ω compares very well with the numerical computations.

In order to numerically compute Ω, we developed a simple method to find the location of elliptic invariant circles of maps. The idea is to consider a slice of the mapping that is transverse to the invariant circle and fit a thin ellipsoid to iterated points returning into the slice. This method converges very fast and appears to be a robust and accurate way to continue the invariant circles. Having found a good approximation of $\mathcal{C}$, the computation of the longitudinal rotation number can be made quadratically convergent by considering its continued fraction expansion. For the transverse rotation number we have a linearly convergent numerical method.

Equipped with the continuation method and the frequency map, an astounding variety of resonant bifurcations of invariant circles can be found whenever $m_1\omega_L + m_2\omega_T = n$. We classified these into three cases: string of pearls bifurcations ($m_2 = 0$), paired circle-saddle-center bifurcations ($m_2 = 1$), and torus link bifurcations ($m_2 \geq 2$). More precisely a torus knot of type (m_1, m_2) is found when $k = \gcd(m_1, m_2) = 1$, while a torus link of type (m_1, m_2) consisting of k circles is found when $k > 1$. For example, there are two types of circle-doublings: either the new circle is a double cover of the original one, or there are two new

invariant circles, each covering the original circle once.

The most spectacular bifurcation is the string of pearls that occurs when the longitudinal frequency becomes rational. Here the invariant circle breaks into a pair of saddle-focus periodic orbits, and each pair recapitulates the original vortex bubble on a smaller scale. In this way, we expect that is is possible to find self-similarity in the SCNS bifurcation: the invariant circle bifurcates to create new bubbles that contain invariant circles linked to the original one and that will themselves undergo string of pearls bifurcations under parameter variation, etc. It seems likely that there exist parameter values with infinite sequences of linked pearls around pearls around pearls ....

Similar bifurcations also occur in dissipative three-dimensional maps. For example, circle doublings can occur, but unlike the infinite sequences of doublings of periodic orbits, these appear to be typically finite in one-parameter families [AA84]. Moreover, a bifurcation similar to the string of pearls also occurs in the dissipative case near the codimension-two saddle-center point. The three-parameter unfolding of this bifurcation is discussed by Broer, Simó, and Vitolo [BSV08]; they also studied a model of the dynamics near $\omega_L = \frac{1}{5}$. It is interesting that they observed that there can be an attracting invariant two-torus containing a repelling period-five bubble. We hope to apply similar methods to study the volume-preserving resonant cases in a future paper. These are both simpler (since there is one fewer parameter) and more complex (since there are no attractors) than the dissipative case.

Appendix A. Parameter reduction. Though the map (5) nominally depends upon the three unfolding parameters $(\varepsilon, \mu_1, \mu_2)$, one of these can generically be eliminated by an affine shift in x. If we set $x \to x_0 + \tilde{x}$, then the new map has the same form (5) with $p(x, y, \varepsilon, \mu)$ replaced by $p(x_0+\tilde{x}, \varepsilon, \mu)$. If there is an x_0 so that $\partial_x p(x_0, 0, \varepsilon, \mu) = 0$, then the linear term $\mu_1 x$ can be eliminated. Near the origin, this equation has a solution providing $\partial_x^2 p(0, 0; 0, 0) \neq 0$.[11] In this paper, we assume this is the case, so that without loss of generality we can set $\mu_1 = 0$. Moreover, even though p is nominally a function of (ε, μ), we will take the coefficients of this function to be independent parameters, thus suppressing this dependence. The normal form (5) then becomes (1).

For the quadratic case (1), one of the parameters in P could be eliminated by scaling the variables; for example, under the scaling $\xi \to a\xi$, the parameter a is effectively replaced by 1. Thus for this case (1) reduces to a four-parameter family.

Appendix B. Basic properties. In this section we study some of the simple dynamical properties of (1), especially the fixed points and low period orbits. At first we let P be an arbitrary smooth function that is $\mathcal{O}(2)$ in the variables (x, y) and such that

$$a \equiv \frac{1}{2} P_{xx}(0, 0) > 0, \tag{33}$$

which we used to eliminate the parameter μ_1; see Appendix A.

When P is smooth, the map (1) is a diffeomorphism; indeed, its inverse can easily be

[11] If this is not the case, but $\partial_y^2 p(0, 0, 0, 0) \neq 0$, then we can choose x_0 so that $\partial_y p(x_0, 0, \varepsilon, \mu) = 0$, eliminating the linear term in y.

computed:

$$f^{-1}(x,y,z) = \begin{pmatrix} x-y+z \\ y-z \\ z+\varepsilon-\mu(y-z)-P(x-y+z,y-z) \end{pmatrix}. \tag{34}$$

The fixed points of (1) have the form $(x^*,0,0)$, where

$$P(x^*,0) = \varepsilon.$$

By (33), this equation has two solutions,

$$x^*_\pm \approx \pm\sqrt{\frac{\varepsilon}{a}}\mathcal{O}(\varepsilon),$$

which emerge from the origin for $\varepsilon > 0$. This corresponds to the saddle-center bifurcation in the normal form. Indeed, the Jacobian of f at these points has characteristic polynomial (3) with

$$\tau_\pm = 3+\mu+P_y(x^*_\pm,0), \quad \sigma_\pm = 3+\mu+P_y(x^*_\pm,0)-P_x(x^*_\pm,0).$$

The saddle-center line, corresponding to a multiplier $\lambda_1 = 1$, occurs on the line $\tau = \sigma$ in Figure 1. Since $a \neq 0$, this corresponds to the line $\varepsilon = 0$.

Recall that a diffeomorphism is *reversible* if it is conjugate to its inverse, $r \circ f = f^{-1} \circ r$. The normal form (1) is only reversible if P satisfies special conditions. For example, if

$$P(x,y) = P(-x-y,y),$$

then there is an involutory reversor

$$r(x,y,z) = (-x,y-z,-z). \tag{35}$$

This condition on P is essentially necessary for reversibility as well. Recall that a reversor maps each period-n orbit into another such orbit with reciprocal multipliers. Orbits that lie on the fixed set of r, in this case the y-axis, are mapped into themselves by r and are called *symmetric*. Orbits that are related by the reversor have reciprocal multipliers, and a symmetric orbit must have self-reciprocal multipliers and therefore one unit multiplier. When $\varepsilon \neq 0$, the fixed points do not lie on the saddle-center line and consequently cannot be symmetric. If there are only two such points, then they must be mapped into each other by the reversor and must have reciprocal multipliers. This means that $\tau_\pm = \sigma_\mp$, or equivalently,

$$P_y(x^*_+,0) - P_y(x^*_-,0) = P_x(x^*_+,0) = -P_x(x^*_-,0).$$

For the quadratic case, this implies that P has the form

$$P = a(x^2+xy)+cy^2$$

and that the reversor is (35). Consequently, the map (1) is only rarely reversible.

For the case that P is a positive definite quadratic form, that is, $a > 0$ and $D \equiv ac - b^2/4 > 0$, it was shown in [LM98] that every bounded orbit of the quadratic map is contained in the cube $|\xi|_\infty \le \kappa_B$, where

$$\kappa_B = \alpha \left(|3+\mu| + 1 + \sqrt{(|3+\mu|+1)^2 + |\varepsilon|/\alpha} \right),$$

where $\alpha = \frac{1}{D}\max(c, a-b+c)$.[12]

Appendix C. Quadratic map fixed points. The stability properties of the fixed points (6) of the quadratic map (1) are easily determined from the characteristic polynomial (3). For the fixed points the trace and second trace become

$$\tau_\pm = 3 + \mu + bx_\pm,$$
$$\sigma_\pm = 3 + \mu + (b-2a)x_\pm.$$

Since $a > 0$, the stability diagram Figure 1 maps diffeomorphically onto the $(\mu, x_\pm)$ plane the half-planes $\tau - \sigma \lessgtr 0$ corresponding to $x_\pm$, respectively. As for the case of a general nonlinear function P, the saddle-center bifurcation curve corresponds to $\varepsilon = 0$. Along this curve, when $-1 < \tau = \sigma < 3$ the other two multipliers have the form $e^{\pm 2\pi i\omega}$, which implies that $-4 < \mu < 0$ and $\omega = \omega_0(\mu)$; see (7). Thus we call this the *saddle-center-Hopf* segment.

Note that for most parameter values, one fixed point is type $(2,1)$ (two-dimensional stable manifold) and the other is type $(1,2)$ (two-dimensional unstable manifold). The exception corresponds to a parabolic region near $\mu = -4$ that happens to coincide with the existence of a period-two orbit. Indeed, the period-doubling bifurcation in Figure 1 corresponds to $\tau + \sigma = -2$. For the quadratic map, this implies $\mu = (a-b)x_\pm - 4$, which is equivalent to the parabola

$$\Delta \equiv (a-b)^2\varepsilon - a(\mu+4)^2 = 0, \tag{36}$$

as shown by the dashed lines in Figure 2. It is easy to see that the quadratic map has at most one period-two orbit, which has the form $(x,y,z) \to (x+y, -y, -z) \to (x,y,z)$, where

$$(x,y,z) = \frac{1}{a-b}\left(\mu + 4 \mp \rho, \pm 2\rho, \pm 4\rho\right).$$

Consequently, this orbit exists only when $a \ne b$ and

$$\rho^2 = \frac{\Delta}{a - 2b + 4c} > 0.$$

For the case that $a - 2b + 4c > 0$, the orbit exists in the interior of the period-doubling parabola (the dashed curves in Figure 2); otherwise, it exists in the exterior of this parabola. The characteristic polynomial for the period-two orbit has the form (3) with

$$\tau - \sigma = \frac{4\Delta}{a-b}.$$

As expected, the period-two orbit is born on the parabola $\Delta = 0$ with a multiplier 1. Thus, for fixed (a,b,c) only four of the eight stability regimes of Figure 1 are reached; an example

[12] This corrects an error in [LM98] in the formula for κ on page 572.

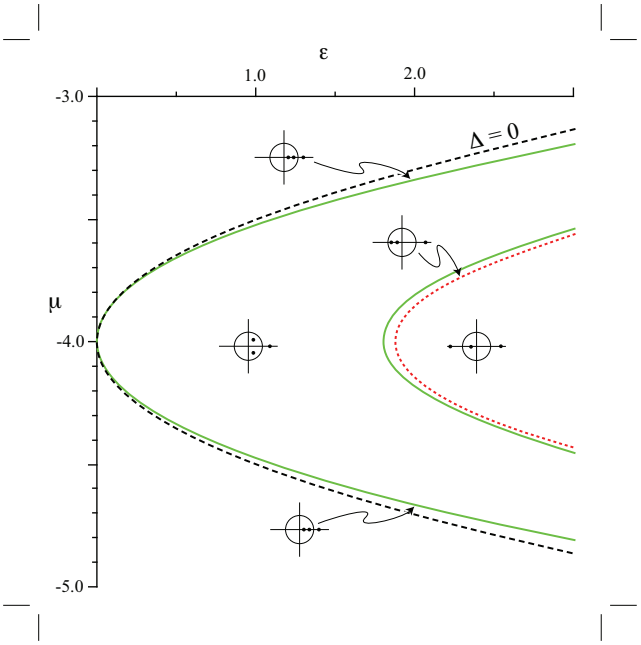


Figure 29. *Stability diagram for the period-two orbit of* (1) *with the polynomial* (2) *and* $a = 1$, $b = c = 0.5$. *The dashed (black) parabola is the curve* $\Delta = 0$, *where one multiplier of the period-two orbit is* 1. *This is the same as the period-doubling curve in Figure* 2. *The period-two orbit exists for* $\Delta > 0$, *to the right of the* $\Delta = 0$ *parabola. The solid (green) curves correspond to a double eigenvalue, and the dotted (red) curve to a multiplier* -1.

for $a > b$ and $a+4c > 2b$ is shown in Figure 29. In this case, the period-two orbit is created on the upper half of the doubling parabola $\Delta = 0$ by being emitted from the fixed point $(x_+, 0, 0)$ and is destroyed on the lower half of the parabola by being absorbed into the second fixed point $(x_-, 0, 0)$.

The two curves of double eigenvalues in Figure 1 correspond to $\lambda_1 = \lambda_2 = r$, giving $\tau = 2r + 1/r^2$ and $\sigma = r^2 + 2/r$. For the fixed point these become

$$2ax_\pm = -r^2 + 2r - 2/r + 1/r^2,$$
$$\mu = -bx_\pm + \left(2r + \frac{1}{r^2}\right) - 3.$$

These are shown as solid curves in Figure 2.

Finally, fixed points with multipliers of the form $\lambda = (re^{2\pi i\omega}, re^{-2\pi i\omega}, \frac{1}{r^2})$ have $\tau = 2r\cos(2\pi\omega) + 1/r^2$ and $\sigma = r^2 + 2\cos(2\pi\omega)/r$, which implies

$$2ax_\pm = \left(r - \frac{1}{r}\right)\left(2\cos(2\pi\omega) - r - \frac{1}{r}\right), \tag{37}$$
$$\mu = 2r\cos(2\pi\omega) + \frac{1}{r^2} - 3 - bx_\pm.$$

Several curves of this type for fixed rational values of ω are shown in Figure 2.

The quadratic map (1) has at most one pair of period-three orbits; this follows because the set of equations $\xi = f^3(\xi)$ can be reduced to a degree-eight polynomial in one variable. This polynomial has a quadratic factor with zeros at the fixed points; the remaining sixth degree equation has zeros on the pair of period-three orbits. The period-three orbits are created in a saddle-center bifurcation along the curve

$$\varepsilon = K(\mu + 3)^2,$$

where

$$K = \frac{a(2a - 3b + 6c)((a - b)^2 + 3c(a - b + c))}{(a^2 - 3ab + 3ac + 3b^2 - 9bc + 9c^2)(a^3 - 3a^2b + 3a^2c + ac^2 + 2b^2c - 5abc + 3ab^2 - b^3)}.$$

For example, when $a = 1$, $b = c = \frac{1}{2}$, $K = 4$. Note that this curve is not the same as the tripling curve of the fixed point.

One of the period-three orbits is type-$(2, 1)$ and the other type-$(1, 2)$. When $\mu < -3$, the period-three orbits are born with real eigenvalues, corresponding to $\tau = \sigma > 3$, and (for $a = 1$, $b = c = \frac{1}{2}$) when $-3 < \mu < -2.24$, they are spiral-saddles (saddle-center-Hopf bifurcation). At $\mu \approx -2.24$ the period-three orbits sit at the point $\tau = \sigma = -1$, the $(-1, -1, 1)$ point. ε just to the right of this parabola, these orbits have real eigenvalues. They then cross a curve of double multipliers and become spiral saddles.

Overall it should be noted that the importance of fixed points for three-dimensional volume-preserving maps is far less than for two-dimensional area-preserving maps, since they are generically unstable. The important objects in the three-dimensional volume-preserving case are elliptic invariant circles.

Appendix D. SCNS normal form. In this section we summarize the computation of the normal form for the SCNS bifurcation of a volume-preserving map that is discussed in section 5.2, and thus the technical part of the proof of Theorem 1. In the complex coordinates (u, v, z) the map takes the form

$$\zeta' = f(\zeta) = M(\zeta + b(\zeta)), \quad M = \mathrm{diag}(e^{2\pi i\omega}, e^{-2\pi i\omega}, 1), \tag{38}$$

where $b = \mathcal{O}(2)$ represents the nonlinear terms.

The formal normal form for f is obtained by constructing a near identity transformation $\eta = \psi(\zeta) = \zeta + h(\zeta)$ such that the conjugacy $g \circ \psi = \psi \circ f$ gives a new map

$$g(\eta) = M(\eta + c(\eta))$$

that is "simpler" than f—for example, it contains fewer nonlinear terms. Assuming that f is in normal form through terms of degree $d - 1$, this means that we must find a degree-d vector function h such that the conjugacy holds to degree d. To do this we must solve the homological equation, $\mathrm{L}_M h(\zeta) = c(\zeta) - b(\zeta)$, at degree d for h, where

$$\mathrm{L}_M h(\zeta) \equiv M^{-1} h(M\zeta) - h(\zeta). \tag{39}$$

The terms in b that are in the range of L_M can be eliminated by the transformation. Thus the nonlinear terms c in the normal form correspond to the parts of b in a complement to the range of L_M.

The normal form is formal in the sense that the map and transformation are expanded in formal power series in the variables (u, v, z). A basis for the space of vectors of polynomials of degree d are the "vector monomials,"

$$p_{i,m} = \zeta^m \hat{e}_i, \quad m \in \mathbb{N}^3, \tag{40}$$

where $\hat{e}_i$, $i = 1, 2, 3$, are the unit vectors $\hat{e}_{i,j} = \delta_{i,j}$, $\zeta^m = u^{m_1} v^{m_2} z^{m_3}$, and $|m| \equiv \Sigma_{i=1}^3 m_i = d$. When M is diagonal, the homological operator (39) has a diagonal representation in the monomial basis (40):

$$\mathrm{L}_M(p_{i,m}) = \left(\frac{\lambda^m}{\lambda_i} - 1\right) p_{i,m}.$$

Consequently $\ker(\mathrm{L}_M)$ is spanned by those monomials that satisfy the "resonance conditions"

$$\lambda^m = \lambda_i, \tag{41}$$

and we can choose the normal form to lie in the kernel of L_M. The elements of the kernel correspond to eigenvectors with zero eigenvalue; if we assume that ω in (16) is irrational, then the resonance condition is satisfied for arbitrary integers m_3, and when

$$m_1 - m_2 = \begin{cases} 1, & i = 1, \\ -1, & i = 2, \\ 0, & i = 3. \end{cases}$$

Thus to arbitrary degree, the normal form can be written

$$\begin{aligned} u' &= \lambda u(1 + c_1(|u|^2, z)), \\ z' &= z + c_3(|u|^2, z), \end{aligned} \tag{42}$$

where c_i are arbitrary polynomials subject to the requirements that $c_i(0, 0) = 0$ and $D_z c_3(0, 0) = 0$.

Since f is volume-preserving, h can be selected to be volume-preserving as well [DM08]; to implement this efficiently, it is better to use Lie series instead of power series. We compute the power series by truncating the exponentials. Through cubic degree (42) becomes

$$\begin{aligned} u' &= \lambda u(1 + Az + Bz^2 + C|u|^2) + \mathcal{O}(4), \\ z' &= z + \alpha z^2 + \beta|u|^2 + \gamma|u|^2 z + \kappa z^3 + \mathcal{O}(4), \end{aligned} \tag{43}$$

where $A, B, C \in \mathbb{C}$ and $\alpha, \beta, \gamma, \kappa \in \mathbb{R}$. This map is volume-preserving to $\mathcal{O}(3)$ when

$$\begin{aligned} A_r &= -\alpha, \\ 2B_r &= 3(\alpha^2 - \kappa) - A_i^2, \\ 4C_r &= -\gamma - 2\alpha\beta, \end{aligned} \tag{44}$$

where we denote $A = A_r + iA_i$, etc.[13]

[13] Alternatively, we could eliminate $\alpha = -A_r$, $\gamma = -4C_r + 2A_r\beta$, and $\kappa = \alpha^2 - \frac{2}{3}B_r - \frac{1}{3}A_i^2$.

The normal form (42) can be unfolded by assuming that the map f depends upon a set of parameters p. Following Elphick et al. [ETB+87], formally expand f, h, and g in double power series in p and ζ; then, for each degree in p, the coefficients of h must still satisfy the homological equation. To zeroth order in p and all orders in ζ, the normal form is given by (42). The new feature occurs at first order in p: there are now terms in f that are constant and terms that are linear in the variables ζ. For the constant terms, the homological equation reduces to

$$(M - I)h_0 = c_0 - b_0. \tag{45}$$

Thus any terms in b not in the null space of $M - I$ can be removed. This implies that there is generically an affine term $(0, 0, \delta(p))$ in the normal form at this order. Since this is the sole term that occurs at this order, we can use δ as a primary parameter; it corresponds to unfolding the saddle-center bifurcation.

The terms of first degree in ζ can be represented by matrices, e.g., $h_1(\zeta) = H\zeta$, etc., and the homological equation becomes the matrix equation

$$M^{-1}HM - H = C - B. \tag{46}$$

Since M is diagonal, this operator is diagonal in the components of H, and its kernel consists of the matrices that commute with M: the diagonal matrices. Thus the normal form must include an added diagonal linear perturbation. For the volume-preserving case when $\omega \neq 0$ or $\frac{1}{2}$, this means that

$$M \to \operatorname{diag}\left(\Lambda e^{2\pi i\tilde{\omega}}, \Lambda e^{-2\pi i\tilde{\omega}}, \frac{1}{\Lambda^2}\right)$$

for a new frequency $\tilde{\omega} = \omega + O(p)$ and expansion/contraction factor $\Lambda = 1 + O(p)$. The three parameters $(\delta, \Lambda, \tilde{\omega})$ unfold the saddle-center Hopf-bifurcation; however, as we will see below, one of them can generically be eliminated.

For all terms of higher degree in ζ, the homological equations are the same as before; consequently, the form of the terms in the normal form is the same as in (43), except that the coefficients are now allowed to depend upon the parameters.

To quadratic degree, the unfolded normal form is thus

$$\begin{aligned} u' &= \Lambda e^{2\pi i\omega} u(1 + Az) + \mathcal{O}(3), \\ z' &= -\delta + z/\Lambda^2 + \alpha z^2 + \beta|u|^2 + \mathcal{O}(3). \end{aligned} \tag{47}$$

Note that when α is nonzero, the implicit function theorem implies the existence of an affine shift $\tilde{z} = z + \Delta$ such that the new map has unit multiplier at the new origin. Hence this shift can be used eliminate Λ, and the final normal form is identical to (43) with the addition of an affine term $-\delta$ in the z' equation. This coordinate transformation modifies the rotation number ω as well.

Transforming (43) to symplectic, cylindrical coordinates gives the final result (18).

Appendix E. Resonance tables. Table 2 shows some parameter values for different string of pearls bifurcations. Table 3 shows low-order resonances encountered in the family with $\mu = -2.4$.

Table 2

Parameters for some of the string of pearls bifurcations.

μ	ε	m_1	m_2	n
-1.95	0.23	4	0	1
-2.4	0.318	7	0	2
-2.6	0.0760	10	0	3
-2.6	0.095	283	0	85
-2.7	0.027	13	0	4
-2.8	0.0354	41	0	13
-3.3	0.1164	14	0	5
-3.5	0.265	8	0	3
-3.8	0.24	12	0	5
-3.85	0.152	7	0	3
-3.85	0.305	19	0	8
-3.9	0.053	9	0	4

Table 3

Approximate resonances, $|m \cdot \omega - n| < 0.001$, with $|m_i| \le 10$ for invariant circles with $\mu = -2.4$; see Table 1 and Figure 16.

ε range	m_1	m_2	n	ε range	m_1	m_2	n
0.016–0.017	7	2	2	0.146–0.148	3	4	1
0.019	4	-10	1	0.157–0.159	8	-7	2
0.023	4	-9	1	0.168–0.17	10	4	3
0.026	3	10	1	0.171–0.173	-1	7	0
0.032	3	9	1	0.183	5	-10	1
0.039	4	-7	1	0.184–0.185	6	7	2
0.041	3	8	1	0.192	2	10	1
0.042	10	9	3	0.200	9	10	3
0.048–0.052	7	1	2	0.210–0.215	4	-3	1
0.053	4	-6	1	0.221	5	-9	1
0.054	7	1	2	0.224–0.226	-1	6	0
0.066–0.067	10	7	3	0.228-0.229	2	9	1
0.07–0.071	3	6	1	0.233–0.237	3	3	1
0.07632–0.07702	4	-5	1	0.250–0.254	10	3	3
0.087	10	6	3	0.267	5	-8	1
0.088	-1	10	0	0.272–0.273	2	8	1
0.095	8	-9	2	0.277	9	8	3
0.098–0.099	3	5	1	0.287–0.288	8	-5	2
0.107–0.108	-1	9	0	0.291	9	-10	2
0.118–0.12	10	5	3	0.293–0.294	-1	5	0
0.121–0.122	4	-4	1	0.296	5	10	2
0.134–0.135	-1	8	0	0.298–0.299	6	5	2

REFERENCES

[AA84] F. Argoul and A. Arneodo, *From quasiperiodicity to chaos—An unstable scenario via period-doubling bifurcations of tori*, Journal De Mecanique Theorique Et Appliquee, Special Issue (1984), pp. 241–288.

[ACS83] A. Arneodo, P. H. Coullet, and E. A. Spiegel, *Cascade of period doublings of tori*, Phys. Lett. A, 94 (1983), pp. 1–6.

[Ada04] C. C. Adams, *The Knot Book: An Elementary Introduction to the Mathematical Theory of*

Knots, AMS, Providence, RI, 2004.

[BC93] T. J. Bridges and R. H. Cushman, *Unipotent normal forms for symplectic maps*, Phys. D, 65 (1993), pp. 211–241.

[BP98] E. Bedford and V. Pambuccian, *Dynamics of shift-like polynomial diffeomorphism of* $\mathbb{C}^n$, Conform. Geom. Dyn., 2 (1998), pp. 45–55.

[Bro81] H. W. Broer, *Formal normal form theorems for vector fields and some consequences for bifurcations in the volume preserving case*, in Dynamical Systems and Turbulence, Warwick 1980 (Coventry, 1979/1980), Lecture Notes in Math. 898, Springer-Verlag, Berlin, 1981, pp. 54–74.

[BSV08] H. W. Broer, C. Simó, and R. Vitolo, *Hopf-saddle-node bifurcation for fixed points of* 3*D-diffeomorphisms: Analysis of a resonance 'bubble,'* Phys. D, 237 (2008), pp. 1773–1799.

[CFS82] I. P. Cornfeld, S. V. Fomin, and Y. G. Sinai, *Ergodic Theory*, Grundlehren Math. Wiss. 245, Springer-Verlag, New York, 1982.

[CJ00] E. Castellà and À. Jorba, *On the vertical families of two-dimensional tori near the triangular points of the bicircular problem*, Celestial Mech. Dynam. Astronom., 76 (2000), pp. 35–54.

[Cow73] S. W. H. Cowley, *A qualitative study of the reconnection between the earth's magnetic field and an interplanetary field of arbitrary orientation*, Radio Sci., 8 (1973), pp. 903–913.

[CS90a] C.-Q. Cheng and Y.-S. Sun, *Existence of invariant tori in three-dimensional measure-preserving mappings*, Celestial Mech. Dynam. Astronom., 47 (1990), pp. 275–292.

[CS90b] C.-Q. Cheng and Y.-S. Sun, *Existence of periodically invariant curves in* 3*-dimensional measure-preserving mappings*, Celestial Mech. Dynam. Astronom., 47 (1990), pp. 293–303.

[DI96] F. Dumortier and S. Ibáñez, *Nilpotent singularities in generic* 4*-parameter families of* 3*-dimensional vector fields*, J. Differential Equations, 127 (1996), pp. 590–647.

[DIK01] F. Dumortier, S. Ibáñez, and H. Kokubu, *New aspects in the unfolding of the nilpotent singularity of codimension three*, Dyn. Syst., 16 (2001), pp. 63–95.

[DLM06] B.-S. Du, M.-C. Li, and M. I. Malkin, *Topological horseshoes for Arneodo–Coullet–Tresser maps*, Regul. Chaotic Dyn., 11 (2006), pp. 181–190.

[DM08] H. R. Dullin and J. D. Meiss, *Normal forms for divergence-free vector fields and volume-preserving maps*, Phys. D, 237 (2008), pp. 156–166.

[DSM00] H. R. Dullin, D. Sterling, and J. D. Meiss, *Self-rotation number using the turning angle*, Phys. D, 145 (2000), pp. 25–46.

[EL03] K. D. Edoh and J. Lorenz, *Numerical approximation of rough invariant curves of planar maps*, SIAM J. Sci. Comput., 25 (2003), pp. 213–223.

[ETB+87] C. Elphick, E. Tirapegui, M. E. Brachet, P. Coullet, and G. Iooss, *A simple global characterization for normal forms of singular vector fields*, Phys. D, 29 (1987), pp. 95–127.

[EV01] K. Efstathiou and N. Voglis, *A method for accurate computation of the rotation and twist numbers for invariant tori*, Phys. D, 158 (2001), pp. 151–163.

[FM89] S. Friedland and J. Milnor, *Dynamical properties of plane polynomial automorphisms*, Ergodic Theory Dynam. Systems, 9 (1989), pp. 67–99.

[FW98] J. E. Fornaess and H. Wu, *Classification of degree* 2 *polynomial automorphisms of* $\mathbb{C}^3$, Publ. Mat., 42 (1998), pp. 195–210.

[GH02] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Appl. Math. Sci. 42, Springer-Verlag, New York, 2002.

[GICH80] G. C. Goldenbam, J. H. Irby, Y. P. Chong, and G. W. Hart, *Formation of a spheromak plasma configuration*, Phys. Rev. Lett., 44 (1980), pp. 393–396.

[GMO06] S. V. Gonchenko, J. D. Meiss, and I. I. Ovsyannikov, *Chaotic dynamics of three-dimensional Hénon maps that originate from a homoclinic bifurcation*, Regul. Chaotic Dyn., 11 (2006), pp. 191–212.

[GOST05] S. V. Gonchenko, I. I. Ovsyannikov, C. Simó, and D. V. Turaev, *Three-dimensional Hénon-like maps and wild Lorenz-like attractors*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 15 (2005), pp. 3493–3508.

[HDLL06a] A. Haro and R. De La Llave, *A parameterization method for the computation of invariant tori and their whiskers in quasi-periodic maps: Numerical algorithms*, Discrete Contin. Dyn. Syst. Ser. B, 6 (2006), pp. 1261–1300.

[HDLL06b] A. Haro and R. De La Llave, *A parameterization method for the computation of invariant tori and their whiskers in quasi-periodic maps: Rigorous results*, J. Differential Equations, 228 (2006), pp. 530–579.

[HDLL07] A. Haro and R. De La Llave, *A parameterization method for the computation of invariant tori and their whiskers in quasi-periodic maps: Explorations and mechanisms for the breakdown of hyperbolicity*, SIAM J. Appl. Dyn. Syst., 6 (2007), pp. 142–207.

[Hén69] M. Hénon, *Numerical study of quadratic area-preserving mappings*, Q. J. Appl. Math., 27 (1969), pp. 291–312.

[Her83] M. R. Herman, *Une méthode pour minorer les exposants de Lyapunov et quelques exemples montrant le caractère local d'un théorème d'Arnold et de Moser sur le tore de dimension* 2, Comment. Math. Helv., 58 (1983), pp. 453–502.

[HM98] G. Haller and I. Mezic, *Reduction of three-dimensional, volume-preserving flows with symmetry*, Nonlinearity, 11 (1998), pp. 319–339.

[Hol84] P. Holmes, *Some remarks on chaotic particle paths in time-periodic, three-dimensional swirling flows*, in Fluids and Plasmas: Geometry and Dynamics, Contemp. Math. 28, AMS, Providence, RI, 1984, pp. 393–404.

[HW79] G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers*, Clarendon Press, Oxford, UK, 1979.

[JK06] T. H. Jäger and G. Keller, *The Denjoy type of argument for quasiperiodically forced circle diffeomorphisms*, Ergodic Theory Dynam. Systems, 26 (2006), pp. 447–465.

[JS06] T. H. Jäger and J. Stark, *Towards a classification for quasiperiodically forced circle homeomorphisms*, J. London Math. Soc. (2), 73 (2006), pp. 727–744.

[Kuz04] Yu. A. Kuznetsov, *Elements of Bifurcation Theory*, 3rd ed., Appl. Math. Sci. 112, Springer-Verlag, New York, 2004.

[Lam45] H. Lamb, *Hydrodynamics*, 6th ed., Dover, New York, 1945.

[LCC06] Y. Lan, C. Chandre, and P. Cvitanovic, *Variational method for locating invariant tori*, Phys. Rev. E, 74 (2006), paper 046206.

[LM98] H. E. Lomelí and J. D. Meiss, *Quadratic volume-preserving maps*, Nonlinearity, 11 (1998), pp. 557–574.

[LM00] H. E. Lomelí and J. D. Meiss, *Heteroclinic primary intersections and codimension one Melnikov method for volume-preserving maps*, Chaos, 10 (2000), pp. 109–121.

[Mac94] R. S. MacKay, *Transport in* 3*D volume-preserving flows*, J. Nonlinear Sci., 4 (1994), pp. 329–354.

[Mae01] K. Maegawa, *Classification of quadratic polynomial automorphisms of* C^3 *from a dynamical point of view*, Indiana Univ. Math. J., 50 (2001), pp. 935–951.

[MH92] K. R. Meyer and G. R. Hall, *Introduction to Hamiltonian Dynamical Systems and the N-Body Problem*, Springer, New York, 1992.

[Mur03] J. Murdock, *Normal Forms and Unfoldings for Local Dynamical Systems*, Springer Monographs in Mathematics, Springer-Verlag, New York, 2003.

[Sim98] C. Simó, *Effective computations in celestial mechanics and astrodynamics*, in Modern Methods of Analytical Mechanics and Their Applications, CISM Courses and Lectures 387, V. V. Rumyantsev and A. V. Karapetyan, eds., Springer, New York, 1998, pp. 55–102.

[SOV05] F. Schilder, H. M. Osinga, and W. Vogt, *Continuation of quasi-periodic invariant tori*, SIAM J. Appl. Dyn. Syst., 4 (2005), pp. 459–488.

[Xia96] Z. Xia, *Homoclinic points in symplectic and volume-preserving diffeomorphisms*, Comm. Math. Phys., 177 (1996), pp. 435–449.

SIAM J. Applied Dynamical Systems
Vol. 8, No. 1, pp. 129–145
© 2009 Society for Industrial and Applied Mathematics

On the Implementation of the 0–1 Test for Chaos*

Georg A. Gottwald† and Ian Melbourne‡

Abstract. In this paper we address practical aspects of the implementation of the 0–1 test for chaos in deterministic systems. In addition, we present a new formulation of the test which significantly increases its sensitivity. The test can be viewed as a method for distilling a binary quantity from the power spectrum. The implementation is guided by recent results from the theoretical justification of the test as well as by exploring better statistical methods for determining the binary quantities. We give several examples to illustrate the improvement.

Key words. deterministic dynamical systems, chaos, time series analysis

AMS subject classifications. 37M10, 65P20, 37D45

DOI. 10.1137/080718851

1. Introduction.

Being able to distinguish between regular and chaotic dynamics in a deterministic system is an important question with applications ranging from cardiac arrhythmias to the stability of our solar system. Much progress has been made in developing tests for chaos [13, 14, 24, 7, 2, 1]. Recently we have introduced a binary test for chaos, the 0–1 test, designed for the analysis of deterministic dynamical systems [8, 9]. The test distinguishes between regular and chaotic dynamics for a deterministic system. The nature of the dynamical system is irrelevant for the implementation of the test; it is applicable to data generated from maps, ordinary differential equations, and partial differential equations. The test has been applied to noisy numerical data [9], experimental data [5], quasi-periodically forced systems and strange nonchaotic attractors [3], Hamiltonian systems [25], nonsmooth systems [15], and fluid dynamics [17].

The usual test of whether a deterministic dynamical system is chaotic or nonchaotic involves the calculation of the maximal Lyapunov exponent λ [13]. A positive maximal Lyapunov exponent indicates chaos: if $\lambda > 0$, then nearby trajectories separate exponentially, and if $\lambda \leq 0$, then nearby trajectories remain in a close neighborhood of each other. This approach has been widely used for dynamical systems whose equations are known. If the equations are not known or one wishes to examine experimental data, then λ may be estimated using the phase space reconstruction method of Takens [26], by approximating the linearization of the evolution operator [23], or by the "direct method" [22].

*Received by the editors March 19, 2008; accepted for publication (in revised form) by T. Sauer September 23, 2008; published electronically January 9, 2009.
http://www.siam.org/journals/siads/8-1/71885.html

†Mathematics and Statistics, University of Sydney, NSW 2006, Australia (gottwald@maths.usyd.edu.au). This author is partly supported by ARC grants DP0452147 and DP0667065.

‡Mathematics and Statistics, University of Surrey, Guildford, Surrey GU2 7XH, UK (ism@math.uh.edu). The research of this author was partly supported by EPSRC grant EP/D055520/1 and by a Leverhulme Research Fellowship.

In contrast, our test does not depend on phase space reconstruction but rather works directly with the time series given. The main advantages of our test are (i) it is binary (minimizing issues of distinguishing small positive numbers from zero), (ii) the nature of the vector field as well as its dimensionality does not pose practical limitations, and (iii) it does not suffer from the difficulties associated with phase space reconstruction [13].

In this paper, we describe in detail how to implement the 0–1 test for chaos. In addition, we carry out modifications to the test that greatly improve the previous versions in [8, 9].

Throughout the paper, we use the logistic map to illustrate our claims, with the exception of section 6, where we use the Lorenz attractor as an example of a continuous time system. The reader can verify that our results apply equally well to other systems, including those considered in our previous papers [8, 9].

1.1. Recipe for the 0–1 test. We briefly review how the test is implemented. Given an observation $\phi(j)$ for $j = 1, \dots, N$, we perform the following sequence of steps:

1. For $c \in (0, \pi)$, we compute the translation variables

$$p_c(n) = \sum_{j=1}^{n} \phi(j) \cos jc, \quad q_c(n) = \sum_{j=1}^{n} \phi(j) \sin jc \tag{1.1}$$

for $n = 1, 2, \dots, N$. Typical plots of p and q for regular and chaotic dynamics are given in Figure 1.

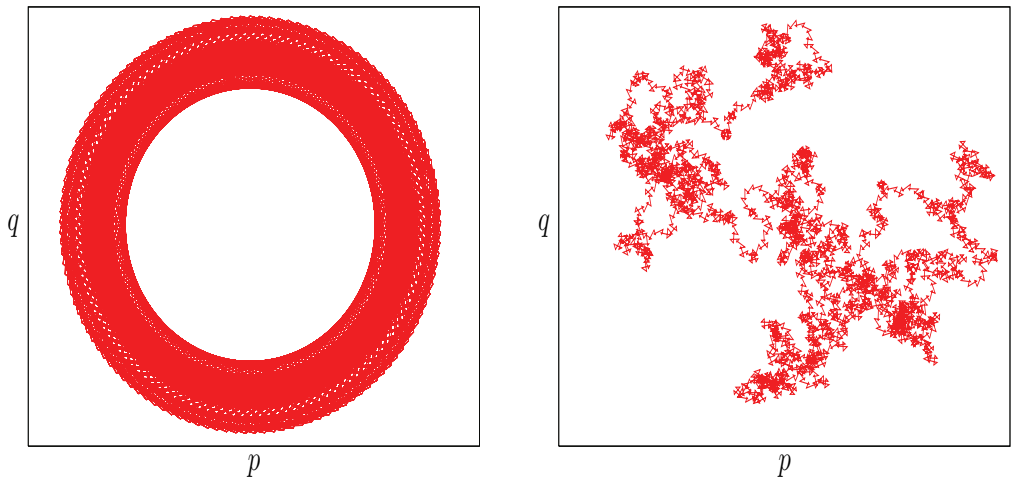


Figure 1. *Plot of p versus q for the logistic map $x_{n+1} = \mu x_n(1 - x_n)$. Left: Regular dynamics at $\mu = 3.55$; Right: Chaotic dynamics at $\mu = 3.9$. We used 5000 data points.*

2. The diffusive (or nondiffusive) behavior of p_c and q_c can be investigated by analyzing the mean square displacement $M_c(n)$. The theory behind our test assures that if the dynamics is regular, then the mean square displacement is a bounded function in time, whereas if the dynamics is chaotic, then the mean square displacement scales linearly with time. In section 2 we look at expressions for the mean square displacement and describe how one may use analytical expressions derived in [11] to conveniently modify the expression for the mean square displacement.

3. We then compute the asymptotic growth rate K_c of the mean square displacement. Methods for the most effective estimation of this are discussed in section 3.
4. Steps 1–3 are performed for N_c values of c chosen randomly in the interval $(0, \pi)$. In practice, $N_c = 100$ is sufficient. The choice of c is discussed further in section 4. We then compute the median of these N_c values of K_c to compute the final result $K = \text{median}(K_c)$. Our test states that a value of $K \approx 0$ indicates regular dynamics, and $K \approx 1$ indicates chaotic dynamics.

In this paper, we explore practical issues arising in the implementation of the above algorithm. Various issues associated with steps 2–4 are discussed in sections 2–4, respectively. In section 5 we examine finite data size effects. In particular we look at weak chaos. In section 6 we consider continuous time systems where oversampled data can lead to small values of K despite an underlying chaotic dynamics. In section 7 we investigate the issue of measurement noise.

Remark 1.1. In the first version of our test, introduced in [8], we defined $p_c(n)$ and $q_c(n)$ by iterating the extended system

$$\begin{aligned} p_c(n+1) &= p_c(n) + \phi(n)\cos(\vartheta_c(n)), \\ q_c(n+1) &= q_c(n) + \phi(n)\sin(\vartheta_c(n)), \\ \vartheta_c(n+1) &= \vartheta_c(n) + c + \alpha\,\phi(n). \end{aligned}$$

The current version of the test corresponds to the case $\alpha = 0$. As shown in [9], the test with $\alpha = 0$ is less sensitive to measurement noise.

Remark 1.2. It can be rigorously shown that (i) $p_c(n)$ and $q_c(n)$ are bounded if the underlying dynamics is regular, i.e., periodic or quasi-periodic, and (ii) $p_c(n)$ and $q_c(n)$ behave asymptotically like Brownian motion for large classes of chaotic dynamical systems.

In [8] we used results of [6, 19, 20] to prove this for the case $\alpha \neq 0$ in Remark 1.1. In the case $\alpha = 0$ these results are not applicable; nevertheless in [11] we cover the case $\alpha = 0$ under even weaker assumptions on the underlying dynamics.

Remark 1.3. In [8, 11] it was shown that the results are valid for almost all observables ϕ. Of course, the choice of the observable ϕ influences the rate of convergence (but not the limiting value $K = 0$ or $K = 1$). From a practical point of view we found that changing the observable does not greatly alter the computed value of K.

2. Computation of the mean square displacement. For a given time series $\phi(j)$ with $j = 1, \ldots, N$, we compute the mean square displacement of the translation variables $p_c(n)$ and $q_c(n)$ defined in (1.1) for several values of $c \in (0, \pi)$. The mean square displacement is defined as

$$M_c(n) = \lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} [p_c(j+n) - p_c(j)]^2 + [q_c(j+n) - q_c(j)]^2. \tag{2.1}$$

Note that this definition requires $n \ll N$. In [9] we calculated the mean square displacement using the definition (2.1) directly. The limit is assured by calculating $M_c(n)$ only for $n \leq n_{\text{cut}}$, where $n_{\text{cut}} \ll N$. In practice we find that $n_{\text{cut}} = N/10$ yields good results.

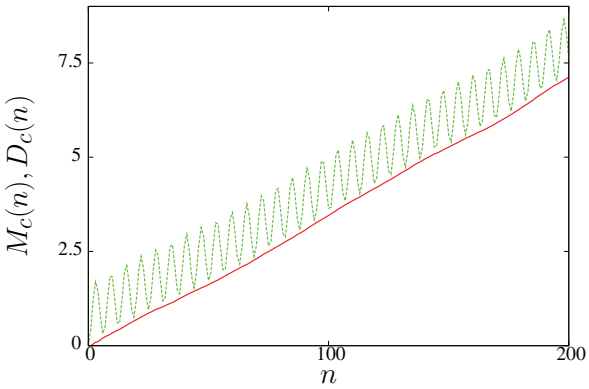


Figure 2. *Plot of mean square displacement versus n for the logistic map with $\mu = 3.91$ corresponding to chaotic dynamics. The oscillating (green) curve is the original mean square displacement $M_c(n)$ as defined in (2.1); the straighter (red) curve is the modified mean square displacement $D_c(n)$ as defined in (2.3). We used 2000 data points and computed $M_c(n)$ and $D_c(n)$ for $n = 1, \ldots, 200$ and $c = 1.0$.*

The test for chaos is based on the growth rate of $M_c(n)$ as a function of n. In the following, we use analytical expressions derived in [11] to formulate a modified mean square displacement $D_c(n)$ which exhibits the same asymptotic growth as $M_c(n)$ but with better convergence properties.

Under mild assumptions on the underlying dynamical system, described in Remark 2.1 below, for each $c \in (0, \pi)$,

$$M_c(n) = V(c)\, n + V_{\text{osc}}(c, n) + e(c, n), \tag{2.2}$$

where $e(c, n)/n \to 0$ as $n \to \infty$ uniformly in $c \in (0, \pi)$ and

$$V_{\text{osc}}(c, n) = (E\phi)^2 \frac{1 - \cos nc}{1 - \cos c}.$$

The expectation $E\phi$ is given by

$$E\phi = \lim_{N \to \infty} \frac{1}{N} \sum_{j=1}^{N} \phi(j).$$

The form (2.2) suggests an improvement for the test: We can subtract the explicit term $V_{\text{osc}}(c, n)$ from the mean square displacement and introduce

$$D_c(n) = M_c(n) - V_{\text{osc}}(c, n). \tag{2.3}$$

Note that the asymptotic growth rates of $M_c(n)$ and $D_c(n)$ are the same.

In Figure 2 we show the two mean square displacements $M_c(n)$ and $D_c(n)$ for the logistic map $x_{n+1} = \mu x_n(1 - x_n)$ with $\mu = 3.91$ (which corresponds to chaotic dynamics) and an arbitrary value of $c = 1.0$. Evidently, the subtraction of the oscillatory term $V_{\text{osc}}(c, n)$ regularizes the linear behavior of $M_c(n)$. This allows a much better determination of the asymptotic growth rate K_c.

Remark 2.1. The autocorrelation function for the observation $\phi(j)$ is given by

$$\rho(k) = E(\phi(1)\phi(k+1)) - (E\phi)^2, \qquad k = 0, 1, 2, \ldots.$$

Provided that the autocorrelations are absolutely summable (that is, $\sum_{k=0}^{\infty} |\rho(k)| < \infty$), then (2.2) is valid, and moreover, the error term $e(c, n)$ decays uniformly in $c \in (0, \pi)$ (see, for example, [11]). It is for this reason that the test based on $D_c(n)$ greatly outperforms the test based on $M_c(n)$.

Furthermore, the absolute summability condition guarantees [11] that

$$V(c) = \sum_{k=-\infty}^{\infty} e^{ikc}\rho(|k|) = \lim_{n\to\infty} \frac{1}{n} E \left| \sum_{j=0}^{n-1} e^{ijc}\phi(j) \right|^2 \tag{2.4}$$

for all $c \in (0, 2\pi)$. This result follows from the Birkhoff ergodic theorem, the Wiener–Khintchine theorem, and standard calculations. In particular, the slope $V(c)$ of the mean square displacement is identified with the power spectrum.

For nonmixing systems, the error term $e(c, n)$ no longer decays to zero, and there are further oscillatory terms in addition to $V_{\text{osc}}(c, n)$. Nevertheless, the identification (2.4) remains valid for nonmixing systems under very weak conditions [18].

More importantly from the point of view of the test for chaos, working with $D_c(n)$ remains highly advantageous even for nonmixing systems. This is illustrated for the logistic map in Figure 4 later in this paper.

3. Computation of K_c.

Having calculated the modified mean square displacement $D_c(n)$ for $n = 1, 2, \ldots, n_{\text{cut}}$, the next step is to estimate the asymptotic growth rate K_c. We have tried out two different methods: a *regression* method and a *correlation* method, described in subsections 3.1 and 3.2 below.

3.1. Regression method.

The regression method consists of linear regression for the log-log plot of the mean square displacement. In [9] we used the original mean square displacement $M_c(n)$, so the asymptotic growth rate K_c is given by the definition

$$K_c = \lim_{n\to\infty} \frac{\log M_c(n)}{\log n}.$$

Numerically, K_c is determined by fitting a straight line to the graph of $\log M_c(n)$ versus $\log n$ through minimizing the absolute deviation [21].

In section 2, we demonstrated the superiority of the modified mean square displacement $D_c(n)$ when compared to $M_c(n)$, so it is natural to apply the regression method to $D_c(n)$. Whereas $M_c(n)$ is strictly positive, $D_c(n)$ may be negative due to the subtraction of the oscillatory term $V_{\text{osc}}(c, n)$. Hence, we set

$$\tilde{D}_c(n) = D_c(n) - \min_{n=1,\ldots,n_{\text{cut}}} D_c(n)$$

and obtain the asymptotic growth rate

$$K_c = \lim_{n\to\infty} \frac{\log \tilde{D}_c(n)}{\log n}.$$

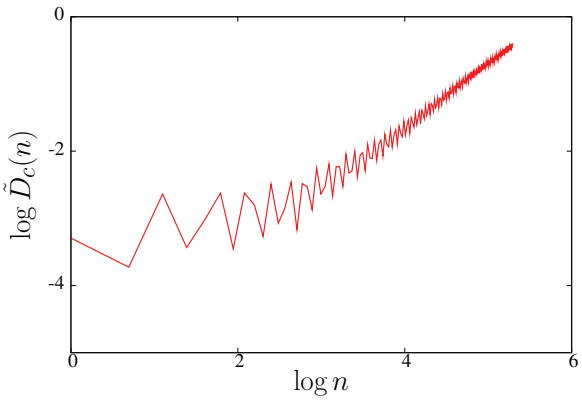


Figure 3. *Plot of* $\log \tilde{D}_c(n)$ *as a function of* $\log n$ *for the logistic map at* $\mu = 3.62$. *We used* $N = 2000$ *and calculated the mean square displacement up to* $n_{\rm cut} = N/10$.

Again, K_c can be determined numerically by regression (minimizing the absolute deviation) for the graph of $\log \tilde{D}_c(n)$ versus $\log n$.

Remark 3.1. Minimizing the absolute deviation is preferable when compared to the usual least square method, as the latter assigns a higher weight to outliers. Since the linear behavior of the mean square displacement is given only asymptotically, one typically encounters outliers for small values of n. We find that it is usually sufficient to use the absolute deviation for estimating K_c, and that it is not necessary to employ more complicated higher-order regression methods such as the method by Yohai [16].

The finite value of the $o(n)$-term $e(c,n)$ in the definition of $M_c(n)$ in (2.2) leads to a distortion for small values of n. In such situations, one typically observes a flattening of the slope of $\log M_c(n)$ or $\log \tilde{D}_c(n)$ as illustrated in Figure 3. It is those values for small n of $\log M_c(n)$ (or $\log \tilde{D}_c(n)$) which would be overestimated in a least squares fit.

3.2. Correlation method. We now present an alternative method for determining K_c from the mean square displacement. (The method is described in terms of $D_c(n)$, but we could use $M_c(n)$ instead.)

Form the vectors $\xi = (1, 2, \ldots, n_{\rm cut})$ and $\Delta = (D_c(1), D_c(2), \ldots, D_c(n_{\rm cut}))$. Given vectors x, y of length q, we define covariance and variance in the usual way:

$$\operatorname{cov}(x,y) = \frac{1}{q}\sum_{j=1}^{q}(x(j)-\bar{x})(y(j)-\bar{y}), \quad \text{where } \bar{x} = \frac{1}{q}\sum_{j=1}^{q} x(j),$$
$$\operatorname{var}(x) = \operatorname{cov}(x,x).$$

Now define the correlation coefficient

$$K_c = \operatorname{corr}(\xi,\Delta) = \frac{\operatorname{cov}(\xi,\Delta)}{\sqrt{\operatorname{var}(\xi)\operatorname{var}(\Delta)}} \in [-1,1].$$

This quantity measures the strength of the correlation of $D_c(n)$ with linear growth. Again, it can be shown rigorously [11] that under weak conditions on the underlying dynamics (as

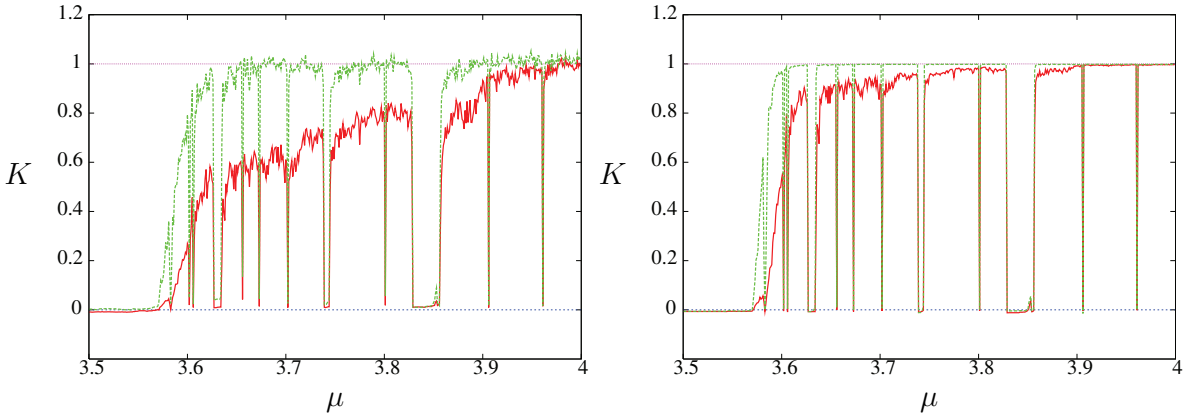


Figure 4. *Plot of K versus μ for the logistic map with $3.5 \le \mu \le 4$ increased in increments of 0.001. We used 2000 data points. The darker (red) lines are obtained by using the original definition of the mean square displacement $M_c(n)$ in (2.1). The lighter (green) lines are obtained by using the modified mean square displacement $D_c(n)$ in (2.3). The resulting values of K are shown for the regression method (left) and the correlation method (right). The horizontal lines (blue and magenta) indicate the cases $K = 0$ and $K = 1$. We used $N_c = 100$ values of c.*

described in Remark 1.2) we obtain $K_c = 0$ for regular dynamics and $K_c = 1$ for chaotic dynamics.

In practical terms, the correlation method greatly outperforms the regression method. This is evident from Figure 4, which compares the regression and correlation methods (using both $M_c(n)$ and $D_c(n)$) for the logistic map.

4. Choice of c and determination of K. In Figure 5 we show the asymptotic growth rate K_c as a function of c for regular and chaotic dynamics. In the case of periodic dynamics, most values of c yield $K_c = 0$ as expected, but there are isolated values of c for which K_c is large. (For the regression method, $K_c \approx 2$ at these resonant points.) These resonances are easily explained as follows: (1.1) shows that if the Fourier decomposition of the observation ϕ contains a term proportional to $\exp(-i\omega k)$, then there is a resonance at $c = \omega$ where $p_c(n) \sim n$, and hence $M_c(n) \sim n^2$, irrespective of whether the dynamics is regular or chaotic. For the plots in Figure 5, we have calculated the asymptotic growth rate using both the regression method described in section 3.1 and the correlation method described in section 3.2.

The occurrence of resonances for isolated values of c suggests using the median of the computed values of K_c. (We use the median rather than the mean, since the median is robust against outliers associated with resonances.)

In Figure 5(right), K_c is shown as a function of c for $\mu = 3.6$ where the dynamics is chaotic but not mixing on the whole interval $[0, 1]$. The actual dynamics in the logistic map oscillates between two disjoint sets, each of which is mixing, and there is a resonance at $c = \pi$. At resonance, $p_c(n) \sim n$ and $M_c(n) \sim n^2$ as before. Close to resonance, the p-q plot eventually behaves like Brownian motion, but in practice one sees only a small part of this motion and so $K_c \approx 0$.

Remark 4.1. Naturally, the choices of c are equally spaced in Figure 5, whereas in applying the test (and throughout the paper with the exception of Figure 5) we choose randomly

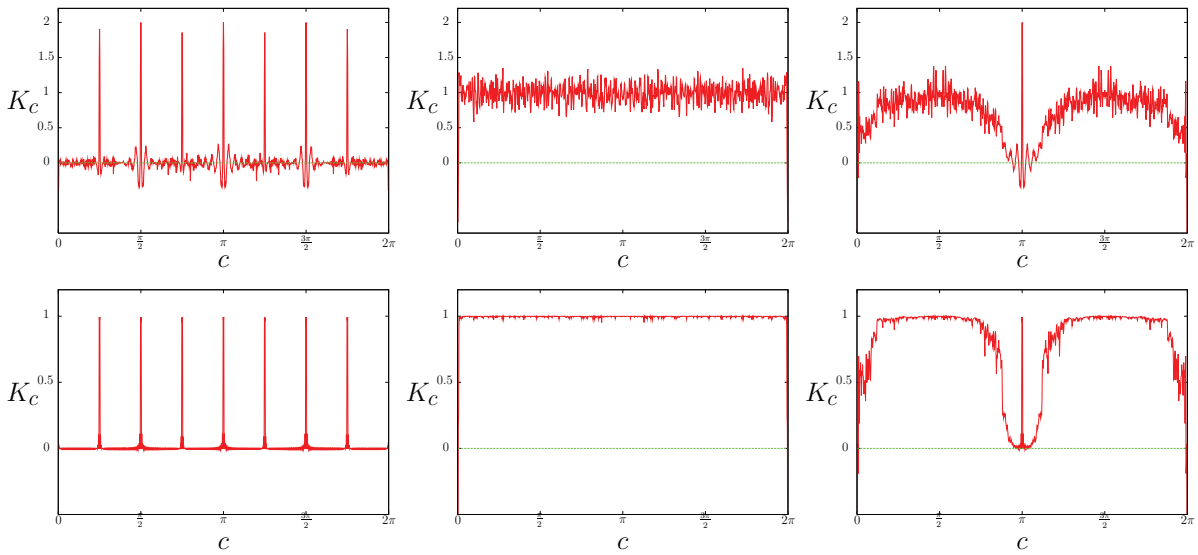


Figure 5. *Plot of K_c versus c for the logistic map calculated using the regression method (top) and correlation method (bottom). We used here $N = 5000$ data points and 1000 equally spaced values for c. Left: $\mu = 3.55$, corresponding to regular dynamics; Middle: $\mu = 3.9$, corresponding to chaotic dynamics; Right: $\mu = 3.6$, corresponding to chaotic but nonmixing dynamics.*

sampled values of c.

To avoid having resonances distort the statistics, we further restrict the range of randomly sampled values for c to $c \in (\pi/5, 4\pi/5)$ for all our computations. The resonance at $c = 0$ is inherent to our test, but it may leak through adjacent values of c, as seen in Figure 5(right). The further restriction to exclude π is not necessary, but we found it helpful. (A typical route to chaos is the Feigenbaum route via period doubling. Here, the parameter ranges for fixed points and period two points are largest.)

In Figure 6, we show how the result for K depends on the number N_c of different values of c. Here we use the correlation version of the test to calculate K_c as described in section 3.2. There is no measurable gain in increasing N_c from 100 to 1000, and we find that generally $N_c = 100$ different values of c are sufficient.

5. Finite size problems. There are three types of finite size effects. First, the time series needs to be long enough to explore and sample the relevant phase space area (i.e., the attractor). This is an inherent problem affecting all tests for chaos. Second, the definition of the mean square displacement involves a limit which requires $n \ll N$. Accordingly, we have chosen $n \leq n_{\text{cut}} = N/10$.

Third, the theory developed in [8, 11] makes statements about the asymptotic behavior of $D_c(n)$ (or $M_c(n)$) and as such requires n_{cut}, and hence N, to be sufficiently large. Finite size effect in this context means that for small n the asymptotic linear growth is not yet dominating; see Figure 3. This finite size effect is explored in the remainder of this section. From now on, we work exclusively with the modified mean square displacement $D_c(n)$ and the correlation method.

In Figure 7 we show how the value of K depends on the amount of data used. We can

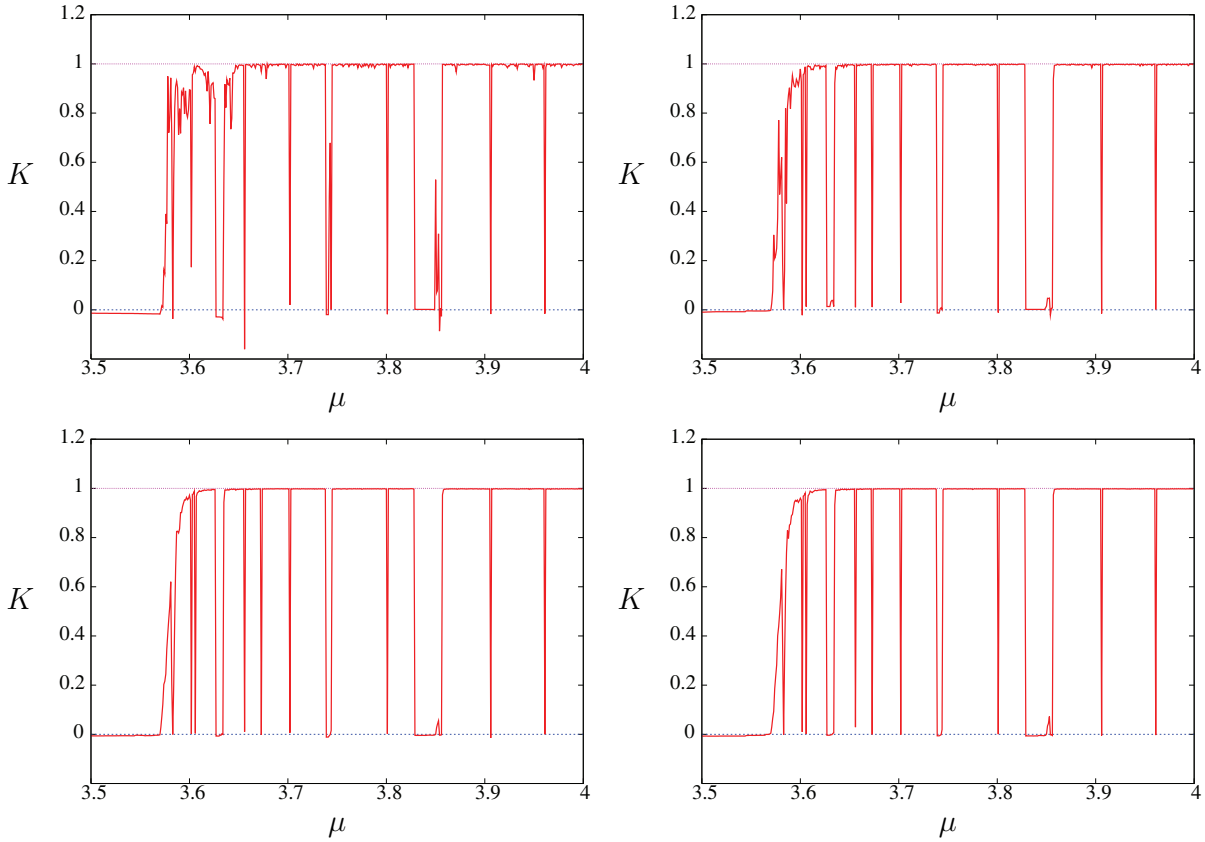


Figure 6. *Plot of K versus μ for the logistic map using the correlation method, with $3.5 \leq \mu \leq 4$ increased in increments of 0.001. We used 2000 data points. Upper left: $N_c = 1$; Upper right: $N_c = 10$; Lower left: $N_c = 100$; Lower right: $N_c = 1000$.*

clearly see the convergence towards the asymptotic values $K = 0$ and $K = 1$ for regular and chaotic underlying dynamics, respectively. (For values of μ corresponding to stronger chaotic dynamics well within the chaotic range, the convergence towards $K = 1$ is even more rapid.)

In the case of "weak chaos," close to the so-called "edge of chaos," longer data sets are required to obtain $K = 1$. Weak chaos is characterized by a slow decay of correlations. This has consequences for the modified mean square displacement $D_c(n) = V(c)n + o(n)$. For systems whose auto-correlation function is slowly decaying, it may be the case that the $o(n)$ term dominates for the available data. We illustrate this problem in the context of the logistic map. The bifurcation parameter μ takes the value $\mu = \mu_\infty = 3.569945672\ldots$ at the edge of chaos, and for $\mu = \mu_\infty + 0.001$ one observes weak chaos.

It has been erroneously claimed that our test cannot detect weak chaos; see [12, 10]. In fact, there are two methods whereby we can distinguish between regular dynamics and weak chaos:

(i) By visual inspection of the plot in the p-q plane generated as in Figure 8. (Note that for longer data sets the dynamics in the p-q plane in Figure 8(right) would look just like Figure 1(right).)

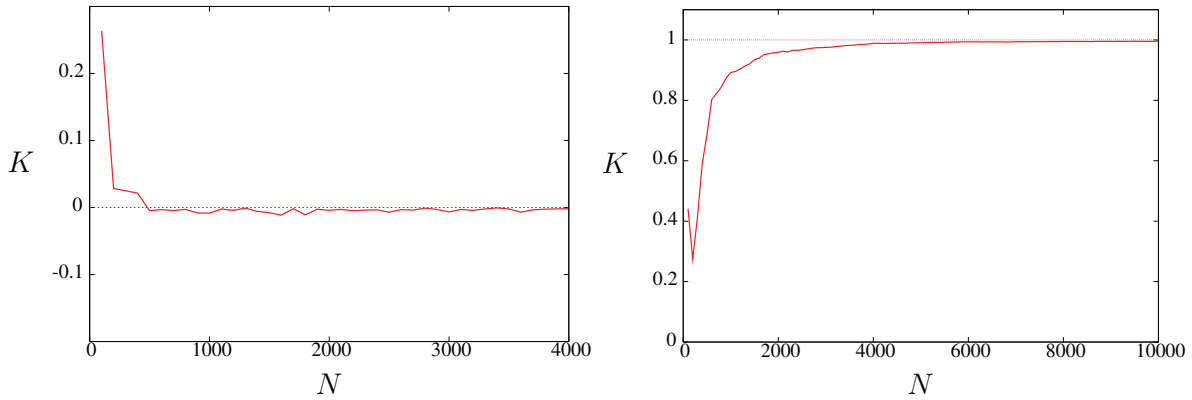


Figure 7. *Plot of K versus the available amount of data N for the logistic map. Left: $\mu = 3.55$, corresponding to regular dynamics; Right: $\mu = 3.6$, corresponding to chaotic dynamics.*

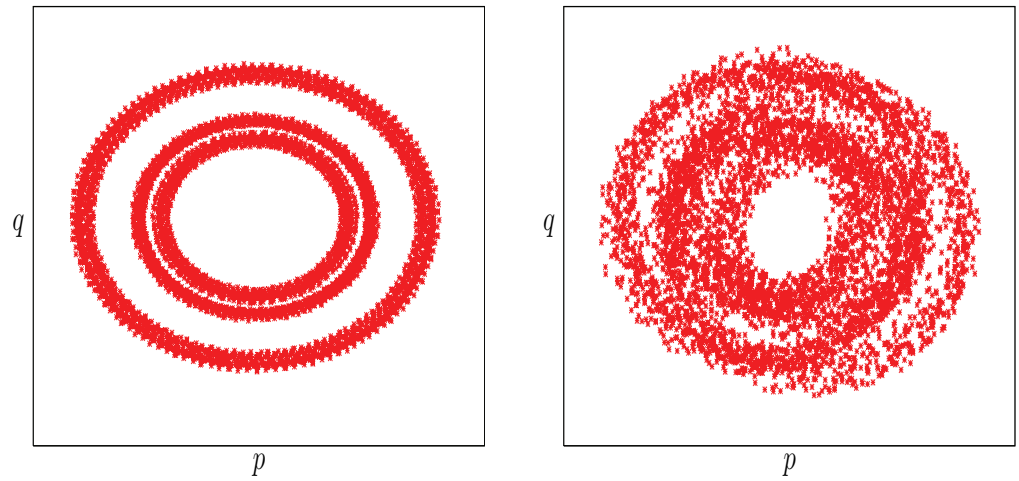


Figure 8. *Plot of p versus q for the logistic map. Left: $\mu = \mu_\infty$; Right: $\mu = \mu_\infty + 0.001$. We used 5000 data points.*

(ii) By looking at the dependence of K as a function of N. As illustrated in Figure 9, we can distinguish weakly chaotic from regular dynamics even when the value of K is very small—note that $K = 0.027$ for $N = 2000$ in the weakly chaotic case.

6. Continuous time systems. In the previous sections, the 0–1 test was formulated for discrete time systems. For continuous time series $\phi(t)$, there is a well-known oversampling issue that must be addressed. In this section, we discuss this difficulty and how to overcome it.

Given $0 < t_1 < t_2 < t_3 < \cdots$, we obtain a discrete time series $\phi(t_1), \phi(t_2), \phi(t_3), \ldots$ to which the test for chaos may be applied as in previous sections. (The sequence t_j, $j \geq 1$, should be chosen in a deterministic manner so that the time series $\phi(t_j)$ is deterministic.) One method of choosing the t_j is as the intersection times with a cross section, so the time series $\phi(t_j)$ corresponds to observing a Poincaré map. In this situation, there are no issues

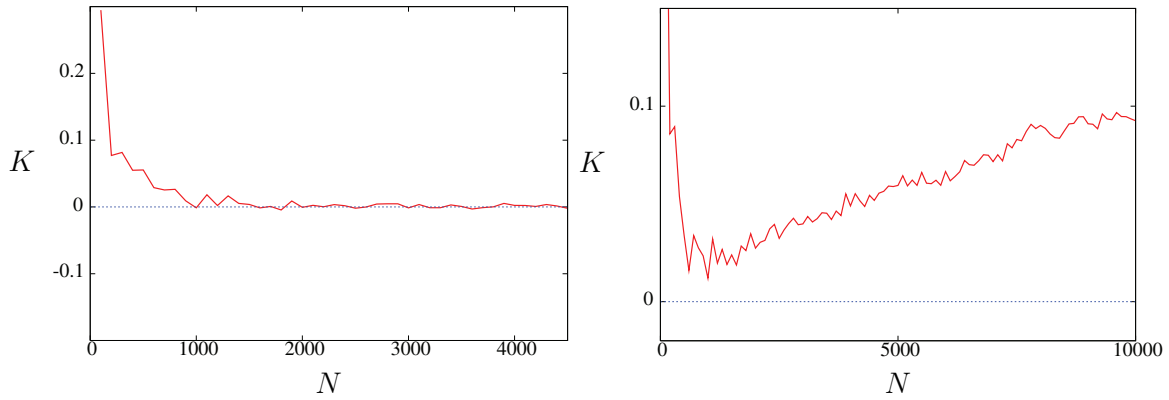


Figure 9. *Plot of K as a function of N for the logistic map at $\mu = \mu_\infty$ (left) and $\mu = \mu_\infty + 0.001$ (right). Although the value of K is small in both cases, the behavior of K as a function of N distinguishes the two cases.*

with oversampling.

A second, perhaps more usual, approach is to take $t_j = j\tau_s$, where $\tau_s > 0$ is the sampling time. The time series $\phi(t_j) = \phi(j\tau_s)$ corresponds to observing the "time-τ_s" map associated with the underlying continuous time system. If τ_s is too small, then the system is *oversampled*, and this often leads to incorrect results. To illustrate the issue of oversampling we study the 3-dimensional Lorenz system

$$\begin{aligned} \dot{x} &= 10(y - x), \\ \dot{y} &= 30\,x - y - xz, \\ \dot{z} &= xy - \tfrac{8}{3}z, \end{aligned} \tag{6.1}$$

which exhibits robust chaos. We have integrated this system with a time step of $\Delta t = 0.001$ and recorded 100,000 data points (i.e., 100 time units).

Figure 10 shows an oversampled and a sufficiently coarsely sampled observable for the Lorenz system (6.1). The finely sampled time series ($\tau_s = 0.005$) yields $K \approx 0$ even for $N = 100{,}000$, whereas the coarsely sampled data ($\tau_s = 0.05$) yields $K \approx 1$ already for $N = 5000$.

A good choice of the sampling time τ_s can often be obtained by visual inspection as in Figure 10. A more refined method is to use the first minimum of the *mutual information* [4, 13]. For the data depicted in Figure 10 this method yields $\tau_s = 0.17$ (roughly a quarter of the oscillation period). Note, however, that in this particular instance the smaller sampling time $\tau_s = 0.05$ already gives $K \approx 1$ and extracts a longer time series from the data in Figure 10. In general, the optimal sampling time will depend on the dynamical system and the time series under consideration. We refer the reader to [13] for a discussion on optimal time delays in the context of phase space reconstruction.

Although oversampling is a practical problem for data series of finite size, it should be emphasized that theoretically the test works for all sampling times τ_s in the limit $N \to \infty$.

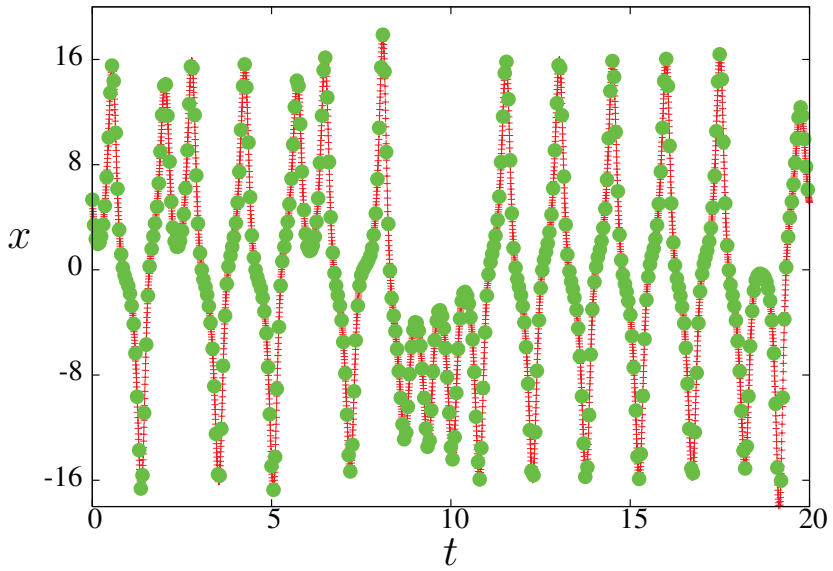


Figure 10. *Plot of the observable $\phi(t) = x(t)$ for the Lorenz system (6.1). The finely sampled data (red plus-signs) are sampled at $\tau_s = 0.005$ time units. The coarsely sampled data (green filled circles) are sampled at $\tau_s = 0.05$ time units.*

6.1. Oversampling and power spectra. For continuous time systems, the mean square displacement is defined as

$$M_c(t) = \lim_{T\to\infty} \frac{1}{T} \int_0^T (p_c(t+\tau) - p_c(\tau))^2 \; + \; (q_c(t+\tau) - q_c(\tau))^2 \, d\tau.$$

For a time series sampled with sample time τ_s this can be approximated by

$$M_c(n) = \lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} \left([p_{c\tau_s}(j+n) - p_{c\tau_s}(j)]^2 \; + \; [q_{c\tau_s}(j+n) - q_{c\tau_s}(j)]^2 \right) \tau_s^2.$$

Similarly the power spectrum for the continuous time case discretizes to

$$S(\nu) = \lim_{n\to\infty} \frac{1}{n} E \left| \sum_{j=0}^{n-1} e^{2\pi i \frac{\nu}{\nu_s} j} \phi(j) \right|^2 \tau_s^2, \tag{6.2}$$

where $\nu_s = 1/\tau_s$ is the sample frequency. The power spectrum consists of discrete peaks if the underlying system is regular, and it is nowhere zero for a large class of chaotic systems [18]. However, for chaotic systems the power spectrum decays for large frequencies ν, and so for frequencies larger than some $\nu_{\max}$ the power spectrum is zero for all practical purposes.

Comparing (6.2) with the power spectrum (2.4) for discrete-time data, we identify

$$c = 2\pi \frac{\nu}{\nu_s}, \quad \nu \in [0, \nu_{\max}] \, .$$

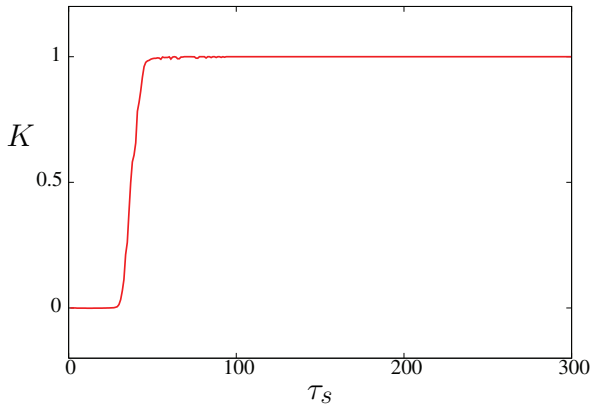


Figure 11. *Plot of K as a function of the sample time τ_s for the Lorenz system (6.1). The sample time is measured in units of $\Delta t = 0.001$.*

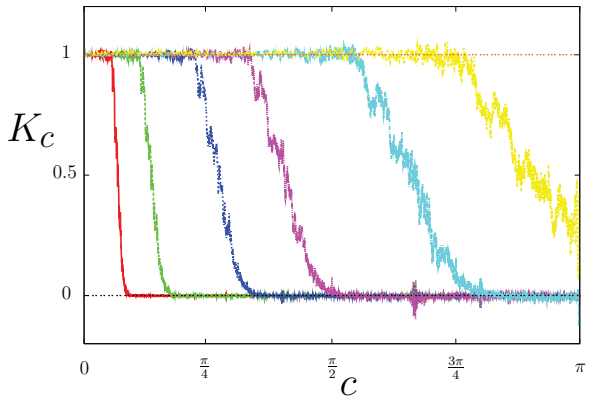


Figure 12. *Plot of K_c as a function of the frequency c for the Lorenz system (6.1). From left to right we used $\tau_s = 5\Delta t$, $\tau_s = 10\Delta t$, $\tau_s = 20\Delta t$, $\tau_s = 30\Delta t$, $\tau_s = 50\Delta t$, $\tau_s = 70\Delta t$. The linear scaling of the range of c for which $K_c \approx 1$ is evident in the relative spacing of the respective lines.*

Sampling at the Nyquist rate with $\nu_s^\star = 2\nu_{\max}$ yields $c \in (0, \pi)$ as before. However, oversampling at a higher frequency $\nu_s > \nu_s^\star$ restricts the effective choices of c to $c \in (0, c^\star)$, where $c^\star = \frac{\nu_s^\star}{\nu_s}\pi < \pi$. There is now a positive probability that the test for chaos will incorrectly yield $K = 0$ since it is possible that more than half of the randomly chosen values of $c \in (0, \pi)$ will lie in $(c^\star, \pi)$.

We illustrate the previous argument using the Lorenz system (6.1) sampled with τ_s ranging from $\tau_s = \Delta t$ up to $\tau_s = 300\Delta t$. In Figure 11 the median of the asymptotic growth rate K is shown as a function of the sample time. For data that is too finely sampled, we obtain $K = 0$, although the dynamics is actually chaotic. Figure 12 illustrates how the range of effective values of c depends on the sampling time τ_s.

7. Noise contaminated data. Real-world data is invariably contaminated with noise. Any method for distinguishing regular from chaotic dynamics can succeed only if the noise-level is sufficiently small. There are various standard noise reduction techniques [13] that may

be applied in advance of applying any given test for chaos. In addition, the test itself may be modified. Below we indicate a modification of the 0–1 test for chaos that makes it more robust to the presence of noise.

In [9] we introduced a version of the 0–1 test that works well for data contaminated with measurement noise. (This is the test as presented in sections 2 and 3, but using $M_c(n)$ instead of $D_c(n)$ and using the regression method instead of the correlation method.) We showed that the 0–1 test clearly outperforms tangent space methods and compares favorably to "direct methods" based on phase space reconstruction. The improvements in this paper have made our test extremely sensitive to weak chaos. However, an unavoidable consequence is an increased sensitivity also to noise (see Figure 13 below).

It turns out that the success of the version of the test in [9] is due to the oscillatory term $V_{\rm osc}(c,n) = (E\phi)^2 \frac{1-\cos nc}{1-\cos c}$ that we subtracted in section 2 to define the modified mean square displacement $D_c(n) = M_c(n) - V_{\rm osc}(c,n)$. This term desensitizes the test and damps the ability to detect slow growth of the mean square displacement for time series data of moderate length. Instead of reintroducing this term we adopt a more flexible approach, defining

$$D_c^\star(n) = D_c(n) + \alpha V_{\rm damp}(n), \quad V_{\rm damp}(n) = (E\phi)^2 \sin(\sqrt{2}n).$$

(The frequency $\sqrt{2}$ was chosen arbitrarily.) For α large, we expect $K = 0$. The amplitude α of the term $V_{\rm damp}(n)$ controls the sensitivity of the test to weak noise and simultaneously to weak chaos. This trade-off is unavoidable in any test for chaos.

As an illustration, we consider the logistic map with measurement noise. Take as observable $\phi(n) = x_n$ and write

$$\tilde{\phi}(n) = \phi(n)\left(1 + \frac{\epsilon}{100}\eta_n\right),$$

where η_n are independently and identically distributed (i.i.d.) random variables drawn from a uniform distribution on $[-1,1]$ and ϵ is the noise-level in percent. Figure 13 shows how the undamped version of the test in this paper copes with a noise level of 10% and the improvement that is obtained by using the damped mean square displacement $D_c^*(n)$. We obtain similar results for normally distributed noise.

Remark 7.1. Under the assumption that the noise is diffusive and not correlated with the dynamics, the mean square displacement for data contaminated with measurement noise may be written as

$$D_c(n) = (V_{\rm dyn}(c) + V_{\rm noise}(c))n + o(n),$$

where for a given value of c, $V_{\rm dyn}(c)$ is the variance associated with the deterministic dynamics and $V_{\rm noise}(c)$ the variance associated with the measurement noise. Consider an idealized situation where the value of $V_{\rm noise}(c)$ is roughly constant as a parameter λ is varied. Suppose further that the dynamics is known to be regular at $\lambda = \lambda_0$. Then we may estimate $V_{\rm noise}(c)$ by making a gauge-measurement at $\lambda = \lambda_0$, applying the correlation method to $D_c(n) - Vn$. The unique value $V = V_c(\lambda_0)$ which yields $K_c = 0$ is our estimate for $V_{\rm noise}(c)$. For other values of λ we may now apply the correlation method to $D_c(n) - V_c(\lambda_0)n$.

8. Discussion. We have presented a guide for the implementation of the 0–1 test for chaos. At the same time, we have introduced an improved version of the test which uses analytical expressions derived in [11]. Issues such as oversampling for continuous time data and the

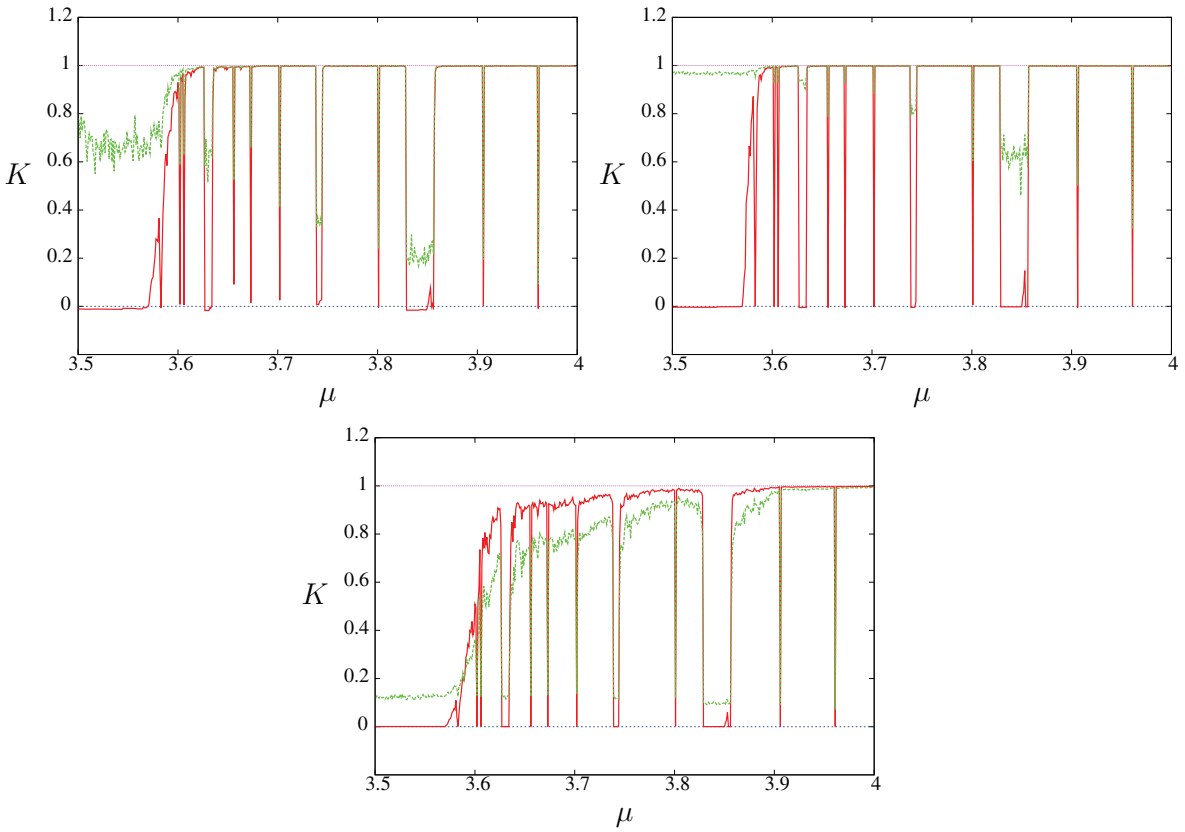


Figure 13. *Plot of K versus μ for the logistic map increased in increments of* 0.001. *The darker (red) lines were computed using clean data. The lighter (green) lines were computed after addition of* 10% *uniformly distributed measurement noise. Both lines were computed using the undamped mean square displacement $D_c(n)$. Left: $N = 1000$ using the undamped mean square displacement $D_c(n)$, Right: $N = 5000$ using $D_c(n)$. Bottom: $N = 5000$ using the damped mean square displacement $D_c^*(n)$ with $\alpha = 2.5$.*

presence of noise have been discussed. We hope that this guide will be helpful for scientists who would like to use the test.

There are numerous methods in the literature for distinguishing between deterministic and chaotic dynamics. In our previous papers [8, 9], we made a careful comparison of the 0–1 test with methods for computing the maximal Lyapunov exponent. Another method is to use the power spectrum, for which there are efficient computational techniques. It should be pointed out, however, that these techniques generally rely on the Wiener–Khintchine theorem, which assumes summable decay of correlations and hence excludes periodic and quasi-periodic dynamics. Hence to use power spectra as a test for chaos, it seems necessary to avoid the Wiener–Khintchine theorem and to work directly with the expression $\lim_{n\to\infty} \frac{1}{n} E|\sum_{j=0}^{n-1} e^{ijc}\phi(j)|^2$. Equation (2.4) shows the relationship between the 0–1 test and power spectra, and our test can be viewed as a way of condensing the information relevant for chaoticity or regularity contained in the power spectrum into a single binary number.

Acknowledgments. We would like to thank Ramon Xulvi-Brunet for pointing us towards the correlation method and for explaining to us the problem of oversampling in power spectra. We would like to thank Michael Breakspear and John Dawes for providing encouraging feedback on an earlier version of the manuscript. IM is grateful to the University of Sydney for its hospitality.

REFERENCES

[1] R. Barrio, *Sensitivity tools vs. Poincaré sections*, Chaos Solitons Fractals, 25 (2005), pp. 711–726.
[2] P. M. Cincotta, C. M. Giordano, and C. Simó, *Phase space structure of multi-dimensional systems by means of the mean exponential growth factor of nearby orbits*, Phys. D, 182 (2003), pp. 151–178.
[3] J. H. P. Dawes and M. C. Freeland, *The '0–1 Test for Chaos' and Strange Nonchaotic Attractors*, preprint, DAMTP, University of Cambridge, Cambridge, UK.
[4] A. M. Fraser and H. L. Swinney, *Independent coordinates for strange attractors from mutual information*, Phys. Rev. A, 33 (1986), pp. 1134–1140.
[5] I. Falconer, G. A. Gottwald, I. Melbourne, and K. Wormnes, *Application of the 0–1 test for chaos to experimental data*, SIAM J. Appl. Dyn. Syst., 6 (2007), pp. 395–402.
[6] M. Field, I. Melbourne, and A. Török, *Decay of correlations, central limit theorems and approximation by Brownian motion for compact Lie group extensions*, Ergodic Theory Dynam. Systems, 23 (2003), pp. 87–110.
[7] M. Fouchard, E. Lega, C. Froeschlé, and C. Froeschlé, *On the relationship between fast Lyapunov indicator and periodic orbits for continuous flows*, Celestial Mech. Dynam. Astronom., 83 (2002), pp. 205–222.
[8] G. A. Gottwald and I. Melbourne, *A new test for chaos in deterministic systems*, Proc. Roy. Soc. A, 460 (2004), pp. 603–611.
[9] G. A. Gottwald and I. Melbourne, *Testing for chaos in deterministic systems with noise*, Phys. D, 212 (2005), pp. 100–110.
[10] G. A. Gottwald and I. Melbourne, *Comment on "Reliability of the 0–1 test for chaos,"* Phys. Rev. E, 77 (2008), paper 028201.
[11] G. A. Gottwald and I. Melbourne, *Validity of the 0–1 test for chaos*, in preparation.
[12] J. Hu, W.-W. Tung, J. Gao, and Y. Cao, *Reliability of the 0–1 test for chaos*, Phys. Rev. E, 72 (2005), paper 056207.
[13] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis*, Cambridge University Press, Cambridge, UK, 1997.
[14] J. Laskar, *Frequency analysis for multi-dimensional systems. Global dynamics and diffusion*, Phys. D, 67 (1993), pp. 257–281.
[15] G. Litak, A. Syta, and M. Wiercigrochc, *Identification of chaos in a cutting process by the 0–1 test*, Chaos Solitons Fractals, to appear.
[16] A. Marazzi, J. Joss, and A. Randriamiharisoa, *Algorithms, Routines, and S-Functions for Robust Statistics: The Fortran Library Robeth with an Interface to S-Plus*, Chapman & Hall, London, 1993.
[17] N. Martinsen-Burrell, K. Julien, M. R. Petersen, and J. B. Weiss, *Merger and alignment in a reduced model for three-dimensional quasigeostrophic ellipsoidal vortices*, Phys. Fluids, 18 (2006), paper 057101.
[18] I. Melbourne and G. A. Gottwald, *Power spectra for deterministic chaotic dynamical systems*, Nonlinearity, 21 (2008), pp. 179–189.
[19] I. Melbourne and M. Nicol, *Statistical properties of endomorphisms and compact group extensions*, J. London Math. Soc., 70 (2004), pp. 427–446.
[20] M. Nicol, I. Melbourne, and P. Ashwin, *Euclidean extensions of dynamical systems*, Nonlinearity, 14 (2001), pp. 275–300.
[21] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C*, Cambridge University Press, Cambridge, UK, 1992.
[22] M. T. Rosenstein, J. J. Collins, and C. J. De Luca, *A practical method for calculating largest*

Lyapunov exponents from small data sets, Phys. D, 65 (1993), pp. 117–134.

[23] M. Sano and Y. Sawada, *Measurement of the Lyapunov spectrum from a chaotic time series*, Phys. Rev. Lett., 55 (1985), pp. 1082–1085.

[24] C. Skokos, *Alignment indices: A new, simple method for determining the ordered or chaotic nature of orbits*, J. Phys. A, 34 (2001), pp. 10029–10043.

[25] C. Skokos, Ch. Antonopoulos, T. C. Bountis, and M. N. Vrahatis, *Detecting order and chaos in Hamiltonian systems by the SALI method*, J. Phys. A, 37 (2004), pp. 6269–6284.

[26] F. Takens, *Detecting strange attractors in turbulence*, in Lecture Notes in Math. 98, Springer, Berlin, 1981, pp. 366–381.

SIAM J. Applied Dynamical Systems
Vol. 8, No. 1, pp. 146–159
© 2009 Society for Industrial and Applied Mathematics

Stability of TCP Dynamics in Large Data Networks*

Russell D. Halper†, Eric J. Harder‡, Brian R. Hunt†, and James A. Yorke†

Abstract. We study the stability of the dynamics of a deterministic model of Transmission Control Protocol (TCP) traffic in a multiple flow, multiple router data network. In this model, flow rates increase continuously until network congestion causes them to decrease discontinuously. In computer simulations of small networks, trajectories appear to approach periodic orbits for most but not all parameter values. We prove that if all the data flows are coupled, periodic orbits must be exponentially attracting and thus persist under parameter changes, regardless of network size. Furthermore, we describe the model as a discontinuous but piecewise affine map and show that trajectories must either approach a periodic orbit or come arbitrarily close to map discontinuities.

Key words. TCP, networks, piecewise linear dynamics

AMS subject classification. 37N99

DOI. 10.1137/080715354

1. Introduction.

The Transmission Control Protocol (TCP) is used to provide reliable and efficient end-to-end communication over the Internet. Most lossless Internet traffic, including electronic mail, file transfer, and web browsing, uses TCP. For each pair of computers that use this protocol to communicate, TCP dynamically governs the rate at which data is sent between them, using what is called an additive increase multiplicative decrease (AIMD) mechanism. Roughly speaking, the computer that is sending data linearly increases its transmission rate over time until it detects congestion in the path of the transmission through the network, at which point the computer decreases its transmission rate by a factor of two [J88]. Nonlinearities due to feedback delays and the methods used by routers to manage congestion can yield complicated dynamics even in small networks (see, e.g., [RAL02]). Nonetheless, we find in a model of TCP dynamics that, in the absence of stochastic inputs, flow rates typically exhibit stable periodic behavior, no matter how complex the network of data flows is.

Data transmitted through the network is sent in small pieces (usually around 1500 bytes) called packets. When a packet is sent, it is forwarded through the network by routers that exist at the nodes of the network. When data arrives at a router faster than the router is able to forward it, congestion occurs. Typically a router has a buffer to store a small amount of incoming data during congestion. However, once the buffer is full, all incoming packets are dropped from the network until the router is able to transmit the contents of its buffer. TCP detects congestion by noticing when packets are dropped from its flow and need to be resent.

*Received by the editors February 8, 2008; accepted for publication (in revised form) by A. Hagberg September 15, 2008; published electronically January 23, 2009. This research was supported by the Air Force Office of Scientific Research (Mathematics, Information and Life Sciences Directorate).
http://www.siam.org/journals/siads/8-1/71535.html
†Department of Mathematics, University of Maryland, College Park, MD 20742 (russ@math.umd.edu, bhunt@glue.umd.edu, yorke@umd.edu).
‡9800 Savage Road, Suite 6534, Fort Meade, MD 20755 (ejharde@tycho.ncsc.mil).

The dynamics of a TCP network are affected by mechanisms that routers employ to mitigate the effects of congestion. The widely used Random Early Detection (RED) algorithm [FJ94] probabilistically drops packets before the input buffer fills to attempt to maintain a high throughput. A proposed alternative to RED is Explicit Congestion Notification (ECN) [F94] in which a router can deterministically request all flows through it to reduce their flow rates before its buffer fills. With this in mind, we study the dynamics of a deterministic model of the AIMD mechanism.

The AIMD-based model that is the topic of this paper was proposed in [BH03] as the asymptotic limit of a stochastic model of TCP as the number of end users and amount of resources available in the network linearly increase to infinity. We consider this deterministic model also to be an idealization of TCP network dynamics with ECN, for which no large-network approximation is needed. The model keeps track of a flow rate for each of N TCP flows. Each flow travels through a subset of the routers in the network, and each router has a predetermined capacity. When the total flow through a router, defined by the sum of the rates of the TCP flows going through that router, reaches its capacity, each of these flow rates is cut by a factor of $\alpha \in (0, 1)$. In the case of a network with a single router, this model is similar to the deterministic models in [SL06] and [SWL06], for which all trajectories approach a unique fixed point. With more than one router, the model we study can yield periodic dynamics, as observed in [BH03], or quasi-periodic dynamics. Related stochastic models for the single router case appear in [BH02] and [SWL06].

In section 2 we present the model, representing the dynamics as a mapping from a manifold to itself. Specifically, the state is in a piecewise linear surface of dimension $N - 1$, the surface of *congestion states*, where at least one router is experiencing congestion. Here the model is a piecewise-affine map that takes one congestion state to the next. In section 3 we consider the question of what assumptions are needed to prove stable periodicity for the model. Numerical results suggest that for small networks, at least, most but not all parameters and initial states yield stable periodic dynamics. Baccelli and Hong [BH03] proved that periodicity implies asymptotically stable periodicity under the assumptions that the network is connected and that each router has a flow that goes through that router only. We are able to prove exponentially stable periodicity, and hence persistence under parameter changes, under much more general hypotheses. In particular, we formulate a coupling condition that is both necessary and sufficient for a periodic orbit to be exponentially attracting. (An interesting feature of this model is that the coupling can depend on a particular trajectory and not just on the network.) We also introduce a geometric description of the local dynamics that provides an intuitive understanding of how small perturbations are transformed by congestion events. We use this approach to prove that every trajectory that satisfies our coupling condition either approaches a periodic orbit or comes arbitrarily close to a degenerate state in which two or more routers experience congestion simultaneously. We conjecture that the latter type of dynamics occurs for an exceptional set of parameters with measure zero. Additionally, we show that when our coupling condition does not hold, the system can be broken down into maximal coupled subsystems, on each of which the dynamics evolve independently and where the results presented in this paper are valid.

2. Introduction to the model. We describe the dynamics of a network consisting of N data flows. Each flow may correspond to a single sender-receiver pair or a collection of

such pairs that follow a single path through the network. Let the flow rate of flow s at time t be $x^s(t)$, where $s \in S = \{1, 2, \ldots, N\}$. Then the state of the system at time t is $\boldsymbol{x}(t) = (x^1(t), x^2(t), \ldots, x^N(t))$. Let the network have M routers, and let each flow s travel through a set of routers $R_s \subset R = \{1, 2, \ldots, M\}$. Let $S_r = \{s \in S \mid r \in R_s\}$ be the set of flows that travel through router r. The amount of data being transmitted through router r at any given time t is $X_r(t) = \sum_{s \in S_r} x^s(t)$. Let each router have a capacity c_r, which is the maximum allowed value for $X_r(t)$ ($X_r(t) \leq c_r$ for all t). We say the router reaches *congestion* when $X_r(T) = c_r$. We refer to such a time T as a *congestion event*.

In the model, each flow s travels through at least one router. Its flow rate $x^s(t)$ increases at a constant rate, denoted by $\dot{x}^s > 0$, except when one of the routers that flow s passes through reaches congestion. Let $\dot{\boldsymbol{x}} = (\dot{x}^1, \dot{x}^2, \ldots, \dot{x}^N)$. When a router r reaches congestion, all flows that travel through router r are signaled, and each instantaneously and discontinuously decreases its flow rate by some factor α, where $\alpha \in (0, 1)$ and is the same for all routers and flows. In an actual TCP network with RED, during a congestion event a router will drop packets only from some (randomly chosen) sender-receiver pairs. The senders that have a packet dropped reduce their transmission rate by a factor of $\frac{1}{2}$, while the other senders do not. In this case, α is considered in [BH03] to represent the average proportional drop in the transmission rates of all sender-receiver pairs in a flow after a congestion event. In a network with ECN, routers instruct all senders to reduce their transmission rates after a congestion event, causing the network to behave more deterministically.

We make the convention that if $x^s(t)$ has a discontinuity at time T, then $x^s(t)$ is continuous from the left and discontinuous from the right; that is, $\lim_{t \nearrow T} x^s(t) = x^s(T)$ and $\lim_{t \searrow T} x^s(t) = \alpha x^s(T)$. As we will see, for almost every initial state there is an infinite sequence of congestion events T_n (where $T_n < T_{n+1}$), and T_n converges to infinity as n goes to infinity.

Discrete time dynamics. We define

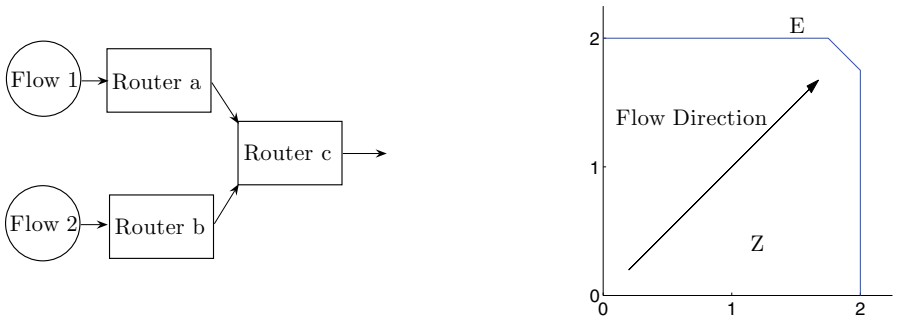
$$Z = \{\boldsymbol{x} = (x^1, \ldots, x^N) : \text{no router exceeds its capacity and all } x^s \geq 0\}$$

and

$$E = \{\boldsymbol{x} \in Z : \text{at least one router is at its capacity}\}.$$

Let $E_1 \subset E$ be the set of states where exactly one router is at its capacity, i.e., experiencing congestion. Furthermore, define E_1^r to be the subset of E_1 where router r is the router experiencing congestion. Let $E_2 = E \backslash E_1$. For a state $\boldsymbol{x} \in Z$, there is a unique point $\texttt{Up}(\boldsymbol{x}) \in E$ that is the state at which the next congestion event occurs. Define $\tau(\boldsymbol{x}) \geq 0$ to be the time until the next congestion event. That is, $\tau(\boldsymbol{x})$ is the unique number for which $\boldsymbol{x} + \tau(\boldsymbol{x})\dot{\boldsymbol{x}} \in E$ and $\texttt{Up}(\boldsymbol{x}) = \boldsymbol{x} + \tau(\boldsymbol{x})\dot{\boldsymbol{x}}$. Note that $\tau(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in E$ and that $\tau(\boldsymbol{x})$ and $\texttt{Up}(\boldsymbol{x})$ are locally affine functions for all $\boldsymbol{x}$ such that $\boldsymbol{x} + \tau(\boldsymbol{x})\dot{\boldsymbol{x}} \in E_1$.

When $\texttt{Up}(\boldsymbol{x})$ causes congestion at precisely one router r, we write $r = \texttt{Router}(\boldsymbol{x})$ to denote the router experiencing congestion. Thus, if $\boldsymbol{x}$ is a state at which a unique router will be the next to experience congestion, then $\texttt{Router}(\boldsymbol{x})$ denotes that router. Notice that $r = \texttt{Router}(\boldsymbol{x})$ if and only if $\texttt{Up}(\boldsymbol{x}) \in E_1^r$. For each $\boldsymbol{x} \in E_1^r$, define the state $\texttt{Drop}_r(\boldsymbol{x})$ coordinatewise by $\texttt{Drop}_r(\boldsymbol{x})^s = \alpha x^s$ for $s \in S_r$ and $\texttt{Drop}_r(\boldsymbol{x})^s = x^s$ otherwise. Then define



(a) (b)

Figure 1. *A network with two flows,* 1 *and* 2*, and three routers, a, b, and c.* (a) *is the router configuration, and* (b) *shows the geometric representation of the model displaying the router constraints, E and Z. In this network, flow* 1 *travels through routers a and c, and flow* 2 *travels through routers b and c. We assume that both flows increase their transmissions at constant rate* 1*. Furthermore, we assume that routers a and b both have capacity* 2 *and router c has capacity* 3.75*. This configuration sends both flows* 1 *and* 2 *through separate routers a and b before both flows jointly traverse router c. This configuration can be written formally as* $S = \{1, 2\}$*,* $R = \{a, b, c\}$*,* $S_a = \{1\}$*,* $S_b = \{2\}$*,* $S_c = \{1, 2\}$*,* $\dot{\boldsymbol{x}} = (1, 1)$*,* $c_a = c_b = 2$*, and* $c_c = 3.75$*. Here,* $E_2 = \{(2, 1.75), (1.75, 2)\}$*.*

$\mathtt{Drop}(\boldsymbol{x}) = \mathtt{Drop}_{\mathtt{Router}(\boldsymbol{x})}(\boldsymbol{x})$ and

$$\mathtt{Next}(\boldsymbol{x}) = \mathtt{Up} \circ \mathtt{Drop}(\boldsymbol{x}).$$

$\mathtt{Next}$ is a piecewise affine mapping from E_1 to E. Given an initial state $\boldsymbol{x}_0 \in E_1$, define $\boldsymbol{x}_n = \mathtt{Next}(\boldsymbol{x}_{n-1})$ for each $n \in \mathbb{N}$, as long as $\boldsymbol{x}_{n-1} \in E_1$. If $\boldsymbol{x}_n \in E_1$ for all $n \in \mathbb{N}$, we call $(\boldsymbol{x}_n)_{n=0}^{\infty}$ the *trajectory* generated by $\boldsymbol{x}_0$. We denote the composition of the $\mathtt{Next}(\boldsymbol{x})$ map with itself k times as $\mathtt{Next}(\boldsymbol{x}, k)$. Thus $\boldsymbol{x}_n = \mathtt{Next}(\boldsymbol{x}_0, n)$. Finally, let $r_n = \mathtt{Router}(\boldsymbol{x}_n)$.

Geometric representation of the model. The piecewise linear path $\boldsymbol{x}(t)$ is confined to the region Z, which is defined by the router configuration. Specifically, each router $r \in R$ defines the *router constraint*, $\sum_{s \in S_r} x^s \leq c_r$. Z is then precisely the subset of $\mathbb{R}^N$ that satisfies the set of router constraints and the additional *nonnegativity constraints*, $x^s \geq 0$ for each $s \in S$. E may be defined as the subset of Z where at least one router constraint holds at equality, and E_1 is then the subset of E where *exactly* one router constraint holds at equality. The subset of E_1 where the router constraint defined by router r holds at equality is the *facet* E_1^r. The model dictates that the path $\boldsymbol{x}(t)$ continues in the direction $\dot{\boldsymbol{x}}$ at constant speed until the path strikes one of the facets in E_1, at which point $\mathtt{Drop}(\boldsymbol{x})$ is applied and the path has a discontinuity.

Consider the network in Figure 1. It is important to note that in this example the capacity of router c is less than the sum of the capacities of routers a and b. Notice that if $c_c > 4$, router c would never experience congestion. Assume that $\alpha = \frac{1}{2}$. Depending on the initial state, two different asymptotic behaviors of the model exist. We classify the asymptotic dynamics of the system by considering the points in E_1. Since all infinite trajectories intersect E_1, this comprehensively classifies the dynamics of the system.

For those initial states $\boldsymbol{x}_0 = (x, y)$ that exist on the slanted open line segment L_1 between $(1.75, 2)$ and $(2, 1.75)$, we may consider what happens to the line segment after each iteration of the map. The map `Drop` contracts L_1 toward the origin by a factor of $\frac{1}{2}$. `Up` then translates this image back up on top of L_1. Thus `Next` is mapping L_1 into itself, contracting the image of L_1 by $\frac{1}{2}$ in the process. The point $(\frac{3.75}{2}, \frac{3.75}{2})$ is fixed by `Next` in this system, producing a periodic orbit with period 1. Furthermore, every initial state in L_1 will never cause congestion at any router other than c. Thus the sequence $(\boldsymbol{x}_n)$ asymptotically approaches the periodic orbit generated by $(\frac{3.75}{2}, \frac{3.75}{2})$ for every initial state in L_1.

For initial states of the form $(x, 2)$ on the open line segment between $(1.25, 2)$ and $(1.75, 2)$, L_2, $\texttt{Next}((x, 2)) = (2, 3-x)$, and thus $\texttt{Next}((x, 2), 2) = (x, 2)$. Consequently, each initial state on this open line segment, and the open line segment between $(2, 1.75)$ and $(2, 1.25)$, produces a periodic orbit. The remaining initial states in E_1 lie either in L_1 or in L_2 after a finite number of iterations and thus follow one of the two described asymptotic behaviors (or map into the points $(1.75, 2)$ or $(2, 1.75)$ and are currently not under consideration).

Coupling. Notice in the previous example that while all initial states in L_1 asymptotically approach a single stable periodic orbit in L_1, each initial state in L_2 lies on a distinct periodic orbit. We will show that the key difference that causes an initial state in L_1 to produce the same asymptotic behavior as nearby initial states is that the trajectory generated by an initial state in L_1 is coupled in the sense described below, while the trajectory generated by an initial state in L_2 is not.

Define a set R' of routers to be *coupled* if for each proper nonempty subset $S' \subset S$ there is a router $r \in R'$, a flow $s' \in S'$, and a flow $s \in S \backslash S'$ such that $s, s' \in S_r$. Define a router r to be a *bottleneck router* of a trajectory $(\boldsymbol{x}_n)$ if router r appears infinitely many times in the router sequence (r_n). Define a trajectory to be *coupled* if the set of bottleneck routers of the trajectory is coupled.

In the example above, a trajectory that starts in L_1 has bottleneck router c and therefore is coupled because both flows are in S_c. The bottleneck routers for a trajectory that starts in L_2 are a and b, which do not form a coupled set because neither S_a nor S_b contains both flows. When a trajectory is not coupled, it is possible to decompose its flows and bottleneck routers into two or more coupled subsystems, where the trajectory evolves independently on each coupled subsystem. Such is the case in the previous example for the dynamics on L_2. In this case the set of flows $\{1, 2\}$ and the uncoupled set $\{a, b\}$ of bottleneck routers can be broken down into coupled subsystems $\{a, 1\}$ and $\{b, 2\}$ where the dynamics on each evolve independently. The different initial states in L_2 do not asymptotically approach the same periodic orbit because each initial state induces a different phase shift for the two coupled subsystems.

While coupling is a hypothesis of our main results (Theorems 3.1 and 3.6), in order to prove directly that a trajectory is exponentially attracting (Proposition 3.4), we use a slightly stronger hypothesis. For $K \geq 1$ we say that a trajectory $(\boldsymbol{x}_n)_{n=0}^{\infty}$ is *K-coupled* if the corresponding sequence of routers (r_n) has the property that the set $\{r_n, \ldots, r_{n+K-1}\}$ is coupled for each $n \geq 0$. The definition strengthens the definition of coupling by assuming that a K-coupled trajectory experiences coupling with some regularity. For a chaotic system, this would be a restrictive assumption, but as shown in [BH03], the model in this paper is (piecewise) nonexpansive in an appropriate metric. We believe that for this model coupled

trajectories are typically K-coupled for some K.

While the K in the definition above refers to a number of iterations of the map `Next` in discrete time, the following proposition demonstrates that it also suffices to assume that each bottleneck router of a trajectory experiences congestion with some known frequency in continuous time. This latter approach was taken in [BH03] when the authors made the assumption that each router r in the model has some flow s that travels only through that router. This guarantees that there are at most $\frac{c_r}{\dot{x}^s}$ units of continuous time between consecutive congestion events at router r.

Proposition 2.1. *Suppose that, given an initial state $\boldsymbol{x}_0$, there is a maximum amount of continuous time Λ_r between consecutive congestion events at each bottleneck router r of the trajectory $\boldsymbol{x}_0$. Furthermore assume that the set of bottleneck routers of the trajectory generated by $\boldsymbol{x}_0$ is coupled. Then there exists a K such that the trajectory generated by $\boldsymbol{x}_0$ is K-coupled.*

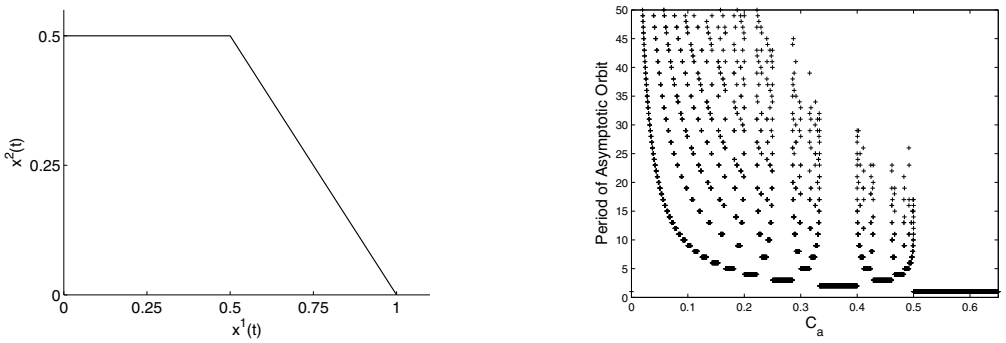
Proof. Let $R_{\boldsymbol{x}_0}$ be the set of bottleneck routers of the trajectory generated by $\boldsymbol{x}_0$, and let $\Lambda = \max_{r \in R_{\boldsymbol{x}_0}} \Lambda_r$. By construction, there is a minimum amount of continuous time l_r between consecutive congestion events at router r. Then every bottleneck router must experience congestion at least once during any time period of length Λ. Furthermore, bottleneck router r may not experience congestion more than $\lfloor \frac{\Lambda}{l_r} \rfloor$ times during any period of length Λ. Let $K = \sum_{r \in R'} \lfloor \frac{\Lambda}{l_r} \rfloor$. K is an upper bound for the number of congestion events that may occur during any time period of length Λ. Since each router r must experience congestion in any time period of length Λ, every bottleneck router r of the trajectory will experience congestion at least once during any K iterations of the map `Next`. Consequently, the trajectory generated by $\boldsymbol{x}_0$ is K-coupled. ■

3. Dynamics of the model. In most of our simulations, we have observed that $(\boldsymbol{x}_n)$ asymptotically approaches a periodic orbit (provided $\boldsymbol{x}_n \in E_1$ for all n), though it appears that the periodic orbit can have an arbitrarily large period. For certain special parameter choices, trajectories approach a set which includes points in E_2, where the process is not defined. Consider the router configuration from the previous example, except let $c_c = 1$ and set $c_b > c_c$. Figure 2(a) shows the router constraints for this model. Since router b will never experience congestion for these parameters, we can eliminate it from the model. In Figure 2, we study the period of the single periodic orbit that the trajectory generated by the initial state $\boldsymbol{x}_0 = (0, 0)$ approaches for values of c_a in the interval $(0, 0.65)$. Figure 2(b) shows the period of the periodic orbit that the trajectory of the initial state $\boldsymbol{x}_0 = (0, 0)$ approaches. When $c_a > \frac{1}{2}$, $\{\frac{1}{|\dot{\boldsymbol{x}}|}\dot{\boldsymbol{x}}\}$ is the unique periodic orbit of the system. In this case, the period is 1, and all trajectories approach it asymptotically. When c_a decreases to 0.5, this attracting period 1 orbit approaches the point of intersection of the two router constraints, $(c_a, 1-c_a)$. As c_a increases to 0.5, the period of the asymptotic orbit appears to go to infinity. The change in behavior as c_a decreases through 0.5 is caused by the attracting periodic orbit colliding with E_2.

When parameters are fixed, we have found that, assuming a few basic hypotheses, periodic orbits will always be exponential attractors. Specifically, we prove the following theorem.

Theorem 3.1. *Let $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{K-1}$ be a periodic orbit of `Next` with period K, where $K \geq 1$. Then the periodic orbit is exponentially attracting if and only if the periodic orbit is coupled.*

Remark 3.2. *Notice that a coupled period K orbit must in fact be K-coupled.*



(a) (b)

Figure 2. (a) *shows the router constraints for the network in Figure 1 when* $c_a = \frac{1}{2}$, $c_c = 1$, *and* $c_b > c_c$. *We have withheld the constraint generated by router b since these router capacities prevent router b from ever experiencing congestion. Panel* (b) *is a plot of the period of the periodic orbit to which the trajectory of the initial state* $x_0 = (0, 0)$ *asymptotes when* $\alpha = \frac{1}{2}$. *Consequently, we observe at most one period for each value of* c_a. *The graph consists of intervals where the period remains constant. The period appears to get arbitrarily large as* c_a *approaches* 0 *from above,* 0.5 *from below, and a variety of other values from one side or the other.*

Theorem 3.1 implies that the linearization of the Kth iterate of `Next` at a coupled period K orbit has all of its eigenvalues inside the unit circle. This and the implicit function theorem imply that each such orbit persists under small perturbations of the parameters. Hence each coupled periodic orbit exists for an open set O of parameter values. For the periodic orbit not to exist on the boundary of O, some point of the periodic orbit must approach E_2 as the boundary of O is approached.

The projection hyperplane. Given an initial state $\boldsymbol{x}_0$, the path $\boldsymbol{x}(t)$ starts at $\boldsymbol{x}_0$ and proceeds in the direction $\dot{\boldsymbol{x}}$ until constraint r_1 is hit at time T_1 and location $\boldsymbol{x}_1$. The instantaneous, multiplicative decrease then causes a discontinuous change of the system state to $\texttt{Drop}(\boldsymbol{x}_1)$. Notice that, regardless of the value of $\sum_{s\in S_{r_1}} x_0^s$, we have $\sum_{s\in S_{r_1}} x_1^s = c_{r_1}$. Furthermore, given a second initial state $\boldsymbol{x}_0'$ such that $\boldsymbol{x}_0 - \boldsymbol{x}_0' = \kappa\dot{\boldsymbol{x}}$, the two paths generated by these different initial states hit the constraint in the same location $\boldsymbol{x}_1$. Consequently, the dynamics generated by the initial state $\boldsymbol{x}_0'$ are the same as the dynamics generated by $\boldsymbol{x}_0$ with time shifted by κ ($\boldsymbol{x}_n = \boldsymbol{x}_n'$, $r_n = r_n'$, and $T_n = T_n' + \kappa$ for all n).

Studying the discrete time dynamics of the system eliminates the flow direction $\dot{\boldsymbol{x}}$. To facilitate our analysis, as in [BH03] we project the discrete time dynamics along $\dot{\boldsymbol{x}}$ from E_1 into a hyperplane $\mathbb{P}$. Let $\mathbf{1}$ be a vector with a 1 in each entry. Let $P(\boldsymbol{x})$ be a linear projection such that $P(\boldsymbol{x}) = \boldsymbol{x} + t\dot{\boldsymbol{x}}$, where t is chosen so $(\boldsymbol{x} + t\dot{\boldsymbol{x}}) \cdot \mathbf{1} = 0$. Define $\overline{\boldsymbol{x}}_n = P(\boldsymbol{x}_n)$. To formally define the dynamics in the *projection hyperplane* $\mathbb{P} = \{\overline{\boldsymbol{x}} \mid \overline{\boldsymbol{x}} \cdot \mathbf{1} = 0\}$ (see Figure 3) we first extend the definition of $\tau(\boldsymbol{x})$ to $\mathbb{R}^N$. Let $\tau(\boldsymbol{x})$ be the unique number such that $\boldsymbol{x} + \tau(\boldsymbol{x})\dot{\boldsymbol{x}}$ satisfies the router constraint $\sum_{s\in s_r} x^s \le c_r$ for each $r \in R$ and holds in at least one router constraint, i.e.,

$$\tau(\boldsymbol{x}) = \frac{c_r - \sum_{s\in S_r} x^s}{\sum_{s\in S_r} \dot{x}^s},$$

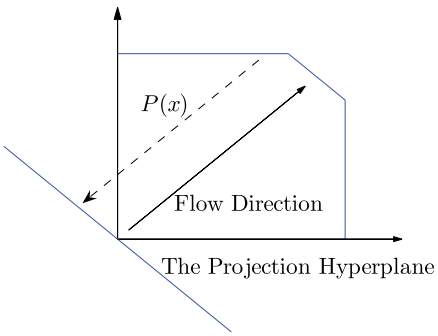


Figure 3. *The projection hyperplane $\mathbb{P}$ in relation to the flow direction $\dot{\boldsymbol{x}}$ and surface of congestion states E.*

where $r = \texttt{Router}(\boldsymbol{x})$. This extends the map $\texttt{Up}(\boldsymbol{x}) = \boldsymbol{x} + \tau(\boldsymbol{x})\dot{\boldsymbol{x}}$ to be well defined on all of $\mathbb{R}^N$. Furthermore, $P(\texttt{Up}(\overline{\boldsymbol{x}})) = \overline{\boldsymbol{x}}$ for all $\overline{\boldsymbol{x}} \in \mathbb{P}$ and $\texttt{Up}(P(\boldsymbol{x})) = \boldsymbol{x}$ for all $\boldsymbol{x} \in E$. We will continue to use an overline to refer to items in the projection hyperplane for the remainder of this paper. Thus we define $\overline{E} = P(E)$ and $\overline{E}_i = P(E_i)$. We define $\overline{\texttt{Next}}(\overline{\boldsymbol{x}}) = P \circ \texttt{Next} \circ \texttt{Up}(\overline{\boldsymbol{x}})$ for $\overline{\boldsymbol{x}} \in \overline{E}_1$.

Let $\overline{\texttt{Next}}_r = \overline{\texttt{Next}}|_{\overline{E}_1^r}$. Then for all $\overline{\boldsymbol{x}} \in \overline{E}_1^r$,

$$\begin{aligned}\overline{\texttt{Next}}_r(\overline{\boldsymbol{x}}) &= P \circ \texttt{Up} \circ \texttt{Drop}_r \circ \texttt{Up}(\overline{\boldsymbol{x}}) \\ &= P \circ \texttt{Drop}_r \circ \texttt{Up}(\overline{\boldsymbol{x}}).\end{aligned}$$

Thus for all $\overline{\boldsymbol{x}} \in \overline{E}_1^r$, $\overline{\texttt{Next}}_r$ is independent of the router that will next experience congestion after r (i.e., $\texttt{Router}(\overline{\texttt{Next}}(\overline{\boldsymbol{x}}))$). Since $\overline{\texttt{Next}}_r$ is an affine map, one may write

$$\overline{\texttt{Next}}_r(\overline{\boldsymbol{x}}) = \overline{A}_r\overline{\boldsymbol{x}} + \overline{\boldsymbol{b}}_r,$$

where $\overline{A}_r$ is a linear map on $\overline{E}_1^r$ that may be extended uniquely to all of $\mathbb{P}$ and $\overline{\boldsymbol{b}}_r$ is a constant vector.

Let $\boldsymbol{e}_i$ be the ith standard unit vector and $\overline{\boldsymbol{e}}_i = P(\boldsymbol{e}_i)$. The vectors $\overline{\boldsymbol{e}}_1, \overline{\boldsymbol{e}}_2, \ldots, \overline{\boldsymbol{e}}_N$ span $\mathbb{P}$ but are linearly dependent; indeed, $\sum_{s\in S} \dot{x}^s \overline{\boldsymbol{e}}_s = P(\dot{\boldsymbol{x}}) = 0$. Thus each vector in $\mathbb{P}$ can be represented as $\sum_{s\in S} a_s \overline{\boldsymbol{e}}_s$ in more than one way. Given a router r, we can make this representation unique by imposing the constraint $\sum_{s\in S_r} a_s = 0$ (this is possible because $\sum_{s\in S_r} \dot{x}^s \neq 0$). We will use this representation of vectors in $\mathbb{P}$ below.

Choose two initial states $\boldsymbol{x}_0, \boldsymbol{x}_0' \in E_1^r$. Then

$$\begin{aligned}\overline{A}_r(\overline{\boldsymbol{x}}_0 - \overline{\boldsymbol{x}}_0') &= \overline{\texttt{Next}}_r(\overline{\boldsymbol{x}}_0) - \overline{\texttt{Next}}_r(\overline{\boldsymbol{x}}_0') \\ &= P \circ \texttt{Drop}_r(\texttt{Up}(\overline{\boldsymbol{x}}_0) - \texttt{Up}(\overline{\boldsymbol{x}}_0')) \\ &= P \circ \texttt{Drop}_r(\boldsymbol{x}_0 - \boldsymbol{x}_0').\end{aligned}$$

Now write $\boldsymbol{x}_0 = \sum_{s\in S} x^s \boldsymbol{e}_s$ and $\boldsymbol{x}_0' = \sum_{s\in S} x'^s \boldsymbol{e}_s$. Let $\boldsymbol{u}_0 = \sum_{s\in S_r} x^s \boldsymbol{e}_s$, $\boldsymbol{u}_0' = \sum_{s\in S_r} x'^s \boldsymbol{e}_s$, $\boldsymbol{v}_0 = \sum_{s\notin S_r} x^s \boldsymbol{e}_s$, and $\boldsymbol{v}_0' = \sum_{s\notin S_r} x'^s \boldsymbol{e}_s$, and let $\overline{\boldsymbol{u}}_0 = P(\boldsymbol{u}_0)$, $\overline{\boldsymbol{u}}_0' = P(\boldsymbol{u}_0')$, $\overline{\boldsymbol{v}}_0 = P(\boldsymbol{v}_0)$, and $\overline{\boldsymbol{v}}_0' = P(\boldsymbol{v}_0')$. Then,

$$\begin{aligned}\overline{A}_r(\overline{\boldsymbol{x}}_0 - \overline{\boldsymbol{x}}'_0) &= P \circ \mathtt{Drop}_r(\boldsymbol{u}_0 - \boldsymbol{u}'_0 + \boldsymbol{v}_0 - \boldsymbol{v}'_0)\\ &= P(\alpha(\boldsymbol{u}_0 - \boldsymbol{u}'_0) + \boldsymbol{v}_0 - \boldsymbol{v}'_0)\\ &= \alpha(\overline{\boldsymbol{u}}_0 - \overline{\boldsymbol{u}}'_0) + \overline{\boldsymbol{v}}_0 - \overline{\boldsymbol{v}}'_0.\end{aligned}$$

Furthermore, since P is a linear map,

$$\boldsymbol{x}_0 = P(\boldsymbol{x}_0) = P\left(\sum_{s\in S} x^s \boldsymbol{e}_s\right) = \sum_{s\in S} x^s \overline{\boldsymbol{e}}_s$$

and likewise for $\overline{\boldsymbol{x}}'_0$. Then $\overline{\boldsymbol{x}}_0 - \overline{\boldsymbol{x}}'_0 = \sum_{s\in S}(x^s - x'^s)\overline{\boldsymbol{e}}_s$. Notice that since router r is congested at $\boldsymbol{x}_0, \boldsymbol{x}'_0 \in E_1^r$, we have that $\sum_{s\in S_r} x^s = \sum_{s\in S_r} x'^s = c_r$, and consequently $\sum_{s\in S_r}(x^s - x'^s) = 0$. Thus, letting $a^s = x^s - x'^s$, we have represented $\overline{\boldsymbol{x}}_0 - \overline{\boldsymbol{x}}'_0$ as $\sum_{s\in S} a_s \overline{\boldsymbol{e}}_s$ with $\sum_{s\in S_r} a_s = 0$. In this representation, $\overline{\boldsymbol{u}}_0 - \overline{\boldsymbol{u}}'_0 = \sum_{s\in S_r} a_s \overline{\boldsymbol{e}}_s$ and $\overline{\boldsymbol{v}}_0 - \overline{\boldsymbol{v}}'_0 = \sum_{s\notin S_r} a_s \overline{\boldsymbol{e}}_s$, and hence

$$\overline{A}_r \sum_{s\in S} a_s \overline{\boldsymbol{e}}_s = \sum_{s\in S_r} \alpha a_s \overline{\boldsymbol{e}}_s + \sum_{s\notin S_r} a_s \overline{\boldsymbol{e}}_s.$$

Notice that the output coordinates above satisfy the same constraint: $\sum_{s\in S_r} \alpha a_s = 0$.

Define

$$\overline{W}_r = \left\{ \sum_{s\in S_r} a_s \overline{\boldsymbol{e}}_s \;:\; \sum_{s\in S_r} a_s = 0 \right\}.$$

The following lemma describes how $\overline{A}_r$ acts on $\mathbb{P}$.

Lemma 3.3. *$\overline{A}_r \boldsymbol{z} = \alpha \boldsymbol{z}$ for all $\boldsymbol{z} \in \overline{W}_r$ and $\overline{A}_r \boldsymbol{z} = \boldsymbol{z}$ for all $\boldsymbol{z} \in \{\sum_{s\notin S_r} a_s \overline{\boldsymbol{e}}_s : a_s \in \mathbb{R}\}$.*

Proof. If $\boldsymbol{z} \in \overline{W}_r$, then $a_s = 0$ for all $s \notin S_r$. Furthermore, by definition, $\sum_{s\in S_r} a_s = 0$. Thus $\overline{A}_r \boldsymbol{z} = \alpha \boldsymbol{z}$. If $\boldsymbol{z} \in \{\sum_{s\notin S_r} a_s \overline{\boldsymbol{e}}_s : a_s \in \mathbb{R}\}$, then $a_s = 0$ for all $s \in S_r$. Thus $\overline{A}_r \boldsymbol{z} = \boldsymbol{z}$. ∎

The previous lemma demonstrates that each iteration of $\overline{\mathtt{Next}}$ contracts and translates a subspace of $\overline{E}_1$. The next proposition uses this result to show that if the trajectory is K-coupled, then for a suitably chosen large C, $\mathtt{Next}(\boldsymbol{x}, C)$ contracts a neighborhood U of an initial state $\boldsymbol{x}_0$ by a factor $a < 1$. To prove this result, given an initial state $\boldsymbol{x}_0 \in E_1$, we define a simplex-shaped neighborhood of $\overline{\boldsymbol{x}}_0 \in \overline{E_1}$. We then show using the previous lemma that each iteration of $\overline{\mathtt{Next}}(\overline{\boldsymbol{x}}_0)$ contracts a face of the simplex. Finally, we use coupling to show that the dynamics contract the image of the simplex under a proper number of iterations. However, we stop short of showing uniform contraction for all coupled trajectories since our contraction estimate depends on K.

Proposition 3.4 (contraction proposition). *There is a neighborhood $\overline{T}$ of $\mathbf{0}$ in $\mathbb{P}$ with the following property. For each $K > 0$, there exists an $a \in (0,1)$ such that, for each $\overline{\boldsymbol{x}}_0 \in \overline{E}_1$ that generates a K-coupled trajectory, if $\epsilon > 0$ is sufficiently small, then $\overline{\mathtt{Next}}(\overline{\boldsymbol{x}}_0 + \epsilon\overline{T}, i) \subset \overline{\boldsymbol{x}}_i + \epsilon\overline{T}$ for $0 \le i \le K(N-1)$ and $\overline{\mathtt{Next}}(\overline{\boldsymbol{x}}_0 + \epsilon\overline{T}, K(N-1)) \subset \boldsymbol{x}_{K(N-1)} + a\epsilon\overline{T}$.*

Remark 3.5. *As will be seen in the proof, ϵ needs only depend on the distance from the points $\{\overline{\boldsymbol{x}}_0, \overline{\boldsymbol{x}}_1, \ldots, \overline{\boldsymbol{x}}_{K(N-1)}\}$ to $\overline{E}_2$; it suffices that $(\overline{\boldsymbol{x}}_i + \epsilon\overline{T}) \cap \overline{E}_2 = \varnothing$ for $0 \le i \le K(N-1)$.*

Proof. Let

$$\overline{T} = \left\{ \sum_{s\in S} \lambda_s \overline{\boldsymbol{e}}_s : \sum_{s\in S} \lambda_s = 1,\ \lambda_s \ge 0,\ \forall s \in S \right\}.$$

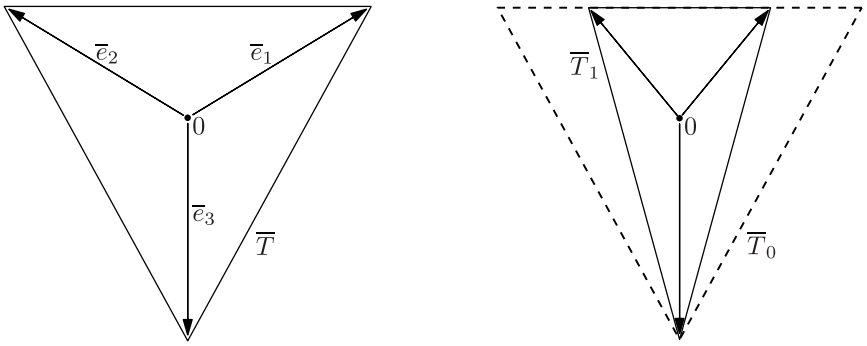


Figure 4. (a) *shows* $\overline{T}$*, the neighborhood in the projection hyperplane of* 0*, which is formed by all convex combinations of the points* $\{\overline{\boldsymbol{e}}_s : s \in S\}$*, when three flows are present in the model.* $\overline{T}_1$ *and* $\overline{T}_0$ *are shown in* (b) *assuming that* $S_{r_0} = \{1, 2\}$*. The arrows in* (b) *are the images of arrows from figure* (a)*. These images preserve orientation of the triangle.*

Choose $\epsilon > 0$ so that $(\overline{\boldsymbol{x}}_i + \epsilon\overline{T}) \cap \overline{E}_2 = \varnothing$ for $0 \le i \le K(N-1)$. Define the neighborhood $\overline{U}_0$ of $\overline{\boldsymbol{x}}_0$ by

$$\overline{U}_0 = \overline{\boldsymbol{x}}_0 + \epsilon\overline{T}.$$

Equivalently, $\overline{U}_0$ may be written as

$$\overline{U}_0 = \left\{ \overline{\boldsymbol{x}}_0 + \sum_{s\in S} \lambda_s \overline{\boldsymbol{e}}_s : \sum_{s\in S} \lambda_s = \epsilon,\ \lambda_s \ge 0,\ \forall s \in S \right\}.$$

Then $\overline{T}$ is an $(N-1)$-simplex containing the origin, and $\overline{U}_0$ is an $(N-1)$-simplex with vertices $\overline{\boldsymbol{x}}_0 + \epsilon\overline{\boldsymbol{e}}_1, \overline{\boldsymbol{x}}_0 + \epsilon\overline{\boldsymbol{e}}_2, \ldots, \overline{\boldsymbol{x}}_0 + \epsilon\overline{\boldsymbol{e}}_N$ (see Figure 4). Notice that this choice of ϵ ensures that, for every $\overline{\boldsymbol{x}} \in \overline{U}_0$, $\texttt{Router}(\texttt{Next}(\texttt{Up}(\overline{\boldsymbol{x}}), i)) = r_i$ for $0 \le i \le K(N-1)$. Define

$$\overline{U}_i = \overline{\texttt{Next}}(\overline{U}_{i-1})$$

for $1 \le i \le K(N-1)$ and

$$\overline{T}_i = \overline{U}_i - \overline{\boldsymbol{x}}_i$$

for $0 \le i \le K(N-1)$.

Let F_i be the map that sends $\overline{T}_i$ to $\overline{T}_{i+1}$. Then for $0 \le i \le K(N-1)$, F_i may be written as

$$\begin{aligned} F_i(\overline{\boldsymbol{x}}) &= \overline{\texttt{Next}}(\overline{\boldsymbol{x}} + \overline{\boldsymbol{x}}_i) - \overline{\boldsymbol{x}}_{i+1} \\ &= \overline{A}_{r_i}\overline{\boldsymbol{x}} + \overline{\texttt{Next}}(\overline{\boldsymbol{x}}_i) - \overline{\boldsymbol{x}}_{i+1} \\ &= \overline{A}_{r_i}\overline{\boldsymbol{x}}. \end{aligned}$$

Claim 1. For all $r \in R$, $\overline{A}_r\overline{T}_0 \subset \overline{T}_0$ and consequently for $0 \le i \le K(N-1)$, $\overline{T}_i \subset \overline{T}_0$.

The vertices of $\overline{T}_0$ are given by the set $\{\epsilon\overline{\boldsymbol{e}}_s : s \in S\}$, and $\overline{T}_0$ is formed by all convex combinations of these vertices. So for each $\overline{\boldsymbol{x}} \in \overline{T}_0$ we can write $\overline{\boldsymbol{x}} = \sum_{s\in S} \lambda_s \overline{\boldsymbol{e}}_s$ with $\sum_{s\in S} \lambda_s = \epsilon$ and $\lambda_s \geq 0$ for all $s \in S$. Let

$$\rho_r(\overline{\boldsymbol{x}}) = \frac{\sum_{s\in S_r} \lambda_s}{\sum_{s\in S_r} \dot{x}^s}. \tag{3.1}$$

Notice that $\sum_{s\in S_r}(\lambda_s - \rho_r(\overline{\boldsymbol{x}})\dot{x}^s) = 0$. Since $P(\dot{\boldsymbol{x}}) = 0$, or equivalently $\sum_{s\in S} \dot{x}^s \overline{\boldsymbol{e}}_s = 0$, by Lemma 3.3 we may write

$$\begin{aligned}\overline{A}_r \sum_{s\in S} \lambda_s \overline{\boldsymbol{e}}_s &= \overline{A}_r \left(\sum_{s\in S} (\lambda_s - \rho_r(\overline{\boldsymbol{x}})\dot{x}^s)\overline{\boldsymbol{e}}_s \right) \\ &= \sum_{s\in S_r} \alpha(\lambda_s - \rho_r(\overline{\boldsymbol{x}})\dot{x}^s)\overline{\boldsymbol{e}}_s + \sum_{s\notin S_r} (\lambda_s - \rho_r(\overline{\boldsymbol{x}})\dot{x}^s)\overline{\boldsymbol{e}}_s.\end{aligned}$$

Adding $\rho_r(\overline{\boldsymbol{x}}) \sum_{s\in S} \dot{x}^s \overline{\boldsymbol{e}}_s = 0$ to the right-hand side, we have

$$\overline{A}_r \sum_{s\in S} \lambda_s \overline{\boldsymbol{e}}_s = \sum_{s\in S_r} (\alpha\lambda_s + (1-\alpha)\rho_r(\overline{\boldsymbol{x}})\dot{x}^s)\overline{\boldsymbol{e}}_s + \sum_{s\notin S_r} \lambda_s \overline{\boldsymbol{e}}_s, \tag{3.2}$$

where $\rho_r(\overline{\boldsymbol{x}})$ is given by (3.1). Since $\sum_{s\in S_r}(\alpha\lambda_s + (1-\alpha)\rho_r(\overline{\boldsymbol{x}})\dot{x}^s) + \sum_{s\notin S_r} \lambda_s = \epsilon$ and $\alpha\lambda_s + (1-\alpha)\rho_r(\overline{\boldsymbol{x}})\dot{x}^s > 0$, this gives a representation of $\overline{A}_r$ acting on $\overline{T}_0$ in the coordinate system with $\sum_{s\in S} \lambda_s = \epsilon$ and $\lambda_s \geq 0$ for all $s \in S$. (In this coordinate system, we have shown that $\overline{A}_r$ is represented by a stochastic matrix; a vector of nonnegative numbers whose sum is 1 is mapped to another such vector.) Thus, $\overline{A}_r\overline{T}_0 \subset \overline{T}_0$ for every r. It follows by induction that if $\overline{T}_i \subset \overline{T}_0$, then $\overline{T}_{i+1} = \overline{A}_{r_i}\overline{T}_i \subset \overline{A}_{r_i}\overline{T}_0 \subset \overline{T}_0$.

Claim 2. Suppose that $\overline{\boldsymbol{y}}_0$ is a point on the boundary of $\overline{T}_0$. Define $\overline{\boldsymbol{y}}_i = \overline{\texttt{Next}}(\overline{\boldsymbol{y}}_0, i)$. Let d_i be the dimension of the minimum-dimensional face of $\overline{T}_0$ containing $\overline{\boldsymbol{y}}_i$, or $N-1$ if $\overline{\boldsymbol{y}}_i$ is in the interior of $\overline{T}_0$. Then if $d_i < N-1$, $d_{i+K} \geq d_i + 1$.

For each i, write $\overline{\boldsymbol{y}}_i = \sum_{s\in S} \lambda_s^i \overline{\boldsymbol{e}}_s$ with $\sum_{s\in S} \lambda_s^i = \epsilon$ and $\lambda_s^i \geq 0$. Then d_i is one less than the number of λ_s^i that are strictly positive. Notice that if $\lambda_s^i > 0$ for some $i \geq 0$, then either $s \notin S_{r_i}$ and $\lambda_s^{i+1} = \lambda_s^i$, or $s \in S_{r_i}$, and by (3.2), $\lambda_s^{i+1} = \alpha\lambda_s^i + (1-\alpha)\rho_r(\overline{\boldsymbol{y}}_i)\dot{x}^s$. In both cases, $\lambda_s^{i+1} > 0$. It follows that if $\lambda_s^i > 0$, then $\lambda_s^{i+n} > 0$ for every $n \geq 0$. Since this is true for all $s \in S$, we have that $d_{i+n} \geq d_i$ for all $n \geq 0$.

Since $(\boldsymbol{x}_i)$ is a coupled trajectory, the set $\{r_i, r_{i+1}, \ldots, r_{i+K-1}\}$ is coupled for each i. Then because $d_i < N-1$, there must exist $s, s' \in S$ such that $\lambda_s^i = 0$, $\lambda_{s'}^i > 0$, and $s, s' \in S_{r_{i+j}}$ for some $j \in \{0, 1, \ldots, K-1\}$. Then $\rho_r(\overline{\boldsymbol{y}}_{i+j}) > 0$, and thus $\lambda_s^{i+j+1} = \alpha\lambda_s^{i+j} + (1-\alpha)\rho_r(\overline{\boldsymbol{y}}_{i+j})\dot{x}^s \geq (1-\alpha)\rho_r(\overline{\boldsymbol{y}}_{i+j})\dot{x}^s > 0$. Because of this, $d_{i+j+1} \geq d_i + 1$, and hence $d_{i+K} \geq d_i + 1$ as well.

Now we prove the proposition. Because the dynamics on $\overline{T}_0$ are generated by applying linear maps, it suffices to show $\overline{A}_{r_{K(N-1)-1}} \cdots \overline{A}_{r_1}\overline{A}_{r_0}\overline{T}_0 \subset a\overline{T}_0$ for some $a \in (0,1)$ that may depend on K, but not the particular sequence $r_1, r_2, \ldots, r_{K(N-1)-1}$. If $\overline{\boldsymbol{x}}$ is a vertex of $\overline{T}_0$, then $d_0 = 0$. Then we must have that $d_{K(N-1)} = (N-1)$. Consequently, the image of each vertex after $K(N-1)$ iterations of `Next` lies in the interior of $\overline{T}_0$. Since $\overline{T}_{K(N-1)}$ is formed by all convex combinations of the images of the vertices under the map $\overline{A}_{r_{K(N-1)-1}} \cdots \overline{A}_{r_1}\overline{A}_{r_0}$, we have

$\overline{T}_{K(N-1)} \subset T_0$. Furthermore, since all the vertices of $\overline{T}_{K(N-1)}$ are in the interior of $\overline{T}_0$, there is an $a' \in (0,1)$ that depends on $\{r_0, r_1, \ldots, r_{K(N-1)}\}$ and the parameters of the model such that $\overline{T}_{K(N-1)} \subset a'\overline{T}_0$. As only a finite number of router sequences of length $K(N-1)$ satisfy the conditions of the theorem, let a be the maximum of a' over such sequences. Then $a \in (0,1)$ and $\overline{T}_{K(N-1)} \subset a\overline{T}_0$. Then since $\overline{\texttt{Next}}(\overline{\boldsymbol{x}}_0 + \epsilon\overline{T}, K(N-1)) = \overline{U}_{K(N-1)} = \overline{\boldsymbol{x}}_{K(N-1)} + \overline{T}_{K(N-1)}$, $\overline{\texttt{Next}}(\overline{\boldsymbol{x}}_0 + \epsilon\overline{T}, K(N-1)) \subset \boldsymbol{x}_{K(N-1)} + a\epsilon\overline{T}$. ■

A metric on E. Since our results so far are proved in a homeomorphic image of E in the projection hyperplane, we define a metric on the projection hyperplane and then lift that metric onto E. For any two points $\boldsymbol{x}, \boldsymbol{y} \in E$, we define

$$d_E(\boldsymbol{x}, \boldsymbol{y}) = |P(\boldsymbol{x}) - P(\boldsymbol{y})|_{L^2}.$$

Although we here use the L^2 metric in $\boldsymbol{P}$, the results below will hold for any chosen metric.

Proof of Theorem 3.1. Assume that the periodic orbit generated by $\boldsymbol{x}_0$ is coupled. Since $\boldsymbol{x}_0$ generates a periodic orbit in E_1, $\overline{\boldsymbol{x}}_0$ generates a periodic orbit in $\mathbb{P}$. Since the metric on E is a lift of the metric on $\mathbb{P}$, it suffices to show that the periodic orbit generated by $\overline{\boldsymbol{x}}_0$ in $\mathbb{P}$ is exponentially attracting. Since the periodic orbit generated by $\overline{\boldsymbol{x}}_0$ is coupled, by Proposition 3.4, there exists a neighborhood $\overline{T}$ of $\mathbf{0}$ and an $\epsilon > 0$ such that $\overline{\texttt{Next}}(\overline{\boldsymbol{x}}_0 + \epsilon\overline{T}, i) \subset \overline{\boldsymbol{x}}_i + \epsilon\overline{T}$ for $0 \le i \le K(N-1)$ and $\overline{\texttt{Next}}(\overline{\boldsymbol{x}}_0 + \epsilon\overline{T}, K(N-1)) \subset \boldsymbol{x}_{K(N-1)} + a\epsilon\overline{T}$. Let $\overline{U} = \overline{\boldsymbol{x}}_0 + \epsilon\overline{T}$ and define

$$d_n = \max_{x \in \texttt{Next}(\overline{\boldsymbol{x}}_0 + \epsilon\overline{T}, nK(N-1))} \{|\boldsymbol{x}_0 - \boldsymbol{x}|_{L^2}\}.$$

Proposition 3.4 implies that since $\overline{\boldsymbol{x}}_{nK(N-1)} = \overline{\boldsymbol{x}}_0$, $\texttt{Next}(\overline{U}, nK(N-1)) - \overline{\boldsymbol{x}}_0 \subset a^n(\overline{U} - \overline{\boldsymbol{x}}_0)$ for every n, $d_n < a^n d_0$. Thus d_n goes to 0 as n goes to infinity, and the periodic orbit is exponentially attracting.

Conversely assume that the periodic orbit generated by $\boldsymbol{x}_0$ is not coupled. Then there exist proper subsets $S' \subset S$ and $R' \subset R$ such that for each $r \in R'$ either $S_r \subset S'$ or $S_r \subset S \backslash S'$. Let $\tilde{R}' = \{r \in R' : S_r \subset S'\}$, and assume without loss of generality that $r_0 \in \tilde{R}'$. Let $\boldsymbol{v} = \sum_{s \in S \backslash S'} \dot{x}^s \boldsymbol{e}_s$. The since $\boldsymbol{x}_0 \in E_1^{r_0}$ and $E_1^{r_0}$ depends only on the coordinates x^s with $s \in S_{r_0} \subset S'$, it follows that $\boldsymbol{x}_0 + \epsilon\boldsymbol{v} \in E_1^{r_0}$ for ϵ sufficiently small. Then for ϵ sufficiently small, the trajectory generated by $\boldsymbol{x}_0 + \epsilon\boldsymbol{v}$ will have congestion events at the same routers as the trajectory generated by $\boldsymbol{x}_0$. In continuous time the kth congestion event will happen at the same time if $r_k \in \tilde{R}'$, or time ϵ earlier if $r_k \notin \tilde{R}'$. In particular, $\boldsymbol{x}_0 + \epsilon\boldsymbol{v}$ also generates a periodic orbit. Therefore, the periodic orbit generated by $\boldsymbol{x}_0$ is not exponentially attracting. ■

Limiting to periodic orbits. The previous results state that periodic orbits are exponentially attracting, but they do not describe which trajectories will approach these period orbits. In fact, as the following theorem demonstrates, all K-coupled trajectories approach a periodic orbit as long as they are bounded away from E_2.

Theorem 3.6. *Assume that an initial state $\boldsymbol{x}_0 \in E_1$ generates a coupled trajectory that is bounded away from E_2. Then the trajectory generated by $\boldsymbol{x}_0$ exponentially approaches a periodic orbit.*

Proof. Choose $\delta > 0$ such that the trajectory generated by $\boldsymbol{x}_0$ stays δ away from E_2 in the d_E metric. Let $\overline{T}$ be the neighborhood of $\mathbf{0}$ from Proposition 3.4 and let $\epsilon > 0$ be small

enough that $\operatorname{diam}(\epsilon\overline{T}) < \delta$. This ensures that ϵ is small enough for Proposition 3.4 to apply not only to $\overline{\boldsymbol{x}}_0$ but to $\overline{\boldsymbol{x}}_n$ for all $n \geq 0$; see Remark 3.5. Without loss of generality assume that r_0 is a bottleneck router. Let $n_0 = 0$, and recursively define n_k to be the first iterate after iterate n_{k-1} when the trajectory generated by $\overline{\boldsymbol{x}}_0$ returns to $\overline{E}_1^{r_0}$ and the set of routers $\{r_{n_{k-1}}, \ldots, r_{n_k-1}\}$ is coupled. Since r_0 is a bottleneck router, n_k exists for all k. Because $\overline{E}_1^{r_0}$ is bounded, for every $\gamma > 0$ there exist two indices k and k' such that $\overline{\boldsymbol{x}}_{n_k}$ and $\overline{\boldsymbol{x}}_{n_{k'}}$ are within γ of each other. Consequently, choosing γ sufficiently small, there exist two indices k and k', with $k < k'$, such that $\overline{\boldsymbol{x}}_{n_{k'}} \in \overline{\boldsymbol{x}}_{n_k} + \epsilon\overline{T}$. Let $K = n_{k'} - n_k$. Then since $\overline{\boldsymbol{x}}_{n_k+K} \in \overline{\boldsymbol{x}}_{n_k} + \epsilon\overline{T}$, it follows by Proposition 3.4 that $\overline{\boldsymbol{x}}_{n_k+K+j} \in \overline{\boldsymbol{x}}_{n_k+j} + \epsilon\overline{T}$ for $0 \leq j \leq K$, and hence by induction, $\overline{\boldsymbol{x}}_{n_k+nK} \in E_1^{r_0}$ for all $n \geq 0$. Furthermore the set $\{r_{n_k+(n-1)K}, \ldots, r_{n_k+nK-1}\} = \{r_{n_k}, \ldots, r_{n_{k'}-1}\}$ for all $n \geq 0$ and is coupled. Thus the trajectory $\{\overline{\boldsymbol{x}}_{n_k}, \overline{\boldsymbol{x}}_{n_k+1}, \overline{\boldsymbol{x}}_{n_k+2}, \ldots\}$ is K-coupled. By Proposition 3.4 it follows that $\overline{\boldsymbol{x}}_{n_k+K(N-1)} \in \overline{\boldsymbol{x}}_{n_k} + a\epsilon\overline{T}$, that $\overline{\boldsymbol{x}}_{n_k+2K(N-1)} \in \overline{\boldsymbol{x}}_{n_k+K(N-1)} + a^2\epsilon\overline{T}$, and so on. Define $U_0 = \mathtt{Up}(\boldsymbol{x}_{n_k} + \epsilon\overline{T})$, and for each $i > 0$ define $U_i = \mathtt{Next}(U_{i-1}, K(N-1))$. Then define $\boldsymbol{y}_0 = \boldsymbol{x}_{n_k}$ and $\boldsymbol{y}_i = \mathtt{Next}(\boldsymbol{x}_{n_k}, iK(N-1))$. Then $(\boldsymbol{y}_i)$ is a subsequence of the trajectory $(\boldsymbol{x}_n)$ in $E_1^{r_0}$ with $\boldsymbol{y}_i, \boldsymbol{y}_{i+1} \in U_i$. Thus by Proposition 3.4,

$$d_E(\boldsymbol{y}_i, \boldsymbol{y}_{i+1}) \leq \operatorname{diam}(U_i) \leq \operatorname{diam}(a^i\epsilon\overline{T}) = a^i \operatorname{diam}(U_0).$$

Thus the sequence $(\boldsymbol{y}_i)$ is Cauchy and converges to a point $\boldsymbol{p}_0 \in E_1^r$ that is a fixed point of the map $\mathtt{Next}(\cdot, K(N-1))$. The orbit generated by $\boldsymbol{p}_0$ then must be a periodic orbit. Consequently, $(\boldsymbol{x}_n)$ exponentially approaches this periodic orbit. ■

4. Conclusion.

We have studied the dynamics of a multiple flow multiple router model of TCP defined by a piecewise affine map `Next` repeatedly applied to a polygonal surface E. Most computer simulations of this model of our own and in [BH03] have trajectories asymptotically approaching periodic orbits, though exceptional cases exist when the trajectory asymptotically approaches E_2, a measure zero subset of E where the dynamics are undefined. Of course, these simulations explore only networks of limited size. We have demonstrated in Theorem 3.1 that, regardless of network size, all coupled periodic orbits are exponentially attracting and thus are stable attractors. This implies that coupled periodic orbits persist on an open set of parameter values. We demonstrated in Theorem 3.6 that every coupled trajectory that is bounded away from E_2 will exponentially approach a periodic orbit. When a trajectory that is bounded away from E_2 is not coupled, the model coordinates can be partitioned into coupled subsets that evolve independently in a neighborhood of the trajectory, and Theorems 3.1 and 3.6 apply to the dynamics of each subset. Though simulations suggest that some parameter values yield trajectories that are not bounded away from E_2, these parameter values appear to form a Cantor set [BH03], which we conjecture to have measure 0. While our results are for a particular model of data network traffic, similar dynamics may occur in other network flow models where flows are coupled via bottleneck nodes. A key feature of the model is that the initial state can determine which nodes are bottlenecks and thus determine the nature of the coupling. Our results and techniques apply to a situation where the uncoupled dynamics are neutrally stable and coupling provides intermittent discontinuous contraction among some of the state variables.

Acknowledgments. We thank A. Finkbiner, E. Ott, and the referees for their helpful comments.

REFERENCES

[BH02] F. Baccelli and D. Hong, *AIMD, fairness and fractal scaling of TCP traffic*, in Proceedings of the IEEE INFOCOM, New York, NY, 2002, pp. 229–238.

[BH03] F. Baccelli and D. Hong, *Interaction of TCP flows as billiards*, in Proceedings of the IEEE INFOCOM, San Francisco, CA, 2003, pp. 895–905.

[F94] S. Floyd, *TCP and explicit congestion notification*, ACM SIGCOMM Comput. Comm. Rev., 24 (1994), pp. 8–23.

[FJ94] S. Floyd and V. Jacobson, *Random early detection gateways for congestion avoidance*, IEEE/ACM Trans. Networking, 1 (1993), pp. 397–413.

[J88] V. Jacobson, *Congestion avoidance and control*, in Proceedings of SIGCOMM, Stanford, CA, ACM, New York, 1988, pp. 157–173.

[RAL02] P. Ranjan, H. Abed, and R. J. La, *Nonlinear instabilities in TCP-RED*, in Proceedings of the IEEE INFOCOM, New York, NY, 2002, pp. 249–258.

[SL06] R. Shorten and D. Leith, *Partitioning and invariance of AIMD dynamics in synchronised networks*, in Proceedings of the International Control Conference, Glasgow, Scotland, 2006, abstract 213.

[SWL06] R. Shorten, F. Wirth, and D. Leith, *A positive system model of TCP-like congestion control: Asymptotic results*, IEEE/ACM Trans. Networking, 14 (2006), pp. 616–629.

SIAM J. Applied Dynamical Systems
Vol. 8, No. 1, pp. 160–179
© 2009 Society for Industrial and Applied Mathematics

Generalized Helmholtz–Kirchhoff Model for Two-Dimensional Distributed Vortex Motion*

Raymond Nagem†, Guido Sandri†, David Uminsky‡, and C. Eugene Wayne‡

Abstract. The two-dimensional Navier–Stokes equations are rewritten as a system of coupled nonlinear ordinary differential equations. These equations describe the evolution of the moments of an expansion of the vorticity with respect to Hermite functions and of the centers of vorticity concentrations. We prove the convergence of this expansion and show that in the zero viscosity and zero core size limit we formally recover the Helmholtz–Kirchhoff model for the evolution of point vortices. The present expansion systematically incorporates the effects of both viscosity and finite vortex core size. We also show that a low-order truncation of our expansion leads to the representation of the flow as a system of interacting Gaussian (i.e., Oseen) vortices, which previous experimental work has shown to be an accurate approximation to many important physical flows [P. Meunier, S. Le Dizès, and T. Leweke, *C. R. Phys.*, 6 (2005), pp. 431–450].

Key words. vortex dynamics, Navier–Stokes equations, fluid dynamics

AMS subject classifications. 76D05, 76D17, 35Q30

DOI. 10.1137/080715056

1. Introduction.

In this paper we represent solutions of the two-dimensional Navier–Stokes equations as a system of interacting vortices. This expansion, which generalizes the Helmholtz–Kirchhoff model of interacting point vortices in an inviscid fluid, systematically incorporates the effects of both vorticity and finite vortex core size. Furthermore, we give conditions which guarantee the convergence of our expansion. Incompressible viscous flow has two standard analytic representations: a formulation in terms of the primitive velocity and pressure variables, and a formulation in terms of the velocity and vorticity variables [8]. The velocity-vorticity representation has particular advantages when boundaries are unimportant, since vorticity cannot be created or destroyed in the interior of a fluid. The vorticity field can also be directly related to physically observed flow structures such as line and ring vortices.

In two space dimensions, the vorticity field has the additional advantage of reducing to a scalar. An early representation of two-dimensional flow in terms of moving point vortices was developed by Kirchhoff [5] and by Von Helmholtz [17]. The point vortex model has been studied extensively—a thorough review of the model and recent developments is provided in the monograph by Newton [13]. While the Helmholtz–Kirchhoff point vortex model captures many of the basic physical phenomena observed in two-dimensional rotational flows, experi-

*Received by the editors February 1, 2008; accepted for publication (in revised form) by T. Kaper October 1, 2008; published electronically January 23, 2009.
http://www.siam.org/journals/siads/8-1/71505.html
†Department of Aerospace and Mechanical Engineering, Boston University, Boston, MA 02215 (nagem@bu.edu, sandri@bu.edu).
‡Department of Mathematics and Statistics and Center for Biodynamics, Boston University, Boston, MA 02215 (duminsky@math.bu.edu, cew@math.bu.edu). The work of these authors was supported in part by DMS-0405724.

ments with even relatively simple vortex configurations exhibit complications far beyond the point vortex predictions [11]. Additionally, the classical point vortex model neglects the effects of *viscosity*. However, these experiments also reinforce the idea that in many circumstances the fluid flow may be well approximated by a collection of interacting vortices, albeit vortices with finite core size, subject to the effects of viscosity. Earlier work to introduce a finite core size model for Euler's equations was done in [9]. In this paper the vortex patches that are modeled are described as disjoint piecewise-constant elliptical finite-area vortex regions (FAVORs). In this model, higher spatial moments of the ellipse are allowed to evolve under the flow. The moment model is truncated to second order, and the corotating vortex example is analyzed. Later in [10] the authors generalize the spatial moment model found in [9] by incorporating weak Newtonian viscosity to derive equations of motions for two like signed vortices.

Other recent studies of the interaction of viscous vortices can be found in [1, 2, 6, 7, 11, 16]. The main focus of most of these papers is the merger of two like signed vortices. A metastable state is found before merger which consists of two rotating near-circular vortices. More recently, it has been conjectured that in a two-vortex system the profiles relax to a pair of Gaussian vortices before merging [7]. An alternative method for incorporating viscous effects was proposed in [1] by adding a random forcing term to the ODEs which describe the motion of the vortex centers. Thus, it is of interest to extend and generalize the Helmholtz–Kirchhoff point vortex model to a model that incorporates nonzero vortex core size and viscous effects while retaining its basic form. Such an extension is the goal of this paper.

The governing equations for the velocity ($\mathbf{u}$) and pressure (p) variables are

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{u} = -\nabla\frac{p}{\rho} + \nu\Delta\mathbf{u}, \tag{1}$$

$$\nabla\cdot\mathbf{u} = 0, \tag{2}$$

where ρ is the fluid density and ν is the kinematic viscosity. Taking the curl of (1) and (2) gives

$$\frac{\partial \boldsymbol{\omega}}{\partial t} + (\mathbf{u}\cdot\nabla)\boldsymbol{\omega} - (\boldsymbol{\omega}\cdot\nabla)\mathbf{u} = \nu\Delta\boldsymbol{\omega}, \tag{3}$$

$$\boldsymbol{\omega} = \nabla\times\mathbf{u}, \quad \nabla\cdot\boldsymbol{\omega} = 0, \tag{4}$$

which are the governing equations for the velocity-vorticity variables. For two-dimensional flows, the vorticity vector $\boldsymbol{\omega}$ is perpendicular to the plane of the flow, and the third term on the left-hand side of (3) vanishes. The condition $\nabla\cdot\boldsymbol{\omega} = 0$ is identically satisfied, and (3) then reduces to the single scalar equation:

$$\frac{\partial \omega}{\partial t} + (\mathbf{u}\cdot\nabla)\omega = \nu\Delta\omega, \tag{5}$$

where ω is the single nonzero component of the vorticity. A drawback of the formulation in (5) is that the velocity of the fluid is still present in the equation. However, assuming that

the vorticity field is sufficiently localized, the velocity vector can be computed in terms of the vorticity ω by the Biot–Savart law

$$\mathbf{u}(\mathbf{x}) = \frac{1}{2\pi} \int_{\mathbb{R}^2} \frac{(\mathbf{x}-\mathbf{y})^{\perp}}{|\mathbf{x}-\mathbf{y}|^2} \omega(\mathbf{y}) d\mathbf{y}, \tag{6}$$

where for a two-vector $\mathbf{z} = (z_1, z_2)$, $\mathbf{z}^{\perp} = (-z_2, z_1)$.

In this paper we use (5) and (6) to develop a vorticity representation of two-dimensional viscous flow. Our representation is based on a decomposition of the vorticity field into a set of moving distributed vortices. Differential equations are derived for the motion of the vortex centers and for the time evolution of the vortex distributions. The evolution of each individual vortex is represented as an expansion with respect to a sequence of Hermite functions. Such expansions have proven useful in theoretical studies of two-dimensional fluid flows [3, 4] and the leading-order term in this expansion is precisely the Gaussian vortex (i.e., Oseen vortex [15, 14]) whose utility as an approximation for vortex interaction was shown in [11]. We show that the coefficients in this expansion satisfy a system of ODEs whose coefficients can be explicitly represented in terms of a fixed, computable kernel function. We also prove the convergence of this expansion. It is shown that our representation reduces in the appropriate limit to the Helmholtz–Kirchhoff model and allows at the same time arbitrarily complex evolution and interaction of the moving vortices.

In the present paper we concentrate on the mathematical formulation of the generalized Helmholtz–Kirchhoff model. In future work we will explore the predictions of this model both numerically and analytically in a number of different physical settings.

2. The "multivortex" expansion. In this section we separate the solution of the vorticity equation into N components and derive separate evolution equations for each component. From a physical point of view this decomposition will be most useful when each of the components corresponds to a localized region of vorticity (e.g., a vortex) well separated from the other lumps, but the mathematical development described below is well defined without regard to these physical considerations. However, with this application in mind, we will often refer to each of the components as a "vortex."

Consider the initial value problem for the two-dimensional vorticity equation

$$\begin{aligned} &\frac{\partial \omega}{\partial t} = \nu \Delta \omega - \mathbf{u} \cdot \nabla \omega, \\ &\omega = \omega(\mathbf{x}, t), \quad \mathbf{x} \in \mathbb{R}^2, \ t > 0, \\ &\omega(\mathbf{x}, 0) = \omega_0(\mathbf{x}), \end{aligned} \tag{7}$$

where $\mathbf{u}$ is the velocity field associated with the vorticity field ω. We begin by decomposing the initial vorticity distribution by writing

$$\omega_0(\mathbf{x}) = \sum_{j=1}^{N} \omega_0^j(\mathbf{x}). \tag{8}$$

Of course, this decomposition is not unique—even the number of pieces, N, into which we decompose the vorticity is up to us to choose. In general the choice we make will be motivated by physical considerations; however, for the development below, all we require of the

decomposition is that the total vorticity of each vortex be nonzero, i.e.,

$$m^j = \int_{\mathbb{R}^2} \omega_0^j(\mathbf{x})\mathrm{d}\mathbf{x} \neq 0, \qquad j = 1, \ldots, N. \tag{9}$$

If (9) is satisfied we define $\mathbf{x}_0^j$ by

$$\int_{\mathbb{R}^2} (\mathbf{x} - \mathbf{x}_0^j)\omega_0^j(\mathbf{x})\mathrm{d}\mathbf{x} = 0, \tag{10}$$

or equivalently,

$$\mathbf{x}_0^j = \frac{1}{m^j}\int_{\mathbb{R}^2} \mathbf{x}\omega_0^j(\mathbf{x})\mathrm{d}\mathbf{x}. \tag{11}$$

We now write the vorticity for $t > 0$ as

$$\omega(\mathbf{x}, t) = \sum_{j=1}^{N} \omega^j(\mathbf{x} - \mathbf{x}^j(t); t) \tag{12}$$

and the velocity field as

$$\mathbf{u}(\mathbf{x}, t) = \sum_{j=1}^{N} \mathbf{u}^j(\mathbf{x} - \mathbf{x}^j(t); t), \tag{13}$$

where $\mathbf{u}^j(\mathbf{y}, t)$ is the velocity field associated with $\omega^j(\mathbf{y}, t)$ by the Biot–Savart law. Of course we still have to define the equations of motion for $\omega^j(\mathbf{y}, t)$ and $\mathbf{x}^j(t)$.

The centers of the vorticity regions, $\mathbf{x}^j(t)$, and the vorticity regions themselves evolve via a coupled system of ordinary-partial differential equations constructed so that in the limit of zero viscosity and when the different components of the vorticity happen to be point vortices (i.e., Dirac-delta functions) we recover the Helmholtz–Kirchhoff point vortex equations. If we take the partial derivative of (12) and use the equation satisfied by the vorticity, we find

$$\begin{aligned} \partial_t\omega(\mathbf{x}, t) &= \sum_{j=1}^{N} \partial_t\omega^j(\mathbf{s} - \mathbf{s}^j(t), t) - \sum_{j=1}^{N} \dot{\mathbf{x}}^j(t) \cdot \nabla\omega^j(\mathbf{x} - \mathbf{x}^j(t), t) \\ &= \sum_{j=1}^{N} \nu\Delta\omega^j(\mathbf{x} - \mathbf{x}^j(t), t) - \sum_{j=1}^{N} \left(\sum_{\ell=1}^{N} \mathbf{u}^\ell(\mathbf{x} - \mathbf{x}^\ell(t), t)\right) \cdot \nabla\omega^j(\mathbf{x} - \mathbf{x}^j(t), t). \end{aligned} \tag{14}$$

Given this equation, it is natural to define ω^j as the solution of

$$\begin{aligned} \frac{\partial\omega^j}{\partial t}(\mathbf{x} - \mathbf{x}^j(t), t) &= \nu\Delta\omega^j(\mathbf{x} - \mathbf{x}^j(t), t) - \left(\sum_{\ell=1}^{N} \mathbf{u}^\ell(\mathbf{x} - \mathbf{x}^\ell(t), t)\right) \cdot \nabla\omega^j(\mathbf{x} - \mathbf{x}^j(t), t) \\ &\quad + \dot{\mathbf{x}}^j(t) \cdot \nabla\omega^j(\mathbf{x} - \mathbf{x}^j(t), t), \qquad j = 1, \ldots, N. \end{aligned} \tag{15}$$

To close this system of equations we must specify how the centers of vorticity $\mathbf{x}^j(t)$ evolve. We impose the condition that the first moment of each vorticity region must vanish at every time $t > 0$; i.e., we require that

$$\int_{\mathbb{R}^2} (\mathbf{x} - \mathbf{x}^j(t))\omega^j(\mathbf{x} - \mathbf{x}^j(t), t)\mathrm{d}\mathbf{x} = \mathbf{0} \qquad \text{for all} \quad t > 0,\ j = 1, \dots, N. \tag{16}$$

(Note that this equation really contains two conditions, one for each component of $\mathbf{x} - \mathbf{x}^\ell(t)$.) We impose this condition to fix the evolution of $\mathbf{x}^j(t)$ because Gallay and Wayne have recently shown [3] that if one considers the evolution of general solutions of (7), the solution will approach an Oseen vortex, and the rate of the approach will be faster if the vorticity distribution has first moment equal to zero. Solutions of (7) preserve the first moment, and hence if the initial conditions have first moment equal to zero, the solution will have first moment zero for all time. The equations (15) no longer preserve the first moment, and thus we impose this condition for all time, which then defines the motion of the center of vorticity.

Note that if we change variables in (16) to $\mathbf{z} = \mathbf{x} - \mathbf{x}^j(t)$, we find

$$\int_{\mathbb{R}^2} \mathbf{z}\omega^j(\mathbf{z}, t)\mathrm{d}\mathbf{z} = \mathbf{0}. \tag{17}$$

Since this equation holds for all $t > 0$, we can differentiate both sides with respect to t to obtain

$$\int_{\mathbb{R}^2} \mathbf{z}\partial_t\omega^j(\mathbf{z}, t)\mathrm{d}\mathbf{z} = \mathbf{0}. \tag{18}$$

Using (15), we can insert the formula for $\partial_t\omega^j$ into this integral, and we obtain

$$\begin{aligned} \nu \int_{\mathbb{R}^2} \mathbf{z}\Delta\omega^j(\mathbf{z}, t)\mathrm{d}\mathbf{z} - \int_{\mathbb{R}^2} \mathbf{z} \left(\sum_{\ell=1}^{N} \mathbf{u}^\ell(\mathbf{z} + \mathbf{x}^j(t) - \mathbf{x}^\ell(t), t) \right) \cdot \nabla\omega^j(\mathbf{z}, t)\mathrm{d}\mathbf{z} \\ + \int_{\mathbb{R}^2} \mathbf{z} \left(\dot{\mathbf{x}}^j(t) \cdot \nabla\omega^j(\mathbf{z}, t) \right) \mathrm{d}\mathbf{z} = \mathbf{0}. \end{aligned} \tag{19}$$

We first note that if we integrate twice by parts, we have

$$\int_{\mathbb{R}^2} \mathbf{z}\Delta\omega^j(\mathbf{z}, t)\mathrm{d}\mathbf{z} = \mathbf{0}. \tag{20}$$

Next if we take the nth component and integrate by parts, we find

$$\int_{\mathbb{R}^2} z_n \left(\dot{\mathbf{x}}^j(t) \cdot \nabla\omega^j(\mathbf{z}, t) \right) \mathrm{d}\mathbf{z} = \sum_{m=1}^{2} \dot{\mathbf{x}}_m^j(t) \int_{\mathbb{R}^2} z_n \partial_{z_m}\omega^j(\mathbf{z}, t)\mathrm{d}\mathbf{z} = -m^j \dot{\mathbf{x}}_n^j(t), \tag{21}$$

where $m^j = \int_{\mathbb{R}^2} \omega^j(\mathbf{z}, t)\mathrm{d}\mathbf{z}$ and $n = 1, 2$.

Remark 2.1. *Note that* (15) *does preserve the total integral ("mass") of the solution, so this definition of* m^j *is consistent with* (9).

Finally, recalling that the velocity field is incompressible, we can rewrite

$$\left(\sum_{\ell=1}^{N}\mathbf{u}^{\ell}(\mathbf{z}+\mathbf{x}^{j}(t)-\mathbf{x}^{\ell}(t),t)\right)\cdot\nabla\omega^{j}(\mathbf{z},t)=\nabla\cdot\left(\sum_{\ell=1}^{N}\mathbf{u}^{\ell}(\mathbf{z}+\mathbf{x}^{j}(t)-\mathbf{x}^{\ell}(t),t)\omega^{j}(\mathbf{z},t)\right),$$

and so, again considering the nth component and integrating by parts, we have

$$\begin{aligned}\int_{\mathbb{R}^2} z_n\left(\sum_{\ell=1}^{N}\mathbf{u}^{\ell}(\mathbf{z}+\mathbf{x}^{j}(t)-\mathbf{x}^{\ell}(t),t)\right)\cdot\nabla\omega^{j}(\mathbf{z},t)\mathrm{d}\mathbf{z}\\ =-\int_{\mathbb{R}^2}\left(\sum_{\ell=1}^{N}\mathbf{u}_n^{\ell}(\mathbf{z}+\mathbf{x}^{j}(t)-\mathbf{x}^{\ell}(t),t)\right)\omega^{j}(\mathbf{z},t)\mathrm{d}\mathbf{z}.\end{aligned}\tag{22}$$

Thus, if we combine (20), (21), and (22), we see that (19) reduces to the system of ODEs for the centers of the vorticity distributions,

$$\frac{\mathrm{d}\mathbf{x}_n^{j}}{\mathrm{d}t}(t)=\frac{1}{m^{j}}\sum_{\ell=1}^{N}\int_{\mathbb{R}^2}\left(\mathbf{u}_n^{\ell}(\mathbf{z}+\mathbf{x}^{j}(t)-\mathbf{x}^{\ell}(t),t)\omega^{j}(\mathbf{z},t)\right)\mathrm{d}\mathbf{z},\tag{23}$$

supplemented by the initial conditions (10), while the N components of the vorticity evolve according to the PDEs (15) with initial conditions

$$\omega^{j}(\mathbf{z},0)=\omega_0^{j}(\mathbf{z}+\mathbf{x}_0^{j}),\tag{24}$$

obtained by combining (8) and (12).

Remark 2.2. *Consider* (23) *in the limit in which the components* ω^j *are all point vortices, i.e.,* $\omega^{j}(\mathbf{z},t)=m^{j}\delta(\mathbf{z})$*, with* $\delta(\mathbf{z})$ *the Dirac-delta function. Recall that the velocity field associated with such a point vortex is*

$$U^{j}(\mathbf{z},t)=m^{j}\frac{\mathbf{n}\times\mathbf{z}}{||\mathbf{z}||^2},$$

where $\mathbf{n}$ *is the unit normal vector to the plane. Then if we ignore the (singular) term with* $\ell=j$ *in the sum on the right-hand side of* (23)*, we find*

$$\frac{\mathrm{d}\mathbf{x}^{j}}{\mathrm{d}t}(t)=\sum_{\ell=1;\,\ell\neq j}^{N}m^{\ell}\frac{\mathbf{n}\times(\mathbf{x}^{j}(t)-\mathbf{x}^{\ell}(t))}{||\mathbf{x}^{j}(t)-\mathbf{x}^{\ell}(t)||^2}.\tag{25}$$

These, of course, are just Helmholtz–Kirchhoff equations for the inviscid motion of a system of point vortices. Thus, our expansion can be regarded as a generalization of this approximation, which allows for both nonzero viscosity and vortices of finite size. To justify omitting the term with $\ell=j$ *on the right-hand side of* (25) *we note that if we approximate the delta function with a narrow Gaussian vorticity distribution, and* $\mathbf{u}^j$ *by the corresponding velocity field, this term will vanish by symmetry.*

3. The moment expansion: Case of a single center. In this section we introduce another idea—an expansion of the vorticity in terms of Hermite functions. Then, in the next section we will combine the Hermite expansion with the multivortex expansion of the previous section.

The moment expansion is an expansion of the solution of the vorticity equation in terms of Hermite functions. Define

$$\phi_{00}(\mathbf{x},t;\lambda)=\frac{1}{\pi\lambda^2}e^{-|\mathbf{x}|^2/\lambda^2}, \tag{26}$$

where $\lambda^2=\lambda_0^2+4\nu t$. Three simple facts that we will use repeatedly are the following:

(i) $\partial_t\phi_{00}=\nu\Delta\phi_{00}$.

(ii) $\int_{\mathbb{R}^2}\phi_{00}(\mathbf{x},t;\lambda)\mathrm{d}\mathbf{x}=1$ for all $t\geq 0$.

(iii) Finally, and crucially for what follows, the vorticity function $\omega(\mathbf{x},t)=\alpha\phi_{00}(\mathbf{x},t)$ is an exact solution (called the Oseen, or Lamb, vortex) of the two-dimensional vorticity equation for all values of α.

Note that we will often suppress the dependence of ϕ_{00} on λ when there is no fear of confusion.

We now define the Hermite functions of order (k_1,k_2) by

$$\phi_{k_1,k_2}(\mathbf{x},t;\lambda)=D_{x_1}^{k_1}D_{x_2}^{k_2}\phi_{00}(\mathbf{x},t;\lambda) \tag{27}$$

and the corresponding moment expansion of a function by

$$\omega(\mathbf{x},t)=\sum_{k_1,k_2=0}^{\infty}M[k_1,k_2;t]\phi_{k_1,k_2}(\mathbf{x},t;\lambda). \tag{28}$$

We define the Hermite polynomials via their generating function:

$$H_{n_1,n_2}(\mathbf{z};\lambda)=\left(D_{t_1}^{n_1}D_{t_2}^{n_2}e^{\left(\frac{2\mathbf{t}\cdot\mathbf{z}-|\mathbf{t}|^2}{\lambda^2}\right)}\right)\Bigg|_{t=0}. \tag{29}$$

Note that the "standard" Hermite polynomials correspond to taking $\lambda=1$.

Then using the standard orthogonality relationship for the Hermite polynomials,

$$\int_{\mathbb{R}^2}H_{n_1,n_2}(\mathbf{z};1)H_{m_1,m_2}(\mathbf{z};1)e^{-z^2}\mathrm{d}\mathbf{z}=\pi 2^{n_1+n_2}(n_1!)(n_2!)\delta_{n1,m1}\delta_{n_2,m_2}, \tag{30}$$

we see that the coefficients in the expansion (28) are defined by the projection operators:

$$M[k_1,k_2;t]=(P_{k_1,k_2}\omega)(t)=\frac{(-1)^{k_1+k_2}\lambda^{2(k_1+k_2)}}{2^{k_1+k_2}(k_1!)(k_2!)}\int_{\mathbb{R}^2}H_{k_1,k_2}(\mathbf{z};\lambda)\omega(\mathbf{z},t)\mathrm{d}\mathbf{z}. \tag{31}$$

Note that if the function $\omega(\mathbf{x},t)$ in (28) is the vorticity field of some fluid, the linearity of the Biot–Savart law implies that we can expand the associated velocity field as

$$\mathbf{V}(\mathbf{x},t)=\sum_{k_1,k_2=0}^{\infty}M[k_1,k_2;t]\mathbf{V}_{k_1,k_2}(\mathbf{x},t;\lambda), \tag{32}$$

where

$$\mathbf{V}_{k_1,k_2}(\mathbf{x},t;\lambda) = D_{x_1}^{k_1} D_{x_2}^{k_2} \mathbf{V}_{00}(\mathbf{x},t;\lambda) \tag{33}$$

and $\mathbf{V}_{00}(\mathbf{x},t;\lambda)$ is the velocity field associated with the Gaussian vorticity distribution ϕ_{00}—explicitly we have

$$\mathbf{V}_{00}(\mathbf{x},t;\lambda) = \frac{1}{2\pi}\frac{(-x_2,x_1)}{|\mathbf{x}|^2}\left(1-e^{-|\mathbf{x}|^2/\lambda^2}\right). \tag{34}$$

3.1. Convergence of the moment expansion. In this subsection we derive a criterion for the convergence of the moment expansion derived above, and we show that if this criterion is satisfied for $t=0$, then it is satisfied for all subsequent times $t>0$.

Our convergence criterion is based on the observation that the Hermite functions $\phi_{k_1,k_2}(\mathbf{x},t;\lambda)$ are, for any value of t, the eigenfunctions of the linear operator

$$\mathcal{L}^\lambda \psi = \frac{1}{4}\lambda^2 \Delta\psi + \frac{1}{2}\nabla\cdot(\mathbf{x}\psi). \tag{35}$$

This fact can be verified by direct computation and is related to the fact that $\mathcal{L}^\lambda$ can be transformed into the Hamiltonian quantum mechanical harmonic oscillator.

The Gaussian function ϕ_{00} plays a crucial role in the convergence proof, and its dependence on the parameter λ is particularly important in this discussion, so for this subsection only we will define

$$\phi_{00}(\mathbf{x},t;\lambda) = \Phi_\lambda(\mathbf{x},t)$$

to emphasize this dependence.

If one now proceeds as in Lemma 4.7 of [4], one can prove the following.

Proposition 3.1. *The operator $\mathcal{L}^\lambda$ is self-adjoint in the Hilbert space*

$$X^\lambda = \{f \in L^2(\mathbb{R}^2) \mid \Phi_\lambda^{-1/2} f \in L^2(\mathbb{R}^2)\}$$

with inner product $(f,g)_\lambda = \int_{\mathbb{R}^2} \Phi_\lambda^{-1}\overline{f}g\mathrm{d}\mathbf{x}$.

An immediate corollary of this proposition and the general theory of self-adjoint operators is the following.

Corollary 3.2. *The eigenfunctions of $\mathcal{L}^\lambda$ form a complete orthogonal set in the Hilbert space X^λ.*

As a corollary of this result and the observation that the eigenfunctions of $\mathcal{L}^\lambda$ are precisely our Hermite functions $\{\phi_{k_1,k_2}\}$, we have finally the next claim.

Proposition 3.3. *Suppose that*

$$\|f\|_\lambda^2 = \int_{\mathbb{R}^2} \Phi_\lambda^{-1}(\mathbf{x})|f(\mathbf{x})|^2\mathrm{d}\mathbf{x} < \infty;$$

then the expansion

$$f(\mathbf{x}) = \sum_{k_1,k_2} M[k_1,k_2]\phi_{k_1,k_2}(\mathbf{x})$$

converges with respect to the norm on the Hilbert space X^λ.

Thus, the following criterion guarantees that the expansion (28) for the vorticity converges:

$$\int_{\mathbb{R}^2} \Phi_\lambda^{-1}(\mathbf{x})(\omega(\mathbf{x},t))^2 \mathrm{d}\mathbf{x} < \infty. \tag{36}$$

The main result of this subsection is the following theorem, which proves that if our initial vorticity distribution satisfies (36) for some $\lambda = \lambda_0$, then the solution of the vorticity equation with that initial condition will satisfy (36) for all time t with $\lambda = \sqrt{4\nu t + \lambda_0^2}$, and hence as a corollary if the initial vorticity distribution satisfies (36), then our moment expansion converges for all times t.

Theorem 3.4. *Define the weighted enstrophy function*

$$\mathcal{E}(t) = \int_{\mathbb{R}^2} \Phi_\lambda^{-1}(\mathbf{x})(\omega(\mathbf{x},t))^2 \mathrm{d}\mathbf{x}.$$

If the initial vorticity distribution ω_0 *is such that* $\mathcal{E}(0) < \infty$ *for some* λ_0, *and* ω_0 *is bounded (in the* L^∞ *norm), then* $\mathcal{E}(t)$ *is finite for all times* $t > 0$.

Proof. The idea of the proof is to derive a differential inequality for $\mathcal{E}(t)$ which guarantees that if $\mathcal{E}(0)$ is finite, then $\mathcal{E}(t)$ will be finite for all t. Differentiating $\mathcal{E}(t)$, we obtain

$$\begin{aligned}
\frac{\mathrm{d}\mathcal{E}}{\mathrm{d}t}(t) &= \frac{4\nu}{\lambda^2}\mathcal{E}(t) - \frac{4\nu}{\lambda^4}\int_{\mathbb{R}^2} |\mathbf{x}|^2 \Phi_\lambda^{-1}(\mathbf{x})(\omega(\mathbf{x},t))^2 \mathrm{d}\mathbf{x} \\
&\quad + 2\int_{\mathbb{R}^2} \Phi_\lambda^{-1}(\mathbf{x})\omega(\mathbf{x},t)\partial_t\omega(\mathbf{x},t)\mathrm{d}\mathbf{x} \\
&= \frac{4\nu}{\lambda^2}\mathcal{E}(t) - \frac{4\nu}{\lambda^4}\int_{\mathbb{R}^2} |\mathbf{x}|^2 \Phi_\lambda^{-1}(\mathbf{x})(\omega(\mathbf{x},t))^2 \mathrm{d}\mathbf{x} \\
&\quad + 2\int_{\mathbb{R}^2} \Phi_\lambda^{-1}(\mathbf{x})\omega(\mathbf{x},t)\left(\nu\Delta\omega - \mathbf{u}\cdot\nabla\omega\right)\mathrm{d}\mathbf{x}.
\end{aligned} \tag{37}$$

We now consider the last term in (37). First note that upon integration by parts we have

$$2\int_{\mathbb{R}^2} \Phi_\lambda^{-1}(\mathbf{x})\omega(\mathbf{x},t)\left(\nu\Delta\omega(\mathbf{x},t)\right)\mathrm{d}\mathbf{x} = -2\nu\int_{\mathbb{R}^2} \Phi_\lambda^{-1}(\mathbf{x})\left(|\nabla\omega|^2 + \frac{2}{\lambda^2}\omega\mathbf{x}\cdot\nabla\omega\right)\mathrm{d}\mathbf{x}. \tag{38}$$

The right-hand side of the expression in (38) can again be broken up into two pieces, and the second can be bounded by

$$2\nu\int_{\mathbb{R}^2} \Phi_\lambda^{-1}(\mathbf{x})\left(\frac{2}{\lambda^2}\omega\mathbf{x}\cdot\nabla\omega\right)\mathrm{d}\mathbf{x} \le \nu\int_{\mathbb{R}^2} \Phi_\lambda^{-1}(\mathbf{x})|\nabla\omega|^2\mathrm{d}\mathbf{x} + \frac{4\nu}{\lambda^4}\int_{\mathbb{R}^2} \Phi_\lambda^{-1}\left(\mathbf{x}^2\omega^2\right)\mathrm{d}\mathbf{x}. \tag{39}$$

Finally, we bound the last term in (37), which comes from the nonlinear term in the vorticity equation. In this estimate we use the fact (see [3, Lemma 2.1]) that the L^∞ norm of the velocity field $\mathbf{u}$ can be bounded by a constant times the sum of the L^1 and L^∞ norms of the vorticity field—i.e., by $C(\|\omega\|_{L^1(\mathbb{R}^2)} + \|\omega\|_{L^\infty(\mathbb{R}^2)})$. This observation, combined with the fact that $\|\omega(\cdot,t)\|_{L^p(\mathbb{R}^2)} \le \|\omega_0\|_{L^p(\mathbb{R}^2)}$, which is a consequence of the maximum principle, implies that

$$\|\mathbf{u}(\cdot,t)\|_{L^\infty(\mathbb{R}^2)} \le C(\|\omega_0\|_{L^1(\mathbb{R}^2)} + \|\omega_0\|_{L^\infty(\mathbb{R}^2)})$$

and hence that

$$2\int_{\mathbb{R}^2}\Phi_\lambda^{-1}(\mathbf{x})\omega(\mathbf{x},t)\,(\mathbf{u}\cdot\nabla\omega)\,\mathrm{d}\mathbf{x}\le 2C(\omega_0)\int_{\mathbb{R}^2}\Phi_\lambda^{-1}(\mathbf{x})|\omega(\mathbf{x},t)||\nabla\omega|\mathrm{d}\mathbf{x}$$
$$\le\frac{4C(\omega_0)}{\nu}\int_{\mathbb{R}^2}\Phi_\lambda^{-1}(\mathbf{x})(\omega(\mathbf{x},t))^2\mathrm{d}\mathbf{x}+\nu\int_{\mathbb{R}^2}\Phi_\lambda^{-1}(\mathbf{x})|\nabla\omega|^2\mathrm{d}\mathbf{x}. \tag{40}$$

If we now combine the inequalities in (38), (39), and (40) with the expression for $\frac{d\mathcal{E}}{\mathrm{d}t}$ in (37), we obtain

$$\frac{\mathrm{d}\mathcal{E}}{\mathrm{d}t}(t)\le\left(\frac{4C(\omega_0)}{\nu}+\frac{4\nu}{\lambda^2}\right)\mathcal{E}(t), \tag{41}$$

from which we see immediately that if $\mathcal{E}(0)$ is bounded, $\mathcal{E}(t)$ remains bounded for all time. ■

3.2. Differential equations for the moments. Assuming that the function $\omega(\mathbf{z},t)$ is a solution of (7), we can derive differential equations satisfied by the moments $M[k_1,k_2,t]$ in (28). Surprisingly the expressions for the coefficients in these expansions are quite simple and explicit.

If we differentiate (28) and assume that ω is a solution of the two-dimensional vorticity equation, we obtain

$$\partial_t\omega=\sum_{k_1,k_2=0}^{\infty}\frac{\mathrm{d}M[k_1,k_2;t]}{\mathrm{d}t}\phi_{k_1,k_2}(\mathbf{x},t;\lambda)+\sum_{k_1,k_2=0}^{\infty}M[k_1,k_2;t]\partial_t\phi_{k_1,k_2}(\mathbf{x},t;\lambda)$$
$$=\sum_{k_1,k_2=0}^{\infty}M[k_1,k_2;t]\,(\nu\Delta\phi_{k_1,k_2}(\mathbf{x},t;\lambda)) \tag{42}$$
$$-\left(\sum_{\ell_1,\ell_2=0}^{\infty}M[\ell_1,\ell_2;t]\mathbf{V}_{\ell_1,\ell_2}(\mathbf{x},t;\lambda)\right)\cdot\nabla\left(\sum_{k_1,k_2=0}^{\infty}M[k_1,k_2;t]\phi_{k_1,k_2}(\mathbf{x},t;\lambda)\right).$$

From the first of the "simple facts" we stated about ϕ_{00}, we see that the last term on the first line cancels the middle line, and hence if we apply the projection operators defined in (31), we are left with the system of ODEs for the moments,

$$\frac{\mathrm{d}M[k_1,k_2;t]}{\mathrm{d}t}=-P_{k_1,k_2}\left[\left(\sum_{\ell_1,\ell_2=0}^{\infty}M[\ell_1,\ell_2;t]\mathbf{V}_{\ell_1,\ell_2}(\mathbf{x},t;\lambda)\right)\right.$$
$$\left.\cdot\nabla\left(\sum_{m_1,m_2=0}^{\infty}M[m_1,m_2;t]\phi_{m_1,m_2}(\mathbf{x},t;\lambda)\right)\right]. \tag{43}$$

The somewhat surprising fact which, in our opinion, makes the preceding straightforward calculations interesting is that the projection on the right-hand side of (43) can be computed explicitly in terms of the derivatives of a relatively simple function.

We now explain how this is done. First recall that

$$\phi_{m_1,m_2}(\mathbf{x},t;\lambda) = D_{x_1}^{m_1} D_{x_2}^{m_2} \phi_{00}(\mathbf{x},\lambda), \qquad \mathbf{V}_{\ell_1,\ell_2}(\mathbf{x},t;\lambda) = D_{x_1}^{\ell_1} D_{x_2}^{\ell_2} \mathbf{V}_{00}(\mathbf{x},\lambda). \tag{44}$$

In order to avoid confusing the two sets of derivatives we will rewrite these formulas as

$$\begin{aligned} \phi_{m_1,m_2}(\mathbf{x},t;\lambda) &= (D_{a_1}^{m_1} D_{a_2}^{m_2} \phi_{00}(\mathbf{x}+\mathbf{a},\lambda))|_{\mathbf{a}=0}, \\ \mathbf{V}_{\ell_1,\ell_2}(\mathbf{x},t;\lambda) &= (D_{b_1}^{\ell_1} D_{b_2}^{\ell_2} \mathbf{V}_{00}(\mathbf{x}+\mathbf{b},\lambda))|_{\mathbf{b}=0}. \end{aligned} \tag{45}$$

Inserting these formulas into the right-hand side of (43) and using the formula for the projection operator P_{k_1,k_2} in terms of the integration against a Hermite polynomial, we obtain

$$\begin{aligned} \frac{\mathrm{d}M}{\mathrm{d}t}[k_1,k_2,t] &= -\frac{(-1)^{(k_1+k_2)}\lambda^{2(k_1+k_2)}}{2^{k_1+k_2}(k_1!)(k_2!)} \sum_{\ell_1,\ell_2} \sum_{m_1,m_2} M[\ell_1,\ell_2,t]M[m_1,m_2,t] \\ &\quad \times \int_{\mathbb{R}^2} H_{k_1,k_2}(\mathbf{x})(D_{x_1}^{m_1} D_{x_2}^{m_2} \mathbf{V}_{00}(\mathbf{x};\lambda)) \cdot \nabla_{\mathbf{x}} (D_{x_1}^{\ell_1} D_{x_2}^{\ell_2} \phi_{00}(\mathbf{x};\lambda))\mathrm{d}\mathbf{x} \\ &= -\frac{(-1)^{(k_1+k_2)}\lambda^{2(k_1+k_2)}}{2^{k_1+k_2}(k_1!)(k_2!)} \sum_{\ell_1,\ell_2} \sum_{m_1,m_2} M[\ell_1,\ell_2,t]M[m_1,m_2,t] \\ &\quad \times D_{t_1}^{k_1} D_{t_2}^{k_2} D_{b_1}^{m_1} D_{b_2}^{m_2} D_{a_1}^{\ell_1} D_{a_2}^{\ell_2} \nabla_{\mathbf{a}} \\ &\quad \cdot \left(\int_{\mathbb{R}^2} e^{\frac{(-t_1^2-t_2^2+2t_1x_1+2t_2x_2)}{\lambda^2}} \mathbf{V}_{00}(\mathbf{x}+\mathbf{b};\lambda)\phi_{00}(\mathbf{x}+\mathbf{a};\lambda)\mathrm{d}\mathbf{x} \right) \Big|_{\mathbf{t}=0,\,\mathbf{a}=0,\,\mathbf{b}=0}. \end{aligned} \tag{46}$$

The last equality in this expression results from rewriting the Hermite polynomial H_{k_1,k_2} in terms of its generating function.

The last step in deriving the equations for the moments is to evaluate the integral in the last line of (46). The key step in this evaluation is to recall that for these incompressible flows the velocity field can be written in terms of the derivatives of the stream function and that the Laplacian of the stream function is minus the vorticity. Thus, we can write

$$\mathbf{V}_{00}(\mathbf{x}+\mathbf{b};\lambda) = -\nabla_{\mathbf{b}}^{*}(\Delta_{\mathbf{b}})^{-1}\phi_{00}(\mathbf{x}+\mathbf{b}), \tag{47}$$

where $\nabla_{\mathbf{b}}^{*} f = (\partial_{x_2} f, -\partial_{x_1} f)$.

Inserting this into the integral in (46), we find

$$\begin{aligned} &\int_{\mathbb{R}^2} e^{\frac{(-t_1^2-t_2^2+2t_1x_1+2t_2x_2)}{\lambda^2}} \mathbf{V}_{00}(\mathbf{x}+\mathbf{b};\lambda)\phi_{00}(\mathbf{x}+\mathbf{a};\lambda)\mathrm{d}\mathbf{x} \\ &\qquad = -\nabla_{\mathbf{b}}^{*}(\Delta_{\mathbf{b}})^{-1} \int_{\mathbb{R}^2} e^{\frac{(-t_1^2-t_2^2+2t_1x_1+2t_2x_2)}{\lambda^2}} \phi_{00}(\mathbf{x}+\mathbf{b};\lambda)\phi_{00}(\mathbf{x}+\mathbf{a};\lambda)\mathrm{d}\mathbf{x}. \end{aligned} \tag{48}$$

Now note that all three factors in the integrand are Gaussians, and thus the integral can be evaluated explicitly, and we find

$$\begin{aligned} &\int_{\mathbb{R}^2} e^{\frac{(-t_1^2-t_2^2+2t_1x_1+2t_2x_2)}{\lambda^2}} \phi_{00}(\mathbf{x}+\mathbf{b};\lambda)\phi_{00}(\mathbf{x}+\mathbf{a};\lambda)\mathrm{d}\mathbf{x} \\ &\qquad = \frac{1}{2\pi\lambda^2} e^{-\frac{1}{2\lambda^2}\left(a_1^2+a_2^2-2a_1b_1+b_1^2-2a_2b_2+b_2^2+2a_1t_1+2b_1t_1+t_1^2+2a_2t_2+2b_2t_2+t_2^2\right)}. \end{aligned} \tag{49}$$

We next compute the expression

$$-\nabla^*_{\mathbf{b}}(\Delta_{\mathbf{b}})^{-1}\frac{1}{2\pi\lambda^2}e^{-\frac{1}{2\lambda^2}\left(a_1^2+a_2^2-2a_1b_1+b_1^2-2a_2b_2+b_2^2+2a_1t_1+2b_1t_1+t_1^2+2a_2t_2+2b_2t_2+t_2^2\right)}. \tag{50}$$

Recall that, given a vorticity field ω, $-(\Delta)^{-1}\omega$ is the associated stream function and $-\nabla^*(\Delta)^{-1}\omega$ the velocity field associated with ω. Since the inverse Laplacian and derivatives in (50) act only on the $\mathbf{b}$-dependent parts of the expression, we need to evaluate

$$\begin{aligned}-\nabla^*_{\mathbf{b}}(\Delta_{\mathbf{b}})^{-1}\frac{1}{2\pi\lambda^2}e^{-\frac{1}{2\lambda^2}\left(-2a_1b_1+b_1^2-2a_2b_2+b_2^2+2b_1t_1+2b_2t_2\right)}\\ = -e^{\frac{1}{2\lambda^2}\left((t_1-a_1)^2+(t_2-a_2)^2\right)}\nabla^*_{\mathbf{b}}(\Delta_{\mathbf{b}})^{-1}\frac{1}{2\pi\lambda^2}e^{-\frac{1}{2\lambda^2}\left((b_1+(t_1-a_1))^2+(b_2+(t_2-a_2))^2\right)}.\end{aligned} \tag{51}$$

However,

$$-\nabla^*_{\mathbf{b}}(\Delta_{\mathbf{b}})^{-1}\frac{1}{\pi\lambda^2}e^{-\frac{1}{2\lambda^2}\left((b_1+(t_1-a_1))^2+(b_2+(t_2-a_2))^2\right)}$$

is just the velocity field associated with a Gaussian vorticity distribution (i.e., an Oseen vortex) centered at the point $-((t_1-a_1),(t_2-a_2))$ which we know explicitly. Hence, the expression on the right-hand side of (48) has the explicit representation

$$\begin{aligned}&\frac{1}{2\pi}e^{-\frac{2}{\lambda^2}(t_1a_1+t_2a_2)}\frac{(-(b_2+(t_2-a_2)),(b_1+(t_1-a_1)))}{(b_1+(t_1-a_1))^2+(b_2+(t_2-a_2))^2}\\ &\qquad\times\left(1-e^{-\frac{1}{2\lambda^2}\left((b_1+(t_1-a_1))^2+(b_2+(t_2-a_2))^2\right)}\right).\end{aligned} \tag{52}$$

If we now return to (46), we see that in order to compute the coefficients in the moment equations we need to evaluate the divergence of this last expression with respect to $\mathbf{a}$, which gives

$$\begin{aligned}&K(a_1,a_2,b_1,b_2,t_1,t_2;\lambda)=\frac{1}{\pi\lambda^2}e^{-\frac{2}{\lambda^2}(t_1a_1+t_2a_2)}\\ &\times\left(\frac{-a_2\,t_1+b_2\,t_1+a_1\,t_2-b_1\,t_2}{((b_1+(t_1-a_1))^2+(b_2+(t_2-a_2))^2)}\right)\left(1-e^{-\frac{1}{2\lambda^2}\left((b_1+(t_1-a_1))^2+(b_2+(t_2-a_2))^2\right)}\right).\end{aligned}$$

Returning to (46), we finally conclude that

$$\frac{\mathrm{d}M}{\mathrm{d}t}[k_1,k_2,t]=-\frac{(-1)^{(k_1+k_2)}\lambda^{2(k_1+k_2)}}{2^{k_1+k_2}(k_1!)(k_2!)}\sum_{\ell_1,\ell_2}\sum_{m_1,m_2}\Gamma[k_1,k_2;\ell_1,\ell_2,m_1,m_2;\lambda]M[\ell_1,\ell_2,t]M[m_1,m_2,t], \tag{53}$$

where

$$\begin{aligned}&\Gamma[k_1,k_2;\ell_1,\ell_2,m_1,m_2;\lambda]\\ &\qquad=D_{t_1}^{k_1}D_{t_2}^{k_2}D_{b_1}^{m_1}D_{b_2}^{m_2}D_{a_1}^{\ell_1}D_{a_2}^{\ell_2}K(a_1,a_2,b_1,b_2,t_1,t_2;\lambda)|_{t=0,\,a=0,\,b=0}.\end{aligned} \tag{54}$$

Thus, we have succeeded in rewriting the two-dimensional vorticity equation as a system of ODEs with simple, quadratic nonlinear terms whose coefficients can be evaluated in terms of derivatives of a single explicit function. Furthermore, we have given a sufficient condition on the initial vorticity distribution to guarantee that the expansion of the vorticity generated by the solution of these ODEs converges for all time.

4. The moment expansion for several vortex centers. In this section we extend the Hermite moment expansion of the previous section to the case in which there are two or more centers of vorticity by combining this expansion with the multivortex representation of section 2. For simplicity of exposition we limit the discussion here to the case of two vortices, but the expansion can be extended to any finite number of vortices.

The basic idea is just to consider (15) for the evolution of each vortex and then expand each of the functions ω^j in Hermite moments as in the previous section. Thus, we define

$$\omega^j(\mathbf{z},t) = \sum_{k_1,k_2=0}^{\infty} M^j[k_1,k_2;t]\phi_{k_1,k_2}(\mathbf{z},t;\lambda) \tag{55}$$

for $j = 1,2$. We make a similar expansion for the velocity field in terms of the functions $\mathbf{V}_{\ell_1,\ell_2}$ and insert the expansions into (15). Letting $\mathbf{z} = \mathbf{x} - \mathbf{x}^j(t)$ and recalling that $\partial_t\phi_{k_1,k_1} = \nu\Delta\phi_{k_1,k_2}$, we obtain

$$\begin{aligned}\frac{\mathrm{d}M^j[k_1,k_2;t]}{\mathrm{d}t} = &-P_{k_1,k_2}\left[\left(\sum_{j'=1}^{2}\sum_{\ell_1,\ell_2=0}^{\infty} M^{j'}[\ell_1,\ell_2;t]\mathbf{V}_{\ell_1,\ell_2}(\mathbf{z}+\mathbf{s}_{j,j'},t;\lambda)\right)\right.\\ &\left.\cdot\nabla\left(\sum_{m_1,m_2=0}^{\infty} M^j[m_1,m_2;t]\phi_{m_1,m_2}(\mathbf{z},t;\lambda)\right)\right]\\ &+P_{k_1,k_2}\left[\dot{\mathbf{x}}^j(t)\cdot\nabla\left(\sum_{m_1,m_2=0}^{\infty} M^j[m_1,m_2;t]\phi_{m_1,m_2}(\mathbf{z},t;\lambda)\right)\right],\end{aligned} \tag{56}$$

where $\mathbf{s}_{j,j'} = \mathbf{x}^{j'}(t) - \mathbf{x}^j(t)$. For the first projection term we proceed as in the previous section, and using (46), we finally conclude that the first term is equal to

$$\begin{aligned}-\frac{(-1)^{(k_1+k_2)}\lambda^{2(k_1+k_2)}}{2^{k_1+k_2}(k_1!)(k_2!)}&\sum_{j'=1}^{2}\sum_{\ell_1,\ell_2}\sum_{m_1,m_2}\Gamma^{j,j'}[k_1,k_2;\ell_1,\ell_2,m_1,m_2;\mathbf{s}_{j,j'},\lambda]\\ &\times M^j[\ell_1,\ell_2,t]M^{j'}[m_1,m_2,t],\end{aligned}$$

where

$$\begin{aligned}&\Gamma^{j,j'}[k_1,k_2;\ell_1,\ell_2,m_1,m_2;\mathbf{s}_{j,j'},\lambda]\\ &\quad= D_{t_1}^{k_1}D_{t_2}^{k_2}D_{b_1}^{m_1}D_{b_2}^{m_2}D_{a_1}^{\ell_1}D_{a_2}^{\ell_2}K^{multi}(a_1,a_2,b_1,b_2,t_1,t_2;\mathbf{s}_{j,j'},\lambda)|_{\mathbf{t}=0,\,\mathbf{a}=0,\,\mathbf{b}=0}\end{aligned} \tag{57}$$

and

$$\begin{aligned}&K^{multi}(a_1,a_2,b_1,b_2,t_1,t_2;\mathbf{s},\lambda)\\ &\quad=\nabla_{\mathbf{a}}\cdot\left(\int_{\mathbb{R}^2} e^{\frac{(-t_1^2-t_2^2+2t_1x_1+2t_2x_2)}{\lambda^2}}\mathbf{V}_{00}(\mathbf{x}-\mathbf{s}+\mathbf{b};\lambda)\phi_{00}(\mathbf{x}+\mathbf{a};\lambda)\mathrm{d}\mathbf{x}\right).\end{aligned} \tag{58}$$

Remark 4.1. *Note that, comparing* (57) *with the calculation leading up to* (53)*, we see that* K^{multi} *can be written in terms of the expression for* K *via the simple formula*

$$K^{multi}(a_1, a_2, b_1, b_2, t_1, t_2; s_1, s_2, \lambda) = K(a_1, a_2, b_1 - s_1, b_2 - s_2, t_1, t_2; \lambda).$$

The second projection term in (56) can be reduced to

$$\dot{x}_1^j(t) M^j[k_1 - 1, k_2, t] + \dot{x}_2^j(t) M^j[k_1, k_2 - 1, t]; \tag{59}$$

thus we can combine the above equations to write

$$\begin{aligned} \frac{\mathrm{d}M^j}{\mathrm{d}t}[k_1, k_2, t] = &-\frac{(-1)^{(k_1+k_2)}\lambda^{2(k_1+k_2)}}{2^{k_1+k_2}(k_1!)(k_2!)} \sum_{j'=1}^{2} \sum_{\ell_1,\ell_2} \sum_{m_1,m_2} \Gamma^{j,j'}[k_1, k_2; \ell_1, \ell_2, m_1, m_2; \mathbf{s}_{j,j'}, \lambda] \\ &\times M^j[\ell_1, \ell_2, t] M^{j'}[m_1, m_2, t] \\ &+ \dot{x}_1^j(t) M^j[k_1 - 1, k_2, t] + \dot{x}_2^j(t) M^j[k_1, k_2 - 1, t]. \end{aligned} \tag{60}$$

We now derive a similar expansion for the evolution of the centers of each vortex. We begin with (23),

$$\frac{\mathrm{d}\mathbf{x}^j}{\mathrm{d}t}(t) = \frac{1}{m^j} \sum_{j'=1}^{N} \int_{\mathbb{R}^2} \left(\mathbf{u}^\ell(\mathbf{z} + \mathbf{x}^j(t) - \mathbf{x}^{j'}(t), t) \omega^j(\mathbf{z}, t) \right) \mathrm{d}\mathbf{z}.$$

(Recall that this is really a pair of equations, one for each component of $\mathbf{x}^j$.) Now insert the moment expansion of $\mathbf{u}^{j'}$ and ω^j into this expression, to obtain

$$\frac{\mathrm{d}\mathbf{x}^j}{\mathrm{d}t}(t) = \sum_{j'=1}^{N} \sum_{\ell_1,\ell_2} \sum_{m_1,m_2} M^j[\ell_1, \ell_2, t] M^{j'}[m_1, m_2, t] \int_{\mathbb{R}^2} \mathbf{V}_{\ell_1,\ell_2}(\mathbf{z} - \mathbf{s}_{j,j'}, t) \phi_{\ell_1,\ell_2}(\mathbf{z}, t) \mathrm{d}\mathbf{z}, \tag{61}$$

where, as before, $\mathbf{s}_{j,j'} = \mathbf{x}^{j'}(t) - \mathbf{x}^j(t)$. The integral of the velocity and vorticity can be evaluated just as in the preceding section, and we find

$$\frac{\mathrm{d}\mathbf{x}^j}{\mathrm{d}t}(t) = \sum_{j'=1}^{N} \sum_{\ell_1,\ell_2} \sum_{m_1,m_2} \Xi^{j,j'}[\ell_1, \ell_2, m_1, m_2; \mathbf{s}_{j,j'}, \lambda] M^j[\ell_1, \ell_2, t] M^{j'}[m_1, m_2, t], \tag{62}$$

where in this case the coefficients $\Xi^{j,j'}[\ell_1, \ell_2, m_1, m_2; \mathbf{s}_{j,j'}, \lambda]$ are given by the expression

$$\begin{aligned} &\Xi^{j,j'}[\ell_1, \ell_2, m_1, m_2; s, \lambda] \\ &\quad = \frac{1}{2\pi} D_{b_1}^{m_1} D_{b_2}^{m_2} D_{a_1}^{\ell_1} D_{a_2}^{\ell_2} \left(\frac{(-(b_2 - s_2 - a_2), (b_1 - s_1 - a_1))}{(b_2 - s_2 - a_2)^2 + (b_1 - s_1 - a_1)^2} \right. \\ &\qquad \left. \times \left(1 - e^{-\frac{1}{2\lambda^2}((b_2 - s_2 - a_2)^2 + (b_1 - s_1 - a_1)^2)} \right) \right) \Big|_{\mathbf{a}=\mathbf{0}, \mathbf{b}=\mathbf{0}}. \end{aligned} \tag{63}$$

Remark 4.2. *Equations* (62) *and* (63) *could be expressed in terms of Cartesian tensors.*

4.1. Convergence of multivortex expansion. We note that it is easy to extend our previous result on the convergence of the moment expansion to the multivortex expansion. To do so, we change variables to recenter each vortex at the origin. Thus if we let

$$\omega^j(\mathbf{x}-\mathbf{x}^j(t),t)=w^j(\mathbf{x},t), \qquad \mathbf{u}^j(\mathbf{x}-\mathbf{x}^j(t),t)=\mathbf{v}^j(\mathbf{x},t),$$

then (15) becomes

$$\frac{\partial w^j}{\partial t}(\mathbf{x},t)=\nu\Delta w^j(\mathbf{x},t)-\left(\sum_{\ell=1}^{N}\mathbf{v}^\ell(\mathbf{x}-\mathbf{x}^\ell(t)+\mathbf{x}^j(t),t)\right)\cdot\nabla w^j(\mathbf{x},t),$$

for $j=1,\dots,N$.

We are now ready to state our result.

Theorem 4.3. *Define*

$$\mathcal{E}^j(t)=\int_{\mathcal{R}^2}\Phi_\lambda^{-1}(\mathbf{x})w^j(\mathbf{x},t)^2\mathrm{d}\mathbf{x}.$$

If the initial vorticity distribution $\omega_0\equiv\sum_{\ell=1}^{N}\omega^\ell(\mathbf{x}-\mathbf{x}^\ell(0),0)$ *is such that* $\mathcal{E}^j(0)<\infty$ *for some* λ_0 *and for all* $j=1,2,\dots,N$*, and if* ω_0 *is bounded (in the* L^∞ *norm), then each* $\mathcal{E}^j(t)$ *is finite for all times* $t>0$.

Proof. We use the same idea as in the single vortex case and differentiate $\mathcal{E}^j(t)$:

$$\begin{aligned}\frac{\mathrm{d}\mathcal{E}^i}{\mathrm{d}t}(t)&=\frac{4\nu}{\lambda^2}\mathcal{E}^i(t)-\frac{4\nu}{\lambda^4}\int_{\mathbb{R}^2}|\mathbf{x}|^2\Phi_\lambda^{-1}(\mathbf{x})w^i(\mathbf{x})^2\mathrm{d}\mathbf{x}\\&\quad+2\int_{\mathbb{R}^2}\Phi_\lambda^{-1}(\mathbf{x})w^i(\mathbf{x},t)\partial_t w^i(\mathbf{x},t)\mathrm{d}\mathbf{x}\\&=\frac{4\nu}{\lambda^2}\mathcal{E}^i(t)-\frac{4\nu}{\lambda^4}\int_{\mathbb{R}^2}|\mathbf{x}|^2\Phi_\lambda^{-1}(\mathbf{x})w^i(\mathbf{x})^2\mathrm{d}\mathbf{x}\\&\quad+2\int_{\mathbb{R}^2}\Phi_\lambda^{-1}w^i\left(\nu\Delta w^i-\left(\sum_{\ell=1}^{N}\mathbf{v}^\ell(\mathbf{x}-\mathbf{x}^\ell(t)+\mathbf{x}^j(t),t)\right)\cdot\nabla w^i\right)\mathrm{d}\mathbf{x}.\end{aligned}$$

From here the proof proceeds identically to the proof of Theorem 3.4 except we must examine the nonlinear term that comes from the vorticity equation a bit more closely. In general, it is unknown whether $w^j(\cdot,t)$ satisfies a maximum principle, but since $\left(\sum_{\ell=1}^{N}\mathbf{v}^\ell(\mathbf{x}-\mathbf{x}^\ell(t)+\mathbf{x}^j(t),t)\right)=\mathbf{u}$, the solution to (1), it does satisfy a maximum principle. Hence, as in subsection 3.1, we can bound the L^∞ norm of $\sum_{\ell=1}^{N}\mathbf{v}^\ell(\mathbf{x}-\mathbf{x}^\ell(t)+\mathbf{x}^j(t),t)$ by a constant depending only on the *initial* vorticity distribution. Hence we proceed to bound the integral

$$\int_{\mathbb{R}^2}\Phi_\lambda^{-1}w^i\left(\left(\sum_{\ell=1}^{N}\mathbf{v}^\ell(\mathbf{x}-\mathbf{x}^\ell(t)+\mathbf{x}^j(t),t)\right)\cdot\nabla w^i\right)\mathrm{d}\mathbf{x}\le 2C(\omega_0)\int_{\mathbb{R}^2}\Phi_\lambda^{-1}|w^i||\nabla w^i|\mathrm{d}\mathbf{x},$$

and thus the rest of the bounds are the same. Again, putting everything together, we arrive at

$$\frac{\mathrm{d}\mathcal{E}^i}{\mathrm{d}t}(t)\le\left(\frac{4\nu}{\lambda^2}+\frac{4C(\omega_0)}{\nu}\right)\mathcal{E}^i(t). \tag{64}$$

■

4.2. Interaction of Gaussian vortices. The experimental and numerical work of [11] has shown that widely separated regions of vorticity can be well approximated by Gaussians for long periods of time. This corresponds to truncating our expansions for the vorticity so that they contain only a single term. In this subsection we analyze the equations that result from this truncation. This approximation can be viewed as a generalization of the Helmholtz–Kirchhoff approximation in which we include the effects of vorticity and finite core size to lowest order. We show that the total vorticity of each of the vortices is constant, while the centers of vorticity evolve along either straight lines or circles.

Remark 4.4. *It should be noted here that the effect of truncating our expansion after one term allows for only viscosity as the driving force in vortex merger. Allowing more terms in the expansion introduces convective forces.*

If we start with two Gaussian distributions for our initial vorticity with the same value of λ_0, and truncate the equations of motion for the moments so that all terms containing higher-order moments are omitted, then we can conclude that the two vortices travel along circular or straight line orbits around the "center of vorticity." Moreover, the leading coefficients $M^j[0,0:t]$ of the expansions are constant. To be precise, we let

$$\omega^1(\mathbf{x},t) = M^1[0,0;t]\phi_{0,0}(x - x^1(t), t; \lambda), \tag{65}$$

$$\omega^2(\mathbf{x},t) = M^2[0,0;t]\phi_{0,0}(x - x^2(t), t; \lambda). \tag{66}$$

Let us also write $\mathbf{s}_{i,j} = \mathbf{x}^j(t) - \mathbf{x}^i(t)$; then using (56)–(58), we first calculate the evolution of $M^1(0,0:t)$:

$$\begin{aligned}\frac{\mathrm{d}M^1}{\mathrm{d}t}[0,0;t] &= -M^1[0,0;t]^2 K(0,0,0,0,0,0,\lambda)\\ &\quad - M^1[0,0;t]M^2[0,0;t]K^{multi}(0,0,0,0,0,0,\mathbf{s}_{1,2},\lambda)\\ &= 0,\end{aligned}$$

since $K(0,0,0,0,0,0,\lambda) = 0$ and $K^{multi}(0,0,0,0,0,\mathbf{s}_{1,2},\lambda) = 0$, respectively. The calculation for $M^2[0,0;t]$ is the same. Thus the evolutions of the coefficients of leading order are constant.

Remark 4.5. *The fact that $M^1[0,0;t]$ is constant in time is not a consequence of the truncation of the moment equations to first order. One can show that the equations of* (14) *conserve the zeroth moment of ω^j, independent of any truncation.*

More interesting, though, is the calculation for the evolution of $\mathbf{x}^j(t)$, the centers of these vortices. Again if we denote $\mathbf{x}^j = (x_1^j, x_2^j)$, then the evolution for each component as defined by (23) can be written as

$$\frac{\mathrm{d}x_i^j}{\mathrm{d}t} = \frac{1}{M^j[0,0,t]}\int V_i^k(\mathbf{y} - \mathbf{s}_{1,2}, t)\omega^j(\mathbf{y},t)d\mathbf{y}. \tag{67}$$

Now using (47) to evaluate $V_i^k(\mathbf{y} - \mathbf{s}_{1,2}, t)$, equation (67) yields the following equations for

each x_i^j, $j = 1, 2$, and $i = 1, 2$:

$$
\begin{aligned}
\dot{x}_1^1 &= -\frac{M^2}{2\pi}\frac{\left(e^{-\frac{(x_1^1-x_1^2)^2+(x_2^1-x_2^2)^2}{2\lambda(t)^2}}-1\right)(x_2^1-x_2^2)}{(x_1^1-x_1^2)^2+(x_2^1-x_2^2)^2},\\
\dot{x}_2^1 &= -\frac{M^2}{2\pi}\frac{\left(e^{-\frac{(x_1^1-x_1^2)^2+(x_2^1-x_2^2)^2}{2\lambda(t)^2}}-1\right)(-x_1^1+x_1^2)}{(x_1^1-x_1^2)^2+(x_2^1-x_2^2)^2},\\
\dot{x}_1^2 &= \frac{M^1}{2\pi}\frac{\left(e^{-\frac{(x_1^1-x_1^2)^2+(x_2^1-x_2^2)^2}{2\lambda(t)^2}}-1\right)(x_2^1-x_2^2)}{(x_1^1-x_1^2)^2+(x_2^1-x_2^2)^2},\\
\dot{x}_2^2 &= \frac{M^1}{2\pi}\frac{\left(e^{-\frac{(x_1^1-x_1^2)^2+(x_2^1-x_2^2)^2}{2\lambda(t)^2}}-1\right)(-x_1^1+x_1^2)}{(x_1^1-x_1^2)^2+(x_2^1-x_2^2)^2},
\end{aligned}
\tag{68}
$$

where $M^j \equiv M^j[0, 0, t]$ is constant and represents the total vorticity of the jth vortex.

Remark 4.6. *If the vortices have different $\lambda(t)$ values, say $\lambda_1(t)$ and $\lambda_2(t)$, then one just needs to replace the $2\lambda(t)^2$ with $\lambda_1(t)^2 + \lambda_2(t)^2$ in the exponential to arrive at the correct system.*

We now state the main result of this section.

Theorem 4.7. *System* (68) *admits only circular or straight-line trajectories.*

Proof. Away from rest points our system (68) can be transformed into

$$
\begin{aligned}
\frac{\partial x_2^1}{\partial x_1^1} &= \frac{-x_1^1+x_1^2}{x_2^1-x_2^2},\\
\frac{\partial x_2^2}{\partial x_1^2} &= \frac{-x_1^1+x_1^2}{x_2^1-x_2^2},\\
\frac{\partial x_2^1}{\partial x_2^2} &= -\frac{M^2}{M^1},\\
\frac{\partial x_1^1}{\partial x_1^2} &= -\frac{M^2}{M^1}.
\end{aligned}
\tag{69}
$$

Integrating out the bottom two equations, we get

$$
\begin{aligned}
x_2^1 &= -\frac{M^2}{M^1}(x_2^2 + k_2),\\
x_1^1 &= -\frac{M^2}{M^1}(x_1^2 + k_1).
\end{aligned}
\tag{70}
$$

First let us assume that $M^2 \neq -M^1$. Then plugging back into the first two equations

of (69), we arrive at

$$\begin{aligned}\frac{\partial x_2^1}{\partial x_1^1} &= \frac{\left(-1-\frac{M^2}{M^1}\right)x_1^1 - k_1}{\left(1+\frac{M^2}{M^1}\right)x_2^1 - k_2} = -\frac{x_1^1-\tilde{k}_1}{x_2^1-\tilde{k}_2},\\ \frac{\partial x_2^2}{\partial x_1^2} &= \frac{\left(1+\frac{M^2}{M^1}\right)x_1^2 - \frac{M^2}{M^1}k_1}{\left(-1-\frac{M^2}{M^1}\right)x_2^1 - \frac{M^2}{M^1}k_2} = -\frac{x_1^1-\hat{k}_1}{x_2^1-\hat{k}_2}\end{aligned} \tag{71}$$

for appropriate constants $\tilde{k}_1, \tilde{k}_2, \hat{k}_1, \hat{k}_2$. Thus we integrate again and get

$$\begin{aligned}(x_2^1(t)-\tilde{k}_2)^2 + (x_1^1(t)-\tilde{k}_1)^2 &= C_1,\\ (x_2^2(t)-\hat{k}_2)^2 + (x_1^2(t)-\hat{k}_1)^2 &= C_2.\end{aligned} \tag{72}$$

If $M^2 = -M^1$, then we have equal but opposite size vortices, and (68) becomes

$$\begin{aligned}\frac{\partial x_2^1}{\partial x_1^1} &= -\frac{k2}{k1},\\ \frac{\partial x_2^2}{\partial x_1^2} &= -\frac{k2}{k1},\end{aligned} \tag{73}$$

which gives us straight-line solutions with slope $-\frac{k2}{k1}$, as desired. ■

If we now consider the case of equal total vorticity for the two vortices ($M^2 = M^1 \equiv M$), the classical point vortex result is that the vortices will rotate around the center of vorticity at a constant frequency, $\frac{2M}{2\pi D^2}$, where D is the distance between the vortex centers. We will now compute the viscous and finite core size effects on the frequency of rotation predicted by our model. For simplicity we will center the vortices at the origin and place them on the circle of radius r. We apply the polar change of variables,

$$x_1^i = r\cos(\theta_i), \quad x_2^i = r\sin(\theta_i), \quad i = 1, 2; \tag{74}$$

then, using the fact that our vortices are out of phase by π, we can compute that

$$(x_1^1 - x_1^2) = r\cos(\theta_1) - r\cos(\theta_2) = 2r\cos(\theta_1), \tag{75}$$

$$(x_2^1 - x_2^2) = r\sin(\theta_1) - r\sin(\theta_2) = 2r\sin(\theta_1), \tag{76}$$

and thus we arrive at the expression for the frequency of rotation

$$\Omega = \frac{M}{4\pi r^2}\left(1 - e^{\frac{-2r^2}{\lambda(t)^2}}\right). \tag{77}$$

Since $\lambda(t)^2 = \lambda_0^2 + 4\nu t$ we notice three things: the first is that viscosity, ν, slows the frequency of rotation down and, second, in the formal limit as $\nu \to 0$ and $\lambda_0^2 \to 0$ we recover the constant frequency, $\frac{2M}{2\pi D^2}$, which is what the Helmholtz–Kirchhoff model for the rotation of two point vortices predicts. The third property we notice is that the frequency predicted

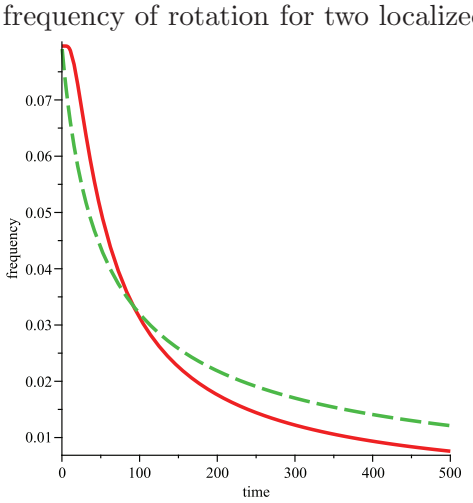


Figure 1. *Here we plot both the frequency of rotation predicted by single vortex expansion up to quadrupole order (dashed) and the frequency predicted by two vortex expansions truncated to leading order (solid). The parameter values used are* $\nu = .01$, $M = 1$, $r = 1$, *and* $\lambda_0 = .01$.

by (77) is different than $\Omega = \frac{M}{4\pi r^2}(1 - e^{-4r^2/\lambda(t)^2})$, which is obtained by naively computing the velocity field of each Gaussian vortex in the center of the other vortex.

With this calculation we may now compare the results of expanding a two-Gaussian initial distribution as a single vortex or as two independent vortices. In [12] a two-Gaussian initial distribution at a distance of $2r$ apart with core size λ_0, each with mass M, is approximated by a single vortex expansion using (28). The authors truncate the expansion to a quadrupole moment ($n = 2$) and calculate the frequency of rotation to be

$$\Omega = \frac{M}{8\pi}\left[\frac{1}{2\nu t}\ln\left(1 + \frac{4\nu t}{r^2}\right) - \frac{\lambda_0^2}{r^4}\frac{1}{1 + 4\nu t/r^2}\right]. \tag{78}$$

This equation is directly comparable to (77). Notice that both equations for frequency of rotation indicate slowing of rotation over time, albeit at different rates. In addition, both equations recover the Helmholtz–Kirchhoff approximation of $\frac{2M}{2\pi D^2}$ in the limit as $\nu \to 0$ and $\lambda_0 \to 0$. In fact, in the sufficiently localized regime, $r \ll \lambda_0$, both equations have similar initial frequencies and remain close asymptotically. A typical example is shown in Figure 1.

5. Conclusions. In this paper we have derived a system of ODEs whose solutions give a representation of solutions of the two-dimensional vorticity equation in terms of a system of interacting vortices. We have also derived a sufficient condition on the initial vorticity distribution which guarantees that this representation in terms of interacting vortices is equivalent to the original solution of the two-dimensional vorticity equation. This model generalizes the classical Helmholtz–Kirchhoff model of interacting, inviscid, point vortices to include the effects of both finite core size and viscosity. A given initial vorticity distribution may be represented in various ways by our model. As an example, we have considered the predicted frequency of rotation by representing a spinning vortex as a pair of interacting Gaussian vortices versus a single vortex but including higher-order moments. While the predictions of the

two approaches were similar in this example, we expect that in general the preferred approach will be governed by the details of the initial conditions.

Acknowledgments. CEW wishes to acknowledge many useful discussions about two-dimensional fluid flows with Th. Gallay. DU and CEW wish also to acknowledge a very helpful discussion of Hermite expansions with G. Van Baalen. Any findings, conclusions, opinions, or recommendations are those of the authors and do not necessarily reflect the views of the NSF.

REFERENCES

[1] O. Agullo and A. Verga, *Effect of viscosity in the dynamics of two point vortices: Exact results*, Phys. Rev. E, 63 (2001), paper 056304.
[2] C. Cerretelli and C. H. K. Williamson, *The physical mechanism for vortex merging*, J. Fluid Mech., 475 (2003), pp. 41–77.
[3] Th. Gallay and C. E. Wayne, *Invariant manifolds and the long-time asymptotics of the Navier-Stokes and vorticity equations on* $\mathbf{R}^2$, Arch. Ration. Mech. Anal., 163 (2002), pp. 209–258.
[4] Th. Gallay and C. E. Wayne, *Global stability of vortex solutions of the two-dimensional Navier-Stokes equation*, Comm. Math. Phys., 255 (2005), pp. 97–129.
[5] G. R. Kirchhoff, *Vorlesungen über Mathematische Physik. Mechanik*, Teubner, Leipzig, 1876.
[6] S. Le Dizès, *Non-axisymmetric vortices in two-dimensional flows*, J. Fluid Mech., 406 (2000), pp. 175–198.
[7] S. Le Dizès and A. Verga, *Viscous interactions of two co-rotating vortices before merging*, J. Fluid Mech., 467 (2002), pp. 389–410.
[8] A. J. Majda and A. L. Bertozzi, *Vorticity and Incompressible Flow*, Cambridge Texts Appl. Math. 27, Cambridge University Press, Cambridge, UK, 2002.
[9] M. V. Melander, A. S. Styczek, and N. J. Zabusky, *Elliptically desingularized vortex model for the two-dimensional Euler equations*, Phys. Rev. Lett., 53 (1984), pp. 1222–1225.
[10] M. V. Melander, N. J. Zabusky, and J. C. McWilliams, *Symmetric vortex merger in two dimensions: Causes and conditions*, J. Fluid Mech., 195 (1988), pp. 303–340.
[11] P. Meunier, S. Le Dizès, and T. Leweke, *Physics of vortex merging*, C. R. Phys., 6 (2005), pp. 431–450.
[12] R. J. Nagem, G. Sandri, and D. Uminsky, *Vorticity dynamics and sound generation in two-dimensional fluid flow*, J. Acoust. Soc. Amer., 122 (2007), pp. 128–134.
[13] P. K. Newton, *The N-Vortex Problem*, Appl. Math. Sci. 145, Springer-Verlag, New York, 2001.
[14] C. W. Oseen, *Über Wirbelbewegung in einer reibenden Flüssigkeit*, Ark. f. Mat. Astron. och Fys., 7 (1911), pp. 1–11.
[15] C. W. Oseen, *Neuere Methoden und Ergebnisse in der Hydrodynamik*, Akademische Verlagsgesellschaft, Leipzig, 1927.
[16] L. Ting and R. Klein, *Viscous Vortical Flows*, Lecture Notes in Phys. 374, Springer-Verlag, Berlin, 1991.
[17] H. von Helmholtz, *Über Integrale der hydrodynamischen Gleichungen, welche den Wirbelbewegungen entsprechen*, J. Reine Angew. Math., 55 (1858), pp. 25–55.

SIAM J. Applied Dynamical Systems
Vol. 8, No. 1, pp. 180–201
© 2009 Society for Industrial and Applied Mathematics

Dynamics of Symplectic Subvolumes*

Jared M. Maruskin†, Daniel J. Scheeres‡, and Anthony M. Bloch§

Abstract. In this paper we analyze the evolution of $2k$-dimensional submanifolds in the $2n$-dimensional Hamiltonian phase space, with $k < n$. We show that parasymplectic subvolumes (to be defined in the paper), which can be thought of as even-dimensional subvolumes that are initially parallel to k of the symplectic planes, have a minimal obtainable $2k$-volume. We define volume expansion factors along the flow, and also the projection factors one obtains by projecting the volumes onto various planes. This structure can be computed by looking at various subdeterminants or Gram determinants of the state transition matrix. We also relate these various subdeterminants to classical integral invariants, giving a practical analytical method for their computation. We then indicate how this can be used to quantify probability expansions in asteroid or space debris tracking problems, with large uncertainty distributions along one of the symplectic planes. We also show how local subvolume expansion leads to the local collapse of phase space along solution trajectories. Despite this general tendency we will also compute a preferred, orthogonal, symplectic basis associated with any linear(ized) symplectomorphism that retains its orthogonality after being mapped to the final state.

Key words. Liouville's theorem, integral invariants, symplectic manifolds, symplectic eigenskeleton

AMS subject classifications. 53D99, 70H05

DOI. 10.1137/070697938

1. Introduction.

1.1. Overview.

The traditional approach for studying the dynamics and control of mechanical systems is to focus on individual trajectories and states in order to determine where they will go and where they can be forced to go. In reality, however, system states are never precisely known and can be determined to exist only within some set of finite volume in the dynamical system's phase space. By treating such systems as sums of individual trajectories, one loses the geometrical insight and deeper results offered by more holistic approaches.

In this paper, we will be concerned with understanding fundamental constraints on the evolution of compact $2k$-dimensional symplectic sets that evolve along a nominal trajectory of the system. Different symplectic constraints arise on such sets, including conservation of the signed $2k$-volume projections on the coupled symplectic planes as well as the constraints implied from Gromov's nonsqueezing theorem (see Scheeres et al. [17] for a discussion of these constraints in relation to orbit uncertainty evolution, and McDuff and Salamon [12], Hofer and Zehnder [7], and Audin, da Silva, and Lerman [2] for recent results on symplectic topology).

*Received by the editors July 20, 2007; accepted for publication (in revised form) by T. Kaper October 2, 2008; published electronically January 23, 2009. This research was supported by National Science Foundation grants CMS-0408542 and DMS-604307.
http://www.siam.org/journals/siads/8-1/69793.html
†Department of Mathematics, San José State University, San José, CA 95192 (maruskin@math.sjsu.edu).
‡Department of Aerospace Engineering, University of Colorado-Boulder, Boulder, CO 80309 (scheeres@colorado.edu).
§Department of Mathematics, University of Michigan, Ann Arbor, MI 48109 (abloch@umich.edu).

We will further present an additional constraint for a minimal obtainable volume that exists on certain classes of $2k$-dimensional symplectic sets and show how such a constraint leads to the local collapse of phase space along solution curves in Hamiltonian phase space. By the "local collapse of phase space" we mean the tendency of basis vectors of the tangent space at the initial point to collapse toward a single direction or subspace as the system evolves. This collapse is especially prevalent in chaotic systems; see, for instance, the discussion in Lichtenberg and Lieberman [9]. This collapse of phase space is fundamentally linked to the expansion of symplectic subvolumes. Finally, for any fixed final time, no matter how long into the future or how chaotic the system, we prove the existence of a distinguished orthogonal symplectic basis that, when mapped by the linearized map to the tangent space at the final point, returns to an orthogonal basis at the final time. The uncertainty of any $2k$-dimensional differential volume initially parallel to the symplectic planes of this basis, even though it may increase dramatically during the course of its evolution, will always return to its initial uncertainty at the final time.

Since the equations produced by applying Pontryagin's maximum principle to optimal control problems are Hamiltonian, the results we discuss here should provide geometric insight to the evolution and control of uncertainty distributions in such systems. This theory provides fundamental limits on dynamical orbits, and hence if one provides a control, it provides limits on the accuracy of the control in the face of uncertainty. It also provides limits on uncertainty propagation in Hamiltonian systems. Applications of these results to control systems will be the topic of future research.

1.2. Outline. In section 2, we introduce Hamiltonian systems and the state transition matrix (STM). The term STM is used in linear systems theory to represent the matrix of the differential of the flow; i.e., for a Hamiltonian flow $\phi_t : Q \to Q$, the STM is the matrix of the linear mapping $d\phi : TQ \to TQ$. We show how classical identities on the Lagrange and Poisson brackets relate to constraints on the STM. Specifically, for any symplectic column of the STM Φ, the sum of the 2×2 symplectic subdeterminants must add up to unity.

In section 3, we consider surfaces that can be explicitly parameterized by one of the symplectic planes. We derive area expansion factors from the parameterization plane to the surface, its image under the Hamiltonian phase flow, and the symplectic projections of its image. If the state of the system is somewhere on the initial surface (with equal a priori probability), we interpret these various expansion factors as a probability map that leads one to understand where the particle, after applying the Hamiltonian phase flow, is likely to be found. In section 5, we provide an application of this to asteroid tracking.

In section 4, we present an expansion property of $2k$-volumes which is closely related to Wirtinger's inequality. We show how this leads to the fact that subvolume expansions in the differential neighborhood of the Hamiltonian flow leads to a collapsing property in systems which exhibit chaos.

In section 5 we discuss how these constraints on subvolume expansions, when considered with Liouville's theorem, lead to the local collapse of the phase space around nominal trajectories. Interestingly we will also show that, given a linear(ized) canonical transformation, there always exists a preferred, orthogonal, symplectic basis that maps into a second orthogonal, symplectic basis by the linearized map. In particular, the volume of a $2k$-dimensional subvolume chosen to be initially parallel to k of the preferred symplectic planes will return to

its initial value at this fixed final time.

2. The STM for Hamiltonian systems. In this section we will discuss how the symplectic structure manifests itself as constraints on the inner structure of the STM for a Hamiltonian system. In general a symplectic structure on an even-dimensional manifold is any closed nondegenerate differential two-form. However, we will restrict our attention to the case of dynamics in the standard Euclidean phase space $\mathbb{R}^{2n}$ with the standard symplectic form $\omega = dp \wedge dq$. A *symplectomorphism* ϕ is a transformation which preserves the symplectic structure, i.e., $\phi^*\omega = \omega$. For additional background on Hamiltonian systems and symplectic manifolds, see Arnold [1], Bloch [3], Greenwood [5], Marsden and Ratiu [10], and Silva [18].

For a Hamiltonian function $H : \mathbb{R}^{2n} \to \mathbb{R}$, Hamilton's equations are given by

$$\dot{q}_i = \frac{\partial H}{\partial p_i} \qquad \text{and} \qquad \dot{p}_i = -\frac{\partial H}{\partial q_i}. \tag{2.1}$$

Solutions curves of the system (2.1) are called the Hamiltonian phase flow and are denoted $\phi_t(q, p)$.

We can represent Hamilton's equations (2.1) using matrix notation in the following way. We will use the ordering convention used by Arnold [1]:

$$x = \langle p_1, q_1, p_2, q_2, \ldots, p_N, q_N \rangle.$$

Define the matrix

$$J_2 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

The matrix J is defined as the $2N \times 2N$ block-diagonal matrix with J_2's down the main diagonal. Grouping the coordinates in symplectic pairs will be useful to us later as we will be looking at various "symplectic columns" of the STM. Given the above definitions, Hamilton's equations can be cast into the following matrix form:

$$\dot{x} = \frac{d}{dt}\phi_t(x_0) = J \cdot \frac{\partial H}{\partial x}, \tag{2.2}$$

where $x(t) = \phi_t(x_0)$. The right-hand side is the so-called *symplectic gradient* of the Hamiltonian function.

The differential of the symplectomorphism $\phi : M \to N$, $\phi(p, q) = (P, Q)$, i.e.,

$$d\phi : T_{(p,q)}M \to T_{(P,Q)}N,$$

is, when represented in matrix form, known as the STM Φ. This terminology originates in linear systems theory. The matrix $\Phi(t)$ maps a variation of the initial state to its corresponding variation along the curve, i.e., $\phi_t(x_0 + \delta x_0) = x(t) + \Phi(t) \cdot \delta x_0 + O((\delta x_0)^2)$ as $\delta x_0 \to 0$.

Given a nominal trajectory $x(t) = \phi_t(x_0)$ originating from the point x_0, we have the following dynamical equations of motion:

$$\frac{d}{dt}\phi_t(x_0) = J \cdot \frac{\partial H}{\partial x}(x(t)).$$

If we perturb the initial conditions to $x_0 + \delta x_0$, we find

$$\frac{d}{dt}\phi_t(x_0 + \delta x_0) = J \cdot \frac{\partial H}{\partial x}(x + \delta x).$$

By expanding this in a Taylor series, one sees

$$\frac{d}{dt}\Phi = J \cdot \frac{\partial^2 H}{\partial x^2} \cdot \Phi. \tag{2.3}$$

This defines a system of $4N^2$ differential equations that can be integrated numerically, simultaneously along with the nominal solution curve $\phi_t(x_0)$.

Liouville's theorem is a statement of volume preservation in Hamiltonian systems and may be represented geometrically by the invariance of the nth exterior power of the standard symplectic form, i.e., $\phi^*\omega^n = \omega^n$. It is well known that Liouville's theorem manifests itself as a constraint on the STM Φ for any Hamiltonian phase flow by the condition that $\det(\Phi) = 1$. However, there is additional structure contained within the STM, corresponding to the symplectic invariants $\phi^*\omega^k = \omega^k$ for $1 \leq k \leq n$. We will now describe how these symplectic invariants manifest themselves as constraints on the internal structure of the STM.

Using the abbreviations $x = \langle p_1, q_1, \ldots, p_N, q_N \rangle^T$ and $X = \langle P_1, Q_1, \ldots, P_n, Q_N \rangle^T$, we can represent the STM Φ as

$$\Phi = \frac{\partial X}{\partial x}.$$

We define the $2n \times 2$ matrix Π_κ as

$$\Pi_\kappa = \begin{bmatrix} \mathbb{O}_2 & \mathbb{O}_2 & \cdots & \mathbb{O}_2 & \mathbb{I}_2 & \mathbb{O}_2 & \cdots & \mathbb{O}_2 \end{bmatrix}^T, \tag{2.4}$$

where the 2×2 identity matrix $\mathbb{I}_2$ appears in the κth 2×2 matrix in the column. This is literally the transpose of the matrix of the differential of the projection map $\pi_\kappa : \mathbb{R}^{2n} \to \mathbb{R}^2$ onto the κth symplectic plane.

The *κth symplectic column* of Φ is the $2n \times 2$ matrix given by $\Phi \cdot \Pi_\kappa$. The *κth symplectic row* of Φ is the $2 \times 2n$ matrix given by $\Pi_\kappa^T \cdot \Phi$. Finally, $\Pi_\kappa^T \cdot \Phi \cdot \Pi_\lambda$ is the 2×2 intersection of the κth symplectic row with the λth symplectic column of Φ.

Define the following subdeterminants:

$$M_{ij} = \det(\Pi_i^T \cdot \Phi \cdot \Pi_j) = \det \begin{pmatrix} \frac{\partial P_i}{\partial p_j} & \frac{\partial P_i}{\partial q_j} \\ \frac{\partial Q_i}{\partial p_j} & \frac{\partial Q_i}{\partial q_j} \end{pmatrix} = \frac{\partial P_i}{\partial p_j}\frac{\partial Q_i}{\partial q_j} - \frac{\partial P_i}{\partial q_j}\frac{\partial Q_i}{\partial p_j}. \tag{2.5}$$

M_{ij} is the subdeterminant of the intersection of the ith *symplectic* row of the STM with its jth *symplectic* column. We can now state how the invariance of the symplectic form $\phi^*\omega = \omega$ manifests itself in terms of these 2×2 subdeterminants.

Theorem 2.1. *Let M_{ij} be the determinants of the 2×2 submatrices of the STM Φ formed by taking the intersection of the ith symplectic row and jth symplectic column, as defined in (2.5). If Φ is the STM of a Hamiltonian phase flow, then the following conditions apply:*

$$\sum_{i=1}^{N} M_{ij} = 1, \tag{2.6}$$

$$\sum_{j=1}^{N} M_{ij} = 1. \tag{2.7}$$

Proof. The differential forms $\{dp_k, dq_k\}$ form a basis of the cotangent bundle $T^*_X\mathbb{R}^{2n}$. In terms of this basis, we have

$$dP_i = \sum_{k=1}^{n} \left(\frac{\partial P_i}{\partial q_k} dq_k + \frac{\partial P_i}{\partial p_k} dp_k \right),$$
$$dQ_i = \sum_{k=1}^{n} \left(\frac{\partial Q_i}{\partial q_k} dq_k + \frac{\partial Q_i}{\partial p_k} dp_k \right).$$

Notice that the entries in the $(2i-1)$th row of Φ are the components of the differential form dP_i with respect to the basis $\langle dp_1, dq_1, \ldots, dp_n, dq_n \rangle$. Similarly the entries of the $(2i)$th row of Φ are the components of the differential form dQ_i with respect to this same basis.

We now give an alternate interpretation of the quantities M_{ij} defined in (2.5). By definition of the wedge product we have

$$dP_i \wedge dQ_i \left(\frac{\partial}{\partial p_j}, \frac{\partial}{\partial q_j} \right) = \det \begin{pmatrix} dP_i\left(\frac{\partial}{\partial p_j}\right) & dP_i\left(\frac{\partial}{\partial q_j}\right) \\ dQ_i\left(\frac{\partial}{\partial p_j}\right) & dQ_i\left(\frac{\partial}{\partial q_j}\right) \end{pmatrix}$$
$$= \det \begin{pmatrix} \frac{\partial P_i}{\partial p_j} & \frac{\partial P_i}{\partial q^j} \\ \frac{\partial Q_i}{\partial p_j} & \frac{\partial Q_i}{\partial q^j} \end{pmatrix} = M_{ij}.$$

However, since the transformation is symplectic, for a fixed j it follows that

$$\sum_{i=1}^{n} dP_i \wedge dQ_i \left(\frac{\partial}{\partial p_j}, \frac{\partial}{\partial q^j} \right) = \sum_{i=1}^{n} dp_i \wedge dq_i \left(\frac{\partial}{\partial p_j}, \frac{\partial}{\partial q_j} \right) = \sum_{i=1}^{n} \delta_{ij} = 1.$$

This shows (2.6). A similar argument shows (2.7). ∎

It follows that invariance of the symplectic form is equivalent to the following set of conditions. For a fixed symplectic column of Φ, the sum of the n determinants of the 2×2 submatrices must add to unity. A similar statement holds for any fixed symplectic row.

3. Symplectic surfaces.

We begin this section by defining a few classes of distinguished surfaces that occur in symplectic spaces. We then give an efficient method for the computation of area expansion factors and area projection factors in terms of the Gram determinant of certain matrices that are derivable from the STM. For a classical exposition of surfaces, see Osserman [15].

3.1. Surface classifications.

We will begin by making the following fairly natural definitions.

Definition 3.1. *A* globally symplectic surface *is a two-dimensional submanifold of the phase space* $\mathbb{R}^{2n}$ *which admits a* 1*-to-*1 *mapping to at least one of the symplectic planes via the projection operator; i.e., it is a surface which can be parameterized in explicit form by one of its symplectic coordinate pairs.*

This characteristic is not an invariant one. It is possible, for example, for a lamina parallel to a symplectic plane to fold under some symplectic map, so that its image under the map is not 1-to-1 with any symplectic plane.

Another surface type we will consider is the following.

Definition 3.2. *A* parametrically symplectic surface *(or* parasymplectic surface*, for short) is a two-dimensional submanifold of* $\mathbb{R}^{2n}$ *that admits a parameterization that is a symplectic one, i.e., one with a parameterization map that is canonical.*

Notice that the restriction of the symplectic form ω to the two-dimensional tangent bundle of any parasymplectic surface is itself a symplectic form on that submanifold.

The parasymplecticity of a surface is an invariant characteristic. Let σ be a lamina on the symplectic plane (u, v) which is the parameterization of the surface $\psi(\sigma) \subset \mathbb{R}^{2n}$, where ψ is the parameterization map. If the surface $\psi(\sigma)$ is a parasymplectic surface, we can find a parameterization map ψ that is also symplectic. Let $\phi : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ be an arbitrary symplectomorphism which takes $\phi : \psi(\sigma) \to \phi(\psi(\sigma))$. Then $\phi(\psi(\sigma))$ is parasymplectic, with symplectic parameterization $\phi \circ \psi : \sigma \to \phi(\psi(\sigma))$.

3.2. Area expansion factors. In this paragraph we discuss a practical approach to computing the time evolution of the area elements of a surface and their symplectic projections in terms of Gram determinants of various submatrices of the STM. We will then indicate how one may use this to compute probability densities for dynamically evolving uncertainty surfaces such as those that arise in asteroid and space debris tracking problems.

3.2.1. Area expansion factors. For convenience we consider a surface τ which is initially *globally symplectic* with respect to the jth symplectic plane; i.e., the projection map $\pi_j : \tau \subset \mathbb{R}^{2n} \to \pi_j(\tau) \subset \mathbb{R}^2$ is 1-to-1. We can parameterize τ by its symplectic shadow on the jth symplectic plane. Now let the surface τ be mapped into the future by the Hamiltonian flow $\phi_t : \langle p_1, q_1, \ldots, p_N, q_N \rangle \to \langle P_1, Q_1, \ldots, P_N, Q_N \rangle$. We will now consider the projection of $\phi_t(\tau)$ onto the ith symplectic plane. For our analysis we will consider a differential area element $d\tau$ of τ (see Figure 1).

The surface τ is described by the parameterization $u = p_j$ and $v = q_j$ by $\langle p_1(u, v), q_1(u, v), \ldots, u, v, \ldots, p_N(u, v), q_N(u, v) \rangle$. We define the matrix $L = \left[\frac{\partial x_i}{\partial u_j}\right]$ by

$$L = \begin{bmatrix} \frac{\partial p_1}{\partial u} & \frac{\partial p_1}{\partial v} \\ \frac{\partial q_1}{\partial u} & \frac{\partial q_1}{\partial v} \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ \frac{\partial p_N}{\partial u} & \frac{\partial p_N}{\partial v} \\ \frac{\partial q_N}{\partial u} & \frac{\partial q_N}{\partial v} \end{bmatrix}, \tag{3.1}$$

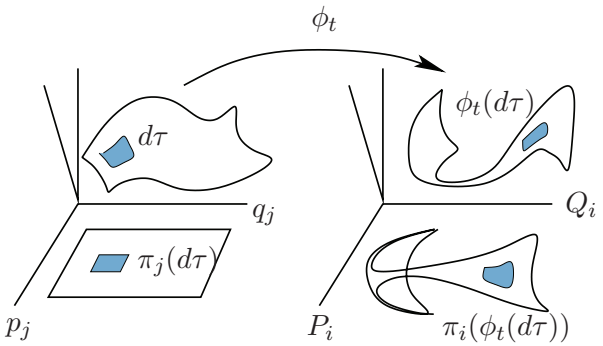


Figure 1. *Area expansion factors.*

where the jth symplectic row is the 2×2 identity matrix $\mathbb{I}_2$. We note that the surface τ may be parameterized by its projection on the $\langle p_j, q_j \rangle$ plane and the parameterization mapping $\psi = \pi_j^{-1} : \pi_j(\tau) \to \tau$, where the image of π_j^{-1} is understood to be restricted to coincide with the surface τ. In this case the matrix L is the simply Jacobian matrix of the parameterization map, i.e., $L = [d\psi]$. For the special case where τ is parallel to the jth symplectic plane, the matrix L reduces to $L = \Pi_j$, defined in (2.4). The metric for the surface τ in terms of variations in the $u - v$ plane is given by

$$[g_{ij}] = \begin{bmatrix} \frac{\partial \mathbf{x}}{\partial u} \cdot \frac{\partial \mathbf{x}}{\partial u} & \frac{\partial \mathbf{x}}{\partial u} \cdot \frac{\partial \mathbf{x}}{\partial v} \\ \frac{\partial \mathbf{x}}{\partial v} \cdot \frac{\partial \mathbf{x}}{\partial u} & \frac{\partial \mathbf{x}}{\partial v} \cdot \frac{\partial \mathbf{x}}{\partial v} \end{bmatrix} = L^T \cdot L,$$

where $\mathbf{x} = \langle p_1(u, v), \ldots, q_N(u, v) \rangle$, as usual. Hence the metric determinant is

$$g = \det(g_{ij}) = \mathfrak{G}(L) = \det(L^T \cdot L),$$

where $\mathfrak{G}(L)$ is the Gram determinant of the matrix L, defined by this equation.

The surface area of τ is thus given by

$$S.A.(\tau) = \iint_{\pi_j(\tau)} \sqrt{g}\, du dv.$$

A simple application of this result gives us the total physical area of the area element $d\tau$ in terms of the area of its projection,

$$\frac{\mathcal{A}(d\tau)}{\mathcal{A}(\pi_j(d\tau))} = \sqrt{\mathfrak{G}(L)}. \tag{3.2}$$

Similarly we have

$$\frac{\mathcal{A}(\phi_t(d\tau))}{\mathcal{A}(\pi_j(d\tau))} = \sqrt{\mathfrak{G}(\Phi \cdot L)}, \tag{3.3}$$

where Φ is the STM associated with ϕ_t and x_0.

The area of the projection $\pi_i(\phi_t(d\tau))$ is given by the Jacobian

$$\det\left(\frac{\partial(P_i,Q_i)}{\partial(u,v)}\right)=\sum_{\kappa=1}^{2N}\sum_{\lambda=1}^{2N}\left(\frac{\partial P_i}{\partial x_\kappa}\frac{\partial Q_i}{\partial x_\lambda}-\frac{\partial Q_i}{\partial x_\kappa}\frac{\partial P_i}{\partial x_\lambda}\right)\frac{\partial x_\kappa}{\partial u}\frac{\partial x_\lambda}{\partial v},$$

which can be represented more concisely with the aid of (2.4) as

$$\frac{\mathcal{A}(\pi_i(\phi_t(d\tau)))}{\mathcal{A}(\pi_j(d\tau))}=\det(\Pi_i^T\cdot\Phi\cdot L). \tag{3.4}$$

3.2.2. Application to uncertainty analysis. Suppose we know that a system can be found anywhere on the surface τ with equal a priori probability. The surface is now mapped into the future by the Hamiltonian flow ϕ_t and we wish to determine where the particle is most likely to be on the P_i-Q_i plane.

We begin by discretizing the $(u=p_j)-(v=q_j)$ plane, each area element with area $\Delta u\Delta v$. Summation will be assumed to be over each district. The probability that the particle is in $d\tau$ is given by

$$\mathcal{P}(d\tau)=\frac{\mathcal{A}(d\tau)}{\mathcal{A}(\tau)}\approx\frac{\sqrt{\mathfrak{G}(L)}\Delta u\Delta v}{\sum\sqrt{\mathfrak{G}(L)}\Delta u\Delta v}=\frac{\sqrt{\mathfrak{G}(L)}}{\sum\sqrt{\mathfrak{G}(L)}}.$$

From (3.4) we see

$$\mathcal{A}(\pi_i(\phi_t(d\tau)))\approx\det(\Pi_i^T\cdot\Phi\cdot L)\Delta u\Delta v,$$

so that the area probability density at $\langle P_i,Q_i\rangle$ is

$$\sigma(P_i,Q_i)\approx\frac{\sqrt{\mathfrak{G}(L)}}{|\det(\Pi_i^T\cdot\Phi\cdot L)|\Delta u\Delta v\sum\sqrt{\mathfrak{G}(L)}}. \tag{3.5}$$

Taking the limit as $\Delta u,\Delta v\to 0$, this becomes

$$\sigma(P_i,Q_i)=\left(\iint_{\pi_j(\tau)}\sqrt{\mathfrak{G}(L)}\,dudv\right)^{-1}\cdot\frac{\sqrt{\mathfrak{G}(L)}}{|\det(\Pi_i^T\cdot\Phi\cdot L)|}.$$

One recognizes that the proportionality constant on the left is simply the inverse of the two-dimensional surface area of the initial two-dimensional uncertainty region.

This approach may be helpful in asteroid tracking, where angular and angular rate information is precisely known, but there is a large amount of initial uncertainty in the topocentric range and range rate variables. This problem is discussed in Milani, Gronchi, and Vitturi [13]. A similar problem exists when correlating optical observations of different pieces of earth-orbiting space debris and the subsequent orbit determination problem; see, for example Maruskin, Scheeres, and Alfriend [11], Milani, Gronchi, and Vitturi [13], and Tommei, Milani, and Rossi [19]. We will illustrate a simple example of these probability maps in section 6 when we consider the example of tracking an asteroid in the earth-moon system.

4. Integral invariants. In this section we will discuss the difference between two fundamental integral invariants defined for an arbitrary $2k$-dimensional subvolume of our $2n$-dimensional phase space, for $k=1,\dots,n$. The first is the well-known integral invariant of Poincaré and Cartan. The second integral invariant is closely related and is tantamount to a global version of Wirtinger's inequality for lower-dimensional subvolumes of phase space.

4.1. Signed and unsigned integrals of differential forms. We will begin our discussion on integral invariants with a brief discussion of the theory of integration of differential forms. Let $\Sigma \subset \mathbb{R}^{2n}$ be a $2k$-dimensional submanifold, parameterized by $\psi : (\sigma \subset \mathbb{R}^{2k}) \to \Sigma$. Let α be a $2k$-form on $\mathbb{R}^{2n}$. Then we define

$$\int_\Sigma \alpha = \int_\sigma \psi^* \alpha.$$

The pullback of α under the map ψ can be expressed in the following form:

$$\psi^* \alpha = f(x_1, \ldots, x_{2k}) dx^1 \wedge \cdots \wedge dx^{2k},$$

where we take $(x_1, \ldots, x_{2k})$ to be coordinates in $\mathbb{R}^{2k}$. Then the integral of $\psi^*\alpha$ over σ reduces to the ordinary Euclidean integral

$$\int_\Sigma \alpha = \int_\sigma \psi^* \alpha = \int_\sigma f(x_1, \ldots, x_{2k}) dx^1 \cdots dx^{2k}.$$

We would like to introduce a further definition as follows. We define the unsigned integral of α over Σ to be

$$\int_\Sigma |\alpha| = \int_\sigma |f(x_1, \ldots, x_{2k})| dx^1 \cdots dx^{2k},$$

where f has been defined above.

A more rigorous definition of the above integrals must involve a partition of unity, but simplicity has been chosen over rigor so as to illustrate the spirit of the definitions.

4.2. The integral invariants of Poincaré and Cartan. Consider the standard symplectic form

$$\omega = \sum_{i=1}^{n} p_i \wedge q_i$$

and its kth exterior product:

$$\frac{1}{k!}\omega^k = \sum_{1 \le i_1 < \cdots < i_k \le n} dp_{i_1} \wedge dq_{i_1} \wedge \cdots \wedge dp_{i_k} \wedge dq_{i_k}.$$

Consider a set of $2k$ vectors $(X^1, \ldots, X^{2k})$ in $\mathbb{R}^{2n}$. Then

$$\frac{1}{k!}\omega^k(X^1, \ldots, X^{2k})$$

represents the sum of the oriented $2k$-volume projections of the parallelepiped spanned by $X^1, \ldots, X^{2k}$ on the symplectic "$2k$-planes."

ω^k is known as the Poincaré–Cartan integral invariant. Given an arbitrary $2k$-dimensional phase volume Ω (in a $2n$-dimensional space) and the Hamiltonian phase flow ϕ_t, we have

$$\frac{1}{k!}\int_\Omega \omega^k = \frac{1}{k!}\int_{\phi_t(\Omega)} \omega^k,$$

so that the sum of the oriented $2k$-volume projections on each symplectic "$2k$-plane" is conserved.

4.3. The Wirtinger-type integral invariants and volume. Given a set of vectors $X_1, \ldots, X_{2k} \in \mathbb{R}^{2n}$, Wirtinger's inequality states that the "$2k$" volume of the parallelepiped spanned by these vectors is bounded by

$$\frac{1}{k!}|\omega^k(X_1, \ldots, X_{2k})| \leq \mathrm{Vol}_{2k}(X_1, \ldots, X_{2k}). \tag{4.1}$$

We make the following two observations. First, it is clear that $|\omega^k|$ is an integral invariant of the Hamiltonian flow ϕ_t, so that, given any $2k$-volume Ω, we have

$$\frac{1}{k!}\int_{\Omega} |\omega^k| = \frac{1}{k!}\int_{\phi_t(\Omega)} |\omega^k|.$$

Moreover, due to (4.1), we have

$$\frac{1}{k!}\int_{\Omega} |\omega^k| \leq \mathrm{Vol}_{2k}(\Omega), \tag{4.2}$$

so that this integral invariant is a lower bound on the set of possible $(2k)$ volumes that the body Ω may attain. For the case $k = n$, the volume of Ω is a constant which equals this invariant quantity (Liouville's theorem).

Equation (4.2) adds to the interesting list of constraints that exist on closed sets in symplectic spaces. There are two types of constraints which act on $2n$-dimensional compact sets, Liouville's theorem and Gromov's nonsqueezing theorem (GNT); see Scheeres et al. [17]. Liouville's theorem states that the $2n$-volume of the set must remain constant under any symplectomorphism. On the other hand, GNT states that the area of the symplectic projection of such a set has a minimum attainable value, known as the Gromov width. Lower-dimensional compact sets also have two types of constraints. The $2k$-volume of these sets does not need to be conserved, since they are not of full dimension. However, the constraint (4.2) introduced above guarantees the existence of a minimum attainable $2k$-volume of such a set under all symplectomorphisms. An alternate interpretation of this constraint is given below in Theorem 4.2. Additionally the integral invariants of Poincaré and Cartan state that the sum of the signed $2k$-volume projections onto any permutation of k symplectic planes must be conserved.

4.4. Parasymplectic $2k$-volumes. In direct analogy with our discussion of parasymplectic surfaces, we define parasymplectic $2k$-volumes as follows.

Definition 4.1. *A* parasymplectic $2k$-volume*, or parametrically symplectic $2k$-volume, is one that admits a symplectic parameterization map.*

An example of a parasympletic volume is the following. Take any $2k$-dimensional volume Ω that is parallel to k of the symplectic planes, i.e., a region defined by

$$\Omega = \langle p_1, q_1, \ldots, p_k, q_k, c_{k+1}, d_{k+1}, \ldots, c_n, d_n \rangle,$$

where the variables $p_1, q_1, \ldots, p_k, q_k$ vary over some region of $\mathbb{R}^{2k}$ and $c_{k+1}, d_{k+1}, \ldots, c_n, d_n$ are constants. Now let ϕ_t be the Hamiltonian phase flow. The $2k$-phase volumes $\phi_t(\Omega) \subset \mathbb{R}^{2n}$ are a one-parameter family of parasymplectic $2k$-volumes. This follows since $\phi_t(\Omega)$ can be symplectically parameterized by the projection of Ω onto the first k symplectic planes.

Theorem 4.2 (volume expansion of parasymplectic $2k$-volumes). *Let $\Omega \subset \mathbb{R}^{2k}$ be the parameterization of a $2k$-dimensional volume $\psi(\Omega) \subset \mathbb{R}^{2n}$ in a symplectic phase space whose parameterization map ψ is a symplectic one. Then*

$$\mathrm{Vol}_{2k}(\Omega) \leq \mathrm{Vol}_{2k}(\psi(\Omega)).$$

Corollary 4.3. *The $2k$-volume of any parasymplectic $2k$-volume is at least as large as the volume of its symplectic parameterization.*

Proof. To prove our theorem, we only need prove it for a differential volume element. The generalization follows via a simple integration argument.

Let $\psi : \mathbb{R}^{2k} \to \mathbb{R}^{2n}$ be a symplectic parameterization of a surface. Let

$$\left\{ \frac{\partial}{\partial u_1}, \frac{\partial}{\partial v_1}, \ldots, \frac{\partial}{\partial u_k}, \frac{\partial}{\partial v_k} \right\}$$

be an orthonormal basis of $\mathbb{R}^{2k}$, and let

$$X_i = \psi_* \left(\frac{\partial}{\partial u_i} \right) \qquad \text{and} \qquad Y_i = \psi_* \left(\frac{\partial}{\partial v_i} \right)$$

be the push forwards of these basis vectors. Applying Wirtinger's inequality (4.1), we have that

$$\begin{aligned} \frac{1}{k!}\omega^k \left(\psi_* \left(\frac{\partial}{\partial u_1} \right), \ldots, \psi_* \left(\frac{\partial}{\partial v_k} \right) \right) &= \frac{1}{k!}\psi^*(\omega^k) \left(\frac{\partial}{\partial u_1}, \ldots, \frac{\partial}{\partial v_k} \right) \\ &= \frac{1}{k!}\omega^k \left(\frac{\partial}{\partial u_1}, \ldots, \frac{\partial}{\partial v_k} \right) \\ &= du_1 \wedge \cdots \wedge dv_k \left(\frac{\partial}{\partial u_1}, \ldots, \frac{\partial}{\partial v_k} \right) \\ &= 1 \leq \mathrm{Vol}_{2k}(X_1, \ldots, Y_k). \end{aligned}$$

Hence any $2k$-dimensional volume measure must be nondecreasing under such a map. ∎

A consequence of this is that any parasymplectic $2k$-dimensional set achieves its minimum $2k$-volume when it is parallel with k of the symplectic planes.

4.5. The volume expansion factor. In this section we provide a practical approach to determining the volume and the integral invariants of $2k$-dimensional subvolumes. We will consider the volume $\Omega \subset \mathbb{R}^{2k}$ to be the parameterization region of a $2k$-dimensional set in the symplectic space $\mathbb{R}^{2n}$ with parameterization map

$$\psi : (\Omega \subset \mathbb{R}^{2k}) \to (\psi(\Omega) \subset \mathbb{R}^{2n}).$$

The Jacobian of the parameterization is the $2n \times 2k$ matrix given by

$$L = [d\psi].$$

We will be interested in computing the total volume of Ω, the sum of its oriented symplectic projections (i.e., the integral invariant of Poincaré and Cartan), and its minimum obtainable volume (i.e., the integral invariant of the Wirtinger type).

In terms of the parameterization coordinates $\langle u_1, v_1, \ldots, u_k, v_k \rangle \in \mathbb{R}^{2k}$, the following metric is induced on the $2k$-dimensional region $\psi(\Omega)$:

$$\mathbf{g} = L^T \cdot L.$$

The metric determinant $g = \det \mathbf{g}$ is therefore the Gram determinant of the Jacobian matrix L,

$$g = \mathfrak{G}(L) = \det(L^T \cdot L),$$

so that

$$\mathrm{Vol}_{2k}(\psi(\Omega)) = \int_\Omega \sqrt{|g|}\, d\Omega.$$

Practically the Gramian of the Jacobian can be identified with the $2k$-volume expansion factor of the parameterization map in the following way:

$$\nu_{2k}(d\Omega; \psi) = \frac{\mathrm{Vol}_{2k}(\psi(d\Omega))}{\mathrm{Vol}_{2k}(d\Omega)} = \sqrt{\mathfrak{G}(L)},$$

where $\nu_{2k}(d\Omega; \psi)$ is the local $2k$-volume expansion factor of $d\Omega$ under the mapping ψ. This is identical to the two-dimensional version (3.2). If one then transforms the $2k$-dimensional set $\psi(\Omega)$ by a symplectomorphism ϕ, with $\Phi = [d\phi]$, one can see that (3.3) generalizes to

$$\frac{\mathrm{Vol}_{2k}(\phi(\psi(d\Omega)))}{\mathrm{Vol}_{2k}(d\Omega)} = \sqrt{\mathfrak{G}(\Phi \cdot L)}.$$

5. Local collapse of phase space and the symplectic eigenskeleton. A fundamental feature of symplectic systems is the local collapse of phase space; i.e., the subspaces of the tangent space at a given point tend to collapse toward a single direction or subspace under the linearized map. This comes about as a result of volume expansion of parasymplectic surfaces (Theorem 4.2). Despite this tendency there always exists a preferred, orthogonal, symplectic basis associated with any linear(ized) symplectomorphism that retains its orthogonality after being mapped to the final state. Additionally any differential parasymplectic surfaces parallel to one of these preferred symplectic planes is mapped into its minimum area configuration by the symplectomorphism ϕ. In this section we will demonstrate the relation between Theorem 4.2 and phase space collapse and then show how one can compute this preferred basis which resists such collapse.

5.1. Volume expansion and the local collapse of phase space. The setting for this subsection will be the evolution of an infinitesimal neighborhood surrounding a Hamiltonian trajectory through phase space. Consider the Hamiltonian flow $\phi_t : x \to X$. Now consider a differential $2n$-"cube" Ω situated at the initial point x whose faces are parallel with the symplectic planes. Let $\Upsilon \subset \Omega$ be a $2k$-dimensional subset that is parallel with k of the symplectic planes, and let $\Upsilon' \subset \Omega$ be a $(2n - 2k)$-dimensional subset that is parallel with the remaining $(n - k)$ symplectic planes such that Ω is a direct sum $\Omega = \Upsilon \oplus \Upsilon'$ and, therefore,

$$\mathrm{Vol}_{2n}(\Omega) = \mathrm{Vol}_{2k}(\Upsilon) \cdot \mathrm{Vol}_{2n-2k}(\Upsilon').$$

The Hamiltonian flow now takes $\phi_t : x \to X$ along with its differential neighborhood. We define

$$\overline{\Omega} = d\phi_t(\Omega),$$
$$\overline{\Upsilon} = d\phi_t(\Upsilon),$$
$$\overline{\Upsilon'} = d\phi_t(\Upsilon').$$

We now define the angle β via the relation

$$\mathrm{Vol}_{2n}(\overline{\Omega}) = \mathrm{Vol}_{2k}(\overline{\Upsilon}) \cdot \mathrm{Vol}_{2n-2k}(\overline{\Upsilon'}) \cdot \sin\beta,$$

so that β is the angle between the subspaces $\overline{\Upsilon}$ and $\overline{\Upsilon'}$. By Liouville's theorem, we have

$$\mathrm{Vol}_{2n}(\overline{\Omega}) = \mathrm{Vol}_{2n}(\Omega),$$

so that

$$1 = \nu_{2k}(\Upsilon;\phi_t)\nu_{2n-2k}(\Upsilon';\phi_t)\sin\beta.$$

As a result of Theorem 4.2 we have

$$\nu_{2k}(\Upsilon;\phi_t) \geq 1,$$
$$\nu_{2n-2k}(\Upsilon';\phi_t) \geq 1.$$

We conclude that the greater the volume expansion of these lower-dimensional differential "slices" Υ and Υ', the greater the inward collapse of their respective subspaces toward each other, signified by a decrease in the relative angle between them. In chaotic Hamiltonian systems, where $\mathrm{Vol}_{2k}(\phi_t(\Upsilon))$ grows exponentially, we see that β is correspondingly decaying at an exponential rate. Thus chaos (for Hamiltonian systems) necessarily implies the collapse of the phase space along certain directions. Despite this (possibly) exponential collapse of the local differential phase space along a trajectory, symplectic systems have the remarkable property that, for a fixed final time, there *always* is an initial symplectic basis of $T_x\mathbb{R}^{2n}$ which returns to an orthogonal basis of $T_X\mathbb{R}^{2n}$ at the final state. We demonstrate the existence of such a basis below.

5.2. The symplectic eigenskeleton. In the previous section we showed that local collapse of phase space at the differential level along solution flows is a general feature of Hamiltonian dynamical systems. In this section we expose a special, orthogonal, symplectic basis associated with any linear(ized) symplectomorphism that retains its orthogonality under the linearized mapping.

Theorem 5.1 (the symplectic eigenskeleton). *Consider a symplectomorphism $\phi : M \to M$ that takes the initial point x to $\phi(x) = X$. Let $\Phi : T_xM \to T_XM$ be the STM of the mapping. Define now the matrix $\Psi = \Phi^T \cdot \Phi$ and let $\{\xi_1, \eta_1, \ldots, \xi_N, \eta_N\}$ be the orthonormal eigenbasis of Ψ. Then the following are true:*

1. *The eigenvectors of Ψ occur in pairs, which are related to each other by*

$$\eta_i = J \cdot \xi_i.$$

Additionally, the associated eigenvalue of η_i is λ_i^{-1} if λ_i is the eigenvalue associated with ξ_i.

2. *The linear transformation T that takes the standard basis to the eigenbasis of Ψ,*

$$T : \{\hat{p}_1, \hat{q}_1, \ldots, \hat{p}_N, \hat{q}_N\} \to \{\xi_1, \eta_1, \ldots, \xi_N, \eta_N\},$$

is symplectic. Moreover, the couples $\{\xi_i, \eta_i\}_{i=1}^N$ make symplectic pairs.

3. *The vectors $\{\Phi \cdot \xi_1, \Phi \cdot \eta_1, \ldots, \Phi \cdot \xi_N, \Phi \cdot \eta_N\}$ are orthogonal. Moreover,*

$$||\Phi \cdot \xi_i|| = \sqrt{|\lambda_i|} \qquad \textit{and} \qquad ||\Phi \cdot \eta_i|| = \sqrt{|\lambda_i^{-1}|}.$$

4. *If a $2k$-dimensional symplectic subvolume Υ is initially parallel to k of the eigenskeleton planes, then the linearized transformation Φ preserves its volume, i.e.,* $\mathrm{Vol}_{2k}(\Upsilon) = \mathrm{Vol}_{2k}(\Phi(\Upsilon))$.

We call the symplectic eigenbasis of the matrix $\Psi = \Phi^T \cdot \Phi$ the symplectic eigenskeleton *of the transformation ϕ_t.*

Proof.

1. Consider the ith eigenvector ξ_i of Ψ with eigenvalue λ_i,

$$\Psi \cdot \xi_i = \lambda_i \xi_i.$$

Taking the transpose of this equation, right-multiplying by $J \cdot \Psi$, and then recognizing the identity $\Psi^T \cdot J \cdot \Psi = J$, we see that

$$\xi_i^T \cdot J = \lambda_i \xi_i^T \cdot J \cdot \Psi.$$

Taking the transpose once more and multiplying by -1 (while noting $\Psi^T = \Psi$ and $J^T = -J$), we have

$$\Psi \cdot (J \cdot \xi_i) = \frac{1}{\lambda_i} J \cdot \xi_i.$$

Hence, the vectors $\eta_i = J \cdot \xi_i$ are also eigenvectors of Ψ, with eigenvalues λ_i^{-1}.

2. We define the matrices Ξ and N as follows:

$$\Xi = \begin{bmatrix} | & & | \\ \xi_1 & \cdots & \xi_N \\ | & & | \end{bmatrix} \qquad \text{and} \qquad \mathrm{N} = \begin{bmatrix} | & & | \\ \eta_1 & \cdots & \eta_N \\ | & & | \end{bmatrix},$$

where $N = J \cdot \Xi$ from part 1. The transformation matrix T can be represented as

$$T = \begin{bmatrix} | & | \\ \Xi & N \\ | & | \end{bmatrix}.$$

We have temporarily reordered our representation of the basis, so that

$$J = \begin{bmatrix} \mathbb{O}_N & -\mathbb{I}_N \\ \mathbb{I}_N & \mathbb{O}_N \end{bmatrix}.$$

Noting again that $N = J \cdot \Xi$, one easily sees

$$T^T \cdot J \cdot T = \begin{bmatrix} \Xi^T \cdot N & -\Xi^T \cdot \Xi \\ N^T \cdot N & -N^T \cdot \Xi \end{bmatrix}.$$

Due to the orthonormality of the eigenvectors (i.e., $\Xi^T \cdot N = N^T \cdot \Xi = \mathbb{O}_N$ and $\Xi^T \cdot \Xi = N^T \cdot N = \mathbb{I}_N$), this expression reduces to

$$T^T \cdot J \cdot T = J,$$

and hence the matrix T is symplectic.

3. Renaming the eigenvectors of Ψ as $\{v_i\}_{i=1}^{2N} = \{\xi_i, \eta_i\}_{i=1}^{N}$, we have from the orthonormality of the eigenbasis of Ψ

$$v_i \cdot v_j = \delta_{ij}.$$

But

$$(\Phi \cdot v_i) \cdot (\Phi \cdot v_j) = v_i^T \cdot \Phi^T \cdot \Phi \cdot v_j = \lambda_j \delta_{ij}.$$

Hence the vectors $\{\Phi \cdot v_i\}_{i=1}^{2N}$ are also orthogonal. Moreover,

$$||\Phi \cdot \xi_i|| = \sqrt{\xi_i^T \cdot \Phi^T \cdot \Phi \cdot \xi} = \sqrt{\lambda_i}.$$

Similarly

$$||\Phi \cdot \eta_i|| = \frac{1}{\sqrt{\lambda_i}}.$$

4. If Υ is a two-dimensional area element spanned by ξ_1 and η_1, then the area expansion is

$$\frac{\mathrm{Vol}_2(\Phi(\Upsilon))}{\mathrm{Vol}_2(\Upsilon)} = (\Phi \cdot \xi_1) \cdot (\Phi \cdot \eta_1) = \sqrt{\lambda_1}\frac{1}{\sqrt{\lambda_1}} = 1.$$

But since this is true of any area element initially parallel to one of the symplectic eigenskeleton planes, and the symplectic eigenskeleton $2k$-volumes are simply direct sums of these area elements, the result follows. ∎

We previously showed, for an arbitrary orthogonal symplectic basis of the tangent space at an initial point (p, q), that the basis vectors tend to collapse toward a particular direction or subspace in a manner commensurate with local differential expansions of subvolumes. Despite this general feature, which is especially prevalent in chaotic Hamiltonian dynamical systems, for a fixed final time there *always* exists a preferred orthogonal symplectic basis of the tangent space whose basis vectors return to an orthogonal configuration at the final time.

We would further like to mention that there is an interesting link between the matrix Ψ and the Lyapunov exponents of the system. If the final endpoint is allowed to vary along the nominal trajectory instead of being held fixed, then $\Psi(t) = \Phi(t)^T\Phi(t)$ can be thought of as a matrix-valued function of time. It has been shown previously that the limit

$$\lim_{t\to\infty} \Psi(t)^{1/2t}$$

exists and, moreover, the logarithm of the eigenvalues of this limit are the Lyapunov exponents for the system. For more details, see Dieci, Russell, and Van Vleck [4], Habib and Ryne [6], Janaki et al. [8], and Rangarajan, Habib, and Ryne [16].

6. Tracking a hyperbolic flyby in the earth-moon system. As an application to the theory of area expansion factors and projections, we will consider the case of the hyperbolic flyby of an asteroid in the earth-moon system, where the asteroid's initial trajectory is not coherent with the plane of the lunar orbit. Let $\mu_\oplus$ and $\mu_\leftmoon$ be the gravitational parameters of the earth and moon, respectively, and a the radius of the moon's (circular) orbit. The angular velocity of the moon is thus taken to be

$$\omega = \sqrt{\frac{\mu_\oplus}{a^3}},$$

so that the moon's position is given by

$$\mathbf{r}_\leftmoon = \langle a\cos(\omega t), a\sin(\omega t), 0\rangle.$$

Consider the position of the asteroid to be given in spherical coordinates as

$$\mathbf{r}_* = \langle r\sin\theta\cos\varphi, r\sin\theta\sin\varphi, r\cos\theta\rangle,$$

where θ is the polar angle and ϕ the azimuthal angle. We now define the function $\psi(r,\theta,\varphi,t)$, for ease of notation, as

$$\psi = |\mathbf{r}_* - \mathbf{r}_\leftmoon| = \sqrt{r^2 + a^2 - 2ra\sin\theta\cos\Omega}, \tag{6.1}$$

where we have defined the selenecentric azimuth as

$$\Omega = \varphi - \omega t. \tag{6.2}$$

The conjugate momenta are given by

$$p_r = \dot{r}, \qquad p_\theta = r^2\dot{\theta}, \qquad p_\varphi = r^2\sin^2\theta\dot{\varphi}, \tag{6.3}$$

so that $\langle p_r, r\rangle$, $\langle p_\theta, \theta\rangle$, and $\langle p_\varphi, \varphi\rangle$ form symplectic pairs. The Hamiltonian for this system is

$$H = \frac{1}{2}\left\{p_r^2 + \frac{p_\theta^2}{r^2} + \frac{p_\varphi^2}{r^2\sin^2\theta}\right\} - \frac{\mu_\oplus}{r} - \frac{\mu_\leftmoon}{\psi(r,\theta,\varphi,t)}. \tag{6.4}$$

Ordering the phase space variables as

$$\mathbf{x} = \langle p_r, r, p_\theta, \theta, p_\varphi, \varphi\rangle,$$

the dynamical equations of motion for the state $\mathbf{x}$ and the STM Φ are given by

$$\dot{\mathbf{x}} = J\cdot\frac{\partial H}{\partial \mathbf{x}} \qquad \text{and} \qquad \dot{\Phi} = J\cdot\frac{\partial^2 H}{\partial \mathbf{x}^2}\cdot\Phi.$$

(These equations are written out explicitly in the appendix.) This is is a system of 42 coupled, first order, nonlinear differential equations which can be integrated numerically. We normalize the problem by taking $\mu_\oplus = 1$, $\mu_\leftmoon = 0.1111$, and $a = 0.5109$. We will then integrate the initial

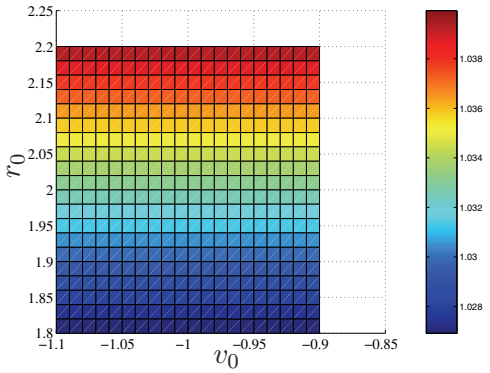


Figure 2. $\sqrt{\mathfrak{G}(L)}$ *plotted against the* $\langle p_r, r\rangle$ *projection of the initial surface.*

uncertainty distribution, given below, to a time $t_f = 3$. Given our normalization of the problem, this final time corresponds to approximately 40 days.

For initial conditions we consider a two-dimensional surface specified by $\langle p_r, r\rangle_0 \in [-1.1, -0.9] \times [1.8, 2.2]$, and $\theta_0 = -2$, $\dot{\theta}_0 = 0.05$, $\varphi = 0$, $\dot{\varphi}_0 = 0.05$. Since it is the initial angular rates and not the initial conjugate momenta we are taking as given, our initial surface will not be parallel to the $\langle p_r, r\rangle$ plane in the symplectic space, due to the dependence of the symplectic momenta on r given by (6.3). Such an initial region arises naturally in asteroid and space debris tracking problems. One can directly measure the angles and angular rates of an asteroid from an earth-based optical telescope. The conjugate momenta cannot be directly observed. The region in the $\langle \dot{r}, r\rangle$ space is limited only by a few physical assumptions, e.g., when tracking space debris the geocentric energy of the observed body must be negative, etc. See Maruskin, Scheeres, and Alfriend [11] and Tommei, Milani, and Rossi [19] for more detail.

We begin by discretizing the uncertainty domain into a 20×20 grid of virtual asteroids (VAs). We will consider the surface as being parameterized by its initial projection on the $\langle p_r, r\rangle$ plane. Given the specification of the initial surface described above, and the definition of the conjugate momenta (6.3), the differential of the parameterization map (3.1) is given by the matrix

$$L(p_r, r) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 2r\dot{\theta}_0 \\ 0 & 0 \\ 0 & 2r\sin^2\theta_0\dot{\varphi}_0 \\ 0 & 0 \end{bmatrix},$$

where θ_0, $\dot{\theta}_0$, and $\dot{\phi}_0$ are the constants given above. The square root of the Gram determinant $\mathfrak{G}(L)$ is plotted against the parameterization domain in Figure 2. This represents the area expansion factor from the parameterization domain to the actual initial surface, which lives in the six-dimensional phase space. Since the asteroid is equally likely to be found anywhere on the initial two-dimensional surface, the likelihood that the actual asteroid is initially inside any given square is proportional to the area expansion factor $\sqrt{\mathfrak{G}(L)}$.

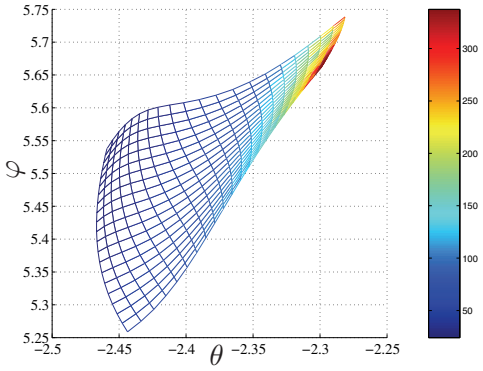


Figure 3. *Probability density in* (θ, ϕ) *observation spaces.*

We can now integrate the trajectory and STM for each VA in the discretization over the time interval $t \in [0, 3]$. Suppose that our initial uncertainty distribution corresponds to an optical sighting of a hyperbolic asteroid at time $t = 0$, and we wish to locate the asteroid in the sky at time $t = 3$ (roughly 40 days later). Similar to (2.4), the transpose of the differential of the projection mapping onto the physical observation space, i.e., the $\langle \theta, \phi \rangle$ plane, is defined as

$$\Pi_{\theta\phi} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

The probability density of sighting the asteroid at a given angular location in the sky follows analogously from (3.2) and the discussion in subsection 3.2.2 and is given by

$$\sigma(\theta, \phi) = \frac{1}{\Delta p_{r0} \Delta r_0 \sum \sqrt{\mathfrak{G}(L)}} \cdot \frac{\sqrt{\mathfrak{G}(L)}}{|\det(\Pi_{\theta\phi}^T \cdot \Phi \cdot L)|},$$

where Δp_{r0} and Δr_0 are the step sizes of the discretization of the parameterization domain, i.e., the initial (p_r, r) plane shown in Figure 2. The term $\Delta p_{r0} \Delta r_0 \sum \sqrt{\mathfrak{G}(L)}$ itself (a constant) is just a Riemann sum for the surface area of the initial two-dimensional uncertainty region in six-dimensional phase space. This probability density is plotted in Figure 3. This can be compared with the results in Milani, Gronchi, and Vitturi [13], in which they outlined the process of computing a Delaunay triangulation of a more general initial uncertainty distribution, mapping the triangulation to the final time. Our approach shows how one may quantify the probability density of an asteroid's angular location, given a previous optical measurement of the asteroid's location.

We now turn our attention to the projection of the uncertainty region onto the symplectic planes. As discussed in section 2, the determinants of the submatrices of the STM Φ, M_{ij}, have a specific geometric meaning. Since we parameterize the initial surface by its projection onto

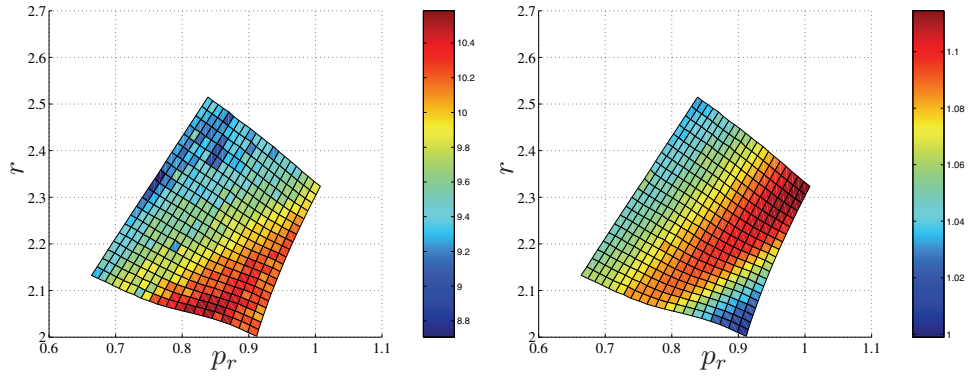


Figure 4. *Probability density (left) and M_{11} (right) in the $\langle p_r, r\rangle$ symplectic plane.*

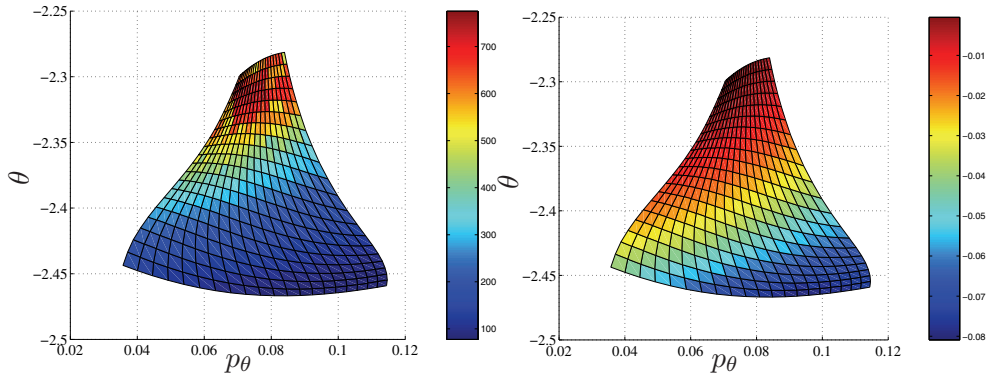


Figure 5. *Probability density (left) and M_{21} (right) in the $\langle p_\theta, \theta\rangle$ symplectic plane.*

the $\langle p_r, r\rangle$ plane, we will focus our attention on $j = 1$. If you attach a differential area element to any point of the initial surface so that the area element is parallel to the $\langle p_r, r\rangle$ plane (and therefore *not* tangent to the surface itself), the area of the projection of the mapped area element onto the ith symplectic plane will be M_{i1} times its initial area. The determinants M_{11}, M_{21}, and M_{31} are plotted against the projection of the final surface onto the $\langle p_r, r\rangle$, $\langle p_\theta, \theta\rangle$, and $\langle p_\phi, \phi\rangle$ symplectic planes in Figures 4(right), 5(right), and 6(right), respectively. As is confirmed numerically, the sum $M_{11} + M_{21} + M_{31} = 1$; see (2.6). Moreover, the signs of these determinants tell whether the projection of the surface onto each of the symplectic planes is positively or negatively oriented. The projection of the surface onto the $\langle p_r, r\rangle$ plane is positively oriented, and the projection onto the $\langle p_\theta, \theta\rangle$ plane is negatively oriented. We also see that the surface has begun to fold over on itself, so that the projection of the surface onto the $\langle p_\phi, \phi\rangle$ plane is no longer 1-to-1. We see that most of the surface projection has positive orientation; however, the small sliver that overlaps has a negative orientation, caused by the fold.

Additionally, one may compute the probability densities of the projection of the surface onto each of the symplectic planes, given by (3.2). In the $\langle p_r, r\rangle$ projection, the asteroid is more likely to be found in the lower corner than in the upper. In the $\langle p_\theta, \theta\rangle$ projection, the

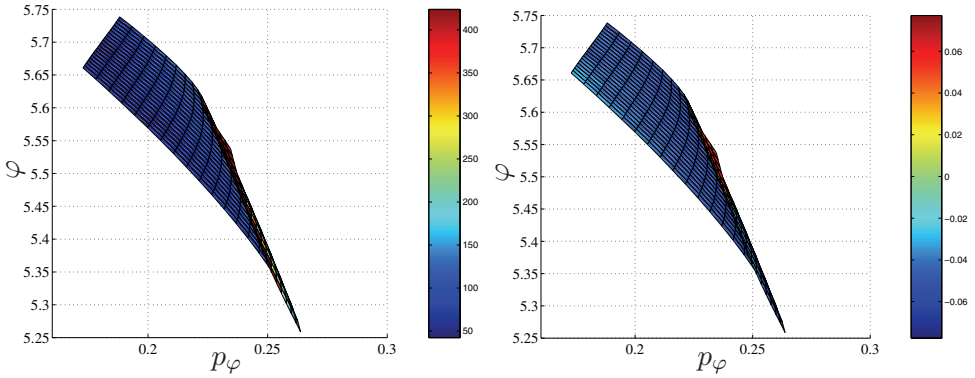


Figure 6. *Probability density (left) and M_{31} (right) in the $\langle p_\phi, \phi \rangle$ symplectic plane.*

higher probability density belongs to the upper area that is more compressed.

Along the fold, the actual surface area elements tangent to the surface along those points are vertical with respect to the $\langle p_\phi, \phi \rangle$ symplectic plane. Due to the $\det(\Pi_3^T \cdot \Phi \cdot L)$ in the denominator of (3.2), one sees that the fold actually causes a singularity in the probability density along the one-dimensional projection of the fold onto that symplectic plane. This is allowable since the probability is infinite only on a set of measure zero, i.e., the projection of the fold line.

7. Conclusion. We showed how the expansion of subvolumes in the local neighborhood of a nominal trajectory leads to the local collapse of the supporting phase space. Moreover, we produced a preferred basis, the symplectic eigenskeleton, which resists collapse and returns uncertainty distributions that are initially parallel to the basis to their minimal uncertainty state at a fixed final time.

Appendix: Equations of motion for the hyperbolic flyby. In section 6 we discussed the motion of a particle in a 3-degree-of-freedom system defined by the time-dependent Hamiltonian (6.4), which models, assuming appropriate initial conditions, the hyperbolic flyby of an asteroid in the earth-moon system. The equations of motion for this system, including the variational equations which determine the evolution of the 6×6 STM, are shown explicitly below. A standard set of spherical coordinates (r, θ, ϕ) is used, where θ and ϕ are the azimuthal and polar angles, respectively, and ψ, Ω, p_r, p_θ, and p_ϕ are defined in (6.1)–(6.3), respectively. The equations of motion work out to be

$$\dot{r} = p_r, \qquad \dot{\theta} = \frac{p_\theta}{r^2}, \qquad \dot{\varphi} = \frac{p_\varphi}{r^2 \sin^2\theta},$$

$$\dot{p}_r = \frac{p_\theta^2}{r^3} + \frac{p_\varphi^2}{r^3 \sin^2\theta} - \frac{\mu_\oplus}{r^2} - \frac{\mu_\leftmoon (r - a \sin\theta \cos\Omega)}{\psi^{3/2}},$$

$$\dot{p}_\theta = \frac{p_\varphi^2 \cos\theta}{r^2 \sin^3\theta} + \frac{\mu_\leftmoon r a \cos\theta \cos\Omega}{\psi^{3/2}},$$

$$\dot{p}_\varphi = \frac{\mu_\leftmoon r a \sin\theta \sin\Omega}{\psi^{3/2}}.$$

The coefficient matrix in the variational equations (2.3) for the STM is given by

$$J \cdot H_{xx} = \begin{pmatrix} 0 & -H_{rr} & -H_{rp_\theta} & -H_{r\theta} & -H_{rp_\varphi} & -H_{r\varphi} \\ H_{p_r p_r} & 0 & 0 & 0 & 0 & 0 \\ 0 & -H_{r\theta} & 0 & -H_{\theta\theta} & -H_{\theta p_\varphi} & -H_{\theta\varphi} \\ 0 & H_{rp_\theta} & H_{p_\theta p_\theta} & 0 & 0 & 0 \\ 0 & -H_{r\varphi} & 0 & -H_{\theta\varphi} & 0 & -H_{\varphi\varphi} \\ 0 & H_{rp_\varphi} & 0 & H_{\theta p_\varphi} & H_{p_\varphi p_\varphi} & 0 \end{pmatrix},$$

with

$$H_{p_r p_r} = 1, \qquad H_{p_\theta p_\theta} = r^{-2}, \qquad H_{p_\varphi p_\varphi} = r^{-2}\mathrm{s}^{-2}\theta,$$

$$H_{rr} = \frac{3p_\theta^2}{r^4} + \frac{3p_\varphi^2}{r^4\mathrm{s}^2\theta} - \frac{2\mu_\oplus}{r^3} - \frac{3\mu_\leftmoon(r - a\mathrm{s}\theta\mathrm{c}\Omega)^2}{\psi^{5/2}} + \frac{\mu_\leftmoon}{\psi^{3/2}},$$

$$H_{rp_\theta} = \frac{-2p_\theta}{r^3}, \qquad H_{rp_\varphi} = \frac{-2p_\varphi}{r^3\mathrm{s}^2\theta},$$

$$H_{r\theta} = \frac{2p_\varphi^2\mathrm{c}\theta}{r^3\mathrm{s}^3\theta} + \frac{3\mu_\leftmoon(r - a\mathrm{s}\theta\mathrm{c}\Omega)ra\mathrm{c}\theta\mathrm{c}\Omega}{\psi^{5/2}} - \frac{\mu_\leftmoon a\mathrm{c}\theta\mathrm{c}\Omega}{\psi^{3/2}},$$

$$H_{r\varphi} = \frac{3\mu_\leftmoon(r - a\mathrm{s}\theta\mathrm{c}\Omega)ra\mathrm{s}\theta\mathrm{s}\Omega}{\psi^{5/2}} - \frac{\mu_\leftmoon a\mathrm{s}\theta\mathrm{s}\Omega}{\psi^{3/2}},$$

$$H_{\theta\theta} = \frac{3p_\varphi^2\mathrm{c}^2\theta}{r^2\mathrm{s}^4\theta} + \frac{p_\varphi^2}{r^2\mathrm{s}^2\theta} - \frac{3\mu_\leftmoon r^2a^2\mathrm{c}^2\theta\mathrm{c}^2\Omega}{\psi^{5/2}} + \frac{\mu_\leftmoon ra\mathrm{s}\theta\mathrm{c}\Omega}{\psi^{3/2}},$$

$$H_{\theta p_\varphi} = \frac{-2p_\varphi\mathrm{c}\theta}{r^2\mathrm{s}^3\theta},$$

$$H_{\theta\varphi} = -\frac{3\mu_\leftmoon r^2a^2\mathrm{c}\theta\mathrm{s}\theta\mathrm{c}\Omega\mathrm{s}\Omega}{\psi^{5/2}} - \frac{\mu_\leftmoon ra\mathrm{c}\theta\mathrm{s}\Omega}{\psi^{3/2}},$$

where we have abbreviated sin and cos with s and c, respectively. We now have a coupled system of 42 first order differential equations which can be integrated numerically for each of a sampling of virtual particles distributed throughout the initial uncertainty region.

Acknowledgments. We would like to thank Professor Mario Bonk, who introduced us to Wirtinger's inequality and who provided the initial proof of the area expansion of parasymplectic surfaces theorem.

REFERENCES

[1] V. I. Arnold, *Mathematical Methods of Classical Mechanics*, 2nd ed., Springer-Verlag, New York, 1989.
[2] M. Audin, A. C. da Silva, and E. Lerman, *Symplectic Geometry of Integrable Hamiltonian Systems*, Birkhäuser Boston, Cambridge, MA, 2003.
[3] A. M. Bloch, *Nonholonomic Mechanics and Control*, Springer-Verlag, New York, 2003.

[4] L. Dieci, R. D. Russell, and E. S. Van Vleck, *On the computation of Lyapunov exponents for continuous dynamical systems*, SIAM J. Numer. Anal., 34 (1997), pp. 402–423.
[5] D. T. Greenwood, *Classical Dynamics*, Dover, New York, 1997.
[6] S. Habib and R. D. Ryne, *Symplectic computation of Lyapunov exponents*, in Nonlinear Evolution Equations and Dynamical Systems: Needs '94 (Los Alamos, NM), World Scientific Publishing, River Edge, NJ, 1995, pp. 283–294.
[7] H. Hofer and E. Zehnder, *Symplectic Invariants and Hamiltonian Dynamics*, Birkhäuser-Verlag, Basel, Switzerland, 1994.
[8] T. M. Janaki, G. Rangarajan, S. Habib, and R. D. Ryne, *Computation of the Lyapunov spectrum for continuous-time dynamical systems and discrete maps*, Phys. Rev. E, 60 (1999), pp. 6614–6626.
[9] A. J. Lichtenberg and M. A. Lieberman, *Regular and Chaotic Dynamics*, 2nd ed., Appl. Math. Sci. 38, Springer, New York, 1992.
[10] J. E. Marsden and T. S. Ratiu, *Introduction to Mechanics and Symmetry*, Springer-Verlag, New York, 1994.
[11] J. M. Maruskin, D. J. Scheeres, and K. T. Alfriend, *Correlation of optical observations of objects in earth orbit*, J. Guidance, Control, and Dynamics, to appear.
[12] D. McDuff and D. Salamon, *Introduction to Symplectic Topology*, Oxford Science Publications, London, 1998.
[13] A. Milani, G. F. Gronchi, M. d'M. Vitturi, *Orbit Determination with Very Short Arcs*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2004.
[14] S. Morita, *Geometry of Differential Forms*, AMS, Providence, RI, 1998.
[15] R. Osserman, *A Survey of Minimal Surfaces*, 2nd ed., Dover, New York, 1986.
[16] G. Rangarajan, S. Habib, and R. D. Ryne, *Lyapunov exponents without rescaling and reorthogonalization*, Phys. Rev. Lett., 80 (1998), pp. 3747–3750.
[17] D. J. Scheeres, F. Y. Hsiao, R. S. Park, B. Villac, and J. M. Maruskin, *Fundamental limits on spacecraft orbit uncertainty and distribution propagation*, J. Astronaut. Sci., 54 (2006), pp. 505–523.
[18] A. Silva, *Lectures on Symplectic Geometry*, Springer, New York, 2001.
[19] G. Tommei, A. Milani, and A. Rossi, *Orbit determination of space debris: Admissible regions*, Celestial Mech. Dynam. Astronom., 97 (2007), pp. 289–304.

SIAM J. Applied Dynamical Systems
Vol. 8, No. 1, pp. 202–221
© 2009 Society for Industrial and Applied Mathematics

Constructing Generalized Synchronization Manifolds by Manifold Equation*

Jie Sun†, Erik M. Bollt†, and Takashi Nishikawa†

Abstract. Full understanding of synchronous behavior in coupled dynamical systems beyond the identical case requires an explicit construction of the generalized synchronization manifold, whether we wish to compare the systems or to understand their stability. Nonetheless, while synchronization has become an extremely popular topic, the bulk of the research in this area has been focused on the identical case, specifically because its invariant manifold is simply the identity function, and there have yet to be any generally workable methods for computing the generalized synchronization manifolds for nonidentical systems. Here, we derive time dependent PDEs whose stationary solution mirrors exactly the generalized synchronization manifold, respecting its stability. We introduce a novel method for dealing with subtle issues with boundary conditions in the numerical scheme to solve the PDE, and we develop first order expansions close to the identical case. We give several examples of increasing sophistication, including coupled nonidentical Van der Pol oscillators. By using the manifold equation, we also discuss the design of coupling to achieve desired synchronization.

Key words. generalized synchronization, computational invariant manifold, coupled dynamical systems

AMS subject classifications. 37D45, 37N35

DOI. 10.1137/080721340

1. Introduction.

Following the fundamental finding in [2] that even chaotic systems can synchronize through coupling, synchronization phenomena in coupled oscillator systems have been studied extensively during the past few years, in the context of dynamical systems [3, 4, 5, 6, 7, 8, 15, 17, 18], control theory [10], complex networks [11, 19, 20], laser physics [15, 18], etc. Besides identical synchronization, studies have been carried out to describe other types of synchronization, such as generalized synchronization [3, 4, 6, 18], phase synchronization [5, 8], and anticipated and lag synchronization [8, 15].

The simplest form of synchronization between two oscillators is complete (or identical) synchronization, meaning that, after a transient time, the difference between the two oscillators approaches zero. However, a more generalized form of synchronization can often occur between two nonidentical systems.

In the combined state space of such systems, an invariant manifold consisting of the synchronized states, termed *generalized synchronization manifold*, is a central object of study in this context, and its form [12] and stability [9, 16] are two of its important properties. The former tells the relationship between corresponding dynamic variables in the synchronization state, and the latter determines whether a trajectory near the manifold evolves closer or farther

*Received by the editors April 15, 2008; accepted for publication (in revised form) by T. Sauer October 21, 2008; published electronically January 23, 2009.
http://www.siam.org/journals/siads/8-1/72134.html
†Department of Mathematics and Computer Science, Clarkson University, Potsdam, NY 13699-5815 (sunj@clarkson.edu, bolltem@clarkson.edu, tnishika@clarkson.edu). The first two authors were supported by the Army Research Office under 51950-MA. The second author was also supported by the National Science Foundation under DMS-0708083 and DMS-0404778.

away. While a great deal of work in synchronization focuses on identical synchronization, primarily because the synchronization manifold in this case is described simply by the identity function, much less work (mainly on the detection of existence and stability of generalized synchronization; see [3, 4, 6]) has been carried out on the ubiquitous generalized synchronization phenomenon, specifically because there has been little work to explicitly construct the corresponding synchronization manifold. One possibility, as reported in [13], is to approximate the manifold using orthogonal functions expansion.

Although one can simply run the system to find a portion of the generalized synchronization manifold that is covered by the attractor of the coupled system, the study of the entire manifold is still necessary and important. One reason is that the transformation from one state variable to the other that defines the manifold contains information on the correlation between the dynamics of the two systems in the whole state space, not just on the attractor. Furthermore, the manifold outside the attractor may embed relevant unstable invariant sets, which can become stable as parameters change.

Our recent work strives to fill this gap. In this paper we present a PDE (manifold equation) which describes the invariant manifold of generalized synchronization between two coupled oscillators as well as a variational equation regarding the transverse stability of the manifold. We also develop a time dependent PDE whose stationary solution is a generalized synchronization manifold, and present straightforward numerical schemes to solve the equation. Several examples, such as 1D and 2D coupled oscillators and coupled Van der Pol oscillators, are used to illustrate our method of constructing the invariant manifold under generalized synchronization. The design of coupling in a coupled system to achieve desired synchronization is also discussed by the use of the manifold equation.

In section 2, we give background regarding generalized synchronization, leading to the manifold equation together with discussion of stability of the manifold and perturbation analysis relative to the identical case. In section 3 we show a specific form of the manifold equation regarding coupled oscillator systems with almost the same individual dynamics and a time dependent PDE used to solve for the original equation. In section 4 we show some examples of invariant manifolds. In section 5 we show how to use the manifold equation to design a coupling function between two oscillators to obtain the desired form of synchronization.

2. Equations for the synchronization manifold. For the discussion in this paper, consider the following equations to describe the dynamics of two nonidentical oscillators w_1 and w_2:

$$\begin{aligned} \dot{w}_1 &= f(w_1, w_2, \mu_1), \\ \dot{w}_2 &= g(w_1, w_2, \mu_2). \end{aligned} \tag{2.1}$$

Here $w_1 \in \Re^m$, $w_2 \in \Re^m$, $f : \Re^m \times \Re^m \times \Re^1 \to \Re^m$, and $g : \Re^m \times \Re^m \times \Re^1 \to \Re^m$, where $f \in C^1$ and $g \in C^1$ (both f and g have continuous first order derivative). Note that the coupling functions which enable the two oscillators to synchronize have been included in the general form of the functions f and g. μ_1 and μ_2 are considered as parameters of the functions.

In this paper, we will focus on perturbations from identical synchronization. In such a case, we require that $f(w_1, w_1, \mu_1) = g(w_1, w_1, \mu_1)$, meaning that when the two systems have the same parameters and same value of variables, they have the same individual dynamics, often with no effective coupling between them.

2.1. Manifold equation. Let (w_{1s}, w_{2s}) denote the synchronization state. The synchronization manifold at synchronization state $w_2 = \Phi(w_1)$ is the set $\{(w_1, w_2) \mid w_2 = \Phi(w_1)\}$, where Φ is a time independent transformation. In this paper we will abbreviate this notation and just use the relationship Φ to indicate the form of this manifold. To find the synchronization manifold of this coupled system, we mean to find some C^1 function Φ ($\Phi : \Re^m \to \Re^m$) such that

$$w_{2s} = \Phi(w_{1s}). \tag{2.2}$$

Suppose such a Φ exists.

Then, along the synchronization manifold,

$$\dot{w_{2s}} = \frac{d\Phi(w_{1s})}{dt} = D\Phi(w_{1s}) \cdot \dot{w_{1s}}. \tag{2.3}$$

Using (2.1) in (2.3) for the synchronization state, and replacing w_{1s} by w_1 for convenience, we obtain the following *manifold equation*:

$$D\Phi(w) \cdot f(w, \Phi(w), \mu_1) = g(w, \Phi(w), \mu_2). \tag{2.4}$$

This is a PDE with unknown function $\Phi = (\Phi_1, \ldots, \Phi_m)$ which is of dimension m, and each component Φ_i has m variables and has been used for the study of generalized synchronization [13] and center manifold approximation [1].

In the case that $\mu_1 = \mu_2 = \mu$ (coupled identical oscillators), a particular solution of (2.3) is $\Phi(w) = w$, since $f(w, w, \mu) = g(w, w, \mu)$, which describes the identical manifold under complete synchronization. However, in general cases, it is much harder to find a solution to this PDE. We will leave it to the following to discuss some appropriate approaches in obtaining the solution.

2.2. Stability of the manifold: Variational equation. In addition to the form of the manifold, the stability plays another important role in the study of synchronization. To measure the stability of the synchronization manifold $w_{2s} = \Phi(w_{1s})$ obtained by (2.4), define

$$\xi \equiv w_2 - \Phi(w_1) \tag{2.5}$$

to be the transverse perturbation from the manifold. Then from (2.1), (2.3), and (2.4), we have

$$\begin{aligned}
\dot{\xi} &= \dot{w_2} - D\Phi(w_1)\dot{w_1} \\
&= g(w_1, w_2, \mu_2) - D\Phi(w_1) f(w_1, w_2, \mu_1) \\
&= g(w_1, \Phi(w_1) + \xi, \mu_2) - g(w_1, \Phi(w_1), \mu_2) + g(w_1, \Phi(w_1), \mu_2) - D\Phi(w_1) f(w_1, \Phi(w_1), \mu_1) \\
&\quad + D\Phi(w_1) f(w_1, \Phi(w_1), \mu_1) - D\Phi(w_1) f(w_1, \Phi(w_1) + \xi, \mu_1).
\end{aligned} \tag{2.6}$$

The term $g(w_1, \Phi(w_1), \mu_2) - D\Phi(w_1) f(w_1, \Phi(w_1), \mu_1) = 0$ by the manifold (2.4), so expanding the other terms to the first order in ξ, we obtain the *local variational equation*

$$\dot{\xi} = [D_{w_2} g(w_1, w_2, \mu_2) - D\Phi(w_1) D_{w_2} f(w_1, w_2, \mu_1)]|_{w_2 = \Phi(w_1)} \xi. \tag{2.7}$$

This variational equation can be used to study the local stability at the points on the invariant manifold. However, we note that when this equation does not give a uniformly asymptotically stable solution, the asymptotic stability of synchronization may not be correctly predicted from this equation; see examples given in [14]. From (2.7) we again see the importance of the knowledge of the manifold $w_{2s} = \Phi(w_{1s})$. Although there are approaches discussing the stability of the synchronization manifold without explicitly knowing the form of the manifold, it will be more natural and convenient to study the stability directly with the knowledge of the form of the synchronization manifold.

In this paper, we will focus on the form of the invariant manifold. Detailed discussion regarding the stability of the invariant manifold under generalized synchronization will be reported in a future work.

2.3. Specific form of the manifold equation for perturbed oscillators. Suppose that the individual dynamics for the two oscillators w_1 and w_2 in isolation are almost the same, but with a small difference. To be explicit, we mean that in (2.1) there is a small mismatch in the parameters: $\mu_2 = \mu_1 + \epsilon$. Then, one would expect the synchronization manifold to be close to the identical synchronization manifold and Φ to approach the identity function as $\epsilon \to 0$. We empirically expect the synchronization manifold to have the form

$$w_{2s} = \Phi(w_{1s}) = w_{1s} + \epsilon H(w_{1s}) + O(\epsilon^2). \tag{2.8}$$

The main interest will be to find H, and thus we will have a first order approximation for the synchronization manifold. Following a procedure similar to the derivation of the manifold (2.4), after neglecting higher order terms we obtain a PDE for H:

$$\begin{aligned} DH(w_1) f(w_1, w_1, \mu_1) = {} & D_{\mu_2} g(w_1, w_1, \mu_2)|_{\mu_2=\mu_1} \\ & + D_{w_2}[g(w_1, w_2, \mu_1) - f(w_1, w_2, \mu_1)]|_{w_2=w_1} H(w_1). \end{aligned} \tag{2.9}$$

3. Solving the manifold equation in practice. In the previous section we derived the general equation for the synchronization manifold of nonidentical oscillators as well as the corresponding variational equation for the study of stability of the manifold. We conclude that in order for such transformation $w_2 = \Phi(w_1)$ to exist when $t \to \infty$ (i.e., Φ is a stationary solution of (3.2) when t is sufficiently large), Φ has to satisfy (2.4). So the problem of finding the invariant manifold reduces to finding the solution of the PDE (2.4) with appropriate boundary conditions.

However, the boundary condition for (2.4) is not easily attainable without actually solving the original ODE system. There may exist more than one synchronization manifold in general, if none of them is globally stable. Our main interest in this paper is to use the perturbed manifold equation (2.8) to find the closest one to the identical manifold, which tells us how the form of synchronization changes if the two oscillators are not perfectly the same.

3.1. A time dependent PDE for the manifold. In order to develop a workable numerical scheme to solve the manifold equation (2.4), we first derive an evolution equation for a mapping relating the coordinates w_1 and w_2 of the two systems obeying (2.1). The idea is that if we start with a reasonable initial guess for the mapping and evolve it according to the dynamics of the coupled system, the mapping will approach the synchronization manifold, provided that

the manifold is asymptotically stable. Indeed, the synchronization manifold turns out to be a stationary solution of the evolution equation, whose stability mirrors exactly the stability of the synchronized systems.

We assume the existence of a smooth time dependent mapping from the phase variable w_1 to w_2, i.e.,

$$w_2 = \phi(w_1, t), \tag{3.1}$$

as an ansatz. Differentiating this with respect to t and using (2.1), we see that $\phi(w_1, t)$ satisfies

$$\frac{\partial \phi}{\partial t} + \frac{\partial \phi}{\partial w_1} f(w_1, \phi, \mu_1) = g(w_1, \phi, \mu_2), \tag{3.2}$$

which is a PDE that describes the time evolution of the functional relationship between the states of the two systems. If we set $\frac{\partial \phi}{\partial t} = 0$, the equation reduces to (2.4), so a synchronization manifold, if it exists, is a stationary solution. Since the evolution equation (3.2) reflects the coevolution of the coupled systems, we expect a synchronization manifold to be asymptotically stable with respect to (3.2), i.e.,

$$\Phi(w_1) = \lim_{t \to \infty} \phi(w_1, t), \tag{3.3}$$

whenever synchronization is asymptotically stable and the initial condition for (3.2) is within the basin of attraction.

If we look for a synchronization manifold of the form (2.8), then we can write the time dependent manifold as

$$\phi(w_1, t) = w_1 + \epsilon h(w_1, t) + O(\epsilon^2) \approx w_1 + \epsilon h(w_1, t). \tag{3.4}$$

Neglecting higher order terms in ϵ we can derive the corresponding time dependent PDE for the function $h(w_1, t)$:

$$\begin{aligned} \frac{\partial h}{\partial t} &= D_{\mu_2} g(w_1, w_1, \mu_2)|_{\mu_2 = \mu_1} + D_{w_2}[g(w_1, w_2, \mu_1) - f(w_1, w_2, \mu_1)]|_{w_2 = w1} h \\ &\quad - \frac{\partial h}{\partial w_1} f(w_1, w_1, \mu_1). \end{aligned} \tag{3.5}$$

The stationary state of h gives H corresponding to the synchronization manifold:

$$H(w_1) = \lim_{t \to \infty} h(w_1, t). \tag{3.6}$$

3.2. An iterative scheme to find the stationary manifold. One approach to solving (3.5) is by successive approximation: start with some initial form and evolve the solution according to the equation until the approximation gets close enough to the true solution. To obtain an iterative scheme to approximate the synchronization manifold, we first discretize t in (3.4) by considering $t = t_0, t_1, \ldots$, where $t_0 = 0$ and $t_{n=1} - t_n := \tau_n$. Let us use $h_n(w_1)$ to represent $h(w_1, t_n)$. Also, for convenience, let

$$b(w_1) := D_{\mu_2} g(w_1, w_1, \mu_2)|_{\mu_2 = \mu_1}, \tag{3.7}$$

$$B(w_1) := D_{w_2}[g(w_1, w_2, \mu_1) - f(w_1, w_2, \mu_1)]|_{w_2 = w_1}. \tag{3.8}$$

Then the iteration scheme can be described as follows:

- Given initial guess h_0.
- For $n = 1, 2, \ldots$, until convergence, do

$$h_n = h_{n-1} + \tau_{n-1}\left[b(w_1) + B(w_1)h_{n-1} - \frac{\partial h_{n-1}}{\partial w_1} f(w_1, w_1, \mu_1)\right]. \tag{3.9}$$

To analyze the convergence of this scheme, let h^* denote a fixed point of our iteration operator, i.e.,

$$h^* = h^* + \tau_{n-1}\left[b(w_1) + B(w_1)h^* - \frac{\partial h^*}{\partial w_1} f(w_1, w_1, \mu_1)\right]. \tag{3.10}$$

Then we have

$$\begin{aligned} h_n - h^* &= h_{n-1} - h^* + \tau_{n-1}\left[B(w_1)(h_{n-1} - h^*) - \frac{\partial (h_{n-1} - h^*)}{\partial w_1} f(w_1, w_1, \mu_1)\right] \\ &= (I_m + \tau_{n-1}B_{w_1})(h_{n-1} - h^*) - \tau_{n-1}\frac{\partial (h_{n-1} - h^*)}{\partial w_1} f(w_1, w_1, \mu_1) \end{aligned} \tag{3.11}$$

so that

$$||h_n - h^*|| \le ||I_m + \tau_{n-1}B_{w_1}|| \cdot ||h_{n-1} - h^*|| + |\tau_{n-1}| \cdot \left|\left|\frac{\partial (h_{n-1} - h^*)}{\partial w_1}\right|\right| \cdot ||f(w_1, w_1, \mu_1)|| \tag{3.12}$$

for appropriate matrix and vector norms.

The relevance of each term in this error propagation equation is interpreted as follows. The first term $||I_m + \tau_{n-1}B_{w_1}|| \cdot ||h_{n-1} - h^*||$ has propagation factor $||I_m + \tau_{n-1}B_{w_1}||$ smaller than 1 if synchronization is stable, while the second term $|\tau_{n-1}| \cdot ||\frac{\partial (h_{n-1} - h^*)}{\partial w_1}|| \cdot ||f(w_1, w_1)||$ corresponds to the total variation in space of the error function, which in general is not guaranteed to converge to 0.

The difficulty of controlling the variation part in the second term above suggests the following adaptive step size scheme:

$$h_n = h_{n-1} + \tau_{n-1}\left[b(w_1) + B(w_1)h_{n-1} - \alpha_{n-1}\frac{\partial h_{n-1}}{\partial w_1} f(w_1, w_1)\right], \tag{3.13}$$

where $\alpha_{n-1} \in [0, 1]$ is a control factor that controls the contribution of the total variation to the next iteration: when the total variation is small, choose α_{n-1} close to 1, and when the total variation is large, a small α_{n-1} will be preferred. A necessary (but not sufficient) condition for the solution to converge is that $\alpha_{n-1} \to 1$ as $n \to \infty$.

3.3. Spatial discretization in two dimensions. To be specific, we will demonstrate the 2D discretization in space. The 1D case can be obtained easily from the 2D scheme, and for higher dimensions our scheme can be modified to work.

Suppose we have the 2D version of (3.9), and denote $w_1 = (x, y)'$ and $h(w_1) = (u(x, y), v(x, y))'$. Also, let

$$\begin{aligned} b(w_1) &= (b_1(x, y), b_2(x, y))', \\ f(w_1, w_1) &= (f_1(x, y), f_2(x, y))' \end{aligned} \tag{3.14}$$

and

$$B(w_1) = \begin{pmatrix} B_{11}(x, y) & B_{12}(x, y) \\ B_{21}(x, y) & B_{22}(x, y) \end{pmatrix}.$$

Then the iterations for u and v will be

$$\begin{aligned} u_n &= u_{n-1} + \tau_{n-1}\left[b_1 + (B_{11}u_{n-1} + B_{12}v_{n-1}) - \left(\frac{\partial u_{n-1}}{\partial x}f_1 + \frac{\partial u_{n-1}}{\partial y}f_2\right)\right], \\ v_n &= v_{n-1} + \tau_{n-1}\left[b_2 + (B_{21}u_{n-1} + B_{22}v_{n-1}) - \left(\frac{\partial v_{n-1}}{\partial x}f_1 + \frac{\partial v_{n-1}}{\partial y}f_2\right)\right], \end{aligned} \tag{3.15}$$

given initial guess u_0 and v_0. Suppose we want to solve the equation in a domain $[a, b] \times [c, d]$; we can form a grid on this domain with mesh size Δ_x and Δ_y, respectively:

$$\begin{aligned} x_i &= a + i\Delta_x, \quad i = 0, 1, 2, \ldots, N_x, \\ y_j &= c + j\Delta_y, \quad j = 0, 1, 2, \ldots, N_y, \end{aligned} \tag{3.16}$$

where $N_x = \frac{b-a}{\Delta x}$ and $N_y = \frac{d-c}{\Delta y}$. By using the central difference operators, we have

$$\begin{aligned} \frac{\partial u_{n-1}(i, j)}{\partial x} &\approx \frac{u_{n-1}(i+1, j) - u_{n-1}(i-1, j)}{2\Delta_x}, \\ \frac{\partial u_{n-1}(i, j)}{\partial y} &\approx \frac{u_{n-1}(i, j+1) - u_{n-1}(i, j-1)}{2\Delta_y}, \\ \frac{\partial v_{n-1}(i, j)}{\partial x} &\approx \frac{v_{n-1}(i+1, j) - v_{n-1}(i-1, j)}{2\Delta_x}, \\ \frac{\partial v_{n-1}(i, j)}{\partial y} &\approx \frac{v_{n-1}(i, j+1) - v_{n-1}(i, j-1)}{2\Delta_y}. \end{aligned} \tag{3.17}$$

Then the iteration scheme for 2D problem (3.15) at each grid point (x_i, y_j) will be

$$\begin{aligned} u_n = u_{n-1} + \tau_{n-1}\Bigg[&b_1 + (B_{11}u_{n-1} + B_{12}v_{n-1}) \\ &- \left(\frac{u_{n-1}(i+1, j) - u_{n-1}(i-1, j)}{2\Delta_x}f_1 + \frac{u_{n-1}(i, j+1) - u_{n-1}(i, j-1)}{2\Delta_y}f_2\right)\Bigg], \\ v_n = v_{n-1} + \tau_{n-1}\Bigg[&b_2 + (B_{21}u_{n-1} + B_{22}v_{n-1}) \\ &- \left(\frac{v_{n-1}(i+1, j) - v_{n-1}(i-1, j)}{2\Delta_x}f_1 + \frac{v_{n-1}(i, j+1) - v_{n-1}(i, j-1)}{2\Delta_y}f_2\right)\Bigg]. \end{aligned} \tag{3.18}$$

Here the functions without arguments are evaluated at grid points (x_i, y_j).

3.4. On boundary condition and computation. There is one issue for implementing the iteration (3.18): the values on the boundary of the domain are difficult to compute in general. One approach to avoid this issue is to use dynamically shrinking domains. Start by assigning values at all points in an initial domain that is much larger than the domain of interest. Then the iteration is defined properly at all interior points but not on the boundary, so we make the domain smaller by discarding all boundary points for the next iteration. Repeating this process, we produce a sequence of domains of decreasing size on which successively better approximation of the synchronization manifold is obtained. By taking the initial domain to be large enough so that the iteration converges before the domain becomes smaller than the original domain of interest, we obtain an approximate solution on the entire domain of interest.

Intuitively, this method works by choosing the boundary of the computational domain to dynamically change with the propagation wave front of the effect of the boundary of the initial domain. Since the solution at a given fixed point in the domain will not be affected by the choice of values on the boundary of the initial domain until this wave front reaches the point, any point that remains in our shrinking domain should not be affected by the choice of boundary condition on the initial domain. The influence of the unknown boundary condition of the outer domain becomes progressively less important as time progresses and the inner domain shrinks.

4. Examples of invariant manifolds. As we discussed in the previous sections, the manifold equation (2.4) is important because it allows further study of the synchronization between coupled oscillators. In this section we show several examples of constructing a generalized synchronization manifold. In particular, we consider examples in which the dynamics of the two oscillators are different but only with minor difference. A motivation for the study of this special case is that in any real dynamical systems no oscillator is physically perfect in the sense that the parameters or even the structure we use to mold these oscillators bear some small error.

4.1. A coupled 1D oscillator. One-dimensional examples provide an excellent starting point of investigation for studying invariant manifolds for higher-dimensional systems. Consider a simple case where

$$\dot{x} = 1, \quad x \in \Re. \tag{4.1}$$

We couple this oscillator with another one, but with some perturbation function on it so that the two oscillators are not identical:

$$\begin{aligned} \dot{x} &= 1, \\ \dot{y} &= 1 + (x - y) + \epsilon(p(x) + p'(x)), \end{aligned} \tag{4.2}$$

where p is any smooth function and p' denotes its derivative. Here the $(x-y)$ term represents the one-way coupling from x to y, and the $\epsilon(p(x)+p'(x))$ term represents some small perturbation on the y oscillator. Note that when $\epsilon = 0$, the synchronization manifold is $y = x$, which is stable. This oscillator is designed to have $\Phi(x) = p(x)$ as a stable invariant manifold, which

can be checked by introducing an appropriate change of coordinates. This type of oscillator always results in an intrinsic relationship between x and y subcomponents, given by $y = p(x)$.

For example, if we let $p(x) = \sin(x)$, then we have

$$\begin{aligned} \dot{x} &= 1, \\ \dot{y} &= 1 + (x - y) + \epsilon(\sin(x) + \cos(x)). \end{aligned} \tag{4.3}$$

Now for $\epsilon \neq 0$ the invariant manifold $\Phi(x)$ is determined by the manifold equation (2.4), which yields

$$\frac{d\Phi(x)}{dx} = -\Phi(x) + (1 + x) + \epsilon(\sin(x) + \cos(x)). \tag{4.4}$$

This equation is a simple ODE that can be solved analytically using the multiplication factor method. The solution is

$$\Phi(x) = x + \epsilon \sin(x) + Ce^{-x}, \tag{4.5}$$

where C is a constant that is determined by the boundary condition. In this family of invariant manifolds, the only one that corresponds to the perturbation of the identical synchronization ($\Phi(x) \to x$ when $\epsilon \to 0$) is obtained when $C = 0$:

$$\Phi(x) = x + \epsilon \sin(x). \tag{4.6}$$

The stability of the manifold can be determined by (2.7); in this case we obtain

$$\dot{\xi} = -\xi, \tag{4.7}$$

so for $t \to \infty$ the invariant manifold is exponentially stable.

The manifold can also be obtained by using the iteration scheme we described in the previous section: suppose the form of the manifold $\Phi(x) = x + \epsilon H(x)$. To find H we look for the corresponding stationary solution of the PDE described by (3.5). Following the iteration scheme (3.9), we start with $h_0 = 0$, choose $\tau_n = \tau = 0.1$, and obtain the following sequence of functions:

$$\begin{aligned} h_0 &= 0, \\ h_1 &= 0.1 \cdot \sin(x) + 0.1 \cdot \cos(x), \\ h_2 &= 0.2 \cdot \sin(x) + 0.18 \cdot \cos(x), \\ h_{10} &= 0.8340089600 \cdot \sin(x) + 0.3315041568 \cdot \cos(x), \\ &\vdots \\ h_{100} &= 0.9999965626 \cdot \sin(x) - 0.00004893563553 \cdot \cos(x). \end{aligned}$$

As we can see, this sequence of iteration functions converges to $\sin(x)$, which is the true form that appears in the manifold. We plot the first 20 iterates in Figure 1 to visualize the convergence of iterations.

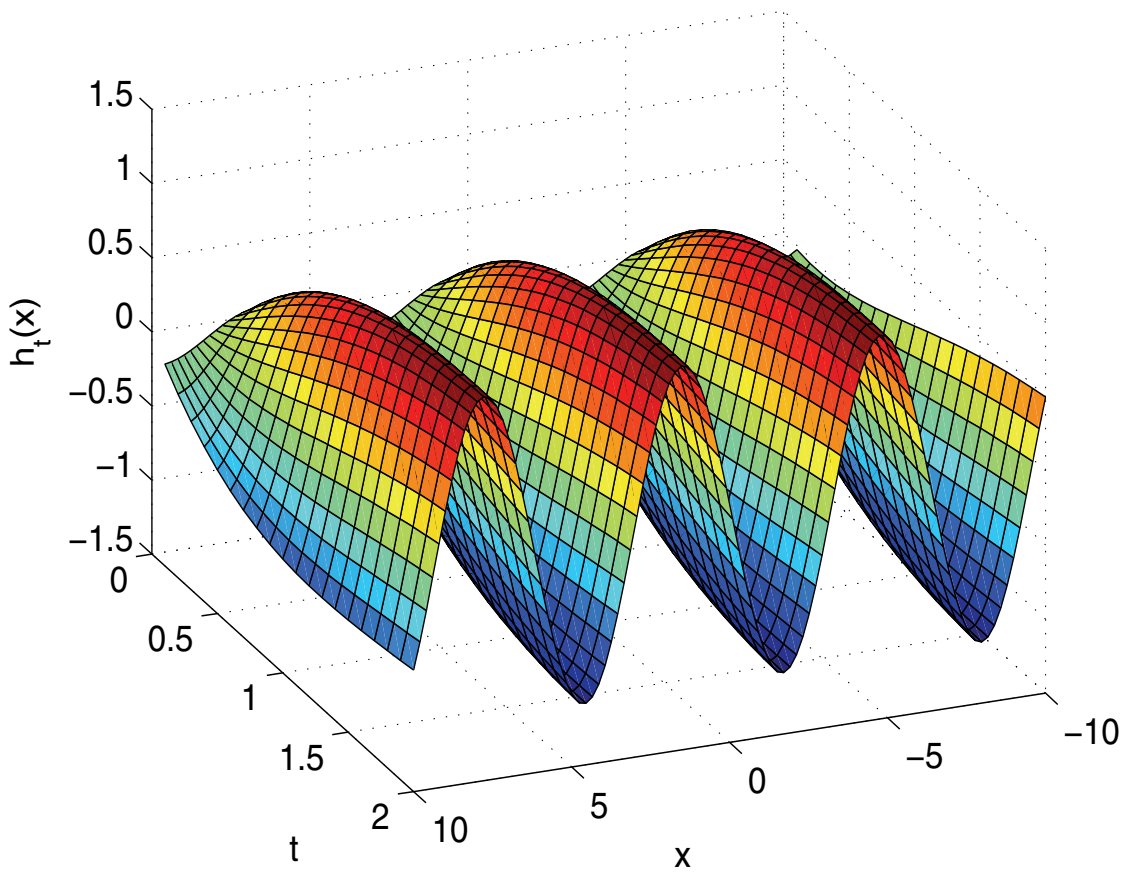


Figure 1. *Successive approximation of the manifold form $H(x)$ corresponding to the system (4.2). Here $h_t(x)$ corresponds to successive functions obtained from (3.9), with step size $\tau = 0.1$. The approximation tends to the true form $H(x) = \sin(x)$.*

4.2. Coupled 2D oscillators. We now consider two coupled oscillators and study the invariant manifold that describes the long-term relationship between them. More specifically, consider a unidirectionally coupled system made up of two different oscillators:

$$\begin{aligned} \dot{x_1} &= 1, \\ \dot{y_1} &= 1 + (x_1 - y_1) + \epsilon_1(p_1(x_1) + p_1'(x_1)), \\ \dot{x_2} &= 1 + k(x_1 - x_2), \\ \dot{y_2} &= 1 + (x_2 - y_2) + \epsilon_2(p_2(x_2) + p_2'(x_2)). \end{aligned} \tag{4.8}$$

Here the first oscillator is sending signals to the second through the x component, where k is the coupling strength. For large enough k we expect the corresponding x components to synchronize identically, while the form of synchronization of the y component can depend on the choice of ϵ and p. However, note that when x synchronizes, the coupling term becomes

small and we have two individual oscillators which we know how to solve. For $t \to \infty$ we expect to have the following relations to hold, at least in the approximate sense:

$$\begin{aligned} x_2 &= x_1, \\ y_1 &= x_1 + \epsilon_1 p_1(x_1), \\ y_2 &= x_2 + \epsilon_2 p_2(x_2). \end{aligned} \tag{4.9}$$

Let $\epsilon := \epsilon_2 - \epsilon_1$ measure the order of mismatch of the two oscillators, and look for the relationship $y_2 = y_1 + \epsilon H(x_1, y_1)$. Plugging (4.9) into this equation, we obtain

$$\epsilon H(x_1, y_1) = \epsilon_2 p_2(x_1) - \epsilon_1 p_1(x_1). \tag{4.10}$$

Thus, we expect the invariant manifold between these two oscillators to be given by

$$\begin{aligned} x_2 &= x_1, \\ y_2 &= y_1 + \epsilon_2 p_2(x_1) - \epsilon_1 p_1(x_1). \end{aligned} \tag{4.11}$$

Indeed, one can show that these equations do give an exponentially stable synchronization manifold by introducing an appropriate change of variables.

As an example, let us consider $p_1(x) = p_2(x) = \sin(x)$, but with $\epsilon_1 \neq \epsilon_2$. Setting $\epsilon = \epsilon_2 - \epsilon_1$, the invariant manifold is given by

$$\begin{aligned} x_2 &= x_1, \\ y_2 &= y_1 + \epsilon \sin(x_1). \end{aligned} \tag{4.12}$$

To check the iteration scheme in this case, let

$$\begin{aligned} x_2 &= x_1 + \epsilon u(x_1, y_1), \\ y_2 &= y_1 + \epsilon v(x_1, y_1) \end{aligned} \tag{4.13}$$

be the form of the invariant manifold, and apply the iteration scheme (3.9) with $h(x_1, y_1) = (u(x_1, y_1), v(x_1, y_1))'$.

Starting with the initial functions $u_0 = v_0 = 0$, we find that

$$\begin{aligned} u_0 = 0, \quad & v_0 = 0, \\ u_1 = 0, \quad & v_1 = 0.1 \cdot \sin(x) + 0.1 \cdot \cos(x), \\ u_2 = 0, \quad & v_2 = 0.2 \cdot \sin(x) + 0.18 \cdot \cos(x), \\ u_{10} = 0, \quad & v_{10} = 0.8340089600 \cdot \sin(x) + 0.3315041568 \cdot \cos(x), \\ & \vdots \\ u_{100} = 0, \quad & v_{100} = 0.9999965626 \cdot \sin(x) - 0.00004893563553 \cdot \cos(x), \end{aligned}$$

and u_n converges to 0 while v_n converges to $\sin(x_1)$, as expected.

4.3. Coupled nonidentical Van der Pol oscillators. The equation of the Van der Pol oscillator can be written as a first order system:

$$\begin{aligned} \dot{x} &= y, \\ \dot{y} &= -x + \mu(1 - x^2)y, \end{aligned} \tag{4.14}$$

where x represents position, y velocity, and μ the strength of the nonlinear damping. For $\mu \ll 1$, there exists a stable limit cycle of radius 2. Now consider two of these oscillators, with slightly different values of the parameter μ, and coupled through the x component from the first to the second. The equations describing this system are

$$\begin{aligned} \dot{x_1} &= y_1, \\ \dot{y_1} &= -x_1 + \mu(1 - x_1^2)y_1, \\ \dot{x_2} &= y_2 + k(x_1 - x_2), \\ \dot{y_2} &= -x_2 + (\mu + \epsilon)(1 - x_2^2)y_2. \end{aligned} \tag{4.15}$$

Here in the second oscillator the $k(x_1 - x_2)$ term represents the one-way coupling from the first oscillator, and $\epsilon \ll \mu$ is the mismatch in the damping strength. Since here we have two different individual systems coupled together, the identical manifold is not invariant under this dynamics. However, for large enough k, when $t \to \infty$ we numerically observe that the difference of x_1 and x_2 is $O(\epsilon^2)$, while the difference of y_1 and y_2 is $O(\epsilon)$. In Figure 2 we show an example of the deviation of trajectory from the identical manifold, with parameters chosen to be $\mu = 0.1$, $\epsilon = 0.01$, and $k = 20$. Generalized synchronization appears instead of complete synchronization in this case.

To demonstrate the effectiveness of our numerical scheme, use the 2D iteration scheme (3.15), start with initial guess close to the theoretical prediction $u(x, y) = 0$, $v(x, y) = x - \frac{1}{3}x^3$ (see the appendix for derivation), and compute the iterations starting on the domain $[-5, 5] \times [-5, 5]$ with mesh size 0.005 and time spacing $\tau = 0.0005$. We terminate when the domain shrinks down to $[-2.5, 2.5] \times [-2.5, 2.5]$ and take the solution at that time to be the approximated stationary solution. In Figure 3 we plot the solutions u and v, with a typical trajectory obtained from the original ODE system.

Figure 4 shows the successive error $e_n = \max(||u_n - u_{n-1}||_\infty, ||v_n - v_{n-1}||_\infty)$ measured by the infinite norm. We see exponential decrease in this error, which indicates that the iterative scheme is converging. In Figure 5 we plot the distance of a typical trajectory to the manifold obtained by this scheme, which suggests that the computed manifold indeed gives a good first order approximation for the synchronization manifold.

5. Design coupling to satisfy desired synchronization. In many real applications it is useful to have the knowledge of how to design a specific coupled system so that the resulting behavior satisfies certain purposes. Mathematically we are seeking a method to design the coupling between two general oscillators so that the resulting invariant manifold will have the prescribed form. In this section we discuss how to achieve the above goal by making use of the manifold equation we derived earlier.

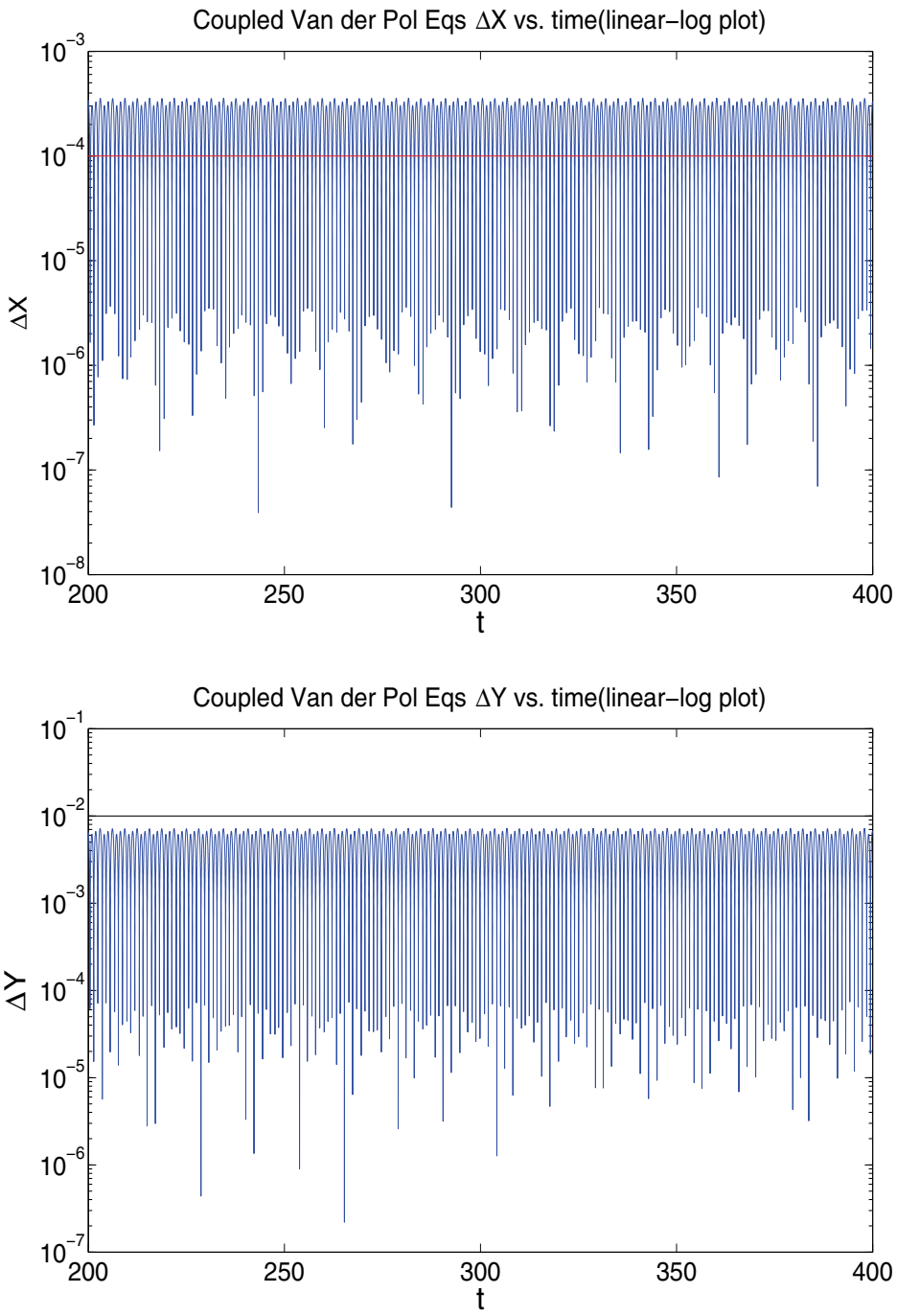


Figure 2. *Deviation from the identical manifold. The time series comes from the system* (4.15) *starting from* $(x_1(0), y_1(0), x_2(0), y_2(0)) = (1.5, 1.5, 1.5006, 1.5107)$. *Here in the two panels we plot the interval* $t \in [200, 400]$. *The curve in the upper panel shows the difference* $\Delta X(t) := |x_1(t) - x_2(t)|$, *and in the lower panel we plot* $\Delta Y(t) := |y_1(t) - y_2(t)|$. *The red line in the upper panel shows the value* $\epsilon^2 = 10^{-4}$, *and the black line in the lower panel indicates* $\epsilon = 10^{-2}$.

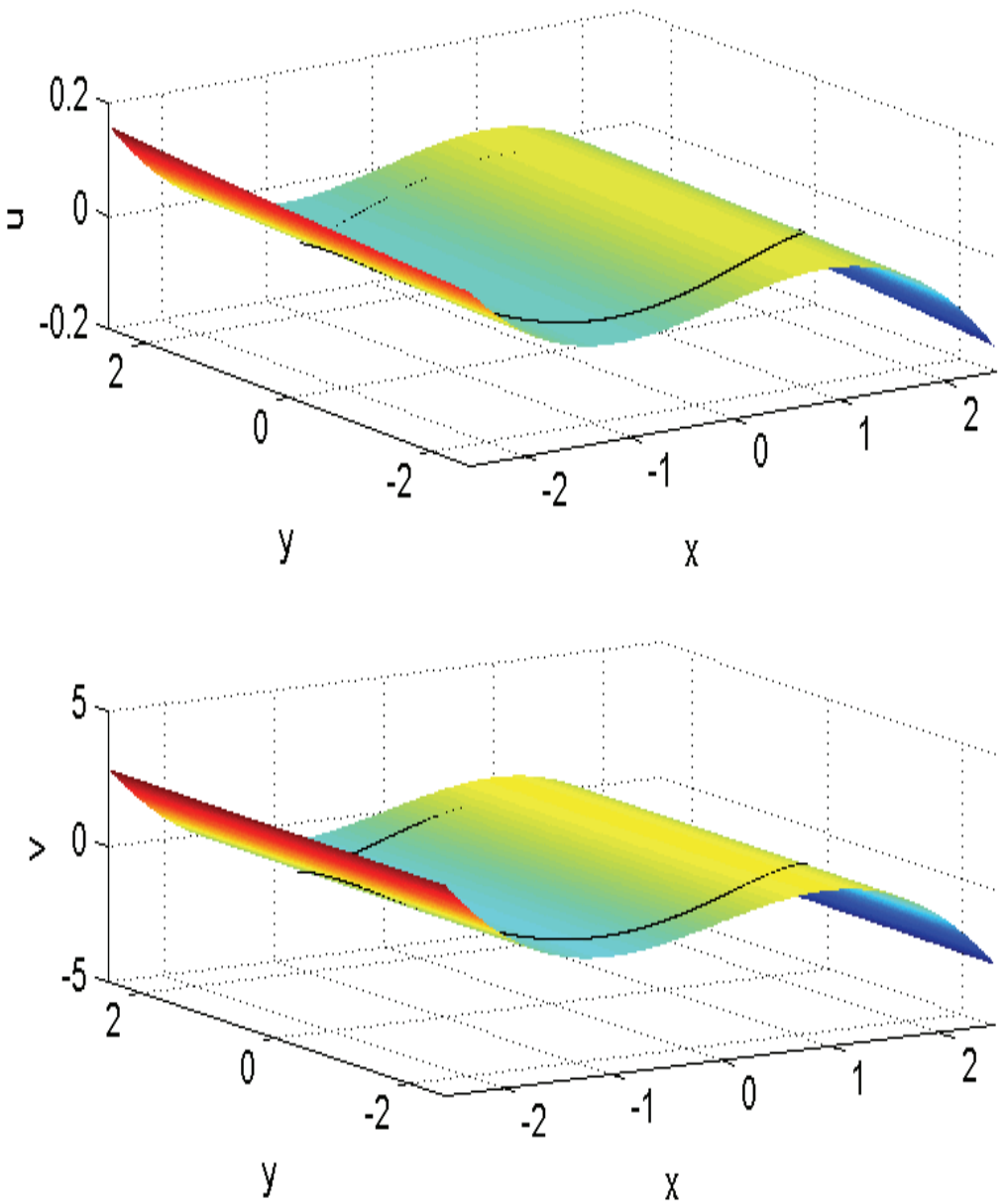


Figure 3. *Numerical solution of the time dependent PDE corresponding to the system* (4.15). *In the upper panel we plot* u, *and the black curve landing on* u *is obtained from plotting time series* $(x_1(t), y_1(t), \frac{x_2(t)-x_1(t)}{\epsilon})$ *obtained from the original ODE. Similarly, in the bottom panel we plot* v *as well as the time series* $(x_1(t), y_1(t), \frac{y_2(t)-y_1(t)}{\epsilon})$.

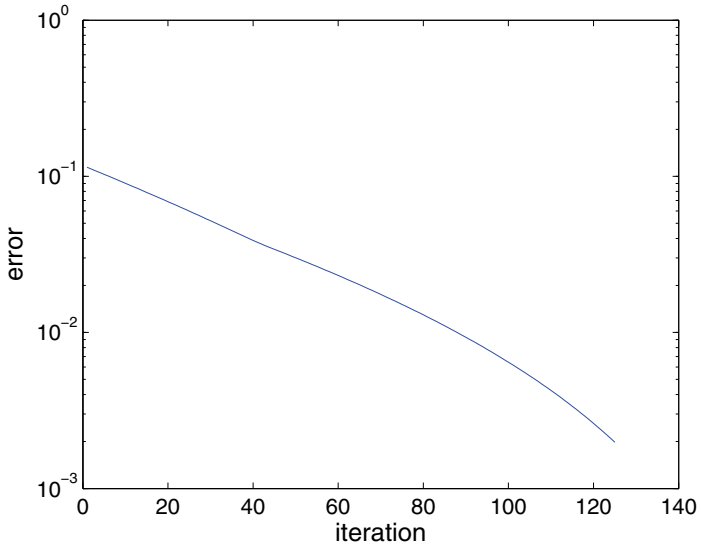


Figure 4. *Successive error e_n in the 2D iteration scheme for the time dependent PDE of the nonidentical coupled Van der Pol system. Here the error is measured by the infinite norm.*

Assume that without coupling we have two separate oscillators described by the following equations:

$$\begin{aligned} \dot{w}_1 &= F(w_1), \\ \dot{w}_2 &= G(w_2), \end{aligned} \tag{5.1}$$

where $w_1 \in \Re^m$, $w_2 \in \Re^m$, $F : \Re^m \to \Re^m$, and $G : \Re^m \to \Re^m$, where $F \in C^1$ and $G \in C^1$. Now suppose that we want w_1 and w_2 to synchronize, and furthermore, the synchronization manifold is prescribed as $w_2 = \Phi(w_1)$; the question is how to couple the two oscillators so that the resulting invariant manifold will have the desired form and be stable.

5.1. Form of the coupling function. We first solve for the coupling function so that the corresponding coupled system will have the invariant manifold satisfying the prescribed form. For convenience, we use a one-way coupling, so that the system becomes

$$\begin{aligned} \dot{w}_1 &= F(w_1), \\ \dot{w}_2 &= G(w_2) + K(w_1, w_2), \end{aligned} \tag{5.2}$$

where $K : \Re^m \times \Re^m \to \Re^m$ represents the coupling from x to y. In order not to make the question too broad, we assume that the coupling function has the form

$$K(w_1, w_2) = \sigma(L(w_1) - H(w_2)), \tag{5.3}$$

where $L : \Re^m \to \Re^m$ and $H : \Re^m \to \Re^m$ both $\in C^1$. Here also σ is a scalar which measures the strength of coupling, often called the *coupling coefficient* or *coupling strength*. Then in

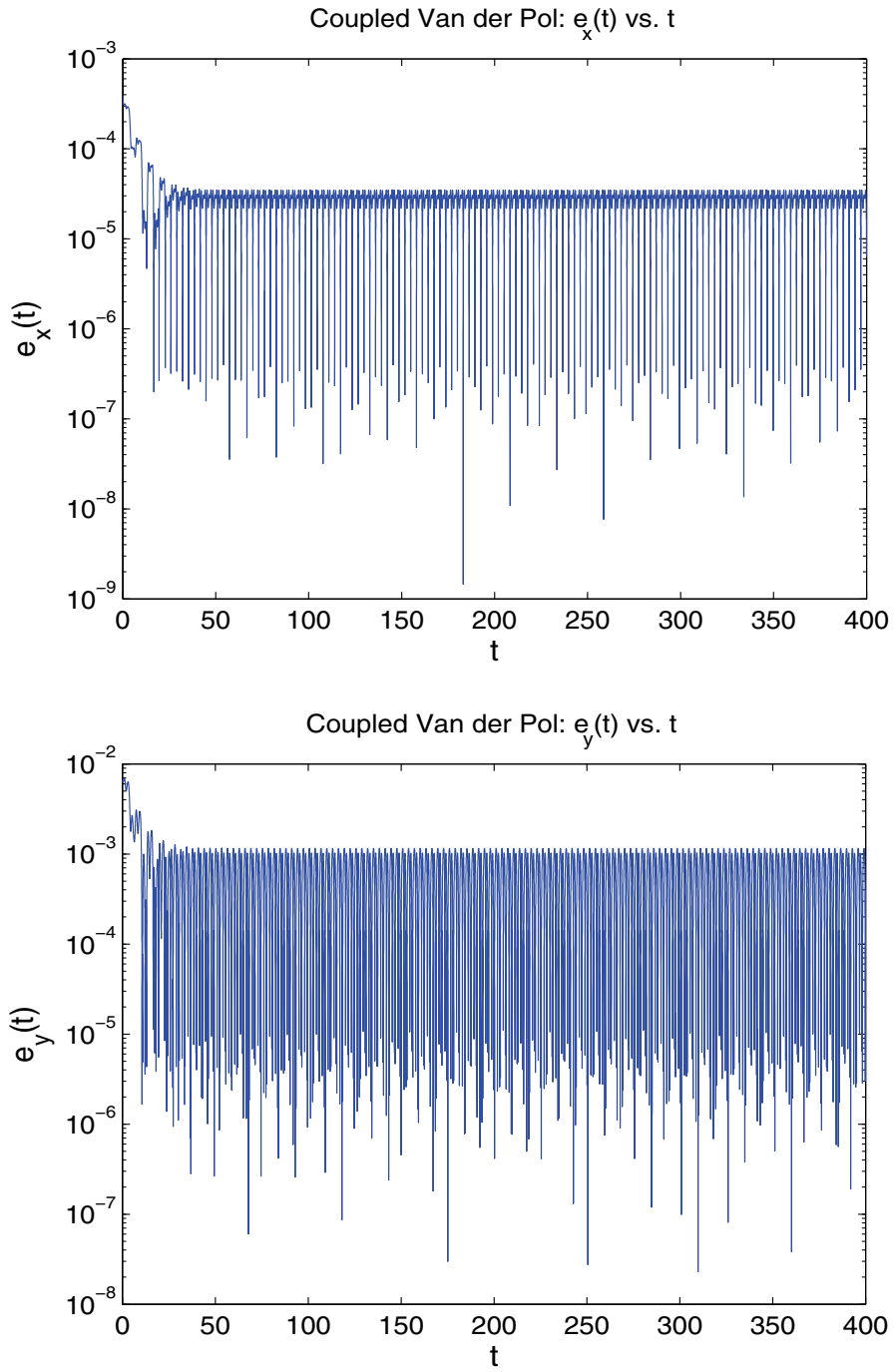


Figure 5. *Distance of a typical trajectory to the manifold obtained by numerical scheme. Here* $e_x(t) := |x_2(t) - x_1(t) - \epsilon u(x_1(t), y_1(t))|$ *and* $e_y(t) := |y_2(t) - y_1(t) - \epsilon v(x_1(t), y_1(t))|$*, where* u *and* v *are solutions by numerical scheme on grid points, and for points in between grid points we use spline interpolation to obtain its corresponding value. Here we see that* $e_x \sim o(\epsilon)$ *and* $e_y \sim o(\epsilon)$*, suggesting that the first order approximation in* ϵ *(*$\epsilon = 0.01$*) is achieved.*

order that the synchronization manifold of the corresponding coupled system

$$\begin{aligned} \dot{w_1} &= F(w_1), \\ \dot{w_2} &= G(w_2) + \sigma(L(w_1) - H(w_2)) \end{aligned} \tag{5.4}$$

has the desired form $w_2 = \Phi(w_1)$, by (2.4), we have

$$D\Phi(w_1)F(w_1) = G(\Phi(w_1)) + \sigma(L(w_1) - H(\Phi(w_1))). \tag{5.5}$$

Note here that we have two functions L and H to choose, while the manifold gives only one constraint, and we need to fix one of the coupling forms. Suppose we are given the form of $H(w_2)$ and want to find $L(w_1)$. Solving for $L(w_1)$ yields

$$L(w_1) = \frac{1}{\sigma}[D\Phi(w_1) \cdot F(w_1) - G(\Phi(w_1))] + H(\Phi(w_1)). \tag{5.6}$$

So for any given form of the synchronization manifold $w_2 = \Phi(w_1)$ we are able to design the corresponding coupling between the two oscillators.

Notice that for identical oscillators ($F = G$), if we want $\Phi(w_1) = w_1$ (identical synchronization), then

$$D\Phi(w_1) = I_m \tag{5.7}$$

is the $m \times m$ identity matrix, so the first term in (5.6) disappears and the second term is just $H(\Phi(w_1)) = H(w_1)$. Thus we have

$$L(w_1) = H(w_1), \tag{5.8}$$

meaning that for the coupled system (5.4) to have the identical synchronization manifold, the coupling functions L and H should have the same form, as expected.

5.2. Adjusting coupling coefficients to obtain stability. In (5.6) we have the form of the coupling function which guarantees that the coupled system described by (5.4) has the invariant manifold $w_2 = \Phi(w_1)$. The corresponding variational equation along the manifold can be obtained using (2.7), which becomes

$$\dot{\xi} = D[G(w_2) - \sigma H(w_2)]|_{w_2=\Phi(w_1)}\xi. \tag{5.9}$$

Stability can be obtained by altering the value of σ, the coupling strength, if it is possible to have stable synchronization. Furthermore, with this variational equation as well as the form of the invariant manifold (known in advance), we are able to analyze the stability of not only the whole manifold but also individual points or subsets on the invariant manifold.

5.3. Examples of application. One case of the design of coupling form appears in linear systems with perturbations. Suppose we have two uncoupled linear systems:

$$\begin{aligned} \dot{w_1} &= Aw_1, \\ \dot{w_2} &= Bw_2, \end{aligned} \tag{5.10}$$

where A and B are linear operators (matrices). It is known that if $A = B$, then by linear diffusive coupling the two oscillators will synchronize and the synchronization is stable for large enough coupling strength. Now in the more general situation A and B can be different, either because the true systems are not made perfectly exact, or for other reasons. But we still want to couple them so that they achieve identical synchronization. The analysis in the previous sections offers us a way to design the coupling form.

Again use one way of coupling and assume the form of a coupling function on w_2. The coupled system becomes

$$\begin{aligned} \dot{w_1} &= Aw_1, \\ \dot{w_2} &= Bw_2 + \sigma(L(w_1) - Hw_2), \end{aligned} \tag{5.11}$$

where H is a coupling matrix (known in advance) and L is the unknown coupling function of w_1 that needs to be determined. By (5.4) and (5.6), we find that

$$L(w_1) = \left(\frac{A-B}{\sigma} + H\right) w_1. \tag{5.12}$$

Here we see that only when $A = B$ can we allow $L(w_1) = H(w_1)$ and obtain identical synchronization; otherwise the coupling form of x and y shall be different to compensate for the difference of the original dynamics on x and y, respectively. The stability of the manifold is described by

$$\dot{\xi} = (B - \sigma H)\xi, \tag{5.13}$$

following (5.9). In order for the synchronization manifold to be stable, we need

$$\rho(B - \sigma H) < 1, \tag{5.14}$$

where $\rho(Q)$ is the spectral radius of the matrix Q.

Equations (5.12), (5.13), and (5.14) give us a way to easily manipulate the coupling functions in the nonidentical linear systems to obtain stable identical synchronization. With minor modifications we can obtain similar results for two-way coupling, and with a few more modifications as well as more analysis we shall be able to control more complicated systems.

6. Discussion and conclusion. In this paper we have developed systematic methods for explicit construction of generalized synchronization manifolds in systems of coupled nonidentical oscillators by means of the manifold equation, a PDE that must be satisfied by the manifold, and the associated variational equation describing its stability.

Although the manifold equation gives necessary conditions to determine the invariant manifold, it is not sufficient, mainly due to the fact that boundary conditions for this PDE are not easily attainable. To obtain a solution of the manifold equation without explicitly knowing the boundary conditions, we have proposed a time dependent PDE, whose stably stationary solution is a solution of the manifold equation, and developed iteration schemes to solve it both symbolically and numerically. Several examples of constructing synchronization manifolds have been given for systems of nearly identical oscillators (considered as a perturbed version

of identical oscillators, which correspond to complete synchronization) with unidirectional coupling, where we obtained the first order approximation of the perturbed manifold. A general technique for using the manifold equation to design the coupling for the purpose of controlling the form and stability of synchronization has also been discussed and illustrated using the simple case of linear systems.

Our numerical algorithm used to obtain the stationary solution of the time dependent PDE is analogous to the Euler method in a functional space, whose convergence is not necessarily global. Thus, finding a systematic method for choosing an appropriate initial condition, as well as a detailed convergence analysis for the scheme, is an important problem that must be addressed in the future. As there exist a variety of different schemes to solve PDEs in the literature, exploring and comparing with other numerical schemes for solving the manifold equation is another important topic of future research.

Appendix. In the coupled Van der Pol system (4.15), in order to obtain the first order approximation of the synchronization manifold in ϵ, we neglect higher order terms in ϵ and write

$$\begin{aligned} x_2 &= x_1 + O(\epsilon^2) \approx x_1, \\ y_2 &= y_1 + \epsilon V(x_1, y_1) + O(\epsilon^2) \approx y_1 + \epsilon V(x_1, y_1). \end{aligned} \tag{A.1}$$

Thus, using (4.15), we have

$$\dot{y_2} - \dot{y_1} = \epsilon(1 - x_1^2)y_1 + O(\epsilon^2). \tag{A.2}$$

We also have, for $\mu \ll 1$,

$$\begin{aligned} \dot{y_2} - \dot{y_1} &= \epsilon \left(\frac{\partial V}{\partial x_1}\dot{x_1} + \frac{\partial V}{\partial y_1}\dot{y_1} \right) + O(\epsilon^2) \\ &= \epsilon \left\{ y_1 \frac{\partial V}{\partial x_1} + [-x_1 + \mu(1 - x_1^2)y_1] \frac{\partial V}{\partial y_1} \right\} + O(\epsilon^2) \\ &= \epsilon \left(y_1 \frac{\partial V}{\partial x_1} - x_1 \frac{\partial V}{\partial y_1} \right) + O(\epsilon^2). \end{aligned} \tag{A.3}$$

Replacing x_1 and y_1 with x and y for convenience and equating the first order ϵ terms in the above two equations, we obtain the following equation:

$$y \frac{\partial V}{\partial x} - x \frac{\partial V}{\partial y} = (1 - x^2)y. \tag{A.4}$$

This equation has a particular solution:

$$V(x, y) = x - \frac{1}{3}x^3. \tag{A.5}$$

Since the difference between x components in this case is already $O(\epsilon^2)$, we take the identical manifold $x_2 = x_1$ as the first order approximation, while for the y component we use the form obtained above: $y_2 = y_1 + \epsilon V(x_1, y_1)$.

Acknowledgment. We would like to thank Naratip Santitissadeekorn for discussion of the numerical scheme part.

REFERENCES

[1] J. Carr, *Applications of Centre Manifold Theory*, Springer, New York, 1981.
[2] L. M. Pecora and T. L. Carroll, *Synchronization in chaotic systems*, Phys. Rev. Lett., 64 (1990), pp. 821–824.
[3] N. F. Rulkov, M. M. Sushchik, and L. S. Tsimring, *Generalized synchronization of chaos in directionally coupled chaotic systems*, Phys. Rev. E, 51 (1995), pp. 980–994.
[4] H. D. I. Abarbanel, N. F. Rulkov, and M. M. Sushchik, *Generalized synchronization of chaos: The auxiliary system approach*, Phys. Rev. E, 53 (1996), pp. 4528–4535.
[5] M. G. Rosenblum, A. S. Pikovsky, and J. Kurths, *Phase synchronization of chaotic oscillators*, Phys. Rev. Lett., 76 (1996), pp. 1804–1807.
[6] L. Kocarev and U. Parlitz, *Generalized synchronization, predictability, and equivalence of unidirectionally coupled dynamical systems*, Phys. Rev. Lett., 76 (1996), pp. 1816–1819.
[7] K. T. Alligood, T. D. Sauer, and J. A. Yorke, *Chaos*, Springer, New York, 1997.
[8] M. G. Rosenblum, A. S. Pikovsky, and J. Kurths, *From phase to lag synchronization in coupled chaotic oscillators*, Phys. Rev. Lett., 78 (1997), pp. 4193–4196.
[9] R. Brown and N. F. Rulkov, *Synchronization of chaotic systems: Transverse stability of trajectories in invariant manifolds*, Chaos, 7 (1997), pp. 395–413.
[10] M. Ding, E.-J. Ding, W. L. Ditto, B. Gluckman, V. In, J.-H. Peng, M. L. Spano, and W. Yang, *Control and synchronization of chaos in high dimensional systems: Review of some recent results*, Chaos, 7 (1997), pp. 644–652.
[11] L. M. Pecora and T. L. Carroll, *Master stability functions for synchronized coupled systems*, Phys. Rev. Lett., 80 (1998), pp. 2109–2112.
[12] K. Josić, *Invariant manifolds and synchronization of coupled dynamical systems*, Phys. Rev. Lett., 80 (1998), pp. 3053–3056.
[13] R. Brown, *Approximating the mapping between systems exhibiting generalized synchronization*, Phys. Rev. Lett., 81 (1998), pp. 4835–4838.
[14] K. Josić, *Synchronization of chaotic systems and invariant manifolds*, Nonlinearity, 13 (2000), pp. 1321–1336.
[15] C. Masoller, *Anticipation in the synchronization of chaotic semiconductor lasers with optical feedback*, Phys. Rev. Lett., 86 (2001), pp. 2782–2785.
[16] J. Chubb, E. Barreto, P. So, and B. J. Gluckman, *The breakdown of synchronization in systems of nonidentical chaotic oscillators: Theory and experiment*, Internat. J. Bif. Chaos Appl. Sci. Engrg., 11 (2001), pp. 2705–2713.
[17] E. Ott, *Chaos in Dynamical Systems*, 2nd ed., Cambridge University Press, Cambridge, UK, 2002.
[18] A. Uchida, R. McAllister, R. Meucci, and R. Roy, *Generalized synchronization of chaos in identical systems with hidden degrees of freedom*, Phys. Rev. Lett., 91 (2003), paper 174101.
[19] A. Arenas, A. Díaz-Guilera, and C. J. Pérez-Vicente, *Synchronization reveals topological scales in complex networks*, Phys. Rev. Lett., 96 (2006), paper 114102.
[20] F. Sorrentino and E. Ott, *Network synchronization of groups*, Phys. Rev. E, 76 (2007), paper 056114.

SIAM J. Applied Dynamical Systems
Vol. 8, No. 1, pp. 222–252
© 2009 Society for Industrial and Applied Mathematics

Bifurcation Analysis of a Spatially Extended Laser with Optical Feedback*

Kirk Green†, Bernd Krauskopf‡, Frank Marten§, and Daan Lenstra¶

Abstract. Vertical cavity surface-emitting lasers (VCSELs) are a new type of semiconductor laser characterized by the spatial extent of their disk-shaped output apertures. As a result, a VCSEL supports several optical modes (patterns of light) transverse to the direction of light propagation. When any laser is coupled to other optical elements, there is unavoidable optical feedback via reflecting surfaces, which influences the stability of the laser output. For a VCSEL, the question is how the transverse optical modes interact dynamically in the presence of optical feedback and how this affects stability of the system. In this paper, we start from a PDE description of the VCSEL. We proceed by using an expansion in suitable eigenfunctions to resolve the spatial dependence. In the presence of optical feedback we obtain a model in the form of a system of delay differential equations (DDEs). As we show with the example of a VCSEL that supports two transverse modes, the spatially expanded DDE model is small enough to allow for a multiparameter bifurcation analysis with numerical continuation tools. Specifically, we present stability regions of steady states and periodic solutions in dependence on the feedback strength and a homotopy parameter that models the amount of self- versus cross-feedback between the two modes. Bifurcations of more complicated spatiotemporal mode dynamics are then discussed.

Key words. PDE with delay, eigenfunction expansion, vertical cavity surface-emitting semiconductor laser (VCSEL), numerical continuation

AMS subject classifications. 37N20, 34K18, 37G10, 37G20

DOI. 10.1137/080712982

1. Introduction.

Modern semiconductor lasers combine a small size, with dimensions of a few micrometers, with a high efficiency of turning electrical energy into coherent light. Their most common applications are as light sources in optical fiber networks and as read/write systems in optical storage systems. Semiconductor lasers are also found in laser printers, sensors, scanners, and even medical applications such as laser eye correction and dentistry. The most common semiconductor lasers today are edge-emitting lasers (EELs), where light is produced in an active region in the shape of a one-dimensional waveguide. The light travels in the plane of the chip on which the EEL is fabricated and exits at one or both end facets, which act as semitransparent mirrors. Due to their elongated geometry, EELs generally exhibit

*Received by the editors January 10, 2008; accepted for publication (in revised form) by T. Kaper October 21, 2008; published electronically February 13, 2009.
http://www.siam.org/journals/siads/8-1/71298.html
†Theoretical Physics, Vrije Universiteit, De Boelelaan 1081, 1081HV Amsterdam, The Netherlands (k.green@few.vu.nl). The research of this author was supported under the Dynamics of Patterns Program of the Netherlands Organization for Scientific Research (NWO) and the Foundation for Fundamental Research on Matter (FOM).
‡Department of Engineering Mathematics, University of Bristol, Bristol BS8 1TR, UK (b.krauskopf@bristol.ac.uk).
§Theoretical Physics, Vrije Universiteit, De Boelelaan 1081, 1081HV Amsterdam, The Netherlands, and Department of Engineering Mathematics, University of Bristol, Bristol BS8 1TR, UK (f.marten@bristol.ac.uk).
¶Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands (d.lenstra@tudelft.nl).

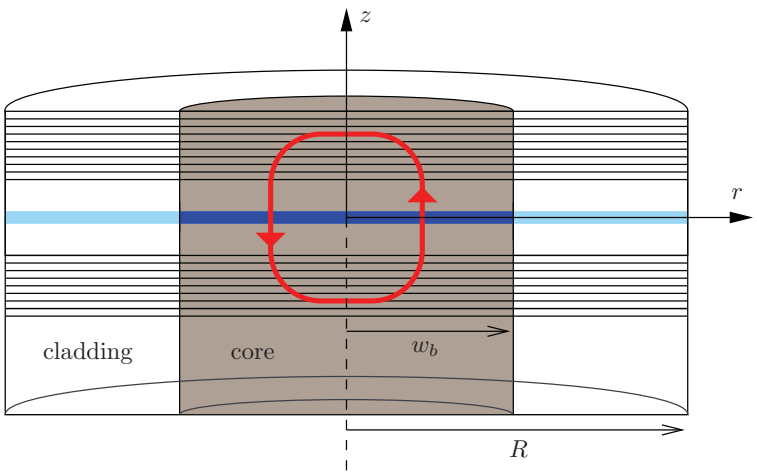


Figure 1. *Schematic of a VCSEL showing the dielectric cylindrical waveguide approximation of an active region (blue) in the core that is surrounded by a layer of cladding. The light is reflected in between two mirror stacks and exits at the top face of the core.*

single transverse mode lasing, which means that a single spot of light exits the laser under any operating conditions. However, an EEL may support several longitudinal modes in the direction of lasing. To ensure single longitudinal mode operation, internal frequency selective elements are included in the design of some EELs.

We are concerned here with vertical cavity surface-emitting lasers (VCSELs)—a second and increasingly important type of semiconductor laser. VCSELs have a cylindrical geometry, where a very thin, spatially extended active region is located in between two stacks of mirrors; see Figure 1. Electrical pumping via circular contacts ensures lasing in a core region. Light exits the VCSEL at a circular aperture in the top face. VCSELs are even more efficient than EELs and produce light already above a threshold current of typically only around 1 mA. They are easily coupled to optical fibers and, in contrast to EELs, can be arranged on-chip into larger arrays or optical circuits. Because their active region is very thin in the direction of lasing, VCSELs provide consistent single longitudinal mode operation. From an applications point of view, the only downside is that VCSELs may lase at several transverse optical modes, which are different "patterns of light" forming on the spatially extended disk-shaped output surface [6]. The number of optical transverse modes that a VCSEL supports increases with the diameter of the aperture. Increasing the diameter of the aperture is the main option for increasing the overall output power level of the VCSEL. We note that an alternative approach would be to increase the number of quantum wells in the active layer [61].

It is now well established that semiconductor lasers are very susceptible to external optical influences; see, for example, [27, 30] as introductions to the extensive literature. In particular, even very small amounts of optical feedback in the form of back-reflected light, on the order of 0.1%, may destabilize the laser [15, 37]. This is why expensive optical isolators must be employed when the laser is coupled to other optical elements in practical applications. A semiconductor laser can be described in a rate equation approach by differential equations for the evolution of the electric field and the density of charge carriers. In the presence of

optical feedback the electric field is coupled back to itself after a single fixed delay τ, which is simply the travel time from the laser to the mirror and back. For the case of a longitudinal single-mode EEL, one employs the famous Lang–Kobayashi (LK) equations [33], which have been shown in numerous studies to describe experimental measurements very well (for weak to moderate feedback); see, for example, [15, 24, 25]. From the mathematical point of view, the LK equations, and indeed other mathematical models of EEL systems with delayed optical feedback, take the form of delay differential equations (DDEs) with a fixed delay of moderate size (from three up to, say, six equations). The analysis of a DDE is complicated by the fact that it has an infinite-dimensional phase space [8, 22]. Numerical continuation tools for DDEs that allow one to find and follow equilibria, periodic solutions, and their bifurcations have become available only quite recently [11, 51, 47]. The desire to understand the dynamics of laser systems with delayed feedback has been one of the driving forces behind the development of these numerical tools [13, 17, 18, 21, 29, 49].

This paper is concerned with the bifurcation analysis of a VCSEL with optical feedback. Compared to the case of EEL systems discussed above, the additional element here is the spatial extent of the VCSEL beam: its two-dimensional lasing aperture gives rise to different transverse optical modes that interact with each other in the presence of optical feedback. Due to their transverse spatial nature, VCSELs need to be described by PDEs for the electric field and the carrier density. In the presence of optical feedback, one is therefore dealing with a mathematical model in the form of delayed partial differential equations (DPDEs). Note that DPDE models are presently also of interest in other application areas, including the control of spatial patterns [45, 52] and dynamics testing methods for mechanical systems [32]. To analyze or even simulate a DPDE model, one generally resolves the spatial part of the system with a suitable expansion method (for example, by spatial discretization or Galerkin projection). After truncation this leads to a system of DDEs. The problem is that numerical continuation tools are not able to cope with the resulting system of DDEs if it is too large. In particular, calculating stability information of equilibria and periodic orbits is extremely time consuming and restricted by the underlying software platform.

The key is therefore to exploit physical properties of the VCSEL in order to obtain a sufficiently small system of DDEs that still describes the spatiotemporal dynamics of the transverse modes under the influence of optical feedback. As opposed to the case of EELs, where there are now well-established rate equation models [30, 27], the modeling of VCSELs is still in development and under some dispute. Two important difficulties are that first, there are many different physical realizations of the VCSEL structure, and, second, it is experimentally challenging to make accurate spatially and time resolved measurements. We follow here the now quite standard modeling approach of expanding the spatial electric field in a hierarchy of transverse, linearly polarized (LP) optical modes inside the cylindrical core of the VCSEL [7, 19, 26, 34, 35, 36, 53, 54, 55, 57, 58, 59]; one also speaks of an index guided VCSEL [39]. Note that the VCSEL geometry also allows the electric field to be expanded in Gauss–Laguerre [3] or Gauss–Hermite [46] modes. An alternative modeling approach is that of fully determining the optical (electric) field profile by the spatial carrier distribution, and vice versa; one speaks of gain-guided VCSELs [40, 39]. The number of LP modes necessary for an accurate description of the VCSEL increases with the size of the aperture. Expansion in LP modes leads to a system of ordinary differential equations, one for the electric field of

each mode. In the presence of an external mirror, each transverse mode is subject to delayed optical feedback from itself and from other transverse modes. The overall DPDE then takes the form of a system of DDEs for the electric fields that are coupled to a PDE for the carrier distribution in the disk-shaped active region. Carriers, provided by electrical pumping, diffuse into the active region. In the references given above, this type of model has been simulated numerically by discretizing the PDE for the spatial charge carriers with a finite difference scheme.

We show here that a much smaller, spatially resolved DDE model can be obtained by expanding the carrier density in the active region of the VCSEL in orthogonal eigenfunctions that satisfy the diffusion equation for the carrier distribution. A similar approach was first taken by Moriki et al. [41], but the use of a finite difference scheme has proved to be the more popular method [59]. A recent return to eigenfunction expansion methods can be found as part of the simulation package `VISTAS` by Jungo, Erni, and Bächtold [26]; however, their choice of expansion functions introduces a singularity at the origin (the center) of the VCSEL. By contrast, our reduced equations are valid throughout. They consist of one complex-valued electric field equation for each optical mode (pattern of light) that is taken into consideration, which are coupled via series expanded carrier (electron-hole pair) equations inside the laser. Our method can be applied for any number of spatial optical modes with or without an azimuthal component. In practice, the number and type of optical modes one considers depends on the precise question one wants to answer. The fewer and simpler the electrical modes are, the lower the dimension of the reduction. Generally, optical modes with an azimuthal component lead to a description of the carriers by several expansion series. The special case of azimuthally independent modes, on the other hand, results in a small reduced model because it suffices to expand the carrier equation by a single series. In either case, the number of terms in each of the expansions of these carrier equations depends only on the value of the diffusion constant that governs how fast carrier sinks are being filled by the pump source. Hence, the number of terms needed in the carrier equation expansion can be determined a priori and does not need to be adjusted during a bifurcation study as performed here. A further advantage of eigenfunction expansions is that the overlap integrals describing the coupling between the optical modes appear as constants that need to be computed only once. Overall, our reduction method results in a substantial increase in computational efficiency, so that a multiparameter bifurcation analysis now becomes feasible.

As a specific example we consider here the case of a VCSEL in which the first two rotationally symmetric LP modes are excited. There are several reasons for this choice. First, this two-mode VCSEL is an ideal test case for our reduction method. It can be described by the least number of governing rate equations (due to the absence of azimuthal terms) while featuring a multimode character. Nevertheless, this case is already an extension of the single-mode model, so that one would expect additional dynamics from the interplay between the two modes (such as the emergence of the mixed oscillations identified later in the paper). But there is also a physical motivation behind studying this two-mode VCSEL. The interaction of the two basic rotationally symmetric modes has been of considerable interest because they are easily coupled into a fiber; see, for example, the experimental studies [43, 44]. We also remark that it is feasible in practice to filter out the frequency of the intermediate LP_{11} mode. In our model the optical feedback terms are modeled in a general form as introduced in [19], where

a homotopy parameter allows us to investigate the influence of the relative amount of self- versus cross-feedback of the modes. Note that all previous studies, for example, [34, 53, 59], assumed pure self-feedback.

As we demonstrate, the resulting spatially expanded DDE system is small enough to be amenable to a full numerical bifurcation analysis. We use the continuation package `DDE-BIFTOOL` [11] to study the bifurcations of steady-state solutions of the VCSEL model; they are called external cavity modes (ECMs) and correspond to solutions where the two fields lase with constant intensities. The accuracy of the spatially expanded model is demonstrated by a comparison with the ECM structure given in [19], obtained with a finite difference scheme discretization of the carrier PDE. A further bifurcation analysis for the difference scheme discretized model proved entirely unpractical. On the other hand, the spatially expanded DDE model of this study provides such a considerable speed-up in the stability calculations that it is now possible to compute stability regions of ECMs in the plane of feedback strength κ and self- versus cross-feedback parameter η. What is more, we are able to continue periodic solutions that are born in Hopf bifurcations of ECMs with the numerical package `PDDE-CONT` [51]. This reveals large stability regions of periodic solutions in the (κ, η)-plane. We find two types of oscillations: in-phase oscillations of both modes at the characteristic relaxation oscillation frequency of the VCSEL, and antiphase oscillations where the two modes suppress each other with a period given by a multiple of the external round trip time. We also consider how these periodic solutions bifurcate to quasi-periodic and even chaotic spatiotemporal dynamics of the two modes. In other words, the bifurcation analysis presented here reveals the influence of the self- versus cross-coupling parameter η on the dynamics of the spatial modes.

This paper is organized as follows. In section 2 we introduce the general spatially extended PDE description of the VCSEL and show how it can be reduced by a suitable mode expansion. The specific mode-expanded DDEs for the two-mode VCSEL with optical feedback are given in section 2.4. The stability regions of the ECMs are presented in section 3 and those of the bifurcating periodic orbits in section 4. Bifurcations to more complicated spatiotemporal dynamics are discussed briefly in section 5. Finally, in section 6 we draw conclusions and point to future work.

2. Rate equation model of a VCSEL. A VCSEL with multiple transverse modes can be described in the semiclassical framework by so-called rate equations for the slowly varying complex amplitudes $E^j_{mn}(t)$ of the contributing transverse modes and for the spatially extended inversion $N(r, \phi, t)$ [59, 61]. Here the profile Ψ^j_{mn} of the mnth optical mode of polarization j is a fixed solution of the Helmholtz equation (which in turn is derived from Maxwell's equations) of the cold (not electrically pumped) laser cavity [61]. The inversion is the spatial distribution of the number of electron-hole pairs in the circular active region, and, hence, it is expressed in polar coordinates for convenience. The VCSEL rate equations are given as [19, 59]

$$\frac{dE^j_{mn}(t)}{dt} = (1 + i\alpha)\xi^j_{mn}(t)E^j_{mn}(t) + F^j_{mn}(t), \tag{2.1}$$

$$T\frac{\partial N(r,\phi,t)}{\partial t} = \frac{d_f}{r}\left[\frac{\partial}{\partial r}\left(r\frac{\partial N(r,\phi,t)}{\partial r}\right)\right] + \frac{d_f}{r^2}\frac{\partial^2 N(r,\phi,t)}{\partial \phi^2} + P(r,\phi,t) - N(r,\phi,t) \tag{2.2}$$

$$- \sum_n \Bigg((1 + 2\xi_{0n}(t))|E_{0n}(t)|^2 |\Psi_{0n}(r,\phi)|^2 + \sum_m \sum_{j=c,s} 2(1 + 2\xi^j_{mn}(t))|E^j_{mn}(t)|^2 |\Psi^j_{mn}(r,\phi)|^2 \Bigg).$$

(This dimensionless form can be obtained from the formulation in [59] by rescaling; see Appendix A for the details.) Material properties of the VCSEL include the line-width enhancement factor α, the ratio between carrier and photon lifetimes T, a diffusion coefficient d_f, and a spatial pump profile P; see below. For reference, time is rescaled with respect to a photon lifetime of $\tau_p = 2$ ps. The function $F^j_{mn}(t)$ represents an external influence on the respective transverse mode that depends on its own set of parameters, such as the feedback strength. Specifically, we consider here optical feedback, as introduced in section 2.4 below. Note, however, that $F^j_{mn}(t)$ could just as easily represent optical injection [35]. The equations for the different transverse optical modes are coupled via the spatial overlaps between the optical modes and the inversion, which are expressed by the modal gain integrals

$$\xi^j_{mn}(t) = \frac{\int_0^1 \int_0^{2\pi} |\Psi^j_{mn}(r,\phi)|^2 N(r,\phi,t) r dr d\phi}{\int_0^1 \int_0^{2\pi} |\Psi^j_{mn}(r,\phi)|^2 r dr d\phi}. \tag{2.3}$$

Physically, these integrals represent the competition of each spatial mode for the reservoir $N(r,\phi,t)$ of carriers. Note that (2.1) does not contain terms involving the frequencies of the individual optical modes. This is because these frequencies are almost the same; they differ only on the scale of hundreds of gigahertz to terahertz [10]. Hence, the resulting beating oscillations fluctuate on a picosecond timescale, which is two to three orders of magnitude faster than the carrier lifetime of a nanosecond [16]. In other words, the carriers cannot respond on such a fast timescale, and, therefore, the contribution from the frequency difference between transverse modes on the dynamics can be neglected [53]. Furthermore, in index-guided VCSELs (which are effectively cylindrical waveguides), the light propagates through the structure as a plane-wave in good approximation [61]. Hence, for this type of VCSEL diffraction losses due to off-axis longitudinal propagation of transverse modes are not included in the model [59]. (Note that this is in contrast to the case of gain-guided VCSELs, where optical diffraction plays a role [39, 40].)

2.1. Spatial optical modes. We adopt the linearly polarized (LP) description for the optical modes. This simplification of the full set of transverse magnetic, transverse electric, and hybrid modes for a cylindrical waveguide is valid under the condition of weak guidance [61]. In other words, we consider a VCSEL in which the dielectric constant in the core region is approximately equal to that in the cladding region; see Figure 1. The LP mode profiles Ψ^j_{mn} are given by roots of the Helmholtz equation [59, 61]

$$\frac{\partial^2 \Psi^j_{mn}(r,\phi)}{\partial r^2} + \frac{1}{r}\frac{\partial \Psi^j_{mn}(r,\phi)}{\partial r} + \frac{1}{r^2}\frac{\partial^2 \Psi^j_{mn}(r,\phi)}{\partial \phi^2} + q_n^2 \Psi^j_{mn}(r,\phi) = 0, \tag{2.4}$$

where

$$q_n^2 \equiv \begin{cases} u_n^2 = \epsilon^{\text{core}} \kappa_n^2 - \beta^2, & r \le a, \\ -w_n^2 = \epsilon^{\text{clad}} \kappa_n^2 - \beta_n^2, & r > a, \end{cases} \tag{2.5}$$

and ϵ^{core} and ϵ^{clad} are the dielectric constants inside the core and cladding regions, respectively. We remark that, due to changes in the inversion, the dielectric constants may vary slightly when the VCSEL is lasing. However, these variations are two orders of magnitude lower than their fixed values, and, as is common in the literature, we treat ϵ^{core} and ϵ^{clad} as constants. The other parameters are the radius a of the active (core) region of the VCSEL, modal vacuum-wavenumbers κ_n, and the longitudinal wavenumber β (which is a separation of variables constant).

Solutions of (2.4) take the form

$$\Psi_{mn}^{j=c,s}(r,\phi) = \Psi_{mn}(r) \times \begin{cases} \sin(m\phi), & j = s, \\ \cos(m\phi), & j = c, \\ 1, & m = 0, \end{cases} \tag{2.6}$$

where

$$\Psi_{mn}(r) = \begin{cases} \dfrac{J_m(u_n r/a)}{J_m(u_n)}, & r \le a, \\ \dfrac{K_m(w_n r/a)}{K_m(w_n)}, & r > a. \end{cases} \tag{2.7}$$

In other words, radial solutions are given by Bessel functions of the first kind J_m inside the core region ($r \le a$) and modified Bessel functions of the second kind K_m inside the cladding region ($r > a$). These solutions conform to the physical structure of the VCSEL; that is, the J_m remain finite with zero slope at $r = 0$, and the K_m for $r > a$ remain finite and rapidly decay to zero with r. Furthermore, continuity is ensured at the boundary $r = a$. Their azimuthal dependence is given in terms of orthogonal sine and cosine functions, which are denoted by the superscript j.

Figure 2 shows the cosine oriented LP mode profiles $\Psi_{mn}^{c}(r,\phi)$ for $0 \le m \le 3$ and $1 \le n \le 4$ in projection onto the (r,ϕ)-plane. (A rotation over 90 degrees yields the sine oriented LP modes $\Psi_{mn}^{s}(r,\phi)$.) Here dark red corresponds to regions of high intensity and dark blue to regions of low intensity. From column to column the radial order n of the mode increases from $n = 1$ to $n = 4$; note that n represents the number of peaks in the radial mode profiles counted from the origin outward. From row to row the azimuthal order m of the mode increases from $m = 0$ to $m = 3$; note that $2m$ represents the number of dark sectors over one full rotation.

2.2. Expansion of the angular dependence. To make (2.1) and (2.2) amenable to a full nonlinear bifurcation analysis with numerical tools such as `DDE-BIFTOOL` [11] and `PDDE-CONT` [51], we now resolve the spatial dependence by expansion in suitable basis functions. A common first step [59] is to expand the angular part of the inversion $N(r,\phi,t)$ as a Fourier series

$$N(r,\phi,t) = \sum_{k=0}^{\infty} \big(N_{ck}(r,t)\cos(k\phi) + N_{sk}(r,t)\sin(k\phi) \big). \tag{2.8}$$

The resulting set of rate equations is premultiplied by both $\sin(k\phi)$ and $\cos(k\phi)$ and integrated over ϕ to obtain the following system of PDEs:

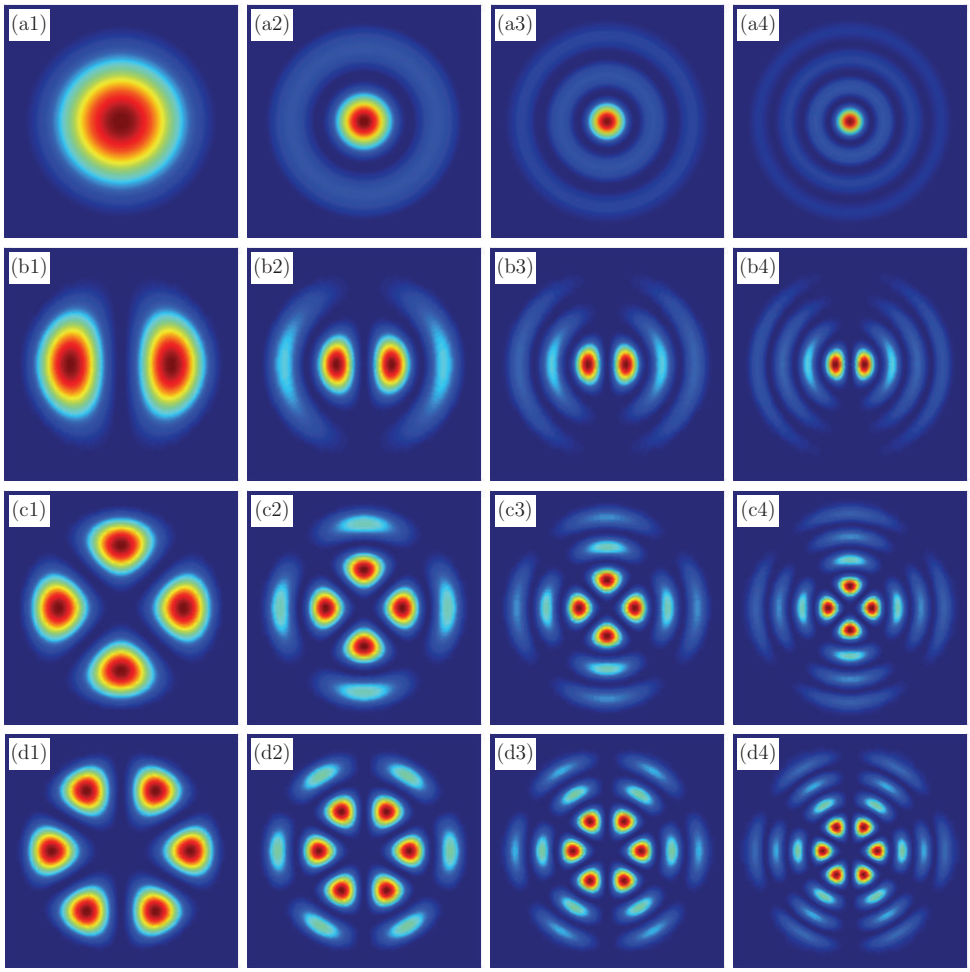


Figure 2. *The linearly polarized optical modes LP_{mn} for $0 \le m \le 3$ and $1 \le n \le 4$; the index m changes from row to row and the index n from column to column.*

$$\frac{dE^j_{mn}}{dt} = (1+i\alpha)\xi^j_{mn}E^j_{mn} + F^j_{mn}(t), \tag{2.9}$$

$$\begin{aligned} T\frac{\partial N_{c0}(r,t)}{\partial t} &= \frac{d_f}{r}\frac{\partial}{\partial r}\left(r\frac{\partial N_{c0}}{\partial r}\right) + \int_0^{2\pi}\frac{P(r,\phi,t)}{2\pi}\,d\phi - N_{c0} \\ &\quad - \sum_n\left((1+2\xi_{0n})|E_{0n}|^2|\Psi_{0n}|^2 + \sum_m\sum_{j=c,s}(1+2\xi^j_{mn})|E^j_{mn}|^2|\Psi_{mn}|^2\right), \end{aligned} \tag{2.10}$$

$$\begin{aligned} T\frac{\partial N_{ck}(r,t)}{\partial t} &= \frac{d_f}{r}\frac{\partial}{\partial r}\left(r\frac{\partial N_{ck}}{\partial r}\right) - \frac{d_f}{r^2}k^2N_{ck} + \int_0^{2\pi}\frac{P(r,\phi,t)}{\pi}\cos(k\phi)\,d\phi - N_{ck} \\ &\quad - \sum_m\sum_n|\Psi_m n|^2\delta_{(2m)k}((1+2\xi^c_{mn})|E^c_{mn}|^2 - (1+2\xi^s_{mn})|E^s_{mn}|^2), \end{aligned} \tag{2.11}$$

$$T\frac{\partial N_{sk}(r,t)}{\partial t} = \frac{d_f}{r}\frac{\partial}{\partial r}\left(r\frac{\partial N_{sk}}{\partial r}\right) - \frac{d_f}{r^2}k^2N_{sk} + \int_0^{2\pi}\frac{P(r,\phi,t)}{\pi}\sin(k\phi)\,d\phi - N_{sk}, \tag{2.12}$$

where $\delta_{(2m)k}$ is the Kronecker symbol meaning that $\delta_{(2m)k} = 1$ for $2m = k$ and $\delta_{(2m)k} = 0$ otherwise. The modal gains $\xi^j_{mn}(t)$ are now given as

$$\begin{aligned}
\xi_{0n}(t) &= \int_0^1 |\Psi_{0n}(r)|^2 N_{c0}(r,t) r dr, \\
\xi^c_{mn}(t) &= \int_0^1 |\Psi_{mn}(r)|^2 \left(N_{c0}(r,t) + \frac{1}{2}\sum_{k=1}^{\infty} \delta_{(2m)k} N_{ck}(r,t) \right) r dr, \\
\xi^s_{mn}(t) &= \int_0^1 |\Psi_{mn}(r)|^2 \left(N_{c0}(r,t) - \frac{1}{2}\sum_{k=1}^{\infty} \delta_{(2m)k} N_{ck}(r,t) \right) r dr.
\end{aligned}$$

We assume that the pump current is rotationally symmetric and constant in time, so that $P(r,\phi,t) \equiv P(r)$ and

$$\int_0^{2\pi} \frac{P(r)}{2\pi}\, d\phi = P(r)$$

in (2.10) and

$$\int_0^{2\pi} \frac{P(r)}{\pi} \cos(k\phi)\, d\phi = \int_0^{2\pi} \frac{P(r)}{\pi} \sin(k\phi)\, d\phi = 0$$

in (2.11) and (2.12).

Finally, we note that for $k \neq 2m$, (2.11) is homogeneous. Moreover, under the assumption of a symmetric pump current, (2.12) is always homogeneous and, thus, the two-dimensional PDE describing the inversion is reduced to the following system of PDEs with a one-dimensional radial dependence:

$$\frac{dE^j_{mn}}{dt} = (1 + i\alpha)\xi^j_{mn} E^j_{mn} + F^j_{mn}(t), \tag{2.13}$$

$$\begin{aligned}
T\frac{\partial N_0(r,t)}{\partial t} &= d_f \left[\frac{1}{r}\frac{\partial}{\partial r}\left(r \frac{\partial N_0}{\partial r} \right) \right] + P(r) - N_0 \\
&\quad - \sum_n \left((1 + 2\xi_{0n}) |E_{0n}|^2 |\Psi_{0n}|^2 - \sum_m \sum_{j=c,s} (1 + 2\xi^j_{mn}) |E^j_{mn}|^2 |\Psi_{mn}|^2 \right),
\end{aligned} \tag{2.14}$$

$$\begin{aligned}
T\frac{\partial N_k(r,t)}{\partial t} &= d_f \left[\frac{1}{r}\frac{\partial}{\partial r}\left(r \frac{\partial N_k}{\partial r} \right) \right] - \frac{d_f}{r^2} k^2 N_k - N_k \\
&\quad - \sum_m \sum_n |\Psi_{mn}|^2 \delta_{(2m)k} ((1 + 2\xi^c_{mn}) |E^c_{mn}|^2 - (1 + 2\xi^s_{mn}) |E^s_{mn}|^2),
\end{aligned} \tag{2.15}$$

where, for ease of presentation, we have written $N_k \equiv N_{ck}$, $k = 0, 2, 4, \ldots$.

It is clear that the number of terms of the expansion that we need to keep depends on the number of optical modes we wish to consider. Here the *azimuthal orders* of these modes are given by the index m. For example, for a VCSEL which supports only rotationally symmetric modes $m = 0$ (Figure 2, row (a), and the case we will consider later), (2.15) is not needed; nor is the last, double summation term of (2.14). In this way, we are left with a single PDE with a

one-dimensional radial dependence describing the inversion. Likewise, for a VCSEL supporting both rotationally symmetric and first-order azimuthal modes $m = 0, 1$ (recall Figure 2, rows (a) and (b)), we need two PDEs $\partial N_0(r,t)/\partial t$ and $\partial N_2(r,t)/\partial t$ for a full description. It follows that a system of $(M+1)$ PDEs, with one-dimensional radial dependencies, describe VCSELs which support modes up to an Mth azimuthal order.

2.3. Expansion of the radial dependence. In the literature, the most common way of dealing with the radial PDEs (2.14) and (2.15) is through the use of a finite difference discretization scheme [19, 59]. This approach works well for rotationally symmetric modes [19]. However, it runs into problems for modes with $m > 0$. Specifically, the $1/r^2$ term in (2.15) introduces a singularity at $r = 0$. This singularity is generally avoided by not placing a mesh point of the finite difference scheme at $r = 0$; accuracy of this approach is typically tested by comparing results for increasingly smaller mesh intervals as $r \to 0$ [59].

A second method for arriving at a spatially resolved model is to expand the radially dependent inversion in terms of another set of orthogonal functions. As the optical modes are written in terms of Bessel functions, it has been proposed to use such a set of functions to expand the radially dependent inversion [26, 41]. In particular, the computation package `VISTAS` [26] employs the use of *zeroth order* Bessel functions of the first kind. However, we note that this expansion still leaves terms depending on $1/r$. Thus, the singularity problem at $r = 0$ is unresolved. (Again, `VISTAS` avoids computing values at $r = 0$.)

In this paper, we take a more general approach. Namely, we expand each inversion coefficient $N_k(r,t)$ in terms of a *kth order* Bessel function expansion of the first kind,

$$N_k(r,t) = \sum_{q=1}^{\infty} N_{kq}(t) J_k(\gamma_{k,q} r), \tag{2.16}$$

where the $\gamma_{k,q}$ are roots of J_k. In this way, noting the scaling of the Bessel function arguments, the boundary conditions of the PDE problem [19] are matched; that is, $\frac{\partial N_k(0,t)}{\partial r} = 0$ and $N_k(1,t) = 0$.

Most importantly, the kth order Bessel functions $J_k(\gamma_{k,q} r)$ are exact eigenfunctions of the kth order operators

$$\nabla_k^2 \equiv \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial}{\partial r}\right) - \frac{k^2}{r^2}, \tag{2.17}$$

which appear in (2.14) and (2.15). We proceed by inserting the expansion (2.16) into (2.14) and (2.15), multiplying each of the expanded rate equations for N_k by a corresponding $rJ_k(\gamma_{k,l} r)$ and integrating over r. Thus, through the orthogonality condition

$$\int_0^1 J_k(\gamma_{k,q} r) J_k(\gamma_{k,l} r) r dr = \frac{\delta_{ql}}{2}[J_{k+1}(\gamma_{k,l})]^2, \tag{2.18}$$

one obtains the following set of *spatially resolved* rate equations:

$$\frac{dE_{mn}^j(t)}{dt} = (1 + i\alpha)\xi_{mn}^j E_{mn}^j + F(t), \tag{2.19}$$

$$T\frac{dN_{0q}(t)}{dt} = -(\gamma_{0,q}^2 d_f + 1)N_{0q} + \rho_q - \sum_n \left((1+2\xi_{0n})|E_{0n}|^2 \beta_{0n}^{0q} + \sum_m \sum_{j=c,s} (1+2\xi_{mn}^j)|E_{mn}^j|^2 \beta_{mn}^{0q} \right), \tag{2.20}$$

$$T\frac{dN_{kq}(t)}{dt} = -(\gamma_{k,q}^2 d_f + 1)N_{kq}(t) - \sum_m \sum_n \delta_{(2m)k} \left((1+2\xi_{mn}^c)|E_{mn}^c|^2 - (1+2\xi_{mn}^s)|E_{mn}^s|^2 \right) \beta_{mn}^{kq}, \tag{2.21}$$

where the two integrals

$$\rho_q = \frac{2}{[J_1(\gamma_{0,q})]^2} \int_0^1 P(r) J_0(\gamma_{0,q} r)\, r dr, \tag{2.22}$$

$$\beta_{mn}^{kq} = \frac{2}{[J_{k+1}(\gamma_{k,q})]^2} \int_0^1 |\Psi_{mn}(r)|^2 J_k(\gamma_{k,q} r)\, r dr \tag{2.23}$$

describe the spatial overlaps between the inversion density with the pump current and the electric field, respectively. These integrals depend on the material properties and geometry of the VCSEL model under consideration. Furthermore, the modal gains are now given as

$$\xi_{0n} = \sum_{q=1}^{\infty} \int_0^1 \left(|\Psi_{0n}(r)|^2 N_{0q} J_0(\gamma_{0,q} r) \right) r dr, \tag{2.24}$$

$$\xi_{mn}^c = \sum_{q=1}^{\infty} \int_0^1 |\Psi_{mn}(r)|^2 \left(N_{0q} J_0(\gamma_{0,q} r) + \frac{1}{2} \sum_{k=1}^{\infty} \delta_{(2m)k} N_{kq} J_k(\gamma_{k,q} r) \right) r dr, \tag{2.25}$$

$$\xi_{mn}^s = \sum_{q=1}^{\infty} \int_0^1 |\Psi_{mn}(r)|^2 \left(N_{0q} J_0(\gamma_{0,q} r) - \frac{1}{2} \sum_{k=1}^{\infty} \delta_{(2m)k} N_{kq} J_k(\gamma_{k,q} r) \right) r dr. \tag{2.26}$$

Importantly, the integrals above need to be evaluated only once and then appear as constants. Hence, their evaluation does not add to the overall computation time. We emphasize that (2.19)–(2.26) describe a spatially resolved VCSEL in which both azimuthally independent ((2.20) and (2.24)) and azimuthally dependent ((2.21), (2.25), and (2.26)) transverse optical modes may exist.

In practice, the Bessel function expansion (2.16) will have to be truncated at some finite order. The arguments for this truncation are more subtle than for the Fourier expansion (2.8). Clearly, there is no upper bound for the sum in (2.16). However, we can consider the timescale at which the higher-order terms in the expansion $N_{kq}(t)$ are damped. Specifically, this timescale is given by the first term in the right-hand side of (2.20) and (2.21) as

$$\tau_{kq} = \frac{1}{1+\gamma_{k,q}^2 d_f}. \tag{2.27}$$

As the values of $\gamma_{k,q}^2$ increase with q, the higher-order terms $N_{kq}(t)$ tend to zero. The convergence of solutions for increasing terms in the Bessel function expansion can also be tested numerically.

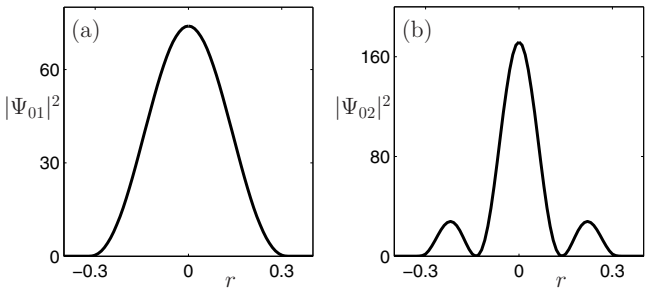


Figure 3. *Radial profile* $|\Psi_{01}|^2$ *of the optical mode* LP_{01} (a), *and* $|\Psi_{02}|^2$ *of* LP_{02} (b).

2.4. Reduced model of a two-mode VCSEL with optical feedback. We are concerned here with a VCSEL that is subject to delayed optical feedback from an external mirror. While this question can be studied, in principle, for any number of modes with the truncated rate equation model presented above, we consider here the case that the VCSEL supports the first two optical modes LP_{01} and LP_{02}. Considering only these two modes is realistic; for example, control of individual modes, in a laser supporting the LP_{01} and LP_{02} modes, has been demonstrated via external injection [43, 44]. Physically, these concentric modes are desirable due to their high efficiencies when coupled to optical fibers. Likewise, different modes may be selected via frequency selective feedback [38]. This two-mode set-up has been studied theoretically in [19, 53, 54, 55].

Importantly, the first two optical modes do not have an azimuthal component, so that they can be described by their radial dependence alone. Figure 3 shows the two radial profiles $|\Psi_{01}|^2$ and $|\Psi_{02}|^2$ of LP_{01} and LP_{02}. For the two modes LP_{01} and LP_{02} the rate equations (2.19)–(2.21) reduce to

$$\frac{dE_1(t)}{dt} = (1+i\alpha)\xi_1 E_1(t) + F_1(t), \tag{2.28}$$

$$\frac{dE_2(t)}{dt} = (1+i\alpha)\xi_2 E_2(t) + F_2(t), \tag{2.29}$$

$$T\frac{dN_q(t)}{dt} = -(\gamma_{0,q}^2 d_f + 1)N_q(t) + \rho_q - \sum_{n=1}^{2}\left((1+2\xi_n)|E_n(t)|^2\beta_n^{0q}\right), \tag{2.30}$$

where, for ease of presentation, we write $E_{0n} \equiv E_n$, $\xi_{0n} \equiv \xi_n$, and $N_{0q} \equiv N_q$. Here (2.30) are the expansion equations of the radial dependence of the inversion. We consider here a total of 14 expansion terms, that is, $q = 1 \ldots 14$, which we found guarantees a relative error of below 0.01% at the steady-state solutions. We stress that this choice of 14 expansion terms depends only on the value d_f of the diffusion, so that it is valid for any feedback. The overlap integrals are now given as

$$\xi_n = \sum_{q=1}^{14}\int_0^1 \left(|\Psi_n(r)|^2 N_q J_0(\gamma_{0,q} r)\right) r dr, \tag{2.31}$$

$$\rho_q = \frac{2}{[J_1(\gamma_{0,q})]^2} \int_0^1 P(r) J_0(\gamma_{0,q} r)\, r dr, \tag{2.32}$$

$$\beta_n^{0q} = \frac{2}{[J_1(\gamma_{0,q}^2)]} \int_0^1 |\Psi_n(r)|^2 J_0(\gamma_{0,q} r)\, r dr. \tag{2.33}$$

As in [19], we consider a disc-shaped profile with pumping confined to a core region of radius 0.3 of the VCSEL, as given by

$$P(r) = P_{\max}\big(1 + \operatorname{erf}(2\sqrt{75}(-r + 0.3))\big)/2. \tag{2.34}$$

The other laser parameters are fixed to the realistic values $\alpha = 3$, $T = 750$, $d_f = 0.05$, and $P_{\max} = 2$.

We are interested here in the influence of self-feedback versus cross-feedback of individual modes of the rate equation model. To this end, we need to specify the optical feedback terms F_1 and F_2 in (2.28) and (2.29). As in [19] we consider

$$F_1(t) = \kappa e^{iC_p}\big[\eta E_1(t-\tau) + (1-\eta)E_2(t-\tau)\big], \tag{2.35}$$

$$F_2(t) = \kappa e^{iC_p}\big[\eta E_2(t-\tau) + (1-\eta)E_1(t-\tau)\big], \tag{2.36}$$

where the feedback enters with a common feedback strength κ and feedback phase C_p. In an experiment the feedback phase C_p can be adjusted by changing the effective length of the external cavity at a subwavelength scale; this can be achieved by a piezo-electric actuator or via small changes of the pump current [15, 25]. In the investigations presented below, we fix the feedback phase at $C_p = 0$; a study into the effects of changing C_p will be reported elsewhere. Note also that C_p and the line-width enhancement factor α are related via the symmetry $(\alpha, C_p) \mapsto -(\alpha, C_p)$ [56], which means that the choice of a positive sign for α in (2.1) does not affect the ECM structure under investigation.

The delay time is set to $\tau = 500$, which corresponds to a distance of approximately 10 cm between the laser and the mirror. The homotopy parameter η represents the amount of self-feedback versus cross-feedback. This allows the two optical modes to couple via the external cavity, that is, via the optical feedback. Indeed, one would expect in general to observe a certain amount of cross-feedback; for example, the modes could easily cross-couple due to deformations in the shape of the external mirror. However, the precise amount of cross-coupling is very hard to estimate or measure. This motivates our bifurcation study with emphasis on how the cross-coupling parameter η affects the dynamics of the VCSEL over the entire range from pure self-feedback $\eta = 1$ to pure cross-feedback $\eta = 0$.

Equations (2.28)–(2.30) with (2.31)–(2.36) are a system of 18 (real-valued) delay differential equations, which provide a good approximation of the delayed PDE model of a VCSEL. This model combines a relatively small system size with the fact that the overlap integrals can be calculated a priori. Compared to the model of dimension 105 considered in [19], where the inversion equation was discretized with a second-order finite difference scheme, the model presented here offers a substantial speed-up. For example, stability calculations of steady states that took 452.46 s (7.54 minutes) per point on a branch in [19] now take only 6.41 s on a 3.4 GHz `Xeon` processor. As a consequence, we are now able to present a detailed multiparameter bifurcation analysis of the ECMs, their stability, and even bifurcating periodic solutions.

3. External cavity modes. The basic solutions of a laser subject to optical feedback are known as external cavity modes (ECMs). They correspond to constant intensity of the laser at a frequency that is selected by the external cavity. In other words, ECMs are due to the fact that the light makes one round trip outside the laser [33]. In the case of model (2.28)–(2.36) of a VCSEL with optical feedback, the ECMs are of the form [19]

$$\big(E_1(t), E_2(t), N_q(t)\big) = (R_1 e^{i\omega_s t}, R_2 e^{i\omega_s t + i\Phi}, N_1, \ldots, N_Q), \tag{3.1}$$

where $R_1, R_2, \omega_s, \Phi, N_1, \ldots, N_Q \in \mathbb{R}$. As with other types of lasers subject to optical feedback, the ECMs are group orbits under the rotational S^1-symmetry of the governing equations [31]. In order to obtain an isolated solution that can be continued numerically, we first transform (2.28)–(2.36) to a rotating frame of reference with frequency b. By setting $b = \omega_s$, we find the ECMs as steady-state solutions; see [13, 21, 31] for further details.

The ECMs of (2.28)–(2.36) lie on closed curves, called ECM-components, that are parametrized by the periodic parameter C_p. As C_p is varied, ECMs are born in pairs in a saddle-node bifurcation, move over the respective ECM-component, and then disappear in pairs in another saddle-node bifurcation. The ECM-components were computed in [19] with the continuation package `DDE-BIFTOOL`, where it was shown that there are several types of ECM-components. There exist ECM-components for which the first field E_1 is dominant, while the second field E_2 provides a negligible contribution to the total power, and vice versa. More interestingly, there exist ECM-components where both fields contribute to the total power, one of which contains stable ECMs [19].

We focus here on the latter ECM-component and study the stability of the corresponding ECMs in dependence on the feedback strength κ and on the self- versus cross-feedback parameter η. Note that κ is a common parameter to consider in the laser dynamics literature. While it is difficult to experimentally determine the exact amount of light that re-enters the laser, the feedback strength can be varied in an experiment by means of density filters. We concentrate here on the physically relevant region of weak feedback strength where $\kappa \leq 0.007$, which corresponds to a maximum of about 0.001% back-reflected light. The exact amount of self- versus cross-feedback, expressed by the parameter η, is generally not known. Nevertheless, one would certainly expect that there is some amount of cross-feedback, which would mean that $\eta < 1$. Moreover, it may be possible to tune η experimentally, for example, via external mirror shaping or frequency selective optical feedback [38]. This makes η an ideal parameter for a bifurcation study that aims at obtaining a global view of the influence of self- versus cross-feedback over the entire possible range.

We start by computing branches of ECMs as a function of the feedback strength κ for fixed $C_p = 0$ and several values of η. Figure 4 was obtained with `DDE-BIFTOOL` and shows ECM-branches of (2.28)–(2.36) in the (κ, P_1)-plane (left) and the (κ, P_2)-plane (right) for $\eta = 1.0$ (row a), $\eta = 0.9$ (row b), $\eta = 0.6$ (row c), and $\eta = 0.0$ (row d). Segments of branches on which one finds stable ECMs are green, while unstable ECM-branches are red. The stability of ECMs is determined in `DDE-BIFTOOL` from the computation of the rightmost characteristic roots using linear multistep methods [12, 60]. (We remark that an alternative method for determining the stability is to consider the infinitesimal operator of the DDE, as is detailed in [4]. However, we are not aware of an associated computational tool for two-parameter continuation with this technique.) Due to the presence of a fixed delay, there are infinitely

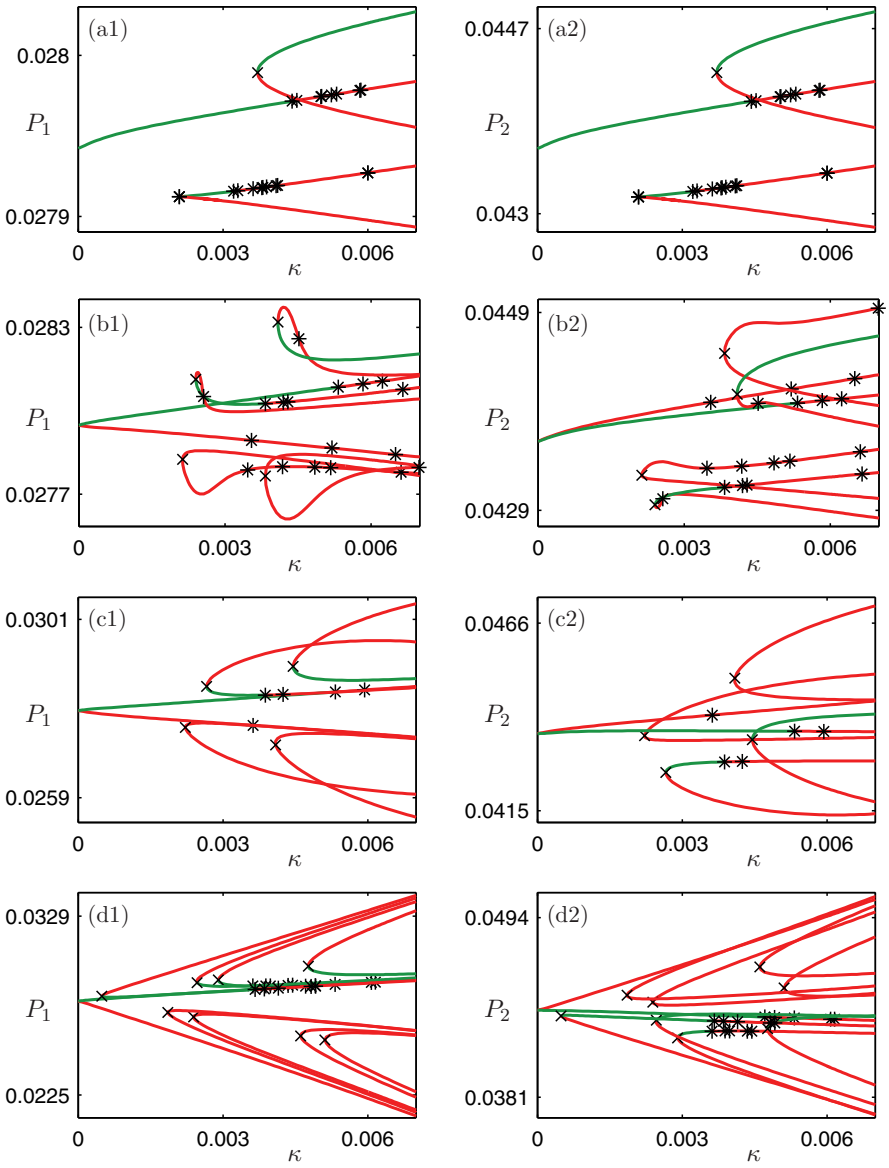


Figure 4. *Branches of ECMs as a function of κ, shown in terms of the powers P_1 of the first mode (left) and P_2 of the second mode (right). From row* (a) *to* (d) *η takes the values* 1.0, 0.9, 0.6, *and* 0.0, *respectively; throughout $C_p = 0$.*

many discrete eigenvalues but only finitely many with positive real part [8, 22]. Hence, as for ODEs, one generically finds saddle-node and Hopf bifurcations when a single parameter is changed. Apart from the ECM-branch that extends from $\kappa = 0$, additional ECMs are born in saddle-node bifurcations ($\times$) as the feedback strength κ is increased. On one of the bifurcating branches, ECMs may be initially stable, but for increasing κ they lose stability in Hopf bifurcations ($*$). This general scenario is typical for a (single-mode) laser with optical feedback [13, 21, 49]. The difference here is that we obtain a projection of the ECM-branches

for each of the two modes, as represented in Figure 4 by the two intensities P_1 and P_2.

Figure 4 illustrates the strong influence of the self- versus cross-feedback parameter η. Namely, for $\eta = 1$, that is, for pure self-feedback, the ECM-branches in panels (a1) and (a2) are double-covered. Any small amount of cross-feedback "unfolds" this situation into twice as many ECM-branches; see Figures 4(b1) and (b2). Notice that there is now a considerable difference in the shape of the ECM-branches of the two modes, which means physically that small changes in κ may cause either an increase or a decrease in the contribution of each mode to the total power. Decreasing η further changes the shape of the ECM-branches; see Figures 4(c1) and (c2). Furthermore, extra ECM-branches enter the region of feedback strength considered here, some of which may be stable; see Figures 4(d1) and (d2).

The ECM-branches in Figure 4 for the mode-expanded model (2.28)–(2.36) agree very well with those in [19, Figures 9 and 10], where a finite difference scheme was used for the carrier distribution N. Of crucial importance is that a stability calculation of a single, fixed ECM shown in Figure 4 took 6.41 s on a 3.4 GHz `Xeon` processor, while the corresponding computation for the finite difference discretized model [19] took 452.46 s (7.54 minutes). This remarkable speed-up in computation time allows us to find and follow bifurcations of the ECMs in parameters, which provides a much broader insight into the stability of ECMs in the two-mode VCSEL. As we will see in section 4, we are even able to continue the bifurcating periodic orbits and determine their stability regions.

3.1. Stability regions of ECMs. Figure 5 shows the curves of saddle-node bifurcations (blue) and Hopf bifurcations (red) in the two-parameter (κ, η)-plane; dark red segments correspond to supercritical and light red segments to subcritical Hopf bifurcations. The bifurcations were identified in Figure 4 and then continued in κ and η with `DDE-BIFTOOL`. Figures 5(a) to (e) are for one particular stable ECM each, whose region of stability is shaded green. The final Figure 5(f) shows the complete ECM stability diagram in the (κ, η)-plane; namely, it is an overlay of panels (a) to (e). Here regions are shown in different shades of green to indicate how many stable ECMs coexist. Specifically, in the region with the lightest shade of green one finds a single stable ECM—for example, directly to the right of $\kappa = 0$. The next darker shade of green corresponds to regions of bistability between two ECMs—for example, directly to the right of the leftmost blue saddle-node curve. The darkest shade of green corresponds to regions in which one finds three coexisting stable ECMs; note that this occurs in the region of small η, that is, for large amounts of cross-feedback. Physically, multistability between ECMs may be exploited for use in, for example, optical flip-flop schemes [28].

More specifically, it can be seen from Figure 5(a) that a stable ECM is born at $\kappa = 0$. As κ is increased, additional stable ECMs are born in saddle-node bifurcations S_2, S_3, S_4, and S_5; see Figures 5(b) to (e). The resulting additional stability regions lead to the large regions in Figure 5(f) where two stable ECMs coexist. Note that some of these regions extend over the entire range of η; see, for example, Figure 5(f) for feedback strengths of $\kappa = 0.0034$ and $\kappa = 0.0051$. Furthermore, for $(\kappa, \eta) \approx (0.006, 0.432)$ and in the range of higher values of cross-feedback, Figure 5(c) shows an additional curve of saddle-node bifurcations entering our κ-range of interest. This curve leads to the birth of the extra (stable) ECM-branches we observed in Figures 4(d1) and (d2). As Figure 5(f) shows, this stable ECM leads to a main region of bistable ECMs at low values of κ and η. A second saddle-node curve enters our κ-range of interest at $\eta \approx 0.283$; see Figure 5(e). The bifurcating stable ECMs give rise to regions

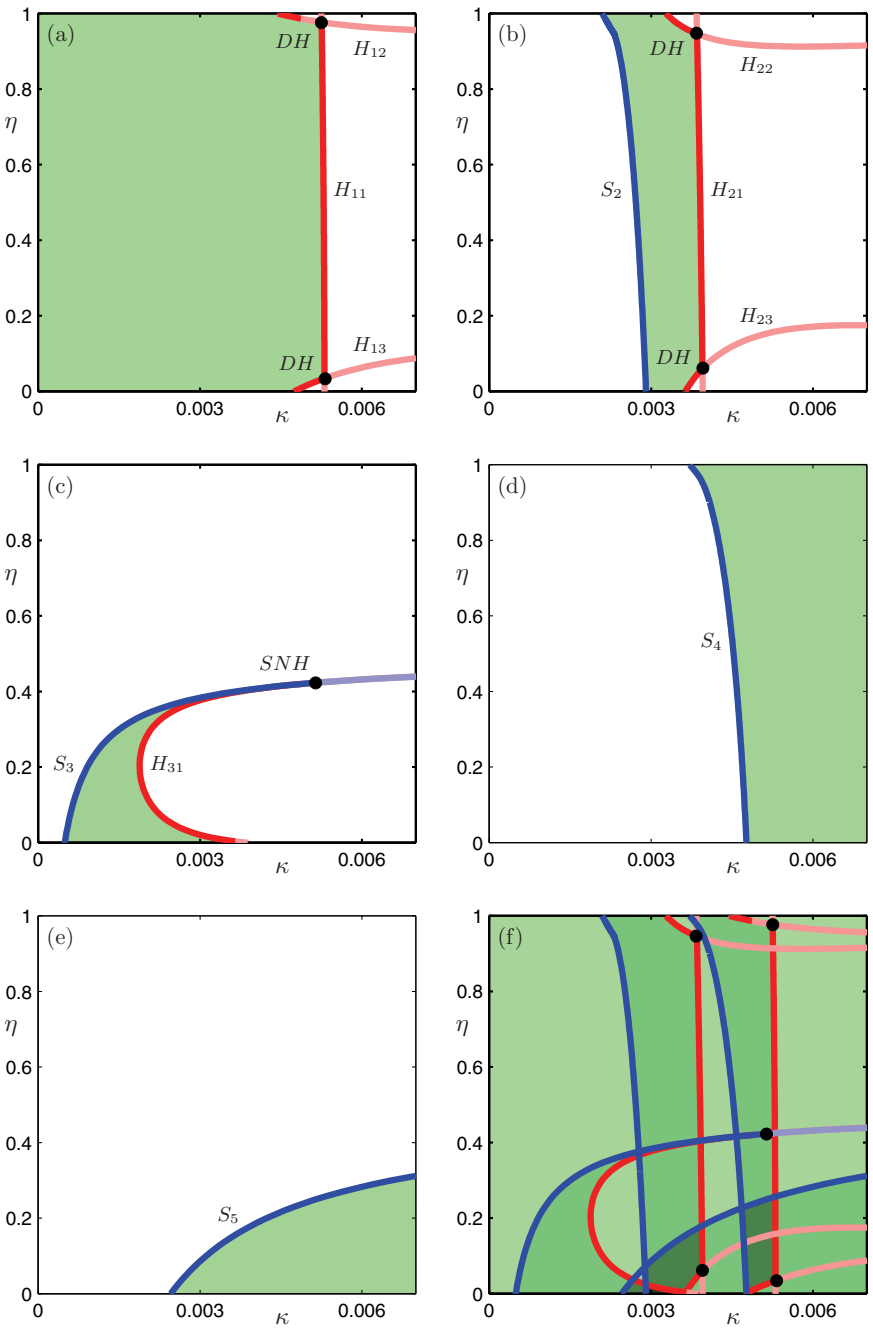


Figure 5. *Stability regions (green) of different ECMs in the (κ, η)-plane, which are bounded by curves of saddle-node bifurcations (blue) and Hopf bifurcations (red). Panel* (f) *is an overlay of the different ECM stability regions in panels* (a)–(e)*, where increasingly darker shading indicates regions of increasing multistability; throughout $C_p = 0$.*

in which one finds three stable ECMs; see, for example, Figure 5(f) for $(\kappa, \eta) = (0.0035, 0.07)$ and $(\kappa, \eta) = (0.0051, 0.1)$. Finally, we note a small and barely visible region in Figure 5(f), just above $(\kappa, \eta) = (0.0032, 0.0)$, in which one finds four coexisting, stable ECM solutions.

The stable ECMs in Figures 5(a) to (c) are destabilized in Hopf bifurcations H_{11}, H_{12}, H_{13}, H_{21}, H_{22}, H_{23}, and H_{31} as κ is increased. For the most part, these Hopf bifurcations are supercritical (dark red curves) and lead to stable periodic solutions, which are the subject of section 4. Moreover, interactions of the bifurcation curves in Figure 5 correspond to codimension-two bifurcation points. We find four double-Hopf bifurcations DH, where two Hopf curves cross and interact. Close to these points the respective Hopf bifurcations change from supercritical (dark red) to subcritical (light red). We also find a codimension-two saddle-node Hopf bifurcation point SNH at $(\kappa, \eta) \approx (4.01 \times 10^{-3}, 0.420)$, where the Hopf curve H_{31} forms a tangency with the saddle-node curve S_3. The two curves bound the stability regions of the associated ECM and are subcritical beyond SNH.

4. Bifurcating periodic solutions. We now address the question of what kind of stable periodic solutions may bifurcate from the different stable ECMs at the supercritical Hopf bifurcation curves shown in Figure 5. To answer this question, we follow periodic orbits and their bifurcations with the `C++` continuation software `PDDE-CONT` [51]. Note that `PDDE-CONT` contains routines that are specifically designed to analyze systems with S^1-symmetry, such as system (2.28)–(2.36). More specifically, we first compute branches of periodic solutions emanating from the Hopf bifurcations of stable ECMs identified in Figures 5(a) to (c). We then find and follow in the (κ, η)-plane the bifurcation curves that form the stability boundaries. Stability of periodic solutions is determined in `PDDE-CONT` from the Floquet multipliers, that is, from the eigenvalues of the Poincaré return map in a suitable transverse section. Because of the fixed delay, there is an infinite number of discrete Floquet multipliers with the origin of the complex plane as their only accumulation point [22]. Hence, as for ODEs, one generically finds saddle-node of limit cycle, period-doubling, and torus (or Neimark–Sacker) bifurcations when a single parameter is changed.

Figures 6 and 7 show two examples of stable periodic solutions that bifurcate from the Hopf bifurcation curve H_{11} in Figure 5(a) and the Hopf bifurcation curve H_{31} in Figure 5(c), respectively. In Figures 6 and 7, panel (a) shows the branch of periodic solutions as it bifurcates from the corresponding Hopf bifurcation point H. The periodic solution is initially stable (green curve) and then destabilizes when κ is increased in a torus bifurcation T and a period-doubling bifurcation PD, respectively. The other panels of Figures 6 and 7 show the periodic dynamics of a stable periodic solution in terms of the two modes, LP_{01} associated with E_1 (black) and LP_{02} associated with E_2 (gray). Specifically, panels (b1) and (b2) are time series of the powers $P_1 = |E_1|^2$ and $P_2 = |E_2|^2$ over five periods, panel (c) shows the projection onto the E_i-plane ($i = 1, 2$), and panel (d) shows that onto $(E_i, \hat{N})$-space. Here $\hat{N}$ is the mean value of the individual inversion coefficients $N_1 \dots N_{14}$, over the orbit. Note that, because both fields feed from the same reservoir of carriers, the inversion for both fields is the same at any given time t. In panels (c) and (d) the black dot and the gray dot on the periodic orbit are passed at the same time in the direction indicated by the arrows. Finally, panels (e1) and (e2) show the time series of the spatial intensities $I_1 = P_1|\Psi_1|^2$ and $I_2 = P_2|\Psi_2|^2$ of the two modes.

As Figures 6(b1) and (b2) show, for a stable periodic solution emanating from the Hopf

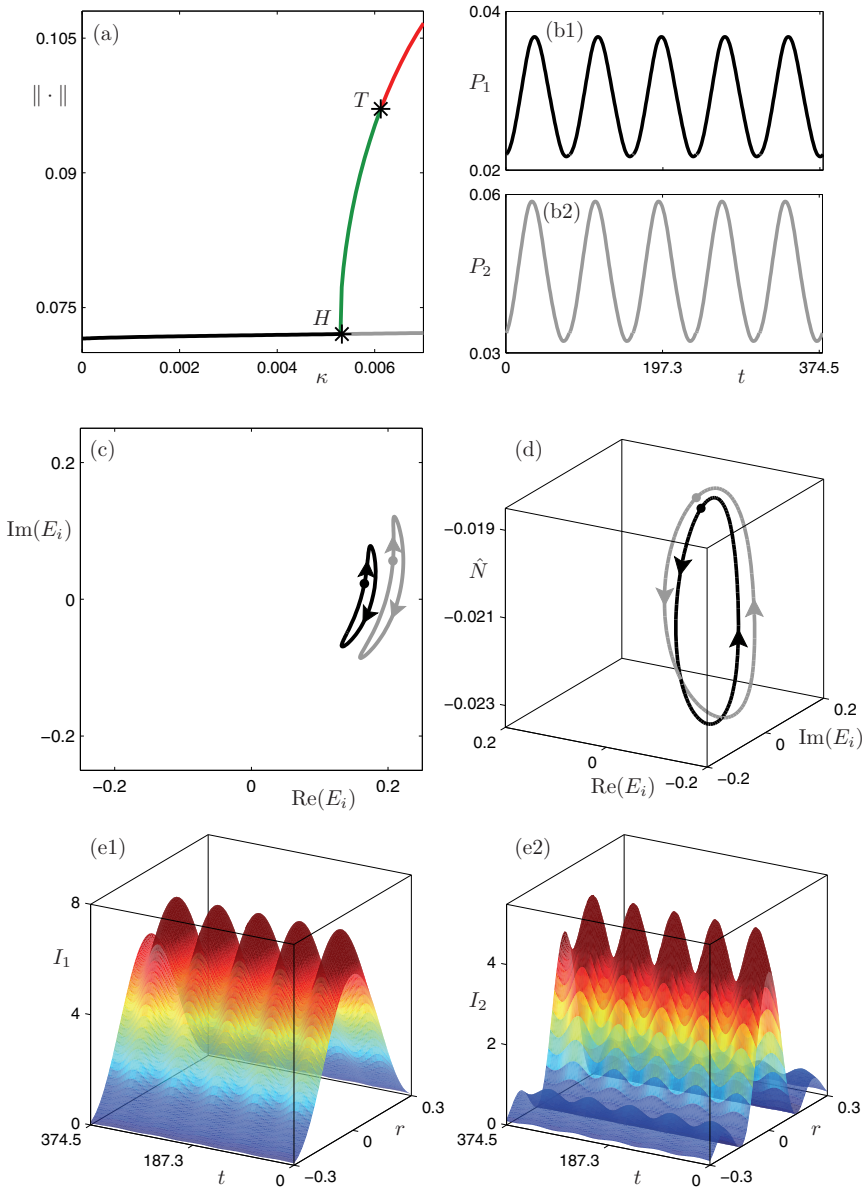


Figure 6. *The branch of periodic solutions for $\eta = 0.6$ emanating from the Hopf bifurcation curve H_{11} in Figure 5(a), and the stable periodic solution for $\kappa = 0.006$ (b)–(e). Panels* (b1) *and* (b2) *show the time series of the powers P_1 and P_2 of the two modes, panels* (c) *and* (d) *are projections onto the E_i-plane and $(E_i, \hat{N})$-space, respectively, and panels* (e1) *and* (e2) *are time series of the spatial intensities I_1 and I_2. Quantities concerning E_1 are plotted black and quantities concerning E_2 are plotted gray; throughout $C_p = 0$.*

curve H_{11} in Figure 5(a), for $\eta = 0.6$ the two electric fields E_1 and E_2 are in phase to very good approximation. The solution moves over the crescent-shaped periodic orbits in panels (c) and (d) in a clockwise direction, as indicated by the arrows. The period is approximately 75 dimensionless time units, which corresponds to a physical frequency of 6.67 GHz. This

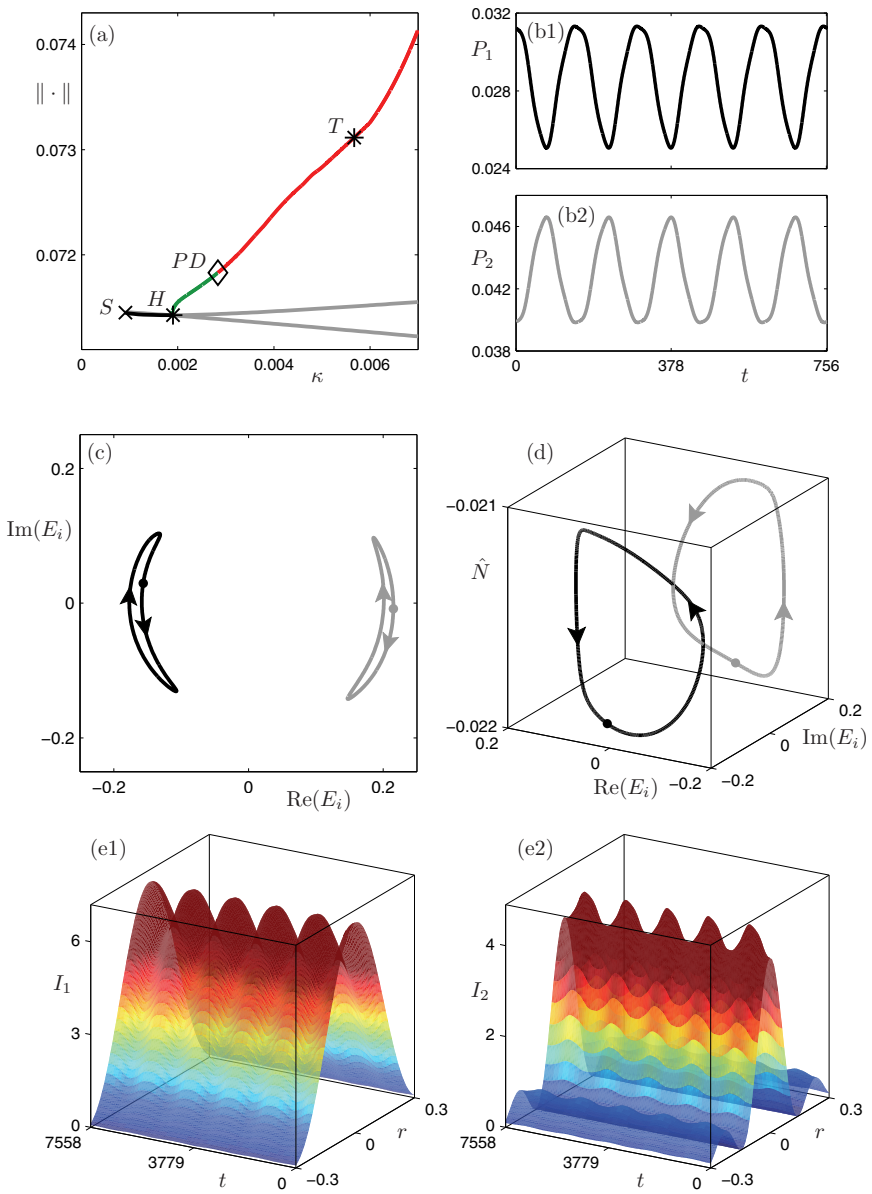


Figure 7. *The branch of periodic solutions for* $\eta = 0.2$ *emanating from the Hopf bifurcation curve* H_{31} *in Figure* 5(c). *The stable periodic solution for* $\kappa = 0.0025$ *is presented in panels* (b)–(e) *in the same projections as in Figure* 6.

identifies the periodic solution physically as a so-called undamped relaxation oscillation (RO). Note that relaxation oscillations of a semiconductor laser are not relaxation oscillations in the sense of slow-fast systems [20]. Rather, they are a damped periodic exchange between the carriers and the electric field of a free-running semiconductor laser, which are observed when the laser is perturbed from steady operation. The spatial nature of damped relaxation

oscillations has been observed experimentally with specialized measurement techniques in [1, 2]. Therefore, it may also be possible to observe the spatiotemporal dynamics of the two modes for the case of sustained, undamped ROs. The in-phase dynamics of the two modes is also apparent from the time series of the spatial intensity profiles I_1 and I_2 in Figures 6(e1) and (e2).

By contrast, as Figures 7(b1) and (b2) show, for a stable periodic solution emanating from H_{31} in Figure 5(c), for $\eta = 0.2$ the two electric fields E_1 and E_2 are in approximate antiphase. Figure 7(c) shows that the solution again moves over crescent-shaped periodic orbits in a clockwise direction, as indicated by the arrows, but the two periodic orbits are now situated about π out of phase; see also panel (d). The periodic solution in Figure 7 has a period of approximately 1512 dimensionless time units, which corresponds to a physical frequency of 331 MHz. This type of oscillation is much slower than ROs and rather takes place on the time-scale of an integer multiple of the external cavity round trip time τ. Therefore, the associated spatiotemporal dynamics of the modes should be easier to observe experimentally [1, 2]. The time series of the spatial intensity profiles I_1 and I_2 in Figures 7(e1) and (e2) clearly show antiphase dynamics. In other words, there is competition between the two modes, with the result that they suppress each other in the total field.

The two types of periodic solutions, in-phase ROs and antiphase external round trip oscillations (EOs), are representative. Specifically, the solution which bifurcates from the Hopf curve H_{21} in Figure 5(b) is an in-phase RO very similar to that shown in Figure 6. On the other hand, solutions born at the remaining Hopf curves, H_{12}, H_{13}, H_{22}, and H_{23} in Figure 5, are in antiphase and have periods on the order of τ. Namely, solutions emanating from H_{12} and H_{22} have a period of approximately 640 time units (781 MHz), and those emanating from H_{13} and H_{23} have a period of approximately 400 time units (1.25 GHz). We remark that a switch between in-phase and antiphase dynamics, as the pump current was increased, was shown to occur in a similar VCSEL model [53, 55].

4.1. Stability regions of periodic solutions. Figure 8 shows the stability regions of periodic solutions in the (κ, η)-plane. They were found with `PDDE-CONT` by detecting, and then continuing in κ and η, saddle-node of limit cycle bifurcations SL (brown), period-doubling bifurcations PD (green), and torus bifurcations T (black). The curves (red) of Hopf bifurcations of ECMs are reproduced from Figure 5. Specifically, Figures 8(a) and (b) show the stability regions (orange) of the in-phase ROs that are born at the Hopf curves H_{11}, H_{12}, H_{13} identified in Figure 5(a), and H_{21}, H_{22}, H_{23} identified in Figure 5(b), respectively. Furthermore, Figure 8(c) shows the stability regions (blue) of the antiphase external round trip oscillations that are born at the Hopf curve H_{31} identified in Figure 5(c). Regions shaded purple indicate bistability between in-phase and antiphase solutions. The final panel, Figure 8(d), shows the complete stability diagram of periodic orbits in the (κ, η)-plane; namely, it is an overlay of panels (a) to (c). Regions are shown in different shades of color to indicate coexistence between different stable periodic solutions.

Figure 8(a) shows that, for most values of cross-coupling η, the stable in-phase ROs emanating from H_{11} are destabilized, as κ is increased, in a torus bifurcation T_{11} (black curve); compare with Figure 6(a). For large η there are additional curves of torus bifurcations emanating from a double-Hopf point, which can be crossed by changing η for fixed κ. The situation for small η is even more complicated, with additional torus, saddle-node of limit

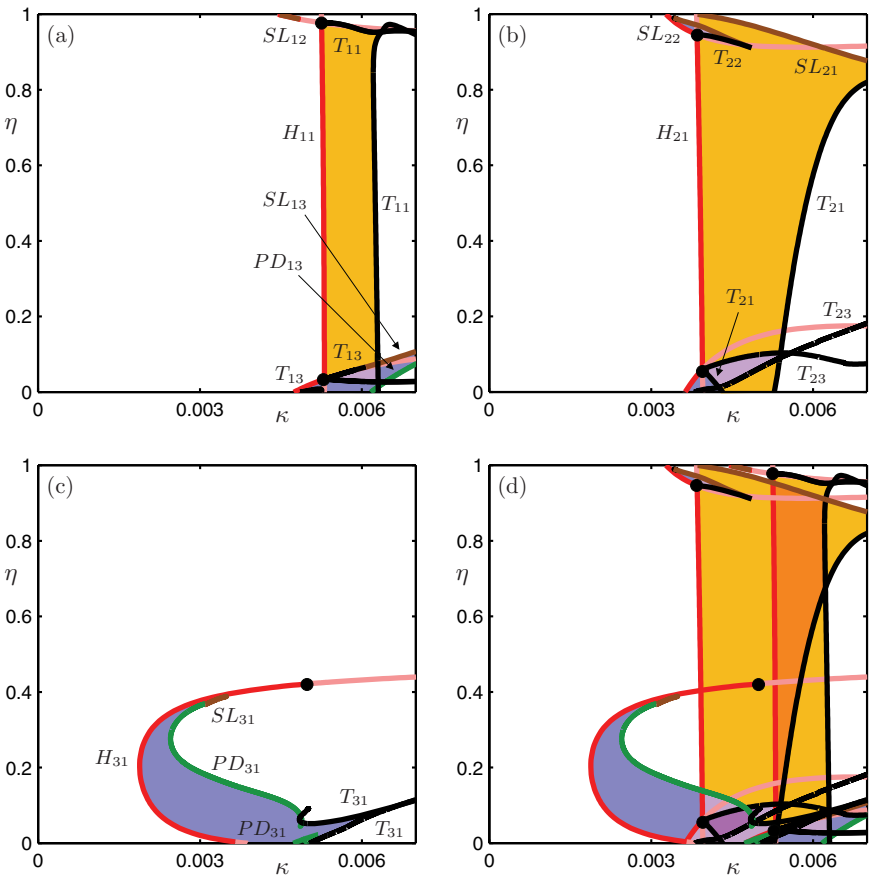


Figure 8. *Stability regions of different periodic solutions in the (κ, η)-plane, which are bounded by curves of Hopf bifurcations (red), saddle-node bifurcations of limit cycles (brown), period-doubling bifurcations (green), and torus bifurcations (black). Stability regions of in-phase ROs are shaded yellow, and that of antiphase EOs are shaded blue. Panel* (d) *is an overlay of the different stability regions in panels* (a)–(c)*, where coloring indicates regions of multistability; throughout $C_p = 0$.*

cycle, and period-doubling curves bounding the stability region of the antiphase oscillations emanating from H_{13}. In fact, the interplay of these bifurcation curves, which are associated with another double-Hopf point, leads to a small region of bistability between the in-phase and antiphase periodic solutions born at the Hopf curves H_{11} and H_{13}, respectively; this region is indicated in Figure 8(a) by purple shading. The situation shown in Figure 8(b), for the in-phase ROs emanating from H_{21}, is very similar. They exist stably in a large region of the (κ, η)-plane, covering effectively the entire η-range. When κ is increased, these periodic solutions are destabilized in torus bifurcations T_{21} (black curve) over almost the entire η-range, namely, for approximately $\eta < 0.8$, and in a saddle-node of limit cycle bifurcation SL_{21} (brown curve) for approximately $\eta > 0.8$. Again, the situation is a bit more complicated for large and small η due to additional bifurcation curves that are connected to double-Hopf points; as a

result a small purple region of bistability between in-phase and antiphase solutions is again created for $\eta < 0.1$.

As Figure 8(c) shows, the blue stability region of the antiphase EOs born at the Hopf curve H_{31} is quite different. The region itself is much smaller and can be found only for large cross-coupling, namely, for $\eta < 0.4$. For most values of η these oscillations are destabilized in a period-doubling bifurcation; compare with Figure 7(a). However, for $\eta < 0.1$ we find destabilization in a torus bifurcation T_{31}, and for $\eta \approx 0.4$ there is a small curve SL_{31} of saddle-node bifurcations of limit cycles. The latter curve is in fact associated with a curve of homoclinic bifurcations that emanates from the saddle-node Hopf point; the homoclinic dynamics of the modes is discussed in the next section.

Overall, Figures 5 and 8 represent the stability information of all ECMs and the periodic solutions emanating from their Hopf bifurcations that can be found for weak feedback κ and any level of self- versus cross-coupling η. The considerable amount of multistability between the different types of solutions can be inferred from the combination of Figures 5(f) and 8(d).

5. Beyond stable periodic solutions. As we have seen, periodic solutions lose their stability when κ is increased beyond the respective stability regions. Interestingly, the in-phase ROs destabilize mainly in torus bifurcations, while the prominent EOs destabilize in period-doubling bifurcations; see Figure 8. We now consider the ensuing dynamics in more detail, where we start from the periodic solutions identified in Figures 6 and 7, respectively.

Figure 9 shows two-frequency dynamics for $\eta = 0.6$. Panel (a) is a bifurcation diagram where for each value of κ the maximum and minimum of the trajectory of the first mode (in black) and the second mode (in gray) are plotted, after transients have died out. The computation was started from the periodic solution for $\kappa = 0.006$ from Figure 6 so that its bifurcations are tracked. We find the Hopf bifurcation at $\kappa \approx 0.0053$, in which this periodic solution is born from the stable ECM, and the torus bifurcation at $\kappa \approx 0.0063$; compare with Figure 5(a). Indeed, for $\kappa > 0.0063$ the dynamics takes place on an invariant torus that manifests itself in the bifurcation diagram of Figure 9(a) as large blocks of points. The dynamics on the torus appears to be quasiperiodic or of very high period. This is confirmed in Figures 9(b) to (d), where the solution for $\kappa = 0.0065$ is shown as time series of the powers, and projections onto the E_i-plane and $(E_i, \hat{N})$-space. Note that the two electric fields E_1 and E_2 still oscillate in phase with the RO period, while their amplitudes are also modulated by a slowly varying envelope with period close to the external cavity round trip time. Figures 9(e1) and (e2) show how this two-frequency dynamics manifests itself in terms of the spatial intensities I_1 and I_2 of the two modes. Notice that the fast oscillations are clearly seen in the flanks of the intensity profiles.

Figure 10 shows chaotic antiphase dynamics for $\eta = 0.2$. The bifurcation diagram in panel (a) was started from the periodic solution for $\kappa = 0.0025$ from Figure 7. One recognizes the Hopf bifurcation from the stable ECM at $\kappa \approx 0.0019$ and the first period-doubling at $\kappa \approx 0.00283$; compare with Figure 7(a). It emerges that this period-doubling is the first in a classic period-doubling route to chaos. Figures 10(b) to (d) show the chaotic solution for $\kappa = 0.00338$ as time series of the powers and projections onto the E_i-plane and $(E_i, \hat{N})$-space. Note from panels (b1) and (b2) that the chaotic dynamics of the individual modes are in antiphase. As for the periodic solution from which it bifurcated, panels (c) and (d) show that the projections onto the two electric fields E_1 and E_2 are located about π out of phase;

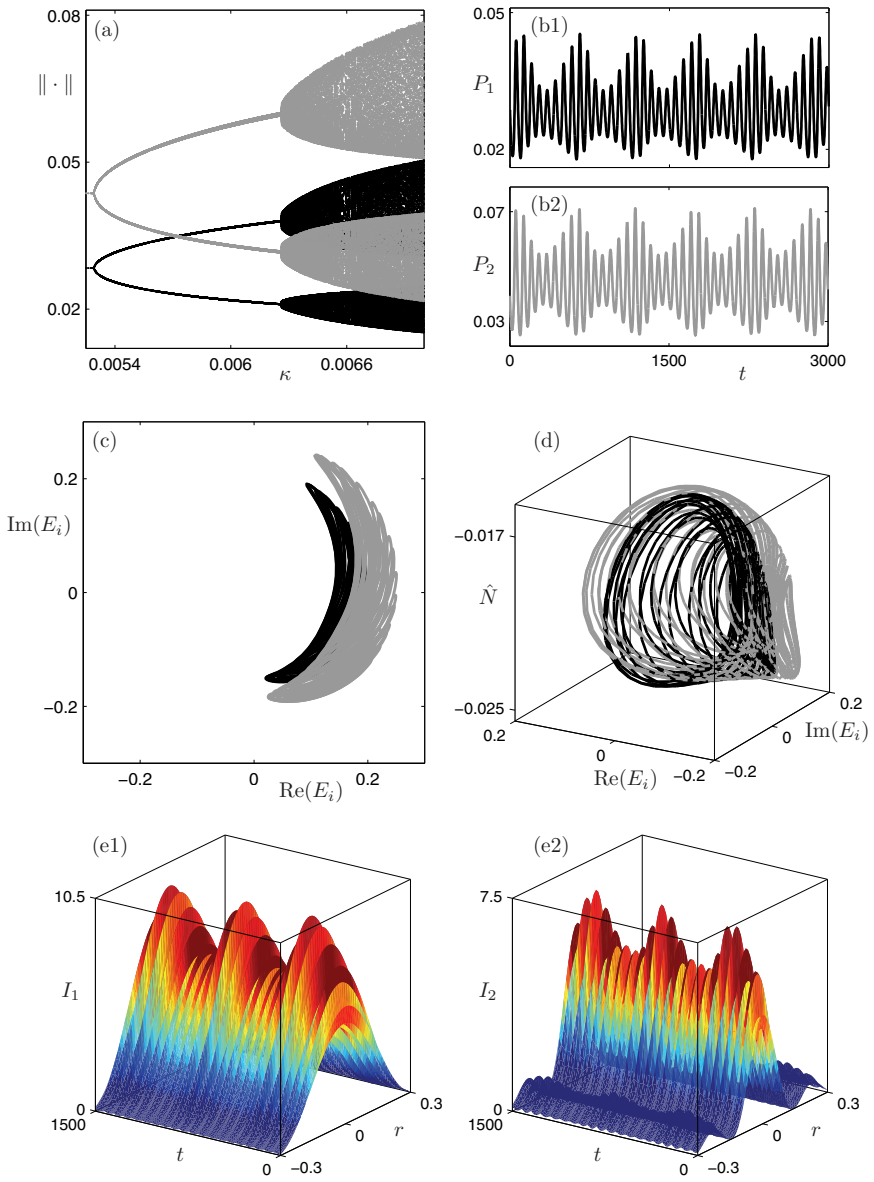


Figure 9. *Bifurcation to an in-phase torus for* $\eta = 0.6$ *with increasing* κ (a). *Stable quasiperiodic dynamics for* $\kappa = 0.065$ *is presented in panels* (b)–(e) *as in Figures* 6 *and* 7.

compare with Figures 7(c) and (d). Note further that the mean fluctuations of the chaotic dynamics are on the timescale of the period of the underlying periodic solution, namely, about three times τ. Figures 10(e1) and (e2) show the chaotic dynamics as time series of the spatial intensities I_1 and I_2 of the two modes. The chaotic dynamics is particularly evident in the flanks of the intensity profiles.

The saddle-node Hopf point SNH in Figures 5(c) and 8(c) gives rise to homoclinic orbits

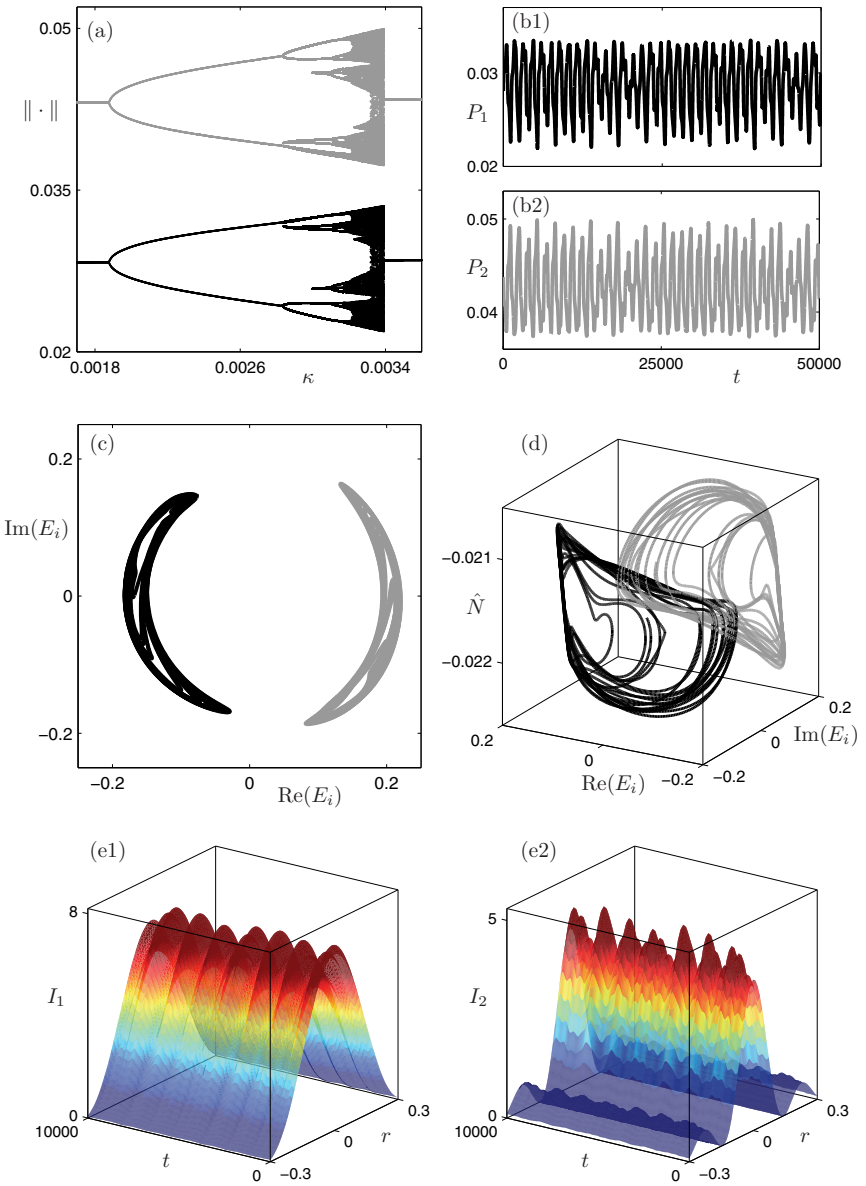


Figure 10. *Period-doubling to an antiphase chaotic attractor for* $\eta = 0.2$ *with increasing* κ (a). *Chaotic dynamics for* $\kappa = 0.00338$ *is presented in panels* (b)–(e) *as in Figures* 6 *and* 7.

of the underlying ECM. Such a homoclinic orbit is shown in Figure 11. It was found by continuing the branch of antiphase EOs for $\eta = 0.39$ in the feedback strength κ. Panel (a) shows this branch plotted in terms of the period T. It folds back and forth while T goes to infinity when κ converges to $\kappa \approx 0.003341$—a clear indication of a homoclinic orbit at this parameter value. This is confirmed by the time series of the powers P_1 and P_2 of the periodic orbit for $\kappa = 0.003341$ that approximates the homoclinic orbit; see Figures 11(b1)

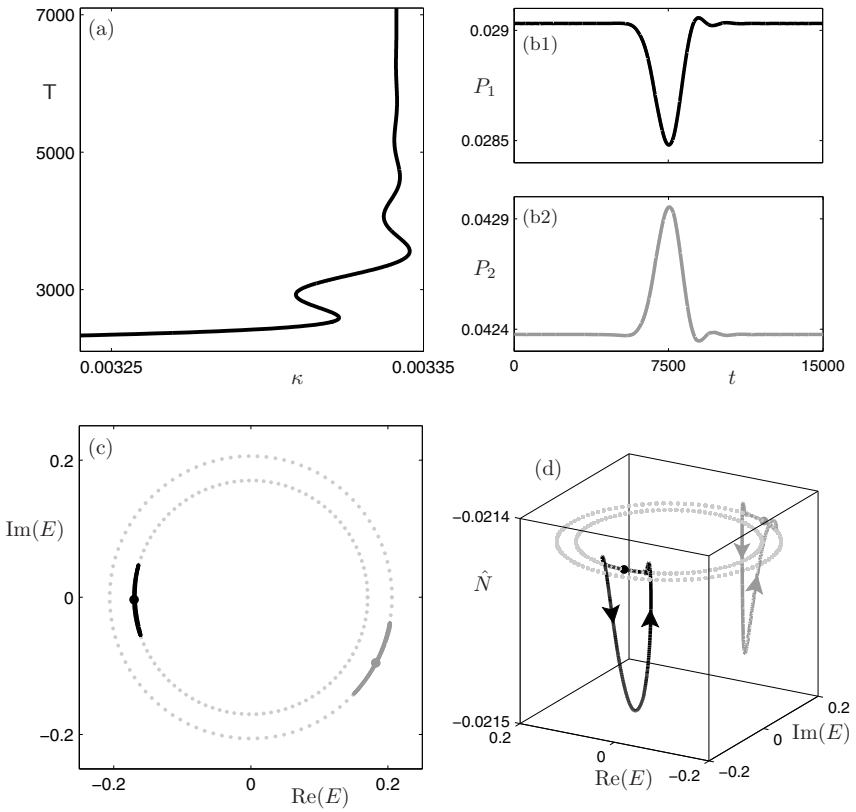


Figure 11. *Branch of periodic solutions converging to a homoclinic orbit for* $\eta = 0.39$ (a)*, and the approximate homoclinic orbit for* $\kappa \approx 0.003341$ *as a series* (b)*, in* E_i*-plane* (c)*, and in* $(E_i, \hat{N})$*-space* (d)*; throughout* $C_p = 0$.

and (b2). The approximate homoclinic orbit itself is shown in projections onto the E_i-plane and $(E_i, \hat{N})$-space in panels (c) and (d), respectively. Here we also show (as dotted circles) the unstable ECM (in terms of both fields) to which the periodic orbit becomes homoclinic. Note the antiphase nature of the time series, which is inherited from the underlying antiphase EO. The power $P_i = |E_i|^2$ of the individual modes is very small; this can be concluded from the fact that the projection of the periodic orbit onto the E_i-plane does not "leave" the ECM circles; see Figure 11(c). Therefore, the spatial mode profiles remain practically constant at the ECM level, which is why they are not shown in Figure 11.

6. Conclusions and outlook. We have considered the dynamics of spatially extended transverse optical modes of a vertical cavity surface-emitting laser (VCSEL) subject to optical feedback as modeled by delayed partial differential equations (DPDEs). By means of an eigenfunction expansion we derived a set of delay differential equations (DDEs) describing the interaction of any number of optical modes. In this general formulation the number of equations that need to be retained after truncation of the expansion depends on which and how many optical modes one wishes to consider, which is determined by the physical size of the VCSEL. The ensuing system of DDEs is quite small when compared with that obtained by

a discretization of the spatial part of the DPDE with a finite difference scheme. Furthermore, the overlap integrals that determine the spatial coupling between the optical modes can be computed off-line and then appear as constants in the system of DDEs. The combination of these two properties resulted in a substantial gain in the numerical efficiency of the model, so that a multiparameter bifurcation analysis of the mode-expanded DDE model with specialized software tools is now feasible.

With this approach we studied the case of a VCSEL that supports the two (rotationally symmetric) fundamental optical modes LP_{01} and LP_{02}, which are coupled via the external feedback. A bifurcation analysis of the resulting mode-expanded 18-dimensional DDE model, describing the spatial dynamics in terms of the radial mode profile of the VCSEL, was performed with the numerical continuation tools `DDE-BIFTOOL` and `PDDE-CONT`. A comparison with a finite difference discretization scheme [19] confirmed the accuracy of this model on the level of one-parameter bifurcation diagrams of the steady-state external cavity modes. We then computed the stability of the ECMs in dependence on the feedback strength κ and the self- versus cross-coupling parameter η. This revealed regions of coexisting, stable ECMs in the (κ, η)-plane, which are bounded by curves of saddle-node and Hopf bifurcations. The bifurcating periodic solutions may be in-phase relaxation oscillations (ROs) or antiphase external round trip oscillations (EOs). Both were shown to occur stably in large regions of the (κ, η)-plane. We finally considered how these periodic solutions may bifurcate to more complicated dynamics. In particular, this revealed that the VCSEL with optical feedback may display chaotic spatiotemporal dynamics, which is of potential interest for chaotic communication schemes [42].

Overall, our bifurcation analysis clearly shows the importance of the self- versus cross-coupling parameter η for the dynamics of the spatial modes. While the case of pure self-feedback is degenerate, for an intermediate level of cross-feedback in the range $\eta \in [0.4, 0.9]$ we identified a typical bifurcation scenario when the feedback strength κ is increased: additional ECMs are born in saddle-node bifurcations; stable ECMs lose their stability in Hopf bifurcations, in which stable in-phase ROs are born; these oscillations then give rise to two-frequency dynamics on an invariant torus, where the new modulation frequency is due to the external round trip of the light. In the range $\eta < 0.4$, that is, for larger levels of cross-feedback, we find stable antiphase EOs where the two fields suppress each other. When κ is increased, the antiphase EOs destabilize in a cascade of period-doublings that eventually results in antiphase spatiotemporal chaos. Physically, one sees a chaotic switching between the two modes. The clear distinction between in-phase ROs and antiphase EOs may provide a means of determining the level of cross-feedback in an experiment. Furthermore, the challenge is to record spatially resolved measurements of VCSEL modes on a picosecond timescale; this may be possible with the techniques developed in [1, 2].

There are several directions for future research. First, the following question arises: What happens when the VCSEL supports more than the two fundamental modes LP_{01} and LP_{02}? The mode-expanded DDE model can be used to describe the interaction between any number of transverse optical modes, but it becomes increasingly involved. The next physically relevant case arises when the first higher-order mode LP_{11} is involved in the dynamics as well. Already in this situation, which still concerns a VCSEL with quite a small aperture, the dynamics can no longer be described solely by the radial mode profile. This effectively doubles the dimension

of the mode-expanded DDE model and is presently under investigation. Second, as soon as higher-order transverse modes are involved, one may also wish to consider the interaction between the spatial modes and the polarization. Again, this doubles the dimension of the model.

Finally, the mode expansion presented here can also be used to model other systems that combine spatial extent with delays. Obvious examples are VCSELs with filtered or phase-conjugate optical feedback or mutual delay-coupling and broad area lasers subject to optical feedback. Application areas outside laser dynamics in which eigenmode expansion might be used include the spatial control of fluid patterns [45, 52] and hybrid (hardware-in-the-loop) mechanical engineering tests involving beams and plates [32].

Appendix A. The dimensionless rate equations (2.1)–(2.2) are derived from the dimensional model given in [59] by means of the following scalings. The dimensionless carrier-density $N(r,\phi,t)$, complex electric fields $E^j_{mn}(t)$, time t, and space variable r are given as

$$N(r,\phi,t) = \frac{1}{2}Av_g\Gamma\tau_p(Z(r,\phi,t) - N_0) - \frac{1}{2},$$
$$E^j_{mn}(t) = \left(\frac{A\tau_n}{2}a_{mn}\right)^{1/2} Y^j_{mn}(t),$$
$$t = t'/\tau_p, \quad r = r'/R.$$

Here $a_{mn} = v_g\Gamma\left(\int_0^{2\pi}\int_0^{\infty}|\psi_{mn}(r,\phi)|^2 r dr d\phi\right)^{-1}$, and $Z(r,\phi,t)$, $Y^j_{mn}(t)$, t', and r' represent the respective dimensional quantities as used in [59]. The dimensionless parameters are thus found as

$$T = \frac{\tau_n}{\tau_p}, \quad d_f = \frac{\tau_n D}{R^2},$$
$$P(r,\phi,t) = \frac{Av_g\Gamma\tau_p\tau_n}{2}\left(\frac{j(r,\phi,t)}{ed} - \frac{\left(N_0 + \frac{1}{Av_g\Gamma\tau_p}\right)}{\tau_n}\right).$$

We refer the reader to [59] for the description of the physical constants used above.

Acknowledgment. We thank Robert Szalai for help with `PDDE-CONT`.

REFERENCES

[1] A. Barchanski, T. Genstry, C. Degen, I. Fischer, and W. Elsässer, *Picosecond emission dynamics of vertical-cavity surface-emitting lasers: Spatial, spectral, and polarization-resolved characterization*, IEEE Quantum Elec., 39 (2003), pp. 850–858.

[2] K. Becker, I. Fischer, and W. Elsässer, *Spatio-temporal emission dynamics of VCSELs: Modal competition in the turn-on behavior*, in Semiconductor Lasers and Laser Dynamics, Proc. SPIE 5452, SPIE, Bellingham, WA, 2004, pp. 452–461.

[3] M. Brambilla, M. Cattaneo, L. A. Lugiato, R. Pirovano, F. Prati, A. J. Kent, G.-L. Oppo, A. B. Coates, C. O. Weiss, C. Green, E. J. D'Angelo, and J. R. Tredicce, *Dynamical transverse laser patterns. I. Theory*, Phys. Rev. A, 49 (1994), pp. 1427–1451.

[4] D. Breda, S. Maset and R. Vermiglio, *Computing the characteristic roots for delay differential equations*, IMA J. Numer. Anal., 24 (2004), pp. 1–19.

[5] W. F. Brinkmann and D. V. Lang, *Physics and the communications industry*, Rev. Modern Phys., 71 (1999), pp. 480–488.

[6] C. Degen, B. Krauskopf, G. Jennemann, I. Fischer, and W. Elsässer, *Polarization selective symmetry breaking in the near-fields of vertical cavity surface emitting lasers*, J. Opt. B Quantum Semiclass. Opt., 2 (2000), pp. 517–525.

[7] J. Dellunde, A. Valle, and K. A. Shore, *Transverse-mode selection in external-cavity vertical-cavity surface-emitting laser diodes*, J. Opt. Soc. Amer. B Opt. Phys., 13 (1996), pp. 2477–2483.

[8] O. Diekmann, S. A. van Gils, S. M. Verduyn Lunel, and H.-O. Walther, *Delay Equations: Functional-, Complex-, and Nonlinear Analysis*, Springer-Verlag, New York, 1995.

[9] E. J. Doedel, R. Paffenroth, A. R. Champneys, T. F. Fairgrieve, Y. A. Kuznetsov, B. E. Oldemann, B. Sandstede, and X. Wang, *AUTO 2000: Continuation and Bifurcation Software for Ordinary Differential Equations (with HomCont)*, http://cmvl.cs.concordia.ca/auto/ (2000).

[10] A. K. Jansen van Doorn, M. P. van Exter, and J. P. Woerdman, *Effects of transverse anisotropy on VCSEL spectra*, Electron. Lett., 30 (1994), pp. 1941–1943.

[11] K. Engelborghs, T. Luzyanina, G. Samaey, D. Roose, and K. Verheyden, *DDE-BIFTOOL v. 2.03: A MATLAB Package for Bifurcation Analysis of Delay Differential Equations*, Department of Computer Science, K.U. Leuven, Leuven, Belgium, 2001, http://ftp.cs.kuleuven.ac.be/~twr/research/software/delay/ddebiftool.shtml.

[12] K. Engelborghs and D. Roose, *On stability of LMS methods and characteristic roots of delay differential equations*, SIAM J. Numer. Anal., 40 (2002), pp. 629–650.

[13] H. Erzgräber, B. Krauskopf, and D. Lenstra, *Bifurcation analysis of a semiconductor laser with filtered optical feedback*, SIAM J. Appl. Dyn. Syst., 6 (2007), pp. 1–28.

[14] H. Erzgräber, D. Lenstra, B. Krauskopf, A. P. A. Fischer, and G. Vemuri, *Feedback Phase Sensitivity of a Semiconductor Laser Subject to Filtered Optical Feedback: Experiment and Theory*, http://hdl.handle.net/1983/916 (2007).

[15] I. Fischer, T. Heil, and W. Elsässer, *Emission dynamics of semiconductor lasers subject to delayed optical feedback: An experimentalist's perspective*, in Fundamental Issues of Nonlinear Lasers Dynamics, B. Krauskopf and D. Lenstra, eds., American Institute of Physics, College Park, MD, 2000, pp. 218–237.

[16] G. E. Giudice, D. V. Kuksenkov, and H. Temkin, *Measurement of differential carrier lifetime in vertical-cavity surface-emitting lasers*, IEEE Photon. Tech. Lett., 10 (1998), pp. 920–922.

[17] K. Green, B. Krauskopf, and G. Samaey, *A two-parameter study of the locking region of a semiconductor laser subject to phase-conjugate feedback*, SIAM J. Appl. Dyn. Syst., 2 (2003), pp. 254–276.

[18] K. Green and B. Krauskopf, *Bifurcation analysis of a semiconductor laser subject to non-instantaneous phase-conjugate feedback*, Opt. Commun., 231 (2004), pp. 383–393.

[19] K. Green, B. Krauskopf, and D. Lenstra, *External cavity mode structure of a two-mode VCSEL subject to optical feedback*, Opt. Commun., 277 (2007), pp. 359–371.

[20] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, New York, 1983.

[21] B. Haegeman, K. Engelborghs, D. Roose, D. Pieroux, and T. Erneux, *Stability and rupture of bifurcation bridges in semiconductor lasers subject to optical feedback*, Phys. Rev. E, 66 (2002), 046216.

[22] J. Hale and S. M. Verduyn Lunel, *Introduction to Functional Differential Equations*, Springer-Verlag, New York, 1993.

[23] R. N. Hall, G. E. Fenner, J. D. Kingsley, T. J. Soltys, and R. O. Carlson, *Coherent light emission from GaAs junctions*, Phys. Rev. Lett., 9 (1962), pp. 366–369.

[24] T. Heil, I. Fischer, W. Elsässer, and A. Gavrielides, *Dynamics of semiconductor lasers subject to delayed optical feedback: The short cavity regime*, Phys. Rev. Lett., 87 (2001), 243901.

[25] T. Heil, I. Fischer, W. Elsässer, B. Krauskopf, K. Green, and A. Gavrielides, *Delay dynamics of semiconductor lasers with short external cavities: Bifurcation scenarios and mechanisms*, Phys. Rev. E, 67 (2003), 066214.

[26] M. X. Jungo, D. Erni, and W. Bächtold, *VISTAS: A comprehensive system-oriented spatiotemporal VCSEL model*, IEEE Sel. Topics in Quantum Elec., 9 (2003), pp. 939–948.

[27] D. M. Kane and K. A. Shore, eds., *Unlocking Dynamical Diversity: Optical Feedback Effects on Semiconductor Lasers*, Wiley, New York, 2005.

[28] H. Kawaguchi, *Bistabilities and Nonlinearities in Laser Diodes*, Artech House, Boston, 1994.
[29] B. Krauskopf, *Bifurcation analysis of lasers with delay*, in Unlocking Dynamical Diversity: Optical Feedback Effects on Semiconductor Lasers, D. M. Kane and K. A. Shore, eds., Wiley, New York, 2005, pp. 147–183.
[30] B. Krauskopf and D. Lenstra, *Fundamental Issues of Nonlinear Lasers Dynamics*, American Institute of Physics, College Park, MD, 2000.
[31] B. Krauskopf, G. H. M. Van Tartwijk, and G. R. Gray, *Symmetry properties of lasers subject to optical feedback*, Opt. Commun., 177 (2000), pp. 347–353.
[32] Y. N. Kyrychko, S. J. Hogan, A. Gonzalez-Buelga, and D. J. Wagg, *Modelling real-time dynamic substructuring using partial delay differential equations*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 463 (2007), pp. 1509–1523.
[33] R. Lang and K. Kobayashi, *External optical feedback effects on semiconductor injection laser properties*, IEEE Quantum Elec., 16 (1980), pp. 347–355.
[34] J. Y. Law and G. P. Agrawal, *Effects of optical feedback on static and dynamic characteristics of vertical-cavity surface-emitting lasers*, IEEE Sel. Topics in Quantum Elec., 3 (1997), pp. 353–358.
[35] J. Y. Law, G. H. M. van Tartwijk, and G. P. Agrawal, *Effects of transverse-mode competition on the injection dynamics of vertical-cavity surface-emitting lasers*, Quantum Semiclass. Opt., 9 (1997), pp. 737–747.
[36] J. Y. Law and G. P. Agrawal, *Feedback-induced chaos and intensity-noise enhancement in vertical-cavity surface-emitting lasers*, J. Opt. Soc. Amer. B Opt. Phys., 15 (1998), pp. 562–569.
[37] D. Lenstra, B. H. Verbeek, and A. J. den Boef, *Coherence collapse in single-mode semiconductor lasers due to optical feedback*, IEEE J. Quantum Electron., 21 (1985), pp. 674–679.
[38] F. Marino, S. Barland, and S. Balle, *Single-mode operation and transverse-mode control in VCSELs induced by frequency-selective feedback*, IEEE Photon. Tech. Lett., 15 (2003), pp. 789–791.
[39] J. Martin-Regalado, S. Balle, M. San Miguel, A. Valle, and L. Pesquera, *Polarization and transverse-mode selection in quantum-well vertical-cavity surface-emitting lasers: Index- and gain-guided devices*, Quantum Semiclass. Opt., 9 (1997), pp. 713–736.
[40] J.-F. Mercier and J. V. Moloney, *Derivation of semiconductor laser mean-field and Swift-Hohenberg equations*, Phys. Rev. E, 66 (2002), 036221.
[41] K. Moriki, H. Nakahara, T. Hattori, and K. Iga, *Single transverse mode condition of surface-emitting injection lasers*, Elec. Commun. Japan, 71 (1988), pp. 81–90.
[42] J. García-Ojalvo and R. Roy, *Spatiotemporal communication with synchronized optical chaos*, Phys. Rev. Lett., 86 (2001), pp. 5204–5207.
[43] Y. Onishi, N. Nishiyama, C. Caneau, F. Koyama, and C.-E. Zah, *All-optical inverter based on long-wavelength vertical-cavity surface-emitting laser*, IEEE J. Sel. Topics Quantum Electron., 11 (2005), pp. 999–1005.
[44] Y. Onishi, N. Nishiyama, C. Caneau, F. Koyama, and C.-E. Zah, *All-optical regeneration using transverse mode switching in long-wavelength vertical-cavity surface-emitting lasers*, Jpn. J. Appl. Phys., 45 (2006), pp. 467–468.
[45] C. M. Postlethwaite and M. Silber, *Spatial and temporal feedback control of traveling wave solutions of the two-dimensional complex Ginzburg-Landau equation*, Phys. D, 236 (2007), pp. 65–74.
[46] F. Prati, M. Travagnin, and L. A. Lugiato, *Logic gates and optical switching with vertical-cavity surface-emitting lasers*, Phys. Rev. A, 55 (1997), pp. 690–700.
[47] D. Roose and R. Szalai, *Continuation and bifurcation analysis of delay differential equations*, in Numerical Continuation Methods for Dynamical Systems: Path Following and Boundary Value Problems, B. Krauskopf, H. M. Osinga, and J. Galán-Vioque, eds., Springer-Verlag, New York, 2007, pp. 359–399.
[48] M. Sciamanna, K. Panajotov, H. Thienpont, I. Veretennicoff, P. Mégret, and M. Blondel, *Optical feedback induces polarization mode hopping in vertical-cavity surface-emitting lasers*, Opt. Lett., 28 (2003), pp. 1543–1545.
[49] M. Sciamanna, T. Erneux, F. Rogister, O. Deparis, P. Mégret, and M. Blondel, *Bifurcation bridges between external-cavity modes lead to polarization self-modulation in vertical-cavity surface-emitting lasers*, Phys. Rev. A, 65 (2004), 041801(R).
[50] G. Stepán, *Retarded Dynamical Systems: Stability and Characteristic Functions*, Longman Scientific and Technical, London, 1989.

[51] R. Szalai, *PDDE-CONT: A Continuation and Bifurcation Software for Delay-Differential Equations*, Budapest University of Technology and Economics, Budapest, Hungary, 2005, http://www.mm.bme.hu/~szalai/pdde.

[52] C. M. Topaz, J. Porter, and M. Silber, *Multifrequency control of Faraday wave patterns*, Phys. Rev. E, 70 (2004), 066206.

[53] M. S. Torre, C. Masoller, and P. Mandel, *Transverse-mode dynamics in vertical-cavity surface-emitting lasers with optical feedback*, Phys. Rev. A, 66 (2002), 053817.

[54] M. S. Torre, C. Masoller, and K. A. Shore, *Synchronization of unidirectionally coupled multi-transverse-mode vertical-cavity surface-emitting lasers*, J. Opt. Soc. Amer. B Opt. Phys., 21 (2004), pp. 1772–1780.

[55] M. S. Torre, C. Masoller, P. Mandel, and K. A. Shore, *Transverse-mode dynamics in directly modulated vertical-cavity surface-emitting lasers with optical feedback*, IEEE J. Quantum Electron., 40 (2004), pp. 620–628.

[56] B. Tromborg, J. H. Osmundsen, and H. Olesen, *Stability analysis for a semiconductor laser in an external cavity*, IEEE J. Quantum Electron., 20 (1984), pp. 1023–1032.

[57] A. Valle, J. Sarma, and K. A. Shore, *Dynamics of transverse mode competition in vertical cavity surface emitting semiconductor lasers*, Opt. Commun., 115 (1995), pp. 297–302.

[58] A. Valle, J. Sarma, and K. A. Shore, *Spatial hole-burning effects on the dynamics of vertical cavity surface emitting semiconductor lasers*, IEEE J. Quantum Electron., 31 (1995), pp. 1423–1431.

[59] A. Valle, *Selection and modulation of high-order transverse modes in vertical-cavity surface-emitting lasers*, IEEE J. Quantum Electron., 34 (1998), pp. 1924–1932.

[60] K. Verheyden, T. Luzyanina, and D. Roose, *Efficient computation of characteristic roots of delay differential equations using LMS methods*, J. Comput. Appl. Math., 214 (2008), pp. 209–226.

[61] S. F. Yu, *Analysis and Design of Vertical Cavity Surface Emitting Lasers*, Wiley, New York, 2003.

SIAM J. Applied Dynamical Systems
Vol. 8, No. 1, pp. 253–278
© 2009 Society for Industrial and Applied Mathematics

Canards, Clusters, and Synchronization in a Weakly Coupled Interneuron Model*

Bard Ermentrout† and Martin Wechselberger‡

Abstract. Applying a low current to a recent biophysical model for inhibitory neurons in the cortex leads to mixed mode oscillations (MMOs), a mixture of spikes and subthreshold oscillations. At higher currents, the neurons fire regularly. We show that a specific slow potassium current underlies this behavior. Next we reduce this five-dimensional biophysical model for an inhibitory neuron to three dimensions with a slow/fast time-scale structure. We then show that there is a range of parameters under which the reduced model shows MMOs which can be explained by the canard phenomenon. Many inhibitory interneurons are coupled with electrical tight junctions (gap junctions). We show that such coupling combined with the underlying MMOs produces clustered solutions in large networks. We explain this using the theory of weakly coupled oscillators and show that the complex dynamics arises from sensitivity near the canard.

Key words. canards, synchrony, weak coupling, clusters, neurons

AMS subject classifications. 34C15, 34C26, 37G15, 92C20

DOI. 10.1137/080724010

1. Introduction.

Electrical coupling via gap junctions is widespread in networks of cortical inhibitory neurons [18, 38]. The effects of such electrical coupling have been the focus of much experimental and theoretical work [18, 38, 9, 41, 37, 15, 33]. Most experimental and theoretical findings show that electrical coupling can help to coordinate synchronous oscillatory behavior in inhibitory networks. Many of the studies of gap junction coupling and oscillations concern pairs of cells (e.g., [38, 37, 17, 41]) rather than larger networks. Synchrony is not the only possibility when coupling via gap junctions and both the shape of spikes [9] and the intrinsic dynamics [41] can destabilize synchronization. Combinations of gap junctions with inhibitory coupling can lead to more complex behavior such as waves [26] and clusters [40, 26] or can enhance the ability to synchronize [37, 33]. Clustered behavior in networks of oscillators often arises when there is multistability in pairwise interactions (although this is not a requirement), and there have been numerous studies of such behavior in oscillator networks [26, 28, 19, 1, 8, 3].

In this paper, we explore the relationship between intrinsic dynamics, coupling, and patterns of synchrony for a recently proposed interneuron model [13] endowed with a slow potassium channel. We will focus on the formation of clusters and synchronization in a network of electrically coupled interneurons. We first show, via numerical simulations, that at low currents and with weak coupling, these networks organize themselves into clustered states

*Received by the editors May 12, 2008; accepted for publication (in revised form) by T. Kaper October 30, 2008; published electronically February 13, 2009.
http://www.siam.org/journals/siads/8-1/72401.html
†Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260 (bard@math.pitt.edu).
‡School of Mathematics & Statistics, Centre of Mathematical Biology, University of Sydney, NSW, Australia (wm@maths.usyd.edu.au).

that persist even in the presence of a small amount of noise. At higher currents, the networks become synchronous. In order to understand the mechanism by which there is a transition from clustering to synchrony, we first examine the single cell model. We find that at low currents, there are so-called mixed mode oscillations (MMOs) [6], an alternation of subthreshold oscillations (STOs) with spiking activity. Thus, we first turn our attention to the analysis of this behavior. By reducing the system of five differential equations to one that has three with a slow/fast time-scale structure, we are able to show that the underlying dynamics is due to the presence of a canard structure [4, 46, 50, 7, 51, 10, 24, 22, 5, 39]. Such canard induced MMOs have been identified in other neuronal models as well [12, 43, 44, 11, 23, 45, 34, 42]. We want to point out that Milik et al. [39] were the first who studied the relationship between canards and MMOs. Next, we use weak coupling theory to analyze the weakly coupled network. The connection between the canards and the clustering is through the complex response of individual neural oscillators to perturbations (the adjoint of the linearized dynamics), which makes the weak interaction between oscillators very complicated. As we discuss later in the paper, the key requirement for the formation of two or more clusters is that the phase-interaction function be dominated by higher order odd Fourier modes [36]. The underlying MMOs cause the emergence of such modes and thus the appearance of robust clusters.

The paper is organized as follows: In section 2 we present the Erisir model and show numerical observations of MMOs, clustering, and synchronization. In section 3 we perform a single cell analysis to explain MMOs via the canard phenomenon. In section 4 we use weak coupling theory to show that the canard structure is responsible for clustering in an inhibitory network. Finally, we conclude in section 5.

2. The Erisir model.

2.1. Single cell. Inhibitory interneurons (fast-spiking (FS) cells) play an important role in the dynamics of the cerebral cortex, particularly in preventing run away excitation. One of the features of these neurons is their ability to produce fast high-frequency spikes. There have been many attempts to model these neurons; one of the best known is the Wang–Buszaki (WB) model [49] which has just a transient sodium, leak, and delayed rectifier potassium channel. Experiments over the last ten years have shown the importance of a class of special potassium channels (Kv3.1 and Kv3.2) enabling the FS cells to maintain a high frequency of firing. Erisir et al. [13] used voltage clamp experiments to create a new model for the fast-spike cells that incorporates the new results on this potassium channel. As this model also has some interesting dynamical properties and has begun to be used in several other computational studies [37, 38], we will use this model in this paper.

The single cell Erisir model [13] is a conductance-based model of an FS cortical interneuron. (We remark that in the original Erisir reference, there are typographical errors in the description of the equations. We are using the version of the model in [38].) The ionic currents of the Erisir model consist of a fast sodium current I_{Na}, two potassium currents, a "moderate" delayed rectifier I_K associated with a Kv3.1 ion channel and a "slow" delayed rectifier I_{Ks} associated with a Kv1.3 ion channel, and a passive leak current I_L. The model equations are

$$\begin{aligned} C\,V' &= I_{app} - (I_{Na} + I_K + I_{Ks} + I_L), \\ x' &= (x_\infty(V) - x)/\tau_x(V), \qquad x = m, h, n, s, \end{aligned} \tag{2.1}$$

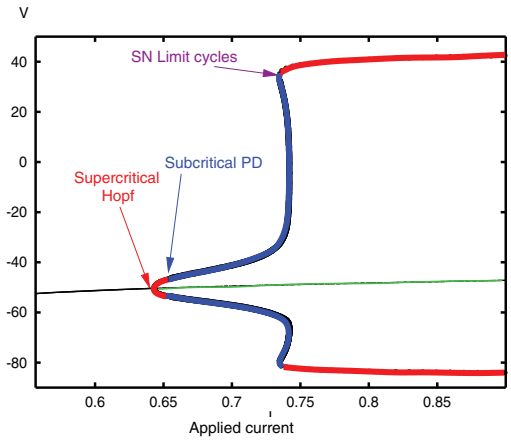


Figure 1. *Bifurcation diagram of Erisir model (2.1): Bifurcation parameter is the applied current I_{app}; supercritical Hopf bifurcation for $I_{app} \approx 0.64$; for $I_{app} \approx 0.65$ the limit cycle loses stability at a subcritical period-doubling bifurcation; for $I_{app} \approx 0.73$ saddle-node of limit cycles occurs. Stable periodic orbits are red, and unstable are blue; stable fixed points are black, and unstable are green.*

where C is the capacitance of the FS interneuron, I_{app} is an applied (external) current, and

$$\begin{aligned} I_{Na} &= g_{Na}\, m^3 h (V - E_{Na}), \\ I_K &= g_K\, n^2 (V - E_K), \\ I_{Ks} &= g_{Ks}\, s^4 (V - E_K), \\ I_L &= g_L\, (V - E_L) \end{aligned} \tag{2.2}$$

are the ionic currents, where m is the activation gate and h is the inactivation gate of the sodium current, n is the activation gate of the I_K potassium current, and s is the activation gate of the I_{Ks} potassium current. The dynamics of each gate $x = m, h, n, s$ is toward the steady state voltage dependent sigmoidal function $x_\infty(V)$ with the characteristic voltage dependent bell-shaped time constant $\tau_x(V)$. The parameters g_y $(y = Na, K, Ks, L)$ and E_z $(z = Na, K, L)$ are conductances and Nernst potentials of the various ionic currents. All parameters and functions are defined in Appendix A.

Figure 1 shows the bifurcation diagram of the single cell Erisir model (2.1) where the bifurcation parameter is the applied current I_{app} (in pA/cm^2). The cell is excitable for $I_{app} < I_{Hopf}$, where $I_{Hopf} \approx 0.64$ denotes a supercritical Hopf bifurcation (HB). The corresponding limit cycle loses stability at $I_{app} \approx 0.65$ due to a subcritical period-doubling bifurcation which opens a parameter window up to $I_{app} = I_{snlc} \approx 0.73$ with apparently no stable simple attractors. I_{snlc} denotes a saddle-node bifurcation of limit cycles. For $I_{app} > I_{snlc}$ there exist large relaxation-type oscillations (ROs).

What are the attractors, if any exist, in this parameter window between small amplitude STOs and large amplitude ROs? Figure 2 shows time traces for several values of applied current I_{app}. STO patterns lose stability for $I_{app} \approx 0.654$, and MMO patterns show up as stable solutions for the Erisir model filling the gap of stable solutions in Figure 1. The simplest periodic MMO patterns are of the form L^s, where L denotes the number of ROs

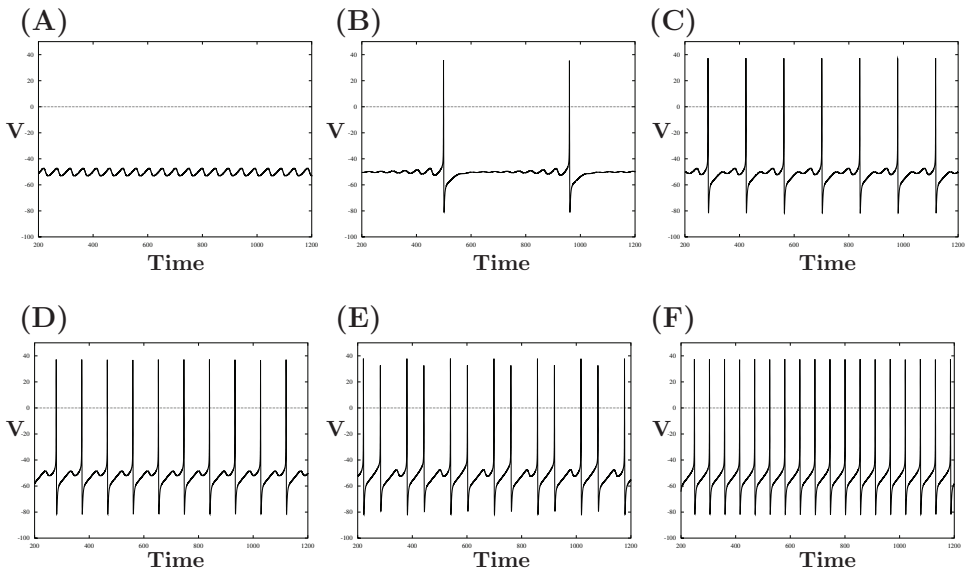


Figure 2. *Voltage time traces of Erisir model* (2.1). (A): *small oscillatory pattern* 0^1 *for* $I_{app} = 0.65$*;* (B): *MMO pattern* 1^9 *for* $I_{app} = 0.66$*;* (C): *MMO pattern* 1^2 *for* $I_{app} = 0.70$*;* (D): *MMO pattern* 1^1 *for* $I_{app} = 0.72$*;* (E): *MMO pattern* 2^1 *for* $I_{app} = 0.73$*;* (F): *relaxation oscillation pattern* 1^0 *for* $I_{app} = 0.74$.

and s denotes the number of STOs. For example, Figure 2 shows MMO patterns of the form 1^9, 1^2, 1^1, and 2^1. STOs can be viewed as a 0^1 MMO pattern, while ROs can be viewed as a 1^0 MMO pattern. Under the variation of the bifurcation parameter I_{app} one observes the following sequence of simple MMO patterns:

$$0^1 \to 1^{s_*} \to \cdots \to 1^1 \to 2^1 \to \cdots \to L_*^1 \to 1^0,$$

where s_* can be an arbitrary large integer and L_* is a finite upper bound. More complicated MMO patterns are also possible and can usually be found in the transition from one to another simple MMO pattern. In section 3 we will show that these MMO patterns can be explained by the canard phenomenon [7, 50].

Before turning to the coupling between such cells and our analysis of the canard phenomenon, we remark that there are several other models for inhibitory interneurons, e.g., the WB model [49], which do not seem to have any MMOs. Notably absent in the WB model is the g_{Ks} channel. If we set this current to zero in the Erisir model, the bifurcation from rest to repetitive oscillations is considerably simpler and resembles the sequence in Izhikevich [31, Figure 6.40, page 197]. The recent model in [20] is like the Erisir model but without the slow potassium current and instead an A-type potassium current. This model, too, has no MMOs, at least in the parameters used by the authors. Figure 3A shows the corresponding diagram to Figure 1 with the norm of the solution plotted for clarity. The HB is subcritical and terminates on a homoclinic. The large amplitude limit cycle (red) is lost via a collision with an unstable limit cycle (blue) (saddle-node of limit cycles (SNLC)). The unstable limit cycle is lost at a homoclinic. Between the SNLC and the HB, the system is bistable.

Figure 3B shows the two-parameter bifurcation diagram. Note that there is only one

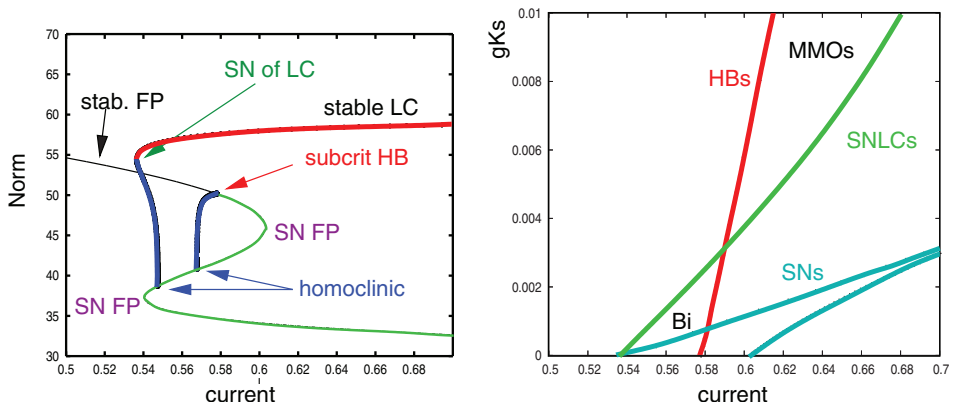


Figure 3. (A) *Bifurcation diagram of Erisir model (2.1) when $g_{Ks} = 0$ as current varies. Instead of membrane potential, the L_2 norm of the solution is plotted as it produces a clearer picture. Colored curves are as in Figure 1.* (B) *Two-parameter diagram showing the branch of Hopf bifurcations (HBs, red), saddle-nodes of fixed points (cyan), and saddle-nodes of limit cycles (SNLCs, green). g_{Ks} for Figure 1 is 0.018, which is out of the picture. In the region labeled "Bi" between the HBs and the SNLCs, there is bistability between fixed points and large amplitude limit cycles. Mixed mode oscillations are found in the region labeled "MMOs."*

fixed point for g_{Ks} larger than about 0.004. At $g_{Ks} \approx 0.0035$, the curve of SNLCs (shown in green) crosses to the right of the curve of HBs. In this region there are neither stable fixed points nor large amplitude limit cycles. For small ranges of the applied current and for g_{Ks} larger than about 0.0045, there are stable small amplitude limit cycles, but the region is primarily dominated by MMOs. Indeed, the absence of a large amplitude limit cycle and a stable fixed point is a necessary condition for MMOs. The Hopf bifurcation occurring on the lower branch has a somewhat complex structure. For g_{Ks} larger than about 0.0095, the bifurcation is supercritical, and thus small amplitude oscillations stably bifurcate from rest. However, in the interval $0.0045 < g_{Ks} < 0.0095$, while the bifurcation is subcritical, the small amplitude unstable branch turns around and gives birth to an interval of stable small amplitude oscillations.

Thus, in this section, we have shown that the slow potassium current appears to be crucial to get MMOs. In the next sections, we see some of the consequences of the inclusion of the slow potassium current for weakly gap-junctionally coupled cells and go on to show that the mechanism of MMOs is a generalized canard phenomenon.

2.2. Clustering and coupling. Cortical interneurons are usually coupled via gap junctions and GABAergic (inhibitory) synapses. To get a better understanding of clustering and synchrony properties of networks of interneurons we focus in this work on networks with gap junctions only, as the issues of time-scales and reversal potential no longer play a role [32, 48, 52].

We study a network of cortical interneurons by using the Erisir model (2.1) with all-to-all coupling via gap junctions:

$$
\begin{aligned}
C\,V_i' &= I_{app} - (I_{Na} + I_K + I_{Ks} + I_L + I_{gap,i}), && i = 1, \ldots, N, \\
x_i' &= (x_{i,\infty}(V_i) - x_i)/\tau_{x_i}(V_i), && x_i = m_i, h_i, n_i, s_i,
\end{aligned} \tag{2.3}
$$

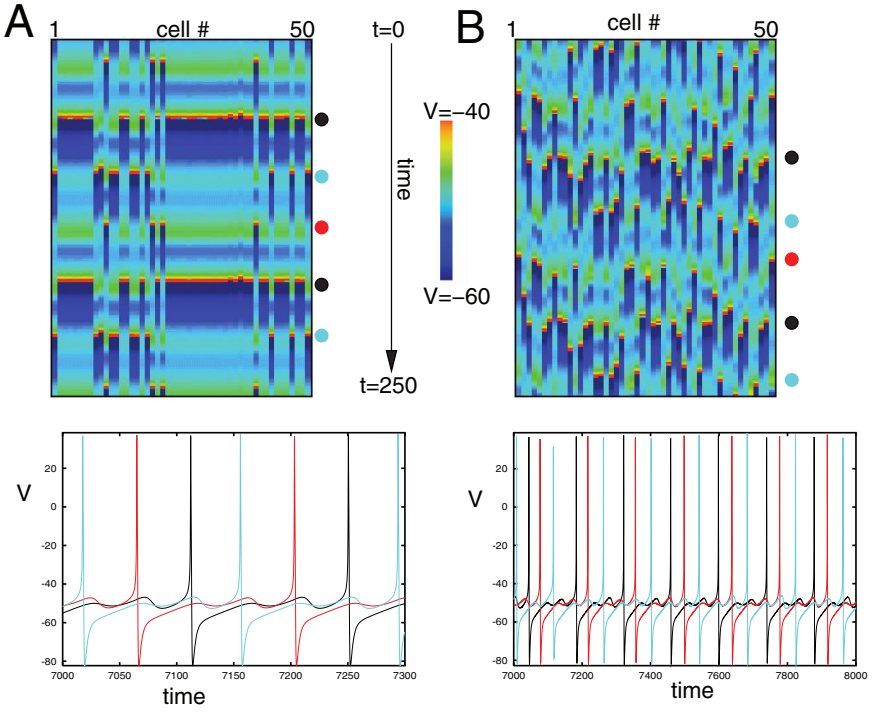


Figure 4. *Simulation of* 50 *globally gap-junctionally coupled cells with* $I_{app} = 0.7$, $g_{Ks} = 0.018$, $g_{gap} = 0.0002$ *with no noise* (A) *and with* $\sigma = 0.002$ *amount of Gaussian noise* (B). *Top figures show the membrane potential of all* 50 *cells over a short time span (time goes down). Filled circles show different groups of cells forming clusters that are nearly synchronized. Lower figures show the voltage traces for representative cells in each cluster. This structure is robust to noise.*

where

$$I_{gap,i} = g_{gap}\frac{1}{N}\sum_{j=1}^{N}(V_i - V_j)$$

is the gap junction current cell i receives from all its $(N-1)$ neighbors j. All the other intrinsic ionic currents depend only on cell i itself. Additionally, to check for robustness, in some simulations, we add a small amount of independent Gaussian noise.

Figure 4 shows the result of a simulation of 50 cells coupled with $g_{gap} = 0.0002$ and $I_{app} = 0.7$. The cells organize themselves into three clusters of differing sizes. There is a range of sizes for the clusters that depends on the initial data (not shown here) and which can be understood in the case of weak coupling by analyzing the associated phase models [28]. In general, clusters which are roughly equal in size have the largest basin of attraction. This is seen in Figure 4B, where a small amount of noise is added. There the cells break into three groups which are much more balanced in size, but whose phases vary around means that are roughly a third of a cycle apart.

Figure 5 shows that STOs are necessary for the clustering to occur. In this figure, $I_{app} = 0.675$, and in the normal model there are three clusters as in Figure 4. However, if we remove

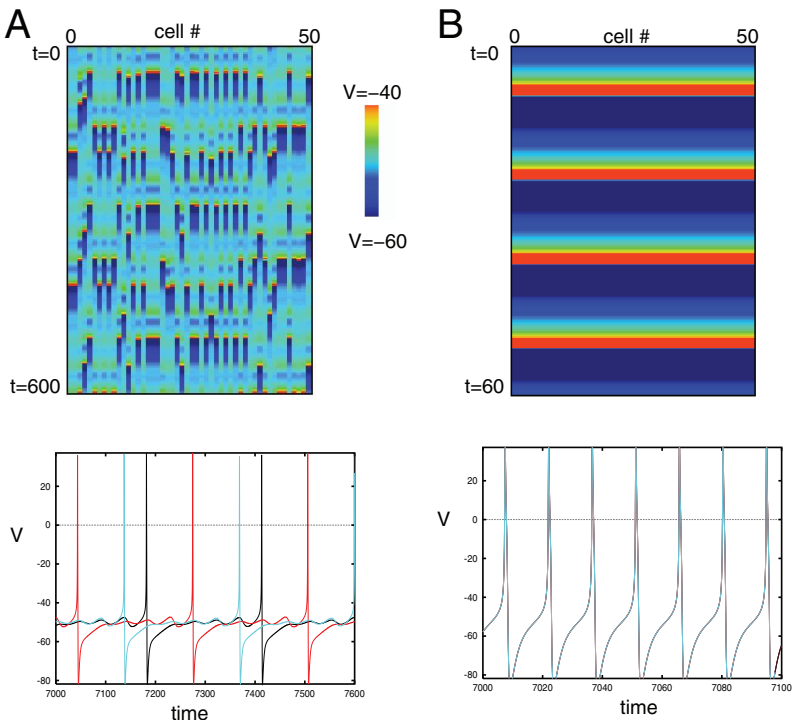


Figure 5. *Clustering depends on the g_{Ks} and the subthreshold oscillations.* (A) $g_{Ks} = 0.018$, $I_{app} = 0.675$, $g_{gap} = 0.0002$. *As in Figure 4, the top shows voltage traces of all the cells, and the bottom shows representative voltages for cells in the clusters.* (B) $g_{Ks} = 0$ *shows synchronization.*

g_{Ks}, the slow potassium channel, then the neurons robustly synchronize. Thus, the presence of the slow potassium current is needed for clustering to occur. If we maintain the slow potassium current, but increase the applied current into the regime where there is regular spiking (e.g., the red curves in Figure 1, say, $I_{app} = 1$), then the clustering disappears. If the coupling is strong enough, then clustering disappears, leaving only synchrony. Thus, it seems that the clusters are a consequence of the interaction of subthreshold behavior with weak coupling. The remainder of the paper explores the mechanism for MMOs and how these impact the interactions for weak coupling.

3. Single cell analysis. In this section we want to uncover the mechanism that generates MMOs in the single cell Erisir model (2.1). In particular, we want to show that the generalized canard phenomenon [7, 50, 46, 39, 22, 51, 10] observed in slow/fast time-scale problems explains the STOs within MMO patterns. The current theory for MMOs via canards is developed for systems with one variable on the fast time-scale and two variables on the slow time-scale. Therefore, the analysis splits into the following parts: (a) identify two different time-scales in the Erisir model (section 3.1); (b) perform a model reduction of the full five-dimensional (5D) model to a three-dimensional (3D) model with one fast and two slow variables (section 3.2); (c) explain MMOs in the reduced problem via the generalized canard phenomenon (sections 3.3–3.4).

3.1. Dimensionless version of the Erisir model. A standard procedure to identify different (time-)scales in a biophysical model is to bring such a system into dimensionless form. Only in dimensionless form can one decide which parameters are small or large. We therefore rescale the dimensional variables V (dependent) and t (independent) with appropriate reference scales k_v and k_t:

$$V = k_v\, v, \qquad t = k_t\, \tau,$$

where v and τ are the dimensionless versions of membrane voltage and time. Note that all the gating variables $x = m, h, n, s$ are dimensionless by definition. After rescaling we obtain the following dimensionless version of the Erisir model (2.1):

$$\begin{aligned} C/(g_{Na}k_t)\, v' &= \hat{I}_{app} - (\hat{I}_{Na} + \hat{I}_K + \hat{I}_{Ks} + \hat{I}_L), \\ x' &= (k_t/\tau_x(v))(x_\infty(v) - x), \qquad x = m, h, n, s, \end{aligned} \tag{3.1}$$

with

$$\begin{aligned} \hat{I}_{Na} &= m^3 h(v - \hat{E}_{Na}), & \hat{I}_K &= \hat{g}_K\, n^2(v - \hat{E}_K), \\ \hat{I}_{Ks} &= \hat{g}_{KS}\, s^4(v - \hat{E}_K), & \hat{I}_L &= \hat{g}_L\,(v - \hat{E}_L), \\ \hat{I}_{app} &= I_{appl}/(g_{Na}\, k_v) \end{aligned} \tag{3.2}$$

and

$$\hat{g}_y = g_y/g_{Na}, \quad y = K, Ks, L, \qquad \hat{E}_z = E_z/k_v, \quad z = Na, K, L. \tag{3.3}$$

If we look at the time traces in Figure 2, then we see that the membrane potential varies in the order of $100\,mV$ (maximum around $120\,mV$) and we choose $k_v = 100\,mV$ as a typical reference voltage scale.[1] Therefore, the right-hand side of the voltage equation v in system (3.1) is basically of order one and the parameter $C/(g_{Na}) \approx 10^{-2}\,ms$ represents a typical time-scale of the voltage equation.

To identify the typical time-scales of the gates x, we have to look at the voltage dependent time-scale functions $\tau_x(v)$ which are shown in Figure 6, left. We observe that the maxima of all the bell-shaped functions are in the subthreshold regime (equivalent to $V \le -40\,mV$) except for the gate n, which has its maximum at $V \approx 0\,mV$. The bell-shaped curve indicates that typical time-scales vary depending on whether the cell is in the subthreshold regime ($V \le -40\,mV$) or in the action potential regime ($V \ge -40\,mV$). In the subthreshold regime, typical time-scale orders are

$$\tau_m \in (10^{-2}, 10^{-1}),\ \tau_h \in (10^0, 10^1),\ \tau_n \in (10^0, 10^1),\ \tau_s \in (10^1, 10^2),$$

while in the action potential regime, typical time-scale orders are

$$\tau_m \sim 10^{-2},\ \tau_h \sim 10^0,\ \tau_n \in (10^0, 10^1),\ \tau_s \in (10^0, 10^1).$$

[1] In *dimensional analysis*, typical time-scales are found as combinations of the model parameters. Since we have a plethora of choices in this model, we decided to make an empirical choice for easier comparison with the original model.

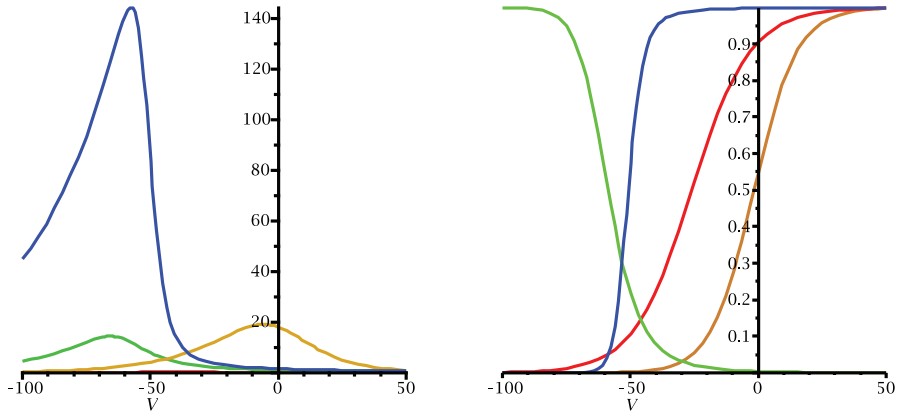


Figure 6. *Voltage dependent time constants $\tau_x(V)$ (left) and voltage dependent quasi steady state $x_\infty(V)$ (right) of Erisir model (2.1): $\tau_m(V)$ and $m_\infty(V)$ (red), $\tau_h(V)$ and $h_\infty(V)$ (green), $\tau_n(V)$ and $n_\infty(V)$ (yellow), and $\tau_s(V)$ and $s_\infty(V)$ (blue). Note that $V = 100 \cdot v$ relates the voltage V to its dimensionless counterpart v. Similarly, $\tau_x(V)$ relates to $\tau_x(v)$ and $x_\infty(V)$ to $x_\infty(v)$. (Note that $\tau_m(V)$ on the left is almost coincident with the axis and is thus difficult to see.)*

We have to choose the minimum order of each τ_x, $x = m, h, n, s$, to obtain the typical time-scale order of the gates. This suggests that m has a typical time-scale of order $10^{-2}\,ms$, while the other gates (h, n, s) are of order $1\,ms$. Altogether, we have identified two different time-scales in system (3.1): a fast time-scale of order $10^{-2}\,ms$ for the fast variables (v, m) and a slow time-scale of order $1\,ms$ for the slow variables (n, h, s). We choose $k_t = 1\,ms$ as the (slow) reference time-scale and we obtain

$$\begin{aligned} \varepsilon\, v' &= \hat{I}_{appl} - (\hat{I}_{Na} + \hat{I}_K + \hat{I}_{Ks} + \hat{I}_L), \\ \varepsilon\, m' &= (1/t_m(v))(m_\infty(v) - m), \\ h' &= (1/t_h(v))(h_\infty(v) - h), \\ n' &= (1/t_n(v))(n_\infty(v) - n), \\ s' &= (1/t_s(v))(s_\infty(v) - s), \end{aligned} \tag{3.4}$$

where $\varepsilon := C/(g_{Na}\, k_t) \approx 10^{-2} \ll 1$ is the singular perturbation parameter and the functions $t_x(v)$ are appropriate dimensionless versions of $\tau_x(v)$. System (3.4) is a slow/fast system with two fast variables (v, m) and three slow variables (h, n, s) on the slow time-scale τ ($' = d/d\tau$).

3.2. Reduction to a 3D model.

Step 1: Reduction of fast variables. In the literature it is a common procedure to reduce fast gating variables to their quasi steady state arguing that they reach these values very fast and the dynamics of these gates are of minor importance. In our case, we set the fast gating variable $m = m_\infty(v)$. This kind of reduction can be justified mathematically by a "center manifold reduction." We will not show this here but refer the reader to [44], which showed such a reduction step rigorously for the m gate of the Hodgkin–Huxley model. With that first reduction step, we reduced the number of fast variables from two to one.

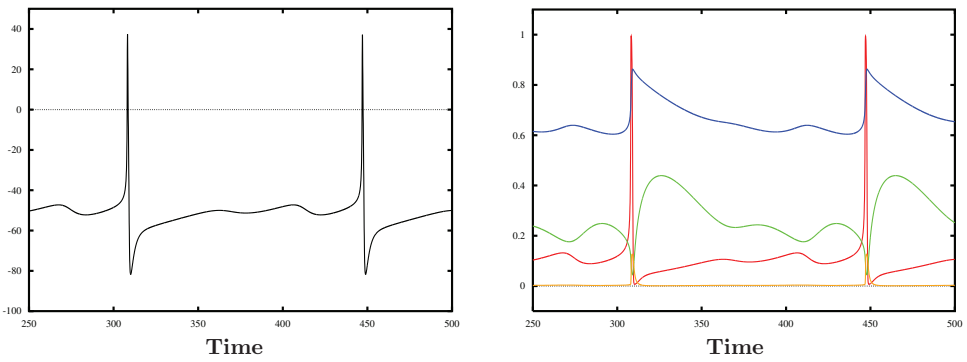


Figure 7. *Time traces in the Erisir model for $I_{app} = 0.7$. Left: Voltage time trace V shows MMO pattern 1^2. Right: Corresponding time traces of gates: m (red); h (green); s (blue); n (yellow). Note that the n-gate is basically zero in the subthreshold regime.*

Step 2: Reduction of slow variables. This step is based on an observation in the subthreshold regime. Figure 7 shows a time trace of the "membrane potential" v (left) as well as the corresponding time traces of the gates x (right), $x = m, h, n, s$, in system (3.4). Note that the n-gate does not influence the subthreshold regime at all, i.e., $n \approx 0$ in this regime. This can be easily explained by looking at the quasi steady state voltage functions $x_\infty(v)$ of the gates x shown in Figure 6. The sigmoidal activation function $n_\infty(v)$, the quasi steady state of n (shown in yellow), is approximately zero for $v < -0.4$ (equivalent to $V < -40\,mV$), the limit of the subthreshold regime. The time constant $\tau_n(v)$ is of order $1\,ms$. Therefore, after an action potential is fired and the cell is repolarized to the subthreshold regime, n will reach its zero equilibrium value after a few milliseconds, and n thereafter plays no role in the creation of STOs. Since we are primarily interested in the creation of the STOs in the MMO pattern, it is reasonable to set $g_K = 0$; i.e., we block the Kv3.1 channel in the Erisir model. We expect to find qualitative similar results (MMOs) without this channel but note that the Kv3.1 channel has a profound role in repolarizing the cell, and we will come back to this point later.

With this second reduction step we reduced the number of slow variables from three to two.

3.3. MMOs in the reduced Erisir model. In the following, we will focus on the reduced Erisir model

$$\begin{aligned} h' &= (1/t_h(v))(h_\infty(v) - h) =: g_1(v,h), \\ s' &= (1/t_s(v))(s_\infty(v) - s) =: g_2(v,s), \\ \varepsilon\, v' &= \hat{I}_{appl} - (\hat{I}_{Na} + \hat{I}_K + \hat{I}_{Ks} + \hat{I}_L) =: f(v,h,s), \end{aligned} \tag{3.5}$$

which is a singularly perturbed system (slow/fast system) with two slow variables (h, s) and one fast variable v. First we show that this model can produce similar patterns as the full Erisir model. Figure 8 shows the bifurcation diagram which has qualitatively the same structure as the bifurcation diagram of the full Erisir model shown in Figure 1: the cell is excitable for $I_{app} < I_{Hopf}$, where $I_{Hopf} \approx 0.615$ denotes a supercritical Hopf bifurcation. The corresponding

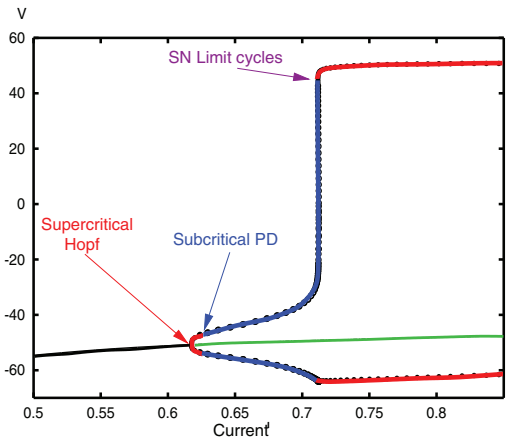


Figure 8. *Bifurcation diagram of reduced Erisir model (3.5). For easier comparison with the original model (2.1) we use the applied current I_{app} as the bifurcation parameter instead of its dimensionless counterpart $\hat{I}_{app}$. Similarly, we plot the phase variable V instead of its dimensionless counterpart v. We observe a supercritical Hopf bifurcation for $I_{app} \approx 0.62$, a limit cycle loses stability for $I_{app} \approx 0.63$, and we observe a saddle-node of limit cycles for $I_{app} \approx 0.71$.*

limit cycle loses stability at $I_{app} \approx 0.625$, which opens a parameter window up to $I_{app} = I_{snlc} \approx 0.715$ with apparently no stable attractors. I_{snlc} denotes a saddle-node bifurcation of limit cycles. For $I_{app} > I_{snlc}$ there exist large ROs. Note that we use I_{app} and not its dimensionless counterpart $\hat{I}_{app}$ as the bifurcation parameter. This allows us to compare the bifurcation structure directly with the bifurcation structure of the original model shown in Figure 1. Figure 9 shows time traces for several values of applied current I_{app}. Similar to the full Erisir model (2.1), MMOs show up as stable solutions for the reduced Erisir model filling the gap of apparently no stable solutions in Figure 8.

Note the qualitative differences between the full and the reduced Erisir model: the cell in the reduced Erisir model does not repolarize as much as in the full model (the difference is approximately $15\,mV$). On the other hand, the cell depolarizes more in the reduced model than in the full model (the difference is approximately $10\,mV$). This is explained by the Kv1.3 channel block which is responsible for the repolarization phase of the cell. Note that the gap between the HB and the SNLC in the bifurcation diagram of the reduced model, Figure 8, is just shifted slightly by $\Delta I_{app} \approx 0.03$ to the left compared to the full model, Figure 1. This is (mainly) because the hyperpolarizing outward current I_K is very weak in the subthreshold regime (approximately zero) and has no influence on the STOs, as explained in the reduction of the slow variables. In summary, it appears that the reduced Erisir model (3.5) covers qualitatively the behavior of the full Erisir model (2.1), at least close to the parameter regime where we observe MMOs.

3.4. Canard analysis of the reduced Erisir model. Here we explain the observed MMOs via the generalized canard phenomenon [7, 44, 50]. The canard theory applies to singular perturbation problems with two slow variables and one fast variable. The reduced Erisir model (3.5) has to fulfill a set of assumptions such that this theory can be applied, which we will show in the following.

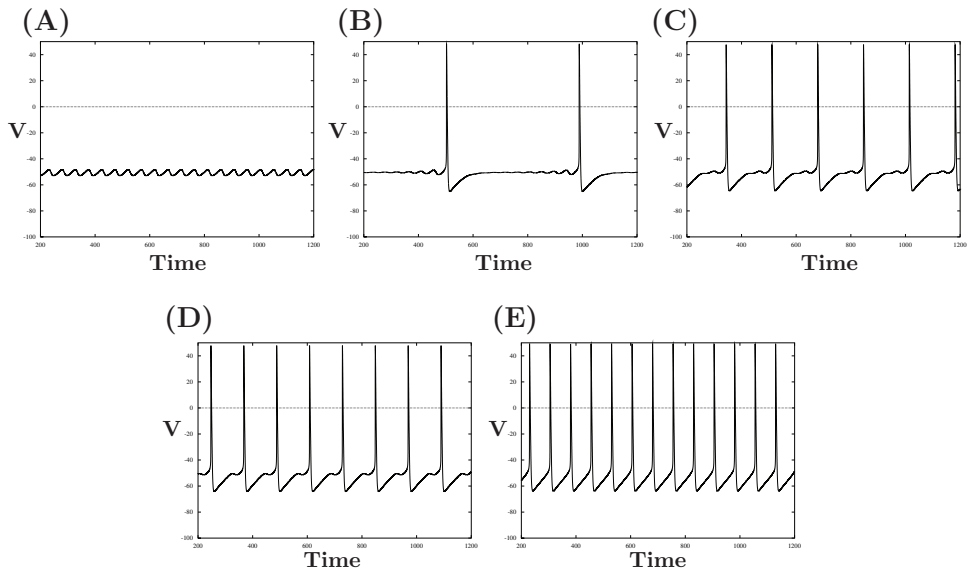


Figure 9. *Voltage time traces of reduced Erisir model* (3.5). (A): *small oscillatory pattern* 0^1 *for* $I_{app} = 0.62$; (B): *MMO pattern* 1^9 *for* $I_{app} = 0.635$; (C): *MMO pattern* 1^2 *for* $I_{app} = 0.67$; (D): *MMO pattern* 1^1 *for* $I_{app} = 0.69$; (E): *relaxation oscillation pattern* 1^0 *for* $I_{app} = 0.72$. *Note that we plot* V *instead of its dimensionless counterpart* v.

We first focus on the limiting problem $\varepsilon \to 0$ in system (3.5) which is called the *reduced problem*. It describes the evolution of the slow variables (h, s) on the *critical manifold* $S := \{(v, s, h) \in \mathbb{R}^3 : f(v, s, h) = 0\}$, i.e., on the v-nullsurface. This critical manifold S can be represented as a graph over, e.g., (s, v)-space given by

$$h(s, v) = \frac{\hat{I}_{app} - \hat{g}_{ks} s^4 (v - \hat{E}_k) - \hat{g}_L (v - \hat{E}_L)}{m_\infty^3(v)(v - \hat{E}_{Na})}. \tag{3.6}$$

The following assumption on the critical manifold S holds.

Assumption 1. *System* (3.5) *possesses (locally) in the subthreshold regime a folded critical manifold.*

Figure 10 (middle) shows the critical manifold for $I_{app} = 0.65$, and it is indeed folded. This is true for all I_{app} parameter values within the observed MMO regime (and beyond).

If we rescale to the fast time $\tau = t/\varepsilon$ in system (3.5) and take the singular limit $\varepsilon \to 0$, then we obtain the one-dimensional *layer problem* which describes the fast dynamics away from the critical manifold S. Trajectories of the layer problem evolve along fast fibers, i.e., one-dimensional sets ($h = \text{const}$, $s = \text{const}$, v), and the critical manifold S is a manifold of equilibria. The folded structure of S implies saddle-node bifurcations of equilibria within the layer problem along the fold-curve L. Therefore, we can describe the critical manifold as the union of subsets $S = S_a \cup L \cup S_r$, where S_a denotes the attracting (stable) branch of S, i.e., $f_v < 0$ on S_a, and S_r denotes the repelling branch of S, i.e., $f_v > 0$ on S_r. The fold-curve L is defined by $f_v = 0$. Therefore, the layer flow is toward S_a, respectively, away from S_r, along fast fibers.

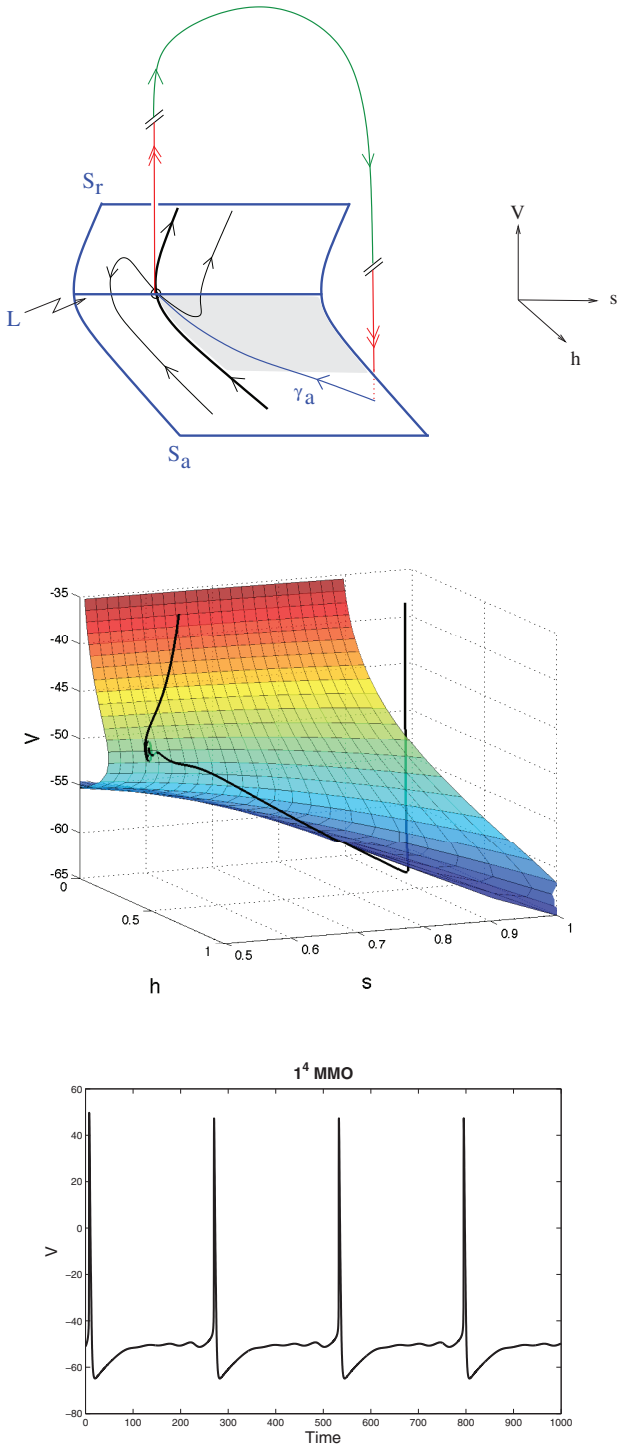


Figure 10. *Explanation of MMOs via generalized canard phenomenon. The upper figure shows a sketch of a singular periodic orbit in phase space which predicts MMOs; the middle figure shows the critical manifold and a trajectory of system* (3.5) *with four STOs in the phase space near the corresponding folded node singularity; the lower figure shows the corresponding time trace of the* 1^4 *MMO pattern for* $I_{app} = 0.65$. *Note that we plot* V *instead of its dimensionless counterpart* v.

To understand the reduced flow on the critical manifold S, we project the limiting system $\varepsilon \to 0$ of (3.5) onto (s, v)-space. This projection can be obtained from implicit differentiation of $f(h, s, v) = 0$ and the equation for $\dot{s}$ which gives

$$\begin{aligned} \dot{s} &= g_2, \\ -f_v \dot{v} &= f_h g_1 + f_s g_2, \end{aligned} \tag{3.7}$$

evaluated along $h = h(s, v)$. Note that this is a natural choice for projection since S is given as a graph over (s, v)-space in (3.6). System (3.7) is singular along the fold-curve L given by $f_v = 0$. We therefore rescale time by the factor $-f_v$ to obtain the *desingularized reduced flow*

$$\begin{aligned} s' &= -f_v g_2, \\ v' &= f_h g_1 + f_s g_2, \end{aligned} \tag{3.8}$$

where "prime" denotes differentiation with respect to the new time $t_1 = t/(-f_v)$. We obtain the phase portrait of the reduced system (3.7) by reversing the orientation of the trajectories of the desingularized system (3.8) on the repelling branch S_r; otherwise the two phase portraits are equivalent. System (3.8) has two types of singularities: regular and folded. Regular singularities are given by $g_1 = g_2 = 0$. They correspond to equilibria of the reduced flow (3.7) away from the fold-curve L. In contrast, folded singularities are given by $f_v = 0$ (the fold-curve condition) and $f_h g_1 + f_s g_2 = 0$ which defines isolated points on the fold-curve L. In general, such folded singularities are not equilibria of the reduced system (3.7). However, folded singularities give the opportunity for the reduced flow to cross from S_a to S_r via folded singularities in finite time. Depending on the classification as a singularity of (3.8), a folded singularity of the reduced system (3.7) is called a folded saddle (the corresponding singularity of (3.8) has two real eigenvalues with opposite signs), a folded node (the corresponding singularity of (3.8) has two real eigenvalues with the same sign), or a folded focus (the corresponding singularity of (3.8) has two complex conjugate eigenvalues).

Assumption 2. *For $0.55 < I_{app} < 0.74$, system (3.7) possesses a folded node singularity.*

Figure 11 shows the reduced flow in the silent phase ($V < -40mV$) for several values of applied current I_{app}. The green curve represents the fold-curve L of the critical manifold. For $I_{app} < 0.55$, there exists a node on the attracting branch S_a and a folded saddle (not shown). For $I_{app} > 0.55$, there exists a saddle on the repelling branch S_r and a folded node (e.g., the green triangle (saddle) and the cyan circle (folded node) shown in Figure 11, $I_{app} = 0.63$). Therefore, at $I_{app} \approx 0.55$ we observe a transcritical bifurcation of a folded and a regular singularity; i.e., the folded saddle and the node for $I_{app} < 0.55$ merge for $I_{app} \approx 0.55$ to split then into a folded node and a saddle for $I_{app} > 0.55$. The corresponding folded singularity at the bifurcation value $I_{app} = I_{fsnII} \approx 0.55$ is called a folded saddle-node type II (the yellow circle shown in Figure 11, $I_{app} = 0.55$). The corresponding saddle-node singularity of (3.8) has one zero eigenvalue and one nonzero eigenvalue. The "type II" refers to the observed transcritical bifurcation of singularities in system (3.8); see [50, 35].

For $I_{app} > I_{fsnII}$, we also observe another folded singularity—a folded saddle (can be seen as a blue circle in, e.g., Figure 11, $I_{app} = 0.63$). This folded saddle also exists for $I_{app} < I_{fsnII}$ but not in the physiological relevant domain ($s > 1$). As we increase the applied current I_{app}, the two folded singularities, the folded node and the folded saddle, approach each

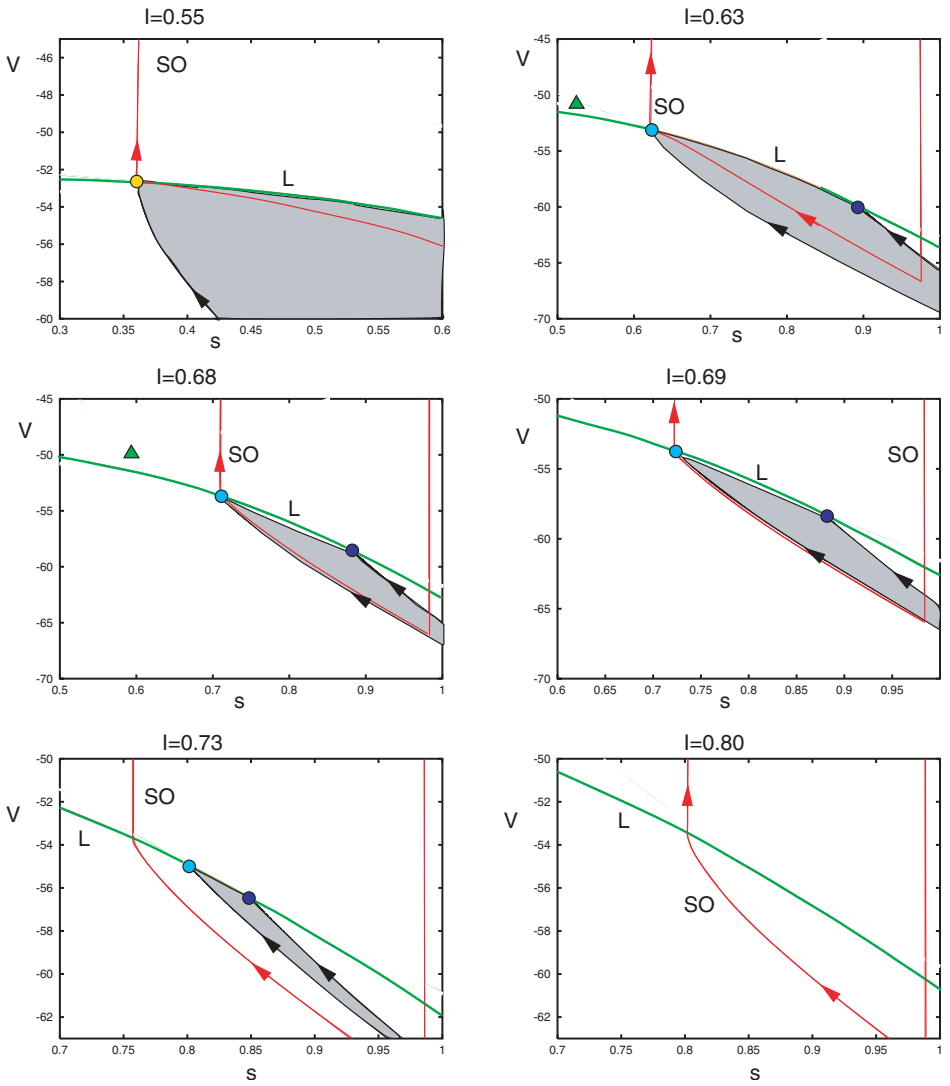


Figure 11. *The reduced flow* (3.7) *of system* (3.5). *Note that we plot* V *instead of its dimensionless counterpart* v. *Applied current* I_{app} *is shown above each panel. Red curves are the singular orbit (SO; computed by solving system* (3.5) *with* $C = 10^{-8}$, *equivalent to very small* $\epsilon \approx 10^{-8}$*); green curves represent the fold-curve* L. *Filled circles are the folded node (cyan), the folded saddle (blue), or the folded saddle-node type II (yellow). The filled triangle (green) is the saddle on* S_r. *Black curves (either alone or leftmost) represent the primary strong canard of the folded node or the folded saddle-node, and rightmost black curves are the canards of the folded saddle. The sector on* S_a *(below the fold-curve* L*) bounded by the two black curves (the strong canard of the folded node and the canard of the folded saddle) is called the* funnel *and is shaded grey. When the SO lies within the funnel, there will be MMOs for sufficiently small values of* ϵ *according to the theorem.*

other. For $I_{app} \approx 0.74$, the folded node (left, cyan circle) and the folded saddle (right, blue circle) merge and subsequently annihilate each other (transition in Figure 11 from $I_{app} = 0.73$ to $I_{app} = 0.80$). Therefore, we observe a saddle-node bifurcation of folded singularities at $I_{app} \approx 0.74$. The corresponding folded singularity at the bifurcation value $I_{app} = I_{fsnI} \approx 0.74$ is called a folded saddle-node type I. The corresponding saddle-node singularity of (3.8) has

one zero eigenvalue and one nonzero eigenvalue. The "type I" refers to the observed saddle-node bifurcation of singularities in system (3.8); see [50, 35]. This explains the difference between the two subtypes of folded saddle-nodes observed in this problem.

For $I_{app} > I_{fsnI}$ there just exists the ordinary saddle on the repelling branch S_r. (Note that this saddle is not shown in Figure 11, $I = 0.69, 0.73, 0.80$, since the corresponding V-value is bigger than $-50\,mV$.) Hence, Assumption 2 is fulfilled for $I_{fsnII} < I_{app} < I_{fsnI}$ as stated.

Note that a folded node creates a *singular funnel* on S_a; i.e., there exists a sector of solutions of the reduced flow (3.7) which is "funneled" from S_a through the folded node to S_r. In Figure 11, the singular funnel (shaded grey) is the area bounded by the fold-curve (green curve) and the two black curves, the primary strong canard of the folded node and the canard of the folded saddle. Identifying this funnel region is essential since solutions of system (3.5) which pass through this funnel region will create STOs. For details we refer the reader to [50].

Finally, we have to construct singular periodic orbits (SOs), concatenated by trajectory segments of both the reduced and the layer problems which explain the observed MMO patterns in (3.5).

Assumption 3. *System* (3.5) *possesses for* $0.55 < I_{app} < 0.69$ *an SO which consists of a segment on the stable folded surface* S_a *within the singular funnel of the folded node singularity as an endpoint.*

Such an SO is illustrated in Figure 10. The main step to find these SOs is to show that the return mechanism projects the fast fiber of the folded node back into the funnel region (see Figure 10). We do not calculate these SOs explicitly but approximate them by calculating periodic orbits in system (3.5) with very small $\varepsilon \approx 10^{-8}$. In the singular limit $\varepsilon \to 0$ these periodic orbits will become the SOs described in Assumption 3. Several examples of SOs are shown in Figure 11. The funnel of the folded node becomes obviously smaller with increasing I_{app} and disappears for $I_{app} = I_{fsnI}$ (the folded saddle-node type I). The global return mechanism projects the fast fiber of the folded node only for $0.55 < I_{app} < I_{rel} \approx 0.69$ into the funnel as shown in Figure 11. Therefore, Assumption 3 is fulfilled only for $I_{fsnII} < I_{app} < I_{rel}$. For $I_{app} \approx 0.68$ the return mechanism projects onto the border of the funnel, the strong canard of the folded node, which represents the separatrix in the phase space for different oscillatory behavior, i.e., between MMOs and relaxation oscillations. It was shown in [7] that if a singularly perturbed system fulfills Assumptions 1–3, then it possesses 1^s MMO patterns for sufficiently small ε. We conclude the following proposition.

Proposition 3.1. *System* (3.5) *possesses MMO-type periodic solutions based on the canard phenomenon for sufficiently small* $\varepsilon \ll 1$.

Note that by Assumptions 2 and 3 we expect to find MMOs for $I_{fsnII} \approx 0.55 < I_{app} < I_{rel} \approx 0.69$. Obviously, the boundaries of this prediction are sensitive to the size of the singular perturbation parameter $\varepsilon \ll 1$. For example, the bifurcation diagram of system (3.5), Figure 8, is calculated for $\varepsilon \approx 10^{-2}$. We observe MMO patterns for $I_{Hopf} \approx 0.615 < I_{app} < I_{snlc} \approx 0.715$. There is a marked shift in the onset of MMOs compared to the prediction $I_{fsnII} = 0.55$. If we decrease the value of ε toward zero, then I_{Hopf} converges to I_{fsnII}; i.e., a folded saddle-node type II is the singular limit representation of the Hopf bifurcation shown in Figure 8. Note further that a stable node on S_a crosses the fold-curve L at I_{fsnII} to become an (unstable) saddle on S_r. These equilibria (saddle and node) persist under small perturbations

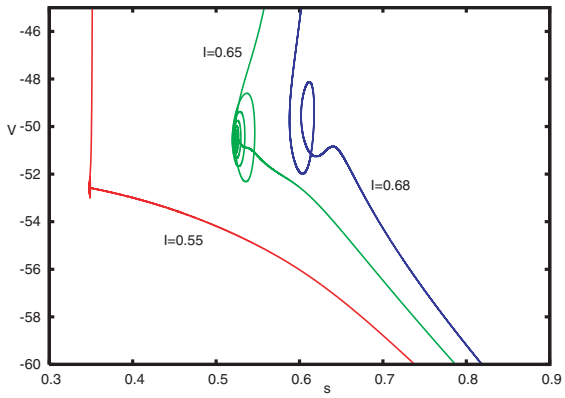


Figure 12. *Trajectories of system (3.5) projected onto (n_s, V)-plane at different currents corresponding to Figure 11.*

ε and are therefore also equilibria of (3.5). Recall the transcritical bifurcation of a folded and an ordinary singularity at $I_{fsnII} \approx 0.55$. These stable and unstable equilibria correspond to the branch of equilibria in the bifurcation diagram, Figure 8, where stability is lost via a Hopf bifurcation. This again emphasizes the relation between the folded saddle-node type II and the Hopf bifurcation. The smaller the singular perturbation ε, the closer the onset of MMOs to $I_{fsnII} \approx 0.55$.

To obtain MMO patterns it is essential that Assumption 3 is fulfilled as well. The return mechanism has to project within the funnel; otherwise we will observe relaxation oscillations but not MMOs. Therefore, the global return mechanism ultimately determines the I_{app} interval where MMOs are observed. In Figure 8, the upper limit I_{snlc} is the parameter value where the return mechanism of system (3.5) with $\varepsilon \approx 10^{-2}$ projects onto the border of the funnel. Again, if we decrease ε toward zero, then I_{snlc} approaches I_{rel} as predicted by the singular limit theory. Note that ε has an increasing effect on the MMO range at the upper limit I_{snlc}, while it has a decreasing effect on the MMO range at the lower limit I_{Hopf}.

We can also give a qualitative prediction on the number of STOs within the MMO pattern: the larger I_{app}, the smaller the number of STOs. Canard theory [50] predicts an upper limit on the number of STOs, $s^* = [(\mu+1)/(2\mu)]$ (the largest integer less than or equal $(\mu+1)/(2\mu)$), based on the eigenvalue ratio $\mu = \lambda_1/\lambda_2 < 1$ of the folded node singularity in system (3.8). This number decreases with increasing applied current[2] for $I_{fsnII} \approx 0.55 < I_{app} < I_{rel} \approx 0.69$. Furthermore, the return mechanism projects the trajectory closer to the strong canard for increased I_{app}. The number of STOs is smaller than the predicted maximal number s^* whenever the return mechanism projects the trajectory $O(\varepsilon^{(1-\mu)/2})$ close to the strong canard. In the limit $I_{app} = I_{rel}$, i.e., the switch from MMOs to relaxation oscillations, the number s^* of STOs becomes zero. This explains the decreased number of STOs within an MMO pattern for increased I_{app} shown in Figures 9 and 12.

[2]The eigenvalue ratio $\mu = \lambda_1/\lambda_2 < 1$ is not a monotone function of I_{app}. The value μ reaches its maximum ($\hat{\mu} \approx 0.27$) for $0.69 < I_{\hat{\mu}} < 0.7$. It increases from zero to $\hat{\mu}$ for $I_{fsnII} < I_{app} < I_{\hat{\mu}}$ and decreases from $\hat{\mu}$ to zero for $I_{\hat{\mu}} < I_{app} < I_{fsnI}$.

We want to point out that the canard theory developed in [7, 50] cannot fully explain the transition from STOs to MMOs near the folded saddle-node type II ($I_{app} = I_{fsnII}$). Note that the limit $\mu \to 0$ predicts infinitely many STOs, i.e., a transition from STOs to MMOs. As pointed out in our analysis, the onset of STOs is associated with an HB which in the singular limit corresponds to a folded saddle-node type II. Milik et al. [39] were the first who studied folded saddle-node type II induced MMOs observed in an autocatalator model. In [35], a rigorous blow-up analysis of a folded saddle-node type II singularity is presented which shows how two phenomena, canards and a singular HB, lead to complicated dynamics including STOs and MMOs. In [25], it is shown how one determines the type of the Hopf bifurcation associated with a folded saddle-node type II singularity (super- or subcritical). It also shows numerical results on nearby bifurcation structures such as period-doubling and torus bifurcation. Other work on folded saddle-nodes can be found in [34].

4. Weak coupling of the Erisir model. We now use weak coupling theory [14] to try to understand how the presence of the canards produces the ability of networks to form clusters. Briefly, weak coupling theory or averaging allows one to reduce systems of N coupled oscillators to sets of differential equations on an N-torus. Consider a coupled pair

$$X_1' = F(X_1) + \epsilon_1 G_1(X_1, X_2),$$
$$X_2' = F(X_2) + \epsilon_1 G_2(X_1, X_2),$$

where $X' = F(X)$ has a stable limit cycle $U(t)$ and ϵ_1 is small and positive. Here ϵ_1 is used to denote the coupling strength with the subscript 1 to distinguish it from the ϵ used to characterize the difference in time-scales in the previous section. In order for averaging and weak coupling theory to be applied, one generally has to assume that $\epsilon_1 \ll \epsilon$. That is, the coupling is "smaller" than any of the other rates characterizing the uncoupled dynamics. If this assumption holds, then the method of weak coupling allows us to conclude that for ϵ_1 sufficiently small, $X_j(t) = U(\theta_j(t)) + O(\epsilon_1)$ and

$$\theta_1 = 1 + \epsilon_1 H_1(\theta_2 - \theta_1) + o(\epsilon_1),$$
$$\theta_2 = 1 + \epsilon_1 H_2(\theta_1 - \theta_2) + o(\epsilon_1),$$

where

$$H_j(\phi) = \frac{1}{T}\int_0^T U^*(t) \cdot G_j(U(t), U(t+\phi))\, dt.$$

Here T is the period of the limit cycle and $U^*(t)$ is the unique periodic solution to the linear adjoint problem:

$$Y' = -D_X F(U(t))^T Y, \qquad Y(t) \cdot U'(t) = 1.$$

For our model, coupling is only through the voltage of the membrane, so that

$$H(\phi) = \frac{1}{T}\int_0^T V^*(t)(V(t+\phi) - V(t))\, dt. \tag{4.1}$$

$V^*(t)$ is the voltage component of the adjoint and is proportional to the phase response curve (PRC) obtained by applying brief small current perturbations to the neuron (see [29]). $V(t)$

is the voltage trace. We note that for gap junctions, $H(0) = 0$. Phase models enable us to predict the possible stable phase-locked patterns when two or more oscillators are coupled. For a pair of identical oscillators, we can write a single scalar equation for the phase-difference $\phi := \theta_2 - \theta_1$:

$$\phi' = \epsilon_1[H(-\phi) - H(\phi)] = -2\epsilon_1 H_{odd}(\phi).$$

Thus, the zeros of the odd part of H are the possible phase-differences, and a phase-difference of $\bar{\phi}$ is stable if $H'(\bar{\phi}) > 0$.

The generalization of the above theory to N oscillators is clear. In particular, for all-to-all coupling with identical oscillators, we obtain the following model for the phases:

$$\theta_i' = 1 + \epsilon_1 \frac{1}{N} \sum_{j=1}^{N} H(\theta_j - \theta_i). \tag{4.2}$$

A clustered solution for this system is a partition of N, that is, a collection of positive integers whose sum is N, say, $k_1, \ldots, k_m$, such that there are m distinct phases and k_l oscillators which are all synchronized in each group. For example, a two-cluster solution would have n oscillators with 0 relative phase and $N - n$ oscillators with $\bar{\phi}$ relative phase. The existence of m-clustered states requires solving m equations. Stability of these clustered states is far more difficult to determine and has been done in general only for the $m = 2$ case and for $m > 2$ for certain symmetric cases [28, 26, 1]. The number of oscillators in each cluster can vary, and, so, there can be multiple stable solutions corresponding to different numbers of oscillators in each cluster. The stability criteria constrains the number of solutions, but for large N there can coexist many stable clustered states (see [28]).

For very large N, one technique that is very useful was developed by Kuramoto [36]. He considers globally connected identical oscillators which are driven by uncorrelated noise. For large enough noise, the density of the phases for the noisy system as $N \to \infty$ is uniform (the asynchronous state). He performs a stability analysis on this state as the amplitude of the noise decreases and under some conditions, the asynchronous state loses stability to a symmetric clustered state. Which state emerges from the noise is very easy to determine. Let a_n denote the Fourier sine coefficients of $H(\phi)$. Then an m-cluster will bifurcate from the asynchronous state when the noise falls below a_m/m. Thus, we need only look at the biggest a_m/m to predict which clusters should emerge as the noise is reduced. While this will not pick out all the attractors, we can think of it as the analogue of simulated annealing in the hopes that the most attractive state is chosen. In this case, the clusters that emerge are similar in size. Thus, we suspect that the clusters with the largest basin of attraction are the equal size clusters. We remark that additive white noise in the original voltage equations becomes white noise that is modulated by the PRC or adjoint of the limit cycle [47]. Like Kuramoto, we replace the phase-dependence by constant noise with amplitude equal to the L^2 norm of the voltage component of the adjoint.

Thus, in the remainder of this section, we will look at the H function for the Erisir model and then try to understand why it looks like it does by considering the subthreshold behavior.

4.1. Phase models for the Erisir model. Using the theory outlined above, the main determinant of the behavior is the function $H(\phi)$ which gives (i) the possible stable pairwise

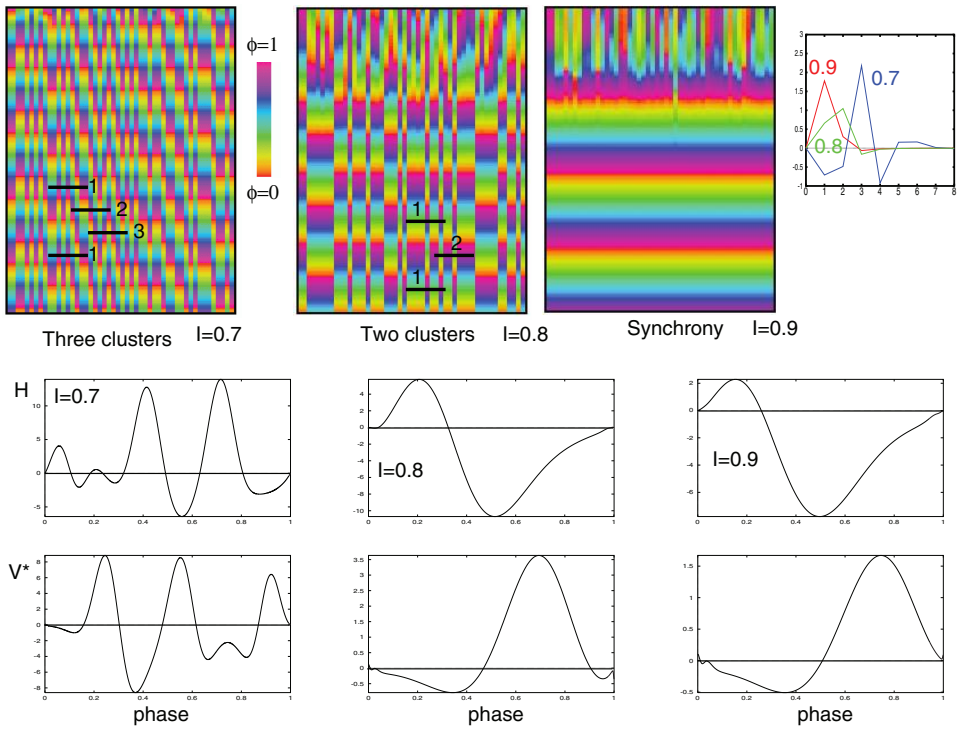


Figure 13. *Behavior of the phase-reduction of the Erisir model for* 50 *coupled oscillators. Each column shows the phases of each oscillator with time running downward and oscillator index across. The function* $H(\phi)$ *is then shown and the adjoint,* $V^*(\phi)$*, is at the bottom. Applied current* I_{app} *from left to right is* 0.7, 0.8, *and* 0.9 $\mu A/cm^2$*. Inset on right depicts* a_m/m*, where* a_m *is the Fourier sine component of* $H(\phi)$*; positive maxima predict the number of clusters.*

phase-differences and (ii) the form of the most stable cluster formation in the presence of noise.

Figure 13 shows the behavior of a phase model of the form (4.2), where H is as defined in (4.1) for the Erisir model with different values of I_{app}. The initial phases are randomly taken from a uniform distribution. In the leftmost panel, the phases organize into a three-cluster solution as in the full model shown in Figure 4, $I_{app} = 0.7$. The two graphs below show the interaction function H along with the adjoint, V^*. It is clear that H contains many higher order Fourier modes, and, in fact, the maximum of a_m/m occurs at $m = 3$, so that the three-cluster solution is predicted to be the most stable one. As the current is increased and the system moves closer to the usual repetitive firing regime, the number of clusters drops to two ($I_{app} = 0.8$) and then to one ($I_{app} = 0.9$), which is the synchronous solution. The Fourier analysis of H for $I_{app} = 0.8$ shows that $a_2/2$ is maximal so that two clusters are predicted to emerge from the asynchronous solution. Similarly, the dominant Fourier sine model is $m = 1$ for $I_{app} = 0.9$ so that there will only be synchrony.

We note that, for the three-cluster case, H is close to the mirror image of V^*. The reason for this can be seen by looking at the definition of H, (4.1). The H function is the convolution of the voltage trace with the adjoint. However, for the long-periodic oscillations associated

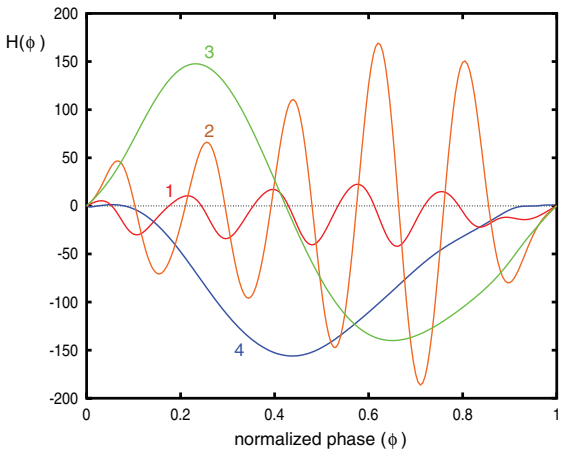


Figure 14. *The adjoint is responsible for the shape of H. Four curves show H computed with the standard model ($I_{app} = 0.657$, $g_{Ks} = 0.018$) or the simpler model ($I_{app} = 0.675$, $g_{Ks} = 0$). 1: standard voltage trace + standard adjoint; 2: simpler voltage trace + standard adjoint; 3: simple voltage trace + simple adjoint; 4: standard voltage trace + simple adjoint.*

with clustering, the voltage trace looks very close to a Dirac delta function, so that a rough approximation of H is

$$H(\phi) \approx \frac{1}{T}\int_0^T V^*(t)[\delta(t+\phi) - \delta(t)]\,dt = V^*(-\phi) - V^*(0), \tag{4.3}$$

which is just the mirror image of $V^*(t)$. This suggests that the key to understanding clustering lies in the adjoint (PRC) rather than the voltage trace. This notion is borne out by looking at Figure 14, in which we use the shape of the adjoint for $I_{app} = 0.675$ (many STOs in V and many oscillations in $V^*(t)$) and for the same current but with g_{Ks} turned off (no STOs and a simple bimodal $V^*(t)$). We compute four different versions of H using all combinations of the complex/simple adjoint and the complex/simple voltage trace. As the figure illustrates in curves 1 and 2, the complex adjoint combined with either of the two voltage traces leads to a highly oscillatory function H, while the simple adjoint yields relatively simple functions H. Thus, the numerics here and the approximation (4.3) provide strong evidence that the multiple oscillations in the interaction function H are due to the complexity of the adjoint/PRC $V^*(t)$ and not to the waveform of the membrane potential.

If we recall that the adjoint is very closely related to the PRC, then we can exploit this to understand the role of the canard structure and STOs in shaping the PRC. Figure 15 shows a blowup of the subthreshold region in the V-n_s phase plane. The unperturbed period is 234 msec. Two different brief (0.1 msec) pulses are delivered at either 70 msec or 90 msec after the neuron spikes. The 70 msec pulse results in a delay of the next spike by 45 msec, while the 90 msec pulse causes the spike to occur 42 msec earlier. The mechanism for this difference is clear from the figure. The earlier pulse makes the neuron undergo an additional oscillation (red loop in the center), while the later pulse causes the neuron to skip a pulse. Thus, the presence of the STOs leads to sensitivity to the timing of small pulses and a PRC

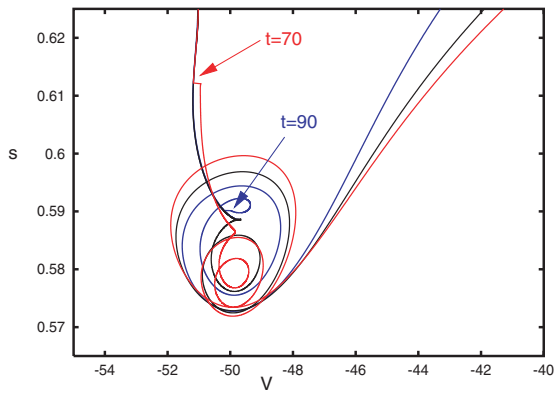


Figure 15. *Explanation for the complexity of the PRC (adjoint). Small differences in the timing of brief input have magnified effects near the subthreshold dynamics. Here* $I_{app} = 0.675$ *and the period is about* 234 *msec. A current pulse lasting* 0.1 *msec with amplitude* 0.25 $\mu A/cm^2$ *is delivered at* $t = 70$ *(red curve) and* $t = 90$ *(blue curve) msec after the spike. The black curve is the unperturbed trajectory.*

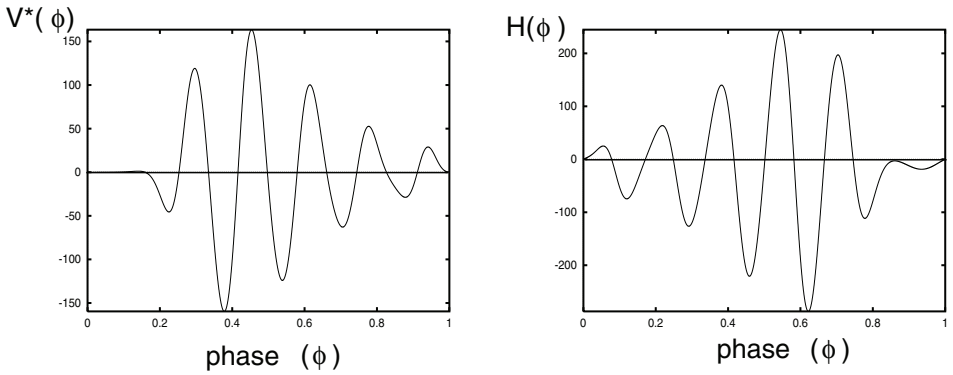


Figure 16. *The adjoint,* $V^*(\phi)$*, and the interaction function,* $H(\phi)$*, for the reduced Erisir model with* $I_{app} = 0.65$*, which is in the regime where there is a folded canard. Compare these to the corresponding functions for the full model in the left panel of Figure* 13.

which has multiple advances and delays during the cycle.

Finally, we note that the same behavior is found in the reduced (three-variable model). Figure 16 shows the adjoint and the interaction function for the reduced model when $I_{app} = 0.65$, which is in the region where there is a folded canard structure. Compare this adjoint/H function with that in Figure 13 on the left. They are qualitatively similar. Thus, the reduction has little qualitative effect on the phase dynamics.

5. Conclusion. In this paper, we have explored the interactions between local dynamics and coupling of neural oscillators. Our choice of coupling, electrical, is motivated by the preponderance of such connections between inhibitory interneurons in the brain (particularly the cortex). Electrical coupling is also the simplest form of coupling and does not introduce the additional complexities of synaptic coupling which involve a variety of time-scales and the kinds of synapses (inhibitory or excitatory). The model used in the present paper has been

used by other authors in order to study the synchronization properties of inhibitory neurons. The novel aspect of the biophysical model is that it contains a slow potassium channel which enables the voltage to undergo mixed mode oscillations (MMOs). This mixture of spikes and subthreshold oscillations (STOs) makes the neuron very sensitive to inputs and leads to a complex phase response curve (PRC). This complex PRC provides the mechanism for multiple clustering in networks of these neurons. There have been few other studies of the interactions between MMOs and coupling. Baba and Krischer [2] found clustering with strong coupling in a chemical model. In their paper, the main tools were fast/slow analysis; weak coupling and phase methods were not applied. Izhikevich [30] looks at weak coupling between elliptic bursters (which are related to MMOs; cf. [12]). Here, the author is interested only in spike versus burst synchrony and does not relate the dynamics to the interaction function.

Biologically, this work provides a novel contribution to the analysis of the effects of slow currents on dynamics and coupling. There have been a number of papers which show that slow currents and other intrinsic currents have profound effects on synchronization [41, 27, 16, 21]; however, all these illustrated transitions between synchrony and "antiphase" oscillations. Furthermore, while the effects were often shown to be a consequence of changing the shape of the PRC, these shape changes were subtle. In the present paper, the slow potassium current I_{Ks} has a profound effect and produces MMOs which, in turn, result in a very complex PRC. Rotstein, Wechselberger, and Kopell [43] identified a slow h-current I_h in an entorhinal cortex layer II stellate cell model mainly responsible for observed MMOs and showed that the canard phenomenon explains these MMOs as well. They also pointed out the effect of MMOs on the PRC.

Strong coupling destroys the clustering ability of these networks (results not shown), and this can be understood by also noting that strong pulses do not have the same subtle effects on the PRC as do weak pulses. Thus, the results described here require "weak" coupling. We point out that the coupling conductance is about 10% of the slow potassium conductance. Finally, we remark that similar results are found with weak inhibitory synaptic coupling as long as the synapses are short-lasting.

Appendix A. Functions and parameters of Erisir model (2.1). The biophysical model has the form

$$\begin{aligned}
C\,V' &= I_{app} - g_L(V - E_L) - g_{Na}m^3h(V - E_{Na}) - g_K n^2(V - E_K) - g_{Ks}s^4(V - e_K),\\
m' &= a_m(V)(1 - m) - b_m(V)m,\\
h' &= a_h(V)(1 - h) - b_h(V)h,\\
n' &= a_n(V)(1 - n) - b_n(V)n,\\
s' &= a_s(V)(1 - s) - b_s(V)s,
\end{aligned}$$

where

$$\begin{aligned}
a_m(V) &= 40(75 - V)/(\exp((75 - V)/13.5) - 1),\\
b_m(V) &= 1.2262\exp(-V/42.248),\\
a_h(V) &= 0.0035\exp(-V/24.186),\\
b_h(V) &= 0.017(-51.25 - V)/(\exp(-(51.25 + V)/5.2) - 1),
\end{aligned}$$

$$
\begin{aligned}
a_n(V) &= (95 - V)/(\exp((95 - V)/11.8) - 1),\\
b_n(V) &= 0.025\exp(-V/22.22),\\
a_s(V) &= -(0.616 + 0.014V)/(\exp(-(44 + V)/2.3) - 1),\\
b_s(V) &= 0.0043\exp(-(44 + V)/34).
\end{aligned}
$$

The kinetics of these variables are exactly as published in [38]. We have made some small changes to the conductances (in particular, our leak g_L is somewhat larger). The current I_{app} is varied as is the slow conductance, g_{KS}. Unless otherwise indicated, the default values are $E_K = -97$, $E_{Na} = 55$, $E_L = -70$ (all in mV), $C = 0.1\mu F/cm^2$, $g_L = 0.041$, $g_K = 18$, $g_{Na} = 9$, and $g_{Ks} = 0.018$ (all in $\mu S/cm^2$).

In the network model, 50 such cells are coupled together via global coupling:

$$
I_{coup} = \frac{g_{gap}}{50}\sum_{j=1}^{50}(V_j - V_i),
$$

which is added to the right-hand side of the voltage equation for each oscillator. In simulations with noise, a normally distributed random variable is added to each voltage equation.

Acknowledgments. B. Ermentrout thanks the National Science Foundation for partial support on this project. Both authors would also like to thank Steve Coombes, Andre Longtin, and Jon Rubin for organizing the conference where this work was conceived.

REFERENCES

[1] P. Ashwin and J. W. Swift, *The dynamics of n weakly coupled oscillators*, J. Nonlinear Sci., 2 (1992), pp. 69–108.

[2] N. Baba and K. Krischer, *Mixed-mode oscillations and cluster patterns in an electrochemical relaxation oscillator under galvanostatic control*, Chaos, 18 (2008), 015103.

[3] M. Banaji, *Clustering in globally coupled oscillators*, Dynamical Systems, 17 (2002), pp. 263–285.

[4] E. Benoit, *Systèmes lents-rapides dans R^3 et leur canards*, in Third Schnepfenried Geometry Conference Vol. 2, Astérisque 109–110, Soc. Math. France, Paris, 1983, pp. 159–191.

[5] K. Bold, C. Edwards, J. Guckenheimer, S. Guharay, K. Hoffman, J. Hubbard, R. Oliva, and W. Weckesser, *The forced van der Pol equation* II: *Canards in the reduced system*, SIAM J. Appl. Dyn. Syst., 2 (2003), pp. 570–608.

[6] M. Brøns, T. Kaper, and H. Rotstein, *Introduction to focus issue: Mixed mode sscillations: Experiment, computation, and analysis*, Chaos, 18 (2008), 015101.

[7] M. Brøns, M. Krupa, and M. Wechselberger, *Mixed mode oscillations due to the generalized canard phenomenon*, in Bifurcation Theory and Spatio-temporal Pattern Formation, Fields Inst. Commun. 49, AMS, Providence, RI, 2006, pp. 39–63.

[8] E. Brown, P. Holmes, and J. Moehlis, *Globally coupled oscillator networks*, in Perspectives and Problems in Nonlinear Science: A Celebratory Volume in Honor of Larry Sirovich, E. Kaplan, J. Marsden, and K. Sreenivasan, eds., Springer-Verlag, New York, 2003, pp. 183–215.

[9] C. C. Chow and N. Kopell, *Dynamics of spiking neurons with electrical coupling*, Neural Comput., 12 (2000), pp. 1643–1678.

[10] M. Desroches, B. Krauskopf, and H. M. Osinga, *The geometry of slow manifolds near a folded node*, SIAM J. Appl. Dyn. Syst., 7 (2008), pp. 1131–1162.

[11] S. Doi, J. Inoue, and S. Kumagai, *Chaotic spiking in the Hodgkin Huxley nerve model with slow inactivation of the sodium current*, J. Integr. Neurosci., 3 (2004), pp. 207–225.

[12] J. Drover, J. Rubin, J. Su, and B. Ermentrout, *Analysis of a canard mechanism by which excitatory synaptic coupling can synchronize neurons at low firing frequencies*, SIAM J. Appl. Math., 65 (2004), pp. 69–92.
[13] A. Erisir, D. Lau, B. Rudy, and S. Leonard, *Function of specific K+ channels in sustained high-frequency firing of fast-spiking neocortical cells*, J. Neurophysiol., 82 (1999), pp. 2476–2489.
[14] G. B. Ermentrout and N. Kopell, *Frequency plateaus in a chain of weakly coupled oscillators,* I., SIAM J. Math. Anal., 15 (1984), pp. 215–237.
[15] B. Ermentrout, *Gap junctions destroy persistent states in excitatory networks*, Phys. Rev. E (3), 74 (2006), 031918.
[16] B. Ermentrout, M. Pascal, and B. Gutkin, *The effects of spike frequency adaptation and negative feedback on the synchronization of neural oscillators*, Neural Comput., 13 (2001), pp. 1285–1310.
[17] J. Gao and P. Holmes, *On the dynamics of electrically-coupled neurons with inhibitory synapses*, J. Comput. Neurosci., 22 (2007), pp. 39–61.
[18] J. R. Gibson, M. Beierlein, and B. W. Connors, *Functional properties of electrical synapses between inhibitory interneurons of neocortical layer* 4, J. Neurophysiol., 93 (2005), pp. 467–480.
[19] D. Golomb, D. Hansel, B. Shraiman, and H. Sompolinsky, *Clustering in globally coupled phase oscillators*, Phys. Rev. A, 45 (1992), pp. 3516–3530.
[20] D. Golomb, K. Donner, L. Shacham, D. Shlosberg, Y. Amitai, and D. Hansel, *Mechanisms of firing patterns in fast-spiking cortical interneurons*, PLoS. Comput. Biol., 3 (8) (2007), e156.
[21] J. A. Goldberg, C. A. Deister, and C. J. Wilson, *Response properties and synchronization of rhythmically firing dendritic neurons*, J. Neurophysiol., 97 (2007), pp. 208–219.
[22] J. Guckenheimer and R. Haiduc, *Canards at folded nodes*, Mosc. Math. J., 5 (2005), pp. 91–103.
[23] J. Guckenheimer, R. Harris-Warrick, J. Peck, and A. Willms, *Bifurcation, bursting, and spike frequency adaptation*, J. Comput. Neurosci., 4 (1997), pp. 257–277.
[24] J. Guckenheimer, *Return maps of folded nodes and folded saddle-nodes*, Chaos, 18 (2008), 015108.
[25] J. Guckenheimer, *Singular Hopf bifurcation in systems with two slow variables*, SIAM J. Appl. Dyn. Syst., 7 (2008), pp. 1355–1377.
[26] F. Gürel Kazanci and G. B. Ermentrout, *Wave formation through the interactions between clustered states and local coupling in arrays of neural oscillators*, SIAM J. Appl. Dyn. Syst., 7 (2008), pp. 491–509.
[27] B. S. Gutkin, G. B. Ermentrout, and A. D. Reyes, *Phase-response curves give the responses of neurons to transient inputs*, J. Neurophysiol., 94 (2005), pp. 1623–1635.
[28] D. Hansel, G. Mato, and C. Meunier, *Clustering and slow switching in globally coupled phase oscillators*, Phys. Rev. E, 48 (1993), pp. 3470–3477.
[29] D. Hansel, G. Mato, and C. Meunier, *Synchrony in excitatory neural networks*, Neural Comput., 7 (1995), pp. 307–337.
[30] E. M. Izhikevich, *Synchronization of elliptic bursters*, SIAM Rev., 43 (2001), pp. 315–344.
[31] E. M. Izhikevich, *Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting*, MIT Press, Cambridge, MA, 2007.
[32] H. Y. Jeong and B. Gutkin, *Synchrony of neuronal oscillations controlled by GABAergic reversal potentials*, Neural Comput., 19 (2007), pp. 706–729.
[33] N. Kopell and B. Ermentrout, *Chemical and electrical synapses perform complementary roles in the synchronization of interneuronal networks*, Proc. Natl. Acad. Sci. USA, 101 (2004), pp. 15482–15487.
[34] M. Krupa, N. Popovic, N. Kopell, and H. Rotstein, *Mixed-mode oscillations in a three time-scale model for the dopaminergic neuron*, Chaos, 18 (2008), 015106.
[35] M. Krupa and M. Wechselberger, *Local analysis near a folded saddle-node*, preprint, University of Sydney, Sydney, Australia, 2008.
[36] Y. Kuramoto, *Chemical Oscillations, Waves, and Turbulence*, Springer-Verlag, Berlin, 1984.
[37] T. J. Lewis and J. Rinzel, *Dynamics of spiking neurons connected by both inhibitory and electrical coupling*, J. Comput. Neurosci., 14 (2003), pp. 283–309.
[38] J. G. Mancilla, T. J. Lewis, D. J. Pinto, J. Rinzel, and B. W. Connors, *Synchronization of electrically coupled pairs of inhibitory interneurons in neocortex*, J. Neurosci., 27 (2007), pp. 2058–2073.

[39] A. Milik, P. Szmolyan, H. Loeffelmann, and E. Groeller, *Geometry of mixed-mode oscillations in the* 3*D autocatalator*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 8 (1998), pp. 505–519.
[40] M. Nomura, T. Fukai, and T. Aoyagi, *Synchrony of fast-spiking interneurons interconnected by GABAergic and electrical synapses*, Neural Comput., 15 (2003), pp. 2179–2198.
[41] B. Pfeuty, G. Mato, D. Golomb, and D. Hansel, *The combined effects of inhibitory and electrical synapses in synchrony*, Neural Comput., 17 (2005), pp. 633–670.
[42] H. Rotstein, T. Opperman, J. White, and N. Kopell, *The dynamic structure underlying subthreshold oscillatory activity and the onset of spikes in a model of medial entorhinal cortex stellate cells*, J. Comput. Neurosci., 21 (2006), pp. 271–292.
[43] H. G. Rotstein, M. Wechselberger, and N. Kopell, *Canard induced mixed-mode oscillations in a medial entorhinal cortex layer* II *stellate cell model*, SIAM J. Appl. Dyn. Syst., 7 (2008), pp. 1582–1611.
[44] J. Rubin and M. Wechselberger, *Giant squid—hidden canard: The* 3*D geometry of the Hodgkin Huxley model*, Biol. Cybernet., 97 (2007), pp. 5–32.
[45] J. Rubin and M. Wechselberger, *The selection of mixed-mode oscillations in a Hodgkin-Huxley model with multiple timescales*, Chaos, 18 (2008), 015105.
[46] P. Szmolyan and M. Wechselberger, *Canards in* $\mathbb{R}^3$, J. Differential Equations, 177 (2001), pp. 419–453.
[47] J. Teramae and D. Tanaka, *Robustness of the noise-induced phase synchronization in a general class of limit cycle oscillators*, Phys. Rev. Lett., 93 (2004), 204103.
[48] C. Van Vreeswijk, L. F. Abbott, and G. B. Ermentrout, *When inhibition not excitation synchronizes neural firing*, J. Comput. Neurosci., 1 (1994), pp. 313–321.
[49] X. J. Wang and G. Buszaki, *Gamma oscillation by synaptic inhibition in a hippocampal interneuronal network model*, J. Neurosci., 16 (1996), pp. 6402–6413.
[50] M. Wechselberger, *Existence and bifurcation of canards in* $\mathbb{R}^3$ *in the case of a folded node*, SIAM J. Appl. Dyn. Syst., 4 (2005), pp. 101–139.
[51] M. Wechselberger, *Canards*, Scholarpedia, 2 (2007), p. 1356.
[52] J. A. White, C. C. Chow, J. Ritt, C. Soto-Trevio, and N. Kopell, *Synchronization and oscillatory dynamics in heterogeneous, mutually inhibited neurons*, J. Comput. Neurosci., 5 (1998), pp. 5–16.

SIAM J. Applied Dynamical Systems
Vol. 8, No. 1, pp. 279–297
© 2009 Society for Industrial and Applied Mathematics

Nonautonomous Hamiltonian Systems and Morales–Ramis Theory I. The Case $\ddot{x} = f(x,t)$*

Primitivo B. Acosta-Humánez†

Abstract. In this paper we present an approach toward the comprehensive analysis of the nonintegrability of differential equations in the form $\ddot{x} = f(x,t)$ which is analogous to Hamiltonian systems with $1+1/2$ degrees of freedom. In particular, we analyze the nonintegrability of some important families of differential equations such as Painlevé II, Sitnikov, and the Hill–Schrödinger equation. We emphasize Painlevé II, showing its nonintegrability through three different Hamiltonian systems, and also Sitnikov, in which two different versions including numerical results are shown. The main tool for studying the nonintegrability of these kinds of Hamiltonian systems is Morales–Ramis theory. This paper is a very slight improvement to the talk with a similar title delivered by the author at the SIAM Conference on Applications of Dynamical Systems in 2007.

Key words. Hill–Schrödinger equation, Morales–Ramis theory, nonautonomous Hamiltonian systems, nonintegrability of Hamiltonian systems, Painlevé II equation, Sitnikov problem, virtually abelian groups

AMS subject classifications. 37J30, 12H05, 70H07

DOI. 10.1137/080730329

1. Introduction.

In this section we present the necessary theoretical background for understanding the rest of the paper.

1.1. Differential Galois theory.

Our theoretical framework consists of a well-established combination of dynamical systems theory, algebraic geometry, and differential algebra. See [22] or [36] for further information and details. Given a linear differential system with coefficients in $\mathbb{C}(t)$,

$$\dot{\boldsymbol{z}} = A(t)\,\boldsymbol{z}, \tag{1.1}$$

a differential field $L \supset \mathbb{C}(t)$ exists, unique up to $\mathbb{C}(t)$-isomorphism, which contains all entries of a fundamental matrix $\Psi = [\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_n]$ of (1.1). Moreover, the group of differential automorphisms of this field extension, called the *differential Galois group* of (1.1), is an algebraic group G acting over the $\mathbb{C}$-vector space $\langle \boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_n \rangle$ of solutions of (1.1) and containing the monodromy group of (1.1).

It is worth recalling that the integrability of a linear system (1.1) is equivalent to the solvability of the identity component G^0 of the differential Galois group G of (1.1)—in other words, equivalent to the *virtual solvability* of G.

*Received by the editors July 14, 2008; accepted for publication (in revised form) by J. Meiss November 1, 2008; published electronically February 13, 2009. This research was partially supported by grant FPI Spanish Government, project BFM2003-09504-C02-02.
http://www.siam.org/journals/siads/8-1/73032.html
†Departament de Matemàtica Aplicada I, Universitat Politècnica de Catalunya, Diagonal 647, E-08028 Barcelona, Spain (primitivo.acosta@upc.edu).

It is well established (e.g., [23]) that any linear differential equation system with coefficients in a differential field K,

$$\frac{d}{dt}\begin{pmatrix}\xi_1\\ \xi_2\end{pmatrix}=\begin{pmatrix}a(t) & b(t)\\ c(t) & d(t)\end{pmatrix}\begin{pmatrix}\xi_1\\ \xi_2\end{pmatrix}, \tag{1.2}$$

by means of an elimination process, is equivalent to the second-order equation

$$\ddot{\xi}-\left(a(t)+d(t)+\frac{\dot{b}(t)}{b(t)}\right)\dot{\xi}-\left(\dot{a}(t)+b(t)c(t)-a(t)d(t)-\frac{a(t)\dot{b}(t)}{b(t)}\right)\xi=0, \tag{1.3}$$

where $\xi:=\xi_1$. Furthermore, any equation of the form $\ddot{z}-2p\dot{z}-qz=0$ can be transformed, through the change of variables $z=ye^{\int p}$, into $\ddot{y}=-ry$, r, satisfying the Riccati equation $\dot{p}=r+q+p^2$. This change is useful since it restricts the study of the Galois group of $\ddot{y}=-ry$ to that of the algebraic subgroups of $\mathrm{SL}_2(\mathbb{C})$.

A natural question which now arises is to determine what happens if the coefficients of the differential equation are not all rational. A new method was developed in [3] to transform a linear differential equation of the form $\ddot{x}=r(t)x$, with transcendental or algebraic nonrational coefficients, into its algebraic form—that is, into a differential equation with rational coefficients. This is called the *algebrization method* and is based on the concept of a *Hamiltonian change of variables* [3]. Such a change is derived from the solution of a one-degree-of-freedom classical Hamiltonian.

Definition 1.1 (Hamiltonian change of variables). *A change of variables $\tau=\tau(t)$ is called* Hamiltonian *if $(\tau(t),\dot{\tau}(t))$ is a solution curve of the autonomous Hamiltonian system X_H with Hamiltonian function*

$$H=H(\tau,p)=\frac{p^2}{2}+\widehat{V}(\tau) \text{ for some } \widehat{V}\in\mathbb{C}(\tau).$$

Theorem 1.1 (Acosta–Blázquez algebrization method [3]). *Equation $\ddot{x}=r(t)x$ is algebrizable by means of a Hamiltonian change of variables $\tau=\tau(t)$ if and only if there exist f,α such that $\frac{d}{d\tau}(\ln\alpha)$, $\frac{f}{\alpha}\in\mathbb{C}(\tau)$, where*

$$f(\tau(t))=r(t),\quad \alpha(\tau)=2(H-\widehat{V}(\tau))=(\dot{\tau})^2.$$

Furthermore, the algebraic form of $\ddot{x}=r(t)x$ is

$$\frac{d^2x}{d\tau^2}+\left(\frac{1}{2}\frac{d}{d\tau}\ln\alpha\right)\frac{dx}{d\tau}-\left(\frac{f}{\alpha}\right)x=0. \tag{1.4}$$

The next intended step, once a differential equation has been algebrized, is studying its Galois group and, as a causal consequence, its integrability. Concerning the latter, and by virtue of the invariance of the identity component of the Galois group by finite branched coverings of the independent variable (Morales-Ruiz and Ramis, [25, Theorem 5]), it was proven in [3, Proposition 1] that the identity component of *the Galois group is preserved in the algebrization mechanism.*

The final step is analyzing the behavior of $t = \infty$ (or $\tau = \infty$) by studying the behavior of $\eta = 0$ through the change of variables $\eta = 1/t$ (or $\eta = 1/\tau$) in the transformed differential equation; i.e., $t = \infty$ (or $\tau = \infty$) is an ordinary point (resp., a regular singular point, an irregular singular point) of the original differential equation if and only if $\eta = 0$ is one such point for the transformed differential equation.

1.2. Morales–Ramis theory. *Everything is considered in the complex analytical setting from now on.* The heuristics of the titular theory rest on the following general principle: if we assume system

$$\dot{\boldsymbol{z}} = X(\boldsymbol{z}) \tag{1.5}$$

"integrable" in some reasonable sense, then the corresponding variational equations along any integral curve $\Gamma = \{\widehat{\boldsymbol{z}}(t) : t \in I\}$ of (1.5), defined in the usual manner

$$\dot{\boldsymbol{\xi}} = X'(\widehat{\boldsymbol{z}}(t))\,\boldsymbol{\xi}, \tag{VE_Γ}$$

must be also integrable—in the Galoisian sense of the last paragraph in section 1.1. We assume Γ, a Riemann surface, may be locally parametrized in a disc I of the complex plane; we may now complete Γ to a new Riemann surface $\overline{\Gamma}$, as detailed in [25, sect. 2.1] (see also [22, sect. 2.3]), by adding equilibrium points, singularities of the vector field, and possible points at infinity.

The aforementioned "reasonable" sense in which to define integrability if system (1.5) is *Hamiltonian* is obviously the one given by the Liouville–Arnold theorem, and thus the above general principle does have an implementation.

Theorem 1.2 (Morales-Ruiz and Ramis [22, 25]). *Let H be an n-degree-of-freedom Hamiltonian having n independent rational or meromorphic first integrals in pairwise involution, defined on a neighborhood of an integral curve $\overline{\Gamma}$. Then, the identity component* $\mathrm{Gal}\left(\mathrm{VE}_{\overline{\Gamma}}\right)^0$ *is an abelian group (i.e.,* $\mathrm{Gal}\left(\mathrm{VE}_{\overline{\Gamma}}\right)$ *is* virtually abelian*).*

The disjunctive between *meromorphic* and *rational* Hamiltonian integrability in Theorem 1.2 is related to the status of $t = \infty$ as a singularity for the normal variational equations. More specifically, and besides the nonabelian character of the identity component of the Galois group, in order to obtain Galoisian obstructions to the *meromorphic* integrability of H the point at infinity must be a regular singular point of (VE_Γ). On the other hand, for there to be an obstruction to complete sets of *rational* first integrals, $t = \infty$ must be an irregular singular point.

See [25, Corollary 8] or [22, Theorem 4.1] for a precise statement and a proof.

1.3. Nonautonomous Hamiltonian systems. *Nonautonomous Hamiltonian systems* on symplectic manifolds have long been the subject of study and appear in a most natural way in classical mechanics and control theory, e.g., [1], [5], [19], [21], [20], [28], [33], [34].

We consider nonautonomous Hamiltonian systems of the form

$$H = H(q_1, p_1, t) = \frac{p_1^2}{2} + V(q_1, p_1, t). \tag{1.6}$$

H is a nonautonomous Hamiltonian system with $1 + 1/2$ degrees of freedom. It is well known (e.g., [28]) that (1.6) can be included as a subsystem of the Hamiltonian system with two

degrees of freedom given by

$$\widehat{H} = \widehat{H}(q_1, q_2, p_1, p_2) = \frac{p_1^2}{2} + V(q_1, p_1, q_2) + p_2, \tag{1.7}$$

where q_2 and p_2 are conjugate variables, i.e., $p_2 = -H + k$, where k is constant, and $q_2 = t$. Furthermore, p_2 is easily seen to be the offset or counterbalancing energy of the system [28], [30].

Also worth mentioning are some recent results on canonical transformations in the extended phase space [30], [31] [32], [35] as well as on definitions and consequences of "integrability" under such circumstances or generalizations thereof, even for non-Hamiltonian systems [11], [9], [14], [18] which we will not delve into further at this point.

2. Main results.

Consider the differential equation

$$\ddot{x} = f(x, t), \tag{2.1}$$

with particular solution $x = x(t)$. We will henceforth order our choice of positions as $q_1 = x$ and $q_2 = t$, thus yielding a Hamiltonian system given by

$$H = \frac{p_1^2}{2} - F(q_1, q_2), \quad F_{q_1}(q_1, q_2) = \frac{\partial F(q_1, q_2)}{\partial q_1} = f(q_1, q_2).$$

Equation (2.1) is obviously equivalent to Hamilton's equations for H,

$$\dot{q}_1 = p_1 = H_{p_1}, \qquad \dot{p}_1 = -H_{q_1} = f(q_1, q_2);$$

this nonautonomous Hamiltonian system is included as a subsystem of $X_{\widehat{H}}$ linked to $\widehat{H} := H + p_2$, such as in (1.7). Assuming $x(t) = q_1(t)$ to be a solution of (2.1) and $q_2(t) = t$, we obtain an integral curve $\Gamma = \{\boldsymbol{z}(t)\}$ of $\widehat{H}$, where

$$\boldsymbol{z}(t) := (q_1(t), q_2(t), p_1(t), p_2(t)) = (q_1(t), t, \dot{q}_1(t), -H(t)).$$

We may now introduce our first main result.

Theorem 2.1. *Let Γ be an integral curve of $X_{\widehat{H}}$ such as the one introduced above. If $X_{\widehat{H}}$ is integrable by means of rational or meromorphic first integrals, then the Galois group of*

$$\ddot{\xi} = (f_{q_1}(q_1, q_2)|_\Gamma)\,\xi \tag{2.2}$$

is virtually abelian.

Proof. The Hamiltonian field $X_{\widehat{H}}$ is given by $X_{\widehat{H}} = (p_1, 1, f(q_1, q_2), F_{q_2}(q_1, q_2))^T$. The variational equation VE_Γ along Γ is

$$\frac{d}{dt}\begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \\ \xi_4 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ f_{q_1}(q_1, q_2) & f_{q_2}(q_1, q_2) & 0 & 0 \\ F_{q_1 q_2}(q_1, q_2) & F_{q_2 q_2}(q_1, q_2) & 0 & 0 \end{pmatrix}\Bigg|_\Gamma \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \\ \xi_4 \end{pmatrix}. \tag{2.3}$$

More precisely, (2.3) is

$$\text{(2.4)}\qquad \begin{cases} \dot{\xi}_1 &= \xi_3, \\ \dot{\xi}_2 &= 0, \\ \dot{\xi}_3 &= (f_{q_1}(q_1,q_2)|_\Gamma)\,\xi_1 + (f_{q_2}(q_1,q_2)|_\Gamma)\,\xi_2, \\ \dot{\xi}_4 &= (F_{q_1q_2}(q_1,q_2)|_\Gamma)\,\xi_1 + (F_{q_2q_2}(q_1,q_2)|_\Gamma)\,\xi_2. \end{cases}$$

Hence $\xi_2 = k$, where k is constant. Assuming $k = 0$, the *normal variational equations* ((NVE$_\Gamma$); see [8, sect. 1], [25, sect. 4.3], [22, sect. 4.1.3]) for $\widehat{H}$ are given by

$$\text{(2.5)}\qquad \frac{d}{dt}\begin{pmatrix} \xi_1 \\ \xi_3 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ f_{q_1}(q_1,q_2)|_\Gamma & 0 \end{pmatrix}\begin{pmatrix} \xi_1 \\ \xi_3 \end{pmatrix},$$

and solving (2.5) we can obtain ξ_4. System (2.5) is equivalent to (2.2), where $\xi = \xi_1$. By virtue of [22, Proposition 4.2], the virtual commutativity of Gal (VE$_\Gamma$) implies that of Gal (NVE$_\Gamma$); this, coupled with Theorem 1.2, implies that if $X_{\widehat{H}}$ is rationally or meromorphically integrable, then the Galois group of (2.2) is virtually abelian. ■

Corollary 2.2. *Suppose the Galois group of the differential equation $\ddot{\xi} = k(t)\xi$ is not virtually abelian. Define $H := \frac{p_1^2}{2} - k(q_2)\frac{q_1^2}{2}$ and $\widehat{H} := H + p_2$. Then $X_{\widehat{H}}$ is not integrable by means of meromorphic or rational first integrals.*

Remark 1. *If the Galois group of the equation $\ddot{\xi} = k(t)\xi$ is the Borel group $G = \mathbb{C}^* \ltimes \mathbb{C}$ (hence connected, solvable, and nonabelian), then $X_{\widehat{H}}$ is neither meromorphically nor rationally integrable, although it is still possible to solve the equation as well as, ostensibly, the nonautonomous Hamiltonian system X_H.*

In anticipation of the following corollary, consider, for any $g(x)$, $a(t)$, and $\alpha = \alpha_0$ given, the equation

$$\text{(2.6)}\qquad \ddot{x} = g_x(x)(g(x) + a(t)) + \alpha, \quad \alpha \in \mathbb{C},$$

having a certain known particular solution $q_1 = q_1(t)$. Let

$$H = \frac{p_1^2}{2} - \frac{(g(q_1) + a(q_2))^2}{2} - \alpha q_1, \quad q_2 = t,$$

be a Hamiltonian linked to (2.6) and $\widehat{H} := H + p_2$ its autonomous completion.

Corollary 2.3. *If $X_{\widehat{H}}$ is integrable through rational or meromorphic first integrals, then, along the integral curve $\Gamma = \{\boldsymbol{z}(t) = (q_1(t), t, \dot{q}_1(t), -H(t))\}$, the Galois group of the equation*

$$\text{(2.7)}\qquad \ddot{\xi} = \big(g_{q_1}^2(q_1) + g_{q_1q_1}(q_1)(g(q_1) + a(q_2))\big)\big|_\Gamma\,\xi$$

is virtually abelian.

Now, keeping the above hypotheses for $g(x)$, $a(t)$, $\alpha = \alpha_0$, (2.6), and $x = q_1(t)$, let us define the Hamiltonian system

$$\text{(2.8)}\qquad \widehat{H} = H + p_2, \quad H = \frac{p_1^2}{2} - (g(q_1) + a(q_2))p_1 - (\alpha + a_{q_2}(q_2))q_1, \quad \alpha \in \mathbb{C}.$$

It takes only the following simple calculation to prove that Hamiltonian H is indeed linked to (2.6) in the manner expected, i.e., as introduced in the beginning of section 2. We have

$$\begin{aligned}\ddot{x} &= g_x(x)(g(x)+a(t))+\alpha \\ &= g_x(x)\dot{x}+g_x(x)g(x)+g_x(x)a(t)+\alpha+\dot{a}(t)-g_x(x)\dot{x}-\dot{a}(t) \\ &= g_x(x)y+\alpha+\dot{a}(t)-\tfrac{d}{dt}(g(x)+a(t)),\end{aligned}$$

where $y=\dot{x}+g(x)+a(t)$, and requiring x and y to be conjugate variables implies the following system, equivalent to (2.6):

$$\begin{aligned}\dot{x} &= y-(g(x)+a(t)) &&= H_y, \\ \dot{y} &= g_x(x)y+\alpha+\dot{a}(t) &&= -H_x.\end{aligned}$$

A straightforward parallel integration,

$$H=\frac{y^2}{2}-(g(x)+a(t))\,y+h_1(x,t)=h_2(y,t)-g(x)y-(\alpha+\dot{a}(t))x,$$

along with the definition of

$$h_1(x,t)=-(\alpha+\dot{a}(t))x, \quad h_2(y,t)=\frac{y^2}{2}-a(t)y,$$

as well as $(q_1,q_2,p_1)=(x,t,y)$, yields the autonomous Hamiltonian system introduced in (2.8).

Theorem 2.4. *If $X_{\widehat{H}}$ is integrable by means of rational or meromorphic first integrals, then, along $\Gamma=\{\boldsymbol{z}\,(t)=(q_1(t),t,p_1(t),-H(t))\}$, the Galois group of the equation*

$$\ddot{\xi}=\left(g_{q_1}^2(q_1)+p_1g_{q_1q_1}(q_1)-g_{q_2q_1}(q_1)\right)\big|_\Gamma\,\xi \tag{2.9}$$

is virtually abelian.

Proof. We may proceed as in the proof of Theorem 2.1. The Hamiltonian field $X_{\widehat{H}}$ is given by

$$X_{\widehat{H}}=\begin{pmatrix} p_1-(g(q_1)+a(q_2)) \\ 1 \\ g_{q_1}(q_1)p_1+(\alpha+a_{q_2}(q_2)) \\ a_{q_2}(q_2)p_1+a_{q_2q_2}(q_2)q_1 \end{pmatrix}.$$

The variational equation VE_Γ along $\Gamma=\{(q_1(t),t,p_1(t),-H(t))\}$ is

$$\dot{\boldsymbol{\xi}}=\left.\begin{pmatrix} -\frac{\partial g(q_1)}{\partial q_1} & -\frac{\partial a(q_2)}{\partial q_2} & 1 & 0 \\ 0 & 0 & 0 & 0 \\ p_1\frac{\partial^2 g(q_1)}{\partial q_1^2} & \frac{\partial^2 a(q_2)}{\partial q_2^2} & \frac{\partial g(q_1)}{\partial q_1} & 0 \\ \frac{\partial^2 a(q_2)}{\partial q_2^2} & \left(p_1\frac{\partial^2 a(q_2)}{\partial q_2^2}+q_1\frac{\partial^3 a(q_2)}{\partial q_2^3}\right) & \frac{\partial a(q_2)}{\partial q_2} & 0 \end{pmatrix}\right|_\Gamma \boldsymbol{\xi}, \tag{2.10}$$

where $\boldsymbol{\xi} = (\xi_1, \xi_2, \xi_3, \xi_4)^T$. $\xi_2 \equiv k \in \mathbb{C}$; in particular, $k = 0$ renders system (2.10) equal to the following:

$$\left\{\begin{array}{rcl} \dot{\xi}_1 &=& -g_{q_1}(q_1)|_\Gamma\, \xi_1 + \xi_3, \\ \dot{\xi}_2 &=& 0, \\ \dot{\xi}_3 &=& p_1 g_{q_1 q_1}(q_1)|_\Gamma\, \xi_1 + -g_{q_1}(q_1)|_\Gamma\, \xi_3, \\ \dot{\xi}_4 &=& a_{q_2 q_2}(q_2)|_\Gamma\, \xi_1 + a_{q_2}(q_2)|_\Gamma\, \xi_3, \end{array}\right. \tag{2.11}$$

NVE_Γ corresponding to

$$\frac{d}{dt}\begin{pmatrix} \xi_1 \\ \xi_3 \end{pmatrix} = \begin{pmatrix} -g_{q_1}(q_1) & 1 \\ p_1 g_{q_1 q_1}(q_1) & g_{q_1}(q_1) \end{pmatrix}\Bigg|_\Gamma \begin{pmatrix} \xi_1 \\ \xi_3 \end{pmatrix}. \tag{2.12}$$

Solving (2.12), we obtain ξ_4. Using (1.2) and (1.3) system (2.12) is equivalent to (2.9), where $\xi = \xi_1$. Again, by virtue of [22, Proposition 4.2] as in Theorem 2.1, the integrability of Hamiltonian $\widehat{H}$ by means of meromorphic or rational first integrals implies the virtual commutativity of the Galois group of (2.9). ■

Remark 2. *We observe that if $\ddot{x} = f(x,t)$, and if the same particular solution $x = x(t)$ occurs in the statements of Theorem 2.1, Corollary 2.3, and Theorem 2.4, then we obtain the same* NVE_Γ *((2.2), (2.7), and (2.9) are equivalent), despite the fact that their respective linked Hamiltonian systems have different expressions.*

3. Examples. In this section, and in the application of Theorems 2.1 and 2.4 as well as of Corollaries 2.2 and 2.3, we analyze the nonintegrability of the Hamiltonian systems corresponding to the following differential equations:

1. Hill–Schrödinger equation: $\ddot{x} = k(t)x$.
2. Sitnikov problem: $\ddot{x} = -\frac{(1-e\cos t)x}{(x^2+r^2(t))^{\frac{3}{2}}}$, $r(t) = \frac{1-e\cos t}{2}$.
3. Painlevé II equation: $\ddot{x} = 2x^3 + tx + \alpha$.
4. An algebraic toy model: $\ddot{x} = -\frac{1}{4x^3} - \frac{t}{x^2} + \alpha$.

In order to analyze normal variational equations, a standard procedure is using Maple, and especially the commands `dsolve` and `kovacicsols`. Whenever the command `kovacicsols` yields an output "`[ ]`," it means that the second-order linear differential equation being considered has no Liouvillian solutions, and thus its Galois group is virtually nonsolvable. For equations of the form $\ddot{y} = ry$ with $r \in \mathbb{C}(x)$, the only virtually nonsolvable group is $\mathrm{SL}_2(\mathbb{C})$. In some cases, moreover, `dsolve` makes it possible to obtain the solutions in terms of special functions such as *Airy functions*, *Bessel functions*, and *hypergeometric functions*, among others [2]. There is a number of second-order linear equations whose coefficients are not rational and whose solutions Maple cannot find by means of the commands `dsolve` and `kovacicsols` alone; this problem, in some cases, can be solved by the stated algebrization procedure.

3.1. Hill–Schrödinger equation $\ddot{x} = k(t)x$. This example corresponds to Corollary 2.2 for $k > 0$, $\epsilon \gg 0$, $\mathcal{P}_n$ a polynomial of degree n with $\mathcal{P}_n(0) = 1$, and $k(t)$ given by

$$\begin{array}{ll} k(t) = k\mathrm{e}^{-\epsilon t}, & k(t) = k\mathcal{P}_n(\epsilon t), \\ k(t) = k(1+\sinh(\epsilon t)), & k(t) = k(1+\sin(\epsilon t)), \\ k(t) = k(1+\cosh(\epsilon t)), & k(t) = k(1+\cos(\epsilon t)). \end{array}$$

$X_{\widehat{H}}$ is nonintegrable by means of rational first integrals.

The integrability of equation $\ddot{x} = k(t)x$ for these examples has been deeply analyzed in [3].

3.2. The Sitnikov problem. The *Sitnikov problem* is a symmetrically configured restricted three-body problem in which two primaries with equal masses move in ellipses of eccentricity e in a plane π_1, and an infinitesimal point mass moves along the line $\pi_1^{\perp}$. See [6], [12], [25], [27], [37], [38], [16] for more details. The motion of the infinitesimal point mass is given by the differential equation

$$\ddot{z} + \frac{z}{\left(r^2(t) + z^2\right)^{3/2}} = 0, \tag{3.1}$$

where $z = z(t)$ is the distance from the infinitesimal mass point to the plane of the primaries and $r(t)$ is half the distance of the primaries,

$$r(t) = \frac{1 - e\cos E(t)}{2},$$

where the *eccentric anomaly* $E(t)$ is the solution of the Kepler equation

$$E = t + e\sin E \tag{3.2}$$

and e is the eccentricity of the ellipses described by the primaries. We will assume $0 \leq e \leq 1$ through subsections 3.2 and 3.3. The Hamiltonian linked to the system is

$$H = \frac{v^2}{2} - \frac{1}{\left(z^2(t) + r^2(t)\right)^{1/2}}, \tag{3.3}$$

provided v stands for $\dot{z}$ in the corresponding equations.

Since (3.3) cannot be solved in explicit form, attempts at a Hamiltonian formulation of (3.1), whether exact or approximate, require one of at least two options: looking for an exact Hamiltonian formulation by means of a change of variables, which we will do in the next paragraph, and searching an approximate Hamiltonian formulation, which will be done in subsection 3.3.

Let us now find a Hamiltonian linked to (3.1). We may express $r(t)$ as

$$r(t) = (R \circ \varphi)(t) := \frac{1 - e^2}{2\left(1 + e\cos\varphi(t)\right)},$$

where φ, the *true anomaly*, is a solution of

$$\frac{d\varphi}{dt} = \frac{(1 + e\cos\varphi)^2}{(1 - e^2)^{3/2}} = \frac{\sqrt{1 - e^2}}{4R^2(\varphi)},$$

and we may follow the procedure introduced in [38] (see also [16]) by taking φ as the new independent variable and $x = \frac{z}{2r(\varphi)}$ as the new dependent variable. Writing t once again to denote φ, we have the following differential equation:

$$\ddot{x} = f(x,t) := -\frac{e\cos t + \left(\frac{1}{4} + x^2\right)^{-3/2}}{1 + e\cos t}x, \tag{3.4}$$

clearly amenable to the hypotheses in Theorem 2.1 and in the first paragraph of section 2. Defining $q_1 = x$, $q_2 = t$, and $p_1 = \dot{x}$, the autonomous Hamiltonian system corresponding to (3.4) is given by

$$\widehat{H}_e = H_e + p_2 := \frac{p_1^2}{2} + \frac{eq_1^2 \cos q_2 - 4\left(1 + 4q_1^2\right)^{-1/2}}{2\left(1 + e\cos q_2\right)} + p_2, \tag{3.5}$$

always assuming $e \in [0, 1]$.

The *circular* Sitnikov problem $\widehat{H}_0$ is meromorphically integrable in the sense of Liouville–Arnold and can be solved using elliptic integrals. The nonintegrability for $e = 1$ was first studied by means of straight Morales–Ramis theory in [26, sect. 5] (see also [22, sect. 5.3]); we will now extend the proof of meromorphic nonintegrability therein to one for every $0 < e \leq 1$ by using Theorem 2.1.

The NVE_Γ derived from the Hamiltonian system given by (3.5) is

$$\ddot{\xi} = \left(\frac{e\left(4q_1^2 + 1\right)^{5/2}\cos q_2 - 64q_1^2 + 8}{\left(e\cos q_2 + 1\right)\left(4q_1^2 + 1\right)^{5/2}}\right)\xi.$$

Taking $q_1 \equiv 0$ we have a solution $\Gamma = \left\{\boldsymbol{z}\left(t\right) = \left(0, t, 0, \frac{2}{1+e\cos t}\right)\right\}$ of $X_{\widehat{H}_e}$ along which the NVE_Γ is given by

$$\ddot{\xi} = \left(\frac{e\cos t + 8}{e\cos t + 1}\right)\xi,$$

which is algebrizable, through the change of variable $\tau = \cos t$, into

$$\frac{d^2\xi}{d\tau^2} - \left(\frac{\tau}{1-\tau^2}\right)\frac{d\xi}{d\tau} - \left(\frac{e\tau + 8}{(e\tau + 1)(1-\tau^2)}\right)\xi = 0. \tag{3.6}$$

This equation can be transformed into the differential equation

$$\frac{d^2\zeta}{d\tau^2} = \left(\frac{5e\tau^3 + 33\tau^2 - 2e\tau - 30}{(e\tau + 1)4(1-\tau^2)^2}\right)\zeta \tag{3.7}$$

by means of $\xi = \frac{\zeta}{\sqrt[4]{1-\tau^2}}$. Equations (3.6) and (3.7) are integrable in terms of Liouvillian solutions only if $e = 0$, whereas for $e \neq 0$ their solutions are given in terms of nonintegrable Heun functions if $e \neq 1$ and nonintegrable hypergeometric functions if $e = 1$, i.e., their Galois groups are virtually nonsolvable and hence virtually nonabelian; furthermore, they have a regular singularity at infinity, which by Theorem 1.2 implies the nonintegrability of $X_{\widehat{H}}$ for $e \in (0, 1]$ by means of *meromorphic* first integrals. In particular, the Galois group of (3.7) is exactly $\mathrm{SL}_2\left(\mathbb{C}\right)$.

3.2.1. Numerical results for the Sitnikov problem. The author is indebted to Sergi Simon in what concerns the following subsection, including the figures. Acknowledgments are also due to Carles Simó for further specific suggestions.

Let $\Sigma = \{\sin\left(q_2\right) = 0\}$. The six figures in this paper show Poincaré sections of the flow with respect to Σ, projected on the (q_1, p_1) plane, for the Hamiltonian system X_{H_e} obtained

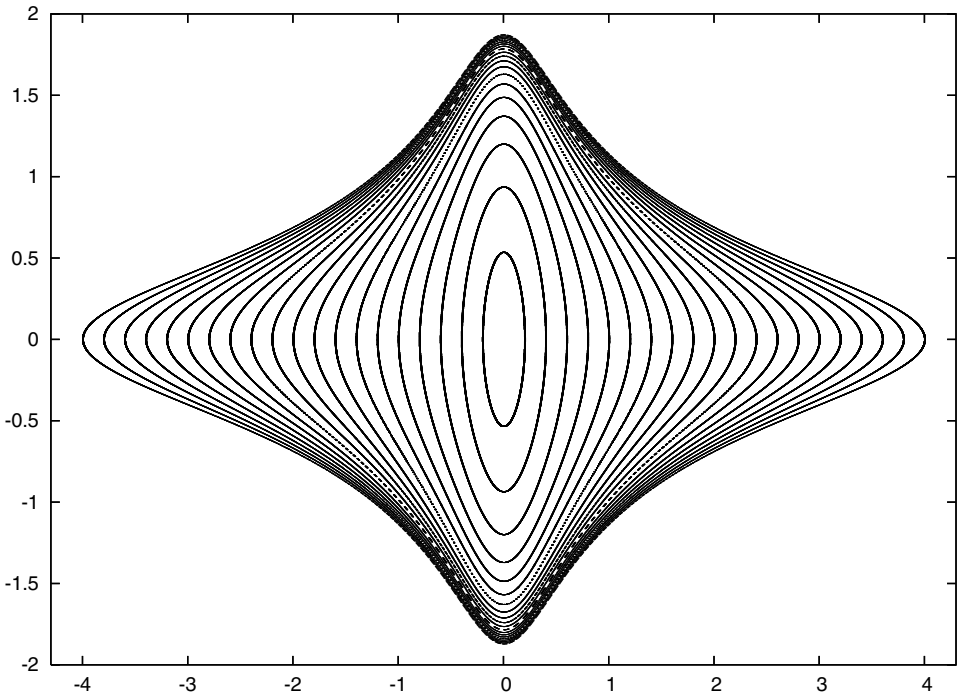


Figure 1. *Poincaré section* $\sin q_2 = 0$ *for the Sitnikov problem:* $e = 0$ *(figure: Sergi Simon).*

from (3.5). As may be easily deduced from said Hamiltonian, all sections are symmetrical with respect to the q_1 and p_1 axes. Different numbers of initial conditions are used for the sake of clarity.

Figure 1 corresponds to $e = 0$. In keeping with what was said after (3.5), the whole subset of Σ transversal to the flow sheds concentric tori (ostensibly, the intersections of the invariant Liouville–Arnold tori with Σ), a typical sign of integrability; the tori shown are only a selection of those therein, as the actual area foliated by them is larger.

A number of these invariant tori break down upon the slightest increase in e, and in the ensuing figures the two most interesting features are those invariant sets (usually called *KAM tori*) whose intersection with Σ prevails in the form of Jordan curves, and the zones of chaotic behavior between them. Sparse zones of the section will account for chaotic zones as well for $e > 0$. Figures 2 and 3 show two different close-up views for the Poincaré section corresponding to $e = 0.01$. The latter figure is actually a detail of the "island" of tori appearing at the right of the general section. For $e = 0.1$, Figures 4 and 5 are, respectively, a general view of the section and a close-up of one of the islands appearing at each side of the central area. As for $e = 0.4$, Figure 6 is an enlarged view of one of the two islands appearing at each side of a central area.

3.3. The approximate Sitnikov problem. As said in subsection 3.2, we now consider an approximation of the Sitnikov problem; see [15] and [12] for more details. As opposed to *meromorphic* nonintegrability, we will prove nonintegrability by means of *rational* first integrals.

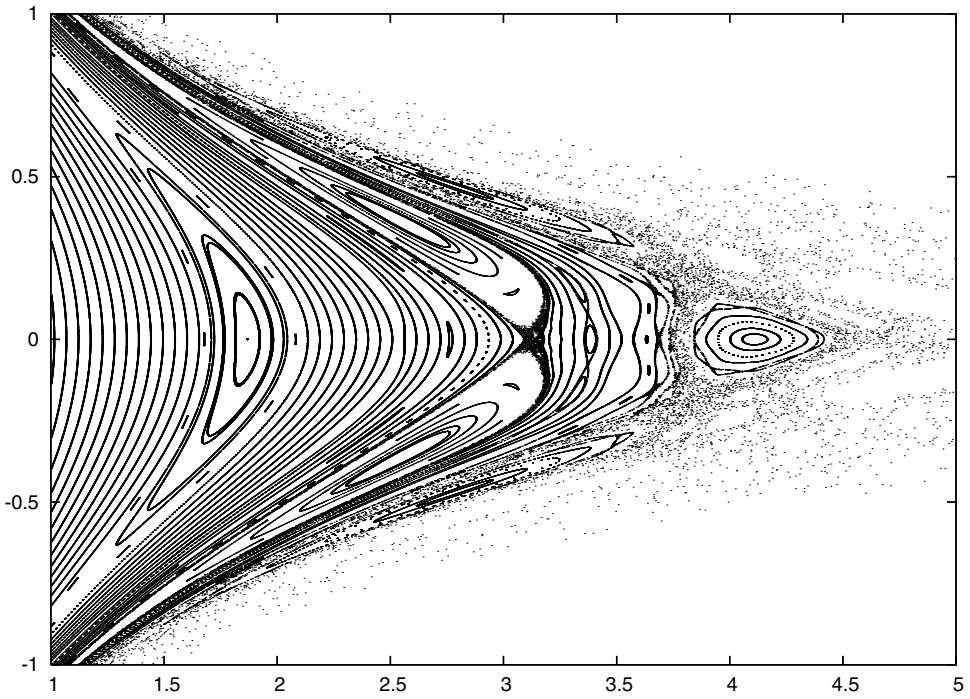


Figure 2. *Poincaré section* $\sin q_2 = 0$ *for the Sitnikov problem,* $e = 0.01$ *(figure: Sergi Simon).*

The fact that $\varphi(t) = t + O(e)$ yields $r(t) = \frac{1-e\cos t}{2} + O\left(e^2\right)$, and thus the Hamiltonian in (3.3) becomes

$$H = \frac{v^2}{2} - \frac{1}{\sqrt{z^2 + \frac{1}{4}}} - e\frac{\cos t}{\left(z^2 + \frac{1}{4}\right)^{3/2}} + O\left(e^2\right)$$

whenever $e \approx 0$. In particular, the first-order approximation of this asymptotic expansion in e yields the Hamiltonian

$$H = \frac{v^2}{2} - \frac{1}{\sqrt{z^2 + \frac{1}{4}}} - e\frac{\cos t}{\left(z^2 + \frac{1}{4}\right)^{3/2}}. \tag{3.8}$$

Considering $q_1 = x$, $q_2 = t$, and $p_1 = v$, the autonomous Hamiltonian system corresponding to this equation is given by

$$\widehat{H} = H + p_2, \quad H = \frac{p_1^2}{2} - e\frac{2\cos q_2}{(4q_1^2+1)^{3/2}} - \frac{2}{\sqrt{4q_1^2+1}}, \tag{3.9}$$

corresponding to the Hamiltonian in Theorem 2.1:

$$f(x,t) = -\frac{8x}{(4x^2+1)^{\frac{3}{2}}} - e\frac{24x\cos t}{(4x^2+1)^{\frac{5}{2}}}.$$

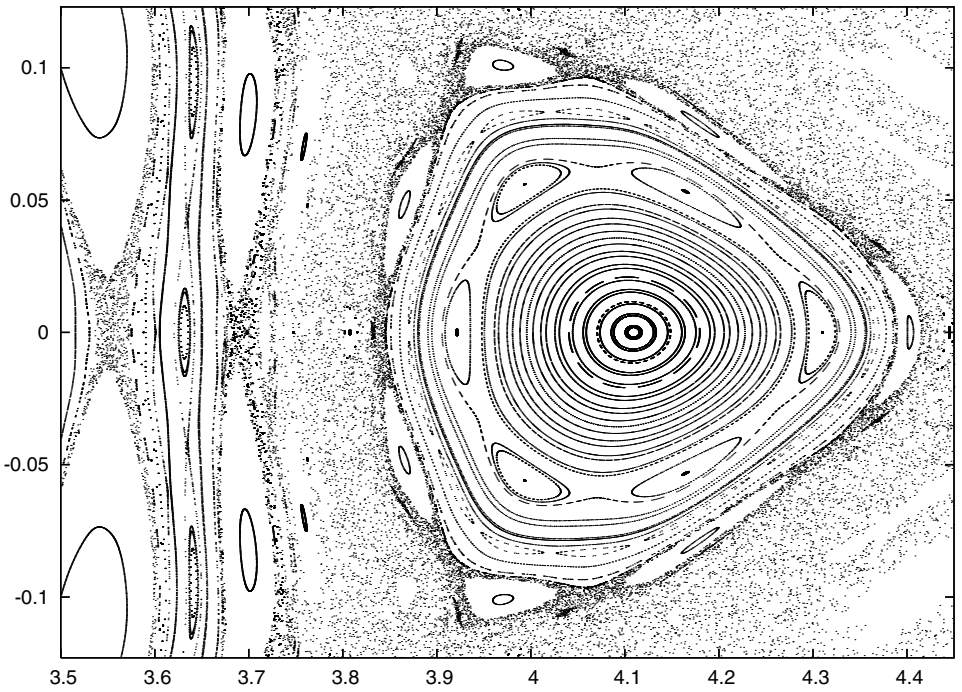


Figure 3. *Poincaré section for* $e = 0.01$ *(figure: Sergi Simon).*

The NVE_Γ for the Hamiltonian (3.9) is given by

$$\ddot{\xi} = \left(e\frac{24(16q_1^2 - 1)\cos q_2}{(4q_1^2 + 1)^{\frac{7}{2}}} + \frac{8(8q_1^2 - 1)}{(4q_1^2 + 1)^{\frac{5}{2}}} \right) \xi.$$

Its general solution may be expressed as

$$\xi\,(t) = K_1 C\left(32, 48e, \frac{t}{2}\right) + K_2 S\left(32, 48e, \frac{t}{2}\right),$$

where the *Mathieu even* (resp., *odd*) *function* $C\,(a, q, t)$ (resp., $S\,(a, q, t)$) is defined as the even (resp., odd) solution to $\ddot{y} + (a - 2q\cos(2t))y = 0$ [2, Ch. 20].

Taking $q_1(t) = 0$, we can see that $\boldsymbol{z}(t) = (0, t, 0, 2e\cos t + 2)$; hence, defining $\boldsymbol{z}(t) = (0, t, 0, 2e\cos t + 2)$ and $\Gamma = \{\boldsymbol{z}(t)\}$, the operator linked to NVE_Γ is given by

$$\ddot{\xi} = (-24e\cos t - 8)\,\xi,$$

which is algebrizable (see Theorem 1.1) through the change of variables $\tau = \cos t$ into

$$\frac{d^2\xi}{d\tau^2} - \left(\frac{\tau}{1 - \tau^2}\right)\frac{d\xi}{d\tau} + \left(\frac{24e\tau + 8}{1 - \tau^2}\right)\xi = 0. \tag{3.10}$$

Now, this equation can be transformed into the differential equation

$$\frac{d^2\zeta}{d\tau^2} = \left(\frac{96e\tau^3 + 31\tau^2 - 96e\tau - 34}{4(1 - \tau^2)^2}\right)\zeta, \quad \xi = \frac{\zeta}{\sqrt[4]{1 - \tau^2}}. \tag{3.11}$$

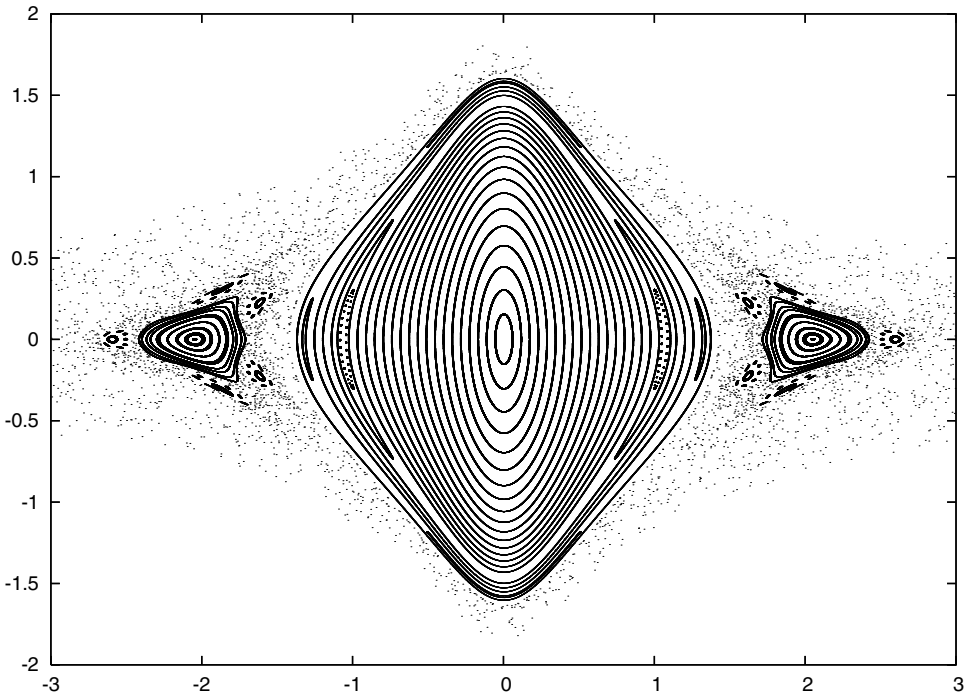


Figure 4. *Poincaré section* $\sin q_2 = 0$ *for the Sitnikov problem,* $e = 0.1$ *(figure: Sergi Simon).*

Equations (3.10) and (3.11) are integrable in terms of Liouvillian solutions only if $e = 0$, since for $e \neq 0$ their solutions are given in terms of nonintegrable Mathieu functions; hence their Galois groups are virtually nonsolvable and thus virtually nonabelian; furthermore, they have an irregular singularity at infinity, implying rational nonintegrability for the Hamiltonian field $X_{\widehat{H}}$ with $\alpha = 1$ by virtue of Theorem 1.2. In particular, the Galois group of (3.11) is exactly $\mathrm{SL}_2(\mathbb{C})$.

Equations (3.10) and (3.11) have been deeply analyzed in [3] using the Hamiltonian change of variables $\tau = e^{it}$, obtaining the same result presented here.

3.4. Painlevé II equation: $\ddot{x} = 2x^3 + tx + \alpha$. The nonintegrability of the second Painlevé equation for integer α was proved by Morales-Ruiz in [23] and later by Stoyanova and Christov in [29]. They used only the Hamiltonian (3.14).

Defining $q_1 = x$, $q_2 = t$, $p_1 = y$, and $\alpha \in \mathbb{C}$, the autonomous Hamiltonian system corresponding to this equation can be given by any of the following three functions:

$$\widehat{H} = H + p_2, \quad H = \frac{p_1^2}{2} - \frac{q_1^4}{2} - q_2\frac{q_1^2}{2} - \alpha q_1, \tag{3.12}$$

$$\widehat{H} = H + p_2, \quad H = \frac{p_1^2}{2} - \frac{1}{2}\left(q_1^2 + \frac{q_2}{2}\right)^2 - \alpha q_1, \tag{3.13}$$

$$\widehat{H} = H + p_2, \quad H = \frac{p_1^2}{2} - \left(q_1^2 + \frac{q_2}{2}\right)p_1 - \left(\alpha + \frac{1}{2}\right)q_1, \tag{3.14}$$

where (3.12), (3.13), and (3.14) correspond to the Hamiltonian of Theorem 2.1 ($f(x,t) =$

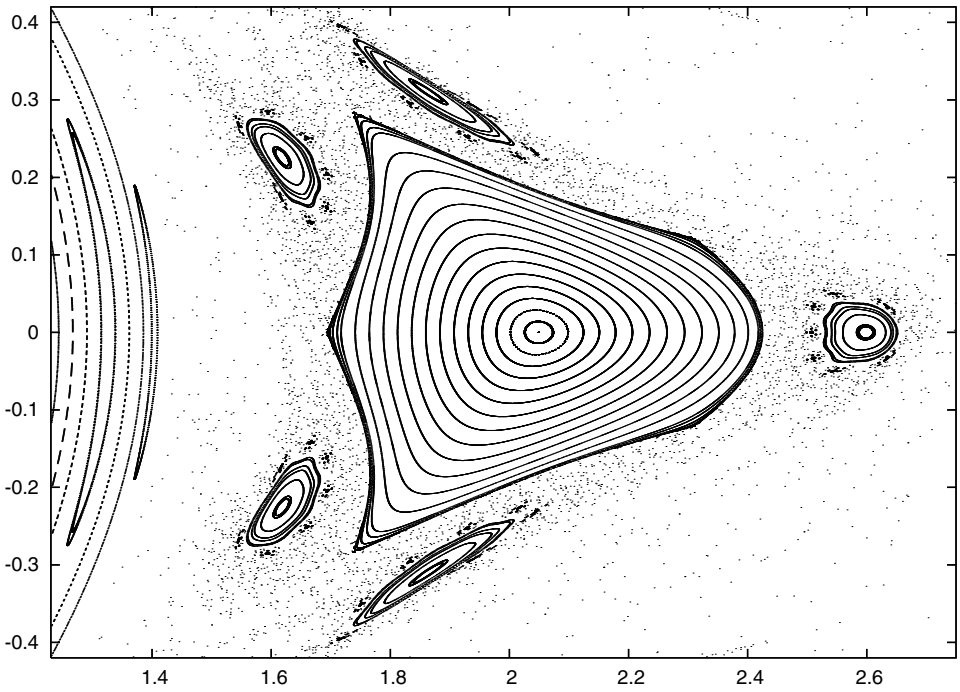


Figure 5. *Poincaré section for* $e = 0.1$ *(figure: Sergi Simon).*

$2x^3 + tx + \alpha$), Corollary 2.3 ($g(x) = x^2$ and $a(t) = t/2$), and Theorem 2.4 ($g(x) = x^2$ and $a(t) = t/2$), respectively. The Hamiltonian system for Painlevé II, studied in [23, 29], corresponds precisely to Hamiltonian (3.14). The NVE_Γ for these Hamiltonians is given by

$$\ddot{\xi} = \left(6q_1^2 + q_2\right)\xi.$$

Taking $\alpha = 0$ and $q_1(t) = 0$, we have particular solutions $\boldsymbol{z}(t) = (0, t, 0, 0)$, $\boldsymbol{z}(t) = \left(0, t, 0, t^2/8\right)$, and $\boldsymbol{z}(t) = \left(0, t, t/2, t^2/8\right)$, respectively, for the Hamiltonians (3.12), (3.13), and (3.14); hence NVE_Γ is given by $\ddot{\xi} = t\xi$, the so-called *Airy equation* [2, sect. 10.4.1], which has an irregular singularity at infinity and is not integrable through Liouvillian solutions; i.e., its Galois group is $\mathrm{SL}_2(\mathbb{C})$ and not virtually abelian; thus, by Theorem 1.2, the Hamiltonian field $X_{\widehat{H}}$ with $\alpha = 0$ is not integrable through rational first integrals.

Now, for $\alpha = 1$ and $q_1(t) = -1/t$, the integral curve $\boldsymbol{z}(t)$ is given by

$$\left(-\frac{1}{t}, t, \frac{1}{t^2}, -\frac{1}{2t}\right), \quad \left(-\frac{1}{t}, t, \frac{1}{t^2}, -\frac{1}{2t} + \frac{t^2}{8}\right), \quad \text{and} \quad \left(-\frac{1}{t}, t, \frac{2}{t^2} + \frac{t}{2}, -\frac{1}{t} + \frac{t^2}{8}\right),$$

respectively, for the Hamiltonians (3.12), (3.13), and (3.14), so that NVE_Γ is given by

$$\ddot{\xi} = \left(\frac{6}{t^2} + t\right)\xi, \quad \Gamma = \{\boldsymbol{z}(t)\},$$

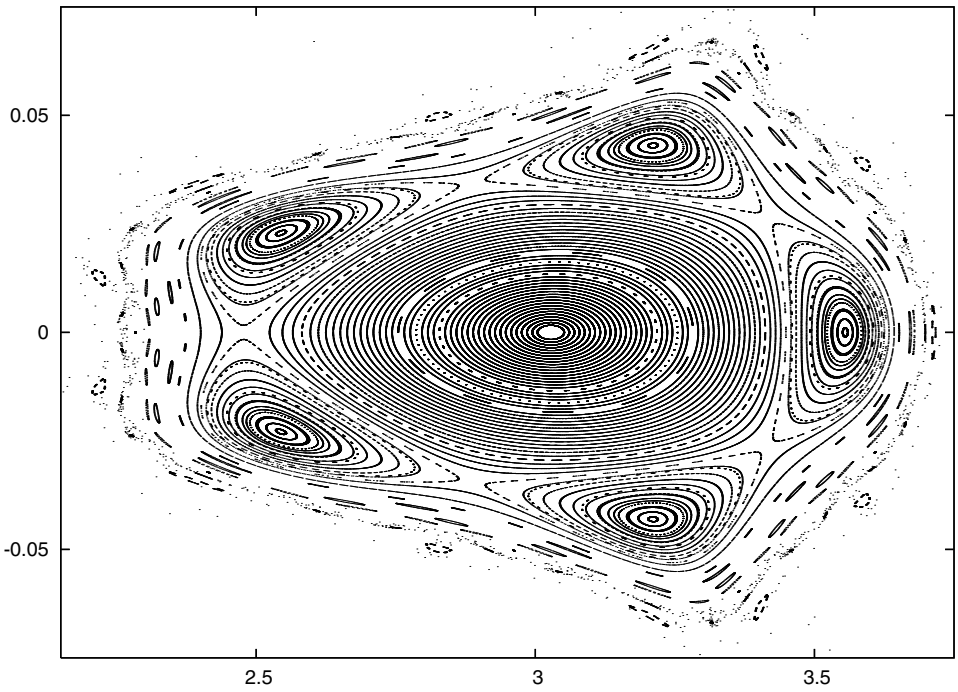


Figure 6. *Poincaré section for the Sitnikov problem, $e = 0.4$ (figure: Sergi Simon).*

whose general solution is

$$\xi(t) = \sqrt{t}\left[K_1 I_{-5/3}\left(\frac{2t^{3/2}}{3}\right) + K_2 I_{5/3}\left(\frac{2t^{3/2}}{3}\right)\right],$$

$I_\alpha = 2^{-\alpha}t^\alpha\left(\frac{1}{\Gamma(1+\alpha)} + \frac{t^2}{2^2\Gamma(2+\alpha)} + O(t^4)\right)$ being, for each α, the *modified Bessel function of the first kind*, i.e., the solution to $t^2\ddot{y} + t\dot{y} - (t^2+\alpha^2)\,y = 0$ [2, sect. 9.6].

The normal variational equation has an irregular singularity at infinity and is not integrable through Liouvillian functions because its solutions are given in term of nonintegrable Bessel functions (see [23, 29]); i.e., its Galois group is $\mathrm{SL}_2(\mathbb{C})$, which is not virtually abelian; again by the Morales–Ramis theorem, Theorem 1.2, the Hamiltonian field $X_{\widehat{H}}$ with $\alpha = 1$ is not integrable through rational first integrals.

3.5. An algebraic toy model: $\ddot{x} = -\frac{1}{4x^3} - \frac{t}{x^2} + \alpha$. Considering $q_1 = x$, $q_2 = t$, $p_1 = y$, and $\alpha \in \mathbb{C}$; the autonomous Hamiltonian systems corresponding to this equation are given by

$$\widehat{H} = H + p_2, \quad H = \frac{p_1^2}{2} - \frac{1}{8q_1^2} - \frac{q_2}{q_1} - \alpha q_1, \tag{3.15}$$

$$\widehat{H} = H + p_2, \quad H = \frac{p_1^2}{2} - \frac{1}{2}\left(\frac{1}{2q_1} + 2q_2\right)^2 - \alpha q_1, \tag{3.16}$$

$$\widehat{H} = H + p_2, \quad H = \frac{p_1^2}{2} + \left(\frac{1}{2q_1} + 2q_2\right)p_1 - (\alpha+2)\,q_1. \tag{3.17}$$

Equations (3.15), (3.16), and (3.17) correspond to Theorem 2.1 ($f(x,t) = -\frac{1}{4x^3} - \frac{t}{x^2} + \alpha$), Corollary 2.3 ($g(x) = -\frac{1}{2x}$ and $a(t) = -2t$), and Theorem 2.4 ($g(x) = -\frac{1}{2x}$ and $a(t) = -2t$), respectively. The NVE_Γ for all three is given by

$$\ddot{\xi} = \left(\frac{3}{4q_1^4} + \frac{2q_2}{q_1^3}\right)\xi.$$

Now, for $\alpha = 1$ and $q_1(t) = \sqrt{t}$, the integral curve $\boldsymbol{z}(t)$ is given by

$$\left(\sqrt{t}, t, \frac{1}{2\sqrt{t}}, 2\sqrt{t}\right), \quad \left(\sqrt{t}, t, \frac{1}{2\sqrt{t}}, 2t^2 + 2\sqrt{t}\right), \quad \text{and} \quad \left(\sqrt{t}, t, -2t, 2t^2\right),$$

respectively, for the Hamiltonians (3.15), (3.16), and (3.17), rendering NVE_Γ equal to

$$\ddot{\xi} = \left(\frac{3}{4t^2} + \frac{2}{\sqrt{t}}\right)\xi, \tag{3.18}$$

having a solution

$$\begin{aligned}\xi_1 &= -\frac{3t^{3/2}}{2}\,{}_0F_1\left(;\frac{7}{3};\frac{8t^{3/2}}{9}\right)\\ &= -\frac{3}{2}t^{3/2} - \frac{4}{7}t^3 - \frac{8}{105}t^{9/2} + O\left(t^6\right),\end{aligned}$$

${}_0F_1\left(;a;t\right) = \lim_{q\to\infty} {}_1F_1\left(q;a;\frac{t}{q}\right) = \sum_{n=0}^{\infty}\frac{t^n}{(a)_n n!}$ being the *confluent hypergeometric limit function* [2, Ch. 13], and an independent new solution $\xi_2 = \xi_1\int \xi_1^{-2}$, satisfying

$$\xi_2 = \frac{1}{3\sqrt{t}} - \frac{8t}{9} - \frac{16}{27}t^{5/2} + O\left(t^4\right).$$

As is the case for the rest of the normal variational operators appearing in this paper, our knowledge of the exponents around 0 of a fundamental set of solutions (in this case, ξ_1 and ξ_2), coupled with the basic result on factorization obtained in [7, Th. 8 (Ch. 5)] (see also [8, Criterion 1]), would suffice to prove nonintegrability. Here, however, we will restrict our attention to Theorems 2.1 and 2.4 and Corollary 2.3.

Equation (3.18) is algebrizable (Theorem 1.1), through the change of variables $\tau = \sqrt{t}$ into

$$\frac{d^2\xi}{d\tau^2} - \left(\frac{1}{\tau}\right)\frac{d\xi}{d\tau} - \left(\frac{8\tau^3+3}{\tau^2}\right)\xi = 0; \tag{3.19}$$

now, this equation can be transformed into the differential equation

$$\frac{d^2\zeta}{d\tau^2} = \left(\frac{32\tau^3+15}{4\tau^2}\right)\zeta, \quad \xi = \zeta\sqrt{\tau}. \tag{3.20}$$

Equations (3.19) and (3.20) have an irregular singularity at $t = \infty$ and are not integrable through Liouvillian solutions due to the presence of Bessel functions; i.e., their Galois groups are virtually nonsolvable and therefore virtually nonabelian, Theorem 1.2 once again settling rational nonintegrability for $\alpha = 1$. In particular, the Galois group of (3.20) is exactly $\mathrm{SL}_2(\mathbb{C})$.

4. Final remarks: Open questions and future work. This paper is the starting point of a project in which the author is involved. The following questions arose during our work:

- In [23, 29] it was proven that the autonomous Hamiltonian system related to Painlevé II is nonintegrable for every $\alpha \in \mathbb{Z}$. Is this also true for (2.6)?
- Does the integrability of (2.6) for arbitrary $\alpha \in \mathbb{Z}$ depend on the choice of $g(x)$ and $a(t)$?
- Assuming the above question has an affirmative answer, in what manner can the choice and form of $g(x)$ and $a(t)$ ensure nonintegrability for every $\alpha \in \mathbb{Z}$? And for every $\alpha \in \mathbb{C}$?
- Is it possible to find transversal sections of the flow, and thus Poincaré maps, for either $\widehat{H}$ or the algebrized equation, even in the absence of nontrivial numerical monodromies? Do Stokes multipliers contribute to the answer in a significant manner?

Among our next goals, the analysis of the following items is due further immediate research:

- the application of Morales–Ramis theory to higher variational equations of nonautonomous Hamiltonian systems;
- differential equations in the form $\ddot{x} = f(x, \dot{x}, t)$;
- the rest of the Painlevé equations: Casale in [10] analyzed Painlevé I, and Horozov and Stoyanova in [17] analyzed some particular cases for Painlevé VI;
- the theoretical aspects of nonautonomous Hamiltonian systems such as their geometry and the feasibility of an analogue to Liouville–Arnold theory;
- the nonintegrability of nonautonomous Hamiltonian systems with two and a half degrees of freedom;
- specific examples of nonautonomous Hamiltonian systems related to control theory, as well as others related to celestial mechanics, such as the restricted three- and four-body problems and Hénon–Heiles systems [4], [13].
- the exact relation, perhaps causal, between separatrix splitting [24] and nonintegrability, whether rational or meromorphic.

Acknowledgments. The author is indebted to Sergi Simon for multiple suggestions concerning organization, style, pictures, and numerical remarks, as well as for preliminary corrections concerning the Sitnikov problem; for his collaboration in the numerical analysis thereof, acknowledgments are also due to Carles Simó. The author acknowledges Juan J. Morales-Ruiz and Jacques-Arthur Weil for their comments and suggestions and for the final proofreading of this paper.

REFERENCES

[1] R. Abraham and J. E. Marsden, *Foundations of Mechanics*, 2nd ed., revised and enlarged, Benjamin/Cummings, Advanced Book Program, Reading, MA, 1978.

[2] M. Abramowitz and I. A. Stegun, eds., *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, John Wiley & Sons, New York, 1984. Reprint of the 1972 edition.

[3] P. B. Acosta-Humanez and D. Blázquez-Sanz, *Non-integrability of some Hamiltonians with rational potentials*, Discrete Contin. Dyn. Syst. Ser. B, 10 (2008), pp. 265–293.

[4] M. A. Andreu, *Dynamics in the center manifold around L_2 in the quasi-bicircular problem*, Celestial Mech. Dynam. Astronom., 84 (2002), pp. 105–133.

[5] V. I. Arnold, *Mathematical Methods of Classical Mechanics*, Graduate Texts in Mathematics 60, Springer-Verlag, New York, 1978.
[6] G. Benettin, L. Galgani, and J.-M. Strelcyn, *Kolmogorov entropy and numerical experiments*, Phys. Rev. A, 14 (1976), pp. 2338–2345.
[7] D. Boucher, *Sur les équations différentielles linéaires paramétrées; une application aux systèmes hamiltoniens*, Thèse de l'Université de Limoges, Limoges, France, 2000.
[8] D. Boucher and J.-A. Weil, *Application of J.-J. Morales and J.-P. Ramis' theorem to test the non-complete integrability of the planar three-body problem*, in From Combinatorics to Dynamical Systems, IRMA Lect. Math. Theor. Phys. 3, de Gruyter, Berlin, 2003, pp. 163–177.
[9] A. Dewisme, S. Bouquet, and P. G. L. Leach, *Symmetries of time dependent Hamiltonian systems*, in Modern Group Analysis: Advanced Analytical and Computational Methods in Mathematical Physics (Acireale, 1992), Kluwer, Dordrecht, The Netherlands, 1993, pp. 173–179.
[10] G. Casale, *Irréductibilité de la première équation de Painlevé*, C. R. Math. Acad. Sci. Paris, 343 (2006), pp. 95–98.
[11] A. Dewisme and S. Bouquet, *First integrals and symmetries of time-dependent Hamiltonian systems*, J. Math. Phys., 34 (1993), pp. 997–1006.
[12] D. Fernández Hevia and A. F. Rañada, *Chaos in the three-body problem: The Sitnikov case*, European J. Phys., 17 (1996), pp. 295–302.
[13] F. Gabern, À. Jorba, and P. Robutel, *On the accuracy of restricted three-body models for the Trojan motion*, Discrete Contin. Dyn. Syst., 11 (2004), pp. 843–854.
[14] G. Giachetta, L. Mangiarotti, and G. Sardanashvily, *Action-angle coordinates for time-dependent completely integrable Hamiltonian systems*, J. Phys. A, 35 (2002), pp. L439–L445.
[15] M. Gidea and F. Deppe, *Chaotic orbits in a restricted three-body problem: Numerical experiments and heuristics*, Commun. Nonlinear Sci. Numer. Simul., 11 (2006), pp. 161–171.
[16] J. Hagel and T. Trenkler, *A computer-aided analysis of the Sitnikov problem*, Celestial Mechanics and Dynamical Astronomy, 56 (1993), pp. 81–98.
[17] E. Horozov and T. Stoyanova, *Non-integrability of some Painlevé* VI*-equations and dilogarithms*, Regul. Chaotic Dyn., 12 (2007), pp. 622–629.
[18] R. S. Kaushal, *Construction of exact invariants for time dependent classical dynamical systems*, Internat. J. Theoret. Phys., 37 (1998), pp. 1793–1856.
[19] R. Kuwabara, *Time-dependent mechanical symmetries and extended Hamiltonian systems*, Rep. Math. Phys., 19 (1984), pp. 27–38.
[20] C. Lanczos, *The Variational Principles of Mechanics*, Mathematical Expositions 4, University of Toronto Press, Toronto, Ontario, Canada, 1949.
[21] A. J. Lichtenberg and M. A. Lieberman, *Regular and Chaotic Dynamics*, 2nd ed., Applied Mathematical Sciences 38, Springer-Verlag, New York, 1992.
[22] J. J. Morales-Ruiz, *Differential Galois Theory and Non-Integrability of Hamiltonian Systems*, Progress in Mathematics 179, Birkhäuser Verlag, Basel, 1999.
[23] J. J. Morales-Ruiz, *A remark about the Painlevé transcendents*, in Théories asymptotiques et équations de Painlevé, Sémin. Congr. 14, Soc. Math. France, Paris, 2006, pp. 229–235.
[24] J. J. Morales-Ruiz and J. M. Peris, *On a Galoisian approach to the splitting of separatrices*, Ann. Fac. Sci. Toulouse Math. (6), 8 (1999), pp. 125–141.
[25] J. J. Morales-Ruiz and J. P. Ramis, *Galoisian obstructions to integrability of Hamiltonian systems* I, Methods Appl. Anal., 8 (2001), pp. 33–96.
[26] J. J. Morales-Ruiz and J. P. Ramis, *Galoisian obstructions to integrability of Hamiltonian systems* II, Methods Appl. Anal., 8 (2001), pp. 97–112.
[27] J. Moser, *Stable and Random Motions in Dynamical Systems. With Special Emphasis on Celestial Mechanics*, Princeton Landmarks in Mathematics, Princeton University Press, Princeton, NJ, 2001. Reprint of the 1973 original.
[28] E. L. Stiefel and G. Scheifele, *Linear and Regular Celestial Mechanics. Perturbed Two-Body Motion, Numerical Methods, Canonical Theory*, Grundlehren Math. Wiss. 174, Springer-Verlag, New York, 1971.
[29] T. Stoyanova and O. Christov, *Non-integrability of the second Painlevé equation as a Hamiltonian system*, C. R. Acad. Bulgare Sci., 60 (2007), pp. 13–18.

[30] J. Struckmeier, *Hamiltonian dynamics on the symplectic extended phase space for autonomous and non-autonomous systems*, J. Phys. A, 38 (2005), pp. 1257–1278.

[31] J. Struckmeier and C. Riedel, *Canonical transformations and exact invariants for time-dependent Hamiltonian systems*, Ann. Phys. (8), 11 (2002), pp. 15–38.

[32] J. Struckmeier and C. Riedel, *Noether's theorem and Lie symmetries for time-dependent Hamilton-Lagrange systems*, Phys. Rev. E (3), 66 (2002), 066605.

[33] D. R. Stump, *Arnol′d's transformation—an example of a generalized canonical transformation*, J. Math. Phys., 39 (1998), pp. 3661–3669.

[34] W. Thirring, *Lehrbuch der mathematischen Physik. Band* 1. *Klassische dynamische Systeme*, Springer-Verlag, Vienna, 1977.

[35] A. V. Tsiganov, *Canonical transformations of the extended phase space, Toda lattices and the Stäckel family of integrable systems*, J. Phys. A, 33 (2000), pp. 4169–4182.

[36] M. van der Put and M. F. Singer, *Galois Theory of Linear Differential Equations*, Grundlehren Math. Wiss. 328, Springer-Verlag, Berlin, 2003.

[37] E. T. Whittaker, *A Treatise on the Analytical Dynamics of Particles and Rigid Bodies. With an Introduction to the Problem of Three Bodies*, Cambridge Mathematical Library, Cambridge University Press, Cambridge, UK, 1988. Reprint of the 1937 edition.

[38] K. Wodnar, *New formulations of the Sitnikov problem*, in Predictability, Stability, and Chaos in N-Body Dynamical Systems (Cortina d'Ampezzo, 1990), NATO Adv. Sci. Inst. Ser. B Phys. 272, Plenum, New York, 1991, pp. 457–466.

SIAM J. Applied Dynamical Systems
Vol. 8, No. 1, pp. 298–347
© 2009 Society for Industrial and Applied Mathematics

Design of Parametrically Forced Patterns and Quasipatterns*

A. M. Rucklidge† and M. Silber‡

Abstract. The Faraday wave experiment is a classic example of a system driven by parametric forcing, and it produces a wide range of complex patterns, including superlattice patterns and quasipatterns. Nonlinear three-wave interactions between driven and weakly damped modes play a key role in determining which patterns are favored. We use this idea to design single and multifrequency forcing functions that produce examples of superlattice patterns and quasipatterns in a new model PDE with parametric forcing. We make quantitative comparisons between the predicted patterns and the solutions of the PDE. Unexpectedly, the agreement is good only for parameter values very close to onset. The reason that the range of validity is limited is that the theory requires strong damping of all modes apart from the driven pattern-forming modes. This is in conflict with the requirement for weak damping if three-wave coupling is to influence pattern selection effectively. We distinguish the two different ways that three-wave interactions can be used to stabilize quasipatterns, and we present examples of 12-, 14-, and 20-fold approximate quasipatterns. We identify which computational domains provide the most accurate approximations to 12-fold quasipatterns and systematically investigate the Fourier spectra of the most accurate approximations.

Key words. pattern formation, quasipatterns, superlattice patterns, mode interactions, Faraday waves

AMS subject classifications. 35B32, 37G40, 52C23, 70K28, 76B15

DOI. 10.1137/080719066

1. Introduction.

The classic Faraday wave experiment consists of a horizontal layer of fluid that spontaneously develops a pattern of standing waves on its surface as it is driven by vertical oscillation with amplitude exceeding a critical value; see [2, 33, 44] for recent reviews and surveys. Many other experimental, biological, and environmental systems also form patterns [14, 31], but Faraday wave experiments have consistently produced patterns with remarkably high degrees of symmetry. One consequence of this is that, over the years, Faraday wave experiments have repeatedly produced new patterns of behavior that go beyond the existing range of theoretical understanding and require the development of new ideas for their explanation. For example, in the early 1990s, *quasipatterns* were discovered in two different Faraday wave experiments, one with a low-viscosity deep layer of fluid with single-frequency forcing [5, 6, 7], and the other with a high-viscosity shallow layer of fluid and forcing with two commensurate temporal frequencies [20]. These patterns are periodic in time but are quasi-periodic in any spatial direction; that is, the amplitude of the pattern (taken along any direction in the plane) can be regarded as the sum of waves with incommensurate

*Received by the editors March 21, 2008; accepted for publication (in revised form) by D. Barkley August 25, 2008; published electronically February 19, 2009. This research was supported by the National Science Foundation (DMS-0309667) and by the Engineering and Physical Sciences Research Council (GR/S45928/01).
http://www.siam.org/journals/siads/8-1/71906.html
†Department of Applied Mathematics, University of Leeds, Leeds LS2 9JT, UK (a.m.rucklidge@leeds.ac.uk).
‡Department of Engineering Sciences and Applied Mathematics, and Northwestern Institute on Complex Systems, Northwestern University, Evanston, IL 60208, USA (m-silber@northwestern.edu).

spatial frequencies. Despite this, the spatial Fourier transforms of quasipatterns have 8-, 10-, or 12-fold rotational order. Quasipatterns are, of course, related to quasicrystals [54], and quasipatterns have been found in nonlinear optical systems [30], in shaken convection [51, 65], and in liquid crystals [37] as well as the Faraday wave experiment [2, 5, 6, 7, 20, 34]. There is, as yet, no satisfactory theoretical understanding of the formation of quasipatterns, owing to the problem of small divisors [52].

Theoretical efforts aimed at understanding the pattern selection problem have centered around *weakly nonlinear theory*. The calculations for the real Faraday wave problem, with finite depth and nonzero viscosity, are difficult (these involve solving the Navier–Stokes equations with a free surface boundary condition [58]). Most calculations aimed at producing superlattice patterns and quasipatterns have focused on simpler equations, such as the Zhang–Viñals [68] equations, which model Navier–Stokes when the depth is infinite and the viscosity is small, or on model equations, such as variations on the Swift–Hohenberg equation [23, 38, 43] or the FitzHugh–Nagumo equations [16].

Nonlinear three-wave resonant interactions have long been recognized as playing a key role in pattern selection in Faraday wave experiments and other situations where complex patterns are found [20, 42, 46, 69]. A series of papers [48, 49, 50, 56, 57, 62, 63] has developed this idea, using symmetry considerations to understand pattern selection in Faraday wave experiments with two-frequency forcing, exploiting the three-wave resonant interactions in the context of weakly broken Hamiltonian structure. This approach was able to explain several of the experimentally observed superlattice patterns and suggested ways of designing multifrequency forcing functions that could be used to control which patterns would emerge [49, 62, 63]. The approach is in principle predictive but has been used only to determine which additional frequencies to add to the forcing function in order to make observed patterns more robust [2, 17, 21]. The theory has not been tested quantitatively against solutions of a pattern-forming system, though weakly nonlinear coefficients have been computed for the Zhang–Viñals equations [50, 62] and, more recently, for the Navier–Stokes equations in the infinite depth case [58].

The main goal of this paper is to come to a greater understanding of some of the complex patterns that are found in high-precision large-aspect-ratio Faraday wave experiments. Much of the complexity arises from using two or more frequencies in the forcing of the experiment, and recent work [50, 62] explains in principle how to connect the amplitudes and phases of the various components of the forcing frequencies to the nonlinear pattern selection problem. The existing theory provides rules of thumb for designing forcing functions that should encourage the appearance of particular patterns, and it seems to work well, at least qualitatively, and at least in some circumstances [2, 17, 21].

We take the point of view that in order to claim convincingly that we understand the pattern selection process in these problems with multifrequency forcing, we should be able to predict *in advance* which patterns will be found for different parameter values, we should be able to predict the amplitudes and range of stability of the patterns, and we should test against a pattern forming system that is not constrained to produce only a limited range of patterns. In order to have a flexible framework for testing predictions, we have devised a partial differential equation (PDE) with multifrequency forcing (3.1) that shares many of the characteristics of the real Faraday wave experiment, but that has easily controllable dissipation

and dispersion relations and simple nonlinear terms. The linear behavior of the PDE reduces to the damped Mathieu equation, with subharmonic and harmonic tongues, and the simple quadratic and cubic nonlinearities allow three-wave interactions. The PDE has a Hamiltonian limit, but it differs from the real situation in the details of its dispersion relation and in the lack of any coupling to a large-scale mean flow. Notwithstanding these differences, the PDE allows us to explore in detail some of the generic issues surrounding pattern selection in very large aspect pattern forming systems with parametric forcing. A preliminary discussion of the question of quasipattern selection in the new model PDE can be found in [53].

In section 2, we review the details of how resonant three-wave interactions influence pattern selection. The main idea is that two pattern-forming modes, with wavevectors separated by an angle θ, are coupled to a weakly damped mode, and this coupling can lead to the angle θ either featuring in the resulting pattern or being eliminated from the resulting pattern [20, 42, 46, 69]. We introduce the model PDE in section 3, and describe its linear and nonlinear features in sections 4 and 5. The appendix gives full details of the weakly nonlinear calculations.

In section 6 we devise a forcing function that stabilizes the 22° superlattice patterns that have been observed in large-aspect-ratio Faraday wave experiments [2, 21, 34]. We compute fully nonlinear solutions of the PDE and compare the computed pattern amplitudes with those predicted by weakly nonlinear theory. This demonstrates that the agreement is quantitatively accurate only very close to onset (within 0.1%). We show that the limit on the range of validity arises because of the presence of the weakly damped modes that are required for the superlattice pattern to be stabilized. Thus parameter regimes that are likely to produce the most interesting patterns, arising from coupling to weakly damped modes, are also parameter regimes where weakly nonlinear theory has the most restricted validity.

Two mechanisms have been proposed for quasipattern formation, both building on ideas of Newell and Pomeau [46], and one aim of this paper is to demonstrate that both proposed mechanisms for quasipattern formation are viable (preliminary work is reported in [53]). One mechanism applies to single-frequency forced Faraday waves [68] and has been tested experimentally [67]. Another was developed to explain the origin of the two length scales in superlattice patterns [50, 63] found in two-frequency experiments [34]. The ideas have not been tested quantitatively but have been used qualitatively to control quasipattern [2, 17] and superlattice pattern [21] formation in two- and three-frequency experiments. We explore the two mechanisms of quasipattern formation in the model PDE in sections 7 and 8.

Also in section 7, we address the distinction between true and approximate quasipatterns, as found in numerical experiments with periodic boundary conditions. Owing to the problem of small divisors, there is as yet no satisfactory mathematical treatment of quasipatterns [52]. Despite this, the weakly nonlinear stability calculations, which are in the framework of a 12-mode amplitude expansion truncated at cubic order, prove to be a reliable guide to finding parameter values where approximate quasipatterns are stable. The fact that stable 12-fold quasipatterns are found where they are expected demonstrates that this approach provides useful information, despite the reservations expressed in [52]. We explore the effect of domain size on the accuracy of the approximation to a true quasipattern and show how certain domains yield particularly accurate approximations.

In section 8, we present examples of *turbulent crystals* [46]: situations in which Fourier modes oriented more than about 30° apart do not affect each other, at the level of a cubic

truncation. In this context, we find examples of 12-, 14-, and 20-fold quasipatterns. These are the first examples of quasipatterns of order greater than 12 found as stable solutions of a PDE. (A preliminary presentation of the 14-fold example is in [53].)

We summarize our results in section 9.

2. Theoretical background. Resonant triads have played a key role in the understanding of pattern formation [20, 42, 46, 69]. This section reviews, somewhat heuristically, the basic selection mechanisms in the case of wave patterns that are parametrically pumped by a two- (or more) frequency forcing function. The details behind this summary can be found in [49, 62]. We write the forcing function as

$$f(t) = f_m \cos(mt + \phi_m) + f_n \cos(nt + \phi_n) + \cdots , \tag{2.1}$$

where m and n are integers, f_m and f_n are real amplitudes, and ϕ_m and ϕ_n are phases. (We could, of course, set $\phi_m = 0$ without loss of generality.) Here we consider m to be the dominant driving frequency and focus on a pair of waves, each with wavenumber k_m and frequency $\Omega(k_m)$ specified by the dispersion relation associated with the linearized problem. These waves naturally oscillate at a frequency that is *subharmonic* to the dominant driving frequency m, so $\Omega(k_m) = m/2$ (subharmonic waves are typically the easiest to excite parametrically). The two waves have wavevectors $\boldsymbol{k}_m^{(1)}$ and $\boldsymbol{k}_m^{(2)}$, where we use the subscript to denote the frequency that excites the waves and the superscript to index the wavevectors. We write the critical modes in the form $z_1 e^{i\boldsymbol{k}_m^{(1)}\cdot\boldsymbol{x}+imt/2}$ and $z_2 e^{i\boldsymbol{k}_m^{(2)}\cdot\boldsymbol{x}+imt/2}$ (together with their complex conjugates), neglecting the additional temporal frequency contributions to the waves. These waves will interact nonlinearly with waves $we^{i\boldsymbol{k}^{(3)}\cdot\boldsymbol{x}+i\Omega(k^{(3)})t}$, where w is a complex amplitude, $\boldsymbol{k}^{(3)} = \boldsymbol{k}_m^{(1)} + \boldsymbol{k}_m^{(2)}$, and $\Omega(k^{(3)})$ is the frequency associated with $k^{(3)}$, provided that either (1) the same resonance condition is met with the temporal frequencies, i.e., $\Omega(k^{(3)}) = \frac{m}{2} + \frac{m}{2}$, as in Figure 1(a,b), or (2) any mismatch $\Delta = |\Omega(k^{(3)}) - \frac{m}{2} - \frac{m}{2}|$ in this temporal resonance condition can be compensated for by the forcing function $f(t)$ (Figure 1(c)). Case (1) corresponds to $1:2$ resonance, which occurs even for single-frequency forcing ($f_n = 0$), and case (2) applies, for example, to two-frequency forcing with the third wave oscillating at the difference frequency: $\Omega(k^{(3)}) = |m - n|$ and $\Delta = n$. Other cases analogous to (2), such as $\Omega(k^{(3)}) = m + n$, are discussed in [62], where the special significance of the difference frequency case is explained. Note that in both cases (1) and (2), the temporal frequency $\Omega(k^{(3)})$ determines the angle θ between the wavevectors $\boldsymbol{k}_m^{(1)}$ and $\boldsymbol{k}_m^{(2)}$ via the dispersion relation and therefore provides a possible selection mechanism for certain preferred angles appearing in the power spectrum associated with the wavepatterns. Selecting an angle of $0°$, with $1:2$ resonance in space and time (Figure 1(b)), is a special case.

The nonlinear interactions of the modes can be understood by considering resonant triad equations describing small amplitude waves, which take the form

$$\begin{aligned} \dot{z}_1 &= \lambda z_1 + q_1 \bar{z}_2 w + (a|z_1|^2 + b|z_2|^2) z_1 + \cdots , \\ \dot{z}_2 &= \lambda z_2 + q_1 \bar{z}_1 w + (a|z_2|^2 + b|z_1|^2) z_2 + \cdots , \\ \dot{w} &= \nu w + q_3 z_1 z_2 + \cdots , \end{aligned} \tag{2.2}$$

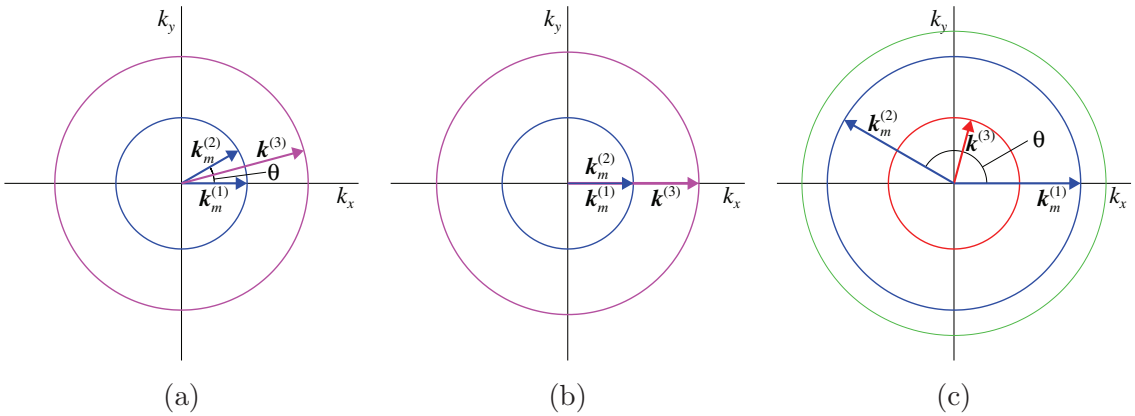


Figure 1. (a) $1:2$ *resonance occurs between two modes with wavevectors* $\boldsymbol{k}_m^{(1)}$ *and* $\boldsymbol{k}_m^{(2)}$ *(blue), with the same wavenumber* k_m *and separated by an angle* θ*, and a third mode with wavevector* $\boldsymbol{k}^{(3)}$ *(magenta), provided that* $\boldsymbol{k}^{(3)} = \boldsymbol{k}_m^{(1)} + \boldsymbol{k}_m^{(2)}$ *and* $\Omega(k^{(3)}) = 2\Omega(k_m)$*.* (b) *A special case of* $1:2$ *resonance in space and time occurs for* $\theta = 0$ *when the dispersion relation satisfies* $\Omega(2k_m) = 2\Omega(k_m)$*.* (c) *With two-frequency* $m:n$ *forcing, the nonlinear combination of two modes with wavevectors* $\boldsymbol{k}_m^{(1)}$ *and* $\boldsymbol{k}_m^{(2)}$*, and with* $\Omega(k_m) = m/2$ *(blue), can, in the presence of a second component of the forcing at frequency* n*, interact with a mode with wavevector* $\boldsymbol{k}^{(3)}$ *(red), provided that* $\boldsymbol{k}^{(3)} = \boldsymbol{k}_m^{(1)} + \boldsymbol{k}_m^{(2)}$ *and* $\Omega(k^{(3)}) = |m-n|$*. Waves driven by frequency* n *(green) do not enter the resonance condition.*

where all coefficients are real and the dot refers to timescales that are long compared to the forcing period. Here the quadratic coupling coefficients q_j are $\mathrm{O}(1)$ in the forcing in the $1:2$ resonance case, and $\mathrm{O}(|f_n|)$ in the difference frequency case [50]. For other angles θ between the wavevectors $\boldsymbol{k}_m^{(1)}$ and $\boldsymbol{k}_m^{(2)}$ we expect $q_1 \approx q_3 \approx 0$ because the temporal resonance condition for the triad of waves is not met.

Since the m frequency is dominant, the z_1 and z_2 modes will be excited (λ increases through zero) while the w mode is damped ($\nu < 0$). In this case, w can be eliminated from (2.2) via center manifold reduction [27] near the bifurcation point ($w \approx \frac{q_3 z_1 z_2}{|\nu|}$), resulting in the bifurcation problem

$$\begin{aligned} \dot{z}_1 &= \lambda z_1 + (a|z_1|^2 + \tilde{b}|z_2|^2)z_1, \\ \dot{z}_2 &= \lambda z_2 + (a|z_2|^2 + \tilde{b}|z_1|^2)z_2. \end{aligned} \tag{2.3}$$

These equations describe the competition between standing waves separated by an angle θ, where $\tilde{b} \equiv b + \frac{q_1 q_3}{|\nu|}$ explicitly includes the contribution from the slaved mode w and hence depends on the angle between the two wavevectors $\boldsymbol{k}_m^{(1)}$ and $\boldsymbol{k}_m^{(2)}$.

The contribution of the damped w mode is significant whenever $q_1 q_3$ is nonnegligible and the damping $|\nu|$ is not too great, and it can be made more important by increasing $q_1 q_3$ and/or by decreasing the damping $|\nu|$. For instance, in the $1:2$ resonance case, for which $q_1 q_3$ is $\mathrm{O}(1)$, the damping can be decreased by taking $n = 2m$ in (2.1), since the w mode, with frequency m, will then be driven subharmonically by the n component of the forcing (as well as harmonically by the m component).

In the difference frequency case, the quadratic interactions rely on the presence of the n component of the forcing to allow the temporal resonance condition to be met, so $q_1 q_3$ is

$O(|f_n|^2)$. Thus the contribution of the damped w mode to $\tilde{b}$ can be made more important in two ways: first, by increasing f_n, or second, by parametrically driving the damped mode so that $|\nu|$ is decreased. This requires a third driving frequency $p = 2|n - m|$ to be added to the forcing (2.1). In both of these instances, the relative phases of the components of the forcing matter, since the modes are being generated nonlinearly with a preferred phase. These ideas are developed systematically in [50, 62], where it is also shown that if there is an underlying Hamiltonian structure, then $q_1 q_3 < 0$ for the $1:2$ resonance, and $q_1 q_3 > 0$ in the difference frequency case, provided that $n > m$. Note that when $q_1 q_3 > 0$ ($q_1 q_3 < 0$) then the $\boldsymbol{k}^{(3)}$-mode makes a positive (negative) contribution to the growth rate of the mode z_2 when z_1 is present.

In order to determine more precisely whether the resonant contribution to $\tilde{b}$ is significant enough to lead to a qualitative change in the resulting pattern, it is useful to rescale the amplitudes z_1 and z_2 in (2.3) by a factor of $1/\sqrt{|a|}$. Then we obtain, for $a < 0$, the rescaled equations

$$\begin{aligned} \dot{z}_1 &= \lambda z_1 - (|z_1|^2 + B_\theta |z_2|^2) z_1, \\ \dot{z}_2 &= \lambda z_2 - (|z_2|^2 + B_\theta |z_1|^2) z_2, \end{aligned} \tag{2.4}$$

where $B_\theta \equiv \frac{\tilde{b}}{a}$. Here the θ subscript indicates that the cross-coupling coefficient between the $\boldsymbol{k}_m^{(1)}$ and $\boldsymbol{k}_m^{(2)}$ modes depends on the angle θ between them.

The B_θ function has important consequences for the stability of regular patterns. As a simple example, note that stripes ($|z_1| = \sqrt{\lambda}$, $z_2 = 0$) are stable if $B_\theta > 1$, while rhombs associated with a given angle θ ($|z_1| = |z_2| = \sqrt{\lambda/(1 + B_\theta)}$) are preferred if $|B_\theta| < 1$. Moreover, if $|B_\theta| < 1$ for any angle θ, then stripes will necessarily be unstable near onset. Since by judicious choice of forcing frequencies we have at least some ability to control both the magnitude and sign of B_θ over some range of angles θ, we have a mechanism for enhancing or suppressing certain combinations of wavevectors in the resulting weakly nonlinear patterns. Alternatively, as suggested by [68], if we choose forcing frequencies that lead to a large $1:2$ resonant contribution at $\theta = 0$ (Figure 1(b)), then the coefficient a can become large compared to b, which in turn will cause the rescaled cross-coupling coefficient B_θ to be small over a broad range of angles away from $\theta = 0$. (As $\theta \to 0$, it can be shown that $B_\theta \to 2$.) These two cases are contrasted in sections 7 and 8, with preliminary work described in [53].

Before continuing, we reiterate the constraint on using this analysis to design forcing functions that will stabilize a given pattern. In eliminating the damped mode w, we performed a center manifold reduction, which is valid provided that all nonneutral modes are linearly damped with decay rates that are bounded away from zero. The domain of validity of the reduced equations depends on the extent to which w is damped near the bifurcation point. If that damping is very weak, then the reduced equations will be quantitatively predictive only in a correspondingly small neighborhood of the bifurcation point. We point this out since, from the discussion above, it is clear that the two ways of influencing the magnitude of $\tilde{b}$, namely increasing the driving force f_n to control $q_1 q_3$ or driving the difference frequency to reduce the damping $|\nu|$, both can lead to situations where the center manifold reduction is no longer valid: the m and n modes could set in together, or the m and difference frequency modes could set in together. In either case, a codimension-two analysis could be performed, but that is beyond the scope of this paper. In practice, the severity of this constraint can be

seen only by comparing predictions for the amplitudes and stability of patterns with solutions of the problem at hand (which we do systematically in section 6).

3. The model PDE. In order to explore these issues in detail, we have devised a phenomenological PDE for which the leading nonlinear coefficients in bifurcation problems such as (2.3) can be calculated relatively easily, and which is reasonably simple to integrate numerically. The equation is

$$\begin{aligned}\frac{\partial U}{\partial t} &= (\mu + i\omega)U + (\alpha + i\beta)\nabla^2 U + (\gamma + i\delta)\nabla^4 U \\ &\quad + Q_1 U^2 + Q_2|U|^2 + C|U|^2 U + i\,\mathrm{Re}(U) f(t),\end{aligned} \tag{3.1}$$

where $f(t)$ is a real 2π-periodic function; $U(x, y, t)$ is a complex-valued function, with $(x, y) \in \mathbb{R}^2$; $\mu < 0$, ω, α, β, γ, and δ are real parameters; and $Q_1 = Q_{1r} + iQ_{1i}$, $Q_2 = Q_{2r} + iQ_{2i}$, and $C = C_r + iC_i$ are complex parameters.

The way the forcing function enters the PDE was chosen so that the linearized problem reduces to the damped Mathieu equation (in much the same way that hydrodynamic models of the Faraday instability reduce to this equation [3]). The PDE has the advantage that the dispersion relation can be controlled easily, and weakly nonlinear theory is relatively straightforward to compute. The linear terms are diagonal in Fourier space, so the PDE is readily amenable to the exponential time differencing numerical methods of [12]. In addition, the nonlinear terms are simple (they do not involve any derivatives), and so numerical solutions are relatively inexpensive.

One of the special features of parametric systems is that, even though pattern selection is a nonlinear process, the position of the linear stability curves often determines which resonant triad interactions are dominant. In turn, it is the resonant triad interactions rather than the details of the particular form of nonlinearity in the equation that drive the pattern selection process. For these reasons, the model PDE is a useful testing ground for results derived from symmetric bifurcation theory.

The model PDE is similar to the complex Ginzburg–Landau equation—but we point out that $U(x, y, t)$ is itself the pattern-forming field and is not the amplitude of some other underlying pattern. With $\mu < 0$, all waves are damped in the absence of driving. Note also that the parametric forcing $f(t)$ is explicitly a function of time, so that we are resolving dynamics on the fast timescale set by the periodic forcing. In contrast, other authors [8, 9, 10, 11, 29] have investigated Ginzburg–Landau equations that describe the slow, large spatial-scale evolution of the amplitude of an otherwise spatially homogeneous oscillatory mode arising through Hopf bifurcation. Instead of resolving the fast oscillations of the subharmonic response to the time dependent forcing, a constant-coefficient $\bar{U}$ term is introduced into the equation, proportional to the amplitude of the parametric forcing [11]. With multifrequency forcing, other terms, such as $\bar{U}^2$, are also introduced [8, 9, 10, 29].

There are important qualitative differences, of course, between the model PDE and the Faraday wave experiment. One difference is the role of the $k = 0$ mode. In the PDE, this mode is damped ($\mu < 0$) and has a nonzero frequency ω; moreover, the $k = 0$ mode can be nonlinearly excited. In the Faraday wave problem, owing to mass conservation, the $k = 0$ mode is neutral and cannot be excited, and this may have important consequences in the dynamics

[13, 40]. (In the Zhang–Viñals model [68] this requirement is met, since all nonlinear terms appear with an overall spatial derivative that prevents the excitation of the $k = 0$ mode.) Another important difference between the model PDE and the Faraday wave experiment is that the dispersion relations have a different structure: in the model PDE, the frequency is a polynomial function of the square of the wavenumber, but the dispersion relation for Faraday waves is more complicated [3, 4, 36]. Nonetheless, the marginal stability curves of the model PDE, especially with multifrequency forcing, are similar to those that are observed in large-aspect-ratio Faraday wave experiments.

The theory developed in [50, 62] applies in the weakly damped weakly forced regime, and certain of those results also require that the undamped problem have a Hamiltonian structure. This limit can be realized for our model (3.1) by setting $\mu = \alpha = \gamma = C_r = 0$ and $Q_2 = -2\bar{Q}_1$. In this case, the Hamiltonian is

$$H(U,\bar{U}) = \iint_R \Big[\omega|U|^2 - \beta|\nabla U|^2 + \delta|\nabla^2 U|^2 + f(t)(\mathrm{Re}(U))^2 - iQ_1U^2\bar{U} + i\bar{Q}_1\bar{U}^2U + \frac{C_i}{2}|U|^4\Big]\,dx\,dy, \tag{3.2}$$

and U evolves according to

$$\frac{\partial U}{\partial t} = i\frac{\delta H}{\delta \bar{U}}. \tag{3.3}$$

The region R corresponds to the domain of integration of the PDE (3.1), where we have assumed that periodic boundary conditions apply on ∂R. In the examples presented below, some of the parameter choices are nearly Hamiltonian (sections 7 and 8) and some are not (section 6).

4. Linear theory. The linear problem associated with (3.1) takes the form of a damped Mathieu equation for each Fourier mode $e^{i\boldsymbol{k}\cdot\boldsymbol{x}}$. Specifically, if we set $U(\boldsymbol{x},t) = e^{i\boldsymbol{k}\cdot\boldsymbol{x}}(p(t)+iq(t))$ in (3.1) linearized about $U = 0$, then we obtain

$$\ddot{p} + \hat{\gamma}\dot{p} + \left(\Omega^2 + \hat{\Omega}f(t)\right)p = 0, \tag{4.1}$$

where

$$\hat{\gamma} = 2\left(-\mu + \alpha k^2 - \gamma k^4\right), \qquad \hat{\Omega} = \omega - \beta k^2 + \delta k^4, \qquad \Omega^2 = \frac{\hat{\gamma}^2}{4} + \hat{\Omega}^2. \tag{4.2}$$

We use the method of [4] to solve this linear problem for multifrequency forcing $f(t)$, which determines the critical forcing amplitude. Further details can be found in the appendix.

We require that the damping $\hat{\gamma}$ be positive for all k, so $\mu < 0$, $\gamma \le 0$, and $\alpha > -\sqrt{4\gamma\mu}$. However, we do not necessarily insist that it be monotonic, which would require $\alpha \ge 0$. Nonmonotonic damping is possible in the Faraday wave experiment in shallow layers: if the viscous boundary layers extend from top to bottom of the experimental container, modes with long wavelengths can be more heavily damped than short-wave modes, and indeed the first harmonic mode can be unstable at lower forcing than the subharmonic mode [35, 45]. We

use nonmonotonic damping for the example in section 7. Likewise, we restrict ourselves to the parameter regime where the dispersion relation $\hat{\Omega}(k)$ is nonnegative and an increasing function of k^2; thus we assume $\omega \geq 0$, $\delta \geq 0$, and $\beta \leq 0$.

We will use both $1:2$ and difference frequency resonances to control how modes interact, so it is important to understand that these resonances impose further constraints on the parameters in the dispersion relation. In the case of single frequency ($m = 1$) forcing $f(t) = a\cos(t+\phi)$, we expect a subharmonic instability to set in first with increasing a at a critical wavenumber k_1, which can be estimated by solving the equation $\hat{\Omega}(k_1) = \frac{1}{2}$. This estimate of k_1 is good, provided the damping is not too large and that the damping is a (nearly) monotonic function of k. The wavenumber associated with the first harmonic instability is determined by solving $\hat{\Omega}(k_2) = 1$. These calculations, together with simple trigonometry, determine that the angle θ associated with a $1:2$ resonant triad satisfies the equation $k_2 = 2k_1\cos\left(\frac{\theta}{2}\right)$, which has a solution, provided that $k_2 \leq 2k_1$. Choosing a scaling of $\boldsymbol{x}$ so that $k_1 = 1$, then we find that ω, β, and δ are related to θ by

$$\begin{aligned} \hat{\Omega}(k = k_1 = 1) &= \omega - \beta + \delta = \frac{1}{2}, \\ \hat{\Omega}\left(k = k_2 = 2\cos\left(\frac{\theta}{2}\right)\right) &= \omega - 4\beta\cos^2\left(\frac{\theta}{2}\right) + 16\delta\cos^4\left(\frac{\theta}{2}\right) = 1. \end{aligned} \tag{4.3}$$

In particular, the $1:2$ resonance will be at $\theta = 0^\circ$ ($k_2 = 2$) if we choose

$$\omega = \frac{1}{3} + 4\delta, \qquad \beta = -\frac{1}{6} + 5\delta, \tag{4.4}$$

where $\delta \in [0, \frac{1}{30}]$ ensures $\beta \leq 0$ and hence a monotonic dispersion relation. The $1:2$ resonance moves to $\theta = 90^\circ$ ($k_2 = \sqrt{2}$) if we choose

$$\omega = 2\delta, \qquad \beta = -\frac{1}{2} + 3\delta, \tag{4.5}$$

where we require $\delta \in [0, \frac{1}{6}]$.

Next we consider the case of two-frequency forcing,

$$f(t) = F\left(a_m\cos(mt + \phi_m) + a_n\cos(nt + \phi_n)\right), \tag{4.6}$$

where m and n are coprime integers, (a_m, a_n) are relative amplitudes scaled by an overall amplitude F, and (ϕ_m, ϕ_n) are phases. We focus on a resonant triad involving two critical modes with dominant frequency $\frac{m}{2}$ and a damped difference frequency mode with dominant frequency $n - m$. We will typically take m even and n odd with $n > m$ and $n - m < \frac{m}{2}$ (i.e., $\frac{n}{m} \in (1, \frac{3}{2})$). These conditions imply that the initial instability is expected to be harmonic, that the difference frequency mode decreases B_θ at the resonance angle in the Hamiltonian limit, and that the difference frequency mode has a wavenumber (k_{diff}) that is smaller than the critical wavenumber, and hence it is not too strongly damped (at least in the case of monotonic dissipation) [50]. We estimate the critical wavenumber of instability as $k \approx k_m$,

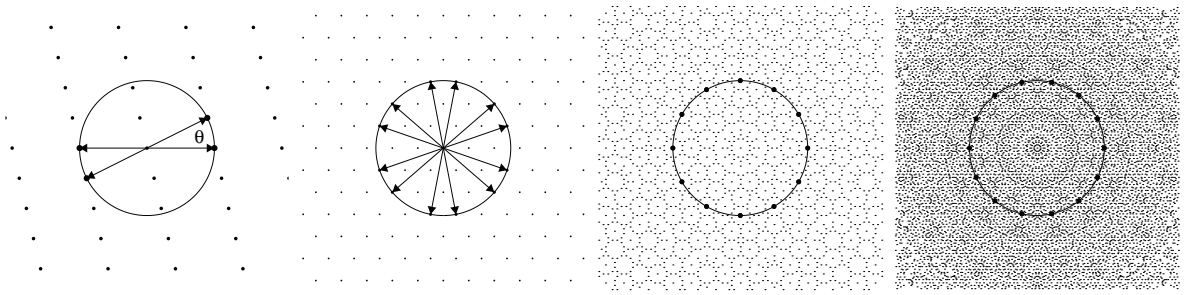


Figure 2. (a) *Rhombic lattice with an angle θ between the primary wavevectors.* (b) *Hexagonal superlattice, with an angle of* 21.8° *between the most closely spaced wavevectors.* (c,d) 12*-fold and* 14*-fold quasilattices, up to* 11*th order and* 7*th order, respectively* [52]. *See section* 7 *for a discussion of how these quasilattices are generated.*

where $\hat{\Omega}(k_m) = \frac{m}{2}$, and we assume a scaling such that $k_m = 1$. We then have that k_m and k_{diff} satisfy

$$\begin{aligned} \hat{\Omega}(k = k_m = 1) &= \omega - \beta + \delta = \frac{m}{2}, \\ \hat{\Omega}(k = k_{\text{diff}}) &= \omega - \beta k_{\text{diff}}^2 + \delta k_{\text{diff}}^4 = n - m, \end{aligned} \tag{4.7}$$

which we can solve for ω and β:

$$\omega = \frac{-k_{\text{diff}}^2 \frac{m}{2} + (n-m)}{1 - k_{\text{diff}}^2} + \delta k_{\text{diff}}^2 \quad \text{and} \quad \beta = \frac{-\frac{m}{2} + (n-m)}{1 - k_{\text{diff}}^2} + \delta(1 + k_{\text{diff}}^2). \tag{4.8}$$

Setting $k_{\text{diff}} = 2\cos\left(\frac{\theta}{2}\right)$, we can relate an angle in the power spectrum of the pattern with the wavenumber k_{diff} of the damped mode associated with the resonant triad. For instance, for $\theta = 150°$ (the complementary angle to 30°, which appears in the 12-fold quasipatterns), we have $k_{\text{diff}} = \sqrt{2 - \sqrt{3}}$. Alternatively, if $\theta = 158.2°$ (the complementary angle to 21.8°, which appears in the simplest hexagonal superlattice patterns), we have $k_{\text{diff}} = \frac{1}{\sqrt{7}}$.

5. Weakly nonlinear theory. Our weakly nonlinear calculations are aimed at determining the coefficients of the leading nonlinear terms in finite-dimensional bifurcation problems associated with certain families of patterns in the plane, and the corresponding lattices of wavevectors. These finite-dimensional bifurcation problems allow us to compute rigorously the relative stability of various simple planforms, e.g., stripes versus squares, rhombs, and hexagons, and also to calculate relative stability of superlattice patterns and hexagons, stripes, and certain rhomb patterns. Moreover, these calculations lead to quantitative predictions of the amplitude of the standing wave patterns as a function of the distance λ from the bifurcation point, where $\lambda \equiv (F - F_c)/F_c$ and F_c is the critical value of the overall forcing amplitude.

A simple example of such a reduction to a finite-dimensional problem was presented in section 2, where we described the bifurcation problem associated with a pair of standing waves oriented at an angle θ relative to each other, where $\theta \in (0, \pi/2]$ was bounded away from $\pi/3$. An example rhombic lattice is shown in Figure 2(a). In that case, the bifurcation problem

consisted of a pair of amplitude equations given by (2.4), and, after appropriate scaling, there was just a single nonlinear coefficient B_θ. The details of the (numerical) computation of this coefficient from the governing PDE (3.1) are given in the appendix. In subsequent sections of this paper we present plots of B_θ as a function of θ for certain parameter sets and forcing functions $f(t)$ used in our numerical simulations of the model PDE.

As an additional, specific example we consider patterns associated with the hexagonal superlattice that has been observed in several Faraday wave experiments [2, 21, 34]. Equivariant bifurcation theory [24] was used to derive the form of the bifurcation problem [18]. This bifurcation problem describes the long-time evolution of the twelve modes on the critical circle that are associated with patterns that tile a plane in hexagonal fashion. The critical Fourier modes associated with the 21.8° superlattice pattern are indicated in Figure 2(b). We label these modes as follows: (z_1, z_3, z_5) are complex amplitudes associated with wavevectors separated by 120°, and, together with their complex conjugates, they comprise the modes associated with one hexagon, while (z_2, z_4, z_6) and their complex conjugates are associated with a second hexagon, rotated by approximately 21.8° relative to the first.

The form of the bifurcation problem associated with this hexagonal superlattice, to cubic order in the amplitudes, is

$$\begin{aligned}
\frac{dz_1}{dt} &= \lambda z_1 + Q\bar{z}_3\bar{z}_5 - \left(|z_1|^2 + B_{60}\left(|z_3|^2 + |z_5|^2\right) + B_{22}|z_4|^2 + B_{38}|z_6|^2 + B_{82}|z_2|^2\right) z_1,\\
\frac{dz_3}{dt} &= \lambda z_3 + Q\bar{z}_5\bar{z}_1 - \left(|z_3|^2 + B_{60}\left(|z_5|^2 + |z_1|^2\right) + B_{22}|z_6|^2 + B_{38}|z_2|^2 + B_{82}|z_4|^2\right) z_3,\\
\frac{dz_5}{dt} &= \lambda z_5 + Q\bar{z}_1\bar{z}_3 - \left(|z_5|^2 + B_{60}\left(|z_1|^2 + |z_3|^2\right) + B_{22}|z_2|^2 + B_{38}|z_4|^2 + B_{82}|z_6|^2\right) z_5,
\end{aligned}$$

with similar equations for z_2, z_4, and z_6, related by symmetry. The label on the coefficients indicates the angle between pairs of modes; e.g., there is an angle of approximately 82° between z_1 and z_2, while z_1 and z_4 are separated by approximately 22°. The nonlinear coefficients Q and B_{60} are computed from the governing PDEs by considering the problem of bifurcation on a simple hexagonal lattice involving a subset of the modes, while B_θ is computed for an arbitrary $\theta \neq 60°$ on a rhombic lattice. We can then read off B_{22}, B_{38}, etc., from a graph of B_θ as required. The details are in the appendix. Note that we have assumed that the bifurcation to a stripe planform is supercritical, so that we can rescale the amplitudes to make the self-coupling coefficient $a = -1$.

The standard planforms, namely stripes, rhombs (associated with each of the angles 22°, 38°, 82°), hexagons, and superlattice patterns, correspond to equilibrium solutions of these equations. The calculation of their linear stability proceeds in a standard fashion and is summarized in [18]. In fact, owing to the presence of the quadratic term, all planforms bifurcate unstably, but because Q is typically very small for multifrequency forcing and sufficiently weak damping [48], the planforms can be stabilized at small amplitude by secondary bifurcations. The superlattice patterns, which satisfy $|z_1| = |z_2| = \cdots = |z_6|$, come in two varieties that bifurcate together, and their relative stability is unresolved at cubic order. Specifically, hexagonal superlattice patterns are associated with z_j being real, while triangular superlattice patterns are of the form $z_j = Re^{i\pi/3}$, where R is the real amplitude [55]. Which of these two superlattice patterns is favored in a given situation depends on higher-order terms in the

bifurcation problem. We do not calculate the coefficients of these terms and merely lump the two types of superlattice patterns together in our bifurcation diagrams.

6. Numerical experiments: Selecting superlattice patterns. Motivated by experimental observations [2, 21, 34] of superlattice patterns in the Faraday wave experiment with 6 : 7 forcing, in this section we use 6 : 7 forcing, with some additional forcing at frequency 2 to drive the difference frequency mode, in order to stabilize a 21.8° hexagonal superlattice pattern. We carry out the linear and weakly nonlinear calculations to find parameter values for which hexagonal superlattice patterns are predicted to be stable, and solve the model PDE numerically to confirm these predictions. In addition, we make quantitative comparisons between the weakly nonlinear predictions and the numerical solutions of the PDEs, comparing the predicted amplitudes and ranges of stability of the patterns. The agreement is not quantitative except at very small amplitude, and we develop an explanation for this at the end of the section.

The PDE was solved numerically using the fourth-order Runge–Kutta exponential time differencing numerical method (ETD4RK) of [12]. This pseudospectral method solves the linear part exactly, excluding the parametric forcing term, which is included with the nonlinear terms. This allows the use of a timestep based on accuracy requirements rather than numerical stability limits. Timestepping takes place in spectral space, and the timestep is chosen to be one-twentieth of the shortest of the periods of the forcing function $f(t)$, in order that the effect of the time dependent forcing be fully resolved by the fourth-order method. The nonlinear terms are evaluated using the FFTW software [22]. The resolution was relatively low for the examples in this section (up to 96×56 Fourier modes), but we used up to 1536^2 Fourier modes for the largest quasipattern examples discussed below. At each timestep, the upper half of the Fourier spectrum was removed in order to dealias the cubic terms.

We choose parameters so that the mode driven subharmonically by the 6 frequency has wavenumber $k_6 = 1$: $\hat{\Omega}(k = k_6 = 1) = 3$. The wavenumber associated with 21.8° is $\frac{1}{\sqrt{7}}$, which we wish to correspond to the difference frequency $(7 - 6)$, so we set $\hat{\Omega}(\frac{1}{\sqrt{7}}) = 1$. This mode responds subharmonically to driving at frequency 2. Furthermore, the wavenumber associated with 81.8° is part of this superlattice pattern, and we can influence this if we set $\hat{\Omega}(\frac{4}{\sqrt{7}}) = 6$, in 1 : 2 resonance with the primary response. This yields $\omega = \frac{2}{3}$, $\beta = -\frac{7}{3}$, and $\delta = 0$. We choose damping coefficients $\mu = -0.1$, $\alpha = 0.01$, and $\gamma = -0.15$ and nonlinear coefficients $Q_1 = 2 + i$, $Q_2 = 1 + 2i$, and $C = -1 + 30i$. Note that these parameters are not close to the Hamiltonian limit: in fact, we have chosen the nonlinear coefficients so that the 1 : 2 interaction *reduces* the cross-coupling coefficient in the range of angles close to 90°.

The linear theory for this problem is shown in Figure 3, confirming that modes with wavenumber close to 1 and $\frac{1}{\sqrt{7}}$ are neutral and weakly damped, respectively, with 6 : 7 : 2 forcing (Figure 3(f)). For this example, we have set the phases of the two main components of the forcing equal to zero, and (after some experimentation) set the phase of the component that drives the difference mode equal to 240°.

The cross-coupling coefficient B_θ is shown in Figure 4: with 6 forcing only, there is a dip in the curve around 90° owing to 1 : 2 resonance at 82°. (Had we chosen parameters sufficiently close to a Hamiltonian limit, this feature would have been a peak rather than a dip [50].) With 6 : 7 forcing, the dip at 22°, corresponding to the difference frequency, is visible, even

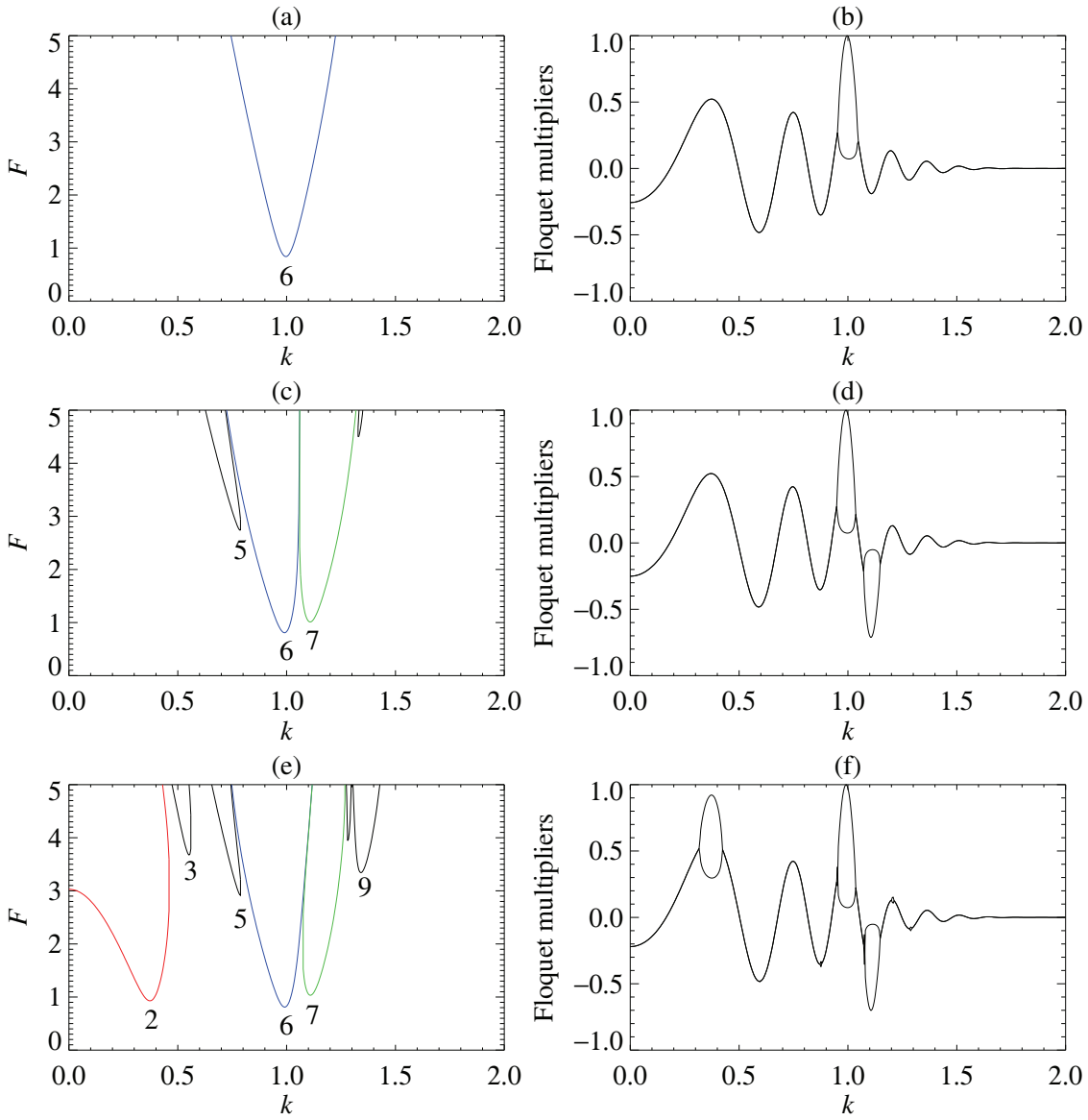


Figure 3. *Linear theory for one-, two-, and three-frequency forcing, with dispersion relation coefficients* $\omega = 2/3$, $\beta = -7/3$, *and* $\delta = 0$, *and damping coefficients* $\mu = -0.1$, $\alpha = 0.01$, *and* $\gamma = -0.1$. (a,b) 6 *forcing, with* $a_6 = 1$ *and* $F_c = 0.83973$. (c,d) $6:7$ *forcing, with* $(a_6, a_7) = (1, 1)$, $(\phi_6, \phi_7) = (0, 0)$, *and* $F_c = 0.80839$. (e,f) $6:7:2$ *forcing, with* $(a_6, a_7, a_2) = (1, 1, 0.45)$, $(\phi_6, \phi_7, \phi_2) = (0, 0, 240^\circ)$, $F_c = 0.80975$, *and* $k_c = 0.9910$. (a,c,e) *show neutral stability curves, whose minima define* F_c *and the critical wavenumber (close to* 1 *in all cases). Curves corresponding to the response to frequency* 6 *are blue, to frequency* 7 *are green, and to twice the difference frequency are red, with the corresponding driving frequency indicated. The minimum of the red curve in* (e) *is close to* $k = \frac{1}{\sqrt{7}}$. (b,d,e) *show the real parts of the Floquet multipliers at the critical forcing, confirming in* (b) *that the instability is harmonic at* $k = 1$, *in* (d) *that modes corresponding to frequency* 7 *are subharmonic but damped, and in* (f) *that modes with wavenumber close to* $\frac{1}{\sqrt{7}}$ *are weakly damped.*

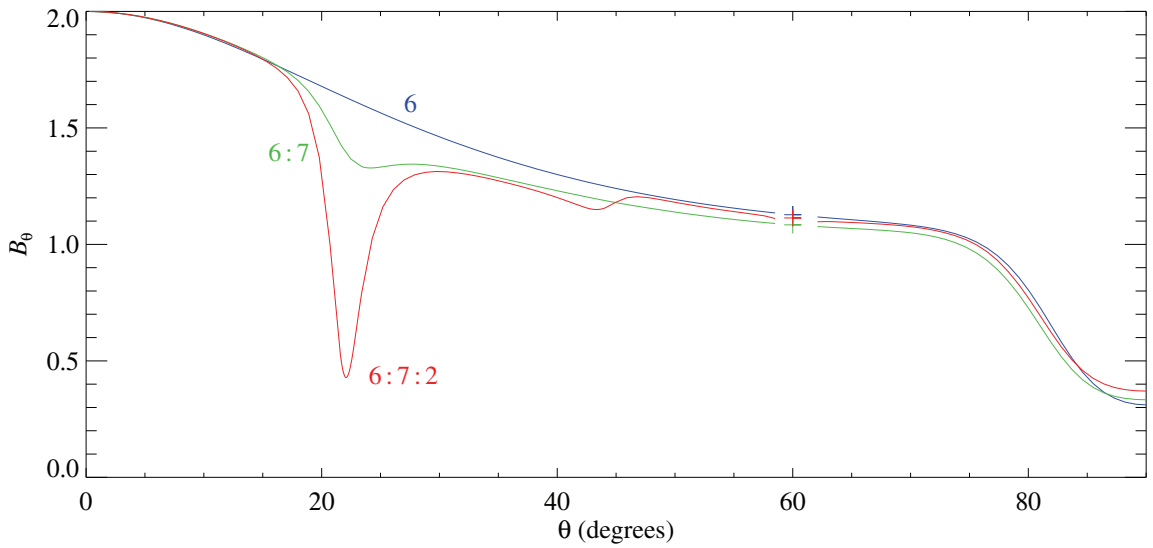
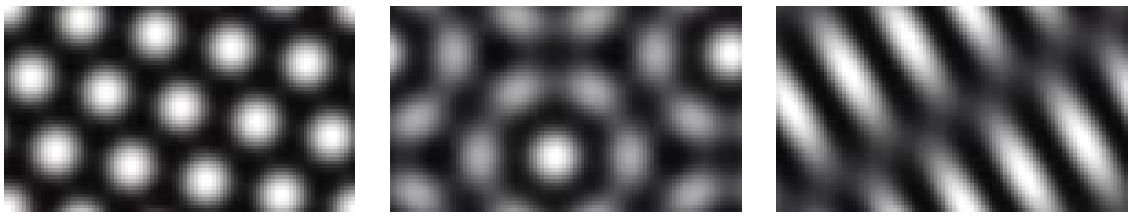


Figure 4. *B_θ, with linear parameters as in Figure 3 and nonlinear coefficients $Q_1 = 2 + i$, $Q_2 = 1 + 2i$, $C = -1+30i$. The blue curve is with 6 forcing only: the dip at 90° is because of $1 : 2$ resonance with frequency 6. The green curve has $6 : 7$ forcing: note the dip starting at 22°, corresponding to the difference frequency, even though this frequency is not forced directly. The red curve has $6 : 7 : 2$ forcing: note the pronounced dip at 22° and the smaller dip around 43°, corresponding to frequency 4 (which is in $1 : 2$ resonance with frequency 2). B_{60} is calculated separately, as described in the appendix. The relevant coefficients are $B_{22} = 0.46$, $B_{38} = 1.23$, $B_{60} = 1.11$, and $B_{82} = 0.63$.*



(a) (b) (c)

Figure 5. *With $6 : 7 : 2$ frequency forcing at* (a,b) 1.004 *and* (c) 1.02 *times critical, we find hexagons, a superlattice pattern, and 22° rhombs, with appropriate initial conditions. The resolution was 96×56 Fourier modes, and the domain is rectangular, $2\sqrt{7} \times 2\sqrt{7/3}$ critical wavelengths, big enough to fit two copies of the superlattice pattern. The grey-scale represents the real part of $U(x, y, t)$ with t equal to an integer multiple of 2π. The linear and nonlinear parameter values are as in Figures 3*(e,f) *and 4.*

though this frequency is not forced directly. Finally, with $6 : 7 : 2$ forcing, we can control the depth of the dip at 22°. An additional feature at 43° is visible, corresponding to frequency 4, in $1 : 2$ resonance with frequency 2.

Solutions of the PDE with $6 : 7 : 2$ forcing are shown in Figure 5, confirming that hexagons, superhexagons, and 22° rhombs are all stable solutions for different parameter values.

In Figure 6, we make quantitative comparison between the amplitudes and stability of these patterns as numerical solutions of the PDE, and the values predicted by weakly nonlinear

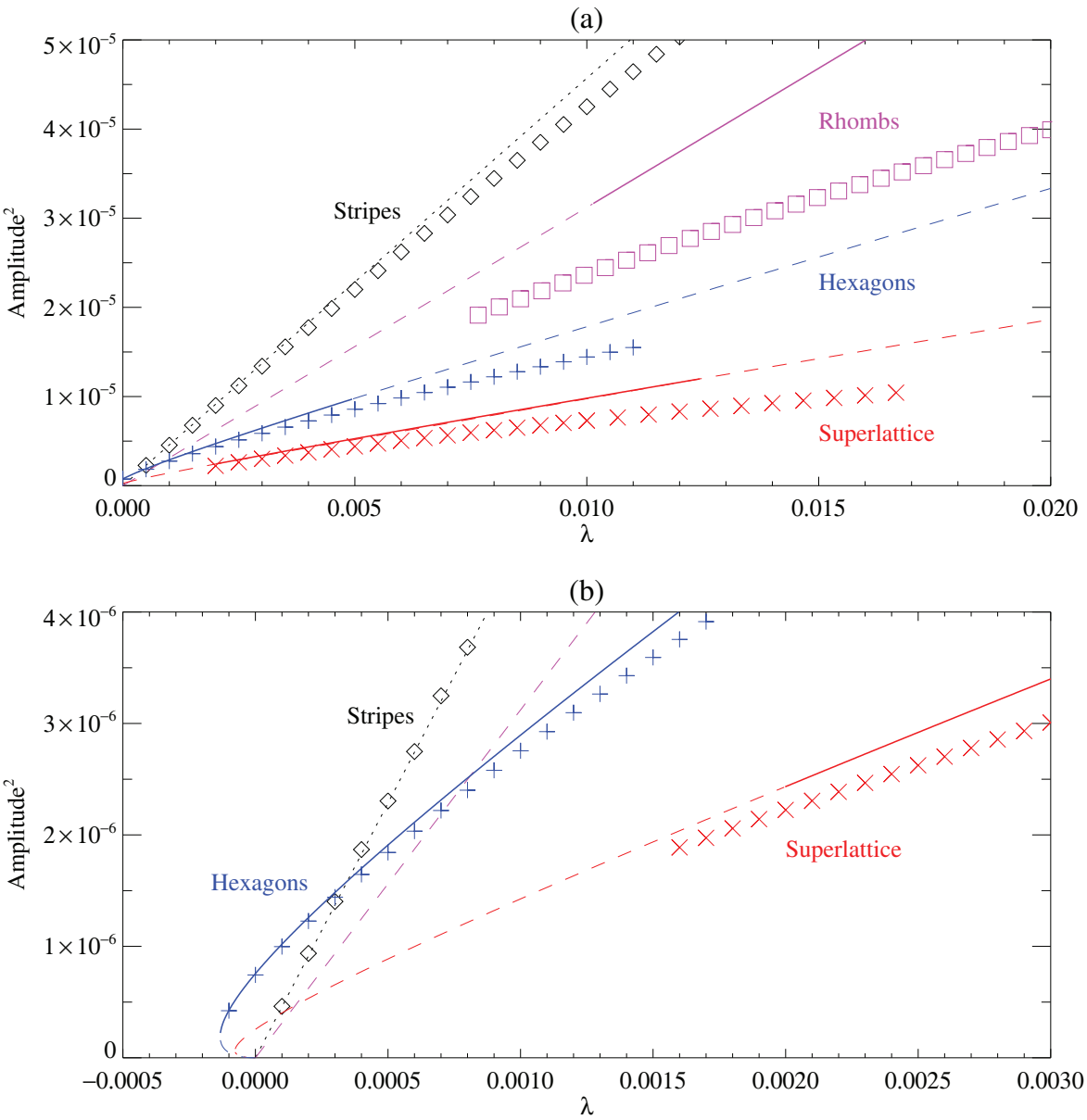


Figure 6. *Amplitudes of hexagons (blue), the superlattice pattern (red), 22° rhombs (magenta), and stripes (black) as a function of* λ *for* $6:7:2$ *forcing. Using weakly nonlinear theory, we denote stable solutions with solid lines and unstable solutions with dashed and dotted lines. We also denote numerically computed stable solutions of the PDEs by* $+$ *for hexagons,* $\times$ *for the superlattice pattern,* $\square$ *for rhombs, and* $\diamond$ *for (unstable) stripes. The amplitudes of the PDE solutions are computed by matching the time evolution of the* $k = 1$ *modes to the linear response functions.* (b) *is a detail of* (a)*, showing that the agreement between weakly nonlinear theory and the PDEs improves very close to onset.*

theory, for forcing up to 1.02 times critical. In addition, we show the results of one-dimensional simulations, which recover the unstable stripe pattern. In all cases, numerical solutions of the PDE agree with the weakly nonlinear prediction, provided we are close enough to onset (we have confirmed this by plotting the data on a logarithmic scale). Secondary bifurcations, which delimit parameter intervals where the patterns are stable, are also recovered, although the agreement is only qualitatively correct.

It is notable that the agreement between the amplitudes predicted by weakly nonlinear theory and measured from PDE simulations is not particularly good for the multimode patterns when compared to the much better agreement in the case of stripes. The reason for this lack of quantitative agreement for the complex patterns, and for the difference between the complex patterns and stripes, can be understood by going to higher order in the center manifold reduction that was performed to go from (2.2) to (2.3). We recall the framework: there are two weakly excited modes with amplitudes z_1 and z_2, and a damped mode w, which evolve according to

$$\begin{aligned} \dot{z}_1 &= \lambda z_1 + q_1 \bar{z}_2 w + (a|z_1|^2 + b|z_2|^2) z_1, \\ \dot{z}_2 &= \lambda z_2 + q_1 \bar{z}_1 w + (a|z_2|^2 + b|z_1|^2) z_2, \\ \dot{w} &= \nu w + q_3 z_1 z_2, \end{aligned} \tag{6.1}$$

where all coefficients are real and $\nu < 0$. We have discarded those higher-order terms that do not play a role in the center manifold reduction in order to emphasize the effect of the higher-order nonlinear terms that appear as a result of the reduction. We express w on the center manifold as a power series in z_1 and z_2 and perform a center manifold reduction to find

$$w = -\frac{q_3}{\nu} z_1 z_2 + \frac{q_3(q_1 q_3 - (a+b)\nu)}{\nu^3} \left(|z_1|^2 + |z_2|^2\right) z_1 z_2 + \cdots, \tag{6.2}$$

which results in

$$\begin{aligned} \dot{z}_1 &= \lambda z_1 + (a|z_1|^2 + \tilde{b}|z_2|^2) z_1 + \frac{q_1 q_3(q_1 q_3 - (a+b)\nu)}{\nu^3} \left(|z_1|^2 + |z_2|^2\right) |z_2|^2 z_1 + \cdots, \\ \dot{z}_2 &= \lambda z_2 + (a|z_2|^2 + \tilde{b}|z_1|^2) z_2 + \frac{q_1 q_3(q_1 q_3 - (a+b)\nu)}{\nu^3} \left(|z_1|^2 + |z_2|^2\right) |z_1|^2 z_2 + \cdots, \end{aligned} \tag{6.3}$$

where $\tilde{b} = b - q_1 q_3/\nu$ as before. In order for the w mode to influence the coupling constant $B_\theta = \tilde{b}/a$, and so produce interesting patterns, the quadratic coefficients q_1 and q_3 must be nonzero (the modes must be in three-wave resonance) and ν must be small (the mode must be weakly damped). However, if ν is small, the ν^3 in the denominator of the quintic terms implies that these high-order terms will be important exactly where the most interesting patterns will be found. Indeed, the graphs of amplitude against λ in Figure 6 are well fit by a quintic polynomial.

In contrast, the stripe pattern involves damped modes at $k = 0$ and $k = 2$, and it can be seen from Figure 3(f) that these modes are well damped, so there is no reason for quantitative agreement not to extend to larger values of λ, and indeed it does.

One can estimate the range of validity of the cubic truncation of (6.3) when λ and ν are both small. In the case of rhombs, the linear and cubic terms balance when $|z_1|^2 = |z_2|^2 =$

$\mathcal{O}(\lambda\nu)$. The quintic terms are thus smaller than the linear and cubic terms when $|\lambda| \ll |\nu|$, which is what one would expect: the center manifold reduction is valid when all modes that are eliminated are heavily damped compared to the modes that are retained.

Therefore, this codimension-one approach to finding interesting patterns has the smallest range of validity exactly where the patterns are most likely to be interesting. A proper treatment would require consideration of the codimension-two problem $(\lambda, \nu) = (0, 0)$, which is beyond the scope of this paper.

A further complication in the Faraday wave situation (and in our model PDE) is that in order for the interaction between the primary harmonic modes (driven by the m forcing) and the weakly damped difference frequency modes $(n-m)$ to take place, the frequency n must be present in the forcing function: the quadratic coefficients q_1 and q_3 increase with the strength of the forcing f_n, so the interaction is strongest when f_n is largest. This implies that the subharmonic mode driven by frequency n is itself only weakly damped and will be excited if f_n is increased beyond its critical value. Experimental evidence suggests that the codimension-two point (or bicritical point), where the primary harmonic (m) and subharmonic (n) modes are both neutral, is an organizing center for the dynamics [2, 20, 34]. (There has been relatively little progress on the theoretical understanding of the bicritical point, apart from a study in the case of a single frequency [66] and in a few particular cases for two-frequency excitation [48, 49].) Therefore, the problem should really be treated as a codimension-three interaction between primary harmonic modes and subharmonic modes, as well as weakly damped difference frequency modes.

Notwithstanding these complications, it is clear that the idea that pattern selection is being influenced by three-wave coupling to weakly damped modes is fundamentally correct, and the codimension-one approach, while having limited quantitative agreement with PDE simulations, is clearly providing the correct qualitative interpretation of the observed patterns.

7. Numerical experiments: 12-fold quasipatterns. In the previous section, we demonstrated how to stabilize simple patterns by driving the difference frequency. In this section, we show how this mechanism can be used to predict parameter values for stable approximate 12-fold quasipatterns, and we demonstrate how well a periodic pattern in a large domain can approximate a quasipattern.

In order to use triad interactions to encourage modes at 30°, we choose $m = 4$, $n = 5$ forcing: $4 : 5$ forcing has been used in several experiments to produce 12-fold quasipatterns [1, 20, 34]. We set $\hat{\Omega}(k = k_4 = 1) = 2$ so that the subharmonic response to frequency 4 comes at wavenumber 1, and we require that a wavenumber involved in 150° mode interactions $(k^2 = 2-\sqrt{3})$ correspond to the difference frequency: $\hat{\Omega}(k) = 1$. One solution is $\omega = 0.633975$, $\beta = -1.366025$, and $\delta = 0$. Twelve-fold quasipatterns also require modes at 90° to be favored, and for this choice of parameters, $\hat{\Omega}(k = \sqrt{2})$ is 3.37. Although this is not particularly close to 4, we can use $1 : 2$ resonance (driving at frequency 8) to control the 90° interaction. The linear theory for these cases is in Figure 7, with $\mu = -0.2$, $\alpha = -0.2$, and $\gamma = -0.15$. Note that the damping is nonmonotonic and has a minimum at $k = 1$.

The resulting B_θ curve with $4 : 5 : 8$ forcing and with near-Hamiltonian choice of nonlinear coefficients (Figure 8) shows pronounced dips at 30° and 90° as required. In Figure 9, we show that within a 12-amplitude cubic truncation, 12-fold quasipatterns are stable between 0.9995 and 1.0085 times critical. We have found stable approximate quasipatterns (marked by crosses

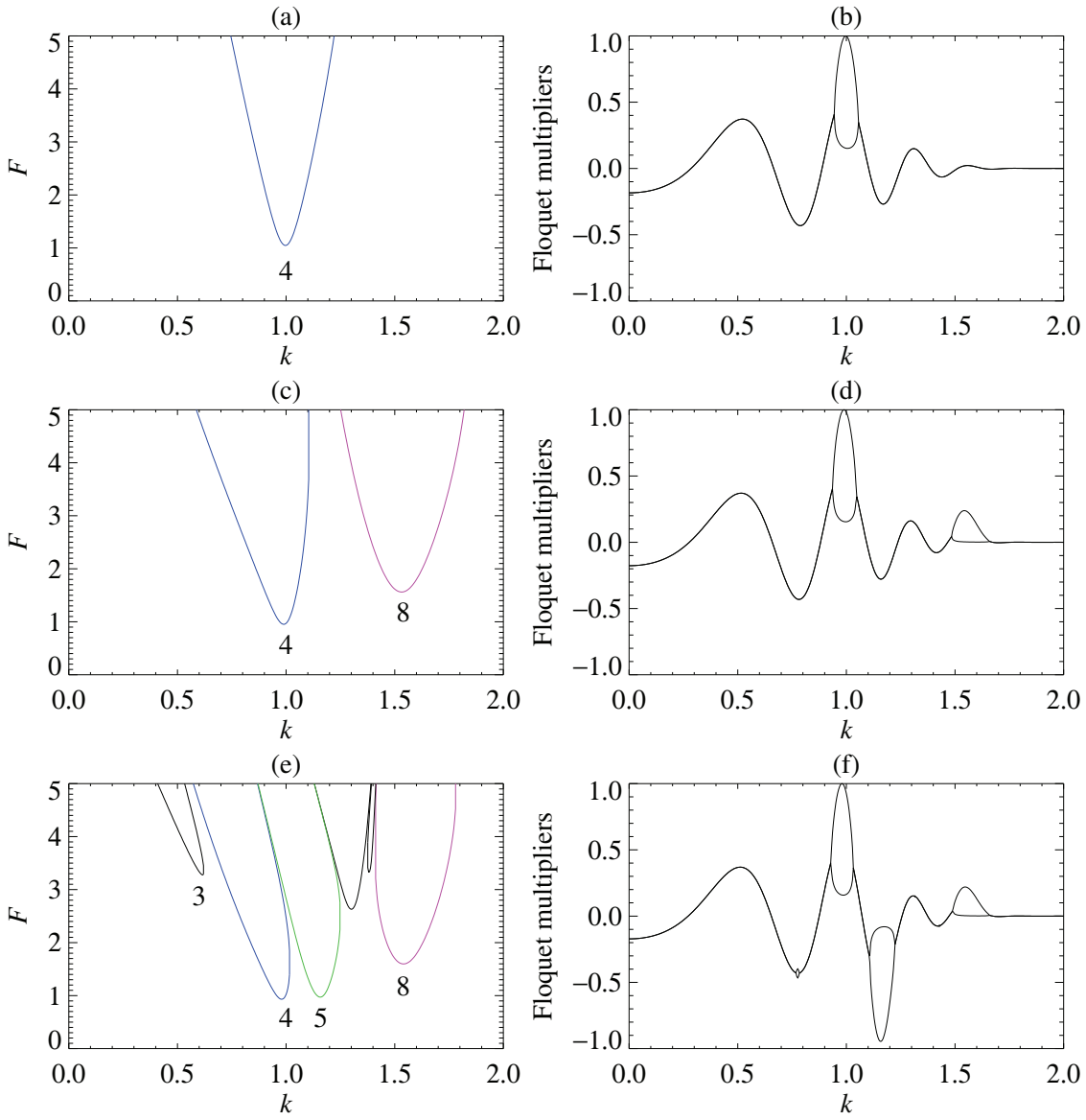


Figure 7. *Linear theory for the quasipattern example. The dispersion relation coefficients are* $\omega = 0.633975$, $\beta = -1.366025$, *and* $\delta = 0$, *and the damping coefficients are* $\mu = -0.2$, $\alpha = -0.2$, *and* $\gamma = -0.15$. (a,b) 4 *forcing, with* $a_4 = 0.57358$ *and* $F_c = 1.04730$. (c,d) $4:8$ *forcing, with* $(a_4, a_8) = (0.57358, 1.6)$, $(\phi_4, \phi_8) = (0, 0)$, *and* $F_c = 0.95214$. (e,f) $4:5:8$ *forcing, with* $(a_4, a_5, a_8) = (0.57358, 0.81915, 1.6)$, $(\phi_4, \phi_5, \phi_8) = (0, 0, 0)$, $F_c = 0.93159$, *and* $k_c = 0.9798$. (a,c,e) *show neutral stability curves, whose minima define* F_c *and the critical wavenumber (close to* 1 *in all cases). The main harmonic curves are blue, subharmonic curves are green, and curves corresponding to response at twice the main forcing frequency are magenta, with the corresponding driving frequency indicated.* (b,d,e) *show the real parts of the Floquet multipliers at the critical forcing, confirming in* (b) *that the instability is harmonic (Floquet multiplier equal to one) at* $k = 1$, *in* (d) *that the modes corresponding to frequency* 8 *are not as strongly damped as in* (b), *and in* (f) *that modes with wavenumber close to* 0.518 *are relatively weakly damped (Floquet multiplier about* 0.5*) even though the difference frequency is not driven directly.*

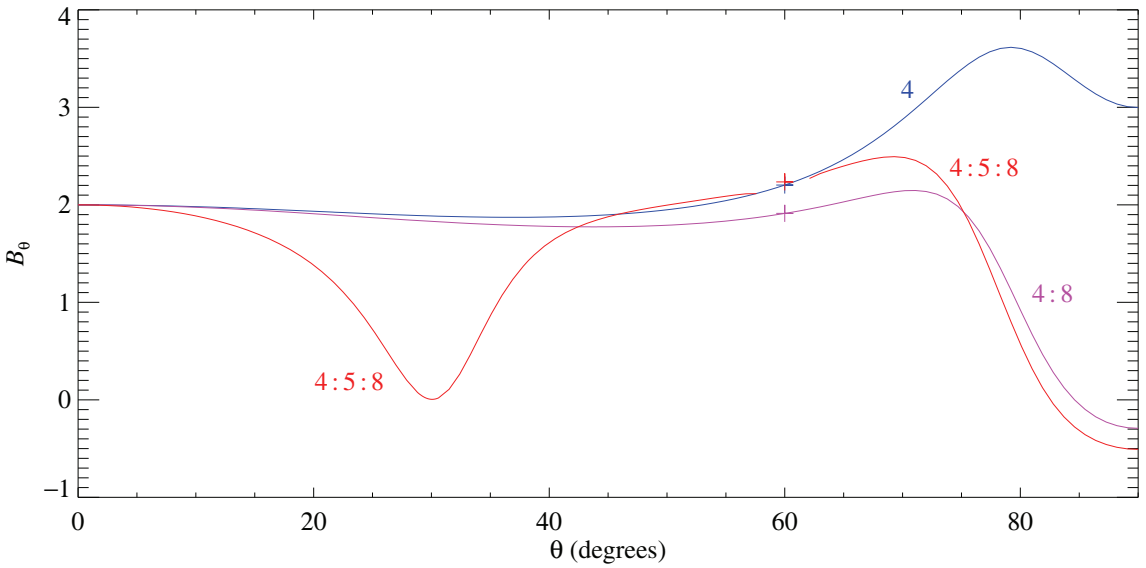


Figure 8. *Cross-coupling coefficient B_θ for the parameters from Figure 7 and nonlinear coefficients $Q_1 = 1+i$, $Q_2 = -2+2i$, $C = -1+10i$. The blue curve is with 4 forcing only: the peak near 80° occurs because of 1 : 2 resonance with frequency 8. The magenta curve has 4 : 8 forcing, using the frequency 8 component to bring down the curve close to 90°. Bringing in the 5 frequency (red curve) gives a pronounced dip at 30°, corresponding to the difference frequency (5 − 4), even though this mode is not driven directly. The relevant coefficients are $B_{30} = -0.01$, $B_{60} = 2.24$, and $B_{90} = -0.51$.*

on the figure, and discussed in more detail below) in the same range. Squares are also stable above 1.0013 times critical. The agreement is good, better than the examples in the previous section (Figure 6), since the difference frequency mode is still fairly well damped (compare the damping for k close to 0.378 in Figure 3(f) and $k = 0.518$ in Figure 7(f)).

7.1. Choice of domain size for approximate quasipatterns. Before presenting numerical solutions in large domains, we discuss the issue of choice of domain for providing accurate approximations to quasipatterns.

Just as hexagonal patterns can be approximated in rectangular domains [39], there are many ways of choosing periodic domains to allow accurate approximations of quasipatterns. Here we discuss three plausible approaches to choosing domains for 12-fold quasipatterns. We show why one approach, based on Pythagorean triplets, does not work at all, while two other approaches both work well.

Reducible symmetry group representations with square periodic domains have been put forward as candidates for producing approximate quasipatterns [15, 59]. In particular, Pythagorean triplets have been identified as of particular interest in this case [15]. Consider a pair of integers (p, q) with $p > q > 0$. Then $(p^2 - q^2, 2pq, p^2 + q^2)$ forms a Pythagorean triplet (i.e., $(p^2-q^2)^2+(2pq)^2 = (p^2+q^2)^2$), and the vectors $\boldsymbol{k}^{(1)} = (1, 0)$ and $\boldsymbol{k}^{(2)} = (2pq, p^2-q^2)/(p^2+q^2)$ have the same length. For example, with $p = 7$, $q = 4$, we have $\boldsymbol{k}^{(1)} = (1, 0)$ and $\boldsymbol{k}^{(2)} = (\frac{56}{65}, \frac{33}{65})$. If $\frac{p}{q}$ is a continued fraction approximation of $\sqrt{3}$ (see Table 1), then the angle between these two vectors, namely $\tan^{-1}((p^2 - q^2)/(2pq))$, tends to $30° + \mathcal{O}(1/q^2)$ as the approximation

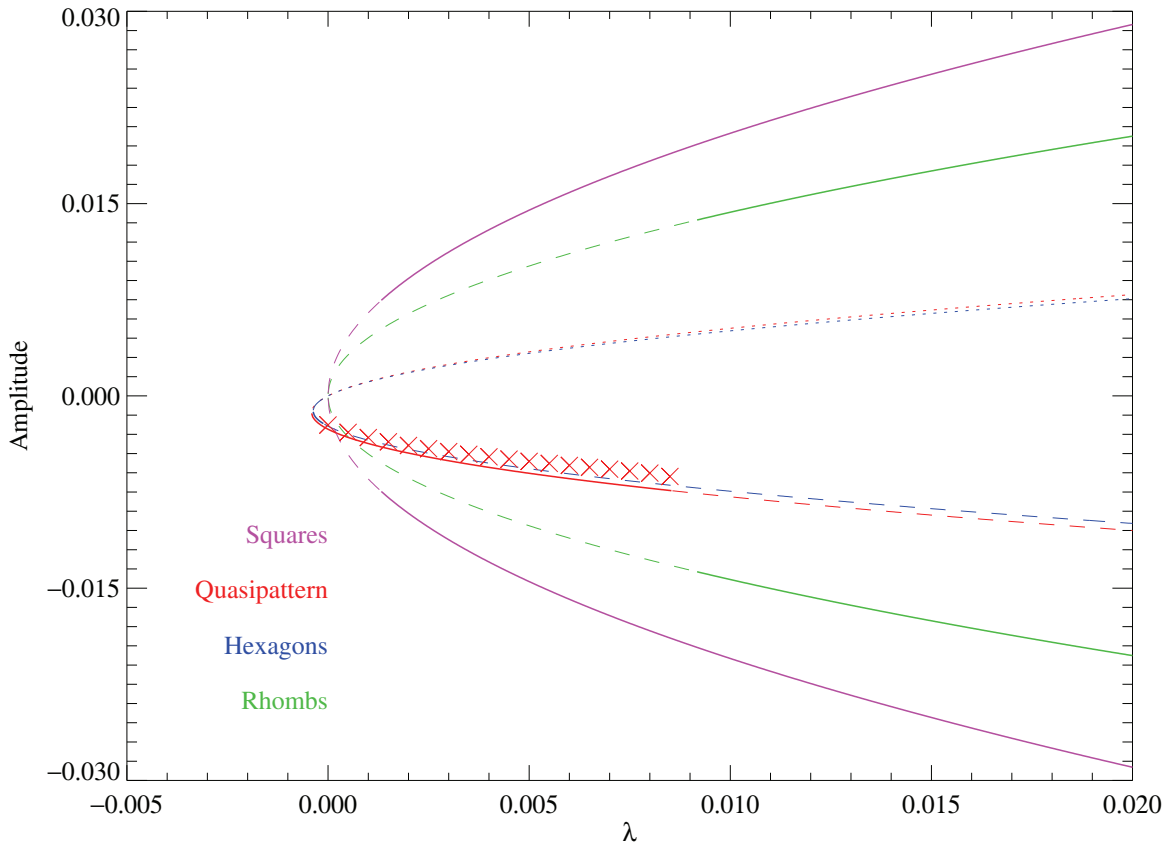


Figure 9. *Bifurcation diagram with 3-frequency* $4:5:8$ *forcing based on a 12-amplitude cubic truncation, with parameter values as in Figure 8. 12-fold quasipatterns (red) are predicted to be stable up to* 1.0085 *times critical, squares (magenta) are stable from* 1.0013 *times critical, and* 30° *rhombs (green) from* 1.0092 *times critical. The crosses show the amplitudes of stable approximate quasipattern solutions of the PDE, calculated in a domain* $2\sqrt{13} \times 2\sqrt{13/3}$ *critical wavelengths.*

Table 1
Continued fraction approximations to $\sqrt{3}$.

$\frac{p}{q}$	$\frac{2}{1}, \frac{5}{3}, \frac{7}{4}, \frac{19}{11}, \frac{26}{15}, \frac{71}{41}, \frac{97}{56}, \frac{265}{153}, \frac{362}{209}, \cdots \to \sqrt{3}$

to $\sqrt{3}$ improves. Thus it might be thought that square $(p^2+q^2)\times(p^2+q^2)$ domains might readily allow approximations to 12-fold quasipatterns.

Unfortunately, these Pythagorean domains do not provide good approximations to quasipatterns, because the essential 60° coupling is not quite correct: consider the wavevector $\boldsymbol{k}^{(1)} = (p^2+q^2, 0)$ (in units of the basic lattice vector), and, at 60° on either side of $\boldsymbol{k}^{(1)}$, the wavevectors $\boldsymbol{k}^{(3)} = (p^2-q^2, 2pq)$ and $\boldsymbol{k}^{(11)} = (p^2-q^2, -2pq)$, where we number wavevectors counterclockwise from $\boldsymbol{k}^{(1)}$. In order for quadratic interactions between these three modes to occur, we need $\boldsymbol{k}^{(1)} = \boldsymbol{k}^{(3)} + \boldsymbol{k}^{(11)}$. However, in this case, $\boldsymbol{k}^{(3)} + \boldsymbol{k}^{(11)} = (2(p^2-q^2), 0) = \boldsymbol{k}^{(1)} + (p^2-3q^2, 0)$, which is close to, but never equal to, $\boldsymbol{k}^{(1)}$: it can be shown that $p^2-3q^2 = 1$

or -2, which is small compared to $|\boldsymbol{k}^{(1)}| = p^2 + q^2$. This small difference means that the important nonlinear interactions between these three waves generate erroneous long-wave modulations in square $(p^2 + q^2) \times (p^2 + q^2)$ domains, and if modes with wavenumber close to zero are not heavily damped, these long-wave modulations can dominate the pattern.

A much better way of generating good approximations to 12-fold quasipatterns is to choose $2q \times 2q$ domains, with vectors $\boldsymbol{k}^{(1)} = (1, 0)$ and $\boldsymbol{k}^{(2)} = (p, q)/2q$, with $\frac{p}{q}$ drawn from Table 1. Again, the angle between these vectors goes as $30^\circ + \mathcal{O}(1/q^2)$, and, though the wavenumbers are not quite equal, we have $|\boldsymbol{k}^{(2)}| \to 1 + \mathcal{O}(1/q^2)$. This approach works because with this choice of wavevectors we have the correct three-wave coupling: $\boldsymbol{k}^{(3)} = (q, p)/2q$, $\boldsymbol{k}^{(11)} = (q, -p)/2q$, and so $\boldsymbol{k}^{(3)} + \boldsymbol{k}^{(11)} = \boldsymbol{k}^{(1)}$.

A third possibility is to approximate the quasipattern as a superlattice pattern, using the 12-dimensional irreducible representations of $D_6 \ltimes T^2$ [18]. If we choose $\alpha = q$ and $\beta = p - q$ (in the notation of [18, Table 2]), with $\frac{p}{q}$ drawn from Table 1, then $\boldsymbol{k}^{(1)} = ((p-q)\sqrt{3}, 3q-p)$, $\boldsymbol{k}^{(2)} = ((2q-p)\sqrt{3}, p)$, $\boldsymbol{k}^{(3)} = ((p-2q)\sqrt{3}, p)$, $\boldsymbol{k}^{(11)} = (q\sqrt{3}, 3q-2p)$ (all these should be divided by their length, $\sqrt{p^2 - 3pq + 3q^2}$). We have $\boldsymbol{k}^{(3)} + \boldsymbol{k}^{(11)} = \boldsymbol{k}^{(1)}$ as required, all wavevectors are the same length (which is an advantage over the second alternative), and the angle between $\boldsymbol{k}^{(1)}$ and $\boldsymbol{k}^{(2)}$ goes as $30^\circ + \mathcal{O}(1/q^2)$. One disadvantage of this approach is that numerical solutions of the PDE must be carried out in $2\sqrt{p^2 - 3pq + 3q^2} \times 2\sqrt{(p^2 - 3pq + 3q^2)/3}$ rectangular domains in order to take advantage of spectral numerical methods. These domains are big enough to contain two repeats of the pattern (as in Figure 5(b)), and so only half the Fourier coefficients are used.

The last two methods are compared in Table 2. The error between the approximation and the 12-fold quasipattern is inversely proportional to the area of the domain in both cases. The hexagonal superlattice approximations have all wavevectors of the correct length, unlike the square approximation. For this reason, the superlattice case is better for computing bifurcation diagrams, as all modes bifurcate at the same value of the forcing.

However, for similar domain areas, the angular separation between wavevectors is about four times closer to 30° in the square case than in the superlattice case. Moreover, the square case is more amenable to efficient use of fast Fourier transforms, since the number of modes can be chosen to be a power of two in each direction, while the superlattice case leads to more awkward choices of numbers of modes.

For these reasons, we choose the rectangular $2\sqrt{13} \times 2\sqrt{13/3}$ domain for the bifurcation diagram in Figure 9, since the quantitative comparison between numerical simulation and analysis is easier when all modes have the same wavenumber and hence bifurcate at the same value of the forcing (see below). We use the more convenient square (8×8, 30×30, and 112×112) domains for the remaining PDE simulations described below. It would be interesting to see how sensitive the quasipattern is to the exact choice of domain size. It would also be interesting to find domains that provide particularly accurate approximations to 8-, 10-, 14-fold and higher-order quasipatterns, but these issues are beyond the scope of this paper.

One consequence of the wavevectors being of unequal length in the square case is that the two wavenumbers concerned have slightly different critical forcings. In the 30×30 and 112×112 cases, this difference is negligible, but in the 8×8 case, the critical forcings for

Table 2

Domains that provide good approximations to 12-fold quasipatterns. The first column gives the rational number $\frac{p}{q}$ that is a continued fraction approximation to $\sqrt{3}$, drawn from Table 1. The second and third columns give the area of a computational domain and two of the wavevectors that will make up an approximate quasipattern. The fourth and fifth columns demonstrate that the errors in the length of $\boldsymbol{k}^{(2)}$ and in the angle $\angle_{12}$ between $\boldsymbol{k}^{(1)}$ and $\boldsymbol{k}^{(2)}$ scale as $1/A$, or equivalently, as q^{-2}. The first set of rows refer to rectangular domains of size $2\sqrt{p^2-3pq+3q^2}\times 2\sqrt{(p^2-3pq+3q^2)/3}$, which allow superlattice patterns that approximate quasipatterns. In these examples, all wavevectors are the same length. The second set of rows refer to square domains of size $2q\times 2q$. In these domains, approximate quasipatterns have wavevectors that have two slightly different lengths, with one wavevector being $\boldsymbol{k}^{(1)}=(1,0)$.

p/q	$A=$ Area	First two wavevectors	$(\lvert\boldsymbol{k}^{(2)}\rvert-1)A$	$(\angle_{12}-30^\circ)A$
$\frac{p}{q}$	$\frac{4(p^2-3pq+3q^2)}{\sqrt{3}}$	$\boldsymbol{k}^{(1)}=\frac{((p-q)\sqrt{3},3q-p)}{2\sqrt{p^2-3pq+3q^2}}$, $\boldsymbol{k}^{(2)}=\frac{((2q-p)\sqrt{3},p)}{2\sqrt{p^2-3pq+3q^2}}$		
$\frac{7}{4}$	30.02	$\frac{(3\sqrt{3},5)}{\sqrt{52}}$, $\frac{(\sqrt{3},7)}{\sqrt{52}}$	0	66.2
$\frac{19}{11}$	224.01	$\frac{(8\sqrt{3},14)}{\sqrt{338}}$, $\frac{(3\sqrt{3},19)}{\sqrt{338}}$	0	−132.3
$\frac{26}{15}$	418.00	$\frac{(11\sqrt{3},19)}{\sqrt{724}}$, $\frac{(4\sqrt{3},26)}{\sqrt{724}}$	0	66.2
$\frac{71}{41}$	3120.0	$\frac{(30\sqrt{3},52)}{\sqrt{5404}}$, $\frac{(11\sqrt{3},71)}{\sqrt{5404}}$	0	−132.3
$\frac{97}{56}$	5822.0	$\frac{(41\sqrt{3},71)}{\sqrt{10084}}$, $\frac{(15\sqrt{3},97)}{\sqrt{10084}}$	0	66.2
⋮				
$\sqrt{3}$	∞	$\frac{(1,1)}{\sqrt{2}}$, $\frac{(\sqrt{3}-1,\sqrt{3}+1)}{2\sqrt{2}}$		
$\frac{p}{q}$	$2q\times 2q$	$\boldsymbol{k}^{(2)}=\left(\frac{p}{2q},\frac{1}{2}\right)$		
$\frac{7}{4}$	64	$\boldsymbol{k}^{(2)}=\left(\frac{7}{8},\frac{1}{2}\right)$	0.498	−16.3
$\frac{19}{11}$	484	$\boldsymbol{k}^{(2)}=\left(\frac{19}{22},\frac{1}{2}\right)$	−1.001	33.2
$\frac{26}{15}$	900	$\boldsymbol{k}^{(2)}=\left(\frac{26}{30},\frac{1}{2}\right)$	0.500	−16.5
$\frac{71}{41}$	6724	$\boldsymbol{k}^{(2)}=\left(\frac{71}{82},\frac{1}{2}\right)$	−1.000	33.1
$\frac{97}{56}$	12544	$\boldsymbol{k}^{(2)}=\left(\frac{97}{112},\frac{1}{2}\right)$	0.500	−16.5
⋮				
$\sqrt{3}$	∞	$\boldsymbol{k}^{(2)}=\left(\frac{\sqrt{3}}{2},\frac{1}{2}\right)$		

the two wavenumbers are appreciably different. We therefore make a small adjustment to the domain size to make the two critical forcings the same, raised by a factor of 1.00066 above F_c.

7.2. Numerical examples of 12-fold quasipatterns. Numerical solutions of the PDE (3.1) with $4:5:8$ forcing at 1.003 times critical are shown in Figure 10, in periodic domains 8×8, 30×30, and 112×112 wavelengths with periodic boundary conditions and using up to 1536^2 Fourier modes. The solutions were followed for at least 10,000 forcing periods in the largest domain. Most initial conditions resulted in square patterns, but minor adjustments to the Fourier amplitudes at an early stage of the calculation resulted in stable approximate 12-fold quasipatterns. Note, however, that the PDE solutions in Figure 10 were not constrained to

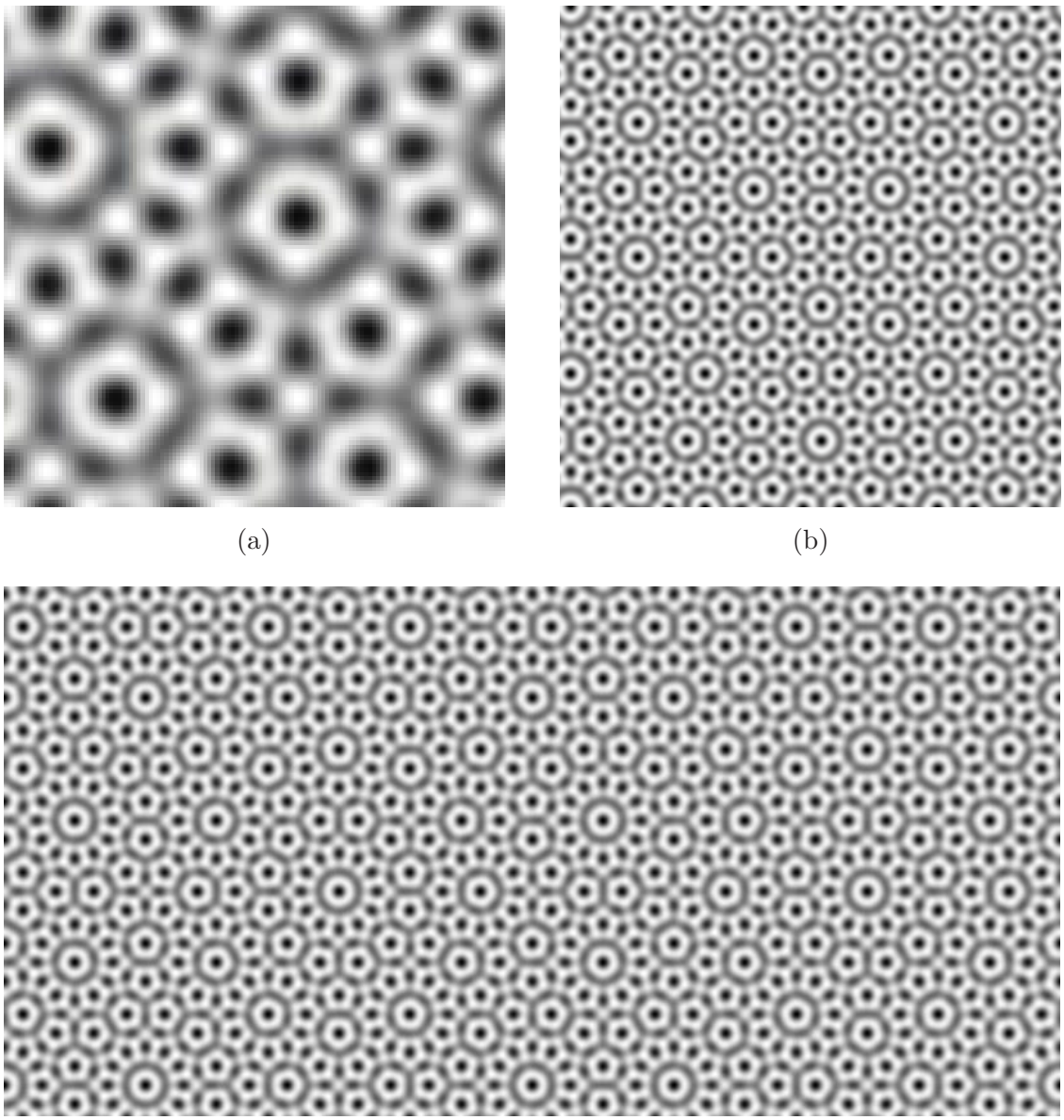


Figure 10. *Numerical solutions of the PDE with* $4:5:8$ *forcing at* 1.003 *times critical.* (a) 8×8 *domain.* (b) 30×30 *domain.* (c) 112×112 *domain (only a* 60×30 *section of the domain is shown). Parameter values are as in Figure* 8.

choose exactly the wavevectors given in Table 2.

The accuracy of the approximation improves with increasing domain size. The modes involved in the 30×30 example are $(30,0)$ and $(26,15)$ and their reflections, which are 29.98° apart and differ in length by 0.05%. The larger 112×112 domain allows an improved approximation to the quasipattern: the important wavevectors are $(112,0)$ and $(97,56)$, which are

29.9987° apart and differ in length by 0.004%. The amplitudes of these modes differ by 1%. We discuss other ways of evaluating the improved approximation to quasiperiodicity in the next section.

7.3. Fourier spectra of quasipatterns. All the numerical PDE solutions presented here have been carried out in periodic domains, so these solutions are only approximations to quasipatterns. In the 12-fold examples (Figures 10 and 18(a) below), the most important twelve wavevectors in the pattern are not exactly 30° apart and do not have exactly the same length (see Table 2). On the other hand, it is possible to generate true quasipatterns using twelve modes with $k = 1$ evenly spaced around the unit circle, but the asymptotic series in the weakly nonlinear approximation are known to diverge [52]. In this section, we compare the detailed Fourier spectra of the approximate quasipatterns to see how these differ from the spectra of true quasipatterns.

We make the comparison by computing the locations of modes generated by nonlinear interactions up to a certain order, in the three cases of approximate quasipatterns in Figure 10 as well as in a true quasipattern. To do this, we first define the *order* of a mode.

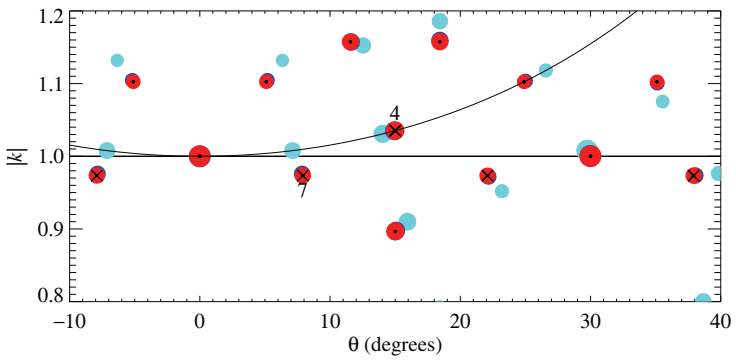
Quadratic interactions between the twelve modes with wavevectors $\boldsymbol{k}^{(1)}, \ldots, \boldsymbol{k}^{(12)}$ generate new modes with wavevectors $2\boldsymbol{k}^{(1)}$, $\boldsymbol{k}^{(1)} + \boldsymbol{k}^{(2)}$, $\boldsymbol{k}^{(1)} + \boldsymbol{k}^{(3)}$ and so on. Likewise, higher-order nonlinear interactions between the twelve modes generate modes with wavevectors $\boldsymbol{k}^{(\boldsymbol{m})}$:

$$\boldsymbol{k}^{(\boldsymbol{m})} = \sum_{j=1}^{12} m_j \boldsymbol{k}^{(j)}, \tag{7.1}$$

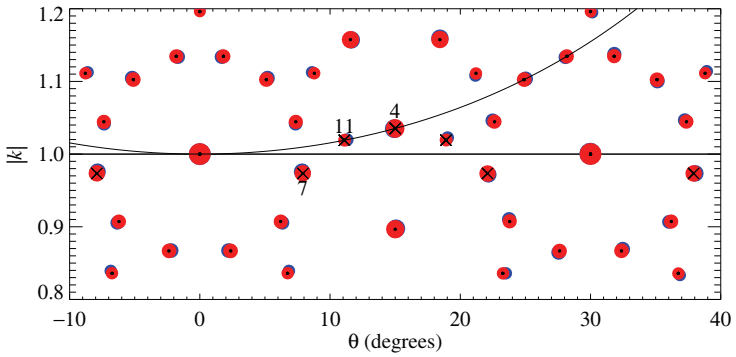
where the m_j's are nonnegative integers adding up to the *order* of the nonlinear interaction $|\boldsymbol{m}|$, defined by $|\boldsymbol{m}| = \sum_{j=1}^{12} m_j$. In the case of a periodic domain, the set of all possible $\boldsymbol{k}^{(\boldsymbol{m})}$ defines a lattice of wavevectors (Figure 2(a,b)), while in the quasipattern case, the set of all possible $\boldsymbol{k}^{(\boldsymbol{m})}$ defines a *quasilattice*: examples of 12- and 14-fold quasilattices with $|\boldsymbol{m}| \leq 11$ ($|\boldsymbol{m}| \leq 7$ in the 14-fold case) are shown in Figure 2(c,d). If a wavevector $\boldsymbol{k}$ is in the lattice or quasilattice, then $\boldsymbol{k} = \boldsymbol{k}^{(\boldsymbol{m})}$ for an infinite number of choices of vector $\boldsymbol{m}$ (corresponding to nonlinear interactions of different orders), since increasing m_1 and m_7 (say) by the same amount does not change $\boldsymbol{k}^{(\boldsymbol{m})}$ but increases $|\boldsymbol{m}|$ by 2. We define the *order* of wavevector $\boldsymbol{k}$ to be the smallest value of $|\boldsymbol{m}|$ with $\boldsymbol{k}^{(\boldsymbol{m})} = \boldsymbol{k}$.

In Figures 11 and 12 we compare the locations of wavevectors in the 8×8 (cyan), 30×30 (blue), and 112×112 (red) examples from the simulation results shown in Figure 10. The 8×8 spectrum is only in Figure 11(a) (up to 7th order), and the 30×30 spectra are only up to 26th order. The amplitude of each Fourier mode is given by the size of the symbol (on a logarithmic scale): the largest symbols are the modes with the largest amplitudes, and the modes with the smallest symbols have amplitudes 10^{17} times smaller. Modes with amplitudes smaller than this are not plotted, as these are in the realm of round-off error (see below).

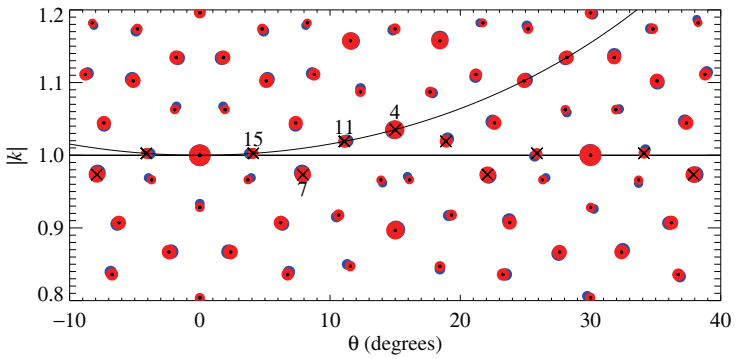
The success with which the periodic patterns approximate a true quasipattern can be judged by the locations of the important modes in the pattern, as compared to the locations of modes on the quasilattice, up to a given order. Up to 7th order (Figure 11(a)), the 30×30 and 112×112 modes (blue and red) overlie each other almost exactly and correspond well with the quasilattice modes (small black dots at the centers of the red markers). However,



(a)



(b)



(c)

Figure 11. *Fourier spectra of* 8×8 *(cyan),* 30×30 *(blue),* 112×112 *(red) approximate* 12*-fold quasipatterns, as well as the* 12*-fold quasilattice (small black dots), up to* (a) 7*th,* (b) 11*th, and* (c) 15*th order. The size of the colored marker indicates the amplitude of the corresponding Fourier mode on a logarithmic scale, with the smallest markers having* 10^{17} *times smaller amplitude than the largest. Only Fourier modes close to the unit circle* $(0.8 \le |\boldsymbol{k}| \le 1.2)$ *are shown, with wavevector angles* $-10^\circ \le \theta \le 40^\circ$*. The horizontal line* $k = 1$ *represents the unit circle, while the curved line represents* $k_x = 1$*. Wavevectors that come closest to the unit circle up to a particular order are labelled with* $\times$*.*

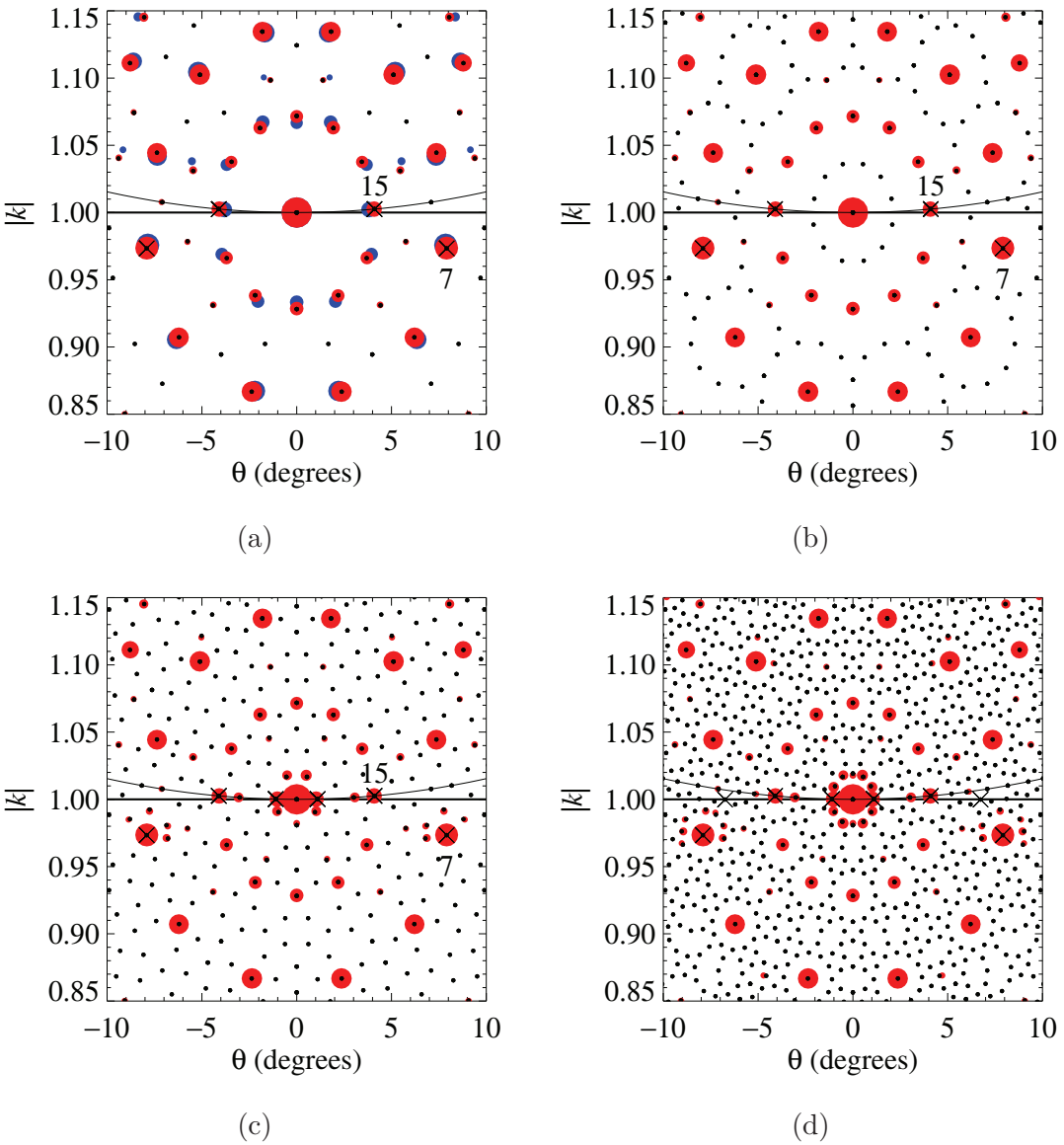


Figure 12. *Fourier spectra of* 30×30 *(blue) and* 112×112 *(red) approximate* 12*-fold quasipatterns, as well as the* 12*-fold quasilattice (small black dots), up to* (a) 26*th,* (b) 39*th,* (c) 56*th, and* (d) 94*th order. Here we show wavevectors within* 10° *of* $(1, 0)$*, and with* $0.85 \le |\boldsymbol{k}| \le 1.15$.

the 8×8 modes (cyan) deviate substantially from the correct locations, and we conclude that the periodic pattern in an 8×8 domain is a poor approximation to a quasipattern (despite appearances in Figure 10(a)).

At 11th order (Figure 11(b)), the agreement between 30×30, 112×112, and the true quasipattern is still good, while at 15th order (Figure 11(c)), the 30×30 modes deviate noticeably from the 112×112 and quasipattern modes. This is more pronounced at 26th order (Figure 12(a)).

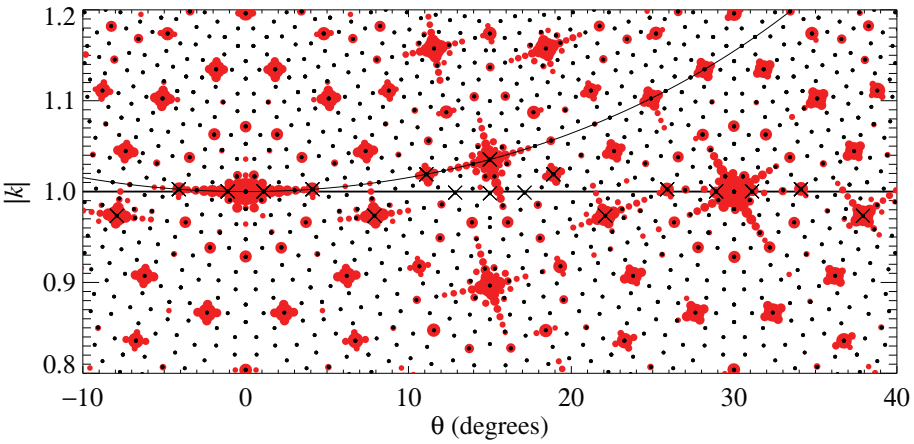


Figure 13. *Fourier spectra of the* 112×112 *(red) approximate* 12*-fold quasipattern, as well as the* 12*-fold quasilattice (small black dots). All modes with amplitude greater than* 10^{-17} *times the maximum amplitude are plotted; the quasilattice is plotted up to* 56*th order. The competing effects of the* 12*-fold order and the square lattice are apparent.*

At 15th order (Figure 11(c)), the agreement between 112×112 and the true quasipattern is excellent: every red marker has a black dot at its center. The agreement is still very good at 26th order (Figure 12(a)): the black dots are not quite in the centers of the smallest red markers, and some black dots do not have corresponding red markers. This means that those modes, present in the true quasipattern at this order, have amplitudes in the 112×112 approximation that are too small to be plotted. The situation at 39th order (Figure 12(b)) is similar, but it is not until 94th order (Figure 12(d)) that the true quasipattern modes miss the centers of the red markers entirely.

In fact, a curious situation arises at 56th and 94th order: it is apparent that modes close to $\boldsymbol{k} = (1, 0)$, and in particular modes on the line $k_x = 1$, have amplitudes that are higher than might be expected, since usually the amplitudes of modes decrease with their order. These modes are discussed in more detail below. However, it should be noted that plotting only modes up to a certain order masks the effect of the underlying lattice in the numerical solutions. In Figure 13, we show all modes in the 112×112 down to round-off error, and quasilattice modes up to 56th order. The underlying square numerical lattice is clearly seen in the strings of red markers emanating from each large-amplitude mode. These give an impression that the large-amplitude modes could be considered to be clusters of modes in Fourier space.

In Figure 14, we show the range of amplitudes of the modes as a function of order, for the 8×8, 30×30, and 112×112 approximate quasipatterns. We note that the 8×8 example deviates significantly from the other two at all orders, while the 30×30 and 112×112 examples are in fairly good agreement until 19th order or so, another indication of how the approximation improves with the domain size. We note that direct comparisons between the 8×8 case and the other two cases are not strictly valid, as the domain size in the 8×8 case had to be adjusted slightly to allow for the different wavenumbers in the pattern, as discussed above.

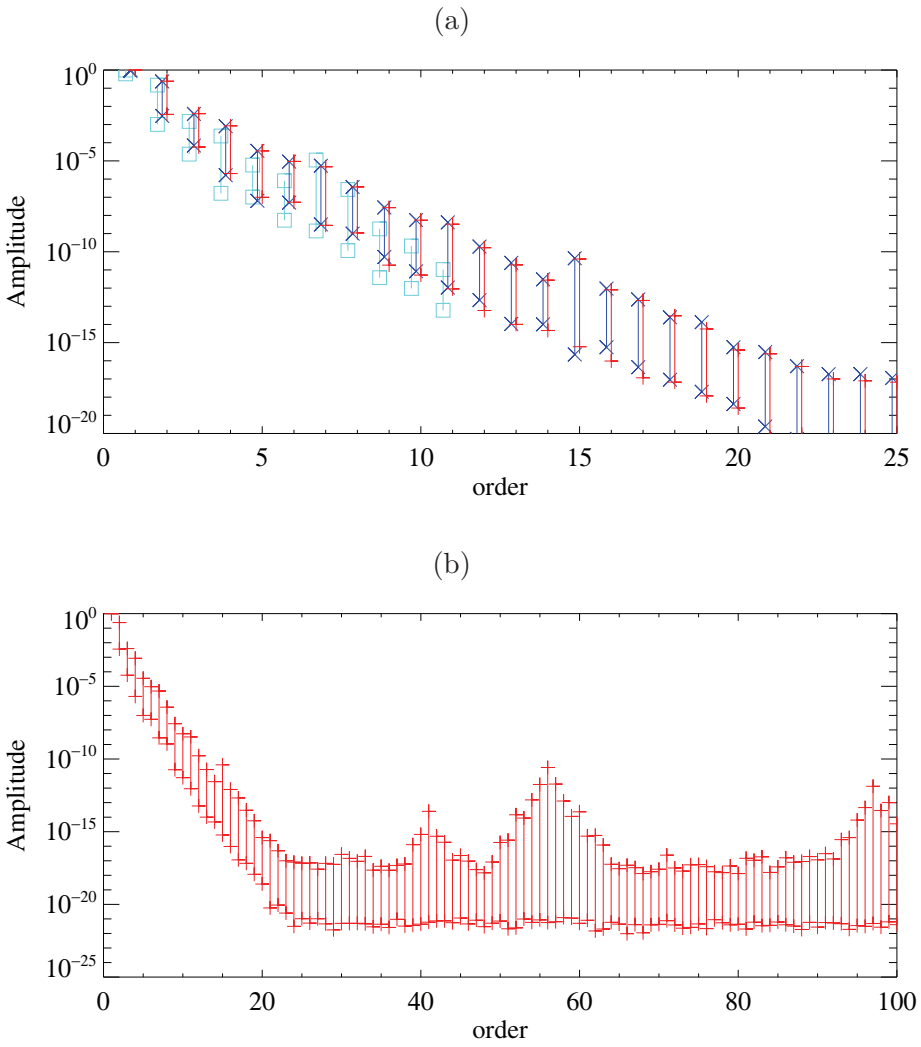


Figure 14. *Fourier spectra of* 8×8 *(cyan),* 30×30 *(blue), and* 112×112 *(red) approximate* 12*-fold quasipatterns, as a function of order,* (a) *up to* 25*th order,* (b) *up to* 100*th order (*112×112 *only). The vertical lines indicate the range of amplitudes of modes at that order, scaled to the maximum amplitude at order* 1.

Figure 14(b) shows the spectrum of the 112×112 example at all orders up to 100. The range of amplitudes at a given order decays exponentially with order up to about 23rd order, but then levels off at the level of the round-off error: about 10^{-17}. On top of this pattern, there is a peak in amplitude at 15th order, and broad peaks around 41st, 56th, and 97th order. We attribute these to the presence of weakly damped modes, with $\boldsymbol{k}$ close to unity, that appear around these orders. With weak damping, these modes will amplify the numerical noise in the PDE solutions, and nonlinear coupling implies that the amplified noise will feed in to nearby modes.

This is shown in Figure 15, where we plot $\left||\boldsymbol{k}^{(\boldsymbol{m})}| - 1\right|$ as a function of order $|\boldsymbol{m}|$ for the different examples. We note the marked drop in this quantity in particular at 15th and 39th

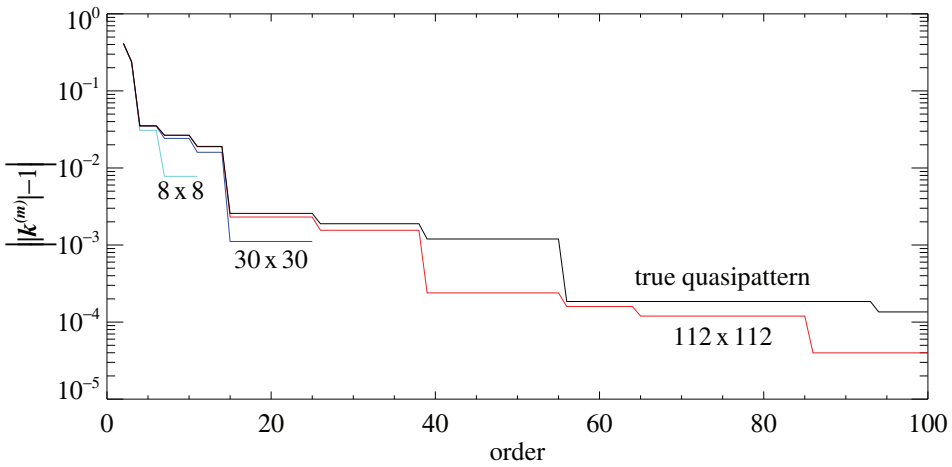


Figure 15. *Minimum value of* $||\boldsymbol{k}^{(m)}| - 1|$*, as a function of order* $|\boldsymbol{m}|$*, for* 8×8 *(cyan),* 30×30 *(blue),* 112×112 *(red) approximate* 12*-fold quasipatterns, and for a true* 12*-fold quasipattern (black) [52]. There are drops in this minimum quantity at* 4*th,* 7*th,* 11*th,* 15*th,* 26*th,* 39*th, and* 56*th order in all cases. Modes that are responsible for these drops are identified in Figures 11 and 12. In the* 30×30 *case, the Pythagorean mode* $(0.6, 0.8)$ *(which is of unit length) is generated at* 26*th order, which is why that curve terminates. The* 8×8 *curve continues horizontally.*

order. While the drops in $\big||\boldsymbol{k}^{(m)}| - 1\big|$ do not line up exactly with the peaks in amplitude, we suspect that it is this marked change in the damping of modes appearing at these orders that is responsible for the amplification of noise at around the same order.

Of course, for the numerical patterns in periodic domains, there is a lower bound to $\big||\boldsymbol{k}^{(m)}| - 1\big|$ that does not depend on $|\boldsymbol{m}|$ (this lower bound is captured within Figure 15), while for the true 12-fold quasipattern, there is no lower bound. If $|\boldsymbol{k}^{(m)}| \neq 1$, then

$$\big||\boldsymbol{k}^{(m)}| - 1\big| > \frac{B}{|\boldsymbol{m}|^2}, \tag{7.2}$$

where B is an order-one constant [52]. For some particular choices of wavevectors, this asymptotic limit is achieved: for example, if

$$\boldsymbol{k}^{(m)} = p\boldsymbol{k}^{(4)} + (q-1)\boldsymbol{k}^{(9)} + (q+1)\boldsymbol{k}^{(11)} = (1, p - \sqrt{3}q), \tag{7.3}$$

where p and q are integers, then $|\boldsymbol{m}| = p + 2q$ and $|\boldsymbol{k}^{(m)}|^2 = 1 + (p - \sqrt{3}q)^2$. When $\frac{p}{q}$ is a continued fraction approximation to $\sqrt{3}$ (Table 1), then $\big||\boldsymbol{k}^{(m)}| - 1\big| \sim \frac{1}{2}(p - \sqrt{3}q)^2 \leq \frac{B_2}{q^2}$, so $\big||\boldsymbol{k}^{(m)}| - 1\big| \leq \frac{B'}{|\boldsymbol{m}|^2}$, where B_2 and B' are order-one constants [52]. These particular vectors are not always the closest ones that can be found at a given order, but they demonstrate that modes approach the unit circle arbitrarily closely (and so are arbitrarily weakly damped) as the order of the mode increases. For the fractions listed in Table 1, we have $|\boldsymbol{m}| = p + 2q = 4, 11, 15, 41, 56, 153, \ldots$, but in fact there are step decreases in the minimum of $\big||\boldsymbol{k}^{(m)}| - 1\big|$ at $|\boldsymbol{m}| = 4, 7, 11, 15, 26, 39, 56, 94$. This is the reason for the choices of orders in Figures 11 and 12.

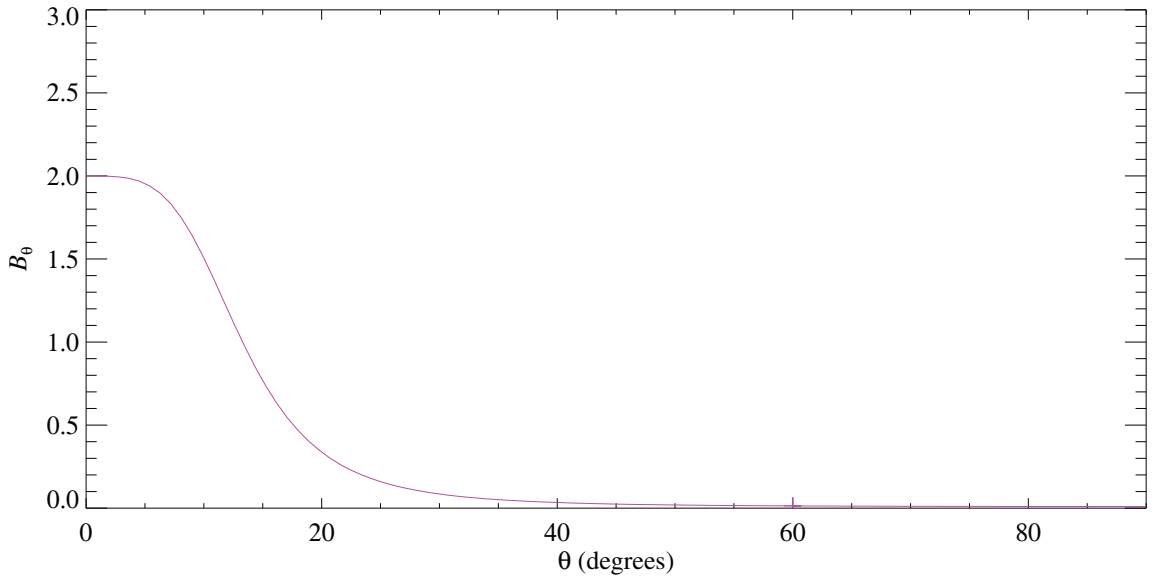


Figure 16. *Cross-coupling coefficient B_θ for single frequency forcing, with $1:2$ resonance in space and time: $\omega = \frac{1}{3}$, $\beta = -\frac{1}{6}$, $\delta = 0$, $\mu = -0.005$, $\alpha = 0.001$, $\gamma = 0$, $Q_1 = 3 + 4i$, $Q_2 = -6 + 8i$, $C = -1 + 10i$, $F_c = 0.024002$, and $k_c = 0.9999$. Note that B_θ drops away sharply as θ increases, and is close to zero for $\theta \geq 30^\circ$. The relevant coefficients are $B_{30} = 0.088$, $B_{60} = 0.014$, and $B_{90} = 0.010$.*

In summary, having identified which sizes result in the most accurate approximations to 12-fold quasipatterns, we find that in the largest example (112×112), the locations of the Fourier modes of the approximation deviate significantly from those of the true quasipattern only beyond 26th order. At this point, the amplitudes of the modes have reached the level of numerical round-off, so going any larger than 112×112 would not lead to any significant improvement in the approximation to a true quasipattern for these parameter values. The small divisors reveal themselves by amplifying the amplitudes of the Fourier modes at (or close to) the order at which the small divisor appears, but they do not appear to cause the amplitudes of the Fourier modes to become excessively large, at least at this level of forcing, and up to the maximum order (153) available in this domain.

8. Turbulent crystals: Quasipatterns using $1:2$ resonance. In order to have $1:2$ resonance in space and time with single frequency forcing ($m = 1$), we impose $\hat{\Omega}(1) = \frac{1}{2}$ and $\hat{\Omega}(2) = 1$, which leads to $\omega = \frac{1}{3} + 4\delta$ and $\beta = -\frac{1}{6} + 5\delta$. We choose $\delta = 0$; small values for the damping coefficients μ, α, and γ; and order one values for the nonlinear coefficients close to the Hamiltonian limit. The resulting cross-coupling curve is shown in Figure 16: as explained in section 2, the $1:2$ resonance in space and time has enhanced the self-coupling coefficient (by about four orders of magnitude compared to the previous cases), and so the cross-coupling coefficient B_θ drops away sharply and is close to zero for $\theta \geq 30^\circ$.

Within the restrictions of a 12-mode expansion, 12-fold quasipatterns are stable (Figure 17). Indeed, at 1.1 times the critical forcing, in a 30×30 domain, the numerical solution of the PDE with random initial conditions is a stable 12-fold approximate quasipattern (Figure 18(a)). As above, the primary modes that make up the pattern are $(30, 0)$ and $(26, 15)$

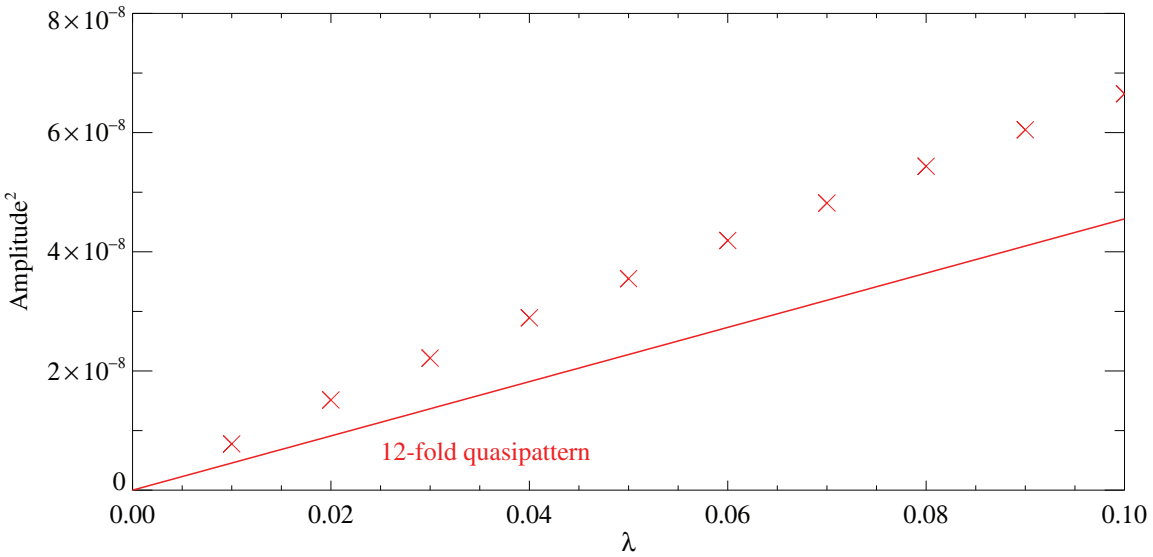


Figure 17. *Bifurcation diagram showing the weakly nonlinear predicted amplitude (solid line) and amplitudes of approximate quasipattern solutions of the PDEs in an 8×8 domain (crosses).*

and their reflections, in units of basic lattice vectors. The amplitudes of the 12 modes differ by 0.5%. The initial condition was not in any invariant subspace, and the PDE was integrated for 160,000 periods of the forcing. The agreement between the weakly nonlinear predictions and the computed amplitudes is not good (Figure 17), which we expect since the $k = 2$ mode is weakly damped (see discussion above).

However, there is no feature in the cross-coupling curve (Figure 16) to indicate that modes at 30° should enjoy a special status. When the forcing is changed to 1.3 times critical with this 12-fold quasipattern as the initial condition, we find that it is unstable and is replaced (after a transient of 70,000 periods) by a stable approximate 14-fold quasipattern (Figures 18(b) and 19, and 71906_01.gif [19.1MB]). In this case, the 14 modes are $(30, 0)$, $(27, 13)$, $(19, 23)$, $(7, 29)$, and their reflections, differing in length by 0.5% and having angular separation within 1.5° of 360°/14, with spacings ranging from 24.7° to 27.1°. The amplitudes differ by about 10%. Both the 12-fold and 14-fold quasipatterns are stable at 1.2 times critical, in a 30×30 domain.

More complex quasipatterns are also possible: calculations done in larger 60×60 and 90×90 domains at 1.3 times critical, starting with random initial conditions, both yield approximate 20-fold quasipatterns (Figure 18(c)). However, these are not particularly accurate approximations to 20-fold quasipatterns: in the 60×60 case, the 20 modes are $(60, 0)$, $(57, 18)$, $(48, 36)$, $(37, 47)$, $(21, 56)$, $(1, 60)$, $(-18, 57)$, $(-33, 50)$, $(-47, 37)$, $(-56, 21)$, and their 180° rotations, which differ in length by 0.4% and which have angular separations within 3° of 360°/20, with spacings ranging from 14.9° to 20.5°. The amplitudes of the 20 modes differ by up to 40%. Figure 20 shows an examination of the Fourier spectrum of Figure 18(c): there are 20 peaks close to $k = 1$, of similar amplitudes and arranged roughly evenly around the unit circle. The 90×90 example is similar: the 20 modes differ in length by 0.2% and also

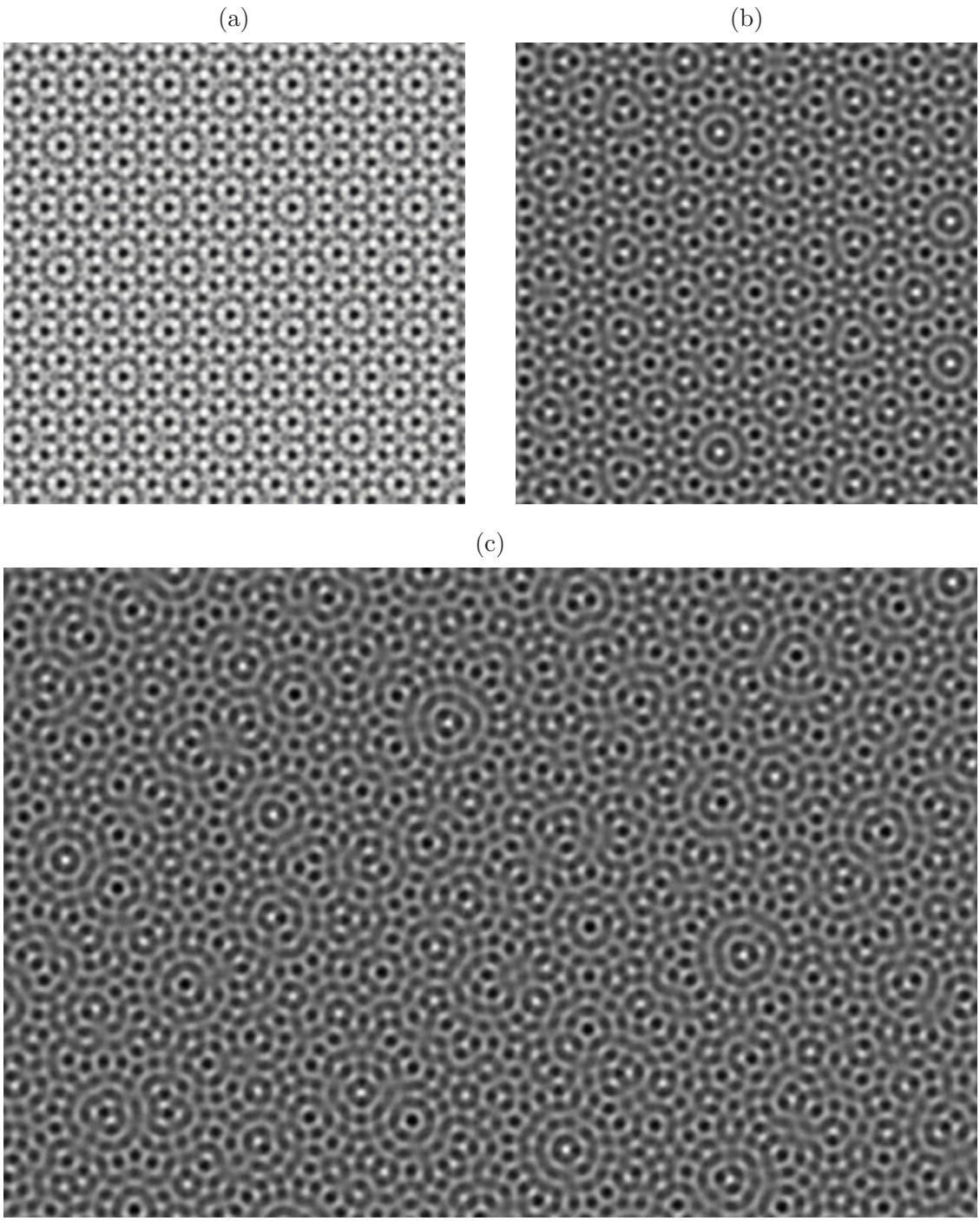


Figure 18. *For the parameter values from Figure 16, we find three different approximate quasipatterns depending on the amplitude of the forcing and the size of the domain.* (a) 1.1 *times critical,* 30×30 *domain: a* 12*-fold quasipattern.* (b) 1.3 *times critical,* 30×30 *domain: a* 14*-fold quasipattern.* (c) 1.3 *times critical,* 60×60 *domain: a* 20*-fold quasipattern (only* $\frac{2}{3}$ *of the domain is shown). See also the accompanying animation* (71906_01.gif [19.1MB]) *of the transition from* (a) *to* (b)*, also showing details of the Fourier spectrum.*

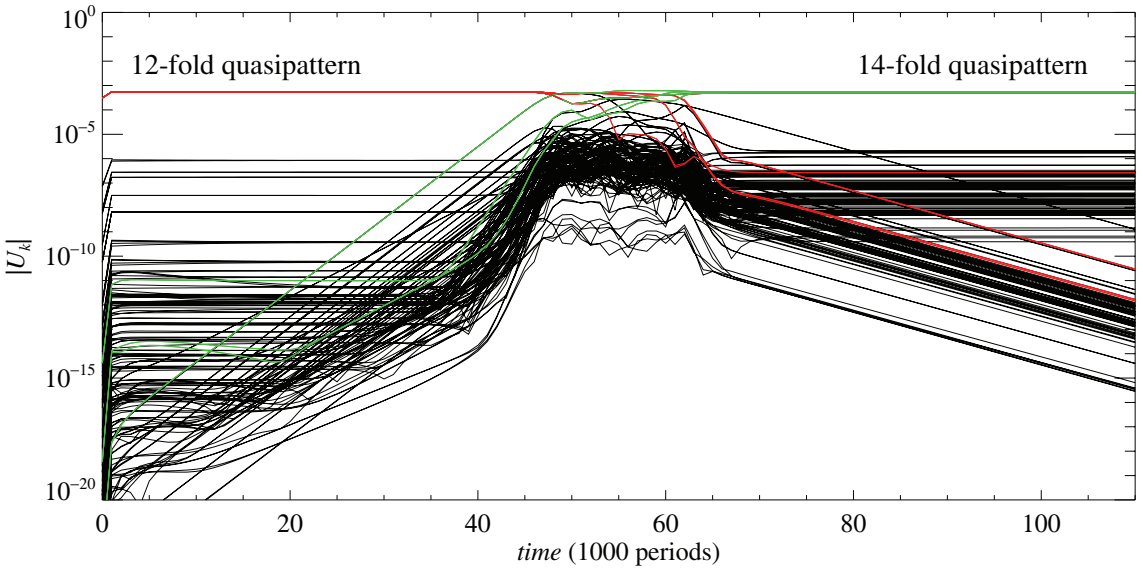


Figure 19. *Amplitudes of Fourier modes as a function of time, at* 1.3 *times critical in a* 30×30 *domain. The initial condition was the* 12*-fold quasipattern from* 1.1 *times critical (red). This is unstable and, after an extended transient of* 70,000 *periods, it is replaced by the* 14*-fold quasipattern (green). Amplitudes of other Fourier modes close to* $k = 1$ *are shown in black. See also the accompanying animation* (71906_01.gif [19.1MB]) *of this transition.*

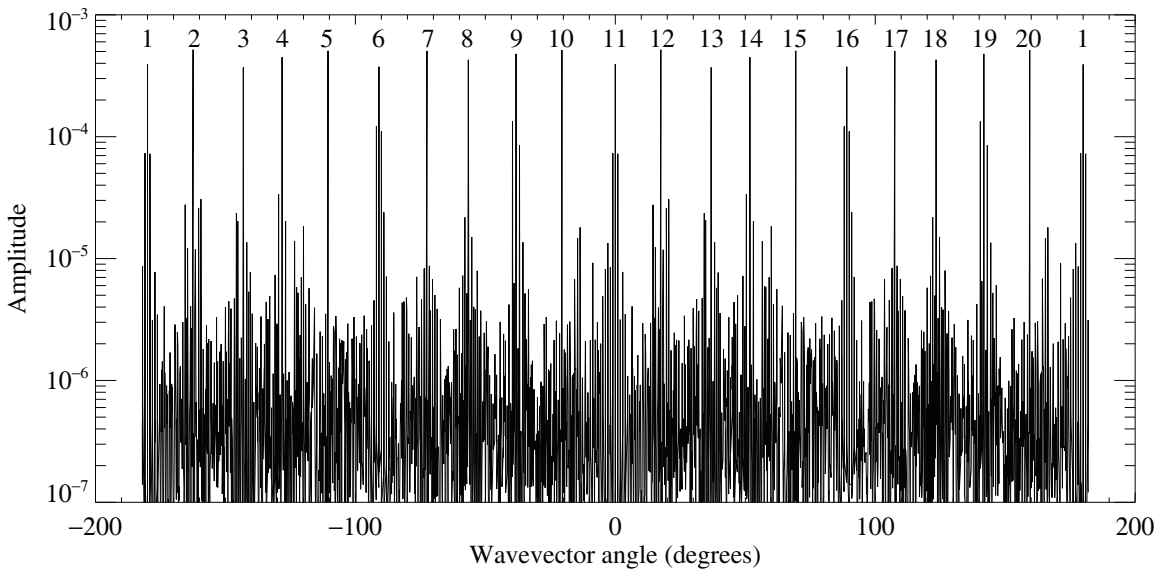


Figure 20. *Amplitudes of Fourier modes of the* 20*-fold quasipattern from Figure* 18(c)*, as a function of wavevector orientation, for* $0.95 \le |\boldsymbol{k}| \le 1.05$*, showing twenty peaks roughly evenly distributed.*

have angular separations within 3° of 360°/20, with spacings ranging from 15.0° to 20.7°. At 1.1 times critical in a 60×60 domain, we have also found a less accurate approximation to a 16-fold quasipattern: the 16 wavevectors differ in length by 0.3%, with spacings ranging from 14.5° to 27.8°. We speculate that in other domains and at other forcings, 18-fold and indeed other approximate quasipatterns could also be observed.

The cross-coupling B_θ curve suggests that modes that are more than about 30° apart do not influence each other (at least in an amplitude equation truncated at cubic order). This suggests that patterns containing many modes at essentially arbitrary angles might be expected—such patterns have been termed *turbulent crystals* by Newell and Pomeau [46]. It is hard to see why one quasipattern should be favored over another, or indeed, why quasipatterns (with wavevectors evenly distributed around the $k = 1$ circle) should be favored over more complex patterns.

We have not attempted to compare the locations of the high-order modes in these 14- and 20-fold approximate quasipatterns with those in the true quasipatterns, as these higher quasilattices are very densely populated (Figure 2(d)), and the approximate solutions are not close enough to the true quasipatterns.

As an aside, there is an interesting connection that can be made between the Fourier spectra of these high-order quasipatterns and the fractal dynamics of the complex ODE $d^2\zeta/dt^2 = -\zeta^{Q-1}$, where Q is an even integer [26]. Solutions of this equation lie on multiply branched Riemann surfaces, with branch points occurring densely at points in a Q-fold quasilattice.

The rotational degeneracy of the plane (in the absence of boundary conditions) implies that any mode with $|\boldsymbol{k}| = 1$ is linearly excited, and it is an open question as to why patterns and quasipatterns, with a finite number of modes evenly distributed around the unit circle, should be the preferred patterns close to onset in so many examples of pattern-forming systems. (Of course, pattern-forming systems are by definition those that produce regular patterns close to onset!) Newell and Pomeau [46] wrote down the evolution equation of Q modes, ignoring quadratic terms (which are absent in the case of subharmonic waves) and truncating at cubic order:

$$\frac{dA_j}{dt} = \lambda A_j - \sum_{k=1}^{Q} B_{\theta_{jk}} |A_k|^2 A_j, \tag{8.1}$$

where A_j is the complex amplitude of mode j ($j = 1, \ldots, Q$), λ is the growth rate, and θ_{jk} is the angle between the wavevectors of modes j and k. Truncated in this way, the phases of the amplitudes do not enter the dynamics: Newell and Pomeau [46] attribute this to the translation symmetry of the underlying problem, but Melbourne [41] points out that this phase-invariance is in fact a normal form symmetry and hence not exact.

This system of ODEs is variational and evolves to minimize a free energy

$$\mathcal{F} = -\lambda \sum_{j=1}^{Q} |A_j|^2 + \frac{1}{2} \sum_{j,k=1}^{Q} B_{\theta_{jk}} |A_j|^2 |A_k|^2. \tag{8.2}$$

Newell and Pomeau [46] claim that in the case that the state that minimizes the free energy $\mathcal{F}$ has many modes, with the magnitudes of all amplitudes equal but with arbitrary phases,

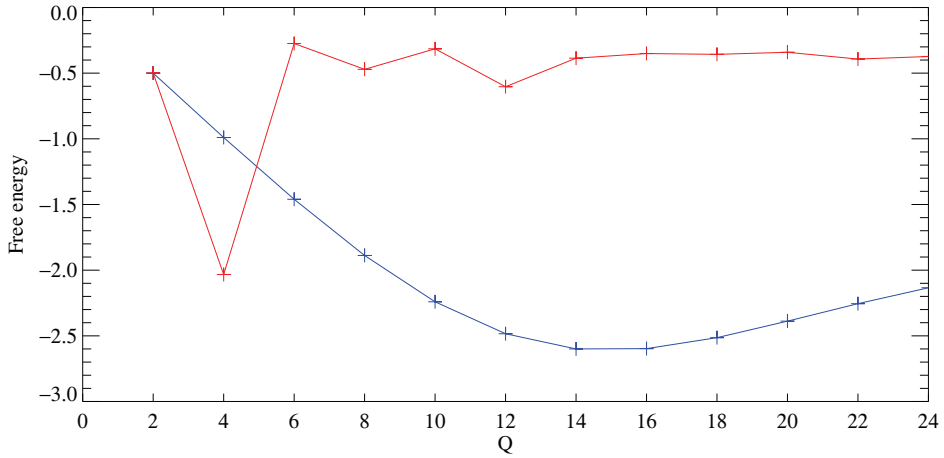


Figure 21. *Value of the free energy $\mathcal{F}$ from (8.2) for Q-fold patterns and quasipatterns, derived from the B_θ curves in Figure 16 with single-frequency forcing (blue) and Figure 8 with multifrequency forcing (red).*

then the pattern will resemble a spatially random field (a turbulent crystal). This case will be realized when $B_\theta < 1$ over a wide range of θ [46, 68] (as in Figure 16, for example). We therefore propose the quasipattern solutions in Figure 18 (12-, 14-, and 20-fold quasipatterns) as examples of turbulent crystals.

It is interesting to consider what determines the number of modes in these turbulent crystals. One argument, explored in more detail by Zhang and Viñals [69], is that the preferred pattern at onset should be the global minimum of the free energy $\mathcal{F}$. Assuming equal amplitudes, the free energy of a Q-fold pattern or quasipattern depends on B_θ evaluated at $360°/Q$, $720°/Q$, etc. Figure 21 shows $\mathcal{F}$ derived from the B_θ curves in Figures 16 and 8, with single- and multifrequency forcing, respectively.

In the single-frequency case (blue curve in Figure 21), where B_θ is close to zero for $\theta \geq 30°$ (Figure 16), there is a broad minimum around 14-fold or 16-fold quasipatterns. However, approximate 12-fold, 14-fold, 16-fold, and 20-fold examples were found (Figure 18), depending on the forcing strength and domain size. No doubt other patterns could also be found with more exploration.

In the multifrequency case (red curve in Figure 21), where B_θ has pronounced dips at 30° and 90° (Figure 8), there are local minima at $Q = 4$ (squares), $Q = 8$, and $Q = 12$. In numerical experiments, most initial conditions found squares, though 12-fold quasipatterns were also stable (Figures 9 and 10). We have not looked for 8-fold quasipatterns.

These results suggest that these free energy arguments provide a useful qualitative tool for understanding pattern selection, but reality is often more complicated than the arguments might suggest.

9. Conclusions.

We have introduced a new model PDE (3.1) for investigating pattern formation and pattern design in parametrically forced systems. The PDE is intended to play the same role for the Faraday wave experiment that the Swift–Hohenberg equation [61] plays for convection: while the model cannot be derived from the fluid mechanics, it has qualitatively

correct linear behavior and the right type of nonlinear interactions in order to provide useful illumination of the processes that are going on. The model produces superlattice patterns (section 6) and quasipatterns (sections 7 and 8) in response to single- and multifrequency forcing, for the same reasons that these complex patterns are found in the Faraday wave experiment, confirming that the mechanisms are generic. The ease of calculating weakly nonlinear coefficients and of computing large-scale numerical solutions has allowed a quantitative exploration of the agreement between the theoretical understanding of the pattern selection mechanism and the patterns that are actually found.

Of course, the model does not capture every detail of the physics of the Faraday wave experiment. In particular, the dispersion relations have different structures, and the model does not include the mean flow effects that are important in nearly inviscid Faraday experiments [32]. The latter could be addressed by coupling the model to a conserved quantity or to a mean-flow equation (as in [25, 64]) or by taking the negative Laplacian of the right-hand side of the PDE (as in [13, 15]).

The Zhang–Viñals [68] equations do not have these drawbacks: these are derived from the Navier–Stokes equations in the limit of infinite depth and zero viscosity. One might ask what is gained by looking at a simpler PDE that is even further from the physics. There are two advantages to the new model: one is that it is very simple—the dispersion relation can be controlled easily for studying any resonant interaction or response to multifrequency forcing; in addition, the weakly nonlinear theory can be computed very easily. A second advantage is that our model is very well suited to the use of efficient numerical methods such as exponential time differencing [12]: the linear terms are diagonal in Fourier space, and the nonlinear terms do not involve any derivatives. In contrast, the Zhang–Viñals equations are considerably more complicated, and the weakly nonlinear computations are more involved. Numerical solutions are also more time-consuming, as the linear term is not diagonal in spectral space, and most of the nonlinear terms involve products of derivatives, resulting in more Fourier transforms for their evaluation. As a result of the relatively low cost of the calculations, we have been able to follow branches of solutions in detail and to go to much larger domains and for much longer times than previous calculations. Of course, in the end it would be desirable to work directly with the Navier–Stokes equations, but for these, the weakly nonlinear theory is very challenging [58], and there are as yet no large-scale numerical simulations.

Like the Zhang–Viñals equations, the new model includes explicit time dependence. In contrast, other approaches, based on developing a description of the slow evolution of the amplitude of an underlying pattern, use the Ginzburg–Landau equation with additional complex conjugate terms to capture the effect of the time dependent forcing [8, 9, 10, 11, 29]. As a result, any complex patterns that are found must be interpreted in terms of slow, long-wavelength amplitude modulations of an underlying pattern, which complicates any effort to make quantitative comparison between theoretical ideas and the behavior of the real system. However, within this framework, [9, 10] show that superlattice patterns and quasipatterns are found only if explicit time dependence is introduced into the Ginzburg–Landau equation by (slowly) modulating the strength of one of the forcing components. The resulting equation is somewhat similar to our model PDE, though the interpretation is different.

The significance of three-wave coupling to weakly damped modes and its role in pattern selection has long been recognized [20, 42, 46, 69]. We have put this idea to a quantitative

test by using it to choose forcing functions that stabilize a desired pattern in large domain calculations. However, the codimension-one approach to this idea, where the weakly damped modes are slaved to the pattern-forming modes, does not provide quantitative predictions of amplitudes of patterns, and of parameter regimes where the desired patterns should be stable, except for very close to onset. The reason is that computation of the cross-coupling coefficient B_θ is valid only when all modes are strongly damped compared to the pattern-forming modes. This poses difficulties because the most interesting patterns occur where the pattern-forming modes are coupled to weakly damped modes, and this is where the theory used to calculate properties of these patterns is of limited validity. As a result, parameters had to be chosen very close to onset in order to find stable numerical examples of the desired patterns in parameter regimes where they were predicted to be stable. A codimension-two approach would extend the range of validity of the theory and will be the subject of future work.

We have investigated two mechanisms for the formation of quasipatterns. One mechanism uses three-wave interactions involving a damped mode associated with the difference of the two frequencies in the forcing to select a particular angle (30° in the example presented here). Using different primary frequencies, or altering the dispersion relation, would allow other angles, or combinations of angles, to be selected. The advantage is that a forcing function can be designed to produce a particular pattern: the mechanism is quite selective and requires some fine-tuning of the parameters.

The second mechanism uses 1 : 2 resonance in space and time to magnify the self-interaction coefficient and thereby, on rescaling, diminish the cross-coupling coefficient B_θ for angles greater than about 30°. This can lead to the formation of turbulent crystals [46]. The mechanism is robust (the patterns are found well above onset) and requires only single-frequency forcing. A dispersion relation that supports 1 : 2 resonance in space and time is needed. Within this framework, an inherent complication is that it is not clear why regular 8-, 10-, 12-, or 14-fold quasipatterns, or indeed any other combination of modes, should be favored. Indeed we have found that 12-, 14-, and 20-fold approximate quasipatterns can be stabilized by altering the level of the forcing or the domain size, without changing other parameters, and we have reported the transition between two different types of quasipattern. The free energy approach [46, 68] cannot make this distinction and would predict that 14- or 16-fold quasipatterns should be found at onset for these parameter values (Figure 21). It remains an open question as to why one turbulent crystal should be favored over another.

The existence of 14-fold (and higher) quasipatterns has been suggested before [52, 60, 62, 68], but we have presented here the first examples of spontaneously formed 14-fold and 20-fold approximate quasipatterns that are stable solutions of a PDE (Figure 18), with preliminary results in [53]. Examples where 14-fold symmetry is imposed externally have been reported in optical experiments [47]. The Fourier spectra of 12-fold and 14-fold quasipatterns are both dense (Figure 2(c,d)), but those of 14-fold quasipatterns are much denser, owing to the difference between quadratic and cubic irrational numbers [52]. This difference may have profound consequences for their mathematical treatment.

We have identified what domain sizes result in the most accurate approximations to 12-fold quasipatterns, based on square and on hexagonal domains (Table 2), and produced exceptionally clean examples of approximate quasipatterns in relatively large computational domains.

Comparing the Fourier spectra of the approximate quasipatterns as a function of domain size, we have identified at what order the locations of Fourier modes in the approximate quasipatterns deviate from those of the true quasipatterns. In the largest example (112×112), the locations of the Fourier modes deviate significantly only beyond 26th order (Figure 12(a)), at which point the amplitudes of the modes have reached the level of numerical round-off (Figure 14). This suggests that going any larger than 112×112 would not lead to any significant improvement in the approximation to a true quasipattern, at least for these parameter values.

We have compared the amplitudes of the Fourier modes of the approximate quasipatterns and the leading-order weakly nonlinear prediction (Figures 9 and 17), and found quantitative agreement very close to onset but only qualitative agreement at larger amplitude, which is what would be expected from the problem of eliminating weakly damped modes, as discussed above. We have not extended this comparison to higher order since the weakly nonlinear calculations are too difficult for this parametrically forced problem. An extension of this work would be to devise a PDE without parametric forcing that also produces stable quasipattern solutions. This would allow the use of a larger timestep (since the fast oscillation of the forcing is not present), and so even larger domain numerical computations could be carried out. The lack of parametric forcing would also allow high-order weakly nonlinear calculations (as in [52], where calculations were carried out to 33rd order), and so a quantitative comparison between computed mode amplitudes (as a function of order) and the weakly nonlinear theory would be possible. Standard weakly nonlinear theory produces amplitudes that diverge at high order because of the presence of small divisors [52], while the PDE solutions have amplitudes that appear to decay exponentially with order—although the small divisors in this problem do make themselves felt by amplifying the magnitudes of the Fourier modes at (or close to) the order at which the small divisor appears (Figure 14). Such a PDE could be based on (for example) the Swift–Hohenberg equation [23, 38, 43], but the Swift–Hohenberg equation itself does not allow the weakly damped modes that are necessary to stabilize quasipatterns.

Other numerical studies of quasipatterns as solutions of a PDE have not made a systematic study of the effect of domain size. Zhang and Viñals [68, 70] report approximate 8-fold quasipatterns in a 64×64 domain in their quasipotential model of the nearly inviscid Navier–Stokes equations, for parameter values close to the 1 : 2 resonance in space and time. The modes involved were separated by 41°, 42°, 48°, and 49° [70], so the approximation was not particularly accurate; our careful choice of domain size allowed much closer approximation. Müller [43] developed a model based on two coupled Swift–Hohenberg equations, with parameters chosen so that the two unstable modes had wavenumbers that would favor 8-fold or 12-fold quasipatterns. Numerical simulations in a 10×10 domain in the second case found approximate 12-fold quasipatterns. The modes involved are not stated, but we estimate them to be $\boldsymbol{k}^{(1)} = (10, -1)$, $\boldsymbol{k}^{(2)} = (9, 4)$, $\boldsymbol{k}^{(3)} = (6, 8)$ in units of the fundamental lattice vector. These have lengths 10.05, 9.85, and 10.00, respectively, and they are separated by 28.48° and 29.17°. With $\boldsymbol{k}^{(11)} = (4, -9)$, they have $\boldsymbol{k}^{(3)} + \boldsymbol{k}^{(11)} = \boldsymbol{k}^{(1)}$, so the 60° resonance condition is satisfied. Frisch and Sonnino [23] present a similar model and report 10-fold quasipatterns. Lifshitz and Petrich [38] found a 12-fold approximate quasipattern in a roughly 30×30 domain, in a model based on a single Swift–Hohenberg equation with a degenerate double minimum in its marginal stability curve. The modes involved appear to be the same as those in the 30×30 examples discussed in section 7.

While we have not discussed the possibility of long-wave instabilities of quasipatterns, the Fourier spectra of the 112×112 example (Figures 12(d) and 13) suggest that long-wave modes that are close to the primary wavevectors in a direction *tangent* to the critical circle are forced by high-order nonlinear interactions. This is also apparent from the locations of modes responsible for the small divisors [52]. Therefore, any treatment of the long-wave stability of quasipatterns should take into account the presence of these modes. This is a delicate question. The only study of the sideband instabilities of quasipatterns [19] focuses on instabilities associated with modes that are perpendicular to the unit circle, using coupled Ginzburg–Landau equations for each of the primary mode directions in the quasipattern. This approach could be extended to include instabilities associated with modes that are tangent to the unit circle by looking at coupled Newell–Whitehead–Segel equations, along the lines suggested by [28], although high-order nonlinear interactions may not be captured in a long-wave analysis truncated at cubic order.

Appendix. Linear and weakly nonlinear theory. In this appendix, we present the linear and weakly nonlinear theory for the PDE (3.1). We will describe the calculation in terms of a harmonic primary bifurcation; the subharmonic case is similar, with the main differences being that the period is 4π rather than 2π and that the quadratic coefficient Q is identically zero.

We start by writing $U = u + iv$, where $u(x, y, t)$ and $v(x, y, t)$ are real functions, and so

$$\begin{aligned}
\frac{\partial u}{\partial t} &= \left(\mu + \alpha\nabla^2 + \gamma\nabla^4\right) u - \left(\omega + \beta\nabla^2 + \delta\nabla^4\right) v \\
&\quad + Q_{1r}(u^2 - v^2) - Q_{1i}(2uv) + Q_{2r}(u^2 + v^2) + C_r(u^2 + v^2)u - C_i(u^2 + v^2)v, \\
\frac{\partial v}{\partial t} &= \left(\omega + \beta\nabla^2 + \delta\nabla^4\right) u + \left(\mu + \alpha\nabla^2 + \gamma\nabla^4\right) v \\
&\quad + Q_{1i}(u^2 - v^2) + Q_{1r}(2uv) + Q_{2i}(u^2 + v^2) + C_r(u^2 + v^2)v + C_i(u^2 + v^2)u \\
&\quad + f(t)u.
\end{aligned}$$

We define differential operators $\mathcal{L}$ and $\mathcal{M}$:

$$\mathcal{L} = \frac{\partial}{\partial t} - \left(\mu + \alpha\nabla^2 + \gamma\nabla^4\right) \qquad \text{and} \qquad \mathcal{M} = \left(\omega + \beta\nabla^2 + \delta\nabla^4\right),$$

so the PDEs for u and v are

$$\begin{aligned}
\mathcal{L}u &= -\mathcal{M}v + \mathrm{NL}_u, \\
\mathcal{L}v &= \quad \mathcal{M}u + \mathrm{NL}_v + f(t)u.
\end{aligned}$$

The nonlinear terms NL_u and NL_v are

$$\begin{aligned}
\mathrm{NL}_u &= Q_{1r}(u^2 - v^2) - Q_{1i}(2uv) + Q_{2r}(u^2 + v^2) + C_r(u^2 + v^2)u - C_i(u^2 + v^2)v, \\
\mathrm{NL}_v &= Q_{1i}(u^2 - v^2) + Q_{1r}(2uv) + Q_{2i}(u^2 + v^2) + C_r(u^2 + v^2)v + C_i(u^2 + v^2)u.
\end{aligned}$$

A.1. Linear theory. We seek solutions of the form $u = e^{i\boldsymbol{k}\cdot\boldsymbol{x}}p_1(t)$ and $v = e^{i\boldsymbol{k}\cdot\boldsymbol{x}}q_1(t)$, where p_1 and q_1 are periodic functions of period T, and defining

$$\hat{\gamma}_1 = 2\left(-\mu + \alpha k^2 - \gamma k^4\right), \qquad \hat{\Omega}_1 = \omega - \beta k^2 + \delta k^4, \qquad \Omega_1 = \sqrt{\left(\frac{\hat{\gamma}_1}{2}\right)^2 + \left(\hat{\Omega}_1\right)^2},$$

we get the linearized equations

$$\begin{aligned} \mathcal{L}_1 p_1 &= -\mathcal{M}_1 q_1, \\ \mathcal{L}_1 q_1 &= \quad \mathcal{M}_1 p_1 + f(t) p_1, \end{aligned}$$

or

$$\mathcal{L}_1^2 p_1 = -\mathcal{M}_1 \mathcal{L}_1 q_1 = -\mathcal{M}_1^2 p_1 - f(t)\mathcal{M}_1 p_1,$$

where $\mathcal{L}_1$ and $\mathcal{M}_1$ act on $p_1(t)$ and $q_1(t)$ as

$$\mathcal{L}_1 = \frac{d}{dt} + \frac{\hat{\gamma}_1}{2} \qquad \text{and} \qquad \mathcal{M}_1 = \hat{\Omega}_1.$$

The linearized PDE thus reduces to a damped Mathieu equation for p_1:

$$\left(\frac{d}{dt} + \frac{\hat{\gamma}_1}{2}\right)^2 p_1 + \hat{\Omega}_1^2 p_1 + f(t)\hat{\Omega}_1 p_1 = 0$$

or (defining a linear operator $\mathbf{L}$)

$$\ddot{p}_1 + \hat{\gamma}_1 \dot{p}_1 + \left(\Omega_1^2 + \hat{\Omega}_1 f(t)\right) p_1 = 0 = \mathbf{L} p_1.$$

The adjoint equation is

$$\ddot{\tilde{p}}_1 - \hat{\gamma}_1 \dot{\tilde{p}}_1 + \left(\Omega_1^2 + \hat{\Omega}_1 f(t)\right) \tilde{p}_1 = 0 = \tilde{\mathbf{L}} \tilde{p}_1,$$

defined with respect to an inner product

$$\langle g, h \rangle = \frac{1}{T} \int_0^T g(t) h(t)\, dt,$$

with $T = 2\pi$ (harmonic case) or $T = 4\pi$ (subharmonic case), so $\langle g, \mathbf{L}h \rangle = \langle \tilde{\mathbf{L}} g, h \rangle$ for T-periodic functions g and h.

For a given value of k, seeking nonzero periodic solutions of $\mathbf{L}p = 0$ yields an eigenvalue problem whose eigenvalue is the amplitude of the forcing function $f(t)$. We use the method of [4] to solve this eigenvalue problem with multifrequency forcing $f(t)$, providing the critical forcing amplitude. Minimizing this critical forcing amplitude over k yields the critical wavenumber k_c, critical forcing function $f_c(t)$, and critical eigenfunction $p_1(t)$. The corresponding $q_1(t)$ is determined by solving $\mathcal{L}_1 q_1 = \hat{\Omega} p_1 + f_c(t) p_1$.

A.2. Rhombs. We consider f close to f_c, writing $f(t) = f_c(t)(1+\epsilon^2 F_2)$, and seek small-amplitude rhombic solutions associated with two wavevectors $\boldsymbol{k}_1$ and $\boldsymbol{k}_2$ at the critical wavenumber: $|\boldsymbol{k}_1| = |\boldsymbol{k}_2| = k_c$, separated by an angle θ. (For ease of notation, in this appendix, we use subscripts to index wavevectors and suppress explicit reference to the forcing frequency.) We formally expand the solution as

$$\begin{aligned} u &= \epsilon u_1 + \epsilon^2 u_2 + \epsilon^3 u_3 + \cdots \\ &= \epsilon \left(z_1(T_2) e^{i\boldsymbol{k}_1\cdot\boldsymbol{x}} + z_2(T_2) e^{i\boldsymbol{k}_2\cdot\boldsymbol{x}} + c.c. \right) p_1(t) \\ &\quad + \epsilon^2 \left(\left(z_1^2 e^{2i\boldsymbol{k}_1\cdot\boldsymbol{x}} + z_2^2 e^{2i\boldsymbol{k}_2\cdot\boldsymbol{x}} + c.c. \right) p_2(t) + \left(|z_1|^2 + |z_2|^2 \right) p_3(t) \right. \\ &\qquad \left. + \left(z_1 z_2 e^{i(\boldsymbol{k}_1+\boldsymbol{k}_2)\cdot\boldsymbol{x}} + c.c. \right) p_4(t) + \left(z_1 \bar{z}_2 e^{i(\boldsymbol{k}_1-\boldsymbol{k}_2)\cdot\boldsymbol{x}} + c.c. \right) p_5(t) \right) + \mathcal{O}(\epsilon^3), \end{aligned}$$

with a similar expression for v in terms of v_1, v_2, and v_3, and $q_1, \ldots, q_5$, where T_2 is a slow time, varying on a scale of ϵ^{-2}, $c.c.$ is the complex conjugate, and the functions $p_2(t), \ldots, q_5(t)$ are to be determined. The form of this solution is chosen by knowing in advance the structure of the nonlinear terms and the modes to be generated by them.

Substituting these expressions for u and v into the PDE and ordering in powers of ϵ, we recover, at leading order in ϵ, the linear theory. At second order in ϵ, we split the PDE into terms that go as $e^{2i\boldsymbol{k}_1\cdot\boldsymbol{x}}$, $e^{2i\boldsymbol{k}_2\cdot\boldsymbol{x}}$, terms without spatial dependence, and terms that go as $e^{i(\boldsymbol{k}_1+\boldsymbol{k}_2)\cdot\boldsymbol{x}}$ and $e^{i(\boldsymbol{k}_1-\boldsymbol{k}_2)\cdot\boldsymbol{x}}$.

The terms like $e^{2i\boldsymbol{k}_1\cdot\boldsymbol{x}}$ lead to equations for p_2 and q_2:

$$\begin{aligned} \mathcal{L}_2 p_2 &= -\mathcal{M}_2 q_2 + \mathrm{NL}_{p2}^{(2)}, \\ \mathcal{L}_2 q_2 &= \quad \mathcal{M}_2 p_2 + \mathrm{NL}_{q2}^{(2)} + f_c(t) p_2, \end{aligned}$$

where the linear operators are

$$\begin{aligned} \mathcal{L}_2 &= \frac{d}{dt} - \left(\mu - 4\alpha k_c^2 + 16\gamma k_c^4 \right) = \frac{d}{dt} + \frac{\hat{\gamma}_2}{2}, \\ \mathcal{M}_2 &= \left(\omega - 4\beta k_c^2 + 16\delta k_c^4 \right) = \hat{\Omega}_2, \end{aligned}$$

and the nonlinear terms $\mathrm{NL}_{p2}^{(2)}$ and $\mathrm{NL}_{q2}^{(2)}$ are

$$\begin{aligned} \mathrm{NL}_{p2}^{(2)} &= Q_{1r}(p_1^2 - q_1^2) - Q_{1i}(2p_1 q_1) + Q_{2r}(p_1^2 + q_1^2), \\ \mathrm{NL}_{q2}^{(2)} &= Q_{1i}(p_1^2 - q_1^2) + Q_{1r}(2p_1 q_1) + Q_{2i}(p_1^2 + q_1^2). \end{aligned}$$

The function q_2 is eliminated by operating with $\mathcal{L}_2$, resulting in a second-order nonconstant coefficient inhomogeneous linear ODE for p_2:

$$\left(\mathcal{L}_2^2 + \mathcal{M}_2^2 + \mathcal{M}_2 f(t) \right) p_2 = \mathcal{L}_2 \, \mathrm{NL}_{p2}^{(2)} - \mathcal{M}_2 \, \mathrm{NL}_{q2}^{(2)}$$

or

$$\ddot{p}_2 + \hat{\gamma}_2 \dot{p}_2 + \left(\Omega_2^2 + \hat{\Omega}_2 f_c(t) \right) p_2 = \left(\frac{d}{dt} + \frac{\hat{\gamma}_2}{2} \right) \mathrm{NL}_{p2}^{(2)} - \hat{\Omega}_2 \, \mathrm{NL}_{q2}^{(2)}.$$

This can be solved numerically for p_2 using Fourier transform methods, and q_2 can then be found. Terms that go as $e^{2i\boldsymbol{k}_2\cdot\boldsymbol{x}}$ result in the same equation.

Terms without spatial dependence, and terms with spatial dependence $e^{i(\boldsymbol{k}_1+\boldsymbol{k}_2)\cdot\boldsymbol{x}}$ and $e^{i(\boldsymbol{k}_1-\boldsymbol{k}_2)\cdot\boldsymbol{x}}$, result in similar equations for p_3, p_4, and p_5, but with linear operators,

$$\begin{aligned}
\mathcal{L}_3 &= \frac{d}{dt} - \mu = \frac{d}{dt} + \frac{\hat{\gamma}_3}{2}, \qquad \text{and} \qquad \mathcal{M}_3 = \omega = \hat{\Omega}_3,\\
\mathcal{L}_4 &= \frac{d}{dt} - \left(\mu - 4\cos^2\left(\frac{\theta}{2}\right)\alpha k_c^2 + 16\cos^4\left(\frac{\theta}{2}\right)\gamma k_c^4\right) = \frac{d}{dt} + \frac{\hat{\gamma}_4}{2},\\
\mathcal{M}_4 &= \left(\omega - 4\cos^2\left(\frac{\theta}{2}\right)\beta k_c^2 + 16\cos^4\left(\frac{\theta}{2}\right)\delta k_c^4\right) = \hat{\Omega}_4,\\
\mathcal{L}_5 &= \frac{d}{dt} - \left(\mu - 4\sin^2\left(\frac{\theta}{2}\right)\alpha k_c^2 + 16\sin^4\left(\frac{\theta}{2}\right)\gamma k_c^4\right) = \frac{d}{dt} + \frac{\hat{\gamma}_5}{2},\\
\mathcal{M}_5 &= \left(\omega - 4\sin^2\left(\frac{\theta}{2}\right)\beta k_c^2 + 16\sin^4\left(\frac{\theta}{2}\right)\delta k_c^4\right) = \hat{\Omega}_5,
\end{aligned}$$

and nonlinear terms:

$$\begin{aligned}
\mathrm{NL}_{p3}^{(2)} &= \mathrm{NL}_{p4}^{(2)} = \mathrm{NL}_{p5}^{(2)} = 2\,\mathrm{NL}_{p2}^{(2)},\\
\mathrm{NL}_{q3}^{(2)} &= \mathrm{NL}_{q4}^{(2)} = \mathrm{NL}_{q5}^{(2)} = 2\,\mathrm{NL}_{q2}^{(2)}.
\end{aligned}$$

Note that $\hat{\gamma}_4$, $\hat{\gamma}_5$, $\hat{\Omega}_4$, and $\hat{\Omega}_5$ depend on θ, the angle between the chosen wavevectors, whereas $\hat{\gamma}_1$, $\hat{\gamma}_2$, $\hat{\gamma}_3$, $\hat{\Omega}_1$, $\hat{\Omega}_2$, and $\hat{\Omega}_3$ do not.

At third order in ϵ, the problem has the following structure:

$$\begin{aligned}
\mathcal{L}u_3 + \frac{\partial u_1}{\partial T_2} &= -\mathcal{M}v_3 + \mathrm{NL}_u^{(3)},\\
\mathcal{L}v_3 + \frac{\partial v_1}{\partial T_2} &= \quad \mathcal{M}u_3 + \mathrm{NL}_v^{(3)} + f_c(t)u_3 + F_2 f_c(t)u_1.
\end{aligned}$$

The nonlinear terms $\mathrm{NL}_u^{(3)}$ and $\mathrm{NL}_v^{(3)}$ are

$$\begin{aligned}
\mathrm{NL}_u^{(3)} &= 2Q_{1r}(u_1u_2 - v_1v_2) - Q_{1i}(2u_1v_2 + 2u_2v_1) + 2Q_{2r}(u_1u_2 + v_1v_2)\\
&\quad + C_r(u_1^2 + v_1^2)u_1 - C_i(u_1^2 + v_1^2)v_1,\\
\mathrm{NL}_v^{(3)} &= 2Q_{1i}(u_1u_2 - v_1v_2) + Q_{1r}(2u_1v_2 + 2u_2v_1) + 2Q_{2i}(u_1u_2 + v_1v_2)\\
&\quad + C_r(u_1^2 + v_1^2)v_1 + C_i(u_1^2 + v_1^2)u_1.
\end{aligned}$$

Eliminating v_3, we obtain

$$\left(\mathcal{L}^2 + \mathcal{M}^2 + \mathcal{M}f_c\right)u_3 = -\mathcal{L}\frac{\partial u_1}{\partial T_2} + \mathcal{M}\frac{\partial v_1}{\partial T_2} - \mathcal{M}F_2 f_c u_1 + \mathcal{L}\,\mathrm{NL}_u^{(3)} - \mathcal{M}\,\mathrm{NL}_v^{(3)}.$$

The operator on the left is the singular operator $\mathbf{L}$ from the linearized problem, so the equation can only be solved for u_3 if a solvability condition is applied to the terms that are proportional

to $e^{i\boldsymbol{k}_1\cdot\boldsymbol{x}}$ and $e^{i\boldsymbol{k}_2\cdot\boldsymbol{x}}$ and complex conjugates. If we take the inner product between $\tilde{p}_1$ and the $e^{i\boldsymbol{k}_1\cdot\boldsymbol{x}}$ component of the above, we find

$$\tau\frac{\partial z_1}{\partial T_2} = \sigma F_2 z_1 + \left(\hat{A}|z_1|^2 + (\hat{B}_{\rm indep} + \hat{B}_{\rm res}(\theta))|z_2|^2\right) z_1,$$

with a similar equation for z_2, where

$$\tau = \left\langle \tilde{p}_1, 2\left(\dot{p}_1 + \frac{\hat{\gamma}_1}{2}p_1\right)\right\rangle \qquad \text{and} \qquad \sigma = \left\langle \tilde{p}_1, -\hat{\Omega}_1 f_c p_1\right\rangle$$

(using $\mathcal{L}_1 p_1 = -\mathcal{M}_1 q_1$) and

$$\begin{aligned}
\hat{A} = \Big\langle \tilde{p}_1, \left(\frac{d}{dt} + \frac{\hat{\gamma}_1}{2}\right)\Big(&2Q_{1r}(p_1p_2 + p_1p_3 - q_1q_2 - q_1q_3)\\
&- 2Q_{1i}(p_1q_2 + p_1q_3 + q_1p_2 + q_1p_3)\\
&+ 2Q_{2r}(p_1p_2 + p_1p_3 + q_1q_2 + q_1q_3)\\
&+ 3C_r(p_1^2 + q_1^2)p_1 - 3C_i(p_1^2 + q_1^2)q_1\Big)\\
&- \hat{\Omega}_1\Big(2Q_{1r}(p_1q_2 + p_1q_3 + q_1p_2 + q_1p_3)\\
&\quad + 2Q_{1i}(p_1p_2 + p_1p_3 - q_1q_2 - q_1q_3)\\
&\quad + 2Q_{2i}(p_1p_2 + p_1p_3 + q_1q_2 + q_1q_3)\\
&\quad + 3C_r(p_1^2 + q_1^2)q_1 + 3C_i(p_1^2 + q_1^2)p_1\Big)\Big\rangle,\\
\hat{B}_{\rm indep} = \Big\langle \tilde{p}_1, \left(\frac{d}{dt} + \frac{\hat{\gamma}_1}{2}\right)\Big(&2Q_{1r}(p_1p_3 - q_1q_3) - 2Q_{1i}(p_1q_3 + q_1p_3)\\
&+ 2Q_{2r}(p_1p_3 + q_1q_3) + 6C_r(p_1^2 + q_1^2)p_1 - 6C_i(p_1^2 + q_1^2)q_1\Big)\\
&- \hat{\Omega}_1\Big(2Q_{1r}(p_1q_3 + q_1p_3) + 2Q_{1i}(p_1p_3 - q_1q_3) + 2Q_{2i}(p_1p_3 + q_1q_3)\\
&\quad + 6C_r(p_1^2 + q_1^2)q_1 + 6C_i(p_1^2 + q_1^2)p_1\Big)\Big\rangle,\\
\hat{B}_{\rm res}(\theta) = \Big\langle \tilde{p}_1, \left(\frac{d}{dt} + \frac{\hat{\gamma}_1}{2}\right)\Big(&2Q_{1r}(p_1p_4 + p_1p_5 - q_1q_4 - q_1q_5)\\
&- 2Q_{1i}(p_1q_4 + p_1q_5 + q_1p_4 + q_1p_5)\\
&+ 2Q_{2r}(p_1p_4 + p_1p_5 + q_1q_4 + q_1q_5)\Big)\\
&- \hat{\Omega}_1\Big(2Q_{1r}(p_1q_4 + p_1q_5 + q_1p_4 + q_1p_5)\\
&\quad + 2Q_{1i}(p_1p_4 + p_1p_5 - q_1q_4 - q_1q_5)\\
&\quad + 2Q_{2i}(p_1p_4 + p_1p_5 + q_1q_4 + q_1q_5)\Big)\Big\rangle.
\end{aligned}$$

For convenience, we have separated the parts of the cross-coupling coefficient that do not depend on the angle between the modes ($\hat{B}_{\rm indep}$) from those that do ($\hat{B}_{\rm res}(\theta)$). We discuss below how these coefficients are then scaled.

A.3. Hexagons. As with rhombs, we look for f close to f_c, writing $f(t) = f_c(t)(1+\epsilon^2 F_2)$, but now we choose three wavevectors, $\boldsymbol{k}_1$, $\boldsymbol{k}_2$, and $\boldsymbol{k}_3$ oriented at 120° to each other, with $\boldsymbol{k}_1 + \boldsymbol{k}_2 + \boldsymbol{k}_3 = \boldsymbol{0}$. We look for small-amplitude solutions with equal amplitudes of the three waves, and write

$$\begin{aligned} u &= \epsilon u_1 + \epsilon^2 u_2 + \epsilon^3 u_3 + \cdots \\ &= \epsilon z(T_1, T_2)\left(e^{i\boldsymbol{k}_1\cdot\boldsymbol{x}} + e^{i\boldsymbol{k}_2\cdot\boldsymbol{x}} + e^{i\boldsymbol{k}_3\cdot\boldsymbol{x}} + c.c.\right) p_1(t) \\ &\quad + \epsilon^2\left(z^2\left(e^{2i\boldsymbol{k}_1\cdot\boldsymbol{x}} + e^{2i\boldsymbol{k}_2\cdot\boldsymbol{x}} + e^{2i\boldsymbol{k}_3\cdot\boldsymbol{x}} + c.c.\right) p_2(t) + 3|z|^2 p_3(t)\right. \\ &\qquad + |z|^2\left(e^{i(\boldsymbol{k}_1-\boldsymbol{k}_2)\cdot\boldsymbol{x}} + e^{i(\boldsymbol{k}_2-\boldsymbol{k}_3)\cdot\boldsymbol{x}} + e^{i(\boldsymbol{k}_3-\boldsymbol{k}_1)\cdot\boldsymbol{x}} + c.c.\right)\tilde{p}_5(t) \\ &\qquad \left. + \bar{z}^2\left(e^{i\boldsymbol{k}_1\cdot\boldsymbol{x}} + e^{i\boldsymbol{k}_2\cdot\boldsymbol{x}} + e^{i\boldsymbol{k}_3\cdot\boldsymbol{x}} + c.c.\right) p_6(t)\right) + \mathcal{O}(\epsilon^3), \end{aligned}$$

with a similar expression for v in terms of $q_1, \ldots, q_6$, where T_1 and T_2 are slow times, varying on scales ϵ^{-1} and ϵ^{-2}, and the functions $p_2(t), \ldots, q_6(t)$ are to be determined. The form of the expression is chosen by knowing in advance the structure of the nonlinear terms.

Substituting these expressions for u and v into the PDE and ordering in powers of ϵ, at leading order in ϵ we recover the linear theory. At second order in ϵ, we split the PDE into terms that go as $e^{2i\boldsymbol{k}_1\cdot\boldsymbol{x}} + e^{2i\boldsymbol{k}_2\cdot\boldsymbol{x}} + e^{2i\boldsymbol{k}_3\cdot\boldsymbol{x}} + c.c.$, terms without spatial dependence, terms that go as $e^{i(\boldsymbol{k}_1-\boldsymbol{k}_2)\cdot\boldsymbol{x}} + e^{i(\boldsymbol{k}_2-\boldsymbol{k}_3)\cdot\boldsymbol{x}} + e^{i(\boldsymbol{k}_3-\boldsymbol{k}_1)\cdot\boldsymbol{x}} + c.c.$, and finally terms that go as $e^{i\boldsymbol{k}_1\cdot\boldsymbol{x}} + e^{i\boldsymbol{k}_2\cdot\boldsymbol{x}} + e^{i\boldsymbol{k}_3\cdot\boldsymbol{x}} + c.c.$, which have to be considered specially.

Terms like $e^{2i\boldsymbol{k}_1\cdot\boldsymbol{x}} + e^{2i\boldsymbol{k}_2\cdot\boldsymbol{x}} + e^{2i\boldsymbol{k}_3\cdot\boldsymbol{x}} + c.c.$ and terms without spatial dependence give the same equations for p_2, q_2, p_3, and q_3 as in the case of rhombs. In particular, the inhomogeneous nonlinear terms are the same. Terms that go as $e^{i(\boldsymbol{k}_1-\boldsymbol{k}_2)\cdot\boldsymbol{x}} + e^{i(\boldsymbol{k}_2-\boldsymbol{k}_3)\cdot\boldsymbol{x}} + e^{i(\boldsymbol{k}_3-\boldsymbol{k}_1)\cdot\boldsymbol{x}} + c.c.$ result in equations for $\tilde{p}_5$ and $\tilde{q}_5$ that are the equations for p_5 and q_5 evaluated for $\theta = 120°$.

Terms that go as $e^{i\boldsymbol{k}_1\cdot\boldsymbol{x}} + e^{i\boldsymbol{k}_2\cdot\boldsymbol{x}} + e^{i\boldsymbol{k}_3\cdot\boldsymbol{x}} + c.c.$ require the use of two timescales and a solvability condition. The linear operators are the same as for the initial linear problem:

$$\begin{aligned} \mathcal{L}_1 p_6 &= -\mathcal{M}_1 q_6 + \mathrm{NL}_{p6}^{(2)} &&- \frac{\partial z/\partial T_1}{\bar{z}^2} p_1, \\ \mathcal{L}_1 q_6 &= \quad \mathcal{M}_1 p_6 + \mathrm{NL}_{q6}^{(2)} + f_c(t) p_6 &&- \frac{\partial z/\partial T_1}{\bar{z}^2} q_1. \end{aligned}$$

The nonlinear terms $\mathrm{NL}_{p6}^{(2)}$ and $\mathrm{NL}_{q6}^{(2)}$ are

$$\mathrm{NL}_{p6}^{(2)} = 2\,\mathrm{NL}_{p2}^{(2)} \qquad \text{and} \qquad \mathrm{NL}_{q6}^{(2)} = 2\,\mathrm{NL}_{q2}^{(2)}.$$

This can be reduced to a second-order nonconstant coefficient inhomogeneous linear ODE for p_6:

$$\left(\mathcal{L}_1^2 + \mathcal{M}_1^2 + \mathcal{M}_1 f(t)\right) p_6 = \mathcal{L}_1\,\mathrm{NL}_{p6}^{(2)} - \mathcal{M}_1\,\mathrm{NL}_{q6}^{(2)} + \frac{\partial z/\partial T_1}{\bar{z}^2}\left(\mathcal{M}_1 q_1 - \mathcal{L}_1 p_1\right).$$

Since the operator on the left-hand side is the singular linear operator $\mathbf{L}$, we must apply a solvability condition:

$$\langle \tilde{p}_1, \mathbf{L} p_6\rangle = 0 = \langle \tilde{p}_1, \mathcal{L}_1\,\mathrm{NL}_{p6}^{(2)} - \mathcal{M}_1\,\mathrm{NL}_{q6}^{(2)}\rangle - 2\frac{\partial z/\partial T_1}{\bar{z}^2}\langle \tilde{p}_1, \mathcal{L}_1 p_1\rangle$$

since $\mathcal{M}_1 q_1 = -\mathcal{L}_1 p_1$. We define

$$\tau = \left\langle \tilde{p}_1, 2\left(\frac{dp_1}{dt} + \frac{\hat{\gamma}_1}{2} p_1\right)\right\rangle$$

as before and

$$\hat{\epsilon} = \left\langle \tilde{p}_1, \mathcal{L}_1 \,\mathrm{NL}_{p6}^{(2)} - \mathcal{M}_1 \,\mathrm{NL}_{q6}^{(2)}\right\rangle,$$

and obtain an equation for the slow (T_1) evolution of the amplitude z:

$$\tau \frac{\partial z}{\partial T_1} = \hat{\epsilon} \bar{z}^2.$$

Once the solvability condition has been imposed, the ODE can be solved for p_6 and q_6. The computed solution p_6 contains an arbitrary amount of p_1; the solution is made unique by specifying that $\langle \tilde{p}_1, p_6 \rangle = 0$.

At third order in ϵ, the problem has the following structure:

$$\begin{aligned}
\mathcal{L} u_3 + \frac{\partial u_2}{\partial T_1} + \frac{\partial u_1}{\partial T_2} &= -\mathcal{M} v_3 + \mathrm{NL}_u^{(3h)}, \\
\mathcal{L} v_3 + \frac{\partial v_2}{\partial T_1} + \frac{\partial v_1}{\partial T_2} &= \quad \mathcal{M} u_3 + \mathrm{NL}_v^{(3h)} + f_c(t) u_3 + F_2 f_c(t) u_1.
\end{aligned}$$

We only need keep track of terms proportional to $e^{i\boldsymbol{k}_1 \cdot \boldsymbol{x}}$ in our derivation of the bifurcation problem, so the $\partial u_2/\partial T_1$ and $\partial v_2/\partial T_1$ terms yield $p_6 \partial \bar{z}^2/\partial T_1$ and $q_6 \partial \bar{z}^2/\partial T_1$. The $e^{i\boldsymbol{k}_1 \cdot \boldsymbol{x}}$ components of the nonlinear terms $\mathrm{NL}_u^{(3h)}$ and $\mathrm{NL}_v^{(3h)}$ are specified below.

Eliminating v_3, we obtain

$$\begin{aligned}
\left(\mathcal{L}^2 + \mathcal{M}^2 + \mathcal{M} f_c\right) u_3 = &-\mathcal{L}\left(\frac{\partial u_2}{\partial T_1} + \frac{\partial u_1}{\partial T_2} - \mathrm{NL}_u^{(3h)}\right) \\
&+ \mathcal{M}\left(\frac{\partial v_2}{\partial T_1} + \frac{\partial v_1}{\partial T_2} - F_2 f_c u_1 - \mathrm{NL}_v^{(3h)}\right).
\end{aligned}$$

The equation can only be solved for u_3 if a solvability condition is applied to the terms that are proportional to $e^{i\boldsymbol{k}_1 \cdot \boldsymbol{x}}$, $e^{i\boldsymbol{k}_2 \cdot \boldsymbol{x}}$, and $e^{i\boldsymbol{k}_3 \cdot \boldsymbol{x}}$, and complex conjugates. If we take the inner product between $\tilde{p}_1$ and the $e^{i\boldsymbol{k}_1 \cdot \boldsymbol{x}}$ component of the above, we find

$$\tau \frac{\partial z}{\partial T_2} = \sigma F_2 z + \left(\hat{A} + 2\hat{B}_{60}\right) |z|^2 z,$$

where τ and σ are unchanged from the rhombic calculations, and

$$\begin{aligned}
\hat{A} + 2\hat{B}_{60} = \Big\langle \tilde{p}_1, &\left(\frac{d}{dt} + \frac{\hat{\gamma}_1}{2}\right) \mathrm{NL}_u^{(3h)} - \hat{\Omega}_1 \,\mathrm{NL}_v^{(3h)} \\
&+ \left(-\mathcal{L}_1 p_6 + \mathcal{M}_1 q_6\right) 2\frac{\hat{\epsilon}}{\tau} \Big\rangle,
\end{aligned}$$

where we have used

$$\frac{\partial \bar{z}^2}{\partial T_1} = 2\bar{z}\frac{\partial \bar{z}}{\partial T_1} = 2\frac{\hat{\epsilon}}{\tau}|z|^2 z$$

and

$$\begin{aligned}
\mathrm{NL}_u^{(3h)} &= 2Q_{1r}(p_1p_2 + 3p_1p_3 + 2p_1\tilde{p}_5 + 2p_1p_6 - q_1q_2 - 3q_1q_3 - 2q_1\tilde{q}_5 - 2q_1q_6) \\
&\quad - 2Q_{1i}(p_1q_2 + 3p_1q_3 + 2p_1\tilde{q}_5 + 2p_1q_6 + q_1p_2 + 3q_1p_3 + 2q_1\tilde{p}_5 + 2q_1p_6) \\
&\quad + 2Q_{1r}(p_1p_2 + 3p_1p_3 + 2p_1\tilde{p}_5 + 2p_1p_6 + q_1q_2 + 3q_1q_3 + 2q_1\tilde{q}_5 + 2q_1q_6) \\
&\quad + 15C_r(p_1^2 + q_1^2)p_1 - 15C_i(p_1^2 + q_1^2)q_1, \\
\mathrm{NL}_v^{(3h)} &= 2Q_{1r}(p_1q_2 + 3p_1q_3 + 2p_1\tilde{q}_5 + 2p_1q_6 + q_1p_2 + 3q_1p_3 + 2q_1\tilde{p}_5 + 2q_1p_6) \\
&\quad + 2Q_{1i}(p_1p_2 + 3p_1p_3 + 2p_1\tilde{p}_5 + 2p_1p_6 - q_1q_2 - 3q_1q_3 - 2q_1\tilde{q}_5 - 2q_1q_6) \\
&\quad + 2Q_{2i}(p_1p_2 + 3p_1p_3 + 2p_1\tilde{p}_5 + 2p_1p_6 + q_1q_2 + 3q_1q_3 + 2q_1\tilde{q}_5 + 2q_1q_6) \\
&\quad + 15C_r(p_1^2 + q_1^2)q_1 + C_i(p_1^2 + q_1^2)p_1.
\end{aligned}$$

From the value of $\hat{A}$ calculated for rhombs, we can recover $\hat{B}_{60}$, which is effectively the cross-coupling coefficient for modes with wavevectors at 60°.

A.4. Reconstitution. At this stage, the pattern formation problem on a hexagonal lattice would take the form

$$\begin{aligned}
\tau\frac{\partial z_1}{\partial T_1} &= \hat{\epsilon}\bar{z}_2\bar{z}_3, \\
\tau\frac{\partial z_1}{\partial T_2} &= \sigma F_2 z_1 + \left(\hat{A}|z_1|^2 + \hat{B}_{60}|z_2|^2 + \hat{B}_{60}|z_3|^2\right) z_1,
\end{aligned}$$

where z_1, z_2, and z_3 are amplitudes of $e^{i\boldsymbol{k}_1\cdot\boldsymbol{x}}$, $e^{i\boldsymbol{k}_2\cdot\boldsymbol{x}}$, and $e^{i\boldsymbol{k}_3\cdot\boldsymbol{x}}$. Similar equations are found for $\partial z_2/\partial T_1$, etc. Recall that the small factor ϵ has been used so that the amplitude of the original amplitude U is explicitly small: $U = \epsilon z_1(T_1, T_2)e^{i\boldsymbol{k}_1\cdot\boldsymbol{x}}p_1(t) + \cdots$.

There is more than one way to combine these equations into a single ODE. Properly, we should consider only the case where the coefficient of the quadratic term $\hat{\epsilon}$ is itself small (order ϵ). This occurs either for small values of Q_1 and Q_2, or near a codimension-one line in (Q_1, Q_2) space:

$$\begin{aligned}
0 = \hat{\epsilon} &= \left\langle \tilde{p}_1, \mathcal{L}_1\,\mathrm{NL}_{p6}^{(2)} - \mathcal{M}_1\,\mathrm{NL}_{q6}^{(2)} \right\rangle \\
&= 2Q_{1r}\left\langle \tilde{p}_1, \mathcal{L}_1(p_1^2 - q_1^2) - \mathcal{M}_1(2p_1q_1)\right\rangle + 2Q_{1i}\left\langle \tilde{p}_1, \mathcal{L}_1(-2p_1q_1) - \mathcal{M}_1(p_1^2 - q_1^2)\right\rangle \\
&\qquad + 2Q_{2r}\left\langle \tilde{p}_1, \mathcal{L}_1(p_1^2 + q_1^2)\right\rangle + 2Q_{2i}\left\langle \tilde{p}_1, -\mathcal{M}_1(p_1^2 + q_1^2)\right\rangle.
\end{aligned}$$

Alternatively, we note that for many of these multifrequency forced problems, $\hat{\epsilon}$ is small anyway [48, 49].

Having decided that $\hat{\epsilon}$ is small, we define a fast time scale t^*, related to the original time scale t by an order-one factor σ/τ:

$$\frac{d}{dt^*} = \epsilon\frac{\tau}{\sigma}\frac{\partial}{\partial T_1} + \epsilon^2\frac{\tau}{\sigma}\frac{\partial}{\partial T_2},$$

where ϵ is the original small parameter. We now scale the z's by $1/\epsilon$, so that the original amplitude U is implicitly small, $U = z_1(T_1, T_2)e^{i\boldsymbol{k}_1\cdot\boldsymbol{x}}p_1(t) + \cdots$, and obtain

$$\frac{dz_1}{dt^*} = \lambda z_1 + \frac{\hat{\epsilon}}{\sigma}\bar{z}_2\bar{z}_3 + \frac{\hat{A}}{\sigma}\left(|z_1|^2 + B_{60}|z_2|^2 + B_{60}|z_3|^2\right)z_1,$$

where $\lambda = \epsilon^2 F_2$, so that the forcing amplitude is $(1+\lambda)$ times the critical amplitude and $B_{60} = \hat{B}_{60}/\hat{A}$.

The advantage of reconstituting in this way is that the quadratic and cubic terms appear at the same order. The disadvantage is that the regime of validity ($\lambda \ll 1$, $\hat{\epsilon} \ll 1$, $z \ll 1$) is not made explicit. In particular, this validity condition will only be satisfied for hexagons when the coefficient of the quadratic term is small—which is precisely the limit required for the quadratic and cubic terms to be of the same order.

Finally, we scale the amplitudes once more by a factor of $\sqrt{|\sigma/\hat{A}|}$, rename the time variable back to t, and obtain

$$\frac{dz_1}{dt} = \lambda z_1 + Q\bar{z}_2\bar{z}_3 + s\left(|z_1|^2 + B_{60}|z_2|^2 + B_{60}|z_3|^2\right)z_1,$$

with similar equations for z_2 and z_3, where

$$Q = \frac{\hat{\epsilon}}{\sigma}\sqrt{\left|\frac{\sigma}{\hat{A}}\right|} \qquad \text{and} \qquad s = \operatorname{sgn}\left(\frac{\hat{A}}{\sigma}\right)$$

(usually, $s = -1$).

Repeating the same reconstitution for the rhombic lattice results in

$$\frac{dz_1}{dt} = \lambda z_1 + s\left(|z_1|^2 + B_\theta|z_2|^2\right)z_1,$$
$$\frac{dz_2}{dt} = \lambda z_2 + s\left(|z_2|^2 + B_\theta|z_1|^2\right)z_2,$$

where $B_\theta = (\hat{B}_{\text{indep}} + \hat{B}_{\text{res}}(\theta))/\hat{A}$.

The equations for other lattices can be found from combinations of the above, as can candidate equations for quasipatterns.

Acknowledgments. We are grateful to many people who have helped shape these ideas: Jessica Conway, Stéphan Fauve, Jay Fineberg, Rebecca Hoyle, Gérard Iooss, Edgar Knobloch, Paul Matthews, Ian Melbourne, Werner Pesch, Michael Proctor, Jeff Porter, Hermann Riecke, Anne Skeldon, Jorge Viñals, and Gene Wayne. We thank Michael Proctor for pointing out the effect of higher-order terms on the center manifold, discussed in section 6. Finally, we are grateful to the Isaac Newton Institute for Mathematical Sciences, where part of this work was carried out.

REFERENCES

[1] H. Arbell and J. Fineberg, *Two-mode rhomboidal states in driven surface waves*, Phys. Rev. Lett., 84 (2000), pp. 654–657.

[2] H. Arbell and J. Fineberg, *Pattern formation in two-frequency forced parametric waves*, Phys. Rev. E, 65 (2002), paper 036224.

[3] T. B. Benjamin and F. Ursell, *The stability of the plane free surface of a liquid in vertical periodic motion*, Proc. Roy. Soc. London Ser. A, 225 (1954), pp. 505–515.
[4] T. Besson, W. S. Edwards, and L. S. Tuckerman, *Two-frequency parametric excitation of surface waves*, Phys. Rev. E, 54 (1996), pp. 507–513.
[5] D. Binks and W. van de Water, *Nonlinear pattern formation of Faraday waves*, Phys. Rev. Lett., 78 (1997), pp. 4043–4046.
[6] D. Binks, M. T. Westra, and W. van de Water, *Effect of depth on the pattern formation of Faraday waves*, Phys. Rev. Lett., 79 (1997), pp. 5010–5013.
[7] B. Christiansen, P. Alstrom, and M. T. Levinsen, *Ordered capillary-wave states: Quasi-crystals, hexagons, and radial waves*, Phys. Rev. Lett., 68 (1992), pp. 2157–2160.
[8] J. M. Conway and H. Riecke, *Multiresonant forcing of the complex Ginzburg–Landau equation: Pattern selection*, Phys. Rev. E, 76 (2007), paper 057202.
[9] J. M. Conway and H. Riecke, *Quasipatterns in a model for chemical oscillations forced at multiple resonance frequencies*, Phys. Rev. Lett., 99 (2007), paper 218301.
[10] J. M. Conway and H. Riecke, *Superlattice patterns in the complex Ginzburg–Landau equation with multi-resonant forcing*, SIAM Appl. Dyn. Syst., submitted.
[11] P. Coullet and K. Emilsson, *Strong resonances of spatially distributed oscillators: A laboratory to study patterns and defects*, Phys. D, 61 (1992), pp. 119–131.
[12] S. M. Cox and P. C. Matthews, *Exponential time differencing for stiff systems*, J. Comput. Phys., 176 (2002), pp. 430–455.
[13] S. M. Cox and P. C. Matthews, *Instability and localisation of patterns due to a conserved quantity*, Phys. D, 175 (2003), pp. 196–219.
[14] M. C. Cross and P. C. Hohenberg, *Pattern formation outside of equilibrium*, Rev. Modern Phys., 65 (1993), pp. 851–1112.
[15] J. H. P. Dawes, P. C. Matthews, and A. M. Rucklidge, *Reducible actions of $D_4 \ltimes T^2$: Superlattice patterns and hidden symmetries*, Nonlinearity, 16 (2003), pp. 615–645.
[16] G. Dewel, M. Bachir, P. Borckmans, and S. Metens, *Superlattice structures and quasipatterns in bistable systems*, C. R. Acad. Sci. Paris Ser. IIb, 329 (2001), pp. 411–416.
[17] Y. Ding and P. Umbanhowar, *Enhanced Faraday pattern stability with three-frequency driving*, Phys. Rev. E, 73 (2006), paper 046305.
[18] B. Dionne, M. Silber, and A. C. Skeldon, *Stability results for steady, spatially periodic planforms*, Nonlinearity, 10 (1997), pp. 321–353.
[19] B. Echebarria and H. Riecke, *Sideband instabilities and defects of quasipatterns*, Phys. D, 158 (2001), pp. 45–68.
[20] W. S. Edwards and S. Fauve, *Patterns and quasi-patterns in the Faraday experiment*, J. Fluid Mech., 278 (1994), pp. 123–148.
[21] T. Epstein and J. Fineberg, *Grid states and nonlinear selection in parametrically excited surface waves*, Phys. Rev. E, 73 (2006), paper 055302(R).
[22] M. Frigo and S. G. Johnson, *The design and implementation of FFTW3*, Proc. IEEE, 93 (2005), pp. 216–231.
[23] T. Frisch and G. Sonnino, 2*-dimensional pentagonal structures in dissipative systems*, Phys. Rev. E, 51 (1995), pp. 1169–1171.
[24] M. Golubitsky, I. Stewart, and D. G. Schaeffer, *Singularities and Groups in Bifurcation Theory, Volume* II, Springer, New York, 1988.
[25] H. S. Greenside and M. C. Cross, *Stability analysis of two-dimensional models of* 3*-dimensional convection*, Phys. Rev. A, 31 (1985), pp. 2492–2501.
[26] P. G. Grinevich and P. M. Santini, *Newtonian dynamics in the plane corresponding to straight and cyclic motions on the hyperelliptic curve $\mu^2 = \nu^n - 1$, $n \in \mathbb{Z}$: Ergodicity, isochrony and fractals*, Phys. D, 232 (2007), pp. 22–32.
[27] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*, Springer, New York, 1983.
[28] G. H. Gunaratne, Q. Ouyang, and H. L. Swinney, *Pattern formation in the presence of symmetries*, Phys. Rev. E, 50 (1994), pp. 2802–2820.
[29] J. Halloy, G. Sonnino, and P. Coullet, *Pattern formation in forced reaction diffusion systems with nearly degenerate bifurcations*, Chaos, 17 (2007), paper 037107.

[30] R. Herrero, E. G. Westhoff, A. Aumann, T. Ackemann, Y. A. Logvin, and W. Lange, *Twelvefold quasiperiodic patterns in a nonlinear optical system with continuous rotational symmetry*, Phys. Rev. Lett., 82 (1999), pp. 4627–4630.
[31] R. B. Hoyle, *Pattern Formation: An Introduction to Methods*, Cambridge University Press, Cambridge, UK, 2006.
[32] E. Knobloch, C. Martel, and J. M. Vega, *Coupled mean flow-amplitude equations for nearly inviscid parametrically driven surface waves*, Ann. New York Acad. Sci., 974 (2002), pp. 201–219.
[33] A. Kudrolli and J. P. Gollub, *Patterns and spatiotemporal chaos in parametrically forced surface waves: A systematic survey at large aspect ratio*, Phys. D, 97 (1996), pp. 133–154.
[34] A. Kudrolli, B. Pier, and J. P. Gollub, *Superlattice patterns in surface waves*, Phys. D, 123 (1998), pp. 99–111.
[35] K. Kumar, *Linear theory of Faraday instability in viscous liquids*, Proc. Roy. Soc. London Ser. A, 452 (1996), pp. 1113–1126.
[36] K. Kumar and L. S. Tuckerman, *Parametric instability of the interface between* 2 *fluids*, J. Fluid Mech., 279 (1994), pp. 49–68.
[37] R. Lifshitz and H. Diamant, *Soft quasicrystals: Why are they stable?*, Philos. Mag., 87 (2007), pp. 3021–3030.
[38] R. Lifshitz and D. M. Petrich, *Theoretical model for Faraday waves with multiple-frequency forcing*, Phys. Rev. Lett., 79 (1997), pp. 1261–1264.
[39] P. C. Matthews, *Hexagonal patterns in finite domains*, Phys. D, 116 (1998), pp. 81–94.
[40] P. C. Matthews and S. M. Cox, *Pattern formation with a conservation law*, Nonlinearity, 13 (2000), pp. 1293–1320.
[41] I. Melbourne, *Derivation of the time-dependent Ginzburg–Landau equation on the line*, J. Nonlinear Sci., 8 (1998), pp. 1–15.
[42] N. D. Mermin and S. M. Troian, *Mean-field theory of quasicrystalline order*, Phys. Rev. Lett., 54 (1985), pp. 1524–1527.
[43] H. W. Müller, *Model-equations for* 2*-dimensional quasipatterns*, Phys. Rev. E, 49 (1994), pp. 1273–1277.
[44] H. W. Müller, R. Friedrich, and D. Papathanassiou, *Theoretical and experimental investigations of the Faraday instability*, in Evolution of Spontaneous Structures in Dissipative Continuous Systems, F. H. Busse and S. C. Müller, eds., Springer, Berlin, 1998, pp. 230–265.
[45] H. W. Muller, H. Wittmer, C. Wagner, J. Albers, and K. Knorr, *Analytic stability theory for Faraday waves and the observation of the harmonic surface response*, Phys. Rev. Lett., 78 (1997), pp. 2357–2360.
[46] A. C. Newell and Y. Pomeau, *Turbulent crystals in macroscopic systems*, J. Phys. A, 26 (1993), pp. L429–L434.
[47] E. Pampaloni, P. L. Ramazza, S. Residori, and F. T. Arecchi, 2*-dimensional crystals and quasicrystals in nonlinear optics*, Phys. Rev. Lett., 74 (1995), pp. 258–261.
[48] J. Porter and M. Silber, *Broken symmetries and pattern formation in two-frequency forced Faraday waves*, Phys. Rev. Lett., 89 (2002), paper 084501.
[49] J. Porter and M. Silber, *Resonant triad dynamics in weakly damped Faraday waves with two-frequency forcing*, Phys. D, 190 (2004), pp. 93–114.
[50] J. Porter, C. M. Topaz, and M. Silber, *Pattern control via multifrequency parametric forcing*, Phys. Rev. Lett., 93 (2004), paper 034502.
[51] J. L. Rogers, W. Pesch, O. Brausch, and M. F. Schatz, *Complex-ordered patterns in shaken convection*, Phys. Rev. E, 71 (2005), paper 066214.
[52] A. M. Rucklidge and W. J. Rucklidge, *Convergence properties of the* 8*,* 10 *and* 12 *mode representations of quasipatterns*, Phys. D, 178 (2003), pp. 62–82.
[53] A. M. Rucklidge and M. Silber, *Quasipatterns in parametrically forced systems*, Phys. Rev. E, 75 (2007), paper 055203(R).
[54] D. Shechtman, I. Blech, D. Gratias, and J. W. Cahn, *Metallic phase with long-range orientational order and no translational symmetry*, Phys. Rev. Lett., 53 (1984), pp. 1951–1953.
[55] M. Silber and M. R. E. Proctor, *Nonlinear competition between small and large hexagonal patterns*, Phys. Rev. Lett., 81 (1998), pp. 2450–2453.
[56] M. Silber and A. C. Skeldon, *Parametrically excited surface waves: Two-frequency forcing, normal form symmetries, and pattern selection*, Phys. Rev. E, 59 (1999), pp. 5446–5456.

[57] M. Silber, C. M. Topaz, and A. C. Skeldon, *Two-frequency forced Faraday waves: Weakly damped modes and pattern selection*, Phys. D, 143 (2000), pp. 205–225.
[58] A. C. Skeldon and G. Guidoboni, *Pattern selection for Faraday waves in an incompressible viscous fluid*, SIAM J. Appl. Math., 67 (2007), pp. 1064–1100.
[59] H. G. Solari and G. B. Mindlin, *Quasicrystals and strong interactions between square modes*, Phys. Rev. E, 56 (1997), pp. 1853–1858.
[60] W. Steurer, *Boron-based quasicrystals with sevenfold symmetry*, Philos. Mag., 87 (2007), pp. 2707–2712.
[61] J. Swift and P. C. Hohenberg, *Hydrodynamic fluctuations at the convective instability*, Phys. Rev. A, 15 (1977), pp. 319–328.
[62] C. M. Topaz, J. Porter, and M. Silber, *Multifrequency control of Faraday wave patterns*, Phys. Rev. E, 70 (2004), paper 066206.
[63] C. M. Topaz and M. Silber, *Resonances and superlattice pattern stabilization in two-frequency forced Faraday waves*, Phys. D, 172 (2002), pp. 1–29.
[64] L. S. Tsimring and I. S. Aranson, *Localized and cellular patterns in a vibrated granular layer*, Phys. Rev. Lett., 79 (1997), pp. 213–216.
[65] U. E. Volmar and H. W. Muller, *Quasiperiodic patterns in Rayleigh–Bénard convection under gravity modulation*, Phys. Rev. E, 56 (1997), pp. 5423–5430.
[66] C. Wagner, H. W. Muller, and K. Knorr, *Pattern formation at the bicritical point of the Faraday instability*, Phys. Rev. E, 68 (2003), paper 066204.
[67] M. T. Westra, D. J. Binks, and W. Van de Water, *Patterns of Faraday waves*, J. Fluid Mech., 496 (2003), pp. 1–32.
[68] W. B. Zhang and J. Viñals, *Square patterns and quasipatterns in weakly damped Faraday waves*, Phys. Rev. E, 53 (1996), pp. R4283–R4286.
[69] W. B. Zhang and J. Viñals, *Pattern formation in weakly damped parametric surface waves*, J. Fluid Mech., 336 (1997), pp. 301–330.
[70] W. B. Zhang and J. Viñals, *Numerical study of pattern formation in weakly damped parametric surface waves*, Phys. D, 116 (1998), pp. 225–243.

SIAM J. Applied Dynamical Systems
Vol. 8, No. 1, pp. 348–370
© 2009 Society for Industrial and Applied Mathematics

A Global Bifurcation and the Appearance of a One-Dimensional Spiral Wave in Excitable Media*

Eric N. Cytrynbaum† and Timothy J. Lewis‡

Abstract. It is well known that excitable media can sustain "fast" stable traveling pulses of excitation. In models for which analysis is tractable, the fast pulse solution is known to appear through a saddle-node bifurcation accompanied by a "slow" unstable traveling pulse. Furthermore, the uniform rest state is also a stable solution. It is generally assumed that the boundary between the basins of attractions of the rest state and fast pulse (i.e., threshold) consists of the stable manifold of the slow pulse. We use numerical experiments to explore this issue. Our results indicate that, near the saddle-node bifurcation, the stable manifold of the slow pulse does indeed act as the threshold between the rest state and fast pulse. However, further away from the saddle-node bifurcation, a global bifurcation involving the heteroclinic connections between slow and fast pulses occurs. This bifurcation gives rise to an unstable periodic solution that has been referred to as a one-dimensional spiral wave. Beyond the bifurcation point, the stable manifold of the one-dimensional spiral wave appears to act as the threshold between the rest state and fast pulse.

Key words. traveling wave, basin of attraction, reflections, shooting method

AMS subject classifications. 34C37, 35K57, 37G15, 37N25

DOI. 10.1137/080722254

1. Introduction.

Spatially distributed excitable media are ubiquitous in nature, appearing in many chemical, physical, and biological systems. These systems have a stable uniform "rest" state, but when perturbed locally above a threshold, they have the ability to generate pulses of elevated activity that propagate without damping (see [22]). Detailed information about the threshold separating the rest state from propagated pulses can be extremely important. For example, it could lead to a better understanding of how aberrant propagated activity can arise in cardiac tissue and induce dangerous arrhythmias [45, 46, 42, 47]. It could also lead to insight into optimal stimuli for annihilating this pathological activity [15, 41, 27, 43, 25]. In this manuscript, we explore the structure of the threshold surface between the basins of attraction of the stable rest state and a stable propagated pulse in an idealized model of excitable media.

1.1. The FitzHugh–Nagumo equations.

The general FitzHugh–Nagumo (FHN) system is a family of two-variable models that capture the essential properties of spatially distributed excitable media and serve as minimal models for excitable systems. The original FitzHugh–Nagumo system [11, 37] was formulated as a reduction of the Hodgkin–Huxley model for nerve

*Received by the editors April 24, 2008; accepted for publication (in revised form) by D. Barkley October 25, 2008; published electronically February 19, 2009.
http://www.siam.org/journals/siads/8-1/72225.html
†Department of Mathematics, University of British Columbia, Vancouver, BC, V6T 1Z2, Canada (cytryn@math.ubc.ca).
‡Department of Mathematics, University of California, Davis, CA 95616 (tjlewis@ucdavis.edu).

axon [18]. Since then, many variants of the general FHN model have been proposed. Some of these models were motivated by specific physiological systems, e.g., the ionic-current–based Morris–Lecar (ML) model [32]; others were proposed for the sake of analytical tractability, e.g., the piecewise linear McKean and Pushchino models [30, 35]. In their general form, the FHN equations are

$$\begin{aligned} \frac{\partial u}{\partial t} &= D\nabla^2 u + f(u,v), \\ \frac{\partial v}{\partial t} &= \epsilon g(u,v), \end{aligned} \tag{1}$$

where u is the excitation or activator variable, and v is the recovery or inhibitor variable. For typical parameter sets, $f(u,v)$ has three zeros as a function of u over a finite interval of fixed values of v and $\frac{\partial f}{\partial u}(u,v) < 0$ at the lower and upper zeros giving the u equation a bistable structure on that interval. Outside that interval of v, $f(u,v)$ has only one zero. $g(u,v)$ is an increasing function of u and decreasing in v so that the kinetic (spatially homogeneous) equations have only one stable steady state which sits on the lowest solution branch of the curve $f(u,v) = 0$. The phase plane for one example of such a system is shown in Figure 1 (bottom half of the inset). We consider dynamics in one spatial dimension with diffusive coupling in the excitation variable u only. We concentrate on the case of a spatially periodic domain (i.e., a ring), but we will also consider an infinite domain when appropriate.

1.2. Steady traveling solutions. Three solutions to the FHN system are well known and well characterized: (i) the spatially uniform stable rest state, i.e., the solution to $f(u,v) = g(u,v) = 0$, (ii) a fast-moving and generally stable traveling pulse, and (iii) a slow-moving unstable traveling pulse [33, 12, 38, 21] (see Figure 1). For the periodic domain, these traveling pulse solutions correspond to activity circulating around a ring with a fixed velocity. Composite solutions consisting of multiple traveling pulses are also possible.

The level of excitability of the FHN system, as reflected in the upstroke velocity and propagation speed of an action potential, is dependent on the parameter ϵ: smaller values of ϵ correspond to dynamics with stronger excitability. The stable rest state exists for all positive ϵ, whereas the fast and slow traveling pulse solutions exist only for positive values of ϵ below a critical value ϵ_{SNB}. These traveling pulses disappear via a saddle-node bifurcation as ϵ is increased through ϵ_{SNB} [38] (Figure 1). The terms "fast" and "slow" arise from the observation that the speed of the fast pulse is asymptotically $O(1)$ with respect to ϵ [17, 5], while the speed of the slow pulse is $O(\sqrt{\epsilon})$ [17, 6, 14]. In the singular limit ($\epsilon \to 0$), the fast pulse can be calculated as a composite structure formed by matched asymptotics of solutions to the singular inner and outer equations [5, 44, 7], and the slow pulse approaches a standing pulse solution to the inner (u) equation of system (1) with v set to the kinetic steady-state constant value [14]. The reduced system, consisting of the inner equation with fixed v, is often referred to as the Nagumo equation [33].

1.3. Threshold solutions. Thresholds that separate two stable states are usually associated with saddle solutions. The unstable manifold of the saddle solution forms heteroclinic connections with the two stable states, and the stable manifold of the saddle solution forms the boundary (separatrix) between the basins of attraction of the two states. Thus, the stable

manifold of the saddle solution acts as a threshold surface between the states. We refer to the stable manifold of the saddle solution as the "threshold" and the saddle solution itself as the "threshold solution."

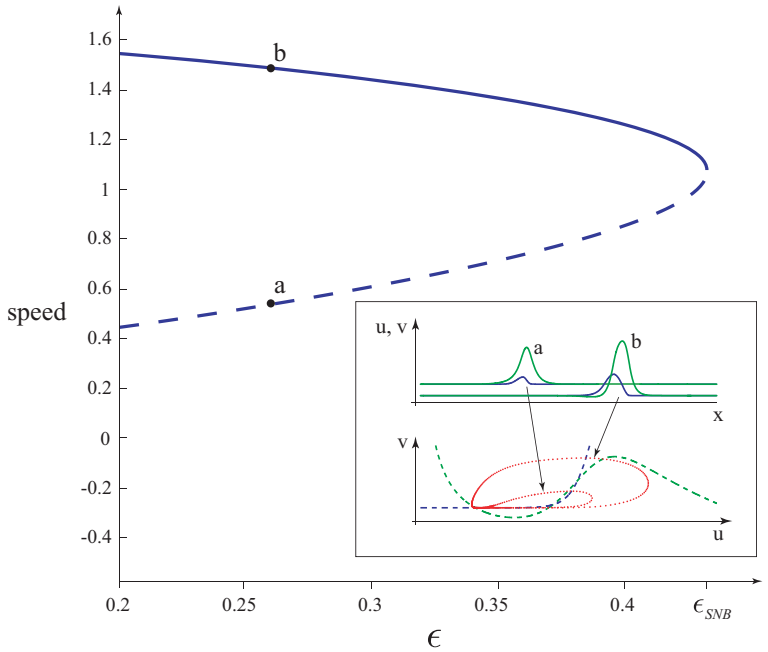


Figure 1. *The speeds of slow (dashed) and fast (solid) pulses plotted as functions of ϵ, calculated numerically for the ML system (see the appendix and Table 1). A saddle-node bifurcation of the fast and slow pulses occurs near $\epsilon = 0.43$. Examples of slow* (a) *and fast* (b) *pulses for $\epsilon = 0.26$ are plotted as functions of x and in the phase plane as curves parametrized by x (inset).*

For the FHN system, it is generally assumed that the slow pulse is the threshold solution between the rest state and the fast pulse. This should indeed be the case for values of ϵ close to the saddle-node bifurcation at ϵ_{SNB}. For $\epsilon > \epsilon_{SNB}$, the rest state is globally attracting, whereas for $\epsilon < \epsilon_{SNB}$, both the rest state and the fast pulse are presumably stable. Consistent with the structure of a saddle-node bifurcation, we infer that the slow pulse is a saddle solution with one unstable mode for values of ϵ near ϵ_{SNB}. The one-dimensional (1D) unstable manifold of the slow pulse is comprised of heteroclinic connections between the slow pulse and the stable rest state and between the slow pulse and the fast pulse. The stable manifold of the slow pulse in turn acts as a threshold surface between the uniform rest state and the fast pulse. Thus, the slow pulse is the threshold solution. Based on the symmetry of the system, the entire structure described above actually appears twice in the phase space, corresponding to both left and right directions of travel.

For small ϵ, there seems to be a fundamentally different scenario. The standing pulse of the Nagumo equation, which is the asymptotic limit of the slow pulse in the singular limit $\epsilon \to 0$, is known to have a saddle structure with one unstable mode [13]; the stable manifold of the standing pulse is the threshold surface between the stable rest state and the excited state [13, 31, 34]. Furthermore, for ϵ small, the slow pulse for the FHN equations also has

a saddle structure with one unstable mode [14]. However, this does not imply that the slow pulse lies on the threshold surface between rest and the corresponding fast pulse. Small excitatory perturbations of the standing pulse in the Nagumo equation evolve into not one wavefront but two counterpropagating wavefronts [10]. When incorporated into a matched asymptotic analysis, this implies that, for ϵ small, perturbations to the slow pulse should lead either directly to the rest state or to two counterpropagating fast pulses rather than a single fast pulse. On a ring, the two counterpropagating fast pulses would eventually collide and annihilate one another, leading the solution to the rest state. As there is only one unstable mode for ϵ small [14], this observation indicates that the slow pulse is *not* on the threshold surface between the rest state and the fast pulse but instead lies completely within the basin of attraction of the rest state. On an infinite domain, the slow pulse appears to lie on the threshold between the rest state and the composite solution of two fast pulses propagating indefinitely in opposite directions.

To summarize, although one half of the unstable manifold of the slow pulse always leads to some form of propagated excitation, the details of the long-term behavior appear to be fundamentally different at opposite ends of the ϵ interval for which traveling pulses exist. Near the saddle-node bifurcation, small perturbations to a slow pulse lead either to rest or to a single fast pulse, and the slow pulse is the threshold solution. For small ϵ, small perturbations to a slow pulse lead either to rest or to two counterpropagating pulses and ultimately to rest for the case of a periodic domain. Thus, for ϵ below a critical value, the threshold solution must be something other than the slow pulse. This critical value is the point at which the global bifurcation, which is the main focus of this paper, occurs.

1.4. Outline. Here, we carry out a series of numerical experiments to examine the changes in threshold structure of the FHN system. First, we use a shooting method to explore the long-term behavior of the slow pulse (i.e., dynamics along its unstable manifold). As predicted, there is a fundamental change in this behavior at a critical value $\epsilon = \epsilon^*$. For $\epsilon > \epsilon^*$, suprathreshold perturbations to the slow pulse lead to a single fast pulse; for $\epsilon < \epsilon^*$, suprathreshold perturbations to the slow pulse lead to two counterpropagating fast pulses. Next, we use a second shooting method to directly examine the threshold between the rest state and the fast pulse for the periodic domain. Consistent with the first study, within the numerical accuracy of the technique, we find that for $\epsilon > \epsilon^*$, the threshold solution is indeed the slow pulse. However, for $\epsilon < \epsilon^*$, the threshold between rest and the fast pulse is not the slow pulse but is instead an unstable solution akin to a 1D transect of a spiral wave, which has been referred to as a 1D spiral wave [26, 9, 28]. Finally, we describe the global bifurcation through which the unstable manifold of the slow pulse redirects itself from single fast pulse to double fast pulse and, simultaneously (or nearly so), through which the 1D spiral wave appears.

2. Technical details. The calculations presented throughout this paper are carried out using ML local dynamics on a periodic domain (see the appendix; parameters are as given in Table 1). As the domain is much larger than the space scale associated with individual pulses, we assume the observed dynamics are representative of those of the infinite domain case as well, with the sole exception of pulse collisions due to the periodicity of the domain. With this in mind, we use *leftward* and *rightward* propagation to refer to *clockwise* and *counterclockwise*

propagation, respectively.

To aid in the description of activity composed of multiple pulses, we introduce the following notation. A rightward propagating fast pulse is denoted as F and a leftward propagating fast pulse as f. Similarly, rightward and leftward propagating slow pulses are referred to as S and s, respectively. The uniform rest state is denoted by R.

Multipulse activity composed of several pulses can be described by sequences of these letters that list the pulses in the order of their appearance along the spatial domain. For example, on the infinite domain, fFF would denote, listing from $x = -\infty$ to $x = \infty$, one fast pulse moving leftward and two fast pulses moving rightward. Although such a solution is not a steady-state-traveling solution because the distances between pulses are time-dependent, it is nonetheless a steady behavior in the more general sense that all three pulses continue to propagate for all time with preserved sequence fFF. In contrast, a sequence of pulses FfF is not steady and converges to F after the collision of the f and leftmost F. On a periodic domain, any sequence of pulses eventually converges to F^n, f^n, or the rest solution as determined by the winding number of the initial condition in the phase plane [24, 23]. Note that the maximum power n is determined by the size of the ring and the dispersion curve for the fast pulse (see [22] for a discussion of dispersion curves for FHN). We make the further assumption that convergence properties of single pulses are preserved in the presence of other pulses. For example, if a perturbation to s drives it to f, then an analogous perturbation to sF drives it to fF.

As we are interested in threshold behavior, we define the following terminology which will be useful throughout the rest of this paper:

- *R/E threshold*: The set of initial conditions which lie on the boundary between initial conditions that converge to rest without any form of propagated excitation and initial conditions that lead to some form of propagated excitation (possibly transient due to collision on a periodic domain).
- *R/E-threshold solution*: The solution to which initial conditions on the R/E threshold converge.
- *R/F threshold*: The set of initial conditions which lie on the boundary between initial conditions that ultimately converge to rest and initial conditions that ultimately converge to the single fast pulse solution.
- *R/F-threshold solution*: The solution to which initial conditions on the R/F threshold converge.

3. Threshold and the long-term behavior of the slow pulse.

In this section, we describe a numerical shooting method used to examine the R/E threshold, i.e., the threshold between the rest state R and *any significant* suprathreshold activity. We are concerned with any suprathreshold activity, whether sustained or transient; so, for example, an fF solution on a periodic domain would count as suprathreshold despite its ultimate convergence to rest.

The shooting method involves a search through a one-parameter family of initial conditions and is designed to converge to initial conditions that lie on the R/E threshold surface. At advanced stages of the shooting method, we are thus able to observe (i) the approach to the threshold solution along its stable manifold, (ii) the divergence from the threshold solution along its unstable manifold, and (iii) the subsequent long-term evolution to the rest state R

or to suprathreshold activity. As expected, the slow pulse S appears to be the R/E-threshold solution for all values of $\epsilon < \epsilon_{SNB}$. However, the long-term behavior of the unstable manifold of the slow pulse depends on the value of ϵ, with a qualitative change at a critical value of ϵ which we define as ϵ^*.

3.1. The R/E-threshold shooting method. The shooting method consists of a series of simulations of the FHN equations (1), using "square pulse" initial conditions $u(x,0) = AP(x,x_1,x_2)$ and $v(x,0) = v_0P(x,x_3,x_4)$, where $x_3 < x_1 < x_4 < x_2$ and P is a plateau function defined in terms of Heaviside functions $P(y,y_1,y_2) = H(y-y_1)H(y_2-y)$ (see Figure 2, panel B0). The parameter v_0 and x_1, x_2, x_3, x_4 are set appropriately so that *initial* leftward propagation is prevented. For sufficiently large A, a single rightward propagating fast pulse F is generated. On the other hand, when $A = 0$, the solution converges to rest R. Thus, a shooting method that systematically varies the parameter A can be used to approach initial conditions that lie on the threshold surface between the rest state R and suprathreshold propagated activity.

From iteration to iteration of the shooting method, the value of A is changed as follows. Initially, A_{min} and A_{max} are chosen so that the solution converges to a rightward fast pulse (F) for $A = A_{max}$ and the solution converges to rest (R) for $A = A_{min}$. Simulations are then run with an intermediate value of $A = (A_{min} + A_{max})/2$. If, at any time, $\max_x u(x,t)$ drops below a value just above the resting value of u, the simulation is halted, and A_{min} is reset to the current value of A. On the other hand, if $\max_x u(x,t)$ rises above a value just below the peak u value in the fast pulse, the simulation is halted, and A_{max} is reset to the current value of A. A new intermediate value of A is then used in the initial condition for the subsequent simulation. This procedure is iterated until A_{min} and A_{max} are indistinguishable in double precision. This final value of A is used as an approximation of A_{RE}, the critical value of A for which the initial conditions lie on the R/E-threshold surface. The shooting method is carried out for a sequence of ϵ values ranging from 0.04 to 0.40 in increments of 0.02; this provides a broad perspective on near-critical behavior (see Figure 2). Additional simulations are carried out for ϵ between 0.28 and 0.29 in increments of 0.001 to narrow down the bounds on the value of ϵ^*.

To determine when solutions approach traveling pulses, the discrete version of the following function is calculated throughout the simulations:

$$d_{tp}(t) = \min_{x_0}(||u(x,t) - u(x-x_0,t-\tau)|| + ||v(x,t) - v(x-x_0,t-\tau)||) + ||u(x,t)||.$$

This is a measure of how close the solution $u(x,t)$ is to an optimal shift in space of $u(x,t-\tau)$, where τ is a fixed time delay. Note that $x_0^{\min}/\tau$ provides an estimate of the traveling speed. When the $||u(x,t)||$ term is omitted, $d_{tp}(t)$ should approach zero for a solution converging to any traveling pulse or the uniform rest state. This additional term is introduced to distinguish between convergence to the fast pulse, the slow pulse, and the uniform rest state.

3.2. Results: A change in long-term slow-pulse behavior at ϵ^*. Figure 2 shows intensity plots of $u(x,t)$ at four different values of A during the shooting method, for each of four values of ϵ. The panels within this figure are described throughout the rest of this section.

At all values of ϵ, for $0 \le A < A_{RE}(\epsilon)$, only subthreshold responses are elicited, and the

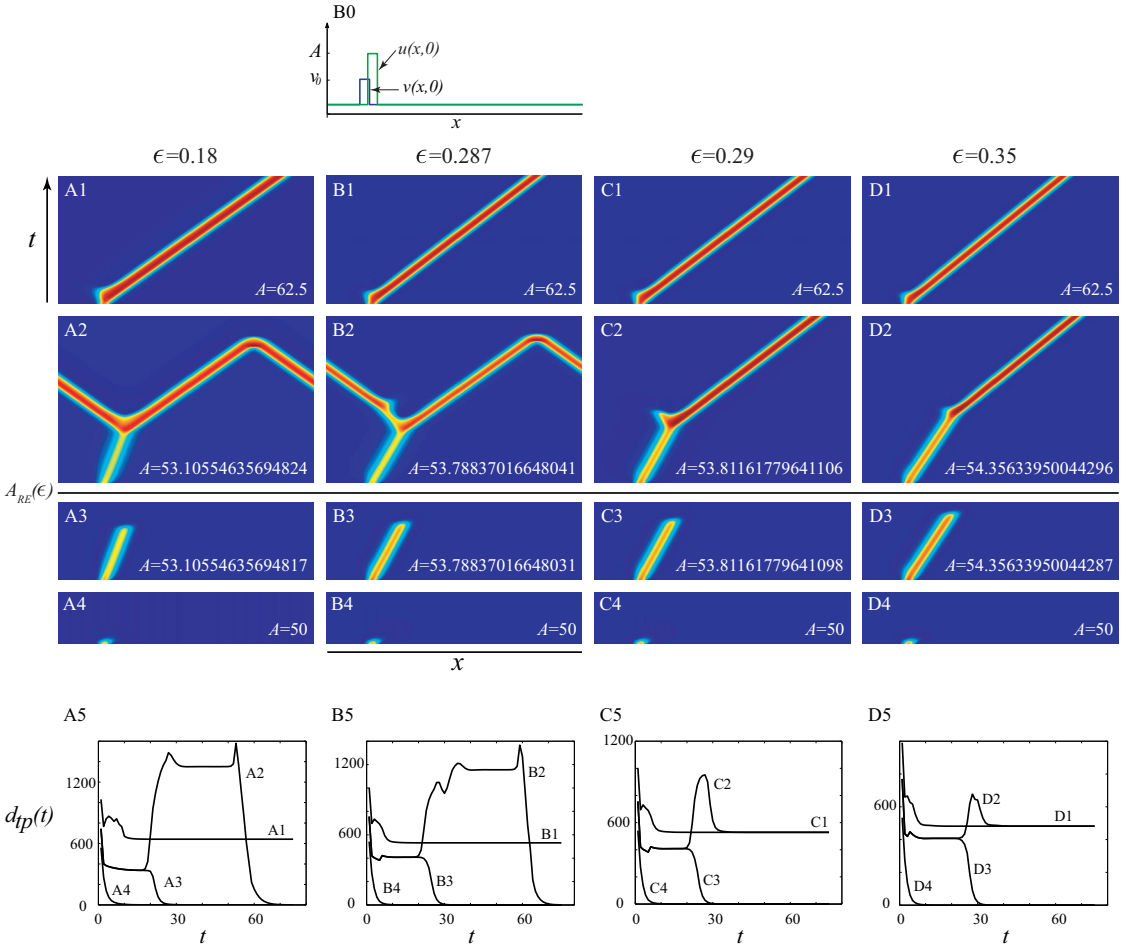


Figure 2. *The results of the shooting method used to find the R/E-threshold solution at four values of ϵ are shown as intensity images ($u(x,t)$ only). A schematic of the initial condition used in all simulations is shown in panel* B0 *(top). All parameters are given in Table* 1 *except for ϵ and A, the amplitude of u in the initial condition, which are specified in the figure for each column and panel, respectively. Panels* A1–D1 *show direct convergence to the fast pulse for A sufficiently large. Panels* A4–D4 *show convergence directly to rest for a sufficiently low value of A. Panels* A2–D2 *and* A3–D3 *show transient convergence to the slow pulse followed by excitation and convergence to rest, respectively. For $\epsilon < \epsilon^*$ (*A2 *and* B2*), excitation consists of two fast pulses but for $\epsilon > \epsilon^*$ (*C2 *and* D2*) only one fast pulse emerges. Approaching ϵ^* from below (*A2→B2*), note that the retrograde fast pulse is gradually pinched down until it finally fails to propagate above ϵ^* (*C2*). Panels* A5–D5 *show $d_{tp}(t)$ for each simulation portrayed in the upper panels. Three or four flat regions are evident in each simulation corresponding to the rest state R, the slow pulse S, the fast pulse F, and, when present, the double fast pulse fF.*

system relaxes to the rest state R (panels A3–D3, A4–D4). Furthermore, values of A slightly below $A_{RE}(\epsilon)$ lead to transient slow pulses S that eventually collapse to rest R (panels A3–D3).

For values of $\epsilon \geq 0.288$, values of $A > A_{RE}(\epsilon)$ produce responses that develop into single fast pulses F (panels C1, C2, D1, D2). For values of A only slightly above the critical value $A_{RE}(\epsilon)$, there are relatively long transients, during which the solution appears to approach

the slow pulse S before eventually converging to the single fast pulse F (panels C2, D2).

At values of $\epsilon \leq 0.287$ and A well above $A_{RE}(\epsilon)$, as seen at larger values of ϵ, a single fast pulse F quickly emerges from the initial condition (panels A1, B1). However, for A only slightly above $A_{RE}(\epsilon)$, the slow pulse S again makes a transient appearance, but, in this range of ϵ, excitation does not lead to a single fast pulse but instead to a composite solution consisting of two counterpropagating fast pulses fF (panels A2, B2).

The above results indicate that there is a fundamental change in the threshold behavior at a critical value $\epsilon = \epsilon^*$, where $0.287 < \epsilon^* < 0.288$. For $\epsilon > \epsilon^*$, the slow pulse S is both the R/E-threshold solution and the R/F-threshold solution. For $\epsilon < \epsilon^*$, the slow pulse S is the R/E-threshold solution, but it is not the R/F-threshold solution; instead, it is the R/fF-threshold solution. This is a particularly important distinction in the case of a periodic domain, where the two fast pulses fF would eventually collide and annihilate one another, with the solution eventually converging to rest. That is, for a periodic domain, both components of the unstable manifold of the slow pulse S are entirely contained within the domain of attraction of the rest state R.

The overall trend in the suprathreshold behavior of the slow pulse for increasing values of ϵ is well illustrated in panels A2–D2. At low values of ϵ, the slow pulse S can blow up into a retrograde and anterograde fast pulse fF. As the value of ϵ approaches ϵ^*, retrograde propagation is gradually inhibited (panel B2) until it is finally cut off altogether above ϵ^* (panel C2). Although we did not verify it numerically, we postulate that at ϵ^*, the slow pulse S grows into a retrograde slow pulse and anterograde fast pulse sF. Of course, this could happen only on the infinite domain. On a periodic domain, a slow and fast pulse would eventually "collide," leading to some other long-term behavior.

To verify that the intensity plots actually correspond to the composite solutions described here in terms of the fast and slow pulses, $d_{tp}(t)$ for each simulation is plotted in panels A5–D5. Initial transients are followed by an approach (i) to rest (labeled A4–D4), (ii) to a transient low level (labeled A3–D3) corresponding to the slow pulse S, (iii) to an intermediate level corresponding to the fast pulse F (labeled A1–D1), or (iv) to a transient high level corresponding to a pair of counterpropagating fast pulses (labeled A2, B2). As described above in the definition of $d_{tp}(t)$, these flat regions indicate that it is in fact meaningful to describe these solutions in terms of the known traveling pulses.

3.3. A qualitative interpretation of the transition at ϵ^*: Phase plane and refractory dynamics. This transition in behavior at ϵ^* can be understood in terms of the effect that the slow pulse has on the refractory variable (v). For low values of ϵ, the slow pulse has a relatively small amplitude in u and has an almost negligible effect on the refractory variable as it propagates through a region. Thus, the "back" of the slow pulse in this regime lies relatively low in the v versus u phase plane as depicted in the left panel of Figure 3. Consequently, as the solution diverges from the slow pulse toward excitation, the refractory barrier behind the slow pulse does not form a strong barrier to retrograde propagation. Therefore, both the "front" and the "back" of the slow pulse grow into true fronts (see [44] for a detailed discussion of front propagation) and develop into two counterpropagating fast pulses. For larger values of ϵ, i.e., closer to the saddle node, the slow pulse has a larger amplitude in u and has a more pronounced effect on v (see the right panel of Figure 3). In this case, as the solution

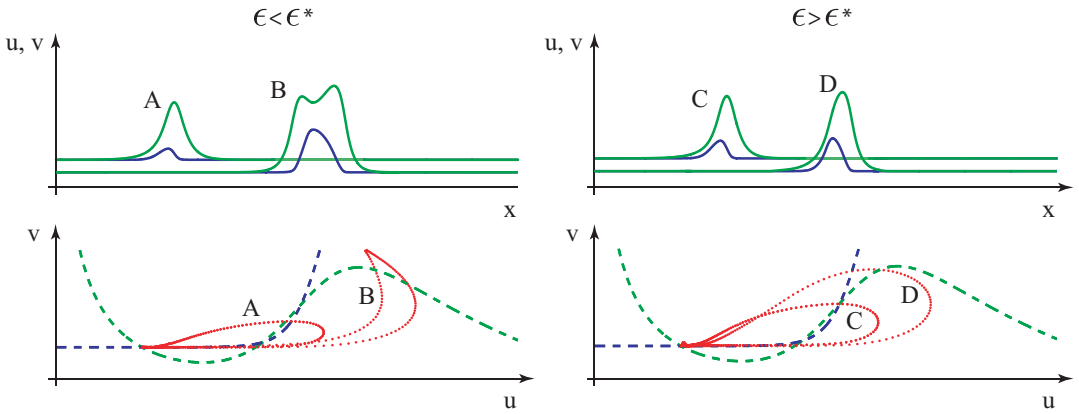


Figure 3. *Comparison of slow-pulse excitation above and below the critical value of ϵ^*. Solutions $u(x,t), v(x,t)$ are plotted as functions of x (top) and parametrized by x in the $u-v$ phase plane (bottom) at two different points in time, when the solutions are close to the slow pulse* (A, C) *and during subsequent excitation* (B, D). *For $\epsilon < \epsilon^*$ (left panel), the "back" of the slow pulse is low in the phase plane and grows into a true front leading to two counter-propagating fast pulses. We use $\epsilon = 0.26$ to illustrate this case. For $\epsilon > \epsilon^*$ (right panel, illustrated with $\epsilon = 0.38$), the "back" of the slow pulse is higher in the phase plane and grows into a true back leading to a single fast pulse.*

diverges from the slow pulse toward excitation, retrograde propagation is no longer possible; the "front" of the slow pulse evolves into a true front, the "back" evolves into a true back, and only a single anterograde fast pulse emerges.

4. Threshold solutions: Slow pulses and 1D spiral waves. The numerical experiments in the previous section indicate that the R/F solution on the periodic domain is the slow pulse for $\epsilon^* < \epsilon < \epsilon_{SNB}$. However, for $0 < \epsilon < \epsilon^*$, the unstable manifold of the slow pulse is embedded in the basin of attraction of the rest state. In order to identify the R/F-threshold solution that replaces the slow pulse for $\epsilon < \epsilon^*$, we implement a second shooting method. We stress that this shooting method is designed to identify the threshold between the rest state R and the sustained single fast pulse F (i.e., the R/F-threshold solution) specifically and not the threshold between the rest state R and any transient suprathreshold activity (i.e., the R/E-threshold solution) as in the previous section.

4.1. The R/F-threshold shooting method. To identify the R/F-threshold solution, simulations are performed starting with square-pulse initial conditions as described in the previous section. However, the value of the shooting parameter A is updated differently. Initially, we set $A_{min} = 0$ and A_{max} to be sufficiently large so that the solution with $A = A_{max}$ converges to a single rightward propagating fast pulse F. A simulation with an intermediate value of $A = (\frac{3}{4}A_{min} + \frac{1}{4}A_{max})$ is then performed (the asymmetric update accelerates convergence due to asymmetric computation times). At every time step, the solution is compared to both the rest state R and the single fast pulse F. As soon as the l_2 distance between the solution and either of these two steady solutions falls below a small tolerance value, the simulation is terminated and the values of A_{max} or A_{min} are updated appropriately. In the case of convergence to rest R, the value of A_{min} is updated to A; for convergence to the fast pulse F, the value of A_{max} is updated to A. The new value of A for the next iteration is taken to

be the weighted average of A_{max} and A_{min} as defined above. This process is repeated until A_{max} and A_{min} are indistinguishable in double precision. This final value of A is used as an approximation of A_{RF}, the critical value of A for which the initial condition lies on the R/F-threshold surface. The shooting method is carried out for the same sequence of ϵ values as was used in the numerical experiments described in section 3.

4.2. Results: A 1D spiral wave replaces the slow pulse as the R/F-threshold solution below the critical value ϵ^*. For $\epsilon^* < \epsilon < \epsilon_{SNB}$, the results for this second shooting method are identical to those obtained with the shooting method from the previous section; i.e., the R/F-threshold solution is the slow pulse.

The sequence of images in Figure 4a shows the results of the shooting method for $\epsilon = 0.18$. These results are characteristic for $\epsilon < \epsilon^*$. Simulations with small values of A (e.g., $A = 50$, bottom panel) approach the rest state R, and those with sufficiently large values of A (e.g., $A = 62.5$, top panel) initiate single fast pulses F. After several iterations of the shooting method (e.g., $A = 53.69$, first panel above A_{RE}), two counterpropagating fast pulses fF transiently arise before the system ultimately returns to the rest state R through pulse collision on the periodic domain. This is the same behavior described in the previous section. However, curious transient spatiotemporal behavior emerges with further iteration of the shooting method. At $A = 54.008923$ (third panel from top), the initial condition elicits a forward-propagating fast pulse and then a retrograde-propagating fast pulse, and then, surprisingly, a new forward-propagating fast pulse arises in its refractory wake. This results in a transient fFF, which eventually evolves into a single rotating fast pulse. As the shooting method progresses further, converging toward $A = A_{RF}$ and the corresponding solution on the R/F-threshold surface, more and more new pulses appear: retrograde fast pulses produce new anterograde pulses in their wakes, and the new anterograde pulses produce new retrograde pulses in their wakes. In each case, the process repeats until the activity completely dies out or a single fast pulse F is left. When the total number of pulses generated is even (e.g., $ffFF$), the system ultimately goes to the rest state R; i.e., all activity circulating around the periodic domain is eventually eliminated due to the pairwise collisions and annihilation of pulses. On the other hand, when the total number of pulses generated is odd (e.g., $ffFFF$), the system evolves into a state with a single rightward-propagating fast pulse F.

Despite these peculiar dynamics, there still appears to be a distinct threshold value $A = A_{RF}$ below which the system evolves to rest R and above which the system evolves to a single rightward-propagating fast pulse F. As A approaches A_{RF}, the total number of pulses increases in a distinctive order. When A is increased toward A_{RF} from below, the number of pulses appears to run through a sequence of even integers $\{0, 2, 4, 6, \dots, 2n, \dots\}$. When A is decreased toward A_{RF} from above, the number of new pulses appears to run through a sequence of odd integers $\{1, 3, 5, 7, \dots, 2n + 1, \dots\}$. $A = A_{RF}$, which yields initial conditions exactly on the boundary between the basins of attraction of the rightward-propagating fast pulse F and the rest state R (i.e., the R/F-threshold surface), appears to be an accumulation point for which infinitely many pulses are generated.

In summary, as A increases, we infer that the full sequence of transient pulse trains generated is

$$R,\ fF,\ ffFF, \dots, f^nF^n, \dots, f^\infty F^\infty, \dots, f^nF^{n+1}, \dots, fFF,\ F. \tag{2}$$

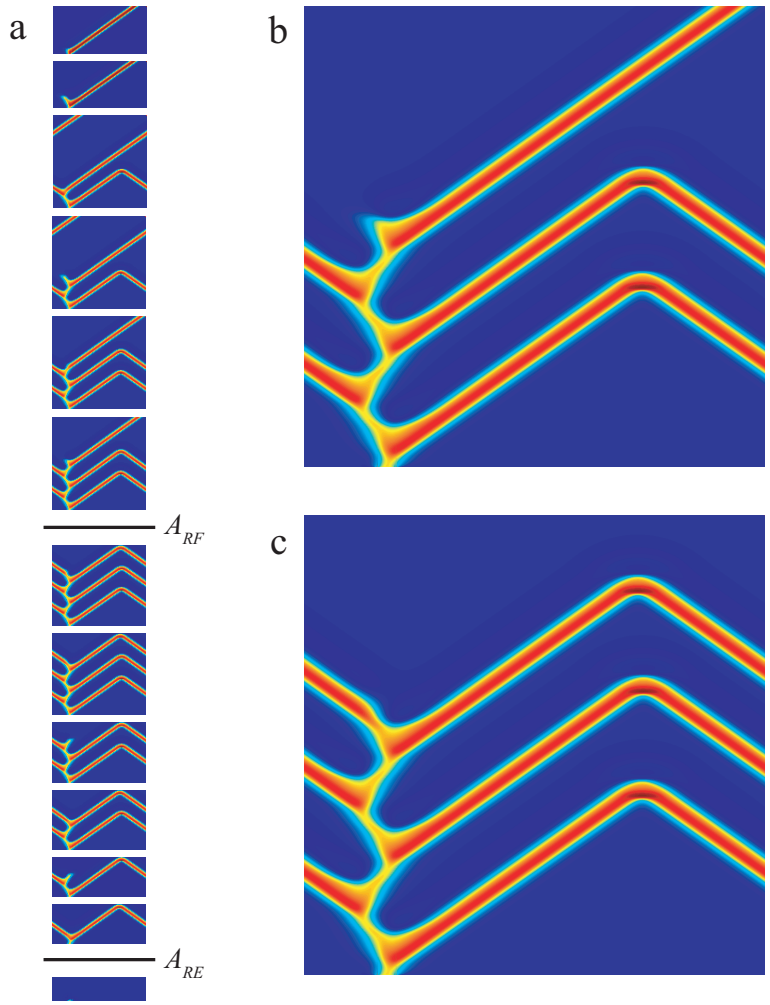


Figure 4. *Results of the shooting method used to find the R/F threshold for $\epsilon = 0.18$ (typical of $\epsilon < \epsilon^*$).* (a) *Space-time plots for a range of values of A used during the shooting method (increasing A from bottom to top, vertical increments not to scale). Below A_{RE}, convergence is directly to rest. Between A_{RE} and A_{RF}, the number of generated pulses increases in pairs giving an even sequence. Notice that in the third image from the bottom excitation of the retrograde pulse is slow enough that a second anterograde fast pulse almost appears (looking briefly like a slow pulse and then dying out). In the fourth image from the bottom, this anterograde excitation succeeds in forming a fast pulse, but because $\epsilon < \epsilon^*$ and the transient is close to a slow pulse, a pair of counterpropagating pulses emerges. This is repeated in each pair-adding stage of the sequence. Above A_{RF} the number of generated pulses decreases in pairs through an odd sequence. Magnified versions of the two solutions closest to and on either side of the R/F threshold are shown in* (b) *(above the R/F threshold) and* (c) *(below the R/F threshold).*

On a periodic domain, counterpropagating pulses eventually collide so that even sequences settle down to rest and odd sequences converge to the fast pulse F. The R/F-threshold solution itself, corresponding to $f^\infty F^\infty$, appears to be an oscillation consisting of a sequence of n, $n+1$, $n+2$, and n fast pulses in succession as retrograde and anterograde fast pulses appear one at a time and then a pair of pulses collide on the far side of the domain. The

number n is determined by the frequency of the pulse generation, the size of the domain, and the dispersion curve for the fast pulse.

On the infinite domain, the solution corresponding to $f^\infty F^\infty$ cannot properly be interpreted as the R/F-threshold solution because, in this regime, there are an infinite number of intervening steady solutions between the basins, i.e., the sequence listed above. This arises from the lack of collisions that occur on the periodic domain.

The above observations suggest that, for $\epsilon < \epsilon^*$, the R/F-threshold solution is an unstable periodic solution consisting of a pulse-generating region that alternately sheds retrograde and anterograde pulses. Furthermore, this periodic orbit must be a saddle structure with a single unstable mode because it can be found using a single-parameter shooting method. As mentioned above, initial conditions exactly on its stable manifold (i.e., the R/F-threshold surface) would produce sustained periodic activity. As the initial conditions get closer to the threshold surface, more and more cycles are seen as the solution winds away from the unstable periodic solution, generating the sequence of pulses before ultimately approaching either the single fast pulse or the uniform rest state.

This type of periodic solution is reminiscent of spiral waves in two spatial dimensions, and thus it has been referred to as a 1D spiral wave [26, 8, 9, 28]. The same sequence of transient behavior described above has been observed in 1D excitable media with a local inhomogeneity [9, 28], and a 1D spiral wave was postulated to be the underlying mechanism. In fact, Ermentrout and Rinzel [9] conjectured that such a 1D spiral structure could exist in homogeneous excitable media and numerically demonstrated its existence in a linear array of six "cells." To our knowledge, our work is the first direct evidence for the existence of the 1D spiral wave in the full homogeneous PDE model.

4.3. Transition at ϵ^* and the link between the slow pulse and the 1D spiral wave. Figure 5 provides a summary of the results of the search for the R/F-threshold solution along with the excitation (R/E) threshold results from section 3. The black dashed curve $A_{RE}(\epsilon)$ lies along the set of initial conditions that converge to the slow pulse. The red curve $A_{RF}(\epsilon)$ represents the set of initial conditions that converge to the R/F-threshold solution. At and above $\epsilon = 0.29$, these two curves are indistinguishable. We note that, for $\epsilon = 0.288$ and $\epsilon = 0.289$, the nature of the threshold solution found using the shooting method could not be clearly distinguished due to numerical limitations. This ambiguous regime just above ϵ^* is addressed in greater detail in the appendix. Below ϵ^*, the R/F-threshold solution appears to be a 1D spiral wave.

These results clarify that there are two qualitatively different threshold solutions on either side of ϵ^*, which raises the issue of the nature of the transition at ϵ^*. As ϵ approaches ϵ^* from below, the period of the unstable 1D spiral appears to increase as longer delays are seen between successive fast pulse initiations. This can be seen in the first five panels of Figure 6a and is shown graphically in Figure 6f, which shows the times between successive right moving fast pulses in the 1D spiral as a function of ϵ. Also, fewer fast pulses were seen numerically at the end of the shooting method. This is likely to be due to the fact that the period of the unstable 1D spiral wave becomes long compared to the time scale at which the solution diverges from the 1D spiral and the R/F-threshold surface.

Interestingly, just below ϵ^*, the activity during the delays between the appearance of

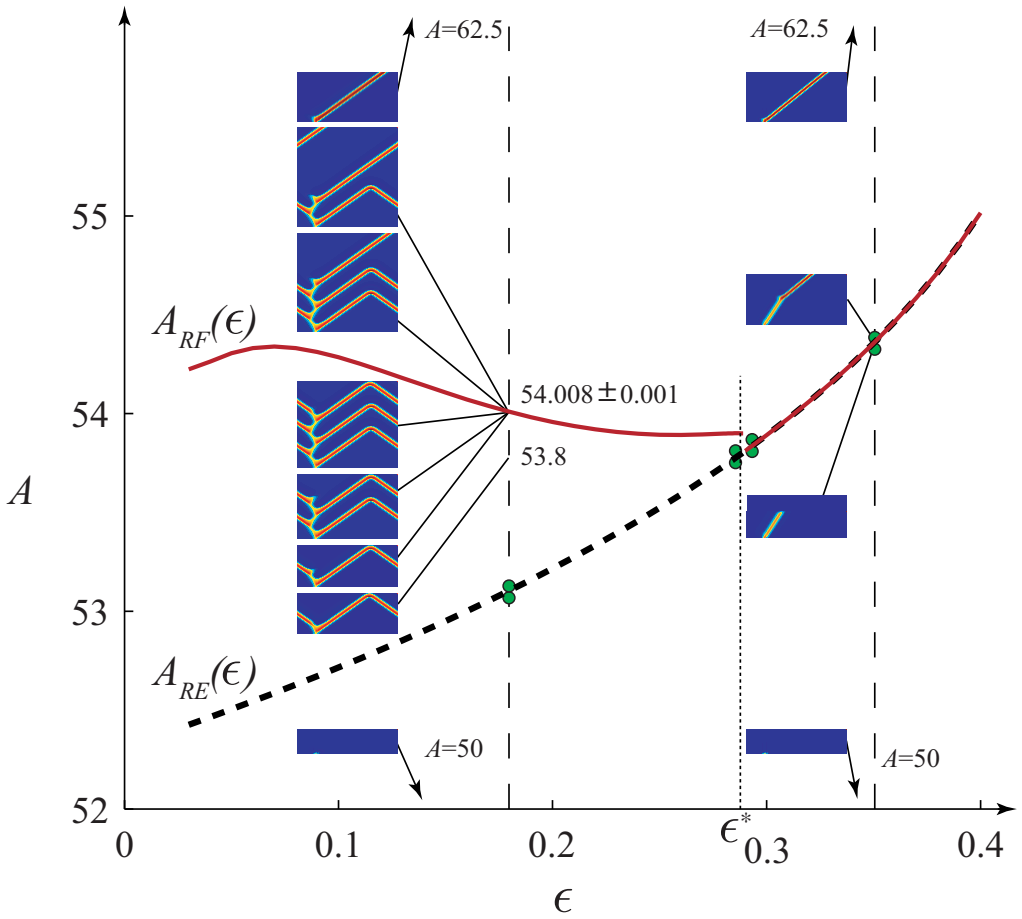


Figure 5. *A parameter/initial condition diagram depicting basins of attraction and their boundaries in the* $\epsilon - A$ *plane. The thick black dashed curve represents initial conditions (ICs) that converge to the slow pulse* ($A_{RE}(\epsilon)$). *The solid red curve represents the boundary between convergence to rest and convergence to the fast pulse* ($A_{RF}(\epsilon)$). *To the accuracy of our calculations, above* ϵ^*, $A_{RE}(\epsilon)$ *and* $A_{RF}(\epsilon)$ *are indistinguishable. Below* ϵ^*, $A_{RE}(\epsilon)$ *and* $A_{RF}(\epsilon)$ *differ, with* $A_{RE}(\epsilon)$ *corresponding to the slow pulse and* $A_{RF}(\epsilon)$ *corresponding to the* 1*D spiral wave. ICs below* $A_{RE}(\epsilon)$ *converge directly to rest. ICs above* $A_{RE}(\epsilon)$ *but below* $A_{RF}(\epsilon)$ *undergo transient excitation before ultimately converging to rest (on a periodic domain). ICs above* $A_{RF}(\epsilon)$ *converge to the fast pulse. Example simulations are shown for* $\epsilon = 0.18$ *and* $\epsilon = 0.35$ *(vertical thin dashed curves). The* $\epsilon = 0.18$ *sequence, also shown in more detail in Figure 4, shows the even and odd sequences described in the text; note that the entire even cascade for* $n > 2$ *exists in a narrow interval near* A_{RF} (54.008 ± 0.001). *The* $\epsilon = 0.35$ *sequence appears in Figure 2 panels* D1–D4 *and is repeated here for context. Simulations from Figure 2 panels* A2–D2 *and* A3–D3 *are represented as green dots.*

successive fast pulses in the 1D spiral resembles the slow pulse. To determine whether the solution is indeed close to the slow pulse, we define the following function that provides a measure of the distance from the slow pulse:

$$d_{sp}(t) = \min_{x_0}(||u_{sp}(x,t) - u(x - x_0, t)||_w + ||v_{sp}(x,t) - v(x - x_0, t)||_w).$$

The "windowed norm" $||\cdot||_w$ used to define d_{sp} is the standard l_2 norm but with the integral

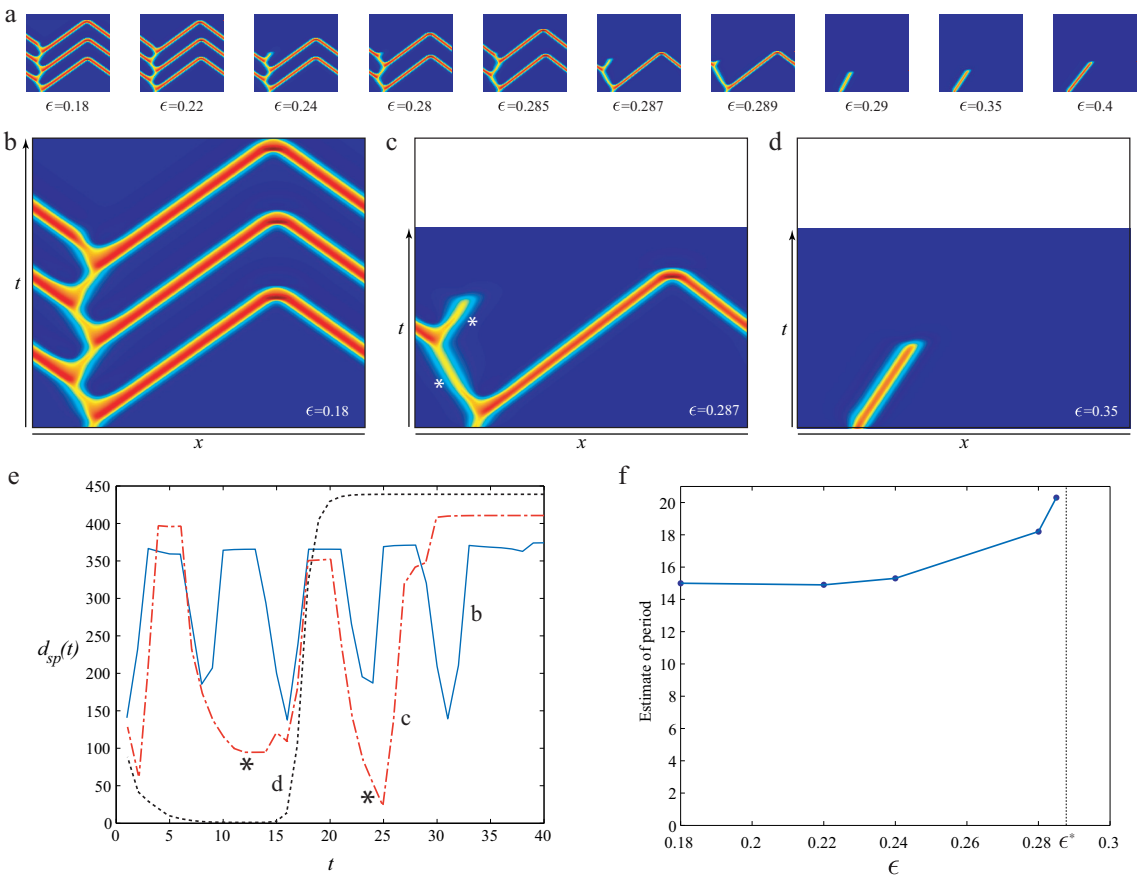


Figure 6. (a) *Simulations corresponding to ICs just below* $A_{RF}(\epsilon)$ *for a selection of* ϵ *values. Magnifications for* $\epsilon = 0.18, 0.287,$ *and* 0.35 *are shown in* (b), (c), *and* (d), *respectively.* (e) *The windowed distance from the slow pulse,* $d_{sp}(t)$, *for each of the highlighted values of* ϵ. *For* $\epsilon = 0.18$ *(b, solid blue), a region of the solution periodically approaches the slow pulse (four distinct minima in* $d_{sp}(t)$*). For* $\epsilon = 0.287$ *(c, dash-dot red), the solution locally approaches the slow pulse to a greater degree and the period is larger (two minima marked with asterisks corresponding to the two apparent slow pulses also marked with asterisks in* (c)*). For* $\epsilon = 0.35 > \epsilon^*$ *(d, dashed black), the approach to the slow pulse is apparently complete and no periodicity is seen.* (f) *Estimated period of the* 1*D spiral as a function of* ϵ. *Data points are obtained from simulations shown in the first five panels of* (a). *Note that the period increases with increasing* ϵ.

restricted to the range of x in which the u component of the slow pulse u_{sp} is more than 10% of the maximum amplitude away from its resting value. That is, the integral is taken in a window around the nonresting portion of the slow pulse. Thus, if $d_{sp}(t)$ is close to zero, then on some interval in x, the solution must be close to the slow pulse even if elsewhere on the domain there are large deviations from the slow pulse. The $d_{sp}(t)$ curve calculated for three values of ϵ are plotted in Figure 6e along with space-time intensity plots of the u component of the solution in 6b, c, and d.

For $\epsilon = 0.18$ (b), the periodic dips in $d_{sp}(t)$ indicate that there is an interval on which the solution periodically approaches the slow pulse, although the approach is incomplete. For $\epsilon = 0.287$ (c), the solution approaches the slow pulse to a greater degree, which leads to

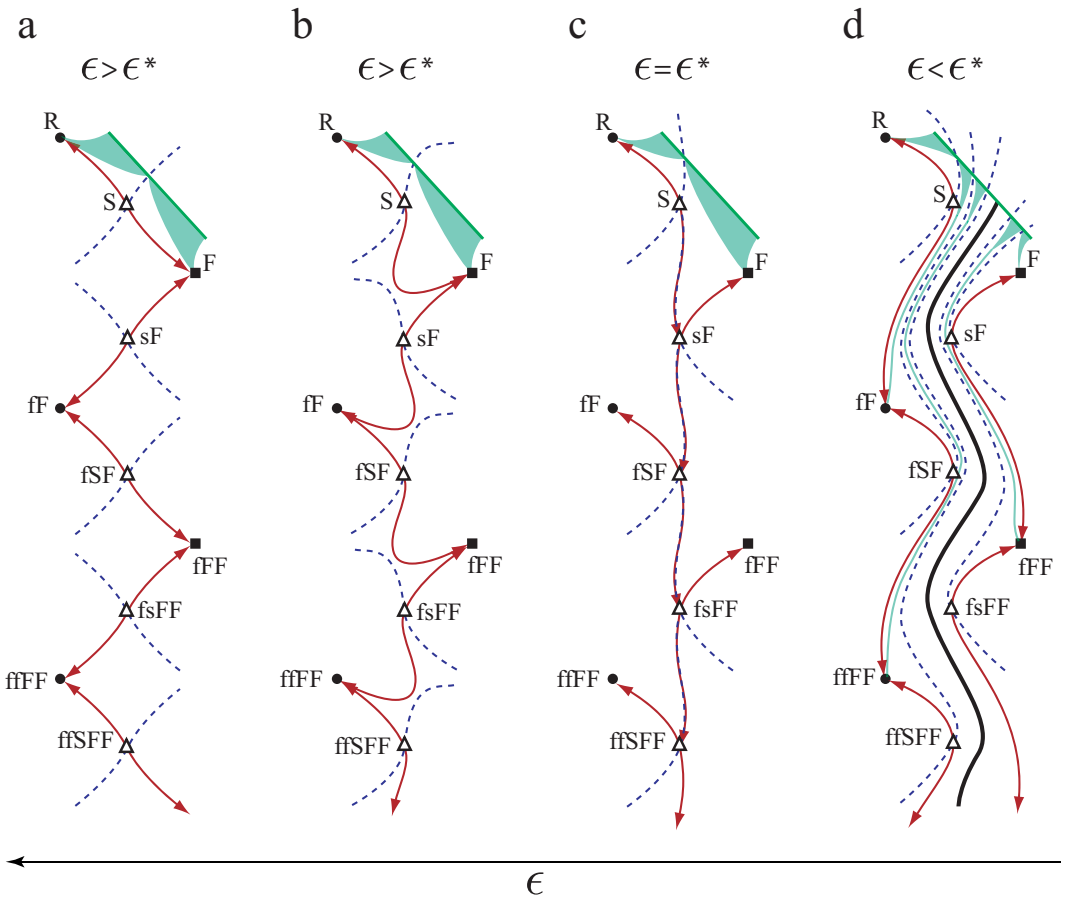


Figure 7. *A sequence of cartoon phase portraits illustrating the proposed heteroclinic bifurcation on the infinite domain. Steady traveling pulse solutions and composites solutions are denoted by circles, triangles, and squares. The top circle is the rest state. Solid shapes denote stable solutions, and empty shapes denote unstable solutions. Red arrows represent the* 1*D unstable manifolds of the slow pulse and composite solutions involving slow pulses. Blue dashed curves represent codimension one stable manifolds. The thick black zig-zag solution in* (d) *converges to the* 1*D spiral wave.* ϵ *decreases from left to right. In* (a) *and* (b)*, the slow pulse is the* R/F*-threshold solution, and its stable manifold is the threshold surface. The dynamics at the proposed bifurcation point are illustrated in* (c)*. For* $\epsilon < \epsilon^*$ (d)*, the* R/F*-threshold consists of the entire region between the stable manifolds of* S *and* sF*. On a finite periodic domain, all solutions to the left of the thick black zig-zag curve, other than the dashed blue curves, converge to rest and the dashed blue curves converge to the slow pulse* S*. All solutions to the right of the thick black zig-zag curve, other than the dashed blue curves, converge to the fast pulse* F*. The long-term behavior of the dashed blue curves to the right of the zig-zag curve (on a periodic domain) remains an open problem.*

fewer periods of the oscillation being revealed before the shooting method terminates at the double-precision limit. At $\epsilon = 0.35$ (d), the approach appears to be "complete," as illustrated by the $d_{sp}(t)$ plot. In other words, at this value of ϵ, the boundary solution is actually the slow pulse rather than a solution that periodically comes close to the slow pulse.

5. A conjectured global bifurcation of heteroclinic orbits. The observations discussed in previous sections and summarized below strongly suggest that the appearance of the 1D

spiral wave at ϵ^* is fundamentally linked to the change in long-term behavior of the slow pulse:

1. The redirection of the unstable manifold of the slow pulse and the appearance of the 1D spiral wave are apparently coincident with regard to the value of ϵ. As ϵ approaches ϵ^*, the initial conditions that lead to the slow pulse ($A_{RE}(\epsilon^*)$) and the 1D spiral wave ($A_{RF}(\epsilon^*)$) appear to approach each other.
2. The local stability of the slow pulse is not altered at ϵ^*. The stable manifold of the slow pulse appears to remain codimension one, as reflected in the fact that our one-parameter shooting method still converges to it. This rules out the possibility that the 1D spiral wave appears through a local bifurcation of the slow pulse.
3. The period of the 1D spiral wave increases as ϵ approaches the critical value ϵ^*.
4. The long-term slow-pulse dynamics appears to progress through the following three stages as ϵ increases: (i) for $\epsilon < \epsilon^*$, $S \to fF$, (ii) for $\epsilon = \epsilon^*$, $S \to sF$, and (iii) for $\epsilon > \epsilon^*$, $S \to F$. As ϵ approaches ϵ^*, the 1D spiral wave is characterized by a core that cycles alternately between $S \to sF$ and $s \to fS$, which is precisely the critical slow-pulse behavior at ϵ^*.

These observations indicate that the 1D spiral wave appears through a global bifurcation that involves a rearrangement of heteroclinic connections of the unstable manifold of the slow pulse.

To describe the nature of the conjectured bifurcation in more detail, it is convenient to consider the infinite domain rather than the periodic domain. For the infinite domain, counterpropagating pulses do not collide and annihilate one another. Instead, the sets of fast counterpropagating pulses (e.g., fF, fFF, etc.) are attracting invariant solutions, each with its own distinct basin of attraction. The boundaries between these basins consist of the stable manifold of composite solutions that include a slow pulse; e.g., the stable manifold of fSF forms the boundary between fF and fFF.

The full conjectured bifurcation is schematically represented in Figure 7. The stable invariant solutions involving only fast pulses and the rest state are represented by filled black circles and squares. The unstable invariant solutions that include a slow pulse are indicated as open triangles. The unstable manifolds of the unstable solutions are red curves, and the dashed blue curves represent the stable manifolds, which act as threshold surfaces. The solid green line near the top of each figure represents the family of rightward pulse-like initial conditions; the green shading represents the corresponding family of trajectories evolving from the initial conditions on the solid green line. Note that the diagram shows only activity that begins with a rightward moving pulse. There is a corresponding set of solutions that begin with leftward moving pulses and solutions involving multiple slow pulses, but these are not shown in the diagram. Also note that, although this discussion focuses on the case of an infinite domain, a similar diagram would hold for the periodic domain with the modification that any solution converging to an even number of pulses (black circles) in this diagram would eventually converge to the rest solution R on a periodic domain. Similarly, the basins of attraction of odd numbers of pulses (black squares) would eventually converge to a single fast pulse F.

For ϵ near the saddle-node bifurcation at ϵ_{SNB} (Figure 7a), all initial conditions on the solid green line evolve to either the rest state R or the single fast pulse F. The stable manifold of the slow pulse acts as the R/F-threshold surface. Although the invariant composite solu-

tions exist, the family of rightward pulse-like initial conditions does not fall into their basins of attraction.

As ϵ gets closer to ϵ^* (Figure 7b), the unstable manifold of S and the stable manifold of Fs (and other pairs of sequentially adjacent unstable solutions) deform and approach one another. This is reflected in Figure 2, panel C2, in which transient retrograde activity appears but quickly dies out. However, the solution space is qualitatively unchanged from Figure 7a.

At ϵ^* (Figure 7c), these manifolds coalesce and form an infinite chain of heteroclinic connections, $S \to sF \to fSF \to fsFF \to \cdots$; i.e., each slow pulse blows up into an anterograde fast pulse and a retrograde slow pulse.

Below ϵ^* (Figure 7d), the heteroclinic connections have completed their rearrangement, creating fundamental changes in the basins of attractions of the stable solutions. The unstable manifold of the slow pulse S now connects to the solution fF, two counterpropagating pulses, and consequently, the stable manifold acts as the threshold between the rest state R and fF. Similarly, the stable manifolds of all composite solutions with a slow pulse in the middle (e.g., fSF) now act as the threshold surfaces for corresponding stable solutions with approximate slow pulses dying out (e.g., fF) or blowing up into two counterpropagating fast pulses (e.g., $ffFF$). Consequently, the stable manifolds of these unstable invariant solutions are interleaved and divide the family of initial conditions into segments that evolve to all of the different sequences of pulses. Due to the topological constraints of the infinite set of heteroclinic crossings (Figure 7c), a solution consisting of the formation of an infinite number of pulses must exist through the middle of this region (thick black curve in Figure 7d). This solution converges to the 1D spiral. In a more general context, pulse-generating solutions of this type are called "sources" [40].

It is possible that these heteroclinic connections do not form simultaneously at a single value of ϵ but instead occur one by one, most likely from the top down or from the bottom up. The latter case might provide an explanation for the ambiguous interval near ϵ^* mentioned earlier and discussed in the appendix. Provided all connections eventually occur for some positive ϵ, this does not cause fundamental problems for the bifurcation scenario proposed here but simply spreads the transition across an interval of ϵ values. Due to the minimal interaction between successive pulses along the domain (particularly between a central slow pulse and outward propagating fast pulses), the appearances of these heteroclinic connections are expected to be concentrated in a fairly narrow interval around ϵ^*; recall that we defined ϵ^* to be the value of ϵ at which a heteroclinic connection between S and sF exists. If this is true, only a minimal change in Figure 7 would be necessary.

For the case of a finite periodic domain, the preceding description must be modified in light of the fact that pairs of fast pulses eventually collide with each other on the far side of the domain, thereby forming a pulse "sink" [40]. Using the f, s, S, F notation, one analogue of the 1D spiral on a periodic domain, for ϵ close to ϵ^*, would cycle through the *temporal* sequence $S, sF, fSF, S, \ldots$, where the first two transitions correspond to movement close to the unstable manifold of S and sF, respectively, while the third corresponds to collision of the f and F pulses. This would be the case on a relatively small domain. On a larger periodic domain, the activity associated with the 1D spiral becomes $f^nSF^n, f^nsF^{n+1}, f^{n+1}SF^{n+1}, f^nSF^n, \ldots$.

6. Implications of the existence of the 1D spiral wave: Reflection and annihilation. Using a set of numerical experiments, we have explored the threshold between the rest state and the traveling fast pulse in a FHN-type model. We find that there is a critical value ϵ^* that separates different types of threshold behavior. For $\epsilon > \epsilon^*$ (lower excitability), the stable manifold of the slow pulse forms the R/E- and the R/F-threshold surfaces. For $\epsilon < \epsilon^*$ (higher excitability), the slow pulse still acts as the R/E-threshold solution. However, between the basins of attraction of the rest state and the single fast pulse, there is a series of regions of initial conditions that lead to solutions consisting of sequences of counterpropagating pulses [9, 28]. At the heart of this series of regions, there is an unstable periodic solution consisting of an infinite sequence of counterpropagating pulses, i.e., the 1D spiral wave [26, 9]. On a periodic domain, the multiple-pulse behavior is transient, and the 1D spiral wave acts as the threshold solution between the rest state and a circulating fast pulse, i.e., the R/F-threshold solution.

In our numerical experiments, initial conditions were chosen in a physically unrealistic manner. However, the same dynamics described in this manuscript can arise in physically realistic situations. Structural heterogeneities such as an abrupt increase in cable diameter [9] or a spatially localized decrease in excitability [28] have been shown to uncover the 1D spiral wave. Similarly, the behavior can be uncovered by transient "functional" heterogeneities. Below, we demonstrate this by examining the effects of (i) transient localized increases in refractoriness in front of a fast pulse and (ii) transient localized stimulation in the wake of a fast pulse. The underlying idea is simply that considering a one-parameter family of either stimulus amplitudes or amplitudes of refractory bumps is analogous to considering a one-parameter family of initial conditions as studied in previous sections.

Consider a pulse traveling toward a spatially localized region with an elevated refractory level, e.g., resulting from localized "ectopic" activity. In simulations, we have implemented this region by initially setting the recovery variable, v, to have an approximately Gaussian profile with amplitude B and fixed width σ (see Figure 8a). As the pulse approaches the refractory region, the region begins to recover. If the pulse runs into the region when it has significantly recovered, then the pulse successfully propagates through the region (Figure 8b). If the pulse runs into the region when refractoriness is still high, the pulse dies out and the system goes to rest (Figure 8e). When $\epsilon > \epsilon^*$, there is a sharp boundary in the refractory level set by B between these two behaviors (not shown). However, when $\epsilon < \epsilon^*$ and the 1D spiral wave exists, as the level of refractoriness is altered by changing B, the system exhibits all of the behavior associated with the existence of the 1D spiral (see sequence (2) and Figure 8c). The most robust of this behavior is the successful propagation of the pulse through the refractory region and the generation of a single retrograde or "reflected" pulse (Ff) (Figure 8d). This type of reflected activity has been seen in numerical simulations of models of excitable media [4, 16, 19, 48] and physiological experiments in both cardiac [1, 20, 39] and neural tissue [2, 36], and it has been suggested to underlie the induction of life-threatening cardiac arrhythmias. Analogous behavior in two and three spatial dimensions has been linked to the onset of spiral and scroll waves [28].

As a second example of how the behavior can arise in a physically realistic setting, we consider the effects of a spatially and temporally localized stimulus delivered in the wake of a fast pulse circulating on a periodic domain. This scenario has been studied in the context of

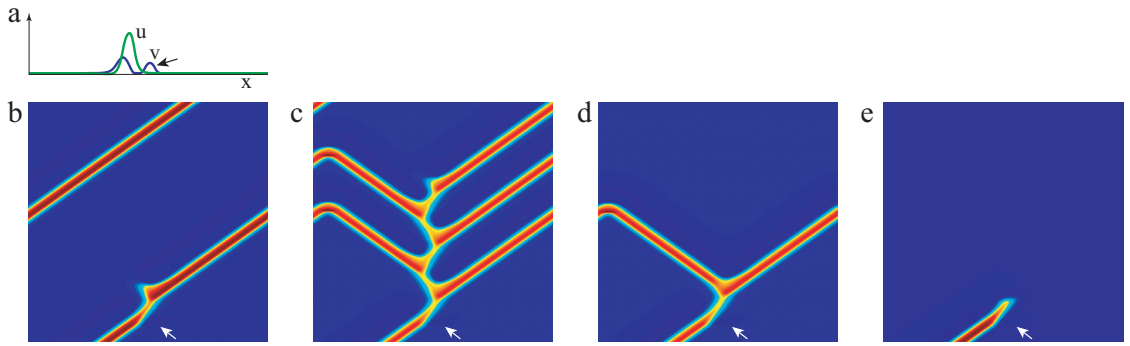


Figure 8. *The effects of propagation through a transient refractory region. A refractory bump, in the form of the addition of a Gaussian* ($B\exp(-(x-x_0)^2/\sigma^2)$) *to the recovery variable* v*, is placed ahead of the fast pulse. This is illustrated in* (a) *(black arrow) and visible as a slightly darker blue region just in front of the fast pulse in all four images (white arrow in* (b)–(e)*).* (b)–(e) *show the effect of changing the amplitude* B *with* x_0 *fixed and* $\sigma = 0.01$ *and* $\epsilon = 0.18 < \epsilon^*$*.* (b) *For sufficiently small* B *(*$B = 0.1838$*), the fast pulse successfully crosses the refractory region.* (c) *By using a shooting method on* B*, all of the transient behavior associated with the existence of the* 1*D spiral can be observed (*$B = 0.183837479709384785109271 8$*).* (d) *For* $B = 0.1839$*, the fast pulse successfully crosses the refractory region with a delay, and a retrograde fast pulse is produced.* (e) *For sufficiently large* B *(*$B = 0.1840$*), the fast pulse hits the refractory region and dies out.*

both the termination [15] and the onset [46, 42] of cardiac reentrant arrhythmias. We take the stimulus to be a square wave with amplitude C and fixed small spatial and temporal widths (see Figure 9a). When an appropriately timed stimulus is delivered in the wake of the pulse, a stimulus of sufficiently large amplitude induces a single retrograde pulse (Figure 9e) [15]. The forward-propagating pulse and the new retrograde pulse eventually collide and all activity is annihilated. On the other hand, when the stimulus has a sufficiently small amplitude, the rotating fast pulse is phase-reset but survives (Figure 9b). As before, when $\epsilon > \epsilon^*$, a sharp threshold is found between the two behaviors as the amplitude C is increased from zero (not shown), whereas, when $\epsilon < \epsilon^*$, the system exhibits all of the behavior associated with the existence of the 1D spiral (see sequence (2) and Figure 9c). Thus, stimuli can elicit new suprathreshold activity without necessarily annihilating all activity. Only when an odd number of new traveling pulses is generated is the activity annihilated; this occurs when stimulus amplitudes are greater than the one that produces the 1D spiral wave (see Figure 9d).

In all of the simulations presented in this manuscript, we use the ML model with all parameters fixed except ϵ (see Table 1). However, we found the same qualitative behavior for the ML model with other parameter sets and for the standard cubic FHN model. Therefore, we expect that the behavior described here is generic for FHN-type models. More work is needed to uncover the details of how parameters affect ϵ^* and the existence of the 1D spiral wave. Furthermore, because of the link between the existence of 1D spiral waves and the induction of cardiac arrhythmias, it is crucial to identify conditions for the existence of a 1D spiral wave and the robustness of its associated transient behavior in physiologically realistic models (for example, [3, 29]).

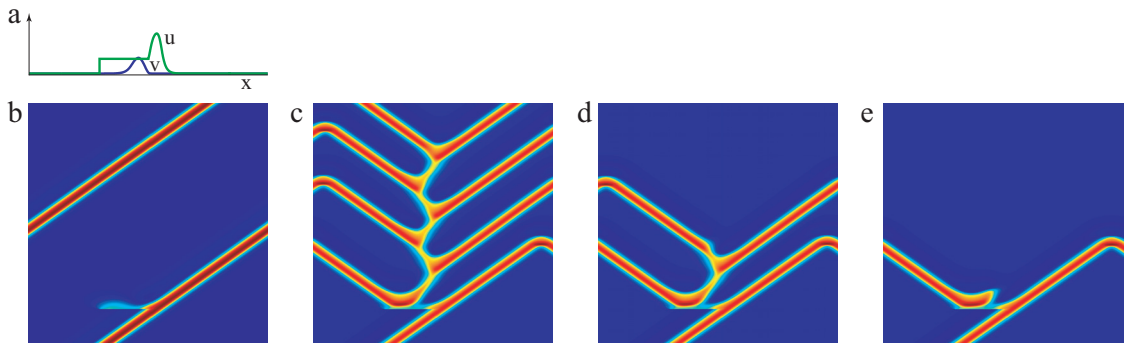


Figure 9. *The effects of stimuli in the wake of a fast pulse rotating on a* 1*D excitable medium. A square-wave stimulus is delivered in the wake of the fast pulse. The stimulus has amplitude* C*, temporal width* $\Delta t = 0.03$*, and spatial width* $\Delta x = 0.2$*, and* $\epsilon = 0.18$ *is in the sub-*ϵ^* *range.* (b) *For sufficiently small* C *(*$C = 1200$*), the stimulus has only a subthreshold effect.* (c) *By adjusting* C*, all of the transient behavior associated with the existence of the* 1*D spiral can be observed (image corresponds to* $C = 1369.42006681669181$*).* (d) *For* $C = 1369.421$*, the stimulus elicits a retrograde fast pulse, an anterograde fast pulse, and subsequently another retrograde fast pulse with the final result of convergence to rest.* (e) *For sufficiently large* C *(*$C = 1406.2$*), the stimulus elicits only a new retrograde fast pulse, eventually eliminating all activity.*

Appendix. Morris–Lecar model and parameters.

The Morris–Lecar (ML) system was first introduced as a physiologically detailed model for barnacle muscle [32]. However, as a model with only two variables, it provides a useful tool for numerical studies of excitable media in general. For such generic studies, it is more common to use the simpler functional forms associated with the proper FHN system which provide analytical tractability. However, the study presented here is a numerical exploration of solutions that lie on boundary of the basins of attraction of two stable solutions and is therefore concerned with solutions that have a saddle structure to them (unstable and stable manifolds). The optimal context for such a study is one in which the positive eigenvalue of the saddle solution is as small as possible, thereby ensuring that it can be approached numerically using a shooting technique. Because the saddle solutions (i.e., slow-pulse and 1D spiral) necessarily traverse the central region of the phase plane, we conjecture that a parameter range in which this central region has slow dynamics is a convenient one for this study in the sense that the slow pulse is easier to find numerically. In fact, it is unclear whether this conjecture holds up to scrutiny, but we use ML dynamics throughout, nonetheless, due to the parametric flexibility it provides in terms of the central region of the phase plane. A more careful study of this conjecture might prove useful in overcoming certain numerical limitations encountered.

The ML system is defined by the following choice of f and g:

$$f(u,v) = -g_{Ca} m_\infty(u)(u - E_{Ca}) - g_K v(u - E_K) - g_l(u - E_l),$$
$$g(u,v) = \frac{v_\infty(u) - v}{\tau(u)},$$

where $m_\infty(u) = 0.5(1+\tanh(\frac{u-u_1}{u_2}))$, $v_\infty(u) = 0.5(1+\tanh(\frac{u-u_3}{u_4}))$, and $\tau(u) = 1/\cosh(\frac{u-u_3}{2u_4})$. Parameter values are given in Table 1. For the spatial domain, we consider primarily the

case of periodic boundary conditions. The domain size is taken to be unity, and we choose a nondimensional diffusion coefficient sufficiently small to ensure that the domain is not a limiting factor in the propagation of excitation.

Table 1

Table of parameters used in simulations.

Name	Value		Name	Value	
g_l	2	leak conductance	D	0.001	diffusion coefficient
g_K	8.0	K^+ conductance	x_1	0.16	slow pulse u IC parameter
g_{Ca}	4.4	Ca^2+ conductance	x_2	0.2	slow pulse u IC parameter
u_1	-1.2	$m_\infty(u)$ parameter	x_3	0.13	slow pulse v IC parameter
u_2	18	$m_\infty(u)$ parameter	x_4	0.17	slow pulse v IC parameter
u_3	2	$v_\infty(u), \tau(u)$ parameter	x_0	0.4	stimulus center
u_4	10	$v_\infty(u), \tau(u)$ parameter	$2w$	0.2	stimulus width
E_l	-60	leak reversal potential	t_{stim}	0.03	stimulus duration
E_K	-84	K^+ reversal potential	τ	0.3	convergence measure delay
E_{Ca}	120	Ca^2+ reversal potential	dx	0.001	numerical space step
I	10	constant current	dt	0.03	numerical time step

A.1. The ambiguous interval near ϵ^*. It is worth noting that the R/F threshold appears to have a discontinuity at ϵ^* when represented in the $\epsilon - A$ plane (see red curve in Figure 5). One explanation for this phenomenon is that the red curve should actually contain an S-shaped portion connecting either side of the discontinuity. This would indicate the existence of a small interval of ϵ between 0.287 and ~ 0.29 in which the family of initial conditions crosses the R/F threshold more than once. The initial conditions at two of these crossings would converge to a 1D spiral, and the initial condition at the third crossing would converge to the slow pulse. If this is true, there is a narrow interval in ϵ in which both the slow pulse and the 1D spiral coexist on the codimension one separatrix between basins of attractions. On the other hand, the apparent discontinuity in the R/F threshold could also be an artifact of the numerical shooting approach used here. A complete understanding of this region requires a more comprehensive numerical continuation approach.

Acknowledgments. The authors would like to thank J. P. Keener, J. Rademacher, J. Rubin, and L. Glass for insight-building discussions regarding various aspects of the results and ideas presented here.

REFERENCES

[1] J. Antzelevitch, C. Jalife, and G. K. Moe, *Characteristics of reflection as a mechanism of reentrant arrhythmias and its relationship to parasystole*, Circulation, 61 (1980), pp. 182–191.

[2] S. A. Baccus, *Synaptic facilitation by reflected action potentials: Enhancement of transmission when nerve impulses reverse direction at axon branch points*, Proc. Natl. Acad. Sci. USA, 95 (1998), pp. 8345–8350.

[3] G. W. Beeler and H. J. Reuter, *Reconstruction of the action potential of ventricular myocardial fibers*, J. Phys., 268 (1977), pp. 177–210.

[4] C. Cabo and R. C. Barr, *Reflection after delayed excitation in a computer model of a single fiber*, Circ. Res., 71 (1992), pp. 260–270.

[5] G. A. Carpenter, *A geometric approach to singular perturbation problems with applications to nerve impulse equations*, J. Differential Equations, 23 (1977), pp. 335–367.
[6] R. G. Casten, H. Cohen, and P. A. Lagerstrom, *Perturbation analysis of an approximation to the Hodgkin-Huxley theory*, Quart. Appl. Math., 32 (1974), pp. 365–402.
[7] E. Cytrynbaum and J. P. Keener, *Stability conditions for the traveling pulse: Modifying the restitution hypothesis*, Chaos, 12 (2002), pp. 788–799.
[8] G. B. Ermentrout and J. Rinzel, *One-dimensional $\lambda - \omega$ target patterns: Empirical stability tests*, J. Math. Biol., 10 (1980), pp. 97–100.
[9] G. B. Ermentrout and J. Rinzel, *Reflected waves in an inhomogeneous excitable medium*, SIAM J. Appl. Math., 56 (1996), pp. 1107–1128.
[10] P. C. Fife and J. B. McLeod, *The approach of solutions of nonlinear diffusion equations to travelling front solutions*, Arch. Ration. Mech. Anal., 65 (1977), pp. 335–361.
[11] R. FitzHugh, *Impulses and physiological states in theoretical models of nerve membrane*, Biophys. J., 1 (1961), pp. 445–466.
[12] R. FitzHugh, *Mathematical models of excitation and propagation in nerve*, in Biological Engineering, H. P. Schwan, ed., McGraw–Hill, New York, 1969, pp. 1–85.
[13] G. Flores, *The stable manifold of the standing wave of the Nagumo equation*, J. Differential Equations, 80 (1989), pp. 306–314.
[14] G. Flores, *Stability analysis for the slow traveling pulse of the FitzHugh–Nagumo system*, SIAM J. Math. Anal., 22 (1991), pp. 392–399.
[15] L. Glass and M. E. Josephson, *Resetting and annihilation of reentrant abnormally rapid heartbeat*, Phys. Rev. Lett., 75 (1995), pp. 2059–2062.
[16] S. S. Goldstein and W. Rall, *Changes of action potential shape and velocity for changing core conductor geometry*, Biophys. J., 14 (1974), pp. 731–757.
[17] S. P. Hastings, *On the existence of homoclinic and periodic orbits for the FitzHugh-Nagumo equations*, Quart. J. Math. Oxford Ser. (2), 27 (1976), pp. 123–134.
[18] A. L. Hodgkin and A. F. Huxley, *A quantitative description of membrane current and its application to conduction and excitation in nerve*, J. Phys., 117 (1952), pp. 500–544.
[19] J. F. Howe, W. H. Calvin, and J. D. Loeser, *Impulses reflected from dorsal root ganglia and from focal nerve injuries*, Brain Research, 116 (1976), pp. 139–144.
[20] J. Jalife and G. K. Moe, *Excitation, conduction, and reflection of impulses in isolated bovine and canine Purkinje fibers*, Circ. Res., 49 (1981), pp. 233–247.
[21] C. K. R. T. Jones, *Stability of the traveling wave solution of the FitzHugh-Nagumo system*, Trans. Amer. Math. Soc., 286 (1984), pp. 431–469.
[22] J. Keener and J. Sneyd, *Mathematical Physiology*, Interdiscip. Appl. Math. 8, Springer-Verlag, New York, 1998.
[23] J. P. Keener, *The topology of defibrillation*, J. Theoret. Biol., 230 (2004), pp. 459–473.
[24] J. P. Keener and E. Cytrynbaum, *The effect of spatial scale of resistive inhomogeneity on defibrillation of cardiac tissue*, J. Theoret. Biol., 223 (2003), pp. 233–248.
[25] J. P. Keener and T. J. Lewis, *The biphasic mystery: Why a biphasic shock is more effective than a monophasic shock for defibrillation*, J. Theoret. Biol., 200 (1999), pp. 1–17.
[26] N. Kopell and L. N. Howard, *Target patterns and horseshoes from a perturbed central-force problem: Some temporally periodic solutions to reaction-diffusion equations*, Stud. Appl. Math., 64 (1981), pp. 1–56.
[27] T. Krogh-Madsen and D. J. Christini, *Resetting and termination of reentry in a loop-and-tail cardiac model*, Phys. Rev. E (3), 77 (2008), 011916.
[28] T. J. Lewis, *The Effects of Nonexcitable Regions on Signal Propagation in Excitable Media: Propagation Failure and Reflection*, Ph.D. thesis, University of Utah, Salt Lake City, UT, 1998.
[29] C. H. Luo and Y. Rudy, *A dynamic model of the cardiac ventricular action potential. I. Simulations of ionic currents and concentration changes*, Circ. Res., 74 (1994), pp. 1071–1096.
[30] H. P. McKean, *Nagumo's equation*, Adv. Math., 4 (1970), pp. 209–223.
[31] V. Moll and S. I. Rosencrans, *Calculation of the threshold surface for nerve equations*, SIAM J. Appl. Math., 50 (1990), pp. 1419–1441.

[32] C. Morris and H. Lecar, *Voltage oscillations in the barnacle giant muscle fiber*, Biophys. J., 35 (1981), pp. 193–213.
[33] J. S. Nagumo, S. Arimoto, and S. Yoshizawa, *An active pulse transmission line simulating nerve axon*, in Proceedings of the Institute of Radio Engineers, 50 (1962), pp. 2061–2071.
[34] J. C. Neu, R. S. Preissig, and W. Krassowska, *Initiation of propagation in a one-dimensional excitable medium*, Phys. D, 102 (1997), pp. 285–299.
[35] A. M. Pertsov, E. A. Ermakova, and A. V. Panfilov, *Rotating spiral waves in a modified FitzHugh-Nagumo model*, Phys. D, 14 (1984), pp. 117–124.
[36] F. Ramon, R. W. Joyner, and J. W. Moore, *Propagation of action potentials in inhomogeneous axon regions*, Federation Proc., 34 (1975), pp. 1357–1363.
[37] J. Rinzel, *Excitation dynamics: Insights from simplified membrane models*, Federation Proc., 44 (1985), pp. 2944–2946.
[38] J. Rinzel and J. B. Keller, *Traveling wave solutions of a nerve conduction equation*, J. Biophys., 13 (1973), pp. 1313–1337.
[39] G. J. Rozanski, J. Jalife, and G. K. Moe, *Reflected reentry in nonhomogeneous ventricular muscle as a mechanism of cardiac arrhythmias*, Circulation, 69 (1984), pp. 163–173.
[40] B. Sandstede and A. Scheel, *Defects in oscillatory media: Toward a classification*, SIAM J. Appl. Dyn. Syst., 3 (2004), pp. 1–68.
[41] S. Sinha and D. J. Christini, *Termination of reentry in an inhomogeneous ring of model cardiac cells*, Phys. Rev. E (3), 66 (2002), 061903.
[42] C. F. Starmer, V. N. Biktashev, D. N. Romashko, M. R. Stepanov, O. N. Makarova, V. I. Krinsky, M. R. Stepanov, O. N. Makarova, and V. I. Krinsky, *Vulnerability in homogeneous excitable media: Analytical and numerical studies of unidirectional propagation*, J. Biophys., 65 (1993), pp. 1775–1787.
[43] N. Trayanova, R. Gray, D. Bourn, and J. Eason, *Virtual electrode induced positive and negative graded response: New insights into fibrillation induction and defibrillation*, J. Cardiovasc. Electrophys., 14 (2003), pp. 756–763.
[44] J. J. Tyson and J. P. Keener, *Singular perturbation theory of traveling waves in excitable media (a review)*, Phys. D, 32 (1988), pp. 327–361.
[45] A. T. Winfree, *When Time Breaks Down: The Three-Dimensional Dynamics of Electrochemical Waves and Cardiac Arrhythmias*, Princeton University Press, Princeton, NJ, 1987.
[46] A. T. Winfree, *Electrical instability in cardiac muscle: Phase singularities and rotors*, J. Theoret. Biol., 138 (1989), pp. 353–405.
[47] A. Xu and M. R. Guevara, *Two forms of spiral-wave reentry in an ionic model of ischemic ventricular myocardium*, Chaos, 8 (1998), pp. 157–174.
[48] Y. Zhou and J. Bell, *Study of propagation along nonuniform excitable fibers*, Math. Biosci., 119 (1994), pp. 169–203.

SIAM J. Applied Dynamical Systems
Vol. 8, No. 1, pp. 371–389
© 2009 Society for Industrial and Applied Mathematics

A Poincaré Section for the General Heavy Rigid Body*

Sven Schmidt†, Holger R. Dullin‡, and Peter H. Richter§

Abstract. A general procedure is developed for the study of rigid body dynamics using Poincaré sections. A section condition is chosen which captures every trajectory on a given energy surface. This section condition is given by extremal heights of the center of mass of the top. The possible topological types of the corresponding sections are orientable surfaces of genus up to 4 and unions thereof. The flow is not transverse to the section, and the set where tangency occurs is not invariant under the flow. In the standard construction of Poincaré sections the section is cut along the tangency set. In the present case this would lead to a complicated presentation of the section. Instead the section is cut into two pieces in such a way that each piece can be projected 1:1 onto a torus. This torus is obtained from the twice-punctured Poisson sphere by adding two circles. Examples of Poincaré maps illustrating the different topologies of the section are presented.

Key words. heavy rigid body, Poincaré section, Euler–Poisson equations, bifurcation diagram

AMS subject classifications. 70E17, 70K99, 70H07, 65P20, 65P30

DOI. 10.1137/080719029

1. Introduction.

The dynamics of a rigid body in a constant field of gravity, with one point held fixed, is in general nonintegrable. The configuration space SO(3) has three degrees of freedom, but except for the special cases of Euler, Lagrange, and Kovalevskaya, the system possesses only two integrals of motion: the energy $H = h$ and the angular momentum component in the direction of gravity, $L_z = l$. Liouville integrability would require a third integral, but this integral does not exist. The parameter space is essentially 4-dimensional (4D): after scaling distance, time, and energy there remain two free parameters for the principal moments of inertia $A = \mathrm{diag}(A_1, A_2, A_3)$ and two for the location of the body's center of mass $\mathbf{r} = (r_1, r_2, r_3)$ relative to the fixed point. Within this 4D set of parameters, Euler's case $\mathbf{r} = 0$ defines a 2-dimensional (2D) subset (two free parameters in the moments of inertia); Lagrange's case defines a 1-dimensional subset (one free ratio of the moments of inertia and no freedom for the direction of $\mathbf{r}$); and Kovalevskaya's case is a single point (no freedom in the moments of inertia or in the direction of $\mathbf{r}$, provided the body-fixed frame of reference is properly chosen). Since these are the only three integrable cases, the vast majority of rigid bodies exhibit some degree of chaotic motion which has remained largely unexplored to this very day.

*Received by the editors March 21, 2008; accepted for publication (in revised form) by J. Meiss November 24, 2008; published electronically February 19, 2009.
http://www.siam.org/journals/siads/8-1/71902.html
†Department of Mathematical Sciences, Loughborough University, LE11 3TU Leicestershire, UK, and Institut für Theoretische Physik, Universität Bremen, D-28334 Bremen, Germany (svenschmidt75@googlemail.com).
‡School of Mathematics and Statistics, University of Sydney, NSW 2006, Australia (hdullin@usyd.edu.au).
§Institut für Theoretische Physik, Universität Bremen, D-28334 Bremen, Germany (prichter@physik.uni-bremen.de).

The three degrees of freedom of configuration space SO(3) may be reduced to effectively two in view of the rotational $\mathbb{S}^1$-symmetry with respect to the axis of gravity. The configuration space thereby reduces to the Poisson sphere $\mathbb{S}^2$, and the corresponding 4D phase spaces $\mathrm{T}^*_l\mathbb{S}^2$ are labeled with the angular momentum constant l. The surfaces $\mathcal{E}^3_{h,l}$ of constant energy $H = h$ are manifolds in $\mathrm{T}^*_l\mathbb{S}^2$, except at values (h, l), where their topological character changes. The first step in the analysis of the phase space structure is the identification of the topology of $\mathcal{E}^3_{h,l}$ and its bifurcation scheme. This requires the study of the energy-momentum map from phase space to the (h, l)-plane and the determination of its critical values. Already this first step is more difficult than one might think. It has been worked out for certain subsets of parameters [7, 10] but not for the entire 4D family of rigid bodies. However, even though the bifurcation schemes have not been resolved for all cases, it is known that the connected components of $\mathcal{E}^3_{h,l}$ come only in four types [11]: sphere $\mathbb{S}^3$, direct product $\mathbb{S}^1 \times \mathbb{S}^2$, real projective space $\mathbb{R}\mathrm{P}^3$, or connected sum $(\mathbb{S}^1 \times \mathbb{S}^2)\#(\mathbb{S}^1 \times \mathbb{S}^2)$.

Once the energy surfaces $\mathcal{E}^3_{h,l}$ have been identified, the problem is to find how they are partitioned into regular and irregular types of motion, i.e., into invariant subsets of one, two, or three dimensions (isolated periodic orbits, invariant tori, and chaotic regions, respectively). The most valuable tool for this kind of study is the method of Poincaré sections. It requires the identification of a 2D surface of section $\mathcal{P}^2_{h,l} \subset \mathcal{E}^3_{h,l}$ which intersects all possible trajectories repeatedly. Finding such a global Poincaré section is a nontrivial matter. It is in general not possible to choose a surface which intersects all trajectories *transversally* [2]. However, it is possible to find a surface (or a set of disjoint surfaces) that is *complete* in the sense that every orbit intersects—or at least touches—it repeatedly. A constructive procedure to obtain such section conditions was given in [5] and will be employed here.

The next question refers to the topology of the sections and how the various types of $\mathcal{P}^2_{h,l}$ may be adequately represented in 2D plots. This is the main concern of the present paper. We show that a single connected component of $\mathcal{P}^2_{h,l}$ may be a sphere $\mathbb{S}^2$, a torus T^2, or a 2D manifold M^2_g of genus $g = 2$, 3, or 4. A single type of graphical representation applies to all cases, namely, the 1:1 projection to a torus constructed from two copies of a twice-punctured Poisson sphere. We call this the "PP-torus" T^2_2 and propose it as a convenient universal tool for investigations of the complex dynamics of rigid bodies. For example, in order to obtain a complete survey of the phase space structure for a given rigid body it may be interesting to vary (h, l) for fixed parameters (A_1, A_2, A_3) and (r_1, r_2, r_3). For the integrable Kovalevskaya top such a survey was presented in [9], but the new tool can be applied to any choice of parameters. Alternatively, one might want to vary parameters at fixed (h, l) and follow the fate of certain conspicuous features (such as isolated periodic orbits or major chaotic regions).

The paper is organized as follows. Section 2 summarizes how the topology of energy surfaces is determined from the effective potential on the Poisson sphere. The new Poincaré section is introduced in section 3 and compared to a proposal made earlier in [6]. Finally, section 4 describes the topology of the sections $\mathcal{P}^2_{h,l}$ and their projection first to the Poisson sphere and then to the PP-torus constructed from it.

2. Topology of energy surfaces. The heavy rigid body on $\mathrm{T}^*\mathrm{SO}(3)$ is symmetric with respect to rotation about the axis of gravity. Reduction by this symmetry gives the Euler–Poisson equations for the motion of a rigid body about a fixed point as seen in a comoving

frame,

$$\begin{aligned} \dot{\boldsymbol{\gamma}} &= \{\boldsymbol{\gamma}, H\} = \boldsymbol{\gamma} \times \boldsymbol{\omega}, \\ \dot{\mathbf{l}} &= \{\mathbf{l}, H\} = \mathbf{l} \times \boldsymbol{\omega} - mg\boldsymbol{\gamma} \times \mathbf{r}. \end{aligned} \tag{2.1}$$

Here $\boldsymbol{\gamma}$ is the unit vector along the spatial z-axis (the axis of gravity), $\boldsymbol{l}$ is the angular momentum vector, $\boldsymbol{\omega} = A^{-1}\boldsymbol{l}$ is the angular velocity vector, $A = \operatorname{diag}(A_1, A_2, A_3)$ is the matrix of principal moments of inertia, and $\mathbf{r}$ is the position of the center of mass in the body. The connection to Euler's angles is given by $(\gamma_1, \gamma_2, \gamma_3) = (\sin\psi \sin\vartheta, \cos\psi \sin\vartheta, \cos\vartheta)$; the angle φ of rotation about the z-axis is eliminated from the equations of motion by the symmetry reduction. The remaining angles (ϑ, ψ) parameterize the Poisson sphere $\mathbb{S}^2$. The constant mg can be absorbed into $\mathbf{r}$ and will be ignored from now on. These equations on the reduced phase space are Hamiltonian with respect to the Poisson bracket

$$\{F, G\} = \langle \nabla_{\boldsymbol{\gamma}} F, \boldsymbol{\gamma} \times \nabla_{\mathbf{l}} G\rangle + \langle \nabla_{\mathbf{l}} F, \boldsymbol{\gamma} \times \nabla_{\boldsymbol{\gamma}} G + \mathbf{l} \times \nabla_{\mathbf{l}} G\rangle \tag{2.2}$$

and Hamiltonian

$$H(\boldsymbol{\gamma}, \boldsymbol{l}) = \frac{1}{2}\langle \boldsymbol{l}, A^{-1}\boldsymbol{l}\rangle - \langle \boldsymbol{\gamma}, \mathbf{r}\rangle, \tag{2.3}$$

where $\langle \cdot, \cdot \rangle$ denotes the standard Euclidean scalar product in $\mathbb{R}^3$. The bracket has two Casimirs,[1]

$$I(\boldsymbol{\gamma}) = \langle \boldsymbol{\gamma}, \boldsymbol{\gamma}\rangle = 1 \quad \text{and} \quad L_z(\boldsymbol{\gamma}, \boldsymbol{l}) = \langle \boldsymbol{\gamma}, \boldsymbol{l}\rangle = l. \tag{2.4}$$

Fixing the Casimirs to the values $I(\boldsymbol{\gamma}) = 1$ and $L_z(\boldsymbol{\gamma}, \boldsymbol{l}) = l$ defines the reduced phase space $\mathrm{T}^*_l\mathbb{S}^2$. Fixing, in addition, the energy $H = h$ defines the reduced energy surface $\mathcal{E}^3_{h,l}$. The effective (or amended) potential on $\mathbb{S}^2$ is (see, e.g., [9])

$$U_l(\boldsymbol{\gamma}) = \frac{l^2}{2\langle \boldsymbol{\gamma}, A\boldsymbol{\gamma}\rangle} - \langle \boldsymbol{\gamma}, \mathbf{r}\rangle. \tag{2.5}$$

The term proportional to l^2 (the kinetic energy of rotation about the axis of gravity) contains the moment of inertia $\langle \boldsymbol{\gamma}, A\boldsymbol{\gamma}\rangle$. The accessible region on $\mathbb{S}^2$ for fixed energy h and angular momentum l is

$$\mathcal{U}_{h,l} = \left\{\boldsymbol{\gamma} \in \mathbb{S}^2 : U_l(\boldsymbol{\gamma}) \le h\right\}. \tag{2.6}$$

The energy-Casimir map is given by

$$\begin{aligned} F : \mathbb{R}^6 &\longrightarrow \mathbb{R}^3(h, l, 1), \\ (\boldsymbol{\gamma}, \boldsymbol{l}) &\longmapsto \big(H(\boldsymbol{\gamma}, \boldsymbol{l}), L_z(\boldsymbol{\gamma}, \boldsymbol{l}), I(\boldsymbol{\gamma})\big). \end{aligned} \tag{2.7}$$

Its critical points $(\boldsymbol{\gamma}, \boldsymbol{l})$, for which $\operatorname{rank} \mathrm{D}F < 3$, are the relative equilibria,[2] and their images under F are the critical values. Since the last component of the energy-Casimir map is

[1] Note that l does not denote the length of $\boldsymbol{l}$ but the value of its z-component $\langle \boldsymbol{\gamma}, \boldsymbol{l}\rangle$.

[2] See, e.g., Arnold [1, Appendix 5C].

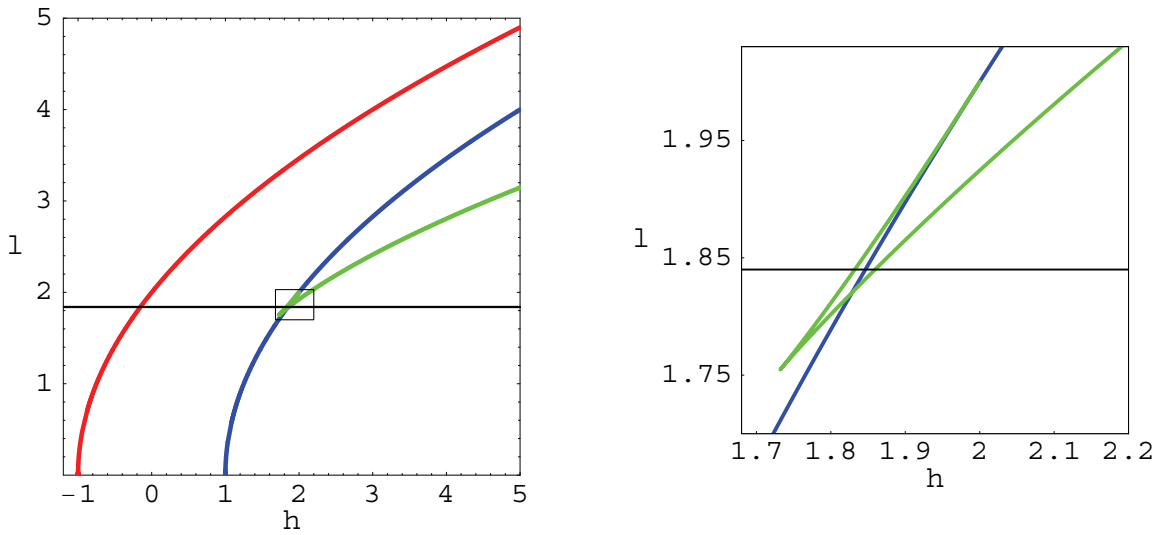


Figure 1. *The bifurcation diagram showing critical values (h, l) of the energy-Casimir map F for the Kovalevskaya case $A = \mathrm{diag}(2, 2, 1)$, $\mathbf{r} = (1, 0, 0)$. Right: Magnification of the rectangle shown on the left.*

constant, it suffices to consider the (h, l) energy-momentum plane as the image. The energy surface $\mathcal{E}^3_{h,l}$ is the preimage of a point (h, l) under F. The set of critical values is called the bifurcation diagram. For critical values the energy surface is not a smooth manifold, and the topology of $\mathcal{E}^3_{h,l}$ in general changes upon crossing critical values.

Instead of computing the rank of $\mathrm{D}F$, the equilibria of (2.1) can be computed directly. The relation between the two approaches is that the gradients of the Casimirs are in the kernel of the Poisson structure, and if $\operatorname{rank}\mathrm{D}F$ is lower than 3, the gradient of the Hamiltonian is zero or a linear combination of the gradients of the Casimirs. A third method to obtain the relative equilibria is to compute the critical points of the effective potential U_l. The energy surface can be viewed as a singular circle bundle over the accessible region $\mathcal{U}_{h,l}$; see, e.g., [2]. At critical points of U_l the topology of $\mathcal{U}_{h,l}$ changes, inducing a change in the topology of the energy surface. Thus a bifurcation diagram, as shown in Figure 1, can be viewed either as a statement about the existence of relative equilibria at the critical values shown or as a statement about the topology of the reduced energy surface for the noncritical values.

For example [9], along the line $l = 1.84$ in Figure 1, the topologies of the accessible regions $\mathcal{U}_{h,l}$ are, from left to right, the empty set, a disk, a disk with two holes, a disk with one hole, and all of $\mathbb{S}^2$. The transitions take place when crossing the lines of critical values in the bifurcation diagram. The corresponding topologies of the energy surface $\mathcal{E}^3_{h,l}$ are determined by the topology of the accessible region; see [2]. They are $\emptyset$, $\mathbb{S}^3$, $(\mathbb{S}^1\times\mathbb{S}^2)\#(\mathbb{S}^1\times\mathbb{S}^2)$, $\mathbb{S}^1 \times \mathbb{S}^2$, $\mathbb{R}\mathrm{P}^3$, again from left to right. More general cases of the bifurcation diagram have been discussed in [7, 11]; see also [6].

It should be remembered that of the three Euler angles, only $\vartheta(t)$ and $\psi(t)$ appear in the Euler–Poisson equations; the angle $\varphi(t)$ must be determined by integration of [9]

$$\dot{\varphi} = \frac{l_1\gamma_1 + l_2\gamma_2}{2(\gamma_1^2 + \gamma_2^2)}. \tag{2.8}$$

Hence, relative equilibria $\boldsymbol{\gamma} = \text{const}$, $\boldsymbol{l} = \text{const}$ of the symmetry-reduced system correspond to periodic motion of the full system.

3. Poincaré section. For systems with two degrees of freedom and 3-dimensional compact energy surfaces, the relevant information about phase space structure (stability of periodic orbits and relative extent and entanglement of regular and chaotic motion) is contained in the 2D maps induced by the motion on suitably chosen Poincaré sections. This has become a standard tool from celestial mechanics to molecular dynamics; yet in the study of rigid bodies there have so far been only a few applications [3, 4, 8]. However, as the Euler–Poisson equations describe a system with effectively two degrees of freedom (at given l), Poincaré sections are the ideal method by which to study the complexity of rigid body dynamics. In the following, we propose a variant which is at the same time general and easy to implement. We begin with a few general considerations.

3.1. General features of Poincaré sections. It would be nice to identify a 2D submanifold $\mathcal{P}^2_{h,l}$ of the energy surface $\mathcal{E}^3_{h,l}$ in such a way that every orbit meets it repeatedly and transversally. As was shown in [5], *complete* sections can indeed be defined, i.e., sections which capture every single trajectory. However, globally transverse sections do not exist in general for topological reasons [2]. Namely, when the section $\mathcal{P}^2_{h,l}$ divides $\mathcal{E}^3_{h,l}$ into an "inner" part and an "outer" part, trajectories that "come in" have to "go out" again, so that part of $\mathcal{P}^2_{h,l}$ is traversed by ingoing and another part by outgoing orbits. The boundary between the two is then a subset of $\mathcal{P}^2_{h,l}$ where the trajectories are tangent.

A complete section $\mathcal{P}^2_{h,l}$ can be obtained with a recipe from [5]: given any smooth function W which maps the energy surface into a bounded set, define $S := \dot{W}$ and take $S(\boldsymbol{\gamma}, \boldsymbol{l}) = 0$ as a section condition:

$$\mathcal{P}^2_{h,l} := \left\{(\boldsymbol{\gamma}, \boldsymbol{l}) \in \mathcal{E}^3_{h,l} : S(\boldsymbol{\gamma}, \boldsymbol{l}) = 0\right\}. \tag{3.1}$$

$\mathcal{P}^2_{h,l}$ is a smooth 2D manifold except when the map

$$\begin{aligned} P : \mathcal{E}^3_{h,l} &\longrightarrow \mathbb{R}^4(h, l, 1, 0), \\ (\boldsymbol{\gamma}, \boldsymbol{l}) &\longmapsto (H(\boldsymbol{\gamma}, \boldsymbol{l}), L_z(\boldsymbol{\gamma}, \boldsymbol{l}), I(\boldsymbol{\gamma}), S(\boldsymbol{\gamma}, l)) \end{aligned} \tag{3.2}$$

has a critical point indicated by $\operatorname{rank} \mathrm{D}P < 4$. Clearly the rank of P drops when the energy surface is not a smooth manifold, but there may be additional singular values.

The tangent set, if it exists, is given by those points on $\mathcal{P}^2_{h,l}$ for which $\dot{S} = 0$. Consider, for example, the Hamiltonian $H = \frac{1}{2}p_1^2 + \frac{1}{2}p_2^2 + V(q_1, q_2)$ and the section condition $S = q_2 = 0$.[3] The surface of section is then given by $\mathcal{P}^2_h = \{(q_1, p_1, p_2) : \frac{1}{2}p_1^2 + \frac{1}{2}p_2^2 + V(q_1, 0) = h\}$. Its usual representation is in terms of its projection to the (q_1, p_1)-plane which is a 2:1 map. To make it unique, only the part $\dot{q}_2 > 0$ is considered; the part $\dot{q}_2 < 0$ is ignored because the two parts are related by time reversal and contain no independent information. The line of tangency $\dot{S} = \dot{q}_2 = p_2 = 0$ lies in the projection plane and is in fact the boundary of the energetically accessible region $\{(q_1, p_1) : \frac{1}{2}p_1^2 + V(q_1, 0) \leq h\}$.

[3] This section condition is not of the form $S = \dot{W}$; hence it need not be complete. There may be orbits which never cross the line $q_2 = 0$ in the (q_1, q_2)-plane.

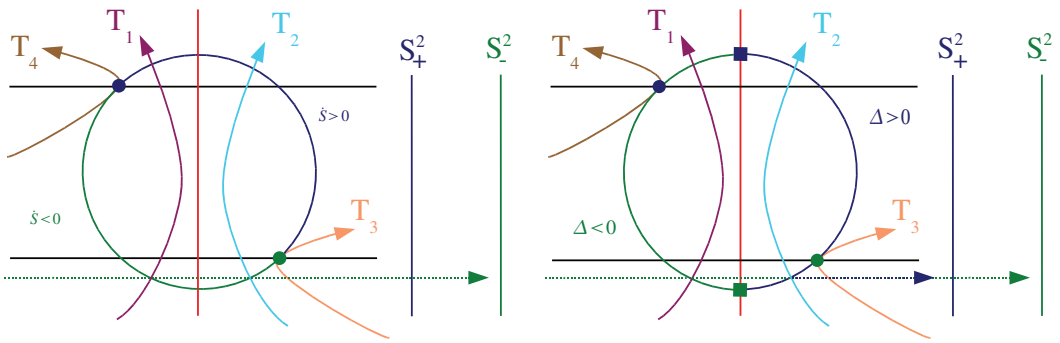


Figure 2. *Schematic representation of the Poincaré surface of section $\mathcal{P}^2_{h,l}$ (circle) with four trajectories, two of which are tangent in the thick dots, and its projections onto two copies of the sphere $\mathbb{S}^2$ (vertical bars). Left: $\mathbb{S}^2_\pm$ defined by the sign of $\dot{S}$; the projection is not invertible. Right: $\mathbb{S}^2_\pm$ defined by the sign of Δ; the projection is invertible.*

The situation is similar but somewhat more complicated in the case of rigid body dynamics. Assume we have a section condition $S(\boldsymbol{\gamma}, \boldsymbol{l}) = 0$. The corresponding surface $\mathcal{P}^2_{h,l} \subset \mathcal{E}^3_{h,l}$ "lives" in the 6-dimensional $(\boldsymbol{\gamma}, \boldsymbol{l})$-space, $\boldsymbol{\gamma} \in \mathbb{S}^2$ and $\boldsymbol{l} \in \mathbb{R}^3$. A convenient projection, as we shall see, maps it onto the Poisson sphere, $\pi : \mathcal{P}^2_{h,l} \to \mathbb{S}^2$. Here it is not generally true that the boundary of the projection coincides with the points of tangency. The situation is rather as shown schematically in the left part of Figure 2. There $\mathcal{P}^2_{h,l}$ corresponds to the circle, and each bar on the right represents $\mathbb{S}^2$. The four trajectories illustrate where $\mathcal{P}^2_{h,l}$ intersects incoming and outgoing orbits ($\dot{S} < 0$ and $\dot{S} > 0$, respectively). In this sketch, the line of tangency reduces to two points which may lie anywhere on the circle. Then, if $\mathbb{S}^2_+$ were chosen to carry the projection of the part where $\dot{S} > 0$, and likewise $\mathbb{S}^2_-$ for $\dot{S} < 0$, these projections would not be 1:1, as illustrated by the green dotted arrow. Intersection points of two different trajectories map to the same point on $\mathbb{S}^2_-$. In this setting, $\boldsymbol{\gamma}$ could not serve as a local coordinate on $\mathcal{P}^2_{h,l}$.

On the other hand, consider the right part of Figure 2. There the circle representing $\mathcal{P}^2_{h,l}$ is divided by the two solid squares in such a way that the two projections to $\mathbb{S}^2_\pm$ are indeed 1:1. The sign of an (as yet unspecified) function Δ determines which side of the circle the intersection is on. A drawback of this division is that both projections contain incoming and outgoing intersections; the points of tangency no longer form the boundaries of the projections. But this turns out not to be serious. The important point is that the projections onto $\mathbb{S}^2_\pm$ provide local coordinates on the two halves of $\mathcal{P}^2_{h,l}$. The intersection points of the two trajectories are now mapped to the same point but on different copies of the sphere.

How is this schematic picture to be implemented? It must be possible to uniquely determine the momenta $\boldsymbol{l}$ on $\mathcal{P}^2_{h,l}$ from the coordinates $\boldsymbol{\gamma}$. With $\mathcal{P}^2_{h,l}$ defined by $H = h$, $L_z = l$, and $S = 0$, the implicit function theorem guarantees that this can be done unless

$$\Delta(\boldsymbol{\gamma}, \boldsymbol{l}) := \det \frac{\partial}{\partial \boldsymbol{l}} (H, L_z, S) = 0. \tag{3.3}$$

The relevant division of $\mathcal{P}^2_{h,l}$ into two parts is therefore given by the sets $\mathcal{P}^+_{h,l} \subset \mathcal{P}^2_{h,l} : \Delta > 0$ and $\mathcal{P}^-_{h,l} \subset \mathcal{P}^2_{h,l} : \Delta < 0$ which project to $\mathbb{S}^2_+$ and $\mathbb{S}^2_-$, respectively. The condition (3.3) contains

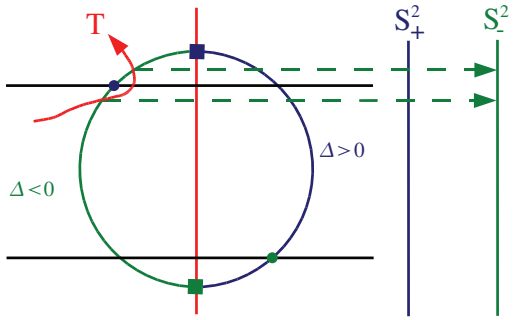


Figure 3. *Schematic representation of the Poincaré surface of section $\mathcal{P}^2_{h,l}$ (circle) with one trajectory near the tangency. Both intersection points are projected onto the same copy of $\mathbb{S}$. Therefore in general both spheres $\mathbb{S}^2_+$ and $\mathbb{S}^2_-$ are necessary to ensure completeness of the section.*

the definition of the dividing line on $\mathcal{P}^2_{h,l}$. It will be seen in (3.7) that there may exist other lines in $\mathcal{P}^2_{h,l}$ where $\Delta = 0$, but these project only to points on $\mathbb{S}^2_\pm$.

Note that even when the system is symmetric under time reversal (which is not the case here, unless $l = 0$), it may not be sufficient to consider only one of the two projections. As can be seen in Figure 3, orbits with intersection points close to tangency tend to have incoming and outgoing intersections with the same sign of Δ; hence, in order to capture all orbits, both projections are needed.

3.2. The section condition. The first general section condition of the type $S = \dot{W}$ was proposed in [6] with $W = \langle \boldsymbol{l}, \boldsymbol{l}\rangle$. It was motivated by a study of the projections of the energy surfaces $\mathcal{E}^3_{h,l}$ to $\boldsymbol{l}$-space. There the preimages of the envelopes of these projections could be characterized by $S = \dot{W} = 0$ or $\langle \boldsymbol{l}, \boldsymbol{\gamma} \times \mathbf{r}\rangle = 0$, i.e., in terms of local extrema of the values of the total angular momentum. Clearly this section condition is an adequate choice when the envelopes in $\boldsymbol{l}$-space are used to represent the Poincaré sections; a number of beautiful examples have been worked out in this setting (Gashenenko, private communication). However, the complex topology of the corresponding surfaces $\mathcal{P}^2_{h,l}$ (being manifolds of genus up to 5 and even higher) makes it hard to study them. Furthermore, the $\boldsymbol{l}$-envelopes are different for each point (h, l) and are difficult to parameterize; hence they do not lend themselves to extended investigation. A possible way out of this last difficulty is to project the surfaces $\mathcal{P}^2_{h,l}$, defined by $W = \langle \boldsymbol{l}, \boldsymbol{l}\rangle$, to the Poisson sphere $\mathbb{S}^2$. However, it turns out [6] that this produces artifacts in the projections which have no relevance in $\mathcal{P}^2_{h,l}$ itself: the condition $W = \langle \boldsymbol{l}, \boldsymbol{l}\rangle$ is well adapted for $\boldsymbol{l}$-space but not for $\boldsymbol{\gamma}$-space.

Therefore, we propose another choice:

$$W = \langle \boldsymbol{\gamma}, \mathbf{r}\rangle \qquad \Rightarrow \qquad S(\boldsymbol{\gamma}, \boldsymbol{l}) = \frac{\mathrm{d}W}{\mathrm{d}t} = \langle A^{-1}\boldsymbol{l}, \mathbf{r} \times \boldsymbol{\gamma}\rangle = 0. \tag{3.4}$$

This is motivated by the behavior of the Lagrange top where $\mathbf{r}$ lies on the body axis so that $\langle \boldsymbol{\gamma}, \mathbf{r}\rangle \propto \cos\vartheta$. The angle ϑ oscillates between $\vartheta_{\min}$ and $\vartheta_{\max}$, which makes $\dot{\vartheta} = 0$ a natural section condition. In addition, when successive points (ϑ_n, ψ_n) $(n = 0, 1, 2, \ldots)$ of the corresponding Poincaré map are plotted on $\mathbb{S}^2$, a natural rotation number may be read off in

terms of the increments of ψ. The section condition (3.4) is an obvious generalization; it picks out extrema of the projection of the center of mass $\mathbf{r}$ to the vertical direction. As we will show, the resulting sections $\mathcal{P}^2_{h,l}$ have genus only up to 4, and they have fewer bifurcations than those studied in [6]. Moreover, the projections $\pi(\mathcal{P}^2_{h,l})$ onto the Poisson sphere coincide with the accessible region $\mathcal{U}_{h,l}$, and the condition $\Delta(\boldsymbol{\gamma}, \boldsymbol{l}) = 0$ defines a line which projects to the boundary $\partial\mathcal{U}_{h,l}$. Finally, this projection is simple because generically it has exactly two preimages; the exceptions are $\boldsymbol{\gamma} \in \partial\mathcal{U}_{h,l}$, which have only one preimage, and $\boldsymbol{\gamma} \parallel \mathbf{r}$, where the preimage is a circle.

To prove these statements, we simultaneously solve the equations $H = h$, $L = l$ (which together define the energy surface $\mathcal{E}^3_{h,l}$), and $S = 0$; i.e., given a point $\boldsymbol{\gamma}$ on the Poisson sphere, we determine the set of corresponding values $\boldsymbol{l}$ such that $(\boldsymbol{\gamma}, \boldsymbol{l}) \in \mathcal{P}^2_{h,l}$. The explicit calculation is given in Appendix A. There we excluded the possibility that $\boldsymbol{\gamma}$ and $\mathbf{r}$ are collinear, $\mathbf{r} \times \boldsymbol{\gamma} = 0$, so let us start with this case here: $\boldsymbol{\gamma} = \pm\mathbf{r}/r =: \hat{\mathbf{r}}$. The condition $S = 0$ is then identically fulfilled and gives no restriction on $\boldsymbol{l}$. The condition $L_z = \langle \boldsymbol{l}, \boldsymbol{\gamma} \rangle = l$ defines a plane in $\boldsymbol{l}$-space, and $H = h$ is the ellipsoid $\langle \boldsymbol{l}, A^{-1}\boldsymbol{l} \rangle = 2(h + \langle \mathbf{r}, \boldsymbol{\gamma} \rangle) = 2(h \pm r)$. Three cases are possible:

(i) The intersection of plane and ellipsoid is a topological circle.
(ii) They do not intersect at all.
(iii) The plane is tangent to the ellipsoid at a point $\boldsymbol{l}^*$.

In the last case the normal to the ellipsoid must be collinear with $\boldsymbol{\gamma}$, i.e., $A^{-1}\boldsymbol{l}^* = \xi\boldsymbol{\gamma}$ or $\boldsymbol{l}^* = \xi A\boldsymbol{\gamma}$. From $\langle \boldsymbol{l}^*, \boldsymbol{\gamma} \rangle = l$ we obtain $\xi = l/\langle \boldsymbol{\gamma}, A\boldsymbol{\gamma} \rangle$; hence the energy equation becomes

$$\frac{l^2}{\langle \boldsymbol{\gamma}, A\boldsymbol{\gamma} \rangle} = 2(h \pm r) \qquad \Leftrightarrow \qquad h = U_l(\boldsymbol{\gamma}) = U_l(\pm\hat{\mathbf{r}})\,. \tag{3.5}$$

So in this case, $\boldsymbol{\gamma} = \pm\hat{\mathbf{r}}$ lies on the boundary of the accessible region, $\boldsymbol{\gamma} \in \partial\mathcal{U}_{h,l}$. In case (i), $\boldsymbol{\gamma}$ lies inside; in case (ii) $\boldsymbol{\gamma}$ lies outside. This will be relevant for the topology of $\mathcal{P}^2_{h,l}$ in the next section.

Let us now consider the points $\boldsymbol{\gamma}$ which are not collinear with $\mathbf{r}$. We show in Appendix A that given $\boldsymbol{\gamma}$, the preimages in $\mathcal{P}^2_{h,l}$ are the points $(\boldsymbol{\gamma}, \boldsymbol{l})$ with

$$\boldsymbol{l} = l\,\frac{A\boldsymbol{\gamma}}{\langle \boldsymbol{\gamma}, A\boldsymbol{\gamma} \rangle} \pm \boldsymbol{v}\sqrt{\frac{2(h - U_l(\boldsymbol{\gamma}))}{\langle \boldsymbol{v}, A^{-1}\boldsymbol{v} \rangle}}\,, \tag{3.6}$$

where $\boldsymbol{v} = (A^{-1}(\mathbf{r} \times \boldsymbol{\gamma})) \times \boldsymbol{\gamma}$. This tells us that for all $\boldsymbol{\gamma}$ with $U_l(\boldsymbol{\gamma}) < h$, i.e., $\boldsymbol{\gamma}$ lies in the interior of $\mathcal{U}_{h,l}$, there are exactly two preimages in $\mathcal{P}^2_{h,l}$. For $\boldsymbol{\gamma}$ on the boundary, $\boldsymbol{\gamma} \in \partial\mathcal{P}^2_{h,l}$, the preimage is unique. No real $\boldsymbol{l}$ exists for $\boldsymbol{\gamma}$ outside $\mathcal{U}_{h,l}$.

Computing the determinant $\Delta(\boldsymbol{\gamma}, \boldsymbol{l})$ according to (3.3), we find

$$\Delta(\boldsymbol{\gamma}, \boldsymbol{l}) = \langle A^{-1}\boldsymbol{l} \times \boldsymbol{\gamma}, A^{-1}(\boldsymbol{\gamma} \times \mathbf{r}) \rangle\,. \tag{3.7}$$

This is zero for $\boldsymbol{\gamma} \in \partial\mathcal{U}_{h,l}$, where $A^{-1}\boldsymbol{l} \times \boldsymbol{\gamma} = 0$, corroborating the assertions made in connection with the right panel of Figure 2: the parts of $\mathcal{P}^2_{h,l}$ where $\Delta > 0$ and $\Delta < 0$ each project to the entire accessible region $\mathcal{U}_{h,l} \subseteq \mathbb{S}^2_\pm$. (Note that $\mathcal{U}_{h,l}$, hence also $\mathcal{E}^3_{h,l}$ and $\mathcal{P}^2_{h,l}$, may consist of several disconnected components.) The other possibility for Δ to become zero, $\boldsymbol{\gamma} \times \mathbf{r} = 0$, is of another kind; its projection to $\mathbb{S}^2_\pm$ gives at most the two points $\boldsymbol{\gamma} = \pm\hat{\mathbf{r}}$. If they are

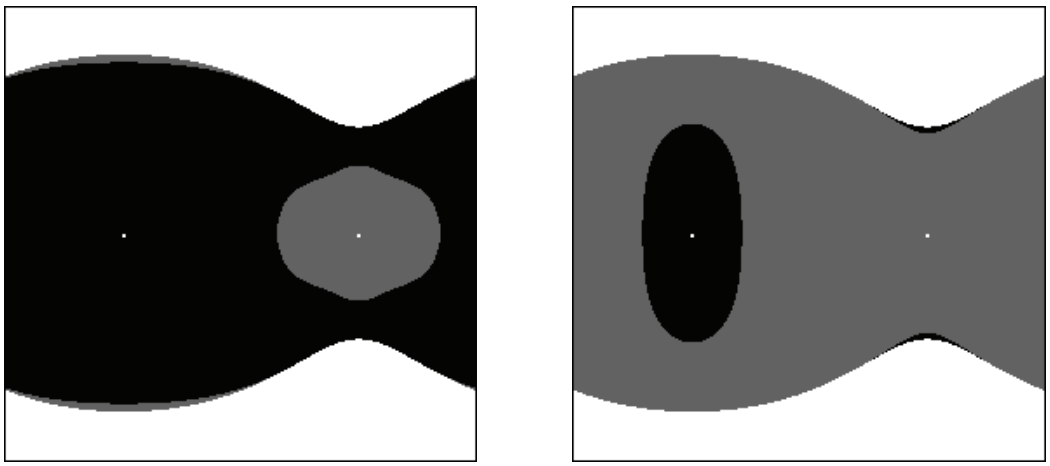


Figure 4. *Projection of $\mathcal{P}^2_{h,l}$ onto the Poisson spheres $\mathbb{S}^2_-$ (left, $\Delta < 0$) and $\mathbb{S}^2_+$ (right, $\Delta > 0$). On the abscissa, ψ is plotted from 0 to 2π; the ordinate ϑ varies from 0 (top) to π (bottom). The black region has $\dot{S} < 0$; the grey region has $\dot{S} > 0$. The parameters are $A = (2, 2, 1)$, $\mathbf{r} = (1, 0, 0)$, $h = 3.4$, $l = 2.87$.*

inside $\mathcal{U}_{h,l}$, their preimages in $\mathcal{P}^2_{h,l}$ are circles; see case (i) above (3.5). The consequences are discussed in the next section.

The condition $\dot{S} = 0$ for the tangency of trajectories with $\mathcal{P}^2_{h,l}$ is obtained from (3.4) by differentiation:

$$\dot{S} = \langle A^{-1}(\boldsymbol{l} \times A^{-1}\boldsymbol{l} - \boldsymbol{\gamma} \times \mathbf{r}), \mathbf{r} \times \boldsymbol{\gamma}\rangle + \langle A^{-1}\boldsymbol{l}, \mathbf{r} \times (\boldsymbol{\gamma} \times A^{-1}\boldsymbol{l})\rangle = 0. \tag{3.8}$$

This must be considered together with (3.6). We do not attempt to derive an explicit expression for the projection of these lines to the Poisson sphere, but they will be shown in the graphical representations of the following sections.

4. Topology of $\mathcal{P}^2_{h,l}$ and its projections. Figure 4 shows a typical projection of $\mathcal{P}^2_{h,l}$ to two copies $\mathbb{S}^2_-$ (left) and $\mathbb{S}^2_+$ (right) of the Poisson sphere for $A = (2, 2, 1)$ and $\mathbf{r} = (1, 0, 0)$. The coordinates of the two panels are $0 \le \psi < 2\pi$ on the abscissa and $0 \le \vartheta \le \pi$ on the ordinate. Recall that θ and ψ are two Euler angles, which after reduction provide spherical coordinates on the Poisson sphere. The white regions at small and large ϑ are inaccessible for the given values of $h = 3.4$ and $l = 2.87$. The grey and black parts together are the accessible region $\mathcal{U}_{h,l}$ which has the topology of an annulus. In the black parts the trajectories are incoming, $\dot{S} < 0$; in the grey region they are outgoing, $\dot{S} > 0$; the boundary between these regions is the tangency where $\dot{S} = 0$. If we were to record only incoming trajectories, only the black parts would need to be considered, but note that they appear on both copies $\mathbb{S}^2_+$ and $\mathbb{S}^2_-$.

The white dots at $(\vartheta, \psi) = (\pi/2, \pi/2)$ represent the "north pole" $\boldsymbol{\gamma} = \hat{\mathbf{r}}$; the white dots at $(\pi/2, 3\pi/2)$ represent the "south pole" $\boldsymbol{\gamma} = -\hat{\mathbf{r}}$. We know from the previous section that these points are special. When they are physically accessible as they are here, then their preimage in $\mathcal{P}^2_{h,l}$ is a circle (instead of a point, as for all other $\boldsymbol{\gamma}$). This suggests representing $\mathcal{P}^2_{h,l}$ in projection not to a sphere but rather to a cylinder: puncture $\mathbb{S}^2$ at the poles and insert "polar

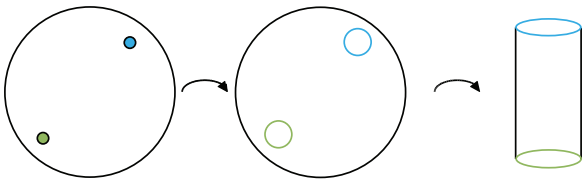


Figure 5. *The projection of $\mathcal{P}^2_{h,l}$ onto the Poisson sphere is singular in two points. For a topologically correct representation of the preimages of these poles $\boldsymbol{\gamma} = \pm\hat{\mathbf{r}}$ in $\mathcal{P}^2_{h,l}$, the poles are replaced by "polar circles," which turns the Poisson sphere into a cylinder, as shown schematically in the figure.*

circles" there, as illustrated schematically in Figure 5. Instead of the two spheres $\mathbb{S}^2_{\mp}$, we then have two cylinders for the projections of $\mathcal{P}^-_{h,l}$ and $\mathcal{P}^+_{h,l}$. But notice that these two cylinders share the same circles at the poles; hence together they form a torus T^2 which we shall call the "PP-torus" T^2_2, the two P's referring to Poisson and Poincaré, or to the two copies of the punctured Poisson sphere (or to *p*unctured *P*oisson) and the subscript 2 indicating that each $\boldsymbol{\gamma}$ is represented twice. We denote its two halves by T^2_-, T^2_+ and remark that each carries one copy of $\mathcal{U}_{h,l}$. $\mathcal{P}^-_{h,l}$ and $\mathcal{P}^+_{h,l}$, respectively, project 1:1 to interior points of the two $\mathcal{U}_{h,l}$, and their boundary, defined by $\Delta(\boldsymbol{\gamma}, \boldsymbol{l}) = 0$, projects to $\partial\mathcal{U}_{h,l}$. Thus, if we identify the two copies of $\partial\mathcal{U}_{h,l}$ on T^2_- and T^2_+, we obtain a 1:1 map $\mathcal{P}^2_{h,l} \to \mathrm{T}^2_2$ which preserves the topology.

The graphical implementation of this idea is automatically achieved if we first transform to new polar coordinates (ϑ', ψ'), where $\boldsymbol{\gamma} = \hat{\mathbf{r}}$ is taken as the polar axis instead of $\boldsymbol{\gamma} = (0, 0, 1)$, and then interpret $\vartheta' = 0$ and $\vartheta' = \pi$ not as points on a sphere but as circles on a cylinder. An example is shown in Figure 6, where the two representations are compared. The top row exhibits $\mathbb{S}^2_{\mp}$, as in Figure 4, and, the bottom row $\mathrm{T}^2_{\mp}$. The north pole $(\vartheta, \psi) = (\pi/2, \pi/2)$ is transformed into the circle $\vartheta' = 0$ (upper boundary) and the south pole into the circle $\vartheta' = \pi$ (lower boundary). Identification of these circles between T^2_- and T^2_+ produces the full torus T^2_2 on which the nonaccessible white regions are now the two holes in the neighborhoods of the points $(\vartheta', \psi') = (\pi/2, 0)$ and $(\pi/2, \pi)$. Since corresponding boundaries $\partial\mathcal{U}_{h,l}$ on T^2_- and T^2_+ must also be identified, each pair of holes on the two sides generates a handle. Hence, the surface of section $\mathcal{P}^2_{h,l}$ for this example has the topology of a torus with two handles; i.e., it is a manifold M^2_3 of genus 3. The lines of tangency, in this particular example, seem to have become lines of constant ϑ'.

One trajectory is shown in Figure 6 in pink. Its incoming part $\dot{S} < 0$ lies entirely on T^2_- and the outgoing part $\dot{S} > 0$ on T^2_+. Note that the incoming part of the trajectory comes close to the pole $\boldsymbol{\gamma} = \hat{\mathbf{r}}$, whereas the outgoing part has points close to the south pole $\boldsymbol{\gamma} = -\hat{\mathbf{r}}$. The comparison of the upper and lower rows illustrates that the torus T^2_2 is the most natural basis for a projection of $\mathcal{P}^2_{h,l}$.

Simpler manifolds occur, e.g., when $\mathcal{U}_{h,l}$ is topologically a disk on the Poisson sphere. If this disk does not include any of the poles $\pm\hat{\mathbf{r}}$, its representation on the PP-torus is two disks, one on each half. The section is then the union of two disks which are glued together at their boundary; in other words, $\mathcal{P}^2_{h,l}$ is topologically a sphere $\mathbb{S}^2$. If the disk contains just one of the poles, its representation on the torus is a cylinder, and $\mathcal{P}^2_{h,l}$ is the torus T^2 obtained by identifying its upper and lower rims. If both poles are contained in the disk, the same kind of

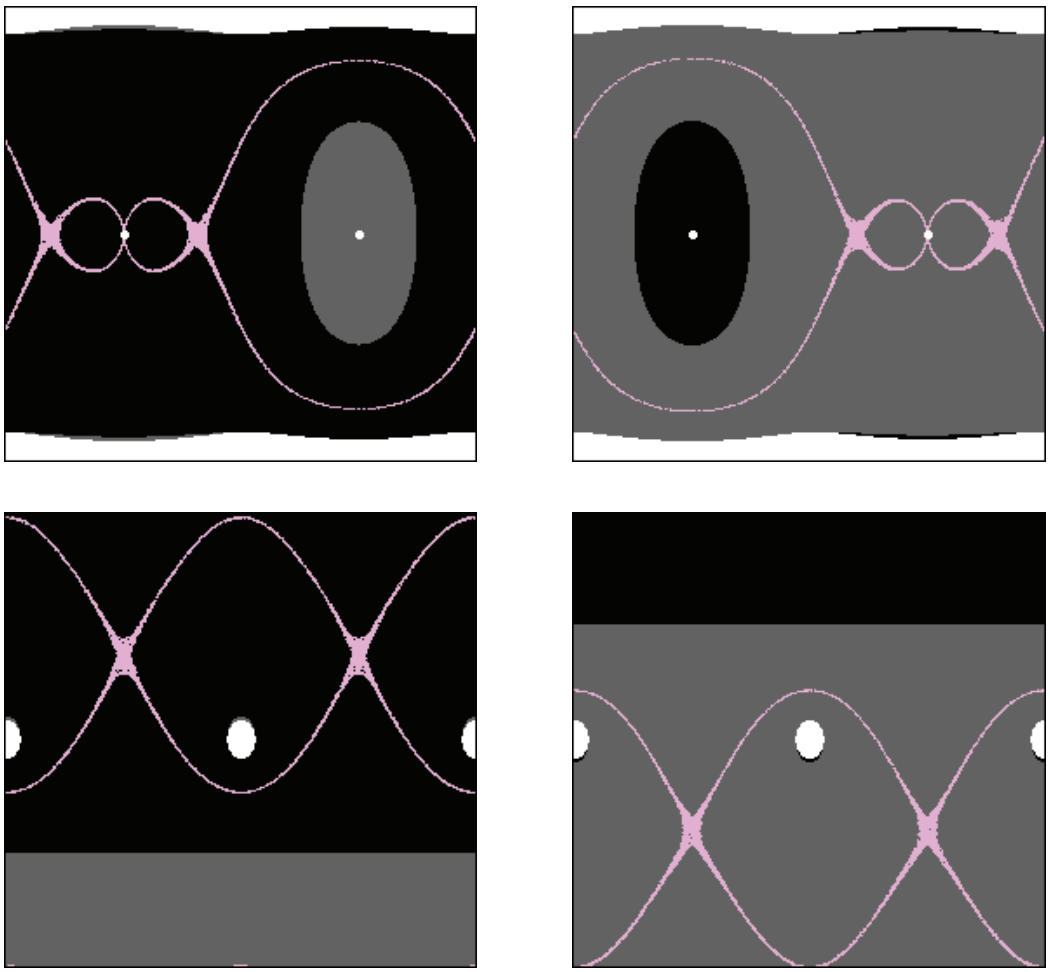


Figure 6. *Comparison of the projection of $\mathcal{P}^2_{h,l}$ onto two copies $\mathbb{S}^2_\mp$ of the Poisson sphere $\mathbb{S}^2$ (top) and onto the two halves $\mathrm{T}^2_\mp$ of the torus T^2_2 (bottom). The poles in the top panels are expanded to circles in the bottom panels, as shown schematically in Figure 5. Incoming intersections of the pink orbit are all on the left in the black region and outgoing on the right in the grey region. The parameters are $A = (2, 1.5, 1)$, $\mathbf{r} = (1, 0, 0)$, $h = 80.5$, $l = 12.80$.*

arguments show that $\mathcal{P}^2_{h,l}$ is a 2-manifold of genus 2.

It is known from the work of Tatarinov [11] and Bolsinov, Dullin, and Wittek [2] that connected components of $\mathcal{U}_{h,l} \subseteq \mathbb{S}^2$ come in only four topological kinds: the entire sphere $\mathbb{S}^2$, a disk $\mathrm{D}^2 = \mathbb{S}^2 \backslash \mathrm{D}^2$, an annulus $\mathbb{S}^2 \backslash 2\mathrm{D}^2$, and a sphere with three holes $\mathbb{S}^2 \backslash 3\mathrm{D}^2$. The corresponding topological types of $\mathcal{P}^2_{h,l}$ are listed in Table 1 for all possible locations of the poles relative to $\mathcal{U}_{h,l}$. They are manifolds of genus g with g ranging from 0 to 4. If we distinguish pairs of $\mathcal{E}^3_{h,l}$ and $\mathcal{P}^2_{h,l}$ when either of the partners is different, then the table shows that there are 10 different possibilities.

To sum up, we propose starting with the PP-torus T^2_2 as the appropriate manifold onto which the Poincaré surface of section $\mathcal{P}^2_{h,l}$ is to be projected 1:1. It is constructed from two copies of the Poisson sphere, where $+\hat{\mathbf{r}}$ serves as reference point for the polar coordinates

Table 1
Possible topologies of the Poincaré surface of section $\mathcal{P}^2_{h,l}$.

	Topology of a connected component of $\mathcal{E}^3_{h,l}$			
	$\mathbb{R}\mathrm{P}^3$	$\mathbb{S}^3$	$\mathbb{S}^1\times\mathbb{S}^2$	$(\mathbb{S}^1\times\mathbb{S}^2)\#(\mathbb{S}^1\times\mathbb{S}^2)$
	Structure of $\mathcal{U}_{h,l}$			
↓ number of poles in $\mathcal{U}_{h,l}$	$\mathbb{S}^2$	D^2	$\mathbb{S}^2\backslash 2\mathrm{D}^2$	$\mathbb{S}^2\backslash 3\mathrm{D}^2$
0	-	$\mathbb{S}^2$	T^2	M^2_2
1	-	T^2	M^2_2	M^2_3
2	T^2	M^2_2	M^2_3	M^2_4

(ϑ',ψ'). The spheres are punctured at the two poles $+\hat{\mathbf{r}}$ and $-\hat{\mathbf{r}}$, which are replaced by circles $(\vartheta',\psi') = (0,\psi')$ and (π,ψ'), respectively. Identifying corresponding circles on these two cylinders produces the torus. It will be seen in the next section that this identification involves a relative shift of the two angles ψ' by π. The PP-torus carries two copies of the accessible region $\mathcal{U}_{h,l}$. Boundaries $\partial\mathcal{U}_{h,l}$ of these two copies are to be identified. The resulting manifold is a 1:1 image of the complete Poincaré surface of section $\mathcal{P}^2_{h,l}$.

We should mention that the representation of $\mathcal{P}^2_{h,l}$ on the PP-torus is not area-preserving when using ϑ,ψ as local coordinates. In principle, area-preserving coordinates can be found, but there seem to be no simple expressions for them. This means that the PP-torus with local coordinates ϑ,ψ is a non–area-preserving representation of an area-preserving map.

5. Examples. In this last section we give nontrivial examples of sections $\mathcal{P}^2_{h,l}$ where the extra bifurcation lines (3.5), given by

$$h = \frac{l^2}{2\langle\hat{\mathbf{r}}, A\hat{\mathbf{r}}\rangle} \mp r, \tag{5.1}$$

do not coincide with bifurcations of the energy surface. This never happens in the Katok family of systems considered in [6] (which includes all cases of Lagrange and Kovalevskaya) because there the center of mass $\mathbf{r}$ lies on a principal axis, and then (5.1) is the bifurcation line corresponding to the relative equilibrium where the body rotates about that axis. In order to make (5.1) different from any of the relative equilibrium conditions, we must choose $\mathbf{r}$ off the principle axes. An arbitrary example is shown in Figure 7, with $A = (2, 1.1, 1)$ and $\mathbf{r} = (0.94868, 0, 0.61623)$. The two orange lines separating from the red and blue relative equilibria correspond to $\boldsymbol{\gamma} = \hat{\mathbf{r}}$ (left) and $\boldsymbol{\gamma} = -\hat{\mathbf{r}}$ (right). When h is increased at fixed l, $\mathcal{U}_{h,l}$ is at first a disk which does not contain any of the poles. Upon crossing the left orange line, $+\hat{\mathbf{r}}$ enters $\mathcal{U}_{h,l}$; then at the right orange line, $-\hat{\mathbf{r}}$ comes in as well.

The six lines in Figure 7 show a partition of the (h,l)-plane (with $l > 0$) into eight regions (plus the nonphysical region to the left of the red line). Seven of them are represented by the black dots on the line $l = 3.25$. The tiny eighth region lies between the yellow and green lines to the left of the orange line (black dot at $(h,l) = (3.1486, 2.72)$). An additional region, not contained in the figure, exists because at large values of h and l the upper branches of the yellow and orange lines intersect; the point $(h,l) = (14.0, 7.2)$ (not shown) is chosen as a representative for this region.

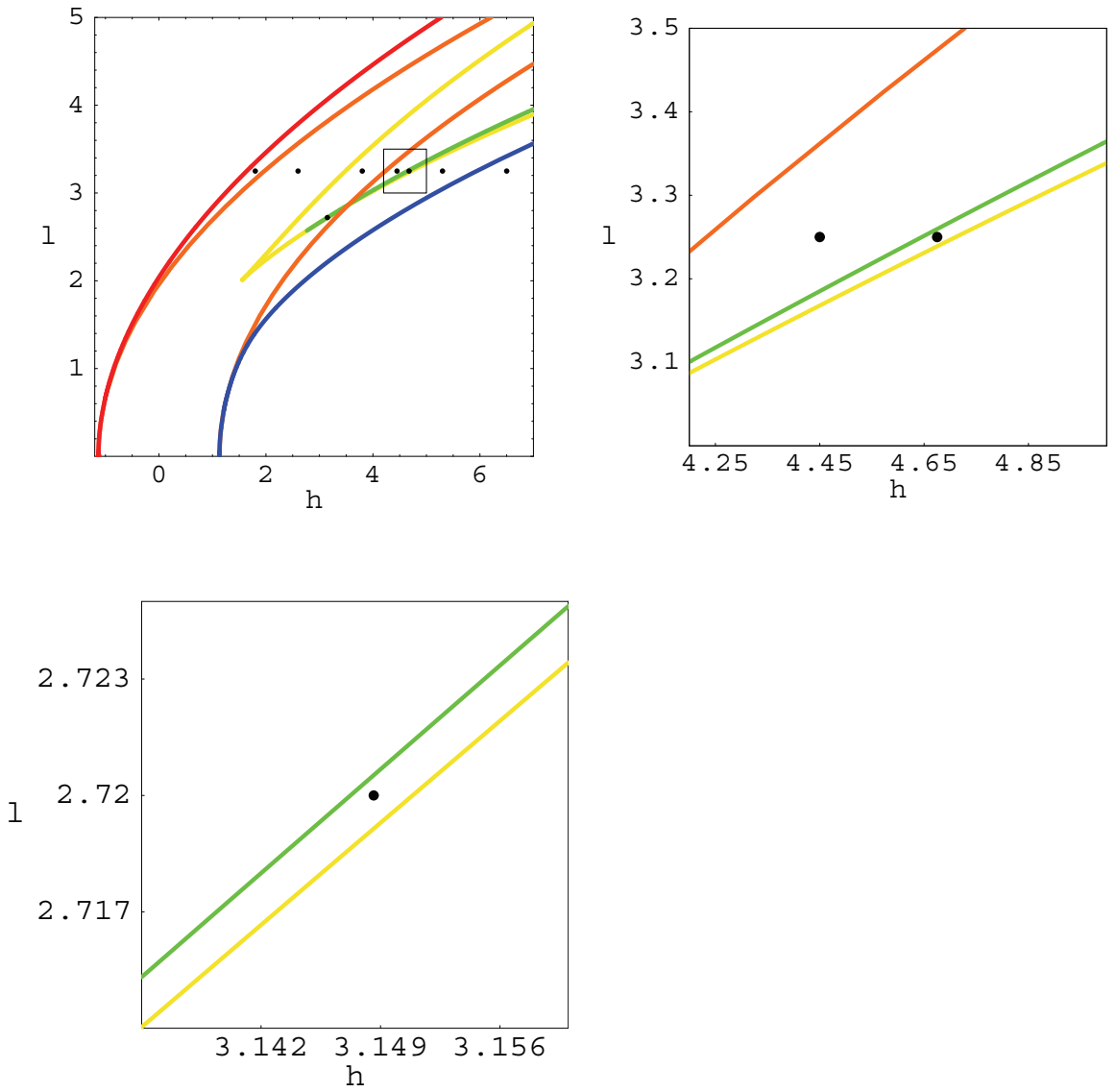


Figure 7. *Bifurcation diagram for the Poincaré section* $\mathcal{P}^2_{h,l}$ *for* $A = (2, 1.1, 1)$ *and* $\mathbf{r} = (0.94868, 0, 0.61623)$. *The red, blue, green, and yellow lines correspond to critical values of the energy-Casimir map* (2.7)*; the two orange lines given by* (5.1) *correspond to changes in the topology of* $\mathcal{P}^2_{h,l}$ *caused by the poles* $\pm\hat{\mathbf{r}} \in \partial\mathcal{U}_{h,l}$. *Top right: Magnification of the rectangle marked on the left. Bottom: Magnification around the point* $(h, l) = (3.1486, 2.72)$ *with lower* l.

These nine regions correspond to nine different topologies of $\mathcal{P}^2_{h,l}$, including six of the types listed in Table 1, and three examples where the section is a disjoint union of surfaces. We discuss them one by one. The following Poincaré sections Figures 8–10 contain typical orbits, with different colors for different orbits. When certain orbits are of the same type (elliptic, hyperbolic, chaotic) and have roughly similar initial conditions for different parameters, they

are given the same color. The pole $\pm\hat{\mathbf{r}}$ is in the section if the upper/lower rim of the picture is accessible.

The first example $(h, l) = (1.8, 3.25)$ represents the leftmost region between the red and orange bifurcation lines. It is shown in the top row of Figure 8. The accessible region $\mathcal{U}_{h,l}$ is a rather small topological disk which does not contain the center of mass $\hat{\mathbf{r}}$. Gluing the two copies together along the boundary of $\mathcal{U}_{h,l}$, we obtain for $\mathcal{P}^2_{h,l}$ a topological sphere $\mathbb{S}^2$.

The second row of Figure 8 shows the case $(h, l) = (2.6, 3.25)$. The north pole (upper rim) has entered $\mathcal{U}_{h,l}$. The orbit structure reveals that when we identify the two circles corresponding to $\hat{\mathbf{r}}$ on $\mathrm{T}^2_{\mp}$, the angle ψ' must be shifted by π. On the Poisson sphere, $\mathcal{U}_{h,l}$ is again a disk D^2, but on the torus $\mathrm{T}^2_+ \cup \mathrm{T}^2_-$ its two copies together form an annulus. Gluing together the two boundaries $\partial\mathcal{U}_{h,l}$ on T^2_2, we find that $\mathcal{P}^2_{h,l}$ is a torus.

Increasing the energy beyond the yellow line in Figure 7, the topology of $\mathcal{E}^3_{h,l}$ changes from $\mathbb{S}^3$ to $2\mathbb{S}^3$, and a new disjoint disk appears as part of $\mathcal{U}_{h,l}$. This disk contains no pole; hence we deduce that $\mathcal{P}^2_{h,l}$ is the union of a torus T^2 and a sphere $\mathbb{S}^2$; see the third row of Figure 8 with $(h, l) = (3.8, 3.25)$. It may be surprising that the trajectories shown indicate only regular behavior with no sign of chaos even though the equations of motion are certainly nonintegrable.

The next point $(h, l) = (4.45, 3.25)$ (see the right part of Figure 7 and the last row of Figure 8) represents a region where both disjoint disks of $\mathcal{U}_{h,l}$ contain a pole. The surface of section is therefore $2\mathrm{T}^2$. It appears that the accessibility of the south pole (corresponding to the center of mass in the upright position) suddenly introduces manifest chaos. This is caused by the hyperbolic equilibrium point nearby. While the component around $\hat{\mathbf{r}}$ still looks very regular, the one containing $-\hat{\mathbf{r}}$ is dominated by chaotic motion. The yellow and purple orbits belong to the same chaotic region but remain separate for a long time, indicating the presence of cantori related to the chain of islands between them.

The next three cases with $l = 3.25$ are shown in Figure 9. The first row with $h = 4.675$ represents the narrow (h, l)-region between the green and yellow bifurcation lines; see the right part of Figure 7. $\mathcal{U}_{h,l}$ is the Poisson sphere minus two disks, i.e., a topological annulus; as this annulus contains both poles, the surface of section $\mathcal{P}^2_{h,l}$ is a manifold M^2_3 of genus 3. Compared to the previous case $h = 4.45$, the two disks about the poles have merged at two points and now allow the chaos to sneak into the upper region. Nevertheless, the neighborhood of $\hat{\mathbf{r}}$ remains mostly regular.

One of the two inaccessible disks disappears as h is increased beyond the yellow line in Figure 7; see the second row of Figure 9 where $h = 5.3$. $\mathcal{U}_{h,l}$ is now a disk which contains both poles; hence $\mathcal{P}^2_{h,l}$ is a manifold M^2_2 of genus 2. It appears that the phase space is fairly distinctly divided into one chaotic and two regular parts. The "outgoing" intersections of the chaotic orbits take place in a region connected to the south pole $\hat{\mathbf{r}}$; the "incoming" lie between the poles. A bigger regular region surrounds the north pole, and a smaller one connects to the south pole at about one half of the ψ'-circle (abscissa). So chaos and order meet at the south pole, depending on the angle ψ'.

Finally, when h crosses the blue bifurcation line, the whole Poisson sphere becomes accessible. This is shown, with $h = 6.5$, in the last row of Figure 9. The section is now isomorphic to the entire PP-torus T^2_2. As to the distribution of regular and chaotic orbits on this torus,

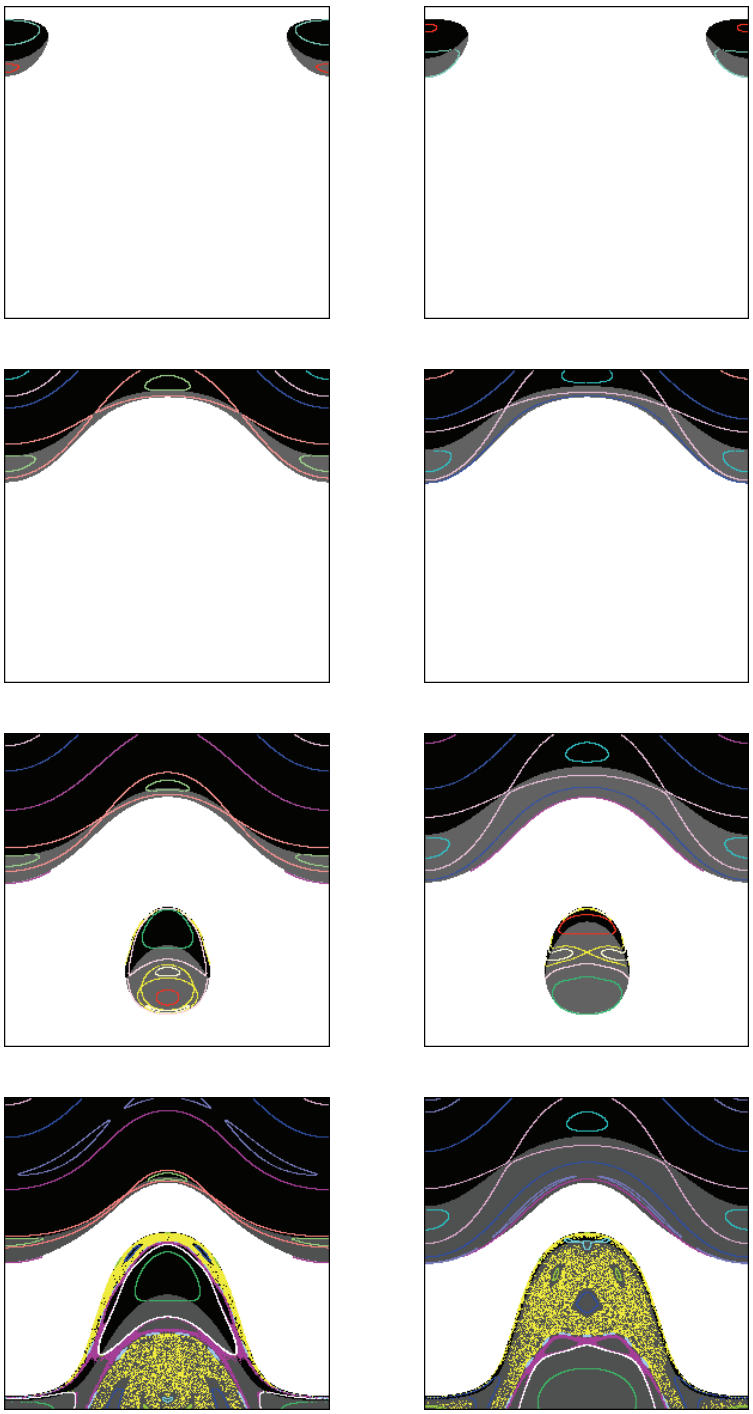


Figure 8. *Projections of the Poincaré section $\mathcal{P}^2_{h,l}$ onto T^2_- (left) and T^2_+ (right). Parameters are as indicated in Figure 7, i.e., $A = (2, 1.1, 1)$ and $\mathbf{r} = (0.94868, 0, 0.61623)$. The angular momentum is $l = 3.25$ in all cases; the energy h from top to bottom is 1.8, 2.6, 3.8, and 4.45. The corresponding topologies of $\mathcal{P}^2_{h,l}$ are $\mathbb{S}^2$, T^2, $\mathbb{S}^2 \cup \mathrm{T}^2$, $2\mathrm{T}^2$. In the first case neither $\pm\hat{\mathbf{r}}$ is in the section, in the two middle cases $+\hat{\mathbf{r}}$ is in the section, and in the last case both $\pm\hat{\mathbf{r}}$ are in the section.*

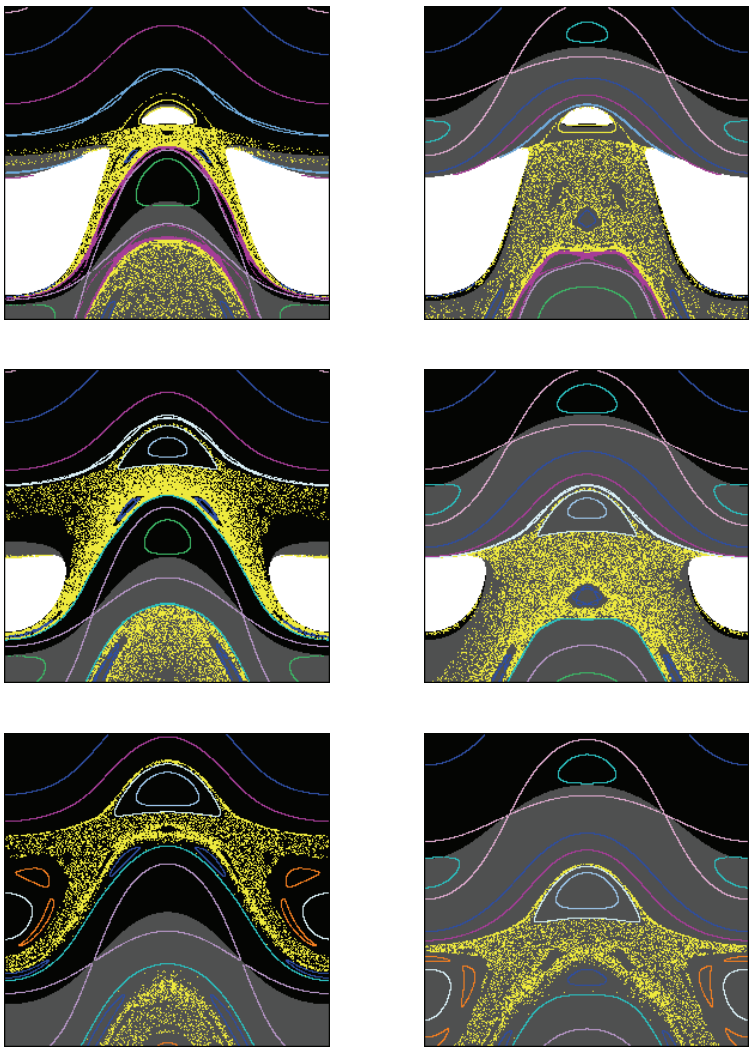


Figure 9. *Projections of the Poincaré section $\mathcal{P}^2_{h,l}$ onto T^2_- (left) and T^2_+ (right). Parameters are as indicated in Figure 7, i.e., $A = (2, 1.1, 1)$ and $\mathbf{r} = (0.94868, 0, 0.61623)$. The angular momentum is $l = 3.25$ in all cases; the energy h from top to bottom is 4.675, 5.3, and 6.5. The corresponding topologies of $\mathcal{P}^2_{h,l}$ are M^2_3, M^2_2, T^2.*

we observe again two regular parts separated by two chaotic regions.

Two more types of sections are shown in Figure 10. They do not exist for $l = 3.25$. The upper row shows the example $(h, l) = (3.1486, 2.72)$; the corresponding dot in Figure 7 lies between the green and yellow lines, but to the left of the orange line. Compared to the first row of Figure 9, the lower part does not contain the south pole, yet there remains a small hole (barely visible in the figure) inside the accessible region, near the center. $\mathcal{U}_{h,l}$ is a topological annulus containing the north pole; hence $\mathcal{P}^2_{h,l}$ is of type M^2_2. In the scenario where this case emerges from the third row of Figure 8, an outbreak of chaos takes place in the lower part.

The last case is $(h, l) = (14.0, 7.2)$; the accessible region consists of two disks, neither of

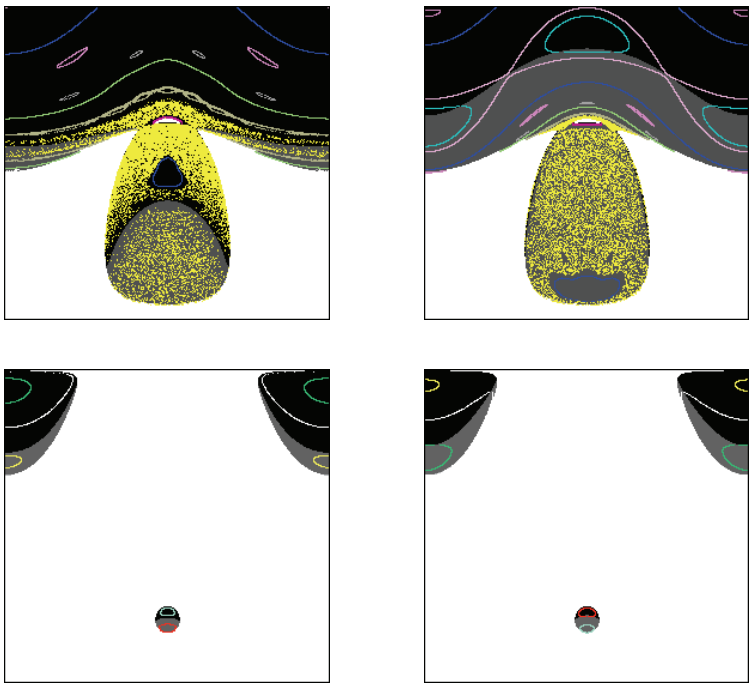


Figure 10. *Projections of the Poincaré section $\mathcal{P}^2_{h,l}$ onto T^2_- (left) and T^2_+ (right). The parameters are $A = (2, 1.1, 1)$ and $\mathbf{r} = (0.94868, 0, 0.61623)$; top row $(h, l) = (3.1486, 2.72)$, bottom row $(h, l) = (14.0, 7.2)$. The corresponding topologies of $\mathcal{P}^2_{h,l}$ are M^2_2, $2\mathbb{S}^2$.*

which contains a pole. Hence $\mathcal{P}^2_{h,l}$ is the union of two topological spheres.

With this series of nine cases we have illustrated six out of the 10 possible topologies of $\mathcal{P}^2_{h,l}$ listed in Table 1. Notice that our choice $A = (2, 1.1, 1)$ represents only one of the seven different classes of moments of inertia that Katok distinguishes with $\hat{\mathbf{r}} = (1, 0, 0)$ (the class K2 in the notation of [6]). With other choices it is possible to also find examples for the other types of $\mathcal{U}_{h,l}$.

6. Summary and outlook. We have presented a new tool for the study of rigid body dynamics: a complete Poincaré section $\mathcal{P}^2_{h,l}$ and its 1:1 representation in projection to a suitably defined torus, the PP-torus T^2_2. The surface $\mathcal{P}^2_{h,l} \subset \mathcal{E}^3_{h,l}$ is defined by local extrema of the vertical component of the center of mass $\mathbf{r}$, namely, by $\mathrm{d}\langle\boldsymbol{\gamma}, \mathbf{r}\rangle/\mathrm{d}t = 0$. The PP-torus is obtained from two copies $\mathbb{S}^2_+$ and $\mathbb{S}^2_-$ of the Poisson sphere by the following procedure: puncture $\mathbb{S}^2_\pm$ at the points $\boldsymbol{\gamma} = \pm\mathbf{r}/r$, replace these points by "polar circles" to obtain two cylinders $\mathrm{T}^2_\pm$, and then identify corresponding polar circles on T^2_+ and T^2_-. The torus T^2_2 so constructed contains two copies of the accessible $\boldsymbol{\gamma}$-region, and $\mathcal{P}^2_{h,l}$ projects 1:1 onto the union of these two copies. Identifying the boundaries $\partial\mathcal{U}_{h,l}$ on T^2_+ and T^2_- produces a manifold which is homeomorphic to $\mathcal{P}^2_{h,l}$ and at the same time readily accessible to intuition.

We have shown examples for nine topologically different surfaces $\mathcal{P}^2_{h,l}$ and propose using this tool for further exploration of the dynamics of rigid bodies in their 4D parameter space. In a forthcoming publication we will apply it to the family of systems $A = (2, 2\eta, 1)$, $\mathbf{r} = (1, 0, 0)$,

which contains the integrable system of Lagrange ($\eta = 1$) and Kovalevskaya ($\eta = 2$). Pictures like those of Figures 8–10 reveal the fate of particular features of the phase space structure under parameter variation, such as location and stability of isolated periodic orbits, their bifurcation schemes, and the extent and entanglement of regular and chaotic regions.

Appendix A. Explicit characterization of $\mathcal{P}^2_{h,l}$. Given an energy h, an angular momentum l, and a point $\boldsymbol{\gamma}$ on the Poisson sphere (not collinear with $\mathbf{r}$), we determine the corresponding vectors $\boldsymbol{l}$ on the Poincaré surface defined by $S = 0$, or $\langle \boldsymbol{l}, A^{-1}(\mathbf{r} \times \boldsymbol{\gamma})\rangle = 0$. For abbreviation we introduce the notation

$$\boldsymbol{u} := A^{-1}(\mathbf{r} \times \boldsymbol{\gamma}), \qquad \boldsymbol{v} := \boldsymbol{u} \times \boldsymbol{\gamma}. \tag{A.1}$$

The equations $S = \langle \boldsymbol{l}, \boldsymbol{u}\rangle = 0$ and $L = \langle \boldsymbol{l}, \boldsymbol{\gamma}\rangle = l$ define planes in $\boldsymbol{l}$-space which intersect in a line with direction $\boldsymbol{v}$. It is easy to check that

$$\boldsymbol{l}_0 = l\,\frac{\boldsymbol{v} \times \boldsymbol{u}}{\langle \boldsymbol{v}, \boldsymbol{v}\rangle} \tag{A.2}$$

is one point on this line; hence we may parameterize the line by $\boldsymbol{l} = \boldsymbol{l}_0 + \boldsymbol{v}t$. The energy equation (2.3) determines the possible values of t. We write it in the form

$$\langle \boldsymbol{l}_0 + \boldsymbol{v}t, A^{-1}(\boldsymbol{l}_0 + \boldsymbol{v}t)\rangle = 2(h + \langle \mathbf{r}, \boldsymbol{\gamma}\rangle) =: c. \tag{A.3}$$

This gives a quadratic equation for t,

$$\langle \boldsymbol{v}, A^{-1}\boldsymbol{v}\rangle t^2 + 2\langle \boldsymbol{l}_0, A^{-1}\boldsymbol{v}\rangle t + \langle \boldsymbol{l}_0, A^{-1}\boldsymbol{l}_0\rangle - c = 0. \tag{A.4}$$

Inserting its solutions into $\boldsymbol{l} = \boldsymbol{l}_0 + \boldsymbol{v}t$, we find $\boldsymbol{l} = \boldsymbol{l}_1 \pm \boldsymbol{l}_2$ with

$$\boldsymbol{l}_1 = \boldsymbol{l}_0 - \frac{\langle \boldsymbol{l}_0, A^{-1}\boldsymbol{v}\rangle}{\langle \boldsymbol{v}, A^{-1}\boldsymbol{v}\rangle}\,\boldsymbol{v} \tag{A.5}$$

and

$$\boldsymbol{l}_2 = \frac{\boldsymbol{v}}{\langle \boldsymbol{v}, A^{-1}\boldsymbol{v}\rangle}\sqrt{\langle \boldsymbol{l}_0, A^{-1}\boldsymbol{v}\rangle^2 - \big(\langle \boldsymbol{l}_0, A^{-1}\boldsymbol{l}_0\rangle - c\big)\langle \boldsymbol{v}, A^{-1}\boldsymbol{v}\rangle}. \tag{A.6}$$

To evaluate further, we use an identity which holds for arbitrary symmetric matrices A and vectors $\boldsymbol{a}, \boldsymbol{b}$:

$$A\boldsymbol{a} \times A\boldsymbol{b} = (\det A)\, A^{-1}(\boldsymbol{a} \times \boldsymbol{b}). \tag{A.7}$$

For example, with $\boldsymbol{v} = \boldsymbol{u} \times \boldsymbol{\gamma}$ we find $A^{-1}\boldsymbol{v} = (A\boldsymbol{u} \times A\boldsymbol{\gamma})/\det A$ and

$$\langle \boldsymbol{v}, A^{-1}\boldsymbol{v}\rangle = \frac{1}{\det A}\langle \boldsymbol{u} \times \boldsymbol{\gamma}, A\boldsymbol{u} \times A\boldsymbol{\gamma}\rangle = \frac{1}{\det A}\langle \boldsymbol{u}, A\boldsymbol{u}\rangle\langle \boldsymbol{\gamma}, A\boldsymbol{\gamma}\rangle; \tag{A.8}$$

in the last step we used Lagrange's identity and $\langle \boldsymbol{\gamma}, A\boldsymbol{u}\rangle = \langle \boldsymbol{\gamma}, \mathbf{r} \times \boldsymbol{\gamma}\rangle = 0$. With similar arguments, and using the explicit form (A.2) for $\boldsymbol{l}_0$, we find

$$\boldsymbol{l}_1 = l\,\frac{A\boldsymbol{\gamma}}{\langle \boldsymbol{\gamma}, A\boldsymbol{\gamma}\rangle}\,. \tag{A.9}$$

To evaluate $\boldsymbol{l}_2$, we first employ Lagrange's identity to obtain

$$\langle \boldsymbol{l}_0, A^{-1}\boldsymbol{v}\rangle^2 - \langle \boldsymbol{l}_0, A^{-1}\boldsymbol{l}_0\rangle\langle \boldsymbol{v}, A^{-1}\boldsymbol{v}\rangle = \langle \boldsymbol{l}_0 \times \boldsymbol{v}, A^{-1}\boldsymbol{v} \times A^{-1}\boldsymbol{l}_0\rangle \tag{A.10}$$

and then (A.7) with A^{-1} instead of A to transform this into

$$\frac{-1}{\det A}\,\langle \boldsymbol{l}_0 \times \boldsymbol{v}, A(\boldsymbol{l}_0 \times \boldsymbol{v})\rangle. \tag{A.11}$$

An easy calculation using $\langle \boldsymbol{v}, \boldsymbol{u}\rangle = 0$ shows that $\boldsymbol{l}_0 \times \boldsymbol{v} = l\boldsymbol{u}$. Inserting this into the radicand of (A.6), we find that it may be written as

$$\frac{\langle \boldsymbol{u}, A\boldsymbol{u}\rangle}{\det A}(c\langle \boldsymbol{\gamma}, A\boldsymbol{\gamma}\rangle - l^2) = 2\langle \boldsymbol{v}, A^{-1}\boldsymbol{v}\rangle\big(h - U_l(\boldsymbol{\gamma})\big) \tag{A.12}$$

with the effective potential of (2.5). The final result for $\boldsymbol{l}$ is

$$\boldsymbol{l} = l\,\frac{A\boldsymbol{\gamma}}{\langle \boldsymbol{\gamma}, A\boldsymbol{\gamma}\rangle} \pm \boldsymbol{v}\sqrt{\frac{2(h - U_l(\boldsymbol{\gamma}))}{\langle \boldsymbol{v}, A^{-1}\boldsymbol{v}\rangle}}\,. \tag{A.13}$$

Acknowledgment. The authors would like to thank Igor Gashenenko for many discussions.

REFERENCES

[1] V. I. Arnold, *Mathematical Methods of Classical Mechanics*, Grad. Texts in Math. 60, Springer-Verlag, Berlin, Heidelberg, New York, 1978.

[2] A. Bolsinov, H. R. Dullin, and A. Wittek, *Topology of energy surfaces and existence of transversal Poincaré sections*, J. Phys. A, 29 (1996), pp. 4977–4985.

[3] H. R. Dullin, *Die Energieflächen des Kowalewskaja-Kreisels*, Mainz Verlag, Aachen, Germany, 1994.

[4] H. R. Dullin, M. Juhnke, and P. H. Richter, *Action integrals and energy surfaces of the Kovalevskaya top*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 4 (1994), pp. 1535–1562.

[5] H. R. Dullin and A. Wittek, *Complete Poincaré sections and tangent sets*, J. Phys. A, 28 (1995), pp. 7157–7180.

[6] I. N. Gashenenko and P. H. Richter, *Enveloping surfaces and admissible velocities of heavy rigid bodies*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 14 (2004), pp. 2525–2553.

[7] S. B. Katok, *Bifurcation sets and integral manifolds in the problem of motion of a heavy rigid body*, Uspekhi Mat. Nauk, 27 (1972), pp. 126–132 (in Russian).

[8] P. H. Richter, *Regular and chaotic rigid body dynamics*, Nonlinear Phenom. Complex Systems, 9 (2006), pp. 115–124.

[9] P. H. Richter, H. R. Dullin, and A. Wittek, *Kovalevskaya Top*, Publikationen zu Wissenschaftlichen Filmen, Sektion Technische Wissenschaften/Naturwissenschaften, 13 (1997), pp. 33–96. Film C1961, Institut für den Wissenschaftlichen Film (IWF), Göttingen, Germany.

[10] Ya. V. Tatarinov, *On the study of the phase space topology of compact configurations with symmetry*, Vestnik Moskov. Univ. Ser. 1 Mat. Mekh., 5 (1973), pp. 70–77.

[11] Ya. V. Tatarinov, *Portraits of classical integrals in the problem of rotation of a rigid body about a fixed point*, Vestnik Moskov. Univ. Ser. 1 Mat. Mekh., 6 (1974), pp. 99–105.

SIAM J. Applied Dynamical Systems
Vol. 8, No. 1, pp. 390–416
© 2009 Society for Industrial and Applied Mathematics

Traveling Wavefronts in a Volume-Filling Chemotaxis Model*

Chunhua Ou† and Wei Yuan‡

Abstract. Chemotaxis is an observed phenomenon in which an organism moves preferentially toward a relatively high concentration, which is contrary to the process of natural diffusion. In this paper, we study a volume-filling model of chemotaxis and provide a valid approach to establishing the existence of traveling wavefronts via the Banach fixed point theorem. Rigorous results hold either when the chemotactic sensitivity is relatively small or when the wave speed is large. Numerical simulations are presented to illustrate the main results, and comparisons of wave patterns in different parameters are demonstrated.

Key words. traveling wavefronts, reaction-diffusion, chemotaxis, volume-filling

AMS subject classifications. 34K15, 35K57, 35R10, 92C17, 92C15

DOI. 10.1137/08072797X

1. Introduction.

An important feature of living organisms is their ability to sense external signals and cues to some particular stimulus. This feature greatly helps them to locate food, search for mates, and avoid being preyed upon. One kind of oriented response to chemical cues is termed *chemotaxis*, or *chemosensitive movement*. Chemotaxis can be either attractive or repulsive. When organisms move preferentially toward a high concentration of a chemical, it is called positive chemotaxis and otherwise negative chemotaxis. Patlak [26, 27] first investigated this kind of movement of organisms in response to an external cue, and Berg [2] extended this investigation to bacterial response to gradients of food with both experimental and theoretical justification. Chemotactic phenomena have also been studied widely in the species *Dictyostelium discoideum*; see [3, 5, 6]. There it was observed that dispersed individual amoeba migrate together toward a high concentration of *cyclic-AMP*, a self-secreted chemical material. Wavelike and other distinctive spatial patterns were also observed experimentally. In 1970, Keller and Segel [15] derived the first continuum-mechanism of pattern formation to describe the aggregation phase of *D. discoideum* and thus made a major contribution to this field. Recently Dallon and Othmer [5] proposed a discrete model which shows how the aggregation phenomenon works from a biological perspective.

Mathematically, the general form of chemotaxis can be described in cellular flux by $J = u(x,t)\chi(v)\nabla v(x,t)$, where $u(x,t)$ represents the density of cell-population and $v(x,t)$ denotes the chemo-attractant (repellent) concentration. Intuitively, the chemotaxis flux is assumed to

*Received by the editors June 20, 2008; accepted for publication (in revised form) by J. Sneyd December 12, 2008; published electronically February 19, 2009.
http://www.siam.org/journals/siads/8-1/72797.html
†Department of Mathematics and Statistics, Memorial University of Newfoundland, St. John's, A1C 5S7 Canada (ou@math.mun.ca). This author's work was supported by the NSERC grant of Canada.
‡Department of Mathematics and Statistics, Memorial University of Newfoundland, St. John's, A1C 5S7 Canada. Current address: Department of Mathematics, Florida State University, Tallahassee, FL 32306-4501 (wyuan@math.fsu.edu).

be proportional to the gradient of chemical concentration, with flux increasing with $u(x,t)$. The factor $\chi(v)$ is the chemotactic sensitivity, and it reflects the response of cells to the chemical $v(x,t)$. Assuming that cell flux obeys the simplest undirected random motion, one can derive the fundamental chemotaxis model

$$\begin{cases} u_t = \nabla \cdot (d_1 \nabla u - u\chi(v)\nabla v) + f(u,v), \\ v_t = d_2 \triangle v + g(u,v), \end{cases} \tag{1.1}$$

where $(t,x) \in (0,-\infty) \times R^n$, $u(x,t)$ denotes the cell density, $v(x,t)$ stands for the concentration of the chemo-attractant, d_1 and d_2 are positive constants, and $f(u,v)$ and $g(u,v)$ are cell's and chemicals' kinetic terms, respectively. Keller and Segel [15] used linear kinetics $g(u,v) = \alpha u - \beta v$ for chemicals and zero kinetics $f(u,v) = 0$ for cells, with all parameters χ, d_1, d_2, α, and β being positive constants. More complicated and realistic models have been developed by Martiel and Goldbeter [17], Monk and Othmer [19], and Hofer, Sherratt, and Maini [12]. For further references, see [28, 33, 34, 36].

A volume-filling chemotaxis model is one of the latest chemotaxis models, introduced recently by Hillen and Painter [9] in 2001. In the volume-filling approach, they assumed that the availability of space in which to move is the main determinant of how the cells make jumps. If cells move preferentially toward a high concentration of chemicals, they will aggregate until the whole finite space is completely packed. Based on this assumption, one of the simplest volume-filling chemotaxis models is given by

$$\begin{cases} u_t = \nabla \cdot (d_1 \nabla u - \chi u(1-u)\nabla v) + f(u,v), \\ v_t = d_2 \triangle v + g(u,v) \end{cases} \tag{1.2}$$

for some functions f and g; see [9, 24, 35]. In this paper we choose $f(u,v) = \mu u(1 - u/u_c)$, where $u_c \leq 1$ represents the carrying capacity corresponding to the density of u, so that the cell kinetics follows the logistic growth equation. Further, we choose $g(u,v) = \alpha u - \beta v$ so that the chemical kinetics is linear in both the cell's and the chemical's concentrations. Here the sensitivity factor χ is assumed to be a positive constant. There are also two other constants in this model we need to mention. The constant 1 in the chemotaxis flux $\chi u(1-u)\nabla v$ is called the crowding capacity, and the other constant $u_c \leq 1$ in f is named the carrying capacity of cell growth density in a particular environment.

Traveling wave behavior in the chemotaxis model has been studied extensively and applied to a variety of problems in bacteria movement [14, 16, 21, 22]. This phenomenon was first observed in experiments with bacteria in a long narrow tube by Adler [1]. Under the conditions of zero chemical diffusion ($d_2 = 0$) and zero cell kinetics $f(u,v) = 0$, Keller and Segel gave an explicit traveling wave solution for (1.1) in [16]; see also [22]. Nagai and Ikeda [21] proved the existence of traveling waves when d_2 is much greater than d_1 and $f(u,v) = 0$. The papers [8, 14, 29, 32] have established the existence of traveling waves for the system in which d_1 and β depend on v, and $\chi(v)$ is a more general function. The stability of chemotactic traveling waves was conjectured by Odell and Keller [22] but not shown, while Gueron and Liron [8] discussed the stability of traveling waves with chemotaxis based on a model of herd grazing. Scribner, Segel, and Rogers [31] indicated that the wave solutions are stable by numerical experiments which also demonstrate an apparent traveling wave formation. Holz

and Chen [11] obtained quantitative data using laser scattering techniques and indicated that the traveling band gradually spreads out rather than maintains speed and shape. For further references, see [32].

However, to the best of our knowledge, no rigorous work has previously been done for the existence of traveling wavefronts when both *the nonlinear chemotaxis term and the nonzero kinetic term f are present*, as shown in system (1.2). The main difficulty is that, on the one hand, the classical phase plane analysis (or the shooting method) seems not to work due to the existence of chemotaxis term; on the other hand, when f is not zero, we cannot develop the ideas in [13, 30] (see formula (2) in [13] and formula (6) in [30]) to express the wave profile u in terms of v. As such, the purpose of this paper is to solve this problem by developing the analytical method in Faria, Huang, and Wu [7] and Ou and Wu [23]. Although here in spirit we use the same idea as that in [7] and [23], there are crucial differences in that the singularly perturbed *chemotaxis term* is nonlinear, which is totally different from that in [23], where the diffusion terms are linear. We prove the existence of traveling wave solutions in two cases: one case is when the chemotactic sensitivity is small, and the other one is when the wave speed is large. For all other cases, numerical simulations for this model are provided to demonstrate the wave patterns.

2. Wave patterns and the minimal wave speed. In this section we use the standard stability analysis to study the existence of traveling wave patterns. The minimal wave speed is provided by this analysis. The selection of the wave speed of propagation from the initial decaying rate is also addressed. For ease of analysis, we shall consider the space domain as $(-\infty, \infty)$.

By a traveling wave solution, we mean a particular solution $u(x,t) = U(z)$, $v(x,t) = V(z)$, $z = x - ct$, where c is the wave speed to (1.2). Specifically, we assume $c \geq 0$ and impose the following boundary conditions:

$$\begin{cases} U(-\infty) = u_c, \\ V(-\infty) = \frac{\alpha u_c}{\beta}, \end{cases} \quad \text{and} \quad \begin{cases} U(+\infty) = 0, \\ V(+\infty) = 0. \end{cases} \tag{2.1}$$

Substituting our proposed solution of the traveling wave form into the system (1.2), we thus have

$$\begin{cases} \frac{d}{dz}\left(d_1 \frac{dU}{dz} - \chi U(1-U)\frac{dV}{dz}\right) + c\frac{dU}{dz} + \mu U\left(1 - \frac{U}{u_c}\right) = 0, \\ d_2 \frac{d^2V}{dz^2} + c\frac{dV}{dz} + \alpha U - \beta V = 0. \end{cases} \tag{2.2}$$

System (1.2) has the spatially homogeneous solutions $(0,0)$ and $(u_c, \alpha u_c/\beta)$, where $u_c \leq 1$. To consider the connection between these two solutions, we first linearize the ODE system (2.2) at the origin $(0,0)$. Let $U' = X$, $V' = Y$; we study the following linear system of (U, X, V, Y):

$$\begin{pmatrix} U' \\ X' \\ V' \\ Y' \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -\frac{\mu}{d_1} & -\frac{c}{d_1} & 0 & 0 \\ 0 & 0 & 0 & 1 \\ -\frac{\alpha}{d_2} & 0 & \frac{\beta}{d_2} & -\frac{c}{d_2} \end{pmatrix} \begin{pmatrix} U \\ X \\ V \\ Y \end{pmatrix}. \tag{2.3}$$

The characteristic equation of the coefficient matrix is

$$\Delta(\lambda) = \left(\lambda^2 + \frac{c}{d_1}\lambda + \frac{\mu}{d_1}\right)\left(\lambda^2 + \frac{c}{d_2}\lambda - \frac{\beta}{d_2}\right), \tag{2.4}$$

where λ is the eigenvalue. To make sure that there exists a positive solution going out from an unstable manifold of the origin and connecting the other equilibrium $(u_c, \alpha u_c/\beta)$, we need to exclude the spiral case to require

$$c \geq c^{**} = 2\sqrt{d_1\mu}. \tag{2.5}$$

Now we assume the existence of wave solutions. Using the same method as in Mollison [18] (see also [20]), we want to derive the selection of the wave speed c from the initial condition at infinity. Suppose the initial condition to system (1.2) is

$$\begin{cases} u(x,0) \sim Ae^{-\xi x}, \\ v(x,0) \sim Be^{-\xi x}, \end{cases}$$

as $x \to \infty$, where $u(x,t)$ and $v(x,t)$ have the same decaying coefficient $\xi > 0$ and A, B are positive constants. If we look for a traveling wave solution of (1.2) in the form (the leading edge form)

$$\begin{cases} u(x,t) = Ae^{-\xi(x-ct)}, \\ v(x,t) = Be^{-\xi(x-ct)} \end{cases} \tag{2.6}$$

and substitute (2.6) into the first equation of system (1.2), this gives the dispersion relation between the wave speed c and the decay rate ξ:

$$c = d_1\xi + \frac{\mu}{\xi}. \tag{2.7}$$

Formula (2.7) provides a direct formula to estimate the wave speed. It is easy to determine from (2.7) that $c_{min} = 2\sqrt{d_1\mu}$ when $\xi = \sqrt{\frac{\mu}{d_1}}$, which agrees with the value obtained by (2.5). Now consider $\min\{e^{-\xi x}, e^{-\sqrt{\frac{\mu}{d_1}}x}\}$ for x large and positive. It follows that

$$\xi < \sqrt{\frac{\mu}{d_1}} \Rightarrow e^{-\xi x} > e^{-\sqrt{\frac{\mu}{d_1}}x}.$$

Then the velocity of propagation with asymptotic initial behavior will depend on the leading edge of the wave, and the wave speed c is given by (2.7). On the other hand, if $\xi > \sqrt{\frac{\mu}{d_1}}$, then $e^{-\xi x}$ is below $e^{-\sqrt{\frac{\mu}{d_1}}x}$ and the front takes the minimal wave speed $c = 2\sqrt{d_1\mu}$. We thus conjecture that the asymptotic wave speed of the traveling wave solution of (2.6) is given by

$$c = \begin{cases} d_1\xi + \frac{\mu}{\xi} & \text{if } 0 < \xi \leq \sqrt{\frac{\mu}{d_1}}, \\ 2\sqrt{d_1\mu} & \text{if } \xi \geq \sqrt{\frac{\mu}{d_1}}. \end{cases} \tag{2.8}$$

Although we cannot theoretically justify this formula, it was confirmed by the simulations in section 5.1 when χ is small.

3. Wave solutions for small sensitivity factor χ. In this section, our purpose is to prove the existence of traveling waves to system (1.2) connecting the two uniform steady-states $(0,0)$ and $(u_c, \alpha u_c/\beta)$ with relatively small chemotactic sensitivity χ. To this end, we first consider the limiting case when $\chi \to 0^+$ and arrive at the following linear diffusion model:

$$\begin{cases} \frac{\partial u}{\partial t} = d_1 \frac{\partial^2 u}{\partial x^2} + \mu u \left(1 - \frac{u}{u_c}\right), \\ \frac{\partial v}{\partial t} = d_2 \frac{\partial^2 v}{\partial x^2} + \alpha u - \beta v. \end{cases} \tag{3.1}$$

Evidently, the first equation of system (3.1) is the well-known Fisher's equation with the logistic kinetics; and it is also decoupled from the second one. Now, constructing a traveling wavefront solution to (3.1) by setting $u(x,t) = U_0(z) = U_0(x-ct)$, $v(x,t) = V_0(z) = V_0(x-ct)$, we obtain the following second order ODE for $U_0(z)$ and $V_0(z)$:

$$\begin{cases} d_1 U_0'' + c U_0' + \mu U_0 \left(1 - \frac{U_0}{u_c}\right) = 0, \\ d_2 V_0'' + c V_0' + \alpha U_0 - \beta V_0 = 0. \end{cases} \tag{3.2}$$

Using the standard phase-plane analysis, Murray [20] has established the existence of traveling wave solutions for the Fisher equation. Once we know the existence of U_0, from the second equation of (3.2) we can obtain V_0 and thus have the following theorem.

Theorem 1. *For any wave speed $c \geq 2\sqrt{d_1 \mu}$, there exists a heteroclinic solution connecting the fixed points $(0,0)$ and $(u_c, \alpha u_c/\beta)$.*

Proof. The first equation of (3.1) admits a traveling wave solution satisfying the boundary conditions $U_0(-\infty) = u_c$ and $U_0(+\infty) = 0$, and it is strictly decreasing; see [20]. From the second equation of (3.2), we have

$$V_0(z) = \frac{1}{d_2(\lambda_2 - \lambda_1)} \left(\int_{-\infty}^{z} e^{\lambda_1 (z-s)} \alpha U_0(s) ds + \int_{z}^{\infty} e^{\lambda_2 (z-s)} \alpha U_0(s) ds \right), \tag{3.3}$$

where

$$\lambda_1 = \frac{-c - \sqrt{c^2 + 4 d_2 \beta}}{2 d_2} < 0, \qquad \lambda_2 = \frac{-c + \sqrt{c^2 + 4 d_2 \beta}}{2 d_2} > 0.$$

We now show that V_0 satisfies the boundary condition (2.1). Applying l'Hôpital's rule to (3.3) and using the fact that $U_0(-\infty) = u_c$ and $U_0(+\infty) = 0$, we obtain

$$\begin{aligned} \lim_{z \to +\infty} V_0(z) &= \lim_{z \to +\infty} \frac{1}{d_2(\lambda_2 - \lambda_1)} \times \left(\frac{\int_{-\infty}^{z} e^{-\lambda_1 s} \alpha U_0(s) ds}{e^{-\lambda_1 z}} + \frac{\int_{z}^{+\infty} e^{-\lambda_2 s} \alpha U_0(s) ds}{e^{-\lambda_2 z}} \right) \\ &= \frac{1}{d_2(\lambda_2 - \lambda_1)} \lim_{z \to +\infty} \left(\frac{\alpha U_0(z)}{-\lambda_1} + \frac{\alpha U_0(z)}{\lambda_2} \right) = 0 \end{aligned}$$

and

$$\begin{aligned} \lim_{z \to -\infty} V_0(z) &= \lim_{z \to -\infty} \frac{1}{d_2(\lambda_2 - \lambda_1)} \times \left(\frac{\int_{-\infty}^{z} e^{-\lambda_1 s} \alpha U_0(s) ds}{e^{-\lambda_1 z}} + \frac{\int_{z}^{+\infty} e^{-\lambda_2 s} \alpha U_0(s) ds}{e^{-\lambda_2 z}} \right) \\ &= \frac{1}{d_2(\lambda_2 - \lambda_1)} \lim_{z \to -\infty} \left(\frac{\alpha U_c(z)}{-\lambda_1} + \frac{\alpha U_c(z)}{\lambda_2} \right) = \frac{\alpha u_c}{\beta}. \end{aligned}$$

Therefore, both U_0 and V_0 satisfy the boundary condition (2.1). This completes the proof.

To obtain the existence of traveling waves when $\chi > 0$, we need the following estimates on the derivative of the wave profile U_0 and V_0.

Lemma 2. *Let V_0 and U_0 be a pair of traveling wavefronts to system (3.1). We have*

$$\frac{-\alpha u_c}{2\sqrt{d_1\mu + d_2\beta}} < V_0'(z) < 0 \quad \text{and} \quad \frac{-(1+\mu)^2 u_c}{8\mu\sqrt{d_1(\mu+1)}} < U_0'(z) < 0.$$

Proof. From the classical result of the Fisher equation we know that $U_0'(z) < 0$; see [20]. It follows from the first equation of system (3.2) that

$$\begin{aligned} U_0(z) = {} & \frac{1}{d_1(\lambda_4 - \lambda_3)} \int_{-\infty}^{z} e^{\lambda_3(z-s)} \left(1 + \mu - \frac{\mu}{u_c} U_0\right) U_0(s) ds \\ & + \frac{1}{d_1(\lambda_4 - \lambda_3)} \int_{z}^{+\infty} e^{\lambda_4(z-s)} \left(1 + \mu - \frac{\mu}{u_c} U_0\right) U_0(s) ds, \end{aligned}$$

where

$$\lambda_3 = \frac{-c - \sqrt{c^2 + 4d_1}}{2d_1} < 0, \qquad \lambda_4 = \frac{-c + \sqrt{c^2 + 4d_1}}{2d_1} > 0.$$

We thus obtain the following inequality for the derivative of U_0:

$$\begin{aligned} U_0'(z) = {} & \frac{\lambda_3}{d_1(\lambda_4 - \lambda_3)} \int_{-\infty}^{z} e^{\lambda_3(z-s)} \left(1 + \mu - \frac{\mu}{u_c} U_0\right) U_0(s) ds \\ & + \frac{\lambda_4}{d_1(\lambda_4 - \lambda_3)} \int_{z}^{+\infty} e^{\lambda_4(z-s)} \left(1 + \mu - \frac{\mu}{u_c} U_0\right) U_0(s) ds \\ > {} & \frac{\lambda_3}{d_1(\lambda_4 - \lambda_3)} \int_{-\infty}^{z} e^{\lambda_3(z-s)} \left(1 + \mu - \frac{\mu}{u_c} U_0\right) U_0(s) ds \\ \geq {} & \frac{\lambda_3}{d_1(\lambda_4 - \lambda_3)} \max\left\{\left(1 + \mu - \frac{\mu}{u_c} U_0\right) U_0\right\} \int_{-\infty}^{z} e^{\lambda_3(z-s)}(s) ds \\ = {} & -\frac{(1+\mu^2)u_c}{4\mu\sqrt{c^2 + 4d_1}} \geq -\frac{(1+\mu)^2 u_c}{8\mu\sqrt{d_1(\mu+1)}}. \end{aligned}$$

For V_0, we rewrite $V_0(z)$ as

$$V_0(z) = \frac{1}{d_2(\lambda_2 - \lambda_1)} \left(\int_{-\infty}^{0} e^{-\lambda_1 t} \alpha U_0(z+t) dt + \int_{0}^{+\infty} e^{-\lambda_2 t} \alpha U_0(z+t) dt \right),$$

and the derivative of $V_0(z)$ is given by

$$V_0'(z) = \frac{1}{d_2(\lambda_2 - \lambda_1)} \left(\int_{-\infty}^{0} e^{-\lambda_1 t} \alpha U_0'(z+t) dt + \int_{0}^{+\infty} e^{-\lambda_2 t} \alpha U_0'(z+t) dt \right).$$

It then follows that $V_0'(z) < 0$ since $U_0'(z+t)$ is negative. On the other hand, by (3.3), we have

$$\begin{aligned} V_0'(z) &= \frac{1}{d_2(\lambda_2-\lambda_1)}\left(\lambda_1\int_{-\infty}^{z} e^{\lambda_1(z-s)}\alpha U_0(s)ds + \lambda_2\int_{z}^{\infty} e^{\lambda_2(z-s)}\alpha U_0(s)ds\right)\\ &> \frac{\lambda_1}{d_2(\lambda_2-\lambda_1)}\int_{-\infty}^{z} e^{\lambda_1(z-s)}\alpha U_0(s)ds\\ &\geq \frac{\lambda_1}{d_2(\lambda_2-\lambda_1)}\max\{\alpha U_0\}\int_{-\infty}^{z} e^{\lambda_1(z-s)}ds\\ &= -\frac{\alpha u_c}{\sqrt{c^2+4d_2\beta}} \geq -\frac{\alpha u_c}{2\sqrt{d_1\mu+d_2\beta}}. \end{aligned}$$

The proof is complete.

Now we are in a position to establish the existence of traveling wave solutions to system (2.2). We will show that the traveling wavefront can be approximated by the corresponding wavefront $(U_0(\xi), V_0(\xi))$ of (3.1) when χ is small. To proceed, we first introduce some notation for later use. Let $C = C(\mathbf{R},\mathbf{R})$ be the Banach space of continuous and bounded functions from $\mathbf{R}$ to $\mathbf{R}$ equipped with the standard norm $\|\phi\|_C = \sup\{|\phi(t)|,\ t \in \mathbf{R}\}$. We denote $C^1 = C^1(\mathbf{R},\mathbf{R}) = \{\phi \in C : \phi' \in C\}$, $C^2 = \{\phi \in C : \phi'' \in C\}$, $C_0 = \{\phi \in C : \lim_{t\to\pm\infty}\phi = 0\}$, and $C_0^1 = \{\phi \in C_0 : \phi' \in C_0\}$, and the corresponding norms are defined by

$$\|\phi\|_{C_0} = \|\phi\|_C, \qquad \|\phi\|_{C^1} = \|\phi\|_{C_0^1} = \|\phi\|_C + \|\phi'\|_C,$$

and

$$\|\phi\|_{C^2} = \|\phi\|_C + \|\phi'\|_C + \|\phi''\|_C.$$

We assume $\binom{U}{V}$ can be approximated by $\binom{U_0}{V_0}$ and thus put

$$\begin{cases} U = U_0 + W_1,\\ V = V_0 + W_2, \end{cases} \tag{3.4}$$

where W_1 and W_2 are two real functions which are subject to the boundary conditions $W_i(\pm\infty) = 0$, $i = 1, 2$.

We first consider the second equation of system (2.2). Substituting (3.4) into the second equation, we have

$$d_2W_2'' + cW_2' - \beta W_2 + \alpha W_1 = 0. \tag{3.5}$$

The solution W_2 to (3.5) can be expressed as

$$W_2(z) = \frac{1}{d_2(\lambda_2-\lambda_1)}\left(\int_{-\infty}^{z} e^{\lambda_1(z-s)}\alpha W_1(s)ds + \int_{z}^{+\infty} e^{\lambda_2(z-s)}\alpha W_1(s)ds\right). \tag{3.6}$$

Taking the first order derivative of W_2 yields

$$W_2'(z) = \frac{1}{d_2(\lambda_2-\lambda_1)}\left(\lambda_1\int_{-\infty}^{z} e^{\lambda_1(z-s)}\alpha W_1(s)ds + \lambda_2\int_{z}^{+\infty} e^{\lambda_2(z-s)}\alpha W_1(s)ds\right), \tag{3.7}$$

and the second order derivative of W_2 is

$$W_2''(z) = \frac{1}{d_2(\lambda_2-\lambda_1)}\left(\lambda_1^2\int_{-\infty}^{z} e^{\lambda_1(z-s)}\alpha W_1 ds + \lambda_2^2\int_{z}^{+\infty} e^{\lambda_2(z-s)}\alpha W_1 ds\right) - \frac{\alpha W_1}{d_2}. \tag{3.8}$$

Similarly, the equation for W_1 is given by

$$d_1 W_1^{''} + cW_1^{'} - W_1 + H_1(W_1, U_0, V_0) = 0, \tag{3.9}$$

where

$$\begin{aligned}
H_1(W_1,U_0,V_0) &= \left(1+\mu-\frac{2\mu U_0}{u_c}\right)W_1 + R_{11} + R_{12} + R_{13},\\
R_{11} &= -\frac{\mu}{u_c}W_1^2,\\
R_{12} &= -\chi\left[(U_0+W_1)(1-U_0-W_1)(V_0''+W_2'') + (1-2U_0-2W_1)(V_0'+W_2')U_0'\right],\\
R_{13} &= -\chi\left[(1-2U_0)(V_0'+W_2')W_1' - (V_0'+W_2')(W_1^2)'\right].
\end{aligned}$$

Thus $W_1(z)$ can be written in the form

$$\begin{aligned}
W_1(z) &= \frac{1}{d_1(\lambda_4-\lambda_3)}\int_{-\infty}^{z} e^{\lambda_3(z-s)}\left[\left(1+\mu-\frac{2\mu U_0}{u_c}\right)W_1(s) + R_{11}+R_{12}+R_{13}\right]ds\\
&\quad + \frac{1}{d_1(\lambda_4-\lambda_3)}\int_{z}^{\infty} e^{\lambda_4(z-s)}\left[\left(1+\mu-\frac{2\mu U_0}{u_c}\right)W_1(s) + R_{11}+R_{12}+R_{13}\right]ds\\
&= \frac{1}{d_1(\lambda_4-\lambda_3)}\int_{-\infty}^{z} e^{\lambda_3(z-s)}\left[\left(1+\mu-\frac{2\mu U_0}{u_c}\right)W_1(s) + R_{11}+R_{12}\right]ds\\
&\quad + \frac{1}{d_1(\lambda_4-\lambda_3)}\int_{z}^{\infty} e^{\lambda_4(z-s)}\left[\left(1+\mu-\frac{2\mu U_0}{u_c}\right)W_1(s) + R_{11}+R_{12}\right]ds + R_1(z,W_1),
\end{aligned} \tag{3.10}$$

where

$$\begin{aligned}
R_1(z,W_1) &= \frac{\chi}{d_1(\lambda_4-\lambda_3)}\int_{-\infty}^{z}[e^{\lambda_3(z-s)}(1-2U_0)(V_0'+W_2')]'W_1 ds\\
&\quad + \frac{\chi}{d_1(\lambda_4-\lambda_3)}\int_{z}^{+\infty}[e^{\lambda_4(z-s)}(1-2U_0)(V_0'+W_2')]'W_1 ds\\
&\quad - \frac{\chi}{d_1(\lambda_4-\lambda_3)}\int_{-\infty}^{z}[e^{\lambda_3(z-s)}(V_0'+W_2')]'W_1^2 ds\\
&\quad - \frac{\chi}{d_1(\lambda_4-\lambda_3)}\int_{z}^{+\infty}[e^{\lambda_4(z-s)}(V_0'+W_2')]'W_1^2 ds,
\end{aligned} \tag{3.11}$$

and W_2' and W_2'' are expressed in terms of W_1 by (3.7) and (3.8).

Define a linear operation $L_1 : C_0 \to C_0$ as follows:

$$\begin{aligned}
L_1(W_1)(z) &= W_1 - \frac{1}{d_1(\lambda_4-\lambda_3)}\int_{-\infty}^{z} e^{\lambda_3(z-s)}\left(1+\mu-\frac{2\mu U_0}{u_c}\right)W_1(s)ds\\
&\quad - \frac{1}{d_1(\lambda_4-\lambda_3)}\int_{z}^{\infty} e^{\lambda_4(z-s)}\left(1+\mu-\frac{2\mu U_0}{u_c}\right)W_1(s)ds.
\end{aligned} \tag{3.12}$$

We claim that $L_1(W_1)(z) \in C_0$ when $W_1(z) \in C_0$. Indeed, we notice that $\lambda_3 < 0$ and $\lambda_4 > 0$; see the definition of them in the proof of Lemma 2. Thus $L_1(W_1)(z)$ is well defined for any $z \in (-\infty, \infty)$. To derive the limit of $L_1(W_1)(z)$ as z tends to $\pm\infty$, we apply l'Hôpital's rule to (3.12). This gives

$$\begin{aligned}
&\lim_{z\to+\infty} L_1(W_1)(z)\\
&= \lim_{z\to+\infty} W_1 - \lim_{z\to+\infty} \frac{1}{d_1(\lambda_4-\lambda_3)}\\
&\quad\times\left(\frac{\int_{-\infty}^{z} e^{-\lambda_3 s}(1+\mu-\frac{2\mu U_0}{u_c})W_1(s)ds}{e^{-\lambda_3 z}} + \frac{\int_{z}^{\infty} e^{-\lambda_4 s}(1+\mu-\frac{2\mu U_0}{u_c})W_1(s)ds}{e^{-\lambda_4 z}}\right)\\
&= \lim_{z\to+\infty} W_1 - \frac{1}{d_1(\lambda_4-\lambda_3)} \lim_{z\to+\infty}\left(1+\mu-\frac{2\mu U_0(z)}{u_c}\right)\left(\frac{W_1(z)}{-\lambda_3} - \frac{W_1(z)}{-\lambda_4}\right)\\
&= 0 - \frac{1}{d_1(\lambda_4-\lambda_3)}\left(1+\mu-\frac{2\mu U_0(\infty)}{u_c}\right)\left(\frac{0}{-\lambda_3} - \frac{0}{-\lambda_4}\right)\\
&= 0.
\end{aligned}$$

The same result holds when $z \to -\infty$.

Before verifying the existence of a solution $W_1(z) \in C_0$ to (3.10), we need to do some preparations to estimate the terms in the right-hand side of (3.10) when $W_1(z) \in C_0$. We have the following lemmas.

Lemma 3. *For each $\rho > 0$, there is a constant $\sigma > 0$ such that*

$$\|R_{11}(\phi) - R_{11}(\varphi)\|_{C_0} \le \rho\|\phi-\varphi\|_{C_0} \tag{3.13}$$

and

$$\begin{aligned}
&\int_{-\infty}^{z} e^{\lambda_3(z-s)}|R_{11}(\phi) - R_{11}(\varphi)|ds + \int_{z}^{+\infty} e^{\lambda_4(z-s)}|R_{11}(\phi) - R_{11}(\varphi)|ds\\
&\le \rho\|\phi-\varphi\|_{C_0}
\end{aligned} \tag{3.14}$$

for all $\phi, \varphi \in B(\sigma)$, where $B(\sigma)$ is the ball in C_0 with radius σ and center at the origin.

Proof. We have

$$\|R_{11}(\phi)\| = O(\|\phi\|_{C_0}^2) \quad \text{as} \quad \|\phi\|_{C_0} \to 0. \tag{3.15}$$

Therefore, (3.13) and (3.14) follow from (3.15).

Lemma 4. *For all $W_1 \in C_0$, we have the estimation*

$$\begin{aligned}
&\int_{-\infty}^{z} e^{\lambda_3(z-s)}R_{12}(z, W_1)ds + \int_{z}^{+\infty} e^{\lambda_4(z-s)}R_{12}(z, W_1)ds\\
&= O(\chi) + O(\chi\|W_1\|) + O(\chi\|W_1\|^2) + O(\chi\|W_1\|^3),
\end{aligned} \tag{3.16}$$

and for any $\rho > 0$, there are constants $\delta, \sigma > 0$ such that

$$\begin{aligned}
&\int_{-\infty}^{z} e^{\lambda_3(z-s)}|R_{12}(z,\phi) - R_{12}(z,\varphi)|ds + \int_{z}^{+\infty} e^{\lambda_4(z-s)}|R_{12}(z,\phi) - R_{12}(z,\varphi)|ds\\
&\le \rho\|\phi-\varphi\|_{C_0}
\end{aligned} \tag{3.17}$$

for all $\phi, \varphi \in B(\sigma)$ and $\chi \in (0, \delta)$, where $B(\sigma)$ is the ball in C_0 with radius σ and center at the origin.

Proof. For later use, we first prove that both $|V_0''(z)|$ and $|W_2''(z)|$ have finite upper bounds. Indeed, we have

$$\begin{aligned} |V_0''(z)| &= \left| \frac{1}{d_2(\lambda_2 - \lambda_1)} \left[\lambda_1^2 \int_{-\infty}^{z} e^{\lambda_1(z-s)} \alpha U_0(s) ds + \lambda_2^2 \int_{z}^{\infty} e^{\lambda_2(z-s)} \alpha U_0(s) ds \right] - \frac{\alpha U_0}{d_2} \right| \\ &< \frac{(\lambda_1^2 + \lambda_2^2)\|\alpha U_0(s)\|}{d_2(\lambda_2 - \lambda_1)} \left| \int_{-\infty}^{z} e^{\lambda_1(z-s)} ds + \int_{z}^{+\infty} e^{\lambda_2(z-s)} ds \right| + \left| \frac{\alpha U_0}{d_2} \right| \\ &< [(\lambda_1^2 + \lambda_2^2) d_2 + \beta] \frac{\alpha u_c}{\beta d_2} \end{aligned} \tag{3.18}$$

and

$$\begin{aligned} |W_2''(z)| &= \left| \frac{1}{d_2(\lambda_2 - \lambda_1)} \left[\lambda_1^2 \int_{-\infty}^{z} e^{\lambda_1(z-s)} \alpha W_1(s) ds + \lambda_2^2 \int_{z}^{\infty} e^{\lambda_2(z-s)} \alpha W_1(s) ds \right] - \frac{\alpha W_1}{d_2} \right| \\ &< [(\lambda_1^2 + \lambda_2^2) d_2 + \beta] \frac{\alpha \|W_1\|}{\beta d_2}. \end{aligned}$$

Thus it is easy to know that

$$\begin{aligned} |R_{12}| &= \left| \chi(U_0 + W_1)(1 - U_0 - W_1)(V_0'' + W_2'') + (1 - 2U_0 - 2W_1)(V_0' + W_2')U_0' \right| \\ &= O(\chi) + O(\chi\|W_1\|) + O(\chi\|W_1\|^2) + O(\chi\|W_1\|^3). \end{aligned}$$

Therefore, we obtain

$$\begin{aligned} &\int_{-\infty}^{z} e^{\lambda_3(z-s)} R_{12}(z, W_1) ds + \int_{z}^{+\infty} e^{\lambda_4(z-s)} R_{12}(z, W_1) ds \\ &= O(\chi) + O(\chi\|W_1\|) + O(\chi\|W_1\|^2) + O(\chi\|W_1\|^3), \end{aligned} \tag{3.19}$$

and (3.17) follows directly from (3.19).

Lemma 5. *For all $W_1 \in C_0$, we have the estimation*

$$|R_1(z, W_1)| \le O(\chi\|W_1\|) + O(\chi\|W_1\|^2) + O(\chi\|W_1\|^3), \tag{3.20}$$

and for any $\rho > 0$, there are constants $\delta, \sigma > 0$ such that

$$|R_1(z, \phi) - R_1(z, \varphi)| \le \rho\|\phi - \varphi\|_{C_0} \tag{3.21}$$

for all $\phi, \varphi \in B(\sigma)$ and $\chi \in (0, \delta)$, where $B(\sigma)$ is the ball in C_0 with radius σ and center at the origin.

Proof. Similarly as in the proofs of Lemmas 3 and 4, (3.21) and (3.20) can be derived from the expressions of (3.7), (3.8), and (3.11).

Now we present our main theorem and its proof.

Theorem 6. *For any $c \ge 2\sqrt{d_1\mu}$, there exists a constant $\delta = \delta(c) > 0$ so that for any $\chi \in [0, \delta]$, system (1.2) has traveling wavefronts $U(x - ct)$ and $V(x - ct)$ satisfying the boundary condition (2.1).*

Proof. Define an operation $T_1 : \Psi \in C^2 \to C$ from the linear part of (3.9) as follows:

$$T_1\Psi(z) = c\Psi'(z) + d_1\Psi''(z) + \left(\mu - \frac{2\mu U_0}{u_c}\right)\Psi(z). \tag{3.22}$$

The formal adjoint equation of $T_1\Psi = 0$ is given by

$$-c\Phi'(z) + d_1\Phi''(z) + \left(\mu - \frac{2\mu U_0}{u_c}\right)\Phi(z) = 0, \quad z \in (-\infty, \infty), \tag{3.23}$$

subject to the boundary condition $\Phi'(\pm\infty) = 0$.

Now, using the method derived by Ou and Wu [23], we divide our proof into five steps.

Step 1. We claim that if $\Phi \in C$ is a solution of (3.23) and Φ is C^2-smooth, then $\Phi = 0$. Moreover, we have $R(T_1) = C$, where $R(T_1)$ is the range space of T_1. Indeed, when $z \to +\infty$, we have $U_0 \to 0$. Then (3.23) tends asymptotically to an equation with constant coefficients

$$-c\Phi'(z) + d_1\Phi''(z) + \mu\Phi(z) = 0. \tag{3.24}$$

The corresponding characteristic equation of (3.24) is

$$d_1\lambda^2 - c\lambda + \mu = 0. \tag{3.25}$$

Both roots of (3.25) have positive real parts as $c \geq 2\sqrt{d_1\mu}$, and thus we conclude that any bounded solution to (3.24) must be the zero solution. So, as $z \to \infty$, any solution to (3.23) other than the zero solution must grow exponentially for large z. Then the only bounded solution is the zero solution. By the Fredholm theory, we have $R(T_1) = C$.

Step 2. Let $\Theta \in C_0$ be given. We claim that if Ψ is a bounded solution of $T_1\Psi = \Theta$, then we have $\lim_{z\to\pm\infty}\Psi(z) = 0$. In fact, when $z \to -\infty$, the equation

$$c\Psi'(z) + d_1\Psi''(z) + \left(\mu - \frac{2\mu U_0}{u_c}\right)\Psi(z) = \Theta \tag{3.26}$$

tends asymptotically to

$$c\Psi'(z) + d_1\Psi''(z) - \mu\Psi(z) = 0. \tag{3.27}$$

Note that for (3.27), the ω-limit set of every bounded solution is just the point $\Psi = 0$. We thus know that every bounded solution to (3.26) also satisfies

$$\lim_{z\to-\infty}\Psi(z) = 0.$$

Similarly, we can obtain that any bounded solution to (3.26) satisfies $\lim_{z\to+\infty}\Psi(z) = 0$. Hence, the claim of Step 2 holds.

Step 3. Considering the linear operation L_1 defined in (3.12), we want to prove that $R(L_1) = C_0$; that is, for any $Z_1 \in C_0$, we have a $W_1 \in C_0$ so that

$$\begin{aligned} &W_1(z) - \frac{1}{d_1(\lambda_4 - \lambda_3)}\int_{-\infty}^{z} e^{\lambda_3(z-s)}\left(1 + \mu - \frac{2\mu U_0}{u_c}\right)W_1 ds \\ &- \frac{1}{d_1(\lambda_4 - \lambda_3)}\int_{z}^{\infty} e^{\lambda_4(z-s)}\left(1 + \mu - \frac{2\mu U_0}{u_c}\right)W_1 ds = Z_1(z). \end{aligned} \tag{3.28}$$

To see this, we assume that $\eta_1(z) = W_1(z) - Z_1(z)$ and obtain an equation for η_1 as follows:

$$\begin{aligned}\eta_1(z) = &\frac{1}{d_1(\lambda_4-\lambda_3)}\int_{-\infty}^{z} e^{\lambda_3(z-s)}\left(1+\mu-\frac{2\mu U_0}{u_c}\right)(\eta_1+Z_1)ds\\ &+\frac{1}{d_1(\lambda_4-\lambda_3)}\int_{z}^{\infty} e^{\lambda_4(z-s)}\left(1+\mu-\frac{2\mu U_0}{u_c}\right)(\eta_1+Z_1)ds.\end{aligned} \tag{3.29}$$

Differentiating both sides twice yields

$$c\eta_1'(z) + d_1\eta_1''(z) + \left(\mu - \frac{2\mu U_0}{u_c}\right)\eta_1(z) = -\left(1+\mu-\frac{2\mu U_0}{u_c}\right)Z_1(z). \tag{3.30}$$

Using the results that $R(T_1) = C$ in Step 1 and $Z_1 \in C_0$, we obtain by Step 2 that there exists a solution $\eta_1(z)$ satisfying (3.30) and $\eta_1(\pm\infty) = 0$. Returning to the variable W_1, we have $W_1 = \eta_1 + Z_1 \in C_0$.

Step 4. Let $N(L_1)$ be the null space of operator L_1 and $N^{\perp}(L_1)$ be the subspace in C_0 so that

$$C_0 = N^{\perp}(L_1) \oplus N(L_1).$$

It is clear that $N^{\perp}(L_1)$ is a Banach space. If we define $S_1 = L_1|_{N^{\perp}(L_1)}$ as an operator which is a restriction of L_1 on $N^{\perp}(L_1)$, then the map $S_1 : N^{\perp}(L_1) \to C_0$ is one to one and onto. By the well-known Banach inverse operator theorem, $S_1^{-1} : C_0 \to N^{\perp}(L_1)$ is a bounded linear operator.

Step 5. When L_1 is restricted to $N^{\perp}(L_1)$, system (3.10) can be written as

$$\begin{aligned}S_1(W_1)(z) = &\frac{1}{d_1(\lambda_4-\lambda_3)}\int_{-\infty}^{z} e^{\lambda_3(z-s)}(R_{11}+R_{12})ds\\ &+\frac{1}{d_1(\lambda_4-\lambda_3)}\int_{z}^{+\infty} e^{\lambda_4(z-s)}(R_{11}+R_{12})ds + R_1(z,W_1).\end{aligned}$$

Since S_1 is invertible, we again rewrite this equation as

$$W_1(z) = F_1(W_1)(z),$$

where

$$F_1(W_1)(z) := S_1^{-1}\begin{pmatrix} \frac{1}{d_1(\lambda_4-\lambda_3)}\int_{-\infty}^{z} e^{\lambda_3(z-s)}[R_{11}+R_{12}]ds \\ +\frac{1}{d_1(\lambda_4-\lambda_3)}\int_{z}^{+\infty} e^{\lambda_4(z-s)}[R_{11}+R_{12}]ds + R_1(z,W_1)\end{pmatrix}.$$

Evidently, from Lemmas 3, 4, and 5, we know that

$$\begin{aligned}\|F_1(W_1)(z)\| &\le \|S_1^{-1}\|\left[O(\chi)+O(\|W_1\|^2)+O(\chi\|W_1\|)+O(\chi\|W_1\|^2)+O(\chi\|W_1\|^3)\right]\\ &= \|S_1^{-1}\|\left[O(\chi)+\|W_1\|\left(O(\chi)+O(\|W_1\|)+O(\chi\|W_1\|)+O(\chi\|W_1^2\|)\right)\right].\end{aligned}$$

Observing that $\|S_1^{-1}\|$ is independent of χ, we want to choose χ small as well as a small ball $B(\sigma)$ so that the operator $F_1(W)$ is self-mapping and contractive for $W \in B(\sigma)$.

By Lemmas 3, 4, and 5 again, we conclude that there exist small constants $\delta, \sigma > 0$ and $0 < \rho < 1$ such that for $W, \varphi, \phi \in B(\sigma) \cap N^{\perp}(L_1)$ and $\chi \in (0, \delta)$, we have the estimates

$$\begin{aligned} \|F_1(W_1)(z)\| &= \|S_1^{-1}\| \left[O(\chi) + O(\|W_1\|^2) + O(\chi\|W_1\|) + O(\chi\|W_1\|^2) + O(\chi\|W_1\|^3)\right] \\ &= \|S_1^{-1}\| \left[O(\chi) + \|W_1\| \left(O(\chi) + O(\|W_1\|) + O(\chi\|W_1\|) + O(\chi\|W_1^2\|)\right)\right] \\ &\le \frac{1}{3}(\sigma + \|W_1\|) \le \sigma, \end{aligned} \tag{3.31}$$

and

$$\|F_1(z, \phi) - F_1(z, \varphi)\| \le \rho\|\phi - \varphi\|.$$

Hence, $F_1(z, W_1)$ is a uniform contractive mapping for $W_1 \in B(\sigma) \cap N^{\perp}(L_1)$. Using the Banach contraction principle, it follows that for $\chi \in (0, \delta)$, system (3.10) has a unique solution $W_1 \in B(\sigma) \cap N^{\perp}(L_1)$; the prototype of this principle in $\mathbf{R}$ can be seen in [39, p. 13]. Returning to the original variable, we have established that $U_0 + W_1$ is a solution connecting the two values 0 and u_c. On the other hand, from (3.6) it follows that the solution $V_0 + W_2$ connects the two values 0 and $\frac{\alpha u_c}{\beta}$. This completes the proof.

4. Traveling wave solutions for large wave speed. In this section we consider how to establish the existence of traveling wavefronts for our model when the wave speed is large. In spirit, the idea originated from Canosa [4] (without a rigorous proof) for the Fisher equation (see also [7] for linear diffusions and delayed kinetics), but we shall verify this approach with mathematical rigor for the volume-filling chemotaxis (*nonlinear flux*) model.

To proceed, we consider the heteroclinic connection between the two equilibria $(u_c, \alpha u_c/\beta)$ and $(0, 0)$ and rescale the traveling wave coordinate by setting $\xi = z/c$, where $z = x - ct$. In terms of variable ξ, the traveling wave equation (2.2) becomes

$$\begin{cases} \frac{d_1}{c^2}\frac{d^2U}{d\xi^2} - \frac{\chi}{c^2}(1-2U)\frac{dU}{d\xi}\frac{dV}{d\xi} - \frac{\chi}{c^2}U(1-U)\frac{d^2V}{d\xi^2} + \frac{dU}{d\xi} + \mu U\left(1 - \frac{U}{u_c}\right) = 0, \\ \frac{d_2}{c^2}\frac{d^2V}{d\xi^2} + \frac{dV}{d\xi} + \alpha U - \beta V = 0. \end{cases} \tag{4.1}$$

Let $\epsilon = \frac{1}{c^2}$, and thus ϵ is small if c is large. Our intention here is to establish the existence of a traveling wave to system (4.1), which connects the two uniform steady-states $(0, 0)$ and $(u_c, \alpha u_c/\beta)$. As shown in section 3, we first present the existence of such wavefronts when $\epsilon = 0$. If so, system (4.1) can be reduced to

$$\begin{cases} \frac{d\tilde{U}_0}{d\xi} = -\mu\tilde{U}_0\left(1 - \frac{\tilde{U}_0}{u_c}\right), \\ \frac{d\tilde{V}_0}{d\xi} = -\alpha\tilde{U}_0 + \beta\tilde{V}_0. \end{cases} \tag{4.2}$$

We find that the first equation of (4.2) is the logistic equation (sometimes called the Verhulst model) which was first investigated by Pierre Verhulst (1838, 1845). If $\tilde{U}_0(0) = u_0$, then the solution of (4.2) can be expressed explicitly as $\tilde{U}_0(\xi) = u_0 u_c/[u_0 + (u_c - u_0)e^{\mu\xi}]$, and it is easy to see that $\tilde{U}_0(-\infty) = u_c$ and $\tilde{U}_0(+\infty) = 0$. On the other hand, $\tilde{V}_0$ can be written in an integral form as

$$\tilde{V}_0(\xi) = \alpha e^{\beta\xi} \int_{\xi}^{+\infty} e^{-\beta t}\tilde{U}_0(t)dt,$$

which also satisfies the boundary condition (2.1). Therefore, the reduced system (4.2) admits a wave profile.

For system (4.1), as before we suppose that (U, V) can be approximated by $(\tilde{U}_0, \tilde{V}_0)$ and set

$$\begin{cases} U = \tilde{U}_0 + W_3, \\ V = \tilde{V}_0 + W_4, \end{cases} \tag{4.3}$$

where W_3 and W_4 are two functions subject to the boundary condition $W_i(\pm\infty) = 0$, $i = 3, 4$. We first consider the equation for W_4. Substitute (4.3) and (4.2) into (4.1) and denote $\frac{d}{d\xi}$ as prime. The equation for W_4 is given by

$$\epsilon d_2 W_4'' + W_4' - \beta W_4 + \alpha W_3 + R_{41} = 0, \tag{4.4}$$

where

$$R_{41} = \epsilon d_2 \left[\beta^2 \tilde{V}_0 - \alpha\beta \tilde{U}_0 + \alpha\mu \tilde{U}_0 \left(1 - \frac{\tilde{U}_0}{u_c} \right) \right].$$

Note that equation $\epsilon d_2 \lambda^2 + \lambda - \beta = 0$ has two real roots λ_7 and λ_8 with

$$\lambda_7 = \frac{-1 - \sqrt{1 + 4d_2\epsilon\beta}}{2\epsilon d_2} < 0, \qquad \lambda_8 = \frac{-1 + \sqrt{1 + 4d_2\epsilon\beta}}{2\epsilon d_2} > 0. \tag{4.5}$$

Using the fact that

$$\epsilon d_2(\lambda_8 - \lambda_7) = \sqrt{1 + 4d_2\epsilon\beta},$$

we can express $W_4(\xi)$ in an integral equation form as

$$W_4(\xi) = \frac{1}{\sqrt{1 + 4d_2\epsilon\beta}} \left[\int_{-\infty}^{\xi} e^{\lambda_7(\xi - t)} (\alpha W_3 + R_{41}) dt + \int_{\xi}^{+\infty} e^{\lambda_8(\xi - t)} (\alpha W_3 + R_{41}) dt \right]. \tag{4.6}$$

For later use, we derive the first and the second derivatives for W_4:

$$\begin{aligned} W_4'(\xi) = {} & \frac{\lambda_7}{\sqrt{1 + 4d_2\epsilon\beta}} \int_{-\infty}^{\xi} e^{\lambda_7(\xi - t)} (\alpha W_3 + R_{41}) dt \\ & + \frac{\lambda_8}{\sqrt{1 + 4d_2\epsilon\beta}} \int_{\xi}^{+\infty} e^{\lambda_8(\xi - t)} (\alpha W_3 + R_{41}) dt \end{aligned} \tag{4.7}$$

and

$$\begin{aligned} W_4''(\xi) = {} & \frac{\lambda_7^2}{\sqrt{1 + 4d_2\epsilon\beta}} \int_{-\infty}^{\xi} e^{\lambda_7(\xi - t)} (\alpha W_3 + R_{41}) dt \\ & + \frac{\lambda_8^2}{\sqrt{1 + 4d_2\epsilon\beta}} \int_{\xi}^{\infty} e^{\lambda_8(\xi - t)} (\alpha W_3 + R_{41}) dt - \frac{(\alpha W_3 + R_{41})}{\epsilon d_2}. \end{aligned} \tag{4.8}$$

On the other hand, the equation for W_3 is written as

$$\epsilon d_1 W_3'' + W_3' - W_3 + H_3(W_3, W_4, \tilde{U}_0, \tilde{V}_0) = 0, \tag{4.9}$$

where

$$
\begin{aligned}
H_3(W_3,\tilde{U}_0,\tilde{V}_0) &= \left(1+\mu-\frac{2\mu\tilde{U}_0}{u_c}\right)W_3+\sum_{i=1}^{6}R_{3i},\\
R_{31} &= -\frac{\mu}{u_c}W_3^2,\\
R_{32} &= \epsilon\mu^2 d_1\tilde{U}_0\left(1-\frac{\tilde{U}_0}{u_c}\right)\left(1-\frac{2\tilde{U}_0}{u_c}\right),\\
R_{33} &= -\epsilon\chi(\tilde{U}_0+W_3)(1-\tilde{U}_0-W_3)\left[\alpha(\mu-\beta)\tilde{U}_0-\frac{\alpha\mu}{u_c}\tilde{U}_0^2+\beta^2\tilde{V}_0\right],\\
R_{34} &= -\epsilon\chi(\tilde{U}_0+W_3)(1-\tilde{U}_0-W_3)W_4'',\\
R_{35} &= -\epsilon\chi(1-2\tilde{U}_0-2W_3)(W_4'-\alpha\tilde{U}_0+\beta\tilde{V}_0)\left(\frac{\mu\tilde{U}_0^2}{u_c}-\mu\tilde{U}_0\right),\\
R_{36} &= -\epsilon\chi(1-2\tilde{U}_0-2W_3)(W_4'-\alpha\tilde{U}_0+\beta\tilde{V}_0)W_3',
\end{aligned}
$$

and W_4' and W_4'' are expressed in terms of W_3 in (4.7) and (4.8). Since the equation $\epsilon d_1\lambda^2+\lambda-1=0$ has two real roots, λ_5 and λ_6, with

$$
\lambda_5=\frac{-1-\sqrt{1+4d_1\epsilon}}{2\epsilon d_1}<0,\qquad \lambda_6=\frac{-1+\sqrt{1+4d_1\epsilon}}{2\epsilon d_1}>0, \tag{4.10}
$$

we can express $W_3(\xi)$ in an integral form as

$$
\begin{aligned}
W_3(\xi) &= \frac{1}{\sqrt{1+4d_1\epsilon}}\int_{-\infty}^{\xi}e^{\lambda_5(\xi-t)}\left[\left(1+\mu-\frac{2\mu U_0}{u_c}\right)W_3+\sum_{i=1}^{6}R_{3i}\right]dt\\
&\quad+\frac{1}{\sqrt{1+4d_1\epsilon}}\int_{\xi}^{\infty}e^{\lambda_6(\xi-t)}\left[\left(1+\mu-\frac{2\mu U_0}{u_c}\right)W_3+\sum_{i=1}^{6}R_{3i}\right]dt\\
&= \frac{1}{\sqrt{1+4d_1\epsilon}}\int_{-\infty}^{\xi}e^{\lambda_5(\xi-t)}\left(1+\mu-\frac{2\mu\tilde{U}_0}{u_c}\right)W_3dt\\
&\quad+\frac{1}{\sqrt{1+4d_1\epsilon}}\int_{\xi}^{\infty}e^{\lambda_6(\xi-t)}\left(1+\mu-\frac{2\mu\tilde{U}_0}{u_c}\right)W_3dt\\
&\quad+\frac{1}{\sqrt{1+4d_1\epsilon}}\left[\int_{-\infty}^{\xi}e^{\lambda_5(\xi-t)}\left(\sum_{i=1}^{5}R_{3i}\right)dt+\int_{\xi}^{\infty}e^{\lambda_5(\xi-t)}\left(\sum_{i=1}^{5}R_{3i}\right)dt\right]\\
&\quad+R_2(\xi,W_3),
\end{aligned} \tag{4.11}
$$

where

$$
\begin{aligned}
R_2(\xi,W_3) &= \frac{\epsilon\chi}{\sqrt{1+4d_1\epsilon}}\int_{-\infty}^{\xi}W_3\left[e^{\lambda_5(\xi-t)}(1-2\tilde{U}_0-2W_3)(W_4'-\alpha\hat{U}_0+\beta\tilde{V}_0)\right]'dt\\
&\quad+\frac{\epsilon\chi}{\sqrt{1+4d_1\epsilon}}\int_{\xi}^{+\infty}W_3\left[e^{\lambda_6(\xi-t)}(1-2\tilde{U}_0-2W_3)(W_4'-\alpha\tilde{U}_0+\beta\tilde{V}_0)\right]'dt.
\end{aligned} \tag{4.12}
$$

Notice that λ_5 tends to $-\infty$ and λ_6 tends to 1 as ε tends to zero. We can take the limiting linear part of (4.11) to define a linear operator which is independent of ε; that is, define a linear operator $L_2 : C_0 \longrightarrow C_0$ by

$$L_2(W_3)(\xi) = W_3(\xi) - \int_{\xi}^{+\infty} e^{(\xi-t)} \left(1 + \mu - \frac{2\mu\tilde{U}_0}{u_c}\right) W_3 dt. \tag{4.13}$$

Then (4.11) can be rewritten as

$$L_2(W_3)(\xi) = W_3(\xi) - \int_{\xi}^{+\infty} e^{(\xi-t)} \left(1 + \mu - \frac{2\mu\tilde{U}_0}{u_c}\right) W_3 dt = A(\epsilon, W_3)(\xi), \tag{4.14}$$

where the remainder $A(\epsilon, W_3)(\xi)$ is given by

$$\begin{aligned} A(\epsilon, W_3)(\xi) = {} & \frac{1}{\sqrt{1+4d_1\epsilon}} \int_{-\infty}^{\xi} e^{\lambda_5(\xi-t)} \left(1 + \mu - \frac{2\mu\tilde{U}_0}{u_c}\right) W_3 dt \\ & + \int_{\xi}^{+\infty} \left(\frac{e^{\lambda_6(\xi-t)}}{\sqrt{1+4d_1\epsilon}} - e^{(\xi-t)}\right) \left(1 + \mu - \frac{2\mu\tilde{U}_0}{u_c}\right) W_3 dt \\ & + \frac{1}{\sqrt{1+4d_1\epsilon}} \left[\int_{-\infty}^{\xi} e^{\lambda_5(\xi-t)} \left(\sum_{i=1}^{5} R_{3i}\right) dt + \int_{\xi}^{+\infty} e^{\lambda_6(\xi-t)} \left(\sum_{i=1}^{5} R_{3i}\right) dt\right] \\ & + R_2(\xi, W_3). \end{aligned} \tag{4.15}$$

To show the existence of a solution $W_3 \in C_0$ to (4.14), we first estimate the terms in (4.15) by the following lemmas.

Lemma 7. *For any given ϵ and $W_3 \in C_0$, there exists the estimation*

$$\|A(W_3, \epsilon)(\xi)\| = O(\epsilon) + O(\epsilon\|W_3\|) + O(\|W_3\|^2) + O(\epsilon\|W_3\|^3), \tag{4.16}$$

and for any $\rho > 0$, there are constants $\delta, \sigma > 0$ such that for any two elements ϕ and $\varphi \in B(\sigma)$, and $\epsilon \in (0, \delta)$, we have that

$$|A(\epsilon, \phi)(\xi) - A(\epsilon, \varphi)(\xi)| \le \rho\|\phi - \varphi\|_{C_0}, \tag{4.17}$$

where $\varphi \in B(\sigma)$ is the ball in C_0 with radius σ and center at the origin.

Proof. It is easy to show that

$$\lim_{\epsilon\to 0^+} \lambda_5 = \frac{-1 - \sqrt{1+4d_1\epsilon}}{2\epsilon d_1} = -\infty, \qquad \lim_{\epsilon\to 0^+} \lambda_6 = \frac{-1 + \sqrt{1+4d_1\epsilon}}{2\epsilon d_1} = 1.$$

Thus we have

$$\left|\frac{1}{\sqrt{1+4d_1\epsilon}} \int_{-\infty}^{\xi} e^{\lambda_5(\xi-t)} \left(1 + \mu - \frac{2\mu\tilde{U}_0}{u_c}\right) W_3 dt\right| = O(\epsilon\|W_3\|),$$

$$\left|\int_{\xi}^{+\infty} \left(\frac{e^{\lambda_6(\xi-t)}}{\sqrt{1+4d_1\epsilon}} - e^{(\xi-t)}\right) \left(1 + \mu - \frac{2\mu\tilde{U}_0}{u_c}\right) W_3 dt\right| = O(\epsilon\|W_3\|),$$

$$\left|\frac{1}{\sqrt{1+4d_1\epsilon}} \int_{-\infty}^{\xi} e^{\lambda_5(\xi-t)} R_{31} dt + \frac{1}{\sqrt{1+4d_1\epsilon}} \int_{\xi}^{\infty} e^{\lambda_6(\xi-t)} R_{31} dt\right| = O(\|W_3\|^2),$$

and

$$\left|\frac{1}{\sqrt{1+4d_1\epsilon}}\int_{-\infty}^{\xi} e^{\lambda_5(\xi-t)}R_{32}dt+\frac{1}{\sqrt{1+4d_1\epsilon}}\int_{\xi}^{\infty} e^{\lambda_6(\xi-t)}R_{32}dt\right|=O(\epsilon).$$

It is easy to see that

$$\begin{aligned}|R_{33}|&=\left|-\epsilon\chi(\tilde{U}_0+W_3)(1-\tilde{U}_0-W_3)\left[\alpha(\mu-\beta)\tilde{U}_0-\frac{\alpha\mu}{u_c}\tilde{U}_0^2+\beta^2\tilde{V}_0\right]\right|\\&=O(\epsilon)+O(\epsilon\|W_3\|)+O(\epsilon\|W_3\|^2).\end{aligned}$$

Thus we can also obtain

$$\begin{aligned}&\left|\frac{1}{\sqrt{1+4d_1\epsilon}}\int_{-\infty}^{\xi} e^{\lambda_5(\xi-t)}R_{33}dt+\frac{1}{\sqrt{1+4d_1\epsilon}}\int_{\xi}^{\infty} e^{\lambda_6(\xi-t)}R_{33}dt\right|\\&=O(\epsilon)+O(\epsilon\|W_3\|)+O(\epsilon\|W_3\|^2).\end{aligned}$$

Before estimating R_{34}, we first estimate W_4''. Actually we have

$$\begin{aligned}|W_4''(\xi)|&\le\frac{\lambda_7^2}{\epsilon d_2(\lambda_8-\lambda_7)}\int_{-\infty}^{\xi} e^{\lambda_7(\xi-t)}|(\alpha W_3+R_{41})|dt\\&\quad+\frac{\lambda_8^2}{\epsilon d_2(\lambda_8-\lambda_7)}\int_{\xi}^{\infty} e^{\lambda_8(\xi-t)}|(\alpha W_3+R_{41})|dt+\frac{|(\alpha W_3+R_{41})|}{\epsilon d_2}\\&\le\frac{2|(\alpha W_3+R_{41})|}{\epsilon d_2}.\end{aligned}$$

Note that

$$\begin{aligned}|(\alpha W_3+R_{41})|&=\left|\alpha W_3+\epsilon d_2\left[\beta^2\tilde{V}_0-\alpha\beta\tilde{U}_0+\alpha\mu\tilde{U}_0\left(1-\frac{\tilde{U}_0}{u_c}\right)\right]\right|\\&=O(\epsilon)+O(\|W_3\|),\end{aligned}$$

$$\begin{aligned}|R_{34}|&=|-\epsilon\chi(\tilde{U}_0+W_1)(1-\tilde{U}_0-W_1)W_4''|\\&=O(\epsilon)+O(\epsilon\|W_3\|)+O(\epsilon\|W_3\|^2)+O(\epsilon\|W_3\|^3).\end{aligned}$$

Then we obtain

$$\begin{aligned}&\left|\frac{1}{\sqrt{1+4d_1\epsilon}}\int_{-\infty}^{\xi} e^{\lambda_5(\xi-t)}R_{34}dt+\frac{1}{\sqrt{1+4d_1\epsilon}}\int_{\xi}^{\infty} e^{\lambda_6(\xi-t)}R_{34}dt\right|\\&=O(\epsilon)+O(\epsilon\|W_3\|)+O(\epsilon\|W_3\|^2)+O(\epsilon\|W_3\|^3).\end{aligned}$$

Before estimating R_{35}, we need to estimate W_4'. In fact, we have

$$\begin{aligned}|W_4'(\xi)|&\le\frac{\lambda_7}{\epsilon d_2(\lambda_8-\lambda_7)}\int_{-\infty}^{\xi} e^{\lambda_7(\xi-t)}|(\alpha W_3+R_{41})|dt\\&\quad+\frac{\lambda_8}{\epsilon d_2(\lambda_8-\lambda_7)}\int_{\xi}^{\infty} e^{\lambda_8(\xi-t)}|(\alpha W_3+R_{41})|dt\\&\le\frac{2|(\alpha W_3+R_{41})|}{\sqrt{1+4\epsilon d_2}}=O(\epsilon)+O(\|W_3\|),\end{aligned}$$

and therefore it follows that

$$|R_{35}| = \left| -\epsilon\chi(1 - 2\tilde{U}_0 - 2W_3)(W_4' - \alpha\tilde{U}_0 + \beta\tilde{V}_0)\left(\frac{\mu\tilde{U}_0^2}{u_c} - \mu\tilde{U}_0\right)\right| \\ \leq O(\epsilon) + O(\epsilon\|W_3\|) + O(\epsilon\|W_3\|^2),$$

and

$$\left|\frac{1}{\sqrt{1+4d_1\epsilon}}\int_{-\infty}^{\xi} e^{\lambda_5(\xi-t)}R_{35}dt + \frac{1}{\sqrt{1+4d_1\epsilon}}\int_{\xi}^{\infty} e^{\lambda_6(\xi-t)}R_{35}dt\right| \\ = O(\epsilon) + O(\epsilon\|W_3\|) + O(\epsilon\|W_3\|^2).$$

From (4.12), we can get

$$R_2(W_3, \xi) = O(\epsilon\|W_3\|) + O(\epsilon\|W_3\|^2) + O(\epsilon\|W_3\|^3).$$

Finally, we arrive at

$$\|A(\epsilon, W_3)(\xi)\| = O(\epsilon) + O(\epsilon\|W_3\|) + O(\|W_3\|^2) + O(\epsilon\|W_3\|^3).$$

For the second part of the lemma, the inequality (4.17) follows from (4.16) directly. This completes the proof.

Now, we are in a position to present and prove the main result.

Theorem 8. *There is a constant $c^* > 0$ so that, for any constant $c > c^*$, system (4.1) has a traveling wave solution $(U(\xi), V(\xi))$, $\xi = \frac{x-ct}{c}$, connecting the two equilibria $(0, 0)$ and $(u_c, \alpha u_c/\beta)$. Moreover, the wave profile $(U(\xi), V(\xi))$ converges to the profile $(\tilde{U}_0(\xi), \tilde{V}_0(\xi))$ when the wave speed $c \to +\infty$.*

Proof. The proof is similar to that of Theorem 6, but for the benefit of the reader, we will give the details.

Define an operator $T_2 : \Psi \in C^1 \to C$ as

$$[T_2\Psi](\xi) = \Psi'(\xi) - \left(\frac{2\mu\tilde{U}_0}{u_c} - \mu\right)\Psi(\xi). \tag{4.18}$$

The formal adjoint equation of $T_2\Psi = 0$ is given by

$$\Phi'(\xi) + \left(\frac{2\mu\tilde{U}_0}{u_c} - \mu\right)\Phi(\xi) = 0, \tag{4.19}$$

subject to the boundary condition $\Phi(\pm\infty) = 0$.

We now divide our proof into five steps.

Step 1. We claim that if $\Phi(\xi) \in C^1$ is a solution of (4.19), then $\Phi = 0$. Moreover, we have $R(T_2) = C$, where $R(T_2)$ is the range space of T_2. Indeed, $\tilde{U}_0 \to u_c$ and $\tilde{V}_0 \to \frac{\alpha u_c}{\beta}$ hold when $\xi \to -\infty$. Equation (4.19) is therefore asymptotic to equation $\Phi'(\xi) + \mu\Phi(\xi) = 0$ with solution $\Phi(\xi) = c_1 e^{-\mu\xi}$, where c_1 is a constant. Therefore, any solution of (4.19) other than

zero must grow exponentially as $\xi \to -\infty$. Then the bounded solution to (4.19) is the zero solution. By the Fredholm theorem, we have $R(T_2) = C$.

Step 2. We conclude that if Ψ is a bounded solution to equation $T_2\Psi = \Theta$, $\Theta \in C_0$, then we have $\Psi \in C_0$. In fact, when $\xi \to -\infty$, this equation asymptotically reduces to $\Psi' = \mu\Psi$, while when $\xi \to +\infty$, equation $T_2\Psi = \Theta$ tends to $\Psi' = -\mu\Psi$. In both cases, we know that any bounded solution of $T_2\Psi = \Theta$ satisfies

$$\lim_{\xi\to\pm\infty} \Psi(\xi) = 0.$$

Hence the claim of Step 2 holds.

Step 3. For the linear operator $L_2 : C_0 \to C_0$ defined by

$$L_2(W_3)(\xi) = W_3(\xi) - \int_{\xi}^{+\infty} e^{(\xi-t)} \left(1 + \mu - \frac{2\mu\tilde{U}_0}{u_c}\right) W_3 dt,$$

we want to prove that $R(L_2)$, the range space of L_2, is equal to C_0; that is, for each $Z_2 \in C_0$ we have a $W_3 \in C_0$ so that

$$W_3(\xi) - \int_{\xi}^{+\infty} e^{(\xi-t)} \left(1 + \mu - \frac{2\mu\tilde{U}_0}{u_c}\right) W_3 dt = Z_2(\xi).$$

To see this, we assume that $\eta_2 = W_3 - Z_2$ and obtain a system for $\eta_2(\xi)$ as follows:

$$\eta_2(\xi) = \int_{\xi}^{+\infty} e^{(\xi-t)} \left(1 + \mu - \frac{2\mu\tilde{U}_0}{u_c}\right) (\eta_2 + Z_2)\, dt.$$

Differentiating both sides yields

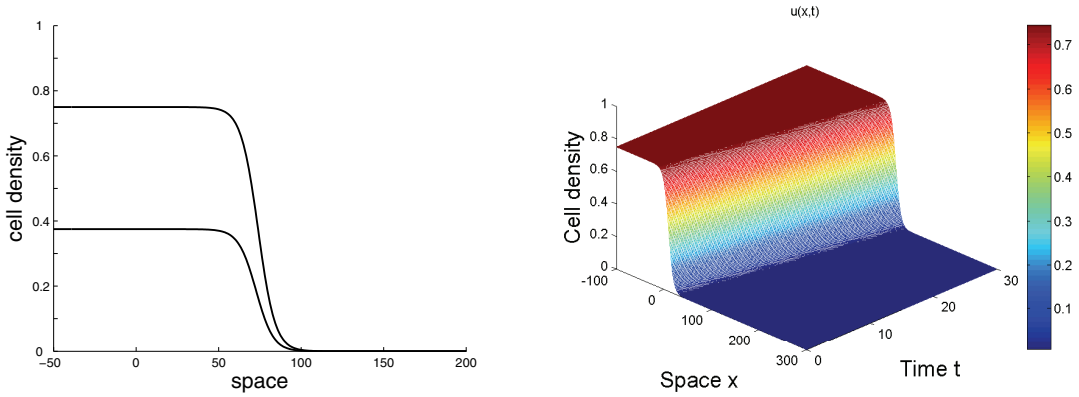
$$\eta_2'(\xi) - \left(\frac{2\mu\tilde{U}_0}{u_c} - \mu\right) \eta_2(\xi) = \left(-1 - \mu + \frac{2\mu\tilde{U}_0}{u_c}\right) Z_2(\xi). \tag{4.20}$$

Using the results that $R(T_2) = C$ in Step 1 and $Z_2 \in C_0$, we obtain that there exists $\eta_2(\xi)$ satisfying (4.20) and $\eta_2(\pm\infty) = 0$. Returning to the variable W_3, we have $W_3 = \eta_2 + Z_2 \in C_0$.

Step 4. Let $N(L_2)$ be the null space of operator L_2 and $N^{\perp}(L_2)$ be the subspace in C_0 so that $C_0 = N^{\perp}(L_2) \oplus N(L_2)$. $N^{\perp}(L_2)$ is a Banach space. If we define $S_2 = L_2|_{N^{\perp}(L_2)}$ as the restriction operator of L_2 on $N^{\perp}(L_2)$, then the map $S_2 : N^{\perp}(L_2) \to C_0$ is one to one and onto. Again using the Banach inverse operator theorem, we have that $S_2^{-1} : C_0 \to N^{\perp}(L_2)$ is a bounded linear operator.

Step 5. When L_2 is restricted to $N^{\perp}(L_2)$, we have $S_2(W_3)(\xi) = A(\epsilon, \xi)$. Define $F_2(\xi, W_3) = S_2^{-1}(A(\epsilon, \xi))$, and (4.14) can be rewritten as $W_3 = F_2(\xi, W_3) = S_2^{-1}(A(\epsilon, \xi))$. Using arguments similar to those in Step 5 of Theorem 6, it follows that there exist constants $\delta, \sigma > 0$ and $0 < \rho < 1$ such that for $W, \varphi, \phi \in B(\sigma) \cap N^{\perp}(L_2)$ and $\epsilon \in [0, \delta)$, we have the estimate

$$\begin{aligned}
\|F_2(\xi, W_3)\| &= \|S_2^{-1}\| \|A(\epsilon, W_3)(\xi)\| \\
&= \|S_2^{-1}\| \left[O(\epsilon) + O(\epsilon\|W_3\|) + O(\|W_3\|^2 + O(\epsilon\|W_3\|^3)\right] \\
&= \|S_2^{-1}\| \left[O(\epsilon) + \|W_3\| \left(O(\epsilon) + O(\|W_3\|) + O(\epsilon\|W_3\|^2)\right)\right] \\
&\le \frac{1}{3}(\|W_3\| + \sigma) \le \sigma
\end{aligned}$$



(a) (b)

Figure 1. (a) *An illustration of traveling wave solution* $u(x,t)$ *(upper line) and* $v(x,t)$ *(lower line) at* $t = 25$*, where* $\xi = 0.2$*,* $\chi = 1$*, and* $\mu = 0.5$*. According to* (2.7)*, the wave speed is around* $c \approx 2.55$*.* (b) *Cell density solution* $u(x,t)$ *for time period* $t \in [0, 30]$ *with the same parametric selection and initial conditions* (5.2)*. The solution stabilizes to a wavefront with speed* $c \approx 2.55$*.*

and

$$\|F_2(\xi, \Phi) - F_2(\xi, \Psi)\| \leq \rho \|\Phi - \Psi\|.$$

Hence, $F_2(\xi, \Phi)$ is a uniform contractive mapping for $W_3 \in B(\sigma) \cap N^{\perp}(L_2)$. By using the Banach contraction principle, it follows that for $\epsilon \in [0, \delta)$ (or $c > c^* = \frac{1}{\sqrt{\delta}}$), (4.14) has a unique solution $W_3 \in B(\sigma) \cap N^{\perp}(L_2)$. Returning to the original variable, we see that $\tilde{U}_0 + W_3$ is a solution connecting the two values 0 and u_c. Thus, according to (4.6), we obtain the formula for $\tilde{V}_0 + W_4$. It is easy to know that when $\varepsilon \to 0$, (W_3, W_4) tends to zero. This completes the proof.

5. Simulations and discussion.

5.1. Simulations. In this section, in order to verify and demonstrate the effectiveness of the proposed theoretical results obtained in both sections 3 and 4, we perform and discuss numerical calculations by using MATLAB.

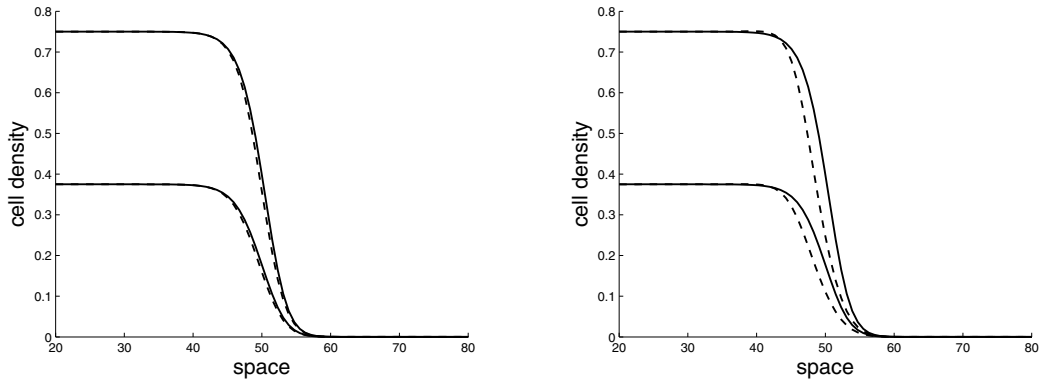
First we simulate the wave patterns by the PDE system. Consider the system

$$\begin{cases} u_t = \frac{\partial}{\partial x}\left(d_1 \frac{\partial u}{\partial x} - \chi u(1-u)\frac{\partial v}{\partial x}\right) + \mu u\left(1 - \frac{u}{u_c}\right), \\ v_t = d_2 \frac{\partial^2 v}{\partial x^2} + \alpha u - \beta v, \end{cases} \tag{5.1}$$

with parameters $d_1 = 0.25$, $d_2 = 1$, $u_c = 0.75$, $\alpha = 1$, and $\beta = 2$ and initial condition

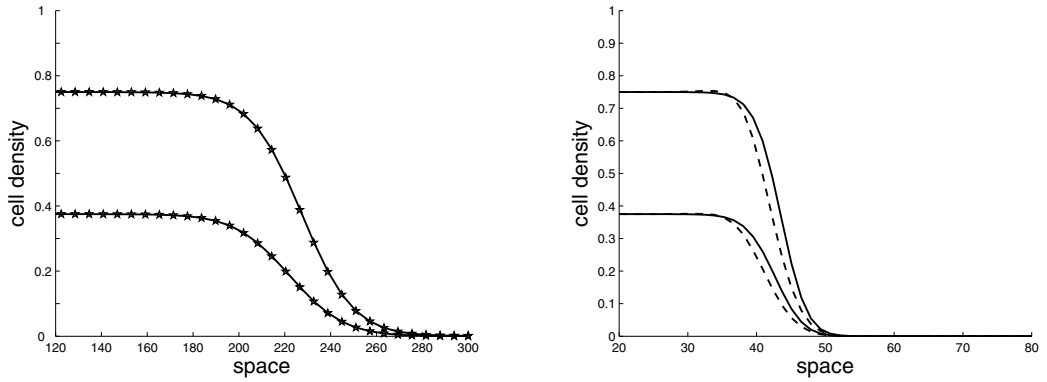
$$\begin{pmatrix} u(x,0) \\ v(x,0) \end{pmatrix} = \begin{pmatrix} \frac{0.75}{1+e^{\xi(x-10)}} \\ \frac{0.75}{2(1+e^{\xi(x-10)})} \end{pmatrix}, \tag{5.2}$$

where the constant ξ is the decay rate of initial wavefront. For particular values $\xi = 0.2$, $\chi = 1$, and $\mu = 0.5$, the numerical solutions are shown in Figure 1. Here the spreading speed



(a) (b)

Figure 2. *Solution $u(x,t)$ (upper lines) and $v(x,t)$ (lower lines) at $t = 30$. The wave (dashed line) represents the solution of nonlinear diffusion, and the solid line represents the wavefront of linear diffusion, where $\mu = 0.5$ and $\xi = 1$. The wave speed $c \approx 0.75$ for both cases. The chemotactic sensitivity $\chi = 1$ in* (a) *and $\chi = 5$ in* (b).



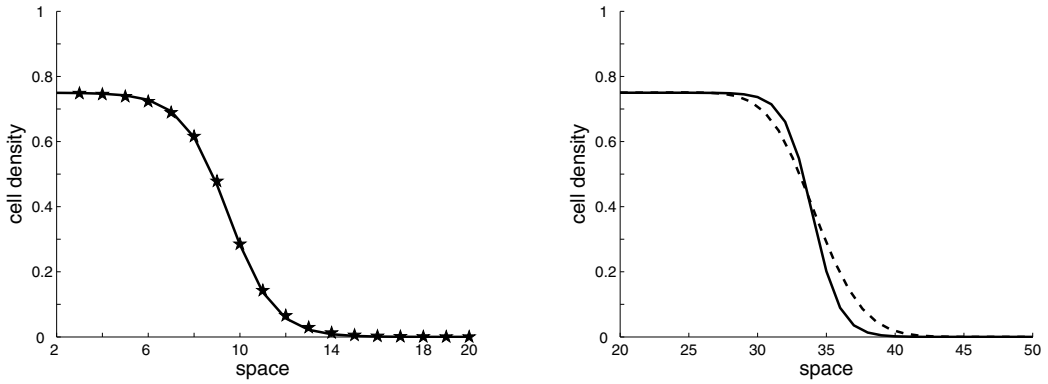
(a) (b)

Figure 3. *Solution $u(x,t)$ (upper lines) and $v(x,t)$ (lower lines) at $t = 25$. The solid line represents the wavefront of linear diffusion for $u(x,t)$, with $\chi = 0$.* (a) *$\chi = 0$, $\xi = 0.1$ (solid) and $\chi = 9$, $\xi = 0.1$ (starred); and in Figure 3*(b)*, $\chi = 0$, $\xi = 1$ (solid) and $\chi = 9$, $\xi = 1$ (dashed).*

agrees with (2.7); that is,

$$c = d_1\xi + \frac{\mu}{\xi} \quad \text{when } \xi < \sqrt{\frac{\mu}{d_1}}.$$

We shall compare the leading-term traveling waves ($\chi = 0$ or $c = \infty$) with the real waves. This comparison involves two cases: (1) $\chi = 0$ and $\chi \neq 0$; (2) $c = \infty$ and c is finite. The first case is presented in Figures 2 and 3. In Figure 2 we take (a) $\chi = 0$ (solid) and $\chi = 1$ (dashed); (b) $\chi = 0$ (solid) and $\chi = 5$ (dashed) with the same decay rate $\xi = 1$. In Figure 3, we take (a) $\chi = 0$, $\xi = 0.1$ (solid) and $\chi = 9$, $\xi = 0.1$ (starred); and in Figure 3(b) $\chi = 0$, $\xi = 1$ (solid) and $\chi = 9$, $\xi = 1$ (dashed).



(a) (b)

Figure 4. *Solid line represents the leading-term wave profile U.* (a) *Starred line represents true solution with $\chi = 5$, $\mu = 1$ and wave speed $c = 10$.* (b) *Dashed line represents the real solution with wave speed $c = 1.25$.*

For the second case, we carry out the simulations by using system (2.2). We apply the finite-difference method to carry out the numerical scheme and solve the ODE problem (boundary value problem (BVP)). See Figure 4(a) for $c = 10$ (starred) and $c = \infty$ (solid), and Figure 4(b) for $c = 1.25$ (dashed) and $c = \infty$ (solid).

The nonlinear finite difference method is shown below. Rewrite system (2.2) as

$$
\begin{cases}
u'' = f(u, u', v', v'') = \frac{\chi}{d_1} u(1-u)v'' + \frac{\chi}{d_1}(1-2u)u'v' - \frac{c}{d_1}u' - \frac{\mu}{d_1}u(1-u/u_c), \\
v'' = g(u, v, v') = \frac{-c}{d_2}v' - \frac{\alpha}{d_2}u + \frac{\beta}{d_2}v, \\
u'(-\infty) = 0, \ u'(+\infty) = 0, \\
v'(-\infty) = 0, \ v'(+\infty) = 0.
\end{cases} \tag{5.3}
$$

Discretize u into $w_1, w_2, \ldots, w_n$ and v into $w_{n+1}, w_{n+2}, \ldots, w_{2n}$ by step size h. Then the $2n \times 2n$ nonlinear system obtained from (5.3) is the system

$$
\begin{aligned}
-w_2 + w_1 + h^2 f\left(w_1, \frac{w_2 - w_1}{2h}, \frac{w_{n+2} - w_{n+1}}{2h}, \frac{w_{n+2} - w_{n+1}}{h^2}\right) &= 0, \\
-w_3 + 2w_2 - w_1 + h^2 f\left(w_2, \frac{w_3 - w_1}{2h}, \frac{w_{n+3} - w_{n+1}}{2h}, \frac{w_{n+3} - 2w_{n+2} + w_{n+1}}{h^2}\right) &= 0, \\
&\vdots \\
-w_n + 2w_{n-1} - w_{n-2} + h^2 f\left(w_{n-1}, \frac{w_n - w_{n-2}}{2h}, \frac{w_{2n} - w_{2n-2}}{2h}, \frac{w_{2n} - 2w_{2n-1} + w_{2n-2}}{h^2}\right) &= 0, \\
w_n - w_{n-1} + h^2 f\left(w_n, \frac{w_n - w_{n-1}}{2h}, \frac{w_{2n} - w_{2n-1}}{2h}, \frac{-w_{2n} + w_{2n-1}}{h^2}\right) &= 0,
\end{aligned}
$$

coupled with

$$
-w_{n+2} + w_{n+1} + h^2 g\left(w_1, w_{n+1}, \frac{w_{n+2} - w_{n+1}}{2h}\right) = 0,
$$

$$
\begin{aligned}
-w_{n+3} + 2w_{n+2} - w_{n+1} + h^2 g\left(w_2, w_{n+2}, \frac{w_{n+3} - w_{n+1}}{2h}\right) &= 0,\\
&\vdots\\
-w_{2n} + 2w_{2n-1} - w_{2n-2} + h^2 g\left(w_{n-1}, w_{2n-1}, \frac{w_{2n} - w_{2n-2}}{2h}\right) &= 0,\\
w_{2n} - w_{2n-1} + h^2 g\left(w_n, w_{2n}, \frac{w_{2n} - w_{2n-1}}{2h}\right) &= 0.
\end{aligned}
$$

Using the Newton method for the nonlinear system, the Jacobian matrix derived from the above $2n \times 2n$ system can be written in a 2×2 block

$$
J_{Jacobian} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}_{2n \times 2n}, \tag{5.4}
$$

where A, B, and D are $n \times n$ tridiagonal matrices, and C is an $n \times n$ diagonal one:

$$
A = \begin{pmatrix}
1 + h^2 f_u - \frac{h}{2} f_{u'} & -1 + \frac{h}{2} f_{u'} & 0 & \cdots & 0 \\
-1 - \frac{h}{2} f_{u'} & 2 + h^2 f_u & -1 + \frac{h}{2} f_{u'} & \ddots & \vdots \\
0 & \ddots & \ddots & \ddots & 0 \\
\vdots & \ddots & \ddots & \ddots & -1 + \frac{h}{2} f_{u'} \\
0 & \cdots & 0 & -1 - \frac{h}{2} f_{u'} & 1 + h^2 f_u + \frac{h}{2} f_{u'}
\end{pmatrix}, \tag{5.5}
$$

$$
B = \begin{pmatrix}
-\frac{h}{2} f_{v'} - f_{v''} & \frac{h}{2} f_{v'} + f_{v''} & 0 & \cdots & 0 \\
-\frac{h}{2} f_{v'} + f_{v''} & -2 f_{v''} & \frac{h}{2} f_{v'} + f_{v''} & \ddots & \vdots \\
0 & \ddots & \ddots & \ddots & 0 \\
\vdots & \ddots & \ddots & \ddots & \frac{h}{2} f_{v'} + f_{v''} \\
0 & \cdots & 0 & -\frac{h}{2} f_{v'} + f_{v''} & \frac{h}{2} f_{v'} - f_{v''}
\end{pmatrix}, \tag{5.6}
$$

$$
C = \begin{pmatrix}
h^2 g_u & 0 & \cdots & \cdots & 0 \\
0 & \ddots & & & \vdots \\
\vdots & & \ddots & & \vdots \\
\vdots & & & \ddots & 0 \\
0 & \cdots & \cdots & 0 & h^2 g_u
\end{pmatrix}, \tag{5.7}
$$

and

$$
D = \begin{pmatrix}
1 + h^2 g_v - \frac{h}{2} g_{v'} & -1 + \frac{h}{2} g_{v'} & 0 & \cdots & 0 \\
-1 - \frac{h}{2} g_{v'} & 2 + h^2 g_v & -1 + \frac{h}{2} g_{v'} & \ddots & \vdots \\
0 & \ddots & \ddots & \ddots & 0 \\
\vdots & \ddots & \ddots & \ddots & -1 + \frac{h}{2} g_{v'} \\
0 & \cdots & 0 & -1 - \frac{h}{2} g_{v'} & 1 + h^2 g_v + \frac{h}{2} g_{v'}
\end{pmatrix}. \tag{5.8}
$$

Using the Newton–Raphson method, we set the error tolerance within 0.001 and obtain Figure 4 after four time iterations.

Our observation from the above numerical simulations is that, when c is large or χ is small, the leading-term wave matches very well with the real waves.

5.2. Discussion about periodic solutions. Now we discuss the existence of periodic waves for system (2.2). Linearizing it around the positive steady-state $(u_c, \alpha u_c/\beta)$, we get the system

$$\begin{pmatrix} U' \\ X' \\ V' \\ Y' \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ \frac{\alpha\chi u_c(u_c-1)+\mu d_2}{d_1 d_2} & -\frac{c}{d_1} & \frac{\beta\chi u_c(1-u_c)}{d_1 d_2} & -\frac{c\chi u_c(1-u_c)}{d_1 d_2} \\ 0 & 0 & 0 & 1 \\ -\frac{\alpha}{d_2} & 0 & \frac{\beta}{d_2} & -\frac{c}{d_2} \end{pmatrix} \begin{pmatrix} U \\ X \\ V \\ Y \end{pmatrix}, \tag{5.9}$$

and the characteristic equation is given by

$$\lambda^4 + \left(\frac{c}{d_1} + \frac{c}{d_2}\right)\lambda^3 + \left(\frac{c^2}{d_1 d_2} - \frac{\chi\alpha u_c(u_c-1)+\mu d_2}{d_1 d_2} - \frac{\beta}{d_2}\right)\lambda^2 - \frac{(\beta+\mu)c}{d_1 d_2}\lambda + \frac{\mu\beta}{d_1 d_2} = 0. \tag{5.10}$$

We suppose that there exist periodic solutions near the positive equilibria and assume that the characteristic equation has purely imaginary roots $\lambda_{1,2} = \pm\omega i$. Plugging them into (5.10) gives a necessary condition $c = 0$. In this case, (5.10) reduces to

$$\lambda^4 - \left(\frac{\chi\alpha u_c(u_c-1)+\mu d_2}{d_1 d_2} + \frac{\beta}{d_2}\right)\lambda^2 + \frac{\mu\beta}{d_1 d_2} = 0. \tag{5.11}$$

Further analysis of (5.11) on the existence of purely imaginary roots requires that

$$\chi > \frac{d_1\beta + \mu d_2}{\alpha u_c(1-u_c)}. \tag{5.12}$$

This means that when $c = 0$ and χ is sufficiently large, system (2.2) may have periodic waves near the positive equilibrium $(u_c, \frac{\alpha u_c}{\beta})$. Actually, the following perturbation analysis indicates the existence of a periodic standing wave. Let $\theta = \frac{z}{\sqrt{\chi}}$; the system (2.2) becomes

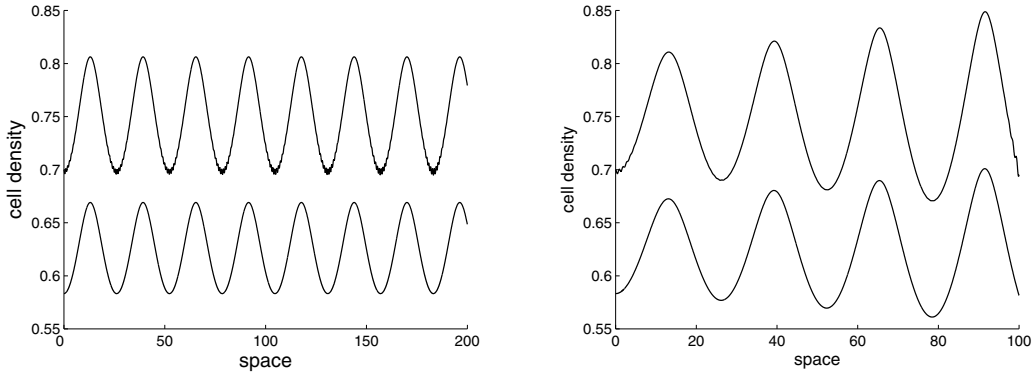
$$\begin{cases} \frac{d_1}{\chi}U'' - (1-2U)U'V' - U(1-U)V'' + \mu U\left(1 - \frac{U}{u_c}\right) = 0, \\ \frac{d_2}{\chi}V'' + \alpha U - \beta V = 0, \end{cases} \tag{5.13}$$

where $U' = \frac{dU}{d\theta}$ and $V' = \frac{dV}{d\theta}$. As χ goes to infinity, the leading term of system (5.13) is given by

$$\begin{cases} -(1-2U)U'V' - U(1-U)V'' + \mu U\left(1 - \frac{U}{u_c}\right) = 0, \\ \alpha U - \beta V = 0. \end{cases}$$

We thus get the following nonlinear second order ODE for U:

$$\big(U(1-U)U'\big)' - \frac{\mu\beta}{\alpha}U\left(1 - \frac{U}{u_c}\right) = 0. \tag{5.14}$$



(a) (b)

Figure 5. (a) *An illustration of periodic solution U (upper line) and V (lower line), where wave speed $c = 0$, $\chi = 120$, $u_c = 0.75$, $\mu = 1$, $d_1 = 0.5$, $\alpha = 1$, and $\beta = 1.2$, with the initial value $U(0) = 0.7$, $V(0) = \frac{7}{12}$.* (b) *An illustration of solutions U (upper line) and V (lower line), where the wave speed $c = 0.1$.*

Let $W = U^2(\frac{1}{2} - \frac{U}{3})$. When U is in the interval $[0, 1]$, $W(U)$ is an increasing function; this gives an inverse function $U(W)$. Therefore, we have

$$W'' - \frac{\mu\beta}{\alpha} U(W) \left(1 - \frac{U(W)}{u_c}\right) = 0.$$

This is a conservative system which apparently exhibits periodic solutions.

When the wave speed is not equal to zero, even if very small, say, $c = 0.1$, the periodic solutions disappear. Figure 5 illustrates how the process works with (a) $c = 0$ and (b) $c = 0.1$.

6. Conclusions. In this paper, we study the existence of the traveling wavefronts for the volume-filling model (1.2) proposed by Hillen and Painter in 2001. The Banach fixed-point theorem is successfully applied to find the existence either when the chemotaxis sensitivity constant χ is small or when the wave speed c is large. Numerical simulations are conducted to confirm our theoretical results, and the comparison between the real waves and the leading-term approximations shows that they agree with each other very well. It has been understood that the existence of positive wavefronts depends on the spreading speed c. In this paper the estimate of the minimal speed is carried out by the standard stability analysis (although not rigorous, but widely used). This gives

$$c \geq c_{\min} = 2\sqrt{d_1 \mu};$$

see also (2.5). The selection of the spreading speed of waves from the initial data is studied by the asymptotic analysis near the wavefront ($x = \infty$), and we obtain

$$c = \begin{cases} d_1\xi + \frac{\mu}{\xi} & \text{if } 0 < \xi < \sqrt{\frac{\mu}{d_1}}, \\ 2\sqrt{d_1\mu} & \text{if } \xi \geq \sqrt{\frac{\mu}{d_1}} \end{cases}$$

if the initial data are taken as

$$\begin{cases} u(x,0) \sim Ae^{-\xi x}, \\ v(x,0) \sim Be^{-\xi x} \end{cases}$$

as $x \to \infty$. This result is interesting and it is confirmed by our simulations.

A brief discussion on the existence of positive periodic waves is presented, and we find that these waves can exist only when the speed c is equal to zero; that is, we can have only the existence of standing periodic patterns; see Figure 5.

For the volume-filling model (1.2) in a bounded domain subject to the Neumann boundary condition, the existence of positive steady-states, pattern formation, and/or the metastable behavior have been studied in [24, 25, 37] when the kinetics $f(u,v) = 0$, and in [35, 38] when f obeys the logistic law $f = \mu u(1 - u/u_c)$; see also the latest review paper [10]. In the latter case, theoretical analysis about the existence of slow motion (or metabehavior) of nonconstant positive patterns seems interesting and challenging and deserves further investigation. This will be a subject of our future study.

Acknowledgments. The authors thank Dr. Zhian Wang for suggesting the study of volume-filling chemotaxis models. This topic proved to be a fruitful one, well worthy of investigation, and we are very grateful to him. We also thank the referees for their valuable comments.

REFERENCES

[1] J. Adler, *Chemotaxis in bacteria*, Ann. Rev. Biochem., 44 (1975), pp. 341–356.
[2] H. C. Berg, *Chemotaxis in bacteria*, Ann. Rev. Biophys. Bioeng., 4 (1975), pp. 119–136.
[3] J. T. Bonner, *The Cellular Slime Moulds*, 2nd ed., Princeton University Press, Princeton, NJ, 1967.
[4] J. Canosa, *On a nonlinear diffusion equation describing population growth*, IBM J. Res. Develop., 17 (1973), pp. 307–313.
[5] J. C. Dallon and H. G. Othmer, *A discrete cell model with adaptive signalling for aggregation of Dictyostelium discoideum*, Phil. Trans. R. Soc. Lond., B325 (1997), pp. 391–417.
[6] P. N. Devreotes, *Dictyostelium discoideum: A model system for cell-cell interactions in development*, Science, 245 (1989), pp. 1054–1058.
[7] T. Faria, W. Huang, and J. Wu, *Travelling waves for delayed reaction-diffusion equations with global response*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 462 (2006), pp. 229–261.
[8] S. Gueron and N. Liron, *A model of herd grazing as a travelling wave, chemotaxis and stability*, J. Theoret. Biol., 27 (1989), pp. 595–608.
[9] T. Hillen and K. J. Painter, *Global existence for a parabolic chemotaxis model with prevention of overcrowding*, Adv. in Appl. Math., 26 (2001), pp. 280–301.
[10] T. Hillen and K. J. Painter, *A user's guide to PDE models for chemotaxis*, J. Math. Biol., 58 (2009), pp. 183–217.
[11] M. Holz and S.-H. Chen, *Spatio-temporal structure of migrating chemotactic band of Escherichia coli. I. Traveling band profile*, Biophys. J., 26 (1979), pp. 243–26l.
[12] T. Hofer, J. Sherratt, and P. K. Maini, *Cellular pattern formation during Dictyostelium aggregation*, Phys. D, 85 (1995), pp. 425–444.
[13] D. Horstmann and A. Stevens, *A constructive approach to traveling waves in chemotaxis*, J. Nonlinear Sci., 14 (2004), pp. 1–25.
[14] E. F. Keller and G. M. Odell, *Necessary and sufficient conditions for chemotactic bands*, Math. Biosci., 27 (1975), pp. 309–317.
[15] E. F. Keller and L. A. Segel, *Initiation of slime mold aggregation viewed as an instability*, J. Theoret. Biol., 26 (1970), pp. 399–415.

[16] E. F. Keller and L. A. Segel, *Traveling bands of chemotactic bacteria: A theoretical analysis*, J. Theoret. Biol., 30 (1971), pp. 235–248.
[17] J. L. Martiel and A. Goldbeter, *A model based on receptor desensitization for cyclic AMP signaling in Dictyostelium cells*, Biophys. J., 52 (1987), pp. 807–828.
[18] D. Mollison, *Spatial contact models for ecological and epidemic spread*, J. Roy. Statist. Soc. Ser. B, 39 (1977), pp. 283–326.
[19] P. B. Monk and H. G. Othmer, *Cyclic AMP oscillations in suspensions of Dictoystelium discoideum*, Proc. Roy. Soc. Lond. B, 240 (1989), pp. 555–589.
[20] J. D. Murray, *Mathematical Biology,* I *and* II, Springer-Verlag, New York, 2002.
[21] T. Nagai and T. Ikeda, *Travelling waves in a chemotactic model*, J. Math. Biol., 30 (1991), pp. 169–184.
[22] G. M. Odell and E. F. Keller, *Travelling bands of chemotactic bacteria revisited*, J. Theoret. Biol., 56 (1976), pp. 243–247.
[23] C. Ou and J. Wu, *Traveling wavefronts in a delayed food-limited population model*, SIAM J. Math. Anal., 39 (2007), pp. 103–125.
[24] K. J. Painter and T. Hillen, *Volume-filling and quorum-sensing in models for chemosensitive movement*, Can. Appl. Math. Q., 10 (2002), pp. 501–542.
[25] A. B. Potapov and T. Hillen, *Metastability in chemotaxis models*, J. Dynam. Differential Equations, 17 (2005), pp. 293–330.
[26] C. S. Patlak, *Random walk with persistence and external bias*, Bull. Math. Biophys., 15 (1953), pp. 311–338.
[27] C. S. Patlak, *A mathematical contribution to the study of orientation of organisms*, Bull. Math. Biophys., 15 (1953), pp. 431–476.
[28] G. Rosen and S. Baloga, *On the stability of steady propagating bands of chemotactic bacteria*, Math. Biosci., 24 (1975), pp. 273–279.
[29] G. Rosen, *On the propagation theory for bands of chemotactic bacteria*, Math. Biosci., 20 (1974), pp. 185–189.
[30] H. R. Schwetlick, *Traveling waves for chemotaxis-systems*, Proc. Appl. Math. Mech., 3 (2003), pp. 476–478.
[31] T. L. Scribner, L. A. Segel, and E. H. Rogers, *A numerical study of the formation and propagation of traveling bands of chemotactic bacteria*, J. Theoret. Biol., 46 (1976), pp. 189–219.
[32] L. A. Segel, *Mathematical Models in Molecular and Cellular Biology*, Cambridge University Press, Cambridge, UK, 1980.
[33] R. Tyson, S. R. Lubkin, and J. D. Murray, *A minimal mechanism for bacterial pattern formation*, P. Roy. Soc. Lond. B Bio., 266 (1999), pp. 299–304.
[34] R. Tyson, S. R. Lubkin, and J. D. Murray, *Model and analysis of chemotactic bacterial patterns in a liquid medium*, J. Math. Biol., 38 (1999), pp. 359–375.
[35] Z. Wang and T. Hillen, *Classical solutions and pattern formation for a volume filling chemotaxis model*, Chaos, 17 (2007), 037108.
[36] D. E. Woodward, R. Tyson, M. R. Myerscough, J. D. Murray, E. O. Budrene, and H. C. Berg, *Spatiotemporal patterns generated by Salmonella typhimurium*, Biophys. J., 68 (1995), pp. 2181–2189.
[37] D. Wrzosek, *Long-time behaviour of solutions to a chemotaxis model with volume-filling effect*, Proc. Roy. Soc. Edinburgh Sect. A, 136 (2006), pp. 431–444.
[38] D. Wrzosek, *Global attractor for a chemotaxis model with prevention of overcrowding*, Nonlinear Anal., 59 (2004), pp. 1293–1310.
[39] E. Zeidler, *Nonlinear Functional Analysis and Its Applications.* I. *Fixed-Point Theorems*, translated from the German by P. R. Wadsack, Springer-Verlag, New York, 1986.

SIAM J. Applied Dynamical Systems
Vol. 8, No. 1, pp. 417–453
© 2009 Society for Industrial and Applied Mathematics

On Computing the Critical Coupling Coefficient for the Kuramoto Model on a Complete Bipartite Graph*

Mark Verwoerd† and Oliver Mason†

Abstract. We extend recent results on the existence of global phase-locked states (GPLS) in the Kuramoto model on a complete graph to the case of a complete bipartite graph. In particular, we prove that, for the Kuramoto model on a complete bipartite graph, the value of the critical coupling coefficient, i.e., the smallest coupling coefficient for which a GPLS can exist, can be determined by solving a system of two nonlinear equations that do not depend on the coupling coefficient itself. We show that said system of equations can be solved using an efficient algorithm described in the paper.

Key words. coupled oscillators, synchronization, phase-locking, Kuramoto model, bipartite graph

AMS subject classifications. 34C15, 35B32, 37C25

DOI. 10.1137/080725726

1. Introduction.

Coordinated or synchronized behavior is a fundamental aspect of a wide array of physical, biological, and engineering systems. It is therefore hardly surprising that the study of such behavior occupies an important place in the general theory of dynamical systems and that it has attracted the attention of researchers from a broad range of scientific disciplines [30, 4, 26, 10, 28].

From a historical point of view, much of the early quantitative work on synchronization was rooted in physical questions such as the functioning of networks of Josephson junctions [51, 50] or coupled pendula [3]. More recently, with the growth of activity and interest in mathematical and systems biology, there has been considerable interest in analyzing synchronization in biological systems. Examples of synchronized behaviors abound in the life sciences. Some of the more important examples to have attracted attention include circadian rhythms [47, 2, 52], the synchronization of heart cells [23], synchronized neuronal activity [22, 48] with particular emphasis on pathological synchronization in disorders such as Parkinson's disease [13, 46, 45], and flocking behavior in populations of animals, birds, or fish [42, 5]. In addition to advancing our understanding of the world around us, a better understanding of the mechanisms of synchronization can guide the design of modern multiagent engineering systems [27, 15, 41, 40, 37].

Of course, the development and analysis of appropriate mathematical models is crucial to clarifying the mechanisms behind coordinated and synchronized behavior. Several general frameworks and approaches have now been proposed. These include the abstract and elegant

*Received by the editors May 30, 2008; accepted for publication (in revised form) by B. Ermentrout November 10, 2008; published electronically March 6, 2009. This work was supported by Science Foundation Ireland (SFI) grants 03/RP1/I382 and 04/IN1/I478 and by the Higher Education PRTLI Network Mathematics grant.
http://www.siam.org/journals/siads/8-1/72572.html
†Hamilton Institute, National University of Ireland, Maynooth, County Kildare, Ireland (mark.verwoerd@nuim.ie, oliver.mason@nuim.ie).

theory of coupled-cell networks [12, 11], which characterizes the possible spatiotemporal behaviors of a system in terms of symmetries in the network describing the system interactions. Much of the early work in this direction focused on symmetries which could be captured by group structure; however, recently there has been considerable effort made to extend this to the more general setting provided by groupoids, which allows for asymmetric network structures. Another approach is provided by the master stability function, which provides a condition for the local stability of the synchronized manifold in terms of the eigenvalues of the Laplacian of the network.

To date, one of the most widely studied frameworks for the analysis of synchronization is the so-called *Kuramoto model*, which is a mathematical model describing the dynamics of a system of weakly coupled (nonlinear) oscillators [17, 18, 44, 43, 24, 1]. The Kuramoto model has been used in numerous applications in biology, physics, and medicine; more recently it has also attracted attention from the engineering community, specifically from the systems and control community [16, 29, 40, 37, 38, 39]. Indeed, recent work in the context of multiagent systems has revealed an interesting connection between the Kuramoto model and the problem of stabilizing collective motion in the plane [40].

The aspect of the Kuramoto model that has attracted most attention to date is the manner in which the onset of synchronization depends on the strength of the coupling between the oscillators. It has been observed numerically and established formally that a connected network has a particular coupling strength below which no phase-locked states can exist, and that increasing the coupling strength beyond this critical value can induce a phase transition in the order parameter of the system. Indeed, the transition from incoherent to fully coherent behavior is marked by *two* critical values of the coupling strength: the first marking the transition from an incoherent to a partially coherent state; the second marking the transition from a partially coherent to a fully coherent (global phase-locked) state. In this paper we shall be concerned exclusively with the second.

The first analytical results on critical coupling were obtained by Kuramoto. Specifically, he obtained a closed formula for the coupling coefficient marking the transition to partially coherent behavior. This was for a setting with infinitely many oscillators, a symmetric frequency distribution, and all-to-all coupling. Notwithstanding the considerable body of numerical evidence supporting the findings of Kuramoto's work, it can still be argued that Kuramoto's analysis is not entirely rigorous [43]. To develop a more rigorous theory, one possible approach is to start with a network of finite size and then prove a convergence result as the number of oscillators tends to infinity, as suggested by Kopell (see [43]). To the best of the authors' knowledge, such a result has yet to be reported. Recently, the authors of the present paper contributed a theorem [49, Theorem 3] towards such a result. We review this and related work in more detail in Appendix A.

Several authors have obtained estimates for the value of the critical coupling for the case of a general graph. Even if it is not clear in every case how good the respective estimate is, the key insight that has emerged from this work (see, e.g., [32, 34, 33] and [16]) is that spectral properties of the Laplacian or adjacency matrix of the graphs play a key role in determining the value of the critical coupling strength. Indeed, the work described in [32] suggests that the critical coupling scales with the largest eigenvalue of the adjacency matrix, while in [16] a lower bound is given which involves the smallest nonzero eigenvalue of the Laplacian.

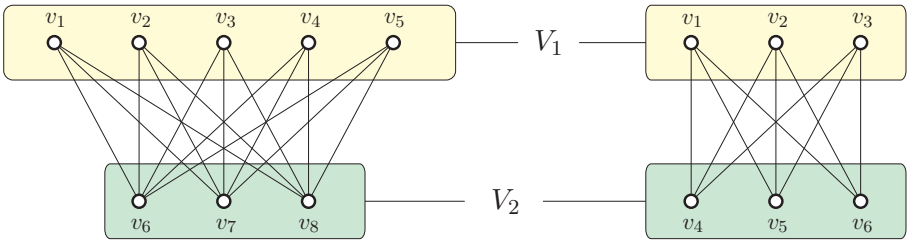
Related work on the Kuramoto model includes the study of desynchronization in finite populations of oscillators, which has been described in considerable detail in [20, 19]. Desynchronization is what takes place when the coupling strength is reduced below its critical value. The authors of the aforementioned work have shown that desynchronization involves a series of *frequency-splitting bifurcations*. At each such bifurcation, the ensemble of oscillators subdivides into smaller and smaller groups of oscillators with identical average frequency, until eventually all oscillators oscillate at their own intrinsic frequency. Another aspect of the Kuramoto model to have attracted attention recently is the emergence of phase chaos [21, 31] in systems of dimension four and higher. A generic feature of coupled oscillator systems, phase chaos in the Kuramoto model is most prominent in systems with relatively low dimension (comprising between ten and fifteen oscillators) [21]. Lastly, we mention closely related recent work on the spectrum and the stability of (partially) locked states [25, 24, 6, 35, 36]. In particular, we note the relation between our own work [49] and results in [24].

In this paper we consider the problem of determining the value of the critical coupling for the Kuramoto model on a complete bipartite graph. The motivation for this work is in no small part mathematical. It turns out that the ideas behind previous work on global phase-locking on complete graphs [49] carry over to the case of complete bipartite graphs, although the actual proofs are very different, as are some of the results. Bipartite graphs have applications in many different fields, including biology, sociology, and engineering, to name but a few [14]. Examples of bipartite graphs include social contact networks, linking people to locations [9] or people to people [8], and metabolic networks, linking compounds (or substrates) to products.

The outline of the paper is as follows. In section 2, we introduce some notation and terminology. In section 3, we introduce the Kuramoto model on the complete bipartite graph and review some of its basic properties. Here, we also give formal definitions of the notion of a global phase-locked state and critical coupling and show that a global phase-locked state always exists for sufficiently strong coupling (essentially proving that the critical coupling is a finite number). In section 4 we derive our main result, which is a condition for criticality. To find the critical coupling we introduce an algorithm which is described in section 5. Section 6 contains some numerical examples to illustrate the results of the paper, and finally, in section 7 we present our concluding remarks.

2. Notation and terminology. Throughout the paper, $\mathbb{R}$ ($\mathbb{C}$) denotes the field of real (complex) numbers, while $\mathbb{R}^n$ ($\mathbb{C}^n$), $\mathbb{R}^{m\times n}$ ($\mathbb{C}^{m\times n}$) denote the vector spaces of n-tuples of real (complex) numbers and the space of $m \times n$ matrices with entries in $\mathbb{R}$ ($\mathbb{C}$), respectively. Frequently throughout the paper, we shall identify $\mathbb{R}^{m+n}$ with $\mathbb{R}^m \times \mathbb{R}^n$ in the obvious fashion. $\mathbb{N}$ denotes the set of positive integers. For $v \in \mathbb{R}^n$ and $1 \le i \le n$, v_i denotes the ith component of v. Also, $\mathbb{R}_{\ge 0}$ denotes the set of all nonnegative real numbers. For any finite set S, $|S|$ denotes the cardinality of S. For any real number x, $|x|$ denotes the absolute value of x, and for a vector $v \in \mathbb{R}^n$, $\|v\|_\infty$ denotes the usual infinity norm given by $\max_{1\le i\le n} |v_i|$.

The graph theoretical terminology and notation adopted here is standard, and for background on basic graph theory, the interested reader should consult [7]. As we shall be concerned only with *undirected* graphs, the term graph shall be used throughout to denote an undirected graph.



(a) $K_{5,3}$ (b) $K_{3,3}$

Figure 1. *Two examples of a complete bipartite graph. As illustrated, the vertex set* $V := \{v_1, \dots, v_8\}$ *of the graph* $K_{5,3} := (V, E)$ *can be divided into subsets* $V_1 := \{v_1, \dots, v_5\}$ *and* $V_2 := \{v_6, v_7, v_8\}$ *such that* $E = V_1V_2 \cup V_2V_1$*. The same holds for* $K_{3,3}$*, whose vertex set can be partitioned into subsets* $V_1 := \{v_1, v_2, v_3\}$ *and* $V_2 := \{v_4, v_5, v_6\}$*.*

Definition 1 (graph). *A graph* G *is an ordered pair* (V, E)*, where* V *is a finite set of vertices and* E *is a set consisting of 2-element subsets of* V*.*

The elements of E are referred to as edges, and we shall usually write uv (or vu) for the edge $\{u, v\}$ and say that u and v are neighbors in G. The *degree* of a vertex $v \in V$ is the number of neighbors of v. A particularly important class of graphs in the context of the Kuramoto model is the class of complete graphs which we define next.

Definition 2 (complete graph). *For any positive integer* $n \geq 2$*, the complete graph* K_n *has* n *vertices* $V = \{v_1, \dots, v_n\}$ *and edges* $E = \{\{v_i, v_j\} : 1 \leq i < j \leq n\}$*.*

So, in K_n the degree of every vertex is equal to $n - 1$. In this paper, our focus shall be on a related class of graphs, namely complete bipartite graphs. First, let us recall the definition of a bipartite graph.

Definition 3 (bipartite graph). *A graph* $G = (V, E)$ *is called* bipartite *if there exists a partition* $V = V_1 \cup V_2$ *of the vertex set so that every edge in* E *is of the form* v_1v_2 *for some* $v_1 \in V_1$ *and* $v_2 \in V_2$*.*

Definition 4 (complete bipartite graph). *For positive integers* $m, n \geq 1$*, the complete bipartite graph* $K_{m,n} = (V_1 \cup V_2, E)$ *is a bipartite graph such that* $|V_1| = m$*,* $|V_2| = n$ *and for any two vertices* $v_1 \in V_1$ *and* $v_2 \in V_2$*,* $v_1v_2 \in E$*.*

Two examples of complete bipartite graphs are given in Figure 1.

3. Global phase-locking on a complete bipartite graph.

Let $(m, n) \in \mathbb{N} \times \mathbb{N}$ be given. We consider a system of coupled oscillators on the complete bipartite graph $K_{m,n} = (V_1 + V_2, E)$. Without loss of generality we assume that the vertices $\{v_i\}$ are labeled in such a way that $v_1, v_2, \dots, v_m$ belong to vertex set V_1, and $v_{m+1}, v_{m+2}, \dots, v_{m+n}$ belong to vertex set V_2. We label oscillators according to their association with a vertex; i.e., we write oscillator i for the oscillator associated with vertex v_i. The coupling dynamics given by the Kuramoto model are as follows:

$$\dot{\theta}_i(t) = \omega_i + \frac{k}{m+n}\sum_{j=1}^{n}\sin(\phi_j(t) - \theta_i(t)), \quad i = 1, \dots, m, \tag{3.1}$$

$$\dot{\phi}_i(t) = \upsilon_i + \frac{k}{m+n}\sum_{j=1}^{m}\sin(\theta_j(t) - \phi_i(t)), \quad i = 1, \dots, n. \tag{3.2}$$

In the above notation, $\theta_i(\cdot) \in \mathbb{R}$ and $\phi_j(\cdot) \in \mathbb{R}$ denote the phase of oscillator i and the phase of oscillator $m + j$, respectively. Likewise, $\omega_i \in \mathbb{R}$ and $\upsilon_j \in \mathbb{R}$ denote the intrinsic frequency of oscillators i and $m + j$, respectively. Lastly, the symbol $k \in \mathbb{R}_{\geq 0}$ denotes the coupling coefficient.

The system of equations (3.1)–(3.2) admits a range of qualitatively different solutions, depending on the value of the (bifurcation) parameter k. In this paper, we are concerned with solutions associated with the existence of global phase-locked states, which we define as follows.

Definition 5 (GPLS, complete bipartite graph). *Let $\Phi^0 \in \mathbb{R}^m \times \mathbb{R}^n$, and let $\Phi(\cdot) := \big(\theta^*(\cdot), \phi^*(\cdot)\big)$, $t \geq t_0$, be a solution of (3.1)–(3.2) with initial condition $\Phi(t_0) = \Phi^0$. We say that Φ^0 is a* global phase-locked state *(GPLS) if*

$$\Phi_i(t) - \Phi_j(t) = \Phi_i^0 - \Phi_j^0, \quad 1 \leq (i, j) \leq m + n, \ t \geq t_0. \tag{3.3}$$

Note that to every global phase-locked solution x^* there corresponds a 1-dimensional manifold

$$\mathcal{M} := \{x \in \mathbb{R}^{m+n} : \ x = x^* + \mathbf{1}_{m+n}\, t, \, t \in \mathbb{R}\}$$

that is invariant under the original system dynamics (3.1)–(3.2). This correspondence is unique up to equivalence in the sense of the following definition.

Definition 6 (equivalence of phase-locked states). *Let $x, x' \in \mathbb{R}^m \times \mathbb{R}^n$ be two GPLS in the sense of Definition 5. We say that x is equivalent to x' (and write $x \simeq x'$) if there exists $c \in \mathbb{R}$ such that $x = x' + c\mathbf{1}_{m+n}$.*

From the above, and from inspection of (3.1)–(3.2), it follows easily that $(\xi^*, \nu^*) \in \mathbb{R}^m \times \mathbb{R}^n$ is a GPLS if and only if

$$\left(\sum_{j=1}^{m}\omega_j + \sum_{j=1}^{n}\upsilon_j\right) - (m+n)\,\omega_i = k\sum_{j=1}^{n}\sin(\nu_j^* - \xi_i^*), \quad i = 1, \dots, m;$$

$$\left(\sum_{j=1}^{m}\omega_j + \sum_{j=1}^{n}\upsilon_j\right) - (m+n)\,\upsilon_i = k\sum_{j=1}^{m}\sin(\xi_j^* - \nu_i^*), \quad i = 1, \dots, n.$$

Let $f : \mathbb{R}^m \times \mathbb{R}^n \mapsto \mathbb{R}^m$ and $g : \mathbb{R}^m \times \mathbb{R}^n \mapsto \mathbb{R}^n$ be given as

$$(m+n)\, f(\xi, \nu) := \left(\sum_{j=1}^{n}\sin(\nu_j - \xi_1), \quad \dots, \quad \sum_{j=1}^{n}\sin(\nu_j - \xi_m)\right)^T, \tag{3.4}$$

$$(m+n)\, g(\xi, \nu) := \left(\sum_{j=1}^{m}\sin(\xi_j - \nu_1), \quad \dots, \quad \sum_{j=1}^{m}\sin(\xi_j - \nu_n)\right)^T. \tag{3.5}$$

Also, let $\mathbf{1}_m$ denote the vector of length m all of whose components are 1, and define $\Omega : \mathbb{R}^m \times \mathbb{R}^n \mapsto \mathbb{R}^m$, and $\Upsilon : \mathbb{R}^m \times \mathbb{R}^n \mapsto \mathbb{R}^n$,

$$\Omega(\omega,\upsilon) := \omega - \frac{1}{m+n}\left(\sum_{j=1}^{m}\omega_j + \sum_{j=1}^{n}\upsilon_j\right)\mathbf{1}_m,$$

$$\Upsilon(\omega,\upsilon) := \upsilon - \frac{1}{m+n}\left(\sum_{j=1}^{m}\omega_j + \sum_{j=1}^{n}\upsilon_j\right)\mathbf{1}_n.$$

Note that $\begin{pmatrix}\Omega & \Upsilon\end{pmatrix}^T$ is a linear projection from $\mathbb{R}^{m+n}$ onto the $(m+n-1)$-dimensional linear subspace $\{x \in \mathbb{R}^{m+n} : \sum_{j=1}^{m+n} x_j = 0\}$. In this paper we present a method for computing the critical coupling associated with the system of equations (3.1)–(3.2), where the critical coupling is defined as follows.

Definition 7 (critical coupling, complete bipartite graph). *Let $(m,n) \in \mathbb{N}\times\mathbb{N}$, and let $\omega \in \mathbb{R}^m$ and $\upsilon \in \mathbb{R}^n$ be given. Consider the Kuramoto model of coupled oscillators on the complete bipartite graph $K_{m,n}$ (3.1)–(3.2). For this model, we define the critical coupling, k_c, as follows:*

$$k_c := \min_k \left\{k \in \mathbb{R}_{\geq 0} : \exists(\xi^*,\nu^*) \in \mathbb{R}^m \times \mathbb{R}^n \text{ such that } \begin{pmatrix}\Omega\\ \Upsilon\end{pmatrix} = -k\begin{pmatrix}f(\xi^*,\nu^*)\\ g(\xi^*,\nu^*)\end{pmatrix}\right\}$$

(omitting the argument (ω,υ) for notational convenience).

The critical coupling is thus the smallest nonnegative coupling for which the system (3.1)–(3.2) admits a GPLS in the sense of Definition 5.

3.1. The order parameter. Let $\mathbb{D}$ denote the closed complex unit disc $\{z \in \mathbb{C} : |z| \leq 1\}$. For $p \in \mathbb{N}$, we define the order parameter $r^p : \mathbb{R}^p \mapsto \mathbb{D}$ as

$$r^p(\xi) := \frac{1}{p}\sum_{j=1}^{p} e^{i\xi_j}. \tag{3.6}$$

Let r_0^p denote the set $\{\xi \in \mathbb{R}^p : r^p(\xi) = 0\}$. For the purposes of this paper, we define the arctangent as a function from $\mathbb{R}^2 \setminus \{(0,0)\}$ onto $[0,2\pi)$, as follows.

Definition 8. *For all $(a,b) \in \mathbb{R}^2 \setminus \{(0,0)\}$ let $\arctan(a,b)$ denote the unique real number $c \in [0,2\pi)$ such that*

$$\cos(c) = \frac{a}{\sqrt{a^2+b^2}}, \qquad \sin(c) = \frac{b}{\sqrt{a^2+b^2}}. \tag{3.7}$$

Using Definition 8, we may may express $r^p(\cdot)$ in polar coordinates:

$$r^p(\xi) = \begin{cases} R^p(\xi)e^{i\psi^p(\xi)}, & \xi \in \mathbb{R}^p \setminus r_0^p,\\ 0, & \xi \in r_0^p. \end{cases} \tag{3.8}$$

Here, $R^p : \mathbb{R}^p \mapsto [0,1]$ and $\psi^p : \mathbb{R}^p \setminus r_0^p \mapsto [0,2\pi)$ are respectively defined as

$$R^p(\xi) := \sqrt{\left(\frac{1}{p}\sum_{j=1}^{p}\cos(\xi_j)\right)^2 + \left(\frac{1}{p}\sum_{j=1}^{p}\sin(\xi_j)\right)^2} \tag{3.9}$$

and

$$\psi^p(\xi) := \arctan\left(\frac{1}{p}\sum_{j=1}^{p}\cos(\xi_j), \frac{1}{p}\sum_{j=1}^{p}\sin(\xi_j)\right). \tag{3.10}$$

The following proposition summarizes some useful properties of the maps R^p and ψ^p.

Proposition 3.1. *The maps R^p and ψ^p defined by (3.9), (3.10) satisfy the following relations:*

1. $R^p(\xi + c\mathbf{1}_p) = R^p(\xi)$ *for all* $\xi \in \mathbb{R}^p$, $c \in \mathbb{R}$;
2. $R^p(\xi) = \frac{1}{p}\sum_{j=1}^{p}\cos(\psi^p(\xi) - \xi_j)$ *for all* $\xi \in \mathbb{R}^p \setminus r_0^p$;
3. $\psi^p(\xi + c\mathbf{1}_p) = \psi^p(\xi) + c \ (mod\ 2\pi)$ *for all* $\xi \in \mathbb{R}^p$, $c \in \mathbb{R}$;
4. $\sum_{j=1}^{p}\sin(\psi^p(\xi) - \xi_j) = 0$ *for all* $\xi \in \mathbb{R}^p \setminus r_0^p$.

Proof. First, note that, for all $\xi \in \mathbb{R}^p$,

$$R^p(\xi + c\mathbf{1}_p) \ := \ \left|\frac{1}{p}\sum_{j=1}^{p} e^{i(\xi_j + c)}\right| \ = \ \left|e^{ic}\right| R^p(\xi) \ = \ R^p(\xi).$$

This proves property 1. Next observe that, for $\xi \in \mathbb{R}^p \setminus r_0^p$,

$$\begin{aligned} R^p(\xi) &= e^{-i\psi^p(\xi)} r^p(\xi) \\ &= \frac{1}{p}\sum_{j=1}^{p} e^{i(\xi_j - \psi^p(\xi))}. \end{aligned} \tag{3.11}$$

Equating real and imaginary parts in (3.11), we see that, for $\xi \in \mathbb{R}^p \setminus r_0^p$,

$$R^p(\xi) = \frac{1}{p}\sum_{j=1}^{p}\cos(\psi^p(\xi) - \xi_j), \tag{3.12}$$

$$\sum_{j=1}^{p}\sin(\psi^p(\xi) - \xi_j) = 0. \tag{3.13}$$

This proves properties 2 and 4. Finally, by definition of $\psi^p(\cdot)$ we have that

$$\begin{aligned} \cos(\psi^p(\xi + c\mathbf{1}_p)) &= \frac{1}{R^p(\xi + c\mathbf{1}_p)}\frac{1}{p}\sum_{j=1}^{p}\cos(\xi_j + c) \\ &= \frac{1}{R^p(\xi)}\frac{1}{p}\sum_{j=1}^{p}\cos(\xi_j)\cos(c) - \sin(\xi_j)\sin(c) \\ &= \cos(\psi^p(\xi))\cos(c) - \sin(\psi^p(\xi))\sin(c). \end{aligned}$$

This implies that

$$\cos(\psi^p(\xi + c\mathbf{1}_p)) = \cos(\psi^p(\xi) + c). \tag{3.14}$$

Similarly we have that

$$\sin(\psi^p(\xi + c\mathbf{1}_p)) = \sin(\psi^p(\xi) + c). \tag{3.15}$$

It follows that $\psi^p(\xi + c\mathbf{1}_p) = \psi^p(\xi) + c \pmod{2\pi}$. This proves property 3. ∎

Using the definition of the order parameter, we can rewrite the functions f and g, previously defined in (3.4) and (3.5), as follows:

$$(m+n)f_i(\xi,\nu) = \begin{cases} n\,R^n(\nu)\sin(\psi^n(\nu) - \xi_i), & \nu \in \mathbb{R}^n \setminus r_0^n, \\ 0, & \nu \in r_0^n, \end{cases} \tag{3.16}$$

and

$$(m+n)g_i(\xi,\nu) = \begin{cases} m\,R^m(\xi)\sin(\psi^m(\xi) - \nu_i), & \xi \in \mathbb{R}^m \setminus r_0^m, \\ 0, & \xi \in r_0^m. \end{cases} \tag{3.17}$$

3.2. GPLS: The homogeneous case. In this section we assume that $\Omega(\omega, \upsilon) = 0$ and $\Upsilon(\omega, \upsilon) = 0$. Thus we assume that there exists $c \in \mathbb{R}$ such that $\omega_i = \upsilon_j = c$ for $1 \le i \le m$ and $1 \le j \le n$. Under this assumption the system of ODEs (3.1)–(3.2) simplifies to

$$\begin{pmatrix} \dot\theta(t) \\ \dot\phi(t) \end{pmatrix} = c\mathbf{1}_{m+n} + k \begin{pmatrix} f(\theta,\phi) \\ g(\theta,\phi) \end{pmatrix}. \tag{3.18}$$

It is easy to see that the homogeneous system (3.18) admits a GPLS for all $k > 0$ (as it does, of course, for the trivial case $k = 0$). Indeed, for $k > 0$, we have that $(\xi,\nu) \in \mathbb{R}^m \times \mathbb{R}^n$ is a GPLS if and only if $f(\xi,\nu) = 0$ and $g(\xi,\nu) = 0$. This motivates the following result.

Proposition 3.2. *Let f and g be defined by (3.4) and (3.5), respectively, and let $(\xi,\nu) \in \mathbb{R}^m \times \mathbb{R}^n$. If (ξ,ν) satisfies the condition that*

$$\sin(\xi_i - \nu_j) = 0, \quad 1 \le i \le m, \quad 1 \le j \le n, \tag{3.19}$$

then (ξ,ν) is a solution of the homogeneous system

$$\begin{pmatrix} f(\xi,\nu) \\ g(\xi,\nu) \end{pmatrix} = 0. \tag{3.20}$$

Moreover, if (ξ,ν) is a solution of (3.20) and if, in addition, $R^m(\xi)R^n(\nu) \neq 0$, then (3.19) holds.

Proof. The first part follows trivially from inspection of (3.4) and (3.5). To prove the second part, suppose $R^m(\xi) \neq 0$. Then $\psi^m(\xi)$ is well defined, and by (3.5) we have that $\sin(\psi^m(\xi) - \nu_i) = 0$ for all i. This implies that there exists $k \in \mathbb{Z}^n$ such that $\nu_i - \nu_j = (k_i - k_j)\pi$ and hence $\sin(\nu_i - \nu_j) = 0$ for all (i,j). By analogy we have that $\sin(\psi^n(\nu) - \xi_i) = 0$ and $\sin(\xi_i - \xi_j) = 0$ for all (i,j). In addition, by Proposition 3.1 (property 3) we have that $\sin(\psi^n(\nu) - \nu_i) = \sin(\psi^n(\nu - \nu_i)) = \frac{1}{R^n(\nu)}\frac{1}{n}\sum_{j=1}^n \sin(\nu_j - \nu_i) = 0$ for all i. It follows that

$$\begin{aligned} \sin(\xi_i - \nu_j) &= \sin(\xi_i - \psi^n(\nu) - (\nu_j - \psi^n(\nu)) \\ &= \underbrace{\sin(\xi_i - \psi^n(\nu))}_{=0}\cos(\nu_j - \psi^n(\nu)) \\ &\quad - \cos(\xi_i - \psi^n(\nu))\underbrace{\sin(\nu_j - \psi^n(\nu))}_{=0} \;=\; 0 \end{aligned}$$

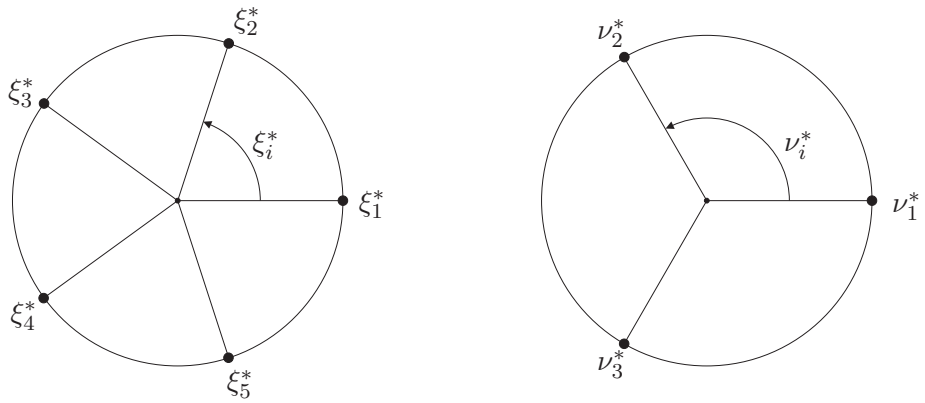


Figure 2. *One solution $(\xi^*, \nu^*) \in \mathcal{S}_{\text{hom}}$ of the system of equations (3.20) for the case of the complete bipartite graph $K_{5,3}$. Note that (ξ^*, ν^*) does not satisfy the conditions of Proposition 3.2. In this particular example, however, we have that $R^m(\xi^*) = 0$ and $R^n(\nu^*) = 0$, and therefore (ξ^*, ν^*) is a GPLS by Remark 1.*

for all (i, j). This concludes the proof. ∎

Remark 1. *Proposition 3.2 provides a partial characterization of the set of solutions to the homogeneous equation (3.20). Condition (3.19) is a sufficient but not a necessary condition. However, it becomes necessary when the extra condition $R^m(\xi)R^n(\nu) \neq 0$ is imposed, as per the second part of the proposition. The complete set of homogeneous solutions can be found by taking the union of the set of solutions defined by conditions (3.19) and the set $\{(\xi, \nu) \in \mathbb{R}^m \times \mathbb{R}^n : R^m(\xi)R^n(\nu) = 0\}$. In general, these sets are not disjoint. In fact, it can be shown that these sets are disjoint if and only if both m and n are odd.*

In the remainder of the paper, we shall use the notation $\mathcal{S}_{\text{hom}}$ to denote the set of all GPLS corresponding to the homogeneous system (3.20). That is,

$$\mathcal{S}_{\text{hom}} := \left\{ (\xi, \nu) \in \mathbb{R}^m \times \mathbb{R}^n : \begin{pmatrix} f(\xi, \nu) \\ g(\xi, \nu) \end{pmatrix} = 0 \right\}.$$

Figure 2 shows a particular GPLS, $(\xi^*, \nu^*) \in \mathcal{S}_{\text{hom}}$, for the case of the complete bipartite graph $K_{5,3}$.

3.3. Global phase-locking under strong coupling. Let $\Pi^1 : \mathbb{R}^{m+n} \mapsto \mathbb{R}^m$ and $\Pi^2 : \mathbb{R}^{m+n} \mapsto \mathbb{R}^n$ be given by

$$\Pi^1(x) := \begin{pmatrix} x_1 & \cdots & x_m \end{pmatrix}^T, \qquad \Pi^2(x) := \begin{pmatrix} x_{m+1} & \cdots & x_{m+n} \end{pmatrix}^T,$$

respectively, and define $F : \mathbb{R}^{m+n} \mapsto \mathbb{R}^{m+n}$,

$$F(x) := \begin{pmatrix} f(\Pi^1(x), \Pi^2(x)) \\ g(\Pi^1(x), \Pi^2(x)) \end{pmatrix}. \tag{3.21}$$

Also, let $\Xi : \mathbb{R}^m \times \mathbb{R}^n \mapsto \mathbb{R}^{m+n}$ be given as

$$\Xi(\omega, \upsilon) := \begin{pmatrix} \Omega(\omega, \upsilon) \\ \Upsilon(\omega, \upsilon) \end{pmatrix}. \tag{3.22}$$

Lastly, define the truncation operator $T : \mathbb{R}^{m+n} \mapsto \mathbb{R}^{m+n-1}$,

$$(Tz) := z_i, \quad i = 1, \ldots, m+n-1.$$

We now state a technical lemma which will play an important role in the proof of Proposition 3.4 below. As the proof of this lemma is rather long and technical, we defer it to Appendix B in the interests of clarity.

Lemma 3.3. *Let $\tilde{F} : \mathbb{R}^{m+n-1} \to \mathbb{R}^{m+n-1}$ be defined by $\tilde{F}(y) = TF(y;0)$, and suppose that $y^* \in \mathbb{R}^{m+n-1}$ satisfies $\tilde{F}(y^*) = 0$ and $R^m(\Pi^1(y^*;0))R^n(\Pi^2(y^*;0)) \neq 0$. Then the Jacobian $J : \mathbb{R}^{m+n-1} \mapsto \mathbb{R}^{m+n-1}$,*

$$J(y) \ := \ \frac{\partial \tilde{F}(\zeta)}{\partial \zeta}(y),$$

is nonsingular at $y = y^*$.

We can now state and prove the following proposition, which establishes that the critical coupling is finite for complete bipartite graphs and any given set of finite natural frequencies.

Proposition 3.4. *Let $(\omega, \upsilon) \in \mathbb{R}^m \times \mathbb{R}^n$ be given, and let $\Xi(\omega, \upsilon)$ be defined by (3.22). Furthermore, let $x^* \in \mathcal{S}_{hom}$, and suppose $R^m(\Pi^1(x^*))R^n(\Pi^2(x^*)) \neq 0$. Then there exist $K > 0$ and an open set $V \subset \mathbb{R}^{m+n-1}$ containing Tx^* such that, for all $k > K$, the system*

$$kF(y_1, \ldots, y_{m+n-1}; x^*_{m+n}) = \Xi(\omega, \upsilon) \tag{3.23}$$

has a unique solution y' on V.

Proof. Observe that $\sum_{j=1}^{m+n} F_j(\cdot) \equiv 0$ and $\sum_{j=1}^{m+n} \Xi_j(\cdot,\cdot) \equiv 0$. This implies that y is a solution of (3.23) if and only if $kTF(y; x^*_{m+n}) = T\Xi(\omega, \upsilon)$ or

$$\begin{aligned} kF_1(y_1, \ldots, y_{m+n-1}; x^*_{m+n}) &= \Xi_1(\omega, \upsilon) \\ &\vdots \\ kF_{m+n-1}(y_1, \ldots, y_{m+n-1}; x^*_{m+n}) &= \Xi_{m+n-1}(\omega, \upsilon). \end{aligned} \tag{3.24}$$

Next, assume without loss of generality that $x^*_{m+n} = 0$, and define $y^* = Tx^*$ and $\tilde{F} : \mathbb{R}^{m+n-1} \mapsto \mathbb{R}^{m+n-1}$, $\tilde{F}(y) := TF(y;0)$. Then under the hypotheses of the proposition we have $\tilde{F}(y^*) = 0$, and it follows from Lemma 3.3 that the Jacobian $J : \mathbb{R}^{m+n-1} \mapsto \mathbb{R}^{m+n-1}$,

$$J(y) \ := \ \frac{\partial \tilde{F}(\zeta)}{\partial \zeta}(y),$$

is nonsingular at $y = y^*$. Then by the inverse function theorem there exists an open set U, containing the origin, such that on U the map $\tilde{F}$ has a unique inverse, $\tilde{F}^{-1}$. Define $V := \tilde{F}^{-1}(U)$, and note that V is open by continuity of $\tilde{F}^{-1}$. Now pick $\epsilon > 0$ such that $\mathcal{B}_\epsilon := \{z \in \mathbb{R}^{n+m-1} : \|z\| < \epsilon\} \subset U$, and let $K = \frac{\|\tilde{\Xi}\|}{\epsilon}$. Then it follows that (3.24) has a unique solution on V for all $k \geq K$. ∎

4. Critical coupling. In the previous section we showed that, for large enough values of k, the heterogeneous system (3.1)–(3.2) will exhibit GPLS in (small) open sets containing a GPLS of the associated homogeneous system (3.20). A particular implication of this result is that the critical coupling k_c is always finite. In the present section we will refine this result and show that the value of the critical coupling can be obtained by solving a particular system of nonlinear equations, the solution of which can be determined very efficiently by means of a bisection algorithm.

Our first result relates the existence of GPLS to solutions of a 2-dimensional system of equations, as follows.

Theorem 4.1. *Let $k > 0$ and $(\omega, \upsilon) \in \mathbb{R}^m \times \mathbb{R}^n$. Then the system (3.1)–(3.2) admits a GPLS if and only if there exist $(x, y) \in [\frac{m+n}{m}\frac{1}{k}\|\Upsilon\|_\infty, 1] \times [\frac{m+n}{n}\frac{1}{k}\|\Omega\|_\infty, 1]$ and $(a, b) \in \{-1, 1\}^m \times \{-1, 1\}^n$ such that*

$$x^2 = \left(\frac{m+n}{mn}\sum_{j=1}^{m}\frac{\Omega_j}{ky}\right)^2 + \left(\frac{1}{m}\sum_{j=1}^{m} a_j\sqrt{1-\left(\frac{m+n}{n}\frac{\Omega_j}{ky}\right)^2}\right)^2, \tag{4.1}$$

$$y^2 = \left(\frac{m+n}{mn}\sum_{j=1}^{n}\frac{\Upsilon_j}{kx}\right)^2 + \left(\frac{1}{n}\sum_{j=1}^{n} b_j\sqrt{1-\left(\frac{m+n}{m}\frac{\Upsilon_j}{kx}\right)^2}\right)^2. \tag{4.2}$$

Proof. We distinguish four cases: (a) $\Omega(\omega, \upsilon) = 0$ and $\Upsilon(\omega, \upsilon) = 0$; (b) $\Omega(\omega, \upsilon) = 0$ and $\Upsilon(\omega, \upsilon) \neq 0$; (c) $\Omega(\omega, \upsilon) \neq 0$ and $\Upsilon(\omega, \upsilon) = 0$; (d) $\Omega(\omega, \upsilon) \neq 0$ and $\Upsilon(\omega, \upsilon) \neq 0$. Case (a) corresponds to the homogeneous system (3.20) that we discussed in the previous section, where we showed that GPLS exist for every $k > 0$. Hence, what we need to show for this case is that for every $k > 0$ there exists $(a, b) \in \{-1, 1\}^m \times \{-1, 1\}^n$ such that (4.1)–(4.2) has a solution $(x, y) \in [0, 1] \times [0, 1]$. Under the given hypotheses it suffices to note that the system $\{x^2 = 1, y^2 = 1\}$ has a nonnegative solution that does not depend on k. Case (b) corresponds to the semihomogeneous system

$$\left.\begin{aligned} 0 &= \tfrac{k}{m+n}\textstyle\sum_{j=1}^{n}\sin(\nu_j - \xi_i), \quad i = 1, \dots, m, \\ -\Upsilon_i &= \tfrac{k}{m+n}\textstyle\sum_{j=1}^{m}\sin(\xi_j - \nu_i), \quad i = 1, \dots, n. \end{aligned}\right\} \tag{4.3}$$

This system has a solution if and only if $k \geq \frac{m+n}{m}\|\Upsilon\|_\infty$. The necessity is obvious. For sufficiency, note that if $\xi_i = c$ for $1 \leq i \leq m$, we can always find $\nu_1, \dots, \nu_n$ satisfying the second equation above. But then $\frac{km}{m+n}\sin(\nu_j - c) = \Upsilon_j$ for $1 \leq j \leq n$ and

$$\frac{k}{m+n}\sum_{j=1}^{n}\sin(\nu_j - c) = m\left(\sum_{j=1}^{n}\Upsilon_j + \sum_{j=1}^{m}\Omega_j\right) = 0.$$

Hence, what we need to show is that there exists $(a, b) \in \{-1, 1\}^m \times \{-1, 1\}^n$ such that the system (4.1)–(4.2) has a solution $(x, y) \in [\frac{m+n}{m}\frac{1}{k}\|\Upsilon\|_\infty, 1] \times [0, 1]$ if and only if $k \geq \frac{m+n}{m}\|\Upsilon\|_\infty$. We note that the condition $x \in [\frac{m+n}{m}\frac{1}{k}\|\Upsilon\|_\infty, 1]$ implies that $k \geq \frac{m+n}{m}\|\Upsilon\|_\infty$, which proves necessity. To prove sufficiency suppose $k \geq \frac{m+n}{m}\|\Upsilon\|_\infty$. Under the given hypotheses it suffices to show that the system

$$\left\{ x^2 = 1,\, y^2 = \left(\frac{m+n}{mn} \sum_{j=1}^{n} \frac{\Upsilon_j}{k} \right)^2 + \left(\frac{1}{n} \sum_{j=1}^{n} \sqrt{1 - \left(\frac{m+n}{m} \frac{\Upsilon_j}{k} \right)^2} \right)^2 \right\}$$

has a solution on $[\frac{m+n}{m}\frac{1}{k}\|\Upsilon\|_\infty, 1] \times [0,1]$. But this amounts to showing that

$$\left(\frac{m+n}{mn} \sum_{j=1}^{n} \frac{\Upsilon_j}{k} \right)^2 + \left(\frac{1}{n} \sum_{j=1}^{n} \sqrt{1 - \left(\frac{m+n}{m} \frac{\Upsilon_j}{k} \right)^2} \right)^2 \leq 1.$$

If $k \geq \frac{m+n}{m}\|\Upsilon\|_\infty$, then for each $j = 1, \ldots, n$ we can write $\frac{m+n}{m}\frac{\Upsilon_j}{k} = \cos(x_j)$ for some $x_j \in [-\pi, \pi]$, and thus the above expression reduces to

$$\left(\frac{1}{n} \sum_{j=1}^{n} \cos(x_j) \right)^2 + \left(\frac{1}{n} \sum_{j=1}^{n} \sin(x_j) \right)^2 = \left| \frac{1}{n} \sum_{j=1}^{n} e^{i x_j} \right|^2 \leq 1,$$

as required. Case (c) is analogous to case (b). We now consider case (d). Let (ξ, ν) be a GPLS in the sense of Definition 5. The conditions $\Omega(\omega, \upsilon) \neq 0$ and $\Upsilon(\omega, \upsilon) \neq 0$ imply that $R^m(\xi) \neq 0$ and $R^n(\nu) \neq 0$, respectively. It follows that $\psi^m(\xi)$ and $\psi^n(\nu)$ are well defined, and by (3.16)–(3.17) we have that

$$\sin(\psi^n(\nu) - \xi_i) = -\frac{(\frac{m+n}{n})\Omega_i}{kR^n(\nu)}, \quad i = 1, \ldots, n,$$
$$\sin(\psi^m(\xi) - \nu_i) = -\frac{(\frac{m+n}{m})\Upsilon_i}{kR^m(\xi)}, \quad i = 1, \ldots, m.$$

Recall that $(R^m(\xi))^2 = (R^m(\xi - \psi^n(\nu)\mathbf{1}_m))^2$ by Proposition 3.1. It follows that

$$(R^m(\xi))^2 = \left(\frac{1}{m} \sum_{j=1}^{m} \sin(\xi_j - \psi^n(\nu)) \right)^2 + \left(\frac{1}{m} \sum_{j=1}^{m} \cos(\xi_j - \psi^n(\nu)) \right)^2,$$
$$(R^n(\nu))^2 = \left(\frac{1}{n} \sum_{j=1}^{n} \sin(\nu_j - \psi^m(\xi)) \right)^2 + \left(\frac{1}{n} \sum_{j=1}^{n} \cos(\nu_j - \psi^m(\xi)) \right)^2.$$

Now let $a \in \{-1, 1\}^m$ and $b \in \{-1, 1\}^n$ be respectively defined as

$$a_i := \begin{cases} +1, & \cos(\xi_i - \psi^n(\nu)) \geq 0, \\ -1, & \cos(\xi_i - \psi^n(\nu)) < 0, \end{cases} \quad 1 \leq i \leq m;$$
$$b_i := \begin{cases} +1, & \cos(\nu_i - \psi^m(\xi)) \geq 0, \\ -1, & \cos(\nu_i - \psi^m(\xi)) < 0, \end{cases} \quad 1 \leq i \leq n.$$

Let $x := R^m(\xi)$, $y := R^n(\nu)$. Then the result follows by substitution. This proves necessity. To prove sufficiency, let $(a', b') \in \{-1, 1\}^m \times \{-1, 1\}^n$ be given, and let (x, y) be a solution of

(4.1)–(4.2) with $(a,b)=(a',b')$. Consider the system of equations

$$\begin{cases} \sin(u_i) &= -\frac{(\frac{m+n}{n})\Omega_i}{ky}, \\ a_i'\cos(u_i) &\geq 0, \end{cases} \qquad i=1,\dots,m; \tag{4.4}$$

$$\begin{cases} \sin(v_i) &= -\frac{(\frac{m+n}{m})\Upsilon_i}{kx}, \\ b_i'\cos(v_i) &\geq 0, \end{cases} \qquad i=1,\dots,n. \tag{4.5}$$

Under the hypotheses of the theorem, the system (4.4)–(4.5) has a unique solution (u',v') on $(-\pi,\pi]^m\times(-\pi,\pi]^n$. We claim that either $\psi^m(u')=-\psi^n(v')$ or $\psi^m(u')-\psi^n(v')=\pi \pmod{2\pi}$. To prove this, recall that $\sum_{j=1}^m\Omega_j+\sum_{j=1}^n\Upsilon_j=0$. By (4.4)–(4.5) we have that $\sum_{j=1}^n\sin(v_j')=-\frac{ny}{mx}\sum_{j=1}^m\sin(u_j')$. If we can show that $|\sum_{j=1}^n\cos(v_j')|=|\frac{ny}{mx}\sum_{j=1}^m\cos(u_j')|$, this claim will be established. By (4.1) and (4.2) we have that

$$\begin{aligned}\left(\sum_{j=1}^n\cos(v_j')\right)^2 &= n^2y^2-\left(\sum_{j=1}^n\sin(v_j')\right)^2\\ &= \frac{n^2y^2}{m^2x^2}\left[m^2x^2-\left(\sum_{j=1}^m\sin(u_j')\right)^2\right]\\ &= \left(\frac{ny}{mx}\sum_{j=1}^m\cos(u_j')\right)^2,\end{aligned}$$

where the last line follows from (4.1). This proves the claim. Indeed, the above implies that either

$$\sum_{j=1}^n\cos(v_j')=\frac{ny}{mx}\sum_{j=1}^m\cos(u_j'), \tag{4.6}$$

in which case $\psi^m(u')=-\psi^n(v')$, or

$$\sum_{j=1}^n\cos(v_j')=-\frac{ny}{mx}\sum_{j=1}^m\cos(u_j'), \tag{4.7}$$

in which case $\psi^m(u')-\psi^n(v')=\pi \pmod{2\pi}$. Suppose that (4.7) holds. Define $(a'',b''):=(a',-b')$ and observe that (x,y) is a solution of (4.1)–(4.2) with $(a,b)=(a'',b'')$. Moreover, if (u'',v'') denotes the solution of (4.4)–(4.5) with (a',b') replaced with (a'',b''), then we have that $\psi^m(u'')=-\psi^n(v'')$. This implies that we can assume without loss of generality that (4.6) holds, and hence that $\psi^m(u')=-\psi^n(v')$. Now define

$$\begin{aligned}\xi_i &:= -u_i', \quad i=1,\dots,m,\\ \nu_i &:= -v_i'-\psi^m(u'), \quad i=1,\dots,n.\end{aligned}$$

Then we have that

$$\begin{aligned}\sin(\psi^n(\nu)-\xi_i)=\sin(-\psi^n(v')-\psi^m(u')+u_i) &= \sin(u_i), \quad i=1,\dots,m,\\ \sin(\psi^m(\xi)-\nu_i)=\sin(-\psi^m(u')+\psi^m(u')+v_i) &= \sin(v_i), \quad i=1,\dots,n.\end{aligned}$$

If we can show that $R^m(\xi) = x$ and $R^n(\xi) = y$, then it follows that (ξ, ν) is a GPLS, and the proof is finished. We have that

$$\begin{aligned}(R^m(\xi))^2 &= \left(\frac{1}{m}\sum_{j=1}^{m}\sin(\xi_j - \psi^n(\nu))\right)^2 + \left(\frac{1}{m}\sum_{j=1}^{m}\cos(\xi_j - \psi^n(\nu))\right)^2 \\ &= \left(\frac{m+n}{mn}\sum_{j=1}^{m}\frac{\Omega_j}{ky}\right)^2 + \left(\frac{1}{m}\sum_{j=1}^{m}a_j\sqrt{1-\left(\frac{m+n}{n}\frac{\Omega_j}{ky}\right)^2}\right)^2 \\ &= x^2, \end{aligned} \tag{4.8}$$

where the last step follows from the hypotheses of the proposition. This implies that $R^m(\xi) = x$, and by analogy we have that $R^n(\nu) = y$. This concludes the proof. ∎

Define

$$\begin{aligned} I_v &:= \left[\left(\tfrac{m+n}{n}\|\Omega\|_\infty\right)^2, \infty\right), \qquad I_v^\circ := \left(\left(\tfrac{m+n}{n}\|\Omega\|_\infty\right)^2, \infty\right), \\ I_w &:= \left[\left(\tfrac{m+n}{m}\|\Upsilon\|_\infty\right)^2, \infty\right), \qquad I_w^\circ := \left(\left(\tfrac{m+n}{m}\|\Upsilon\|_\infty\right)^2, \infty\right), \end{aligned} \tag{4.9}$$

and let $p : I_v \times \{-1,1\}^m \mapsto [0,1]$ and $q : I_w \times \{-1,1\}^n \mapsto [0,1]$ be given as

$$p(v,a) := \left(\frac{m+n}{mn}\sum_{j=1}^{m}\Omega_j\right)^2\frac{1}{v} + \left(\frac{1}{m}\sum_{j=1}^{m}a_j\sqrt{1-\left(\frac{m+n}{n}\Omega_j\right)^2\frac{1}{v}}\right)^2, \tag{4.10}$$

$$q(w,b) := \left(\frac{m+n}{mn}\sum_{j=1}^{n}\Upsilon_j\right)^2\frac{1}{w} + \left(\frac{1}{n}\sum_{j=1}^{n}b_j\sqrt{1-\left(\frac{m+n}{m}\Upsilon_j\right)^2\frac{1}{w}}\right)^2. \tag{4.11}$$

In terms of the above notation, Theorem 4.1 states that the system (3.1)–(3.2) admits a GPLS if and only if there exist $(v,w) \in I_v \times I_w$ and $(a,b) \in \{-1,1\}^m \times \{-1,1\}^n$ such that $w = k^2 p(v,a)$ and $v = k^2 q(w,b)$. It follows that the critical coupling is the smallest k for which such a solution exists. Before we proceed, let us illustrate the result of Theorem 4.1 with an example. We consider the case of the complete bipartite graph $G = K_{5,4}$. Thus, let $(m,n) = (5,4)$, and let ω, υ be given as

$$\omega := \begin{pmatrix}1 & 3 & 5 & 7 & 9\end{pmatrix}^T, \qquad \upsilon := \begin{pmatrix}2 & 4 & 6 & 8\end{pmatrix}^T.$$

Then by definition, we have that

$$\Omega(\omega,\upsilon) = \begin{pmatrix}-4 & -2 & 0 & 2 & 4\end{pmatrix}^T, \qquad \Upsilon(\omega,\upsilon) = \begin{pmatrix}-3 & -1 & 1 & 3\end{pmatrix}^T.$$

We note that Theorem 4.1 implies the weaker result that the system (4.1)–(4.2) admits a GPLS if there exists $(v,w) \in I_v \times I_w$ such that $w = k^2 p(v, \mathbf{1}_m)$ and $v = k^2 q(w, \mathbf{1}_n)$. In Figure 3 we have plotted three sets of two graphs for three different values of k. The graphs correspond to the sets of points

$$\left\{\left(\frac{v}{k^2}, p(v,\mathbf{1}_m)\right) \;:\; \frac{v}{k^2} \in \left[\left(\frac{m+n}{n}\|\Omega\|_\infty\right)^2, 1\right]\right\}$$

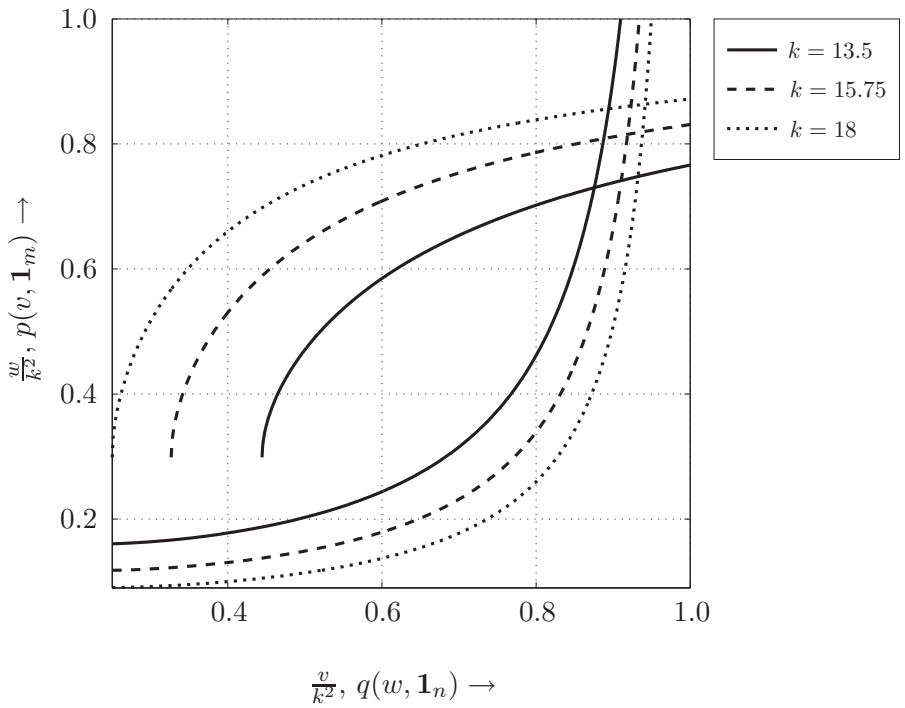


Figure 3. *The graphs of* $(\frac{v}{k^2}, p(v, \mathbf{1}_m))$ *and* $(q(w, \mathbf{1}_n), \frac{w}{k^2})$ *for different values of the coupling coefficient* k*. An intersection is a point* (v^*, w^*) *such that* $k^2 p(v^*, \mathbf{1}_m) = w^*$ *and* $k^2 q(w^*, \mathbf{1}_n) = v^*$*. By Theorem 4.1, such an intersection corresponds to a GPLS of the system (3.1)–(3.2).*

and

$$\left\{ \left(q(w, \mathbf{1}_n), \frac{w}{k^2} \right) \ : \ \frac{w}{k^2} \in \left[\left(\frac{m+n}{m} \|\Upsilon\|_\infty \right)^2, 1 \right] \right\},$$

respectively. We observe that for each value of k there is a unique point of intersection. By Theorem 4.1 such an intersection corresponds to a GPLS.

In Proposition 4.4 below, we shall show that if, for a given value of k, the system (3.1)–(3.2) admits a solution for some $(a, b) \in \{-1, 1\}^m \times \{-1, 1\}^n$, then it will also admit a solution for $(a, b) = (\mathbf{1}_m, \mathbf{1}_n)$. This suggests that, without loss of generality, the critical coupling can be found by determining the smallest value of k for which the system of equations

$$\left. \begin{array}{rcl} w & = & k^2 p(v, \mathbf{1}_m) \\ v & = & k^2 q(w, \mathbf{1}_n) \end{array} \right\} \tag{4.12}$$

has a solution. To prove this, we will make use of some special properties of the functions p and q, namely, monotonicity (Lemma 4.2) and concavity (Lemma 4.3).

Lemma 4.2 (monotonicity). *Let* $(\omega, \upsilon) \in \mathbb{R}^m \times \mathbb{R}^n$ *be given and let* p, q *be defined by (4.10) and (4.11), respectively. Then the following hold:*

1. $p(\cdot, \mathbf{1}_m)$ *is monotone increasing (on* I_v*);*
2. $p(\cdot, \mathbf{1}_m)$ *is strictly monotone if and only if* $\Omega_i(\omega, \upsilon) \neq \Omega_j(\omega, \upsilon)$ *for some* (i, j)*;*
3. $p(\cdot, \mathbf{1}_m)$ *is either strictly monotone or constant.*

Moreover, the statement remains true if we replace $p(\cdot, \mathbf{1}_m)$ *with* $q(\cdot, \mathbf{1}_m)$, I_v *with* I_w, *and* $\Omega(\omega, v)$ *with* $\Upsilon(\omega, v)$.

Lemma 4.3 (concavity). *Let* $(\omega, v) \in \mathbb{R}^m \times \mathbb{R}^n$ *be given and let* p, q *be defined by* (4.10) *and* (4.11), *respectively. Then the following hold:*

1. $p(\cdot, \mathbf{1}_m)$ *is concave (on* I_v*);*
2. $p(\cdot, \mathbf{1}_m)$ *is strictly concave if and only if* $\Omega_i(\omega, v) \neq \Omega_j(\omega, v)$ *for some* (i, j)*;*
3. $p(\cdot, \mathbf{1}_m)$ *is either strictly concave or constant.*

Moreover, the statement remains true if we replace $p(\cdot, \mathbf{1}_m)$ *with* $q(\cdot, \mathbf{1}_m)$, I_v *with* I_w, *and* $\Omega(\omega, v)$ *with* $\Upsilon(\omega, v)$.

The proofs of Lemmas 4.2 and 4.3 are given in Appendix B. We have the following result.

Proposition 4.4. *Let* $k > 0$, *and let* p *and* q *be given by* (4.10) *and* (4.11), *respectively. Suppose there exists* (v', w') *such that* $w' = k^2 p(v', a)$ *and* $v' = k^2 q(w', b)$ *for some* $(a, b) \in \{-1, 1\}^m \times \{-1, 1\}^n$. *Then there exists* (v'', w'') *such that* $w'' = k^2 p(v'', \mathbf{1}_m)$ *and* $v'' = k^2 q(w'', \mathbf{1}_n)$.

Proof. Let $D \subset I_v$ denote the largest (open) subinterval of I_v on which $q(k^2 p(\cdot))$ is defined. Note that if $k^2 p(v^0) \in I_w$ for some $v^0 \in I_v$, then $k^2 p(v) \in I_w$ for all $v \geq v^0$. It follows that

$$D \;=\; \begin{cases} \emptyset & \text{if } \{v \in I_v : k^2 p(v) \in I_w\} = \emptyset, \\ [\min\{v \in I_v : k^2 p(v) \in I_w\}, \infty) & \text{otherwise.} \end{cases}$$

Let $s : D \mapsto \mathbb{R}_{\geq 0}$ be given as

$$s(v) \;:=\; k^2 q(k^2 p(v, \mathbf{1}_m), \mathbf{1}_n).$$

Then there exists $(v, w) \in I_v \times I_w$ such that

$$w \;=\; k^2 p(v, \mathbf{1}_m), \qquad v \;=\; k^2 q(w, \mathbf{1}_n) \tag{4.13}$$

if and only if s has a fixed point on D. Indeed, if v is a fixed point of s, then $(v, w) = (v, k^2 p(v, \mathbf{1}_m))$ is a solution of (4.13) and vice versa. Inspection of (4.10)–(4.11) shows that

$$\begin{aligned} p(v, \mathbf{1}_m) &\geq p(v, a) \quad \forall a \in \{-1, 1\}^m, \quad \forall v \in I_v, \\ q(w, \mathbf{1}_n) &\geq q(w, b) \quad \forall b \in \{-1, 1\}^n, \quad \forall w \in I_w. \end{aligned} \tag{4.14}$$

Let (a', b') be such that $v' = k^2 q(w', b')$ and $w' = k^2 p(v', a')$. As an immediate consequence of (4.14) we have that

$$s(v') \geq k^2 q(k^2 p(v', \mathbf{1}_m), b'). \tag{4.15}$$

Moreover, monotonicity of $p(\cdot, \mathbf{1}_m)$ and $q(\cdot, \mathbf{1}_n)$ (see Lemma 4.2) implies that

$$k^2 q(k^2 p(v', \mathbf{1}_m), b') \geq k^2 q(k^2 p(v', a'), b') \;=\; v'. \tag{4.16}$$

Combining (4.15) and (4.16), we see that $s(v') \geq v'$. If $s(v') = v'$, we are done. Suppose therefore that $s(v') > v'$. If we can show that there exists $v^+ \in (v', \infty)$ such that $s(v^+) < v^+$, then by continuity of s there exists $v'' \in [v', v^+)$ such that $s(v'') = v''$. Note that $s(v) \leq k^2$ for all $v \geq v'$. Take $v^+ = 2k^2$. It follows that $s(v^+) \leq k^2 < v^+$. This concludes the proof. ∎

Theorem 4.5. *Let $(\omega, \upsilon) \in \mathbb{R}^m \times \mathbb{R}^n$, and let p and q be given by (4.10) and (4.11), respectively. If $\Omega_i(\omega,\upsilon) \neq \Omega_j(\omega,\upsilon)$ for some (i,j) and $\Upsilon_l(\omega,\upsilon) \neq \Upsilon_m(\omega,\upsilon)$ for some (l,m), then there exists a unique pair $(v,w) \in I_v \times I_w$ such that*

$$\left.\begin{aligned} \tfrac{\partial q(\zeta,\mathbf{1}_n)}{\partial \zeta}(w) \cdot \tfrac{\partial p(\zeta,\mathbf{1}_m)}{\partial \zeta}(v) &= \tfrac{p(v,\mathbf{1}_m)}{v}\tfrac{q(w,\mathbf{1}_n)}{w} \\ w \cdot q(w,\mathbf{1}_n) - v \cdot p(v,\mathbf{1}_m) &= 0 \end{aligned}\right\}, \tag{4.17}$$

and we have that

$$k_c = \sqrt{\frac{v}{q(w,\mathbf{1}_n)}} = \sqrt{\frac{w}{p(v,\mathbf{1}_m)}}.$$

In all other cases, k_c is given by

$$k_c = \begin{cases} \frac{m+n}{m}\|\Upsilon\|_\infty, & \Omega_i = \Omega_j \ \forall (i,j), \\ \frac{m+n}{n}\|\Omega\|_\infty, & \Upsilon_l = \Upsilon_m \ \forall (l,m). \end{cases}$$

Proof. First we prove the case $\Omega_i = \Omega_j$ for all (i,j). Suppose that there exists $c \in \mathbb{R}$ such that $\Omega_i = c$ for all i. Then $c = -\frac{1}{m}\sum_{j=1}^n \Upsilon_j$ by definition of Ω and Υ. We claim that the system

$$\left.\begin{aligned} -c &= \tfrac{k}{m+n}\textstyle\sum_{j=1}^n \sin(\nu_j - \xi_i), \quad i = 1,\dots,m, \\ -\Upsilon_i &= \tfrac{k}{m+n}\textstyle\sum_{j=1}^m \sin(\xi_j - \nu_i), \quad i = 1,\dots,n, \end{aligned}\right\} \tag{4.18}$$

admits a solution (ξ,ν) if and only if $k \geq \frac{m+n}{m}\|\Upsilon\|_\infty$. Necessity follows by inspection. To prove sufficiency, suppose $k \geq \frac{m+n}{m}\|\Upsilon\|_\infty$. Then there exist $\nu' \in \mathbb{R}^n$ such that $\sin(\nu_i') = \frac{m+n}{km}\Upsilon_i$. Now let $\xi' = 0$. Then by construction $-\Upsilon_i = \frac{k}{m+n}\sum_{j=1}^m \sin(\xi_j' - \nu_i')$, $1 \leq i \leq n$. Moreover, for $1 \leq i \leq m$, we have that $\frac{k}{m+n}\sum_{j=1}^n \sin(\nu_j' - \xi_i') = \frac{1}{n}\sum_{j=1}^n \Upsilon = -c$, and it follows that (ξ',ν') is a solution of (4.18). This proves the claim. The case $\Upsilon_l = \Upsilon_m$ for all (l,m) follows by analogy. Now suppose $\Omega_i(\omega,\upsilon) \neq \Omega_j(\omega,\upsilon)$ for some (i,j) and $\Upsilon_l(\omega,\upsilon) \neq \Upsilon_m(\omega,\upsilon)$ for some (l,m). Let $D := \{v,k) \in I_v \times \mathbb{R}_{\geq 0} : k^2 p(v,\mathbf{1}_m) \in I_w\}$, and let $t : D \mapsto \mathbb{R}_{\geq 0}$ be given as

$$t(v,k) := k^2 q(k^2 p(v,\mathbf{1}_m),\mathbf{1}_n). \tag{4.19}$$

Then k_c is the smallest k for which $t(\cdot,k)$ is defined on a nonempty interval and has a fixed point on that interval. We claim that the fixed point equation $t(v,k_c) = v$ has precisely one solution. To prove this, observe first that, by definition of the critical coupling, it has *at least* one solution. Now suppose by contradiction that there exist $v^{(1)}, v^{(2)}$, $v^{(1)} \neq v^{(2)}$, such that $t(v^{(1)},k_c) = v^{(1)}$ and $t(v^{(2)},k_c) = v^{(2)}$. Without loss of generality, assume that $v^{(2)} > v^{(1)}$. It is easy to verify that the set of points on which $t(\cdot,k_c)$ is defined is convex (i.e., defines an interval). This implies that $t(v,k_c)$ is defined for all $v \in V := [v^{(1)},v^{(2)}]$. Under the given hypotheses, t is strictly concave on V by Lemma 4.3. This implies that there exists $v' \in V^\circ := (v^{(1)},v^{(2)})$ such that $t(v',k_c) > v'$ (in fact, $t(v,k_c) > v$ for all $v \in V^\circ$). As t is both strictly monotone and continuous with respect to its second argument, this implies that there exists $k' < k_c$ such that $t(v',k') = v'$, which is a contradiction. We conclude that the

fixed point equation $t(v, k_c) = v$ has a unique solution, and we denote this solution by v^*. Let $w^* := k_c^2 p(v^*)$. Then by definition

$$\left.\begin{array}{rcl} (k_c)^2 q(w^*, \mathbf{1}_n) & = & v^* \\ (k_c)^2 p(v^*, \mathbf{1}_m) & = & w^* \end{array}\right\}. \tag{4.20}$$

Equation (4.20) implies, first, that

$$k_c \;=\; \sqrt{\frac{v^*}{q(w^*, \mathbf{1}_n)}} \;=\; \sqrt{\frac{w^*}{p(v^*, \mathbf{1}_m)}}$$

and, second, that

$$w^* \cdot q(w^*, \mathbf{1}_n) - v^* \cdot p(v^*, \mathbf{1}_m) \;=\; 0.$$

Next, we will show that

$$\left.\frac{\partial t(\zeta, k_c)}{\partial \zeta}\right|_{\zeta = v^*} \;=\; 1. \tag{4.21}$$

As a first step towards proving this we will establish the fact that there exists an open subinterval $S \subset I_v$, containing v^*, such that $t(\cdot, k_c)$ is defined on the whole of S. It suffices to show that $v^* > v_{\min} := \min\{v \in I_v : (v, k_c) \in D\}$. Suppose by contradiction that $v^* = v_{\min}$. Then, by monotonicity, it must be the case that $w^* = (\frac{m+n}{m}\|\Upsilon\|_\infty)^2$. But this implies that there exists $\delta > 0$ such that

$$\frac{\partial t(\zeta, k_c)}{\partial \zeta}(v) \;=\; \left((k_c)^2 \left.\frac{\partial q(\zeta, \mathbf{1}_n)}{\partial \zeta}\right|_{\zeta = k^2 p(v, \mathbf{1}_m)}\right) \cdot \left((k_c)^2 \left.\frac{\partial p(\zeta, \mathbf{1}_m)}{\partial \zeta}\right|_{\zeta = v}\right) \;>\; 1$$

for all $v \in (v^*, v^* + \delta)$. Indeed, by continuity of the partial derivative and by the fact that p is strictly concave, there exist $\delta > 0$ and $c > 0$ such that $(k_c)^2 \left.\frac{\partial p(\zeta, \mathbf{1}_m)}{\partial \zeta}\right|_{\zeta = v} > \frac{1}{c}$ and $(k_c)^2 \left.\frac{\partial q(\zeta, \mathbf{1}_n)}{\partial \zeta}\right|_{\zeta = k^2 p(v, \mathbf{1}_m)} > c$ for all $v \in (v^*, v^* + \delta)$. However, this implies that there exists $v' \in (v^*, v^* + \delta)$ such that $t(v', k_c) > v'$. Noting that $t(v, k_c) < v$ for large enough v, it follows by continuity that there exists $v^{**} > v^*$ such that $t(v^{**}, k_c) = v^{**}$. But this contradicts the uniqueness of the fixed point, and we conclude that $v^* > v_{\min}$.

We are now ready to prove equality (4.21). Suppose that $\frac{\partial t(\zeta, k_c)}{\partial \zeta}(v^*) > 1$. Then by continuity there exists $\delta > 0$ such that $t(v, k_c) > v$ for all $v \in (v^*, v^* + \delta)$. But, as we have seen above, this leads to a contradiction, and we conclude that $\frac{\partial t(\zeta, k_c)}{\partial \zeta}(v^*) \leq 1$. Now suppose that $\frac{\partial t(\zeta, k_c)}{\partial \zeta}(v^*) < 1$. Then by continuity there exists $v' < v^*$ such that $t(v', k_c) > v'$. But this implies that there exists $k' < k_c$ such that $t(v', k') = v'$, which is a contradiction by definition of k_c. We conclude that $\frac{\partial t(\zeta, k_c)}{\partial \zeta}(v^*) = 1$. Straightforward manipulation and substitution of (4.20) gives us

$$\begin{aligned} \frac{\partial t(\zeta, k_c)}{\partial \zeta}(v^*) &= \left((k_c)^2 \left.\frac{\partial q(\zeta, \mathbf{1}_n)}{\partial \zeta}\right|_{\zeta = k^2 p(v^*, \mathbf{1}_m)}\right) \cdot \left((k_c)^2 \left.\frac{\partial p(\zeta, \mathbf{1}_m)}{\partial \zeta}\right|_{\zeta = v^*}\right) \\ &= \left(\frac{v^*}{q(w^*, \mathbf{1}_n)} \left.\frac{\partial q(\zeta, \mathbf{1}_n)}{\partial \zeta}\right|_{\zeta = w^*}\right) \cdot \left(\frac{w^*}{p(v^*, \mathbf{1}_m)} \left.\frac{\partial p(\zeta, \mathbf{1}_m)}{\partial \zeta}\right|_{\zeta = v^*}\right). \end{aligned}$$

It follows that

$$\frac{\partial q(\zeta,\mathbf{1}_n)}{\partial \zeta}(w^*)\cdot\frac{\partial p(\zeta,\mathbf{1}_m)}{\partial \zeta}(v^*) \;=\; \frac{p(v^*,\mathbf{1}_m)}{v^*}\frac{q(w^*,\mathbf{1}_n)}{w^*}. \tag{4.22}$$

This concludes the proof. ∎

5. Algorithm. Theorem 4.5 states that the critical coupling of a system of coupled oscillators on a complete bipartite graph can be determined from the unique solution of the system of equations (4.17). In the present section we will show how this solution may be computed using numerical techniques.

Let $P : I_v^\circ \mapsto \mathbb{R}$ and $Q : I_w^\circ \mapsto \mathbb{R}$ be given as

$$m^2P(v) := \left(\sum_{j=1}^{m}\sqrt{1-\left(\tilde{\Omega}_j\right)^2\frac{1}{v}}\right)\left(\sum_{j=1}^{m}\frac{1}{\sqrt{1-\left(\tilde{\Omega}_j\right)^2\frac{1}{v}}}\right), \tag{5.1}$$

$$n^2Q(w) := \left(\sum_{j=1}^{n}\sqrt{1-\left(\tilde{\Upsilon}_j\right)^2\frac{1}{w}}\right)\left(\sum_{j=1}^{n}\frac{1}{\sqrt{1-\left(\tilde{\Upsilon}_j\right)^2\frac{1}{w}}}\right), \tag{5.2}$$

where, as before, $\tilde{\Omega} := \frac{m+n}{n}\Omega$ and $\tilde{\Upsilon} := \frac{m+n}{m}\Upsilon$. Using the notation

$$z_j(v) \;:=\; 1-\frac{\left(\frac{m+n}{n}\Omega_j\right)^2}{v}, \quad j=1,\ldots,m, \tag{5.3}$$

we can rewrite $\frac{\partial p(\zeta,\mathbf{1}_m)}{\partial \zeta}(v)$ as follows:

$$\frac{\partial p(\zeta,\mathbf{1}_m)}{\partial \zeta}(v) = -\frac{1}{v}\left(p(v,\mathbf{1}_m)-\frac{1}{m}\sum_{i=1}^{m}\sqrt{z_i(v)}\frac{1}{m}\sum_{j=1}^{m}\frac{1}{\sqrt{z_j(v)}}\right). \tag{5.4}$$

An analogous formula also holds for the function $q(\zeta,\mathbf{1}_n)$. Let $(v,w)\in I_v\times I_w$. Then it follows from (5.4) that (v,w) is a solution of (4.17) if and only if (v,w) satisfies

$$\left(\frac{P(v)}{p(v,\mathbf{1}_m)}-1\right)\left(\frac{Q(w)}{q(w,\mathbf{1}_n)}-1\right)=1. \tag{5.5}$$

As a first step, we introduce the function $\Gamma : I_v\times I_w \mapsto \{(\mathrm{i}),(\mathrm{ii}),(\mathrm{iii}),(\mathrm{iv})\}$, which partitions the set $I_v\times I_w$ into four disjoint subsets, as follows:

$$\Gamma(v,w) \;:=\; \begin{cases} (\mathrm{i}) & vp(v,\mathbf{1}_m)-wq(w,\mathbf{1}_n)\le 0,\ (\frac{P(v)}{p(v,\mathbf{1}_m)})-1)(\frac{Q(w)}{q(w,\mathbf{1}_n)})-1)\le 1,\\ (\mathrm{ii}) & vp(v,\mathbf{1}_m)-wq(w,\mathbf{1}_n)> 0,\ (\frac{P(v)}{p(v,\mathbf{1}_m)})-1)(\frac{Q(w)}{q(w,\mathbf{1}_n)})-1)\le 1,\\ (\mathrm{iii}) & vp(v,\mathbf{1}_m)-wq(w,\mathbf{1}_n)> 0,\ (\frac{P(v)}{p(v,\mathbf{1}_m)})-1)(\frac{Q(w)}{q(w,\mathbf{1}_n)})-1)> 1,\\ (\mathrm{iv}) & vp(v,\mathbf{1}_m)-wq(w,\mathbf{1}_n)\le 0,\ (\frac{P(v)}{p(v,\mathbf{1}_m)})-1)(\frac{Q(w)}{q(w,\mathbf{1}_n)})-1)> 1. \end{cases} \tag{5.6}$$

Note that, by definition of Γ, the solution of the system of equations (4.17) is contained in the set $\{(v,w):\Gamma(v,w)=(\mathrm{i})\}$. Figure 4 shows the sets corresponding to the different values of Γ for a particular realization of the complete bipartite graph $K_{24,17}$. We have the following result.

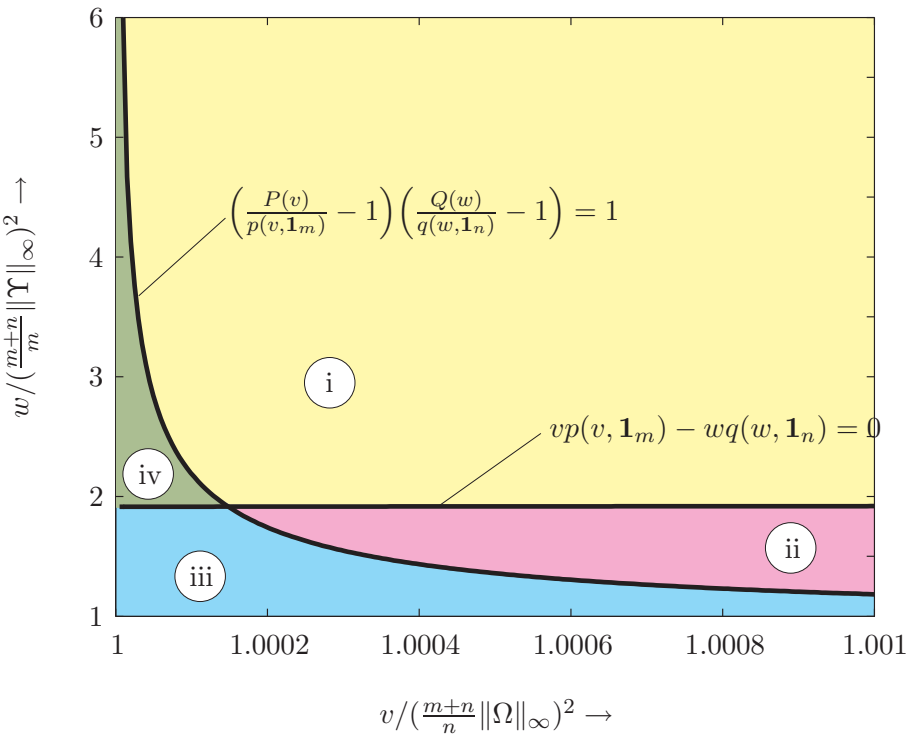


Figure 4. *Illustration of the action of the map* Γ*, given by* (5.6)*:* Γ *induces a partition on the set* $I_v^\circ \times I_w^\circ$*. In the figure, the different subsets are represented by colors, and a lowercase roman numeral, e.g.,* (i) *or* (ii)*, is used to denote the value of* Γ *on that particular subset. The bold lines correspond to the isoclines* $\{(v,w) : vp(v,\mathbf{1}_m) - wq(w,\mathbf{1}_n) = 0\}$ *and* $\{(v,w) : (\frac{P(v)}{p(v,\mathbf{1}_m)} - 1)(\frac{Q(w)}{q(w,\mathbf{1}_n)} - 1) = 1\}$ *separating the subsets.*

Proposition 5.1. *Let* (v', w') *denote the unique solution of the system of equations* (4.17)*, and let* Γ *be given by* (5.6)*. Let* $(v,w) \in I_v^\circ \times I_w^\circ$*. Then the following implications hold:*

1. *If* $\Gamma(v,w) = (\mathrm{i})$*, then* $w \geq w'$*.*
2. *If* $\Gamma(v,w) = (\mathrm{ii})$*, then* $v > v'$*.*
3. *If* $\Gamma(v,w) = (\mathrm{iii})$*, then* $w < w'$*.*
4. *If* $\Gamma(v,w) = (\mathrm{iv})$*, then* $v < v'$*.*

Proof. As a first step we show that the functions $\frac{P(\cdot)}{p(\cdot,\mathbf{1}_m)} - 1$ and $\frac{Q(\cdot)}{q(\cdot,\mathbf{1}_n)} - 1$ are monotone decreasing and nonnegative on their respective domains. The nonnegativity follows immediately from (5.4). It is also possible to show, using arguments very similar to those used in the proof of Lemma 4.3, that $\frac{\partial P(\zeta)}{\partial \zeta}(v) \leq 0$ for all $v \in I_v^\circ$ and that $\frac{\partial Q(\zeta)}{\partial \zeta}(w) \leq 0$ for all $w \in I_w^\circ$. Since p and q are monotone increasing by Lemma 4.2, it follows that $\frac{P(\cdot)}{p(\cdot,\mathbf{1}_m)}$ and $\frac{Q(\cdot)}{q(\cdot,\mathbf{1}_n)}$ are monotone decreasing. Since p and q are monotone increasing by Lemma 4.2, it follows that $\frac{P(\cdot)}{p(\cdot,\mathbf{1}_m)}$ and $\frac{Q(\cdot)}{q(\cdot,\mathbf{1}_n)}$ are monotone decreasing, and hence so are $\frac{P(\cdot)}{p(\cdot,\mathbf{1}_m)} - 1$ and $\frac{Q(\cdot)}{q(\cdot,\mathbf{1}_n)} - 1$. Now let $(v,w) \in I_v^\circ \times I_w^\circ$. It follows that

$$\left(\frac{P(v)}{p(v,\mathbf{1}_m)} - 1\right)\left(\frac{Q(w)}{q(w,\mathbf{1}_n)} - 1\right) > 1$$

if $v < v'$ and $w < w'$. Similarly, we have that

$$v \cdot p(v, \mathbf{1}_m) - w \cdot q(w, \mathbf{1}_n) \; > \; 0$$

if $v > v'$ and $w < w'$ (by monotonicity of p and q). Now suppose, for example, that $\Gamma(v, w) =$ (i). Then it follows from the above that ($v \geq v'$ OR $w \geq w'$) AND ($v \leq v'$ OR $w \geq w'$), which implies that $v = v'$ OR $w \geq w'$. Inspection shows that if $v = v'$ and $w < w'$, then necessarily

$$v \cdot p(v, \mathbf{1}_m) - w \cdot q(w, \mathbf{1}_n) \; > \; 0,$$

which contradicts the assumption that $\Gamma(v, w) =$ (i). We conclude that $w \geq w'$, which proves implication 1. The other cases follow similarly. This concludes the proof. ∎

Proposition 5.1 suggests that the solution of the system of equations (4.17) can be found numerically using a bisection algorithm. The idea is as follows. To initialize the algorithm, we pick $(v^0, w^0) \in I_v^\circ \times I_w^\circ$ such that $v^0 \geq v'$ and $w^0 \geq w'$ (more about this shortly), and we define $\Delta v := \frac{1}{2}(v^0 - a_v)$ and $\Delta w := \frac{1}{2}(w^0 - a_w)$, where a_v and a_w are given by

$$a_v := \left(\frac{m+n}{n}\|\Omega\|_\infty\right)^2, \qquad a_w := \left(\frac{m+n}{m}\|\Upsilon\|_\infty\right)^2. \tag{5.7}$$

We set $v := v^0 - \Delta_v$ and $w := w^0 - \Delta_w$. Now we invoke Proposition 5.1 to obtain information about the location of (v, w) relative to (v', w'). There are four possibilities:

(1) $\Gamma(v, w) =$ (i). In this case $w \geq w$, and we update w and Δw according to the update rule $w \mapsto w - \Delta w$, $\Delta w \mapsto \frac{1}{2}\Delta w$.
(2) $\Gamma(v, w) =$ (ii). In this case $v > v'$, and we update v and Δv according to the update rule $v \mapsto v - \Delta v$, $\Delta v \mapsto \frac{1}{2}\Delta v$.
(3) $\Gamma(v, w) =$ (iii). In this case $w < w'$, and we update w and Δw according to the update rule $w \mapsto w + \Delta w$, $\Delta w \mapsto \frac{1}{2}\Delta w$.
(4) $\Gamma(v, w) =$ (iv). In this case $v < v'$, and we update v and Δv according to the update rule $v \mapsto v + \Delta v$, $\Delta v \mapsto \frac{1}{2}\Delta v$.

We repeat until the algorithm terminates (which it does when the error first drops below a certain tolerance level). The error can be defined in different ways. We have defined it as follows:

$$\texttt{Err}(v, w) := \max\left\{|vp(v, \mathbf{1}_m) - wq(w, \mathbf{1}_n)|, \left|\frac{P(v)}{p(v, \mathbf{1}_m)} - 1\right| \cdot \left|\frac{Q(w)}{q(w, \mathbf{1}_n)} - 1\right|\right\}. \tag{5.8}$$

Before we present the actual algorithm, let us discuss how to pick v^0, w^0 such that $v^0 \geq v'$ and $w^0 \geq w'$. One approach is to try to derive analytic upper bounds on v' and w' in terms of $\|\Omega\|_\infty$ and $\|\Upsilon\|_\infty$. We propose a more practical approach, as follows. We pick any pair $(v, w) \in I_v^\circ \times I_w^\circ$. For concreteness, let us say $(v, w) := (2a_v, 2a_w)$. Next we fix $w = 2a_w$ and repeatedly increase v by a factor of 2 until $\Gamma(v, 2a_w) =$ (ii). It is easy to show that this procedure always terminates in a finite number of steps. If we let v^0 denote the value of v after termination of this procedure, then by Proposition 5.1 we have that $v^0 \geq v'$. Now we fix $v = v^0$

and repeatedly increase w by a factor of 2 until $\Gamma(v^0, w) =$ (i). Again, this procedure will always terminate in a finite number of steps. Let w^0 denote the value of w after termination of this procedure. Then by Proposition 5.1 we have that $w^0 \geq w'$. Effectively, if ρ_1 is the smallest nonnegative integer i for which $\Gamma(2^{i+1}a_v, 2a_w) =$ (ii) and ρ_2 is the smallest nonnegative integer j for which $\Gamma(2^{\rho_1+1}a_v, 2^{j+1}a_w) =$ (i), then we have that $(v^0, w^0) = (2^{\rho_1+1}a_v, 2^{\rho_2+1}a_w)$. We now have the following algorithm.

Algorithm 1.

1. $\mathtt{a_v} := (\frac{\mathtt{m}+\mathtt{n}}{\mathtt{n}}\|\Omega\|_\infty)^2$;
2. $\mathtt{a_w} := (\frac{\mathtt{m}+\mathtt{n}}{\mathtt{m}}\|\Upsilon\|_\infty)^2$;
3. $\mathtt{v} := 2 \cdot \mathtt{a_v}$;
4. $\mathtt{w} := 2 \cdot \mathtt{a_w}$;
5. $\mathtt{AbsTol} := 10^{-6}$;
6. $\mathtt{Err} := 1$;
7. `while` $\Gamma(\mathtt{v}, \mathtt{w}) <>$ (ii)
   $\mathtt{v} := 2 \cdot \mathtt{v}$;
   $\mathtt{w} := \mathtt{w}$;
8. `end`
9. `while` $\Gamma(\mathtt{v}, \mathtt{w}) <>$ (i)
   $\mathtt{v} := \mathtt{v}$;
   $\mathtt{w} := 2 \cdot \mathtt{w}$;
10. `end`
11. $\Delta_\mathtt{v} := \frac{1}{2}(\mathtt{v} - \mathtt{a_v})$;
12. $\Delta_\mathtt{w} := \frac{1}{2}(\mathtt{w} - \mathtt{a_w})$;
13. `while` $\mathtt{Err} > \mathtt{AbsTol}$
    13.1. `switch` $\Gamma(\mathtt{v}, \mathtt{w})$
      13.1.1. `case` (i)
        $\mathtt{w} := \mathtt{w} - \Delta_\mathtt{w}$;
        $\Delta_\mathtt{w} := \frac{1}{2}\Delta_\mathtt{w}$;
      13.1.2. `case` (ii)
        $\mathtt{v} := \mathtt{v} - \Delta_\mathtt{v}$;
        $\Delta_\mathtt{v} := \frac{1}{2}\Delta_\mathtt{v}$;
      13.1.3. `case` (iii)
        $\mathtt{w} := \mathtt{w} + \Delta_\mathtt{w}$;
        $\Delta_\mathtt{w} := \frac{1}{2}\Delta_\mathtt{w}$;
      13.1.4. `case` (iv)
        $\mathtt{v} := \mathtt{v} + \Delta_\mathtt{v}$;
        $\Delta_\mathtt{v} := \frac{1}{2}\Delta_\mathtt{v}$;
    13.2. `end`
    13.3. $\mathtt{Err} := \max\left\{|\mathtt{v} \cdot \mathtt{p}(\mathtt{v}, \mathbf{1}_\mathtt{m}) - \mathtt{w} \cdot \mathtt{q}(\mathtt{w}, \mathbf{1}_\mathtt{n})|,\ \left|\frac{\mathtt{P}(\mathtt{v})}{\mathtt{p}(\mathtt{v}, \mathbf{1}_\mathtt{m})} - 1\right| \cdot \left|\frac{\mathtt{Q}(\mathtt{w})}{\mathtt{q}(\mathtt{w}, \mathbf{1}_\mathtt{n})} - 1\right|\right\}$;
14. `end`

By construction of the algorithm and Proposition 5.1, v and w exponentially converge to the unique solution (v', w') of the system of equations (4.17). In the next section we will apply this algorithm in a number of different examples.

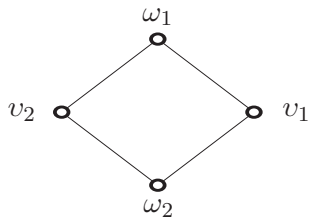


Figure 5. *The complete bipartite graph $K_{(2,2)}$. The symbol next to each node represents the intrinsic frequency of the corresponding oscillator.*

6. Numerical examples.

6.1. Example 1: $K_{2,2}$. In this first example, we consider the very simple complete bipartite graph $K_{(2,2)}$ depicted in Figure 5, which is also known as the cycle graph C_4. We pick natural frequencies $(\omega, \upsilon) \in \mathbb{R}^2 \times \mathbb{R}^2$, as follows: $\omega := (1.0 \quad 1.3)^T$, $\upsilon := (0.9 \quad 1.2)^T$. By definition of Ω and Υ we have that

$$\begin{pmatrix} \Omega(\omega,\upsilon) \\ \Upsilon(\omega,\upsilon) \end{pmatrix} := \begin{pmatrix} \omega \\ \upsilon \end{pmatrix} - \frac{1}{4}(\omega_1+\omega_2+\upsilon_1+\upsilon_2)\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -0.1 \\ 0.2 \\ -0.2 \\ 0.1 \end{pmatrix}.$$

Our objective is to determine the smallest $k \geq 0$ for which the system of differential equations

$$\left.\begin{array}{rcl} \dot{\theta}_1(t) &=& \omega_1 + \frac{k}{4}\sin(\phi_1(t)-\theta_1(t)) + \frac{k}{4}\sin(\phi_2(t)-\theta_1(t)) \\ \dot{\theta}_2(t) &=& \omega_2 + \frac{k}{4}\sin(\phi_1(t)-\theta_2(t)) + \frac{k}{4}\sin(\phi_2(t)-\theta_2(t)) \\ \dot{\phi}_1(t) &=& \upsilon_1 + \frac{k}{4}\sin(\theta_1(t)-\phi_1(t)) + \frac{k}{4}\sin(\theta_2(t)-\phi_1(t)) \\ \dot{\phi}_2(t) &=& \upsilon_2 + \frac{k}{4}\sin(\theta_1(t)-\phi_2(t)) + \frac{k}{4}\sin(\theta_2(t)-\phi_2(t)) \end{array}\right\} \tag{6.1}$$

admits a GPLS. To investigate the phase-locking behavior of this system, we integrate (6.1) twice, first with $k = 0.55$ and then with $k = 0.65$. The results are given in Figures 6 and 7. When $k = 0.55$, we observe that the magnitudes of the order parameters do not converge but rather oscillate between the extrema 0 and 1. If a GPLS would exist for this value of k, and if it were stable in the sense that its equivalence class was a stable manifold, then we might expect that at least for some initial conditions the order parameters would converge. This is, in fact, what we observe in Figure 7. On the other hand, convergence of the magnitude of the order parameters is a necessary but not a sufficient condition for the existence of a GPLS. In other words, these plots are not conclusive insofar as the (non)existence of GPLS is concerned. To determine whether a GPLS exists in each case we examine the graph of $t(\cdot, k)$ for $k = 0.55$ and $k = 0.65$ to see whether, for one value of k or the other, it has a fixed point on the interval I_v. If it does, then we know that a GPLS exists for the corresponding value of k (and, in fact, for any k greater than or equal to it); if does not, then we know that no GPLS exists for $k \leq 0.65$. The graphs of $t(\cdot, 0.55)$ and $t(\cdot, 0.65)$ are given in Figure 8. We observe that $t(\cdot, 0.55)$ does not have a fixed point (in the given interval), whereas $t(\cdot, 0.65)$ does. From this we deduce that $0.55 < k_c \leq 0.65$. To compute the exact value of k_c, we

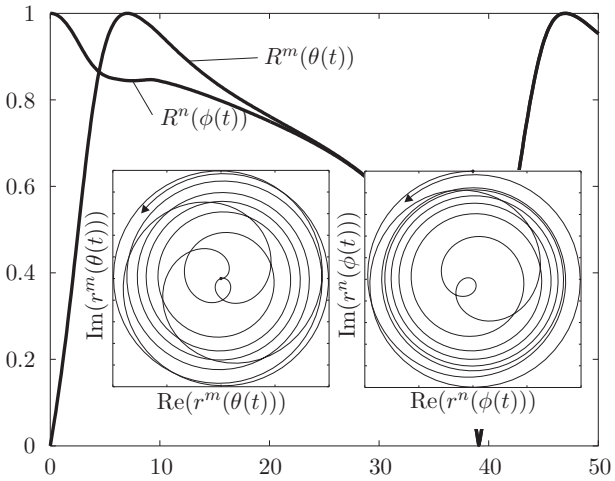


Figure 6. *Case $k = 0.55$: No GPLS exists. The graph shows the time evolution of $R^m(\theta)$ and $R^n(\phi)$, where $(\theta(\cdot), \phi(\cdot))$ is the solution to the system (6.1) subject to the initial condition $(\theta(0), \phi(0)) = ((\frac{1}{2}\pi\ -\frac{1}{2}\pi), (\frac{1}{2}\pi\ \frac{1}{2}\pi))$. Inset: The evolution of the order parameters $r^m(\theta)$ (left panel) and $r^n(\phi)$ (right panel) in the complex plane.*

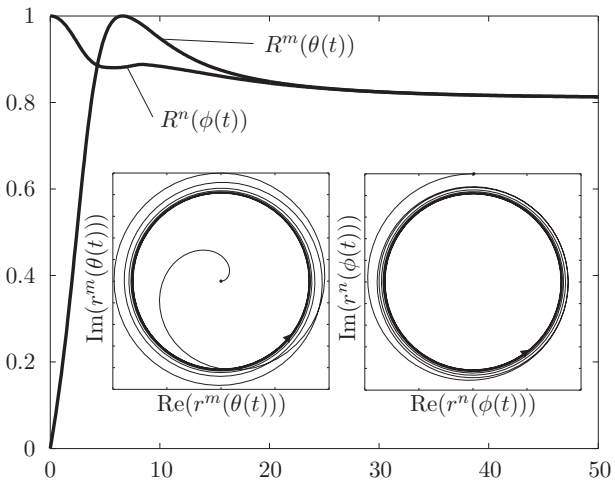


Figure 7. *Case $k = 0.65$: The system converges to a GPLS. The graph shows the time evolution of $R^m(\theta)$ and $R^n(\phi)$, where $(\theta(\cdot), \phi(\cdot))$ is the solution to the system (6.1) subject to the initial condition $(\theta(0), \phi(0)) = ((\frac{1}{2}\pi\ -\frac{1}{2}\pi), (\frac{1}{2}\pi\ \frac{1}{2}\pi))$. Inset: The evolution of the order parameters $r^m(\theta)$ (left panel) and $r^n(\phi)$ (right panel) in the complex plane.*

Table 1
The error term $\texttt{Err}(\texttt{v}, \texttt{w})$ *vs. iteration number.*

1	2	3	4	5	6	7	8	9	10	11
0.859	0.750	0.571	0.278	0.179	0.110	0.012	0.053	0.518	0.158	0.040

invoke Algorithm 2 (see Appendix A). It turns out that for this example $k_c = 0.6298$. Table 1 shows the error $\texttt{Err}(\texttt{v}, \texttt{w})$, which was defined in (5.8) as a function of the number of iterations.

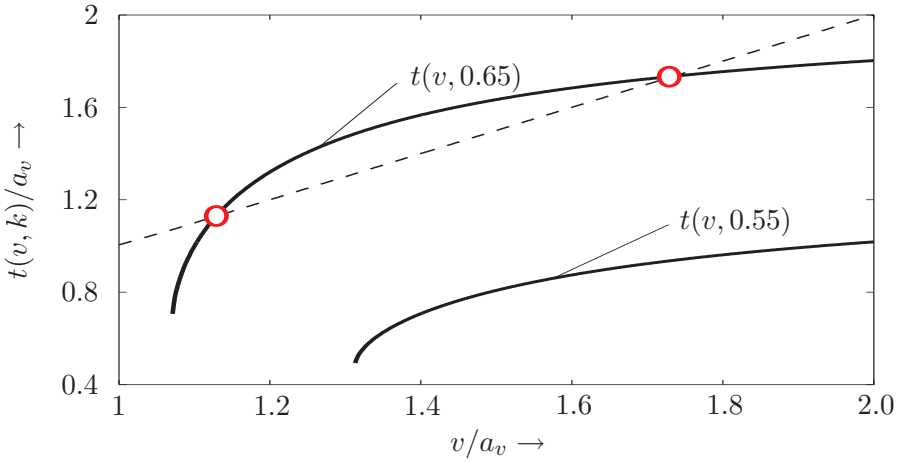


Figure 8. *The graphs of* $t(v, 0.55)$ *and* $t(v, 0.65)$ *for selected values of* v *in the interval* $[a_v, 2a_v]$. *The dashed line corresponds to the identity map* $\mathrm{Id} : \mathbb{R} \mapsto \mathbb{R}$. *The red circles marking the intersections of* $t(v, 0.65)$ *with* $\mathrm{Id}(v)$ *correspond to GPLS.*

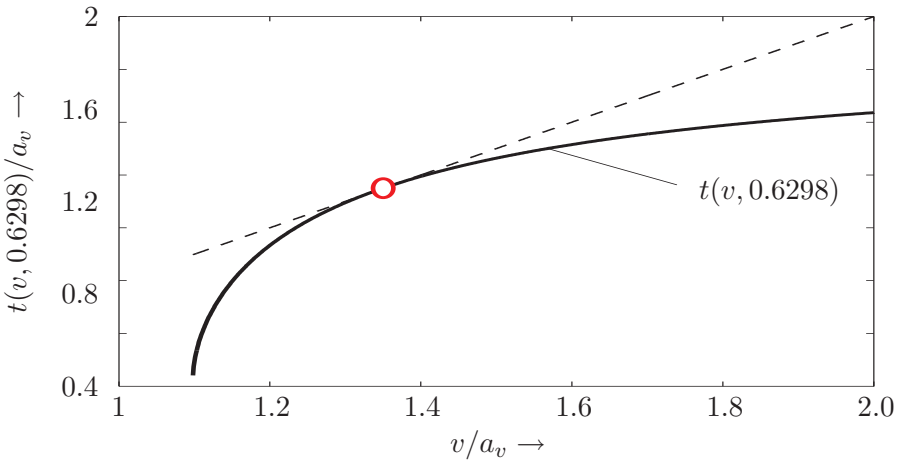


Figure 9. *The graph of* $t(v, 0.6298)$ *for selected values of* v *in the interval* $[a_v, 2a_v]$. *The dashed line corresponds to the identity map* $\mathrm{Id} : \mathbb{R} \mapsto \mathbb{R}$. *The red circle marking the intersection of* $t(v, 0.6298)$ *with* $\mathrm{Id}(v)$ *defines a unique fixed point* v' *of* $t(\cdot, 0.6298)$. *Note that* $\frac{\partial t(\zeta, 0.6298)}{\partial \zeta}\big|_{\zeta = v'} = 1$.

We observe that the error decreases rapidly, as expected, though not monotonically. To check whether the given answer is correct we examine the graph of $t(v, 0.6298)$, shown in Figure 9. Inspection shows that for this value of k the map $t(\cdot, k)$ has exactly one fixed point v^* and $\frac{\partial t(\zeta, k)}{\partial \zeta}\big|_{\zeta = v'} = 1$, as required.

6.2. Example 2a: $K_{27,13}$, unimodal distribution. In this example we compute the critical coupling for a particular realization of the complete bipartite graph $K_{27,13}$. The individual frequencies ω_i, $i = 1, \ldots, 27$, and v_i, $i = 1, \ldots, 13$, used in this example are drawn from a normal distribution with zero mean and unit variance. For the given vector of frequencies, the value of the critical coupling predicted by our method is 6.2943. To check whether this

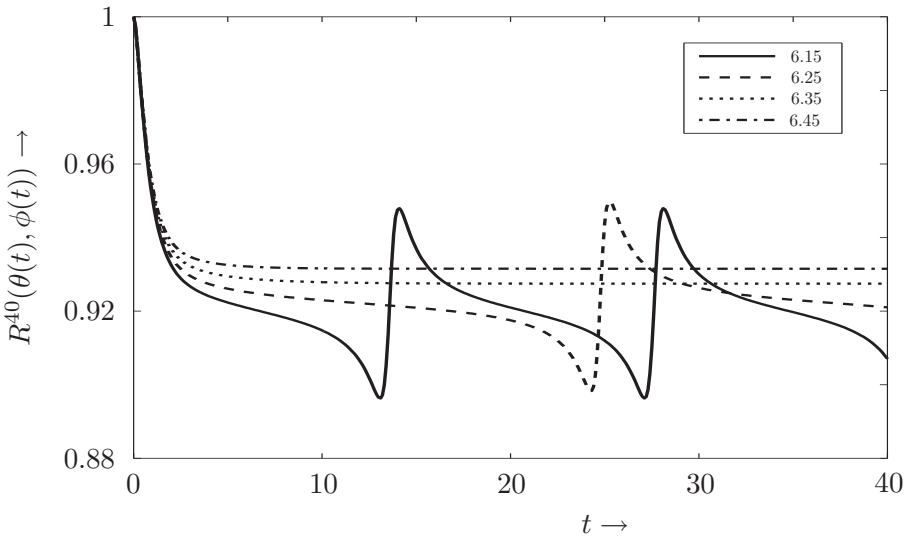


Figure 10. *The graph of* $R^{40}(\theta^{(k)}(t), \phi^{(k)}(t))$ *for* $k \in \{6.15, 6.25, 6.35, 6.45\}$*, where* $(\theta^{(k)}, \phi^{(k)})$ *denotes the solution of the bipartite system* (3.1)–(3.2) *for a given value of* k *subject to the initial condition* $(\theta(0), \phi(0)) = (0_{27\times 1}, 0_{13\times 1})$.

value is consistent with numerical evidence, we integrate the system (3.1)–(3.2) four times, using a different value of k each time ($k = 6.15, 6.25, 6.35, 6.45$) while starting from the same initial condition $(\theta(0), \phi(0)) = (0_{27\times 1}, 0_{13\times 1})$. For each of the trajectories that we obtain, we compute the corresponding order parameter $r^{40}(\theta, \phi)$, the magnitude of which is shown as a function of time in Figure 10. We observe that for the two subcritical values of the coupling ($k = 6.15, 6.25$) the magnitude of the order parameter does not converge to a constant, while for the two supercritical values ($k = 6.35, 6.45$), it does. This agrees well with our expectations, which suggests that our estimate of the critical coupling is correct.

6.3. Example 2b: $K_{231,117}$, bimodal distribution. This example is similar to Example 2a, except that in this case the frequencies are picked by drawing from a bimodal distribution, as follows. First we form the auxiliary vector z_i, $i = 1, \ldots, 348$, by drawing samples from a uniform distribution on $[0, 2\pi]$. Then we form the vectors ω, υ by defining $\omega_i := \sin(z_i)$, $i = 1, \ldots, 231$, and $\upsilon_i := \sin(z_{i+231})$, $i = 1, \ldots, 117$. Figure 11 shows the distribution of the oscillator frequencies. For this case, the value of the critical coupling predicted by our method is 3.5299. Like in the previous example, we integrate the system multiple times from the same initial condition, using a different value of k each time. The results are shown in Figure 12. As expected, the order parameter converges only for supercritical values of the coupling coefficient ($k = 3.58$, $k = 3.68$). This concludes the example.

6.4. Example 3: $K_{m,n}$ and K_{m+n}. In this example we use Algorithms 1 and 2 to compare the critical coupling of a complete bipartite graph $K_{m,n}$ with the critical coupling of the associated complete graph K_{m+n}. More specifically, we compare the critical coupling of $K_{2i,3i}$ with that of K_{5i}; that of $K_{2i,4i}$ with that of K_{6i}; and that of $K_{3i,4i}$ with that of K_{7i} for selected values of i in the interval $[1, 100]$. The intrinsic frequencies of the respective graphs are picked as follows. We construct the vectors $\alpha \in \mathbb{R}^{300}$ and $\beta \in \mathbb{R}^{400}$ by drawing from a

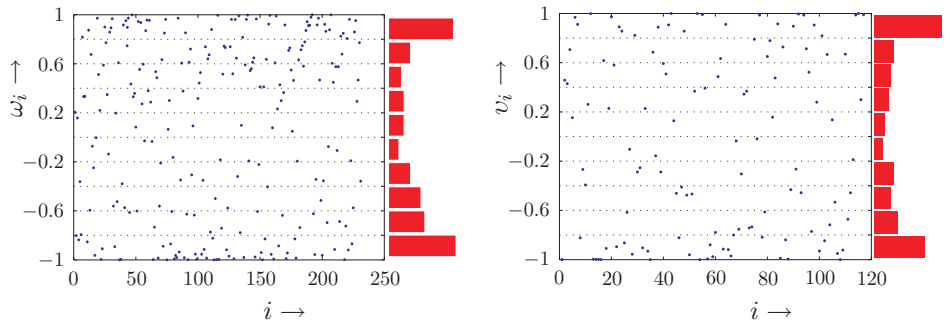


Figure 11. *The intrinsic frequencies in Example* 2b. *The panel on the left shows the distribution of* $\{\omega_i, i = 1, \ldots, 231\}$*; the panel on the right the distribution of* $\{\upsilon_i, i = 1, \ldots, 117\}$. *The corresponding histograms are shown in red.*

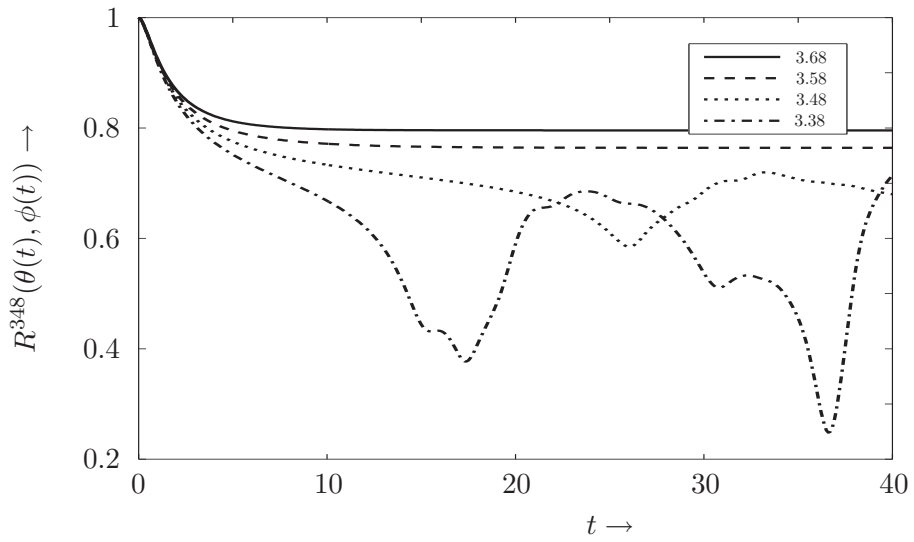


Figure 12. *The graph of* $R^{348}(\theta^{(k)}(t), \phi^{(k)}(t))$ *for* $k \in \{3.38, 3.48, 3.58, 3.68\}$*, where* $(\theta^{(k)}, \phi^{(k)})$ *denotes the solution of the bipartite system* (3.1)–(3.2) *for a given value of* k *subject to the initial condition* $(\theta(0), \phi(0)) = (0_{231\times1}, 0_{117\times1})$.

uniform distribution on $[0, 1]$, respectively 300 and 400 times. Then for the set of complete bipartite graphs $\{K_{m,n} : (m, n) \in \{(2, 4), (2, 3), (3, 4)\}\}$ we define intrinsic frequencies $\omega \in \mathbb{R}^m$ and $\upsilon \in \mathbb{R}^n$,

$$\begin{aligned} \omega_i &:= \alpha_i, \quad i = 1, \ldots, m, \\ \upsilon_i &:= \beta_i, \quad i = 1, \ldots, n, \end{aligned}$$

and for the set of complete graphs $\{K_{m+n} : (m, n) \in \{(2, 4), (2, 3), (3, 4)\}\}$ we define $\omega \in \mathbb{R}^{m+n}$,

$$\omega_i := \begin{cases} \alpha_i, & i = 1, \ldots, m, \\ \beta_{i-m}, & i = m + 1, \ldots, m + n. \end{cases}$$

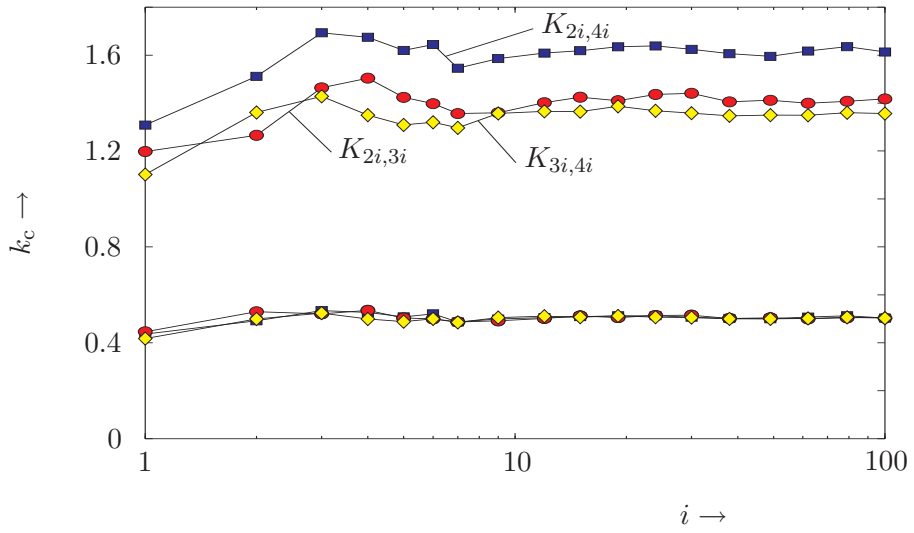


Figure 13. *The figure shows the values of the critical coupling of the complete bipartite graphs $K_{2i,4i}$ (blue square), $K_{2i,3i}$ (red circle), and $K_{3i,4i}$ (yellow diamond) along with those of the corresponding complete graphs K_{6i} (blue square), K_{5i} (red circle), and K_{7i} (yellow diamond) for selected values of i.*

Finally, for each pair of graphs $(K_{m,n}, K_{m+n})$ we compute the critical coupling. The results are depicted in Figure 13. We observe that, in all cases, the critical coupling of the Kuramoto model on the complete graph is strictly smaller than that of its counterpart on a complete bipartite graph. The data also suggest—though we do not have proof to show this—that in each case the value of the critical coupling tends to a limit as the number of oscillators tends to infinity, with the limit depending on the ratio of m and n. As regards the first observation, on the critical coupling of a complete graph being smaller than that of its complete bipartite counterpart, we remark that, as it is, the comparison is not entirely fair because the normalization constant in the bipartite system (3.1)–(3.2) scales with the total number of oscillators $(n+m)$ and not with the number of neighbors (n or m), as it does in the case of a complete graph (where the ratio between the two tends to 1 as $n+m$ tends to infinity). To ensure a fair comparison we consider the modified complete bipartite system given below, where, consistent with the Kuramoto model on a complete graph, the normalization constant is directly proportional to the number of neighbors:

$$\dot{\theta}_i(t) = \omega_i + \frac{k}{n+1}\sum_{j=1}^{n}\sin(\phi_j(t)-\theta_i(t)), \quad i=1,\dots,m, \tag{6.2}$$

$$\dot{\phi}_i(t) = \upsilon_i + \frac{k}{m+1}\sum_{j=1}^{m}\sin(\theta_j(t)-\phi_i(t)), \quad i=1,\dots,n. \tag{6.3}$$

Computing the critical coupling for this system turns out to be equivalent to solving the familiar system of equations

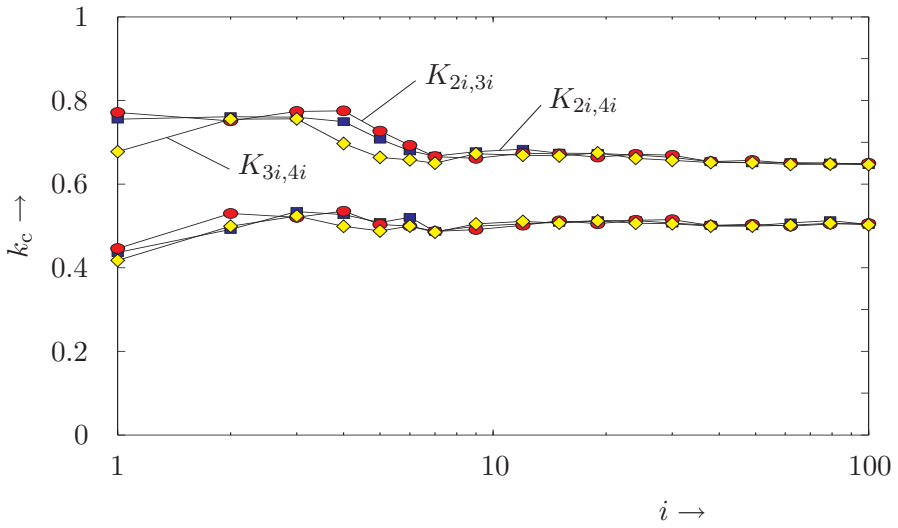


Figure 14. *The same as Figure 13, but the data in this figure apply to the modified bipartite system (6.2)–(6.3) rather than the original system (3.1)–(3.2). Note the difference in scale.*

$$\frac{k}{m+n}\sum_{j=1}^{n}\sin(\xi_i-\nu_j)=-\Omega_i,\quad i=1,\dots,m,$$
$$\frac{k}{m+n}\sum_{j=1}^{m}\sin(\nu_i-\xi_j)=-\Upsilon_i,\quad i=1,\dots,n,$$

where, in this case, Ω and Υ are given by

$$\Omega:=\left(\frac{n+1}{m+n}\right)\left(\omega-\frac{(m+1)\sum_{j=1}^{n}v_j+(n+1)\sum_{j=1}^{m}\omega_j}{(m+1)n+(n+1)m}\mathbf{1}_m\right),$$
$$\Upsilon:=\left(\frac{m+1}{m+n}\right)\left(v-\frac{(m+1)\sum_{j=1}^{n}v_j+(n+1)\sum_{j=1}^{m}\omega_j}{(m+1)n+(n+1)m}\mathbf{1}_n\right),$$

When we compare the critical coupling of this system (Figure 14) with that of the original system (Figure 13) we observe that (a) the values of the critical coupling of the modified system are closer to those of the corresponding complete system than the values of the original system are, yet the critical coupling of the complete system is still significantly smaller in all cases; (b) as was the case for the original system, the data suggests that the values of the critical coupling of the modified system tend to a limit as the size of the system tends to infinity. Here, however, the limit appears to be independent of the ratio of m and n.

7. Conclusion. In this paper, we have derived new results on the critical coupling of the Kuramoto model for the case of a complete bipartite graph. In particular, we have shown that the value of the critical coupling can be obtained exactly by solving a system of two nonlinear equations that do not depend on the coupling coefficient. We showed that said system of equations can be solved using efficient numerical techniques, and we proposed an

algorithm based on the same. Using this algorithm, we were able to investigate numerically the relation between the critical coupling of a complete bipartite graph and the critical coupling of the associated complete graph for networks with large numbers of nodes (oscillators). After renormalization of the coupling strength, we found that the critical coupling of the complete graph is significantly smaller than that of the bipartite graph. We also found that the critical coupling of the appropriately renormalized bipartite graph tends to a limit as the number of oscillators tends to infinity. The numerical results suggest that this limit is independent of the ratio of the number of oscillators in each partition, but further analysis is required to show that this is indeed the case.

Appendix A. Global phase-locking on complete graphs. In this section we briefly review some relevant results on global phase-locking on complete graphs. The Kuramoto model for a system of coupled oscillators on a complete graph K_n, $n \geq 2$, is given as

$$\dot{\theta}_i = \omega_i + \frac{k}{n} \sum_{j=1}^{n} \sin(\theta_j - \theta_i), \quad i = 1, \ldots, n, \tag{A.1}$$

where $\theta_i(\cdot) \in \mathbb{R}$ and $\omega_i \in \mathbb{R}$ denote the phase and intrinsic (or natural) frequency of oscillator i, respectively, and $k \in \mathbb{R}_{\geq 0}$ denotes the coupling coefficient. One of the central problems in the literature on the Kuramoto model [17, 18, 1] has been to find conditions on the coupling strength k and the density function of intrinsic frequencies, sometimes denoted by $g(\omega)$, under which the system (A.1) exhibits synchronous behavior. This problem was essentially solved by Kuramoto himself, although subsequent to his work numerous others have contributed to a more complete solution in different ways. For an overview of classical results, see [43]; for a more recent survey, consult [1]. Although very powerful, Kuramoto's original approach has two obvious drawbacks. First, it requires the density function $g(\omega)$ to be symmetric about the mean, which excludes certain distributions of practical interest. Second, it works only for settings with infinitely many oscillators (the so-called thermodynamic limit case) since the Fokker–Planck equation describing the evolution of the oscillator density function does not, by its very nature, capture finite-size effects. The extension of classical results, on the value of the critical coupling and the instability of the incoherent state, for example, to the finite-dimensional case has proven difficult, and papers on the same are few [35, 36, 24, 25]. (See also [43] for a discussion of open problems in the finite-dimensional theory.) In recent work [49], the authors of the present paper showed that the problem of determining the critical coupling of the Kuramoto model (A.1) boils down to solving a particular nonlinear equation, the solution of which can be obtained very efficiently using appropriate numerical tools. Aside from numerical implications, this enables us to prove a convergence result which may help bridge the gap between the finite-dimensional and the infinite-dimensional theory, as we hope to demonstrate in a future publication. We shall now briefly review the main contribution of the aforementioned work. But first we introduce some terminology.

Definition 9 (GPLS, complete graph). *Let $\theta^0 \in \mathbb{R}^n$, and let $\{\theta(t) : t \geq t_0\}$ be the solution of the system (A.1) with initial condition $\theta(t_0) = \theta^0$. We say θ^0 is a GPLS if*

$$\theta_i(t) - \theta_j(t) \;=\; \theta_i^0 - \theta_j^0$$

for all $t \geq t_0$.

The existence of GPLS in the Kuramoto model (A.1) is dependent on the value of the coupling strength k. The smallest value of k for which phase-locked states exist defines the critical coupling, as follows.

Definition 10 (critical coupling, complete graph). *Let $n \in \mathbb{N}$, $n \geq 2$, and let $\omega \in \mathbb{R}^n$ be given. Consider the Kuramoto model of coupled oscillators on the complete graph K_n (A.1). For this model, the critical coupling, k_c, is defined as follows:*

$$k_c := \min_k \{k \in \mathbb{R}_{\geq 0} : \textit{the system (A.1) admits a GPLS}\}.$$

Define $\Omega : \mathbb{R}^n \mapsto \mathbb{R}^n$,

$$\Omega_i(\omega) := \omega_i - \frac{1}{n}\sum_{j=1}^{n} \omega_j, \quad 1 \leq i \leq n. \tag{A.2}$$

We are now ready to reformulate the main result of [49], as follows.

Theorem A.1. *Let $\omega \in \mathbb{R}^n$ be given, and suppose $\Omega(\omega) \neq 0$. Then the equation*

$$2\frac{1}{n}\sum_{j=1}^{n}\sqrt{1-\left(\frac{\Omega_j(\omega)}{u}\right)^2} = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{\sqrt{1-\left(\frac{\Omega_j(\omega)}{u}\right)^2}} \tag{A.3}$$

has a unique solution $u^* \in (\|\Omega(\omega)\|_\infty, 2\|\Omega(\omega)\|_\infty]$, and we have that

$$k_c = \frac{u^*}{\frac{1}{n}\sum_{j=1}^{n}\sqrt{1-\left(\frac{\Omega_j(\omega)}{u^*}\right)^2}}. \tag{A.4}$$

The critical point u^* can be computed numerically to within user-defined precision `AbsTol` using the following algorithm.

Algorithm 2.

1. $\mathtt{a} := \|\Omega\|_\infty$;
2. $\mathtt{u} := 2 \cdot \mathtt{a}$;
3. $\mathtt{AbsTol} := 10^{-6}$;
4. $\mathtt{Err} := 1$;
5. $\Delta_\mathtt{u} := \frac{1}{2}(\mathtt{u} - \mathtt{a})$;
6. `while` $|\mathtt{Err}| > \mathtt{AbsTol}$
   - 6.1. $\mathtt{Err} := \sum_\mathtt{j}\sqrt{1-\left(\frac{\Omega_\mathtt{j}}{\mathtt{u}}\right)^2} - \frac{1}{2}\sum_\mathtt{j}\frac{1}{\sqrt{1-\left(\frac{\Omega_\mathtt{j}}{\mathtt{u}}\right)^2}}$;
   - 6.2. `if` $\mathtt{Err} \geq 0$
     $\mathtt{u} := \mathtt{u} - \Delta_\mathtt{u}$;
     $\Delta_\mathtt{u} := \frac{1}{2}\Delta_\mathtt{u}$;
   - 6.3. `else`
     $\mathtt{u} := \mathtt{u} + \Delta_\mathtt{u}$;
     $\Delta_\mathtt{u} := \frac{1}{2}\Delta_\mathtt{u}$;
   - 6.4. `end`
7. `end`

Appendix B. Proofs.

B.1. Proof of Lemma 3.3. Let $x^* = (y^*; 0)$. Then as $\sum_{j=1}^{m+n} F_j(\cdot) \equiv 0$, it follows that $F(x^*) = 0$ and, under the hypotheses of the lemma, $R^m(\Pi^1(x^*))R^n(\Pi^2(x^*)) \neq 0$. Thus, from Proposition 3.2 we have that $\sin(x_j^* - 0) = 0$ for $1 \leq j \leq m + n - 1$.

Inspection shows that $J(y^*)$ is of the form

$$J(y^*) = \frac{1}{m+n}\begin{pmatrix} A & B \\ B^T & C \end{pmatrix},$$

where A and C are diagonal matrices of size $m \times m$ and $(n-1) \times (n-1)$, respectively, whose entries are given as

$$\begin{aligned} A_{ii} &:= -\sum_{j=1}^{n} \cos(x_{m+j}^* - x_i^*) \\ &= -\sum_{j=1}^{n} \cos(x_{m+j}^*)\cos(x_i^*), \\ C_{ii} &:= -\sum_{j=1}^{m} \cos(x_j^* - x_{m+i}^*) \;=\; -\sum_{j=1}^{m} \cos(x_j^*)\cos(x_{m+i}^*). \end{aligned}$$

The submatrix B is a matrix of dimension $m \times (n-1)$ whose entries satisfy the following relation:

$$\begin{aligned} B_{ij} &:= \cos(x_{m+j}^* - x_i^*) \\ &= \cos(x_{m+j}^*)\cos(x_i^*). \end{aligned}$$

Using the identities

$$R^m(\Pi^1(x^*))^2 = \left(\frac{1}{m}\sum_{j=1}^{m} \cos(x_j^*)\right)^2, \tag{B.1}$$

$$R^n(\Pi^2(x^*))^2 = \left(\frac{1}{n}\sum_{j=1}^{n} \cos(x_{m+j}^*)\right)^2, \tag{B.2}$$

we rewrite A_{ii} and C_{ii} as

$$A_{ii} \;=\; -nR^n(\Pi^2(x^*))\cos(x_i^*), \qquad C_{ii} \;=\; -mR^m(\Pi^1(x^*))\cos(x_{m+i}^*).$$

This shows that, under the hypotheses of the proposition, all diagonal entries of A and C are nonzero. If we define the nonsingular diagonal matrix $D = \mathrm{diag}(\cos(x_1^*), \ldots, \cos(x_{m+n-1}^*))$, then $J(y^*)$ is nonsingular if and only if $S = D^{-1}J(y^*)$ is nonsingular. We shall show that S is nonsingular by contradiction. Suppose there is some nonzero vector $z \in \mathbb{R}^{m+n-1}$ with

$Sz = 0$. Then by inspection the components of z must satisfy

$$az_i + \sum_{j=1}^{n-1} b_j z_{m+j} = 0 \quad \text{for } 1 \le i \le m, \tag{B.3}$$

$$cz_i + \sum_{j=1}^{m} d_j z_j = 0 \quad \text{for } m \le i \le m+n-1, \tag{B.4}$$

where $a = -\sum_{j=1}^{n} \cos(x^*_{m+j})$, $b_j = \cos(x^*_{m+j})$, $c = -\sum_{j=1}^{m} \cos(x^*_j)$, $d_j = \cos(x^*_j)$. From (B.1), (B.2) it follows that a and c are nonzero. Now note the following easily verifiable facts:

(i) If $z_i = 0$ for any i with $1 \le i \le m+n-1$, then $z = 0$. Thus we can assume that each component of z is nonzero.
(ii) There are constants $\kappa_1, \kappa_2 \in \mathbb{R}$ such that $z_i = \kappa_1$ for $1 \le i \le m$ and $z_i = \kappa_2$ for $m \le i \le m+n-1$.
(iii) $a + \sum_{j=1}^{n-1} b_j = -1$.
(iv) $c + \sum_{j=1}^{m} d_j = 0$.

We can assume without loss of generality that $\kappa_1 = 1$. Then it follows from (B.4) that $c + \kappa_2(\sum_{j=1}^{m} d_j) = 0$. Thus (iv) implies that $\kappa_2 = 1$. But then (B.3) implies that $a + \sum_{j=1}^{n-1} b_j = 0$, which contradicts (iii) above. Thus, there is no nonzero $z \in \mathbb{R}^{m+n-1}$ satisfying $Sz = 0$, and hence S and $J(y^*)$ are nonsingular as claimed. ∎

B.2. Proof of Lemma 4.2. We prove the result for p. The result for q follows by analogy. Note that $p(\cdot, \mathbf{1}_m)$ is continuous on I_v and differentiable on I_v°. We claim that $\frac{\mathrm{d}p(\xi, \mathbf{1}_m)}{\mathrm{d}\xi}(v) \ge 0$ on I_v°. Note that

$$m^2 \frac{\mathrm{d}p(\xi, \mathbf{1}_m)}{\mathrm{d}\xi}(v) = \frac{1}{v^2} \left(\underbrace{\left(\sum_{j=1}^{m} \sqrt{1 - \frac{\tilde{\Omega}_j^2}{v}} \right) \left(\sum_{j=1}^{m} \frac{\tilde{\Omega}_j^2}{\sqrt{1 - \frac{\tilde{\Omega}_j^2}{v}}} \right)}_{A} - \left(\sum_{j=1}^{m} \tilde{\Omega}_j \right)^2 \right) \tag{B.5}$$

for $v \in I_v^\circ$, where $\tilde{\Omega} := \left(\frac{m+n}{n}\right) \Omega$. Using the notation

$$y_i(v) := \sqrt{1 - \frac{\tilde{\Omega}_i^2}{v}}, \quad 1 \le i \le m,$$

we rewrite the term A, as follows:

$$\left(\sum_{i=1}^{m} y_i(v) \right) \left(\sum_{j=1}^{m} \frac{\tilde{\Omega}_j^2}{y_j(v)} \right) = \left(\sum_{i=1}^{m} \left(\sqrt{y_i(v)} \right)^2 \right) \left(\sum_{j=1}^{m} \left(\frac{\tilde{\Omega}_j}{\sqrt{y_j(v)}} \right)^2 \right).$$

By the Cauchy–Schwarz inequality we have that

$$\left(\sum_{i=1}^{m} \left(\sqrt{y_i(v)} \right)^2 \right) \left(\sum_{j=1}^{m} \left(\frac{\tilde{\Omega}_j}{\sqrt{y_j(v)}} \right)^2 \right) \ge \left(\sum_{j=1}^{m} \sqrt{y_j(v)} \frac{\tilde{\Omega}_j}{\sqrt{y_j(v)}} \right)^2 = \left(\sum_{j=1}^{m} \tilde{\Omega}_j \right)^2$$

for all $v \in I_v^\circ$, where equality holds (for a fixed v) if and only if there exists $\alpha \in \mathbb{R}$ such that $y_i(v) = \alpha\,\Omega_i$ for all i. Inspection shows that this condition is satisfied if and only if $\Omega_i = \Omega_j$ for all (i,j), in which case $p(\cdot, \mathbf{1}_m)$ is a constant function ($p(\cdot, \mathbf{1}_m) \equiv 1$). This concludes the proof. ∎

B.3. Proof of Lemma 4.3. We prove the result for p. The proof for q follows by analogy. Suppose $\Omega_i(\omega, v) = \Omega_j(\omega, v)$ for all (i,j). Then $p(\cdot, \mathbf{1}_m)$ is a constant function by Lemma 4.2. Hence $p(\cdot, \mathbf{1}_m)$ is concave but not strictly concave. Now suppose $\Omega_i(\omega, v) \neq \Omega_j(\omega, v)$, for some (i,j). We claim that $\frac{\partial^2 p(\zeta, \mathbf{1}_m)}{\partial \zeta^2}(v) < 0$ for all $v \in I_v^\circ$.

It follows easily from (B.5) that

$$m^2 \frac{\partial p(\xi, \mathbf{1}_m)}{\partial \xi}(v) = \frac{1}{v^2} \sum_{i,j=1,\, i\neq j}^{m} \left(\tilde{\Omega_j}^2 \sqrt{\frac{v - \tilde{\Omega_i}^2}{v - \tilde{\Omega_j}^2}} - \tilde{\Omega_i}\tilde{\Omega_j} \right), \tag{B.6}$$

and hence we may write

$$m^2 \frac{\partial^2 p(\xi, \mathbf{1}_m)}{\partial \xi^2}(v) = -\frac{2}{v} m^2 \frac{\partial p(\xi, \mathbf{1}_m)}{\partial \xi}(v) + \frac{1}{2v^2} B, \tag{B.7}$$

where the term B is given by

$$\sum_{i,j=1,\, i\neq j}^{m} \tilde{\Omega}_j^2 \left((v - \tilde{\Omega}_i^2)^{-1/2} (v - \tilde{\Omega}_j^2)^{-1/2} - (v - \tilde{\Omega}_i^2)^{1/2} (v - \tilde{\Omega}_j^2)^{-3/2} \right). \tag{B.8}$$

As $\Omega_i(\omega, v) \neq \Omega_j(\omega, v)$, for some (i,j), it follows from Lemma 4.2 that the first term on the right-hand side of (B.7) is negative for all $v \in I_v^\circ$. Thus, to complete the proof, it is enough to show that $B \leq 0$ for $v \in I_v^\circ$.

Note first that, by rearranging terms, we can write

$$\begin{aligned} B &= \sum_{i,j=1,\, i\neq j}^{m} \tilde{\Omega}_j^2 (v - \tilde{\Omega}_i^2)^{-1/2} (v - \tilde{\Omega}_j^2)^{-1/2} \left(1 - \frac{v - \tilde{\Omega}_i^2}{v - \tilde{\Omega}_j^2} \right) \\ &= \sum_{i,j=1,\, i<j}^{m} (v - \tilde{\Omega}_i^2)^{-1/2} (v - \tilde{\Omega}_j^2)^{-1/2} T_{ij}, \end{aligned} \tag{B.9}$$

where

$$T_{ij} = \tilde{\Omega}_j^2 \left(1 - \frac{v - \tilde{\Omega}_i^2}{v - \tilde{\Omega}_j^2} \right) + \tilde{\Omega}_i^2 \left(1 - \frac{v - \tilde{\Omega}_j^2}{v - \tilde{\Omega}_i^2} \right) \tag{B.10}$$

$$= \tilde{\Omega}_j^2 \left(\frac{\tilde{\Omega}_i^2 - \tilde{\Omega}_j^2}{v - \tilde{\Omega}_j^2} \right) + \tilde{\Omega}_i^2 \left(\frac{\tilde{\Omega}_j^2 - \tilde{\Omega}_i^2}{v - \tilde{\Omega}_i^2} \right) \tag{B.11}$$

$$= (\tilde{\Omega}_i^2 - \tilde{\Omega}_j^2) \left(\frac{\tilde{\Omega}_j^2}{v - \tilde{\Omega}_j^2} - \frac{\tilde{\Omega}_i^2}{v - \tilde{\Omega}_i^2} \right). \tag{B.12}$$

It can be readily verified that

$$(\tilde{\Omega}_i^2 - \tilde{\Omega}_j^2) \;\geq\; 0 \;\Leftrightarrow\; \frac{\tilde{\Omega}_j^2}{v - \tilde{\Omega}_j^2} - \frac{\tilde{\Omega}_i^2}{v - \tilde{\Omega}_i^2} \;\leq\; 0$$

and hence $T_{ij} \leq 0$ for $1 \leq i < j \leq m$. But it then follows immediately from (B.9) that $B \leq 0$ for $v \in I_v^\circ$, and hence that $\frac{\partial^2 p(\zeta, \mathbf{1}_m)}{\partial \zeta^2}(v) < 0$ for all $v \in I_v^\circ$, as required. ■

Acknowledgments. The authors would like to thank the anonymous reviewers for their thorough and insightful reviews which have helped to improve the paper.

REFERENCES

[1] J. Acebrón, L. Bonilla, C. Pérez Vicente, F. Ritort, and R. Spigler, *The Kuramoto model: A simple paradigm for synchronization phenomena*, Rev. Modern Phys., 77 (2005), pp. 137–185.
[2] S. J. Aton and E. D. Herzog, *Come together, right now: Synchronization of rhythms in a mammalian circadian clock*, Neuron, 48 (2005), pp. 531–534.
[3] M. Bennett, M. Schatz, H. Rockwood, and K. Wiesenfield, *Huygen's clocks*, Proc. R. Soc. Lond. A, 458 (2002), pp. 563–579.
[4] A. Beuter, L. Glass, M. C. Mackey, and M. S. Titcombe, eds., *Nonlinear Dynamics in Physiology and Medicine*, Springer, New York, 2003.
[5] F. Cucker and S. Smale, *Emergent behavior in flocks*, IEEE Trans. Automat. Control, 52 (2007), pp. 852–862.
[6] F. De Smet and D. Aeyels, *Partial entrainment in the finite Kuramoto–Sakaguchi model*, Phys. D, 234 (2007), pp. 81–89.
[7] R. Diestel, *Graph Theory*, Springer-Verlag, Berlin, 2000.
[8] G. Ergun, *Human sexual contact network as a bipartite graph*, Phys. A, 308 (2002), pp. 483–488.
[9] S. Eubank, H. Guclu, V. S. Kumar, M. V. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang, *Modelling disease outbreaks in realistic urban social networks*, Nature, 429 (2004), pp. 180–184.
[10] L. Glass, *Synchronization and rhythmic processes in physiology*, Nature Rev., 410 (2001), pp. 277–284.
[11] M. Golubitsky and I. Stewart, *The Symmetry Perspective: From Equilibrium to Chaos in Phase Space and Physical Space*, Birkhäuser-Verlag, Basel, Switzerland, 2000.
[12] M. Golubitsky and I. Stewart, *Nonlinear dynamics of networks: The groupoid formalism*, Bull. Amer. Math. Soc., 43 (2006), pp. 305–364.
[13] C. Hammond, H. Bergman, and P. Brown, *Pathological synchronization in Parkinson's disease: Networks, models and treatments*, Trends in Neurosci., 30 (2007), pp. 357–364.
[14] W. Huber, V. Carey, L. Long, S. Falcon, and R. Gentleman, *Graphs in molecular biology*, BMC Bioinformatics, 8 (Suppl. 6) (2007), pp. 1–14.
[15] A. Jadbabaie, J. Lin, and A. S. Morse, *Coordination of groups of mobile autonomous agents using nearest neighbor rules*, IEEE Trans. Automat. Control, 48 (2003), pp. 988–1001.
[16] A. Jadbabaie, N. Motee, and M. Barahona, *On the stability of the Kuramoto model of coupled nonlinear oscillators*, in Proceedings of the IEEE American Control Conference, Boston, 2004, IEEE Press, Piscataway, NJ, pp. 4296–4301.
[17] Y. Kuramoto, *Self-entrainment of a population of coupled nonlinear oscillators*, in International Symposium on Mathematical Problems in Theoretical Physics, Lecture Notes in Phys. 39, H. Araki, ed., Springer, New York, 1975, pp. 420–422.
[18] Y. Kuramoto, *Chemical Oscillations, Waves and Turbulence*, Springer, New York, 1984.
[19] Y. Maistrenko, O. Popovych, and P. Tass, *Desynchronization and Chaos in the Kuramoto Model*, Lecture Notes in Phys. 671, Springer, New York, 2005.
[20] Y. Maistrenko, V. Popovych, O. Burylko, and P. Tass, *Mechanism of desynchronization in the finite-dimensional Kuramoto model*, Phys. Rev. Lett., 93 (2004), paper 084102.

[21] Y. Maistrenko, V. Popovych, and P. Tass, *Chaotic attractor in the Kuramoto model*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 15 (2005), pp. 3457–3466.
[22] L. Melloni, C. Molina, M. Pena, D. Torres, W. Singer, and E. Rodriguez, *Synchronization of neural activity across cortical areas correlates with conscious perception*, J. Neurosci., 27 (2007), pp. 2858–2865.
[23] D. C. Michaels, E. P. Matyas, and J. Jalife, *Mechanisms of sinoatrial pacemaker synchronization: A new hypothesis*, Circ. Res., 61 (1987), pp. 704–714.
[24] R. Mirollo and S. Strogatz, *The spectrum of the locked state for the Kuramoto model of coupled oscillators*, Phys. D, 205 (2005), pp. 249–266.
[25] R. Mirollo and S. Strogatz, *The spectrum of the partially locked state for the Kuramoto model*, J. Nonlinear Sci., 7 (2007), pp. 309–347.
[26] J. D. Murray, *Mathematical Biology*, 3rd ed., Springer, New York, 2002.
[27] R. Olfati-Saber and R. M. Murray, *Consensus problems in networks of agents with switching topology and time-delays*, IEEE Trans. Automat. Control, 49 (2004), pp. 1520–1533.
[28] G. Osipov, J. Kurths, and C. Zhou, *Synchronization in Oscillatory Networks*, Springer Ser. Synergetics, Springer-Verlag, New York, 2007.
[29] D. Paley, N. E. Leonard, R. Sepulchre, D. Grünbaum, and J. K. Parrish, *Oscillator models and collective motion: Spatial patterns in the dynamics of engineered and biological networks*, IEEE Control Systems Magazine, 27 (2007), pp. 89–105.
[30] A. Pikovsky, M. Rosenblum, and J. Kurths, *Synchronization: A Universal Concept in Nonlinear Sciences*, Cambridge University Press, London, 2003.
[31] V. Popovych, Y. Maistrenko, and P. Tass, *Phase chaos in coupled oscillators*, Phys. Rev. E, 71 (2005), paper 065201.
[32] J. Restrepo, E. Ott, and B. Hunt, *Onset of synchronization in large networks of coupled oscillators*, Phys. Rev. E, 71 (2005), paper 036151.
[33] J. G. Restrepo, E. Ott, and B. R. Hunt, *Emergence of coherence in complex networks of heterogeneous dynamical systems*, Phys. Rev. Lett., 96 (2006), paper 254103.
[34] J. G. Restrepo, E. Ott, and B. R. Hunt, *Synchronization in large directed networks of coupled phase oscillators*, Chaos, 16 (2006), paper 015107.
[35] J. Rogge and D. Aeyels, *Existence of partial entrainment and stability of phase locking behavior of coupled oscillators*, Progr. Theoret. Phys., 112 (2004), pp. 921–942.
[36] J. Rogge and D. Aeyels, *Stability of phase locking in a ring of unidirectionally coupled oscillators*, J. Phys. A, 37 (2004), pp. 11135–11148.
[37] L. Scardovi, A. Sarlette, and R. Sepulchre, *Synchronization and balancing on the n-torus*, Systems Control Lett., 56 (2007), pp. 335–341.
[38] R. Sepulchre, *Oscillators as systems and synchrony as a design principle*, in Current Trends in Nonlinear Systems and Control, Lecture Notes in Control and Inform. Sci., L. Menini, L. Zaccarian, and C. T. Abdallah, eds., Birkhäuser Boston, Cambridge, MA, 2006, pp. 123–142.
[39] R. Sepulchre, D. Paley, and N. E. Leonard, *Group coordination and cooperative control of steered particles in the plane*, in Group Coordination and Cooperative Control, Lecture Notes in Control and Inform. Sci., K. Y. Pettersen, J. T. Gravdahl, and H. Nijmeijer, eds., Springer-Verlag, London, 2006, pp. 217–232.
[40] R. Sepulchre, D. Paley, and N. E. Leonard, *Stabilization of planar collective motion with all-to-all communication*, IEEE Trans. Automat. Control, 52 (2007), pp. 811–824.
[41] R. Sepulchre, D. Paley, and N. E. Leonard, *Stabilization of planar collective motion with limited communication*, IEEE Trans. Automat. Control, 53 (2008), pp. 706–719.
[42] J. Shen, *Cucker–Smale flocking under hierarchical leadership*, SIAM J. Appl. Math., 68 (2007), pp. 694–719.
[43] S. Strogatz, *From Kuramoto to Crawford: Exploring the onset of synchronization in populations of coupled oscillators*, Phys. D, 143 (2000), pp. 1–20.
[44] S. Strogatz and R. Mirollo, *Stability of incoherence in a population of coupled oscillators*, J. Statist. Phys., 63 (1991), pp. 613–635.
[45] P. A. Tass, *Desynchronization of brain rhythms with soft phase-resetting techniques*, Biol. Cybernet., 87 (2002), pp. 102–115.

[46] P. A. Tass, *A model of desynchronizing deep brain stimulation with a demand-controlled coordinated reset of neural subpopulations*, Biol. Cybernet., 89 (2003), pp. 81–88.
[47] T. To, M. Henson, E. Herzog, and F. Doyle, *A molecular model for intercellular synchronization in the mammalian circadian clock*, Biophys. J., 92 (2007), pp. 3792–3803.
[48] F. Varela, J.-P. Lachaux, E. Rodriguez, and J. Martinerie, *The brainweb: Phase synchronization and large-scale integration*, Nature Rev. Neurosci., 2 (2001), pp. 229–239.
[49] M. Verwoerd and O. Mason, *Global phase-locking in finite populations of phase-coupled oscillators*, SIAM J. Appl. Dyn. Syst., 7 (2008), pp. 134–160.
[50] K. Wiesenfeld, P. Colet, and S. H. Strogatz, *Synchronization transitions in a disordered Josephson series array*, Phys. Rev. Lett., 76 (1996), pp. 404–407.
[51] K. Wiesenfeld, P. Colet, and S. H. Strogatz, *Frequency locking in Josephson arrays: Connection with the Kuramoto model*, Phys. Rev. E, 57 (1998), pp. 1563–1569.
[52] S. Yamaguchi, H. Isejima, T. Matsuo, R. Okura, K. Yagita, M. Kobayashi, and H. Okamura, *Synchronization of cellular clocks in the suprachiasmatic nucleus*, Science, 302 (2003), pp. 1408–1412.

SIAM J. Applied Dynamical Systems
Vol. 8, No. 1, pp. 454–479
© 2009 Society for Industrial and Applied Mathematics

On a Hamiltonian Version of the Rikitake System*

Răzvan M. Tudoran†, Anania Aron‡, and Ştefan Nicoară†

Abstract. In this paper we will describe some relevant dynamical and geometrical properties of the Rikitake dynamo system from the Poisson geometry point of view.

Key words. Hamiltonian dynamics, Rikitake two-dynamo system, stability analysis, periodic orbits, homoclinic and heteroclinic orbits

AMS subject classifications. 70H05, 70H14, 70K44

DOI. 10.1137/080728822

1. Introduction.

The Rikitake two-disk dynamo system (see, e.g., [20], [6]) has been widely investigated in the last fifty years. This system is a model which attempts to explain the irregular polarity switching of Earth's magnetic field. Among the topics related to the Rikitake system, we recall a few of them together with a partial list of references: chaotic behavior [22], [18], [4], [13], integrals and invariant manifolds [3], [8], [14], [23], structural stability [12], symmetries and periodic motions [21], [9], time delay [5], KCC theory [24], and many others.

In this paper we consider a particular case of the Rikitake two-dynamo system, realizing this system as a Hamiltonian system, and then study the system from some standard and nonstandard Poisson geometry points of view. By standard Poisson geometry point of view we mean the classical study of the Lyapunov stability of equilibria by using energy-Casimir-type stability tests [2], [11]; the study of the existence of periodic solutions by using the Lyapunov center theorem on the symplectic leaves of the Poisson configuration manifold, or the Weinstein–Moser theorem [25]; bifurcations phenomena; integrability, etc. By nonstandard Poisson geometry point of view we understand the connection between the standard dynamical elements enumerated above and some pure geometrical objects that we have identified, for example, in the case of the Rikitake system, as elements related to the geometry of Whitney semialgebraic stratifications of the image of the energy-Casimir mapping associated with the Rikitake system. This nonstandard Poisson geometry approach is trying to answer, in the particular case of the Rikitake two-dynamo system, the following open questions. *Is there any connection between the dynamical properties of a given dynamical system (the Rikitake system, in our case) and the geometry of the image of a vector valued constant of motion (the*

*Received by the editors June 30, 2008; accepted for publication (in revised form) by T. Kaper November 10, 2008; published electronically March 6, 2009.
http://www.siam.org/journals/siads/8-1/72882.html
†Faculty of Mathematics and Computer Science, Department of Mathematics, The West University of Timişoara, B-dul. Vasile Pârvan, No. 4, 300223 - Timişoara, Romania (tudoran@math.uvt.ro, nicoara_matcad@yahoo.com). The first author was partially supported by grants CNCSIS AT60/2007 and PN II 1081/2008.
‡Department of Mathematics, "Politehnica" University of Timişoara, Piaţa Victoriei nr. 2, 300006 - Timişoara, România (anania.girban@gmail.com).

energy-Casimir mapping, in our case), and if yes, how can one detect as many as possible dynamical elements (e.g., equilibria, periodic orbits, homoclinic and heteroclinic connections) and dynamical behavior (e.g., stability, bifurcation phenomena for equilibria, periodic orbits, homoclinic and heteroclinic connections) by just looking at the image of this mapping?

An attempt to answer these questions in the case of the Rikitake system makes the object of the main part of this work. More exactly, in the case of the Rikitake system, we get that dynamical objects, such as equilibria, periodic orbits, and homoclinic and heteroclinic connections, are associated with the fibers (both regular and singular) of the energy-Casimir mapping (note that the celebrated integrability theorem of Liouville [1] considers only the case of regular fibers in the symplectic compact and connected case); dynamical behavior such as detection of Lyapunov stable equilibria appears like extremal points of the image of the energy-Casimir mapping viewed as a convex set; bifurcations of homoclinic connections into heteroclinic connections are associated with the decay of the semialgebraic dimension of the principal stratum of a Whitney stratification naturally associated with a piece of the energy-Casimir mapping image, etc. Note that the study of the common level sets of the Hamiltonian and the Casimir is also considered in the case of the rigid body dynamics, e.g., [10].

The structure of this paper is as follows. In the second section of this work, we prepare the framework of our study by writing the Rikitake system as a Lie–Poisson system on the dual of the Lie algebra $\mathfrak{o}(Q)$ of the real orthogonal group generated by a special nondegenerate real diagonal matrix Q, and we also find an $SL(2,\mathbb{R})$ parameterized family of Hamilton–Poisson realizations. In the third section of the paper we characterize the equilibria of the Rikitake system as the critical points of the so-called energy-Casimir mapping and then perform their complete Lyapunov stability analysis. The fourth section is dedicated to the study of the existence of periodic solutions of the system, the main result summarizing that each regular coadjoint orbit of $\mathfrak{o}(Q)^*$ is locally foliated by periodic orbits, around some special Lyapunov stable equilibria. In the fifth section of the article we present some convexity properties of the image of the energy-Casimir mapping and define some naturally associated semialgebraic splittings of the image. More precisely, we discuss the relation between the image through the energy-Casimir mapping of the stable and respectively unstable families of equilibria of the Rikitake system and the canonical Whitney stratifications of the semialgebraic splittings of the image of the energy-Casimir mapping. In the sixth part of the paper we give a topological classification of the fibers of the energy-Casimir mapping, classification that follows naturally from the stratifications introduced in the previous section. Note that in our approach we are considering fibers over the regular and also over the singular values of the energy-Casimir mapping. In the seventh part of the article we give a method for explicitly obtaining homoclinic and respectively heteroclinic orbits of the Rikitake system. The described method is closely related to the Whitney stratifications introduced in the fifth section. In the eighth part of the paper we give an explicit integration of the Rikitake system via elliptic functions. In the ninth section of the paper we give two Lax formulations of the system, and finally the last part of this article discusses some numerics associated with the Poisson geometrical structure of the system. For details on Poisson geometry and Hamiltonian dynamics, see, e.g., [15], [16], [7], [19].

Note to readers: Figures and tables are not always placed within the text at the point of their discussion, but may "float" to subsequent pages in the paper.

2. Hamilton–Poisson realizations of the Rikitake system. The particular case of the Rikitake system we consider for our study is governed by the equations

$$\begin{cases} \dot{x} = yz + \beta y, \\ \dot{y} = xz - \beta x, \\ \dot{z} = -xy, \end{cases} \tag{2.1}$$

where $\beta \in \mathbb{R}$ is a parameter. Note that the above system is the Rikitake system studied in [14] in the particular case $\alpha = \mu = 0$.

As the purpose of this paper is to study the above system from the Poisson geometry point of view, the first step in this approach is to give a Hamilton–Poisson realization of the system. Because the main Hamilton–Poisson realization we are going to give is a Lie–Poisson realization, let us introduce first the geometric framework of the problem.

Let $Q := \mathrm{diag}(-1,-1,-2) \in GL(3,\mathbb{R})$ be a real diagonal matrix and $\mathcal{O}(Q) = \{A \in GL(3,\mathbb{R}) : A^T Q A = Q\}$ be the real orthogonal group generated by Q. Recall that $\mathcal{O}(Q)$ is a Lie group whose Lie algebra is $\mathfrak{o}(Q) = \{A \in \mathfrak{gl}(3,\mathbb{R}) : A^T Q + QA = O_3\}$. Note that, as a real vector space, $\mathfrak{o}(Q)$ is generated by

$$\mathfrak{o}(Q) = \mathrm{span}_{\mathbb{R}} \left\{ \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & -2 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 2 \\ 0 & -1 & 0 \end{bmatrix} \right\}.$$

The following proposition describes a Lie algebra isomorphism between $\mathfrak{o}(Q)$ and $\mathbb{R}^3$.

Proposition 2.1. *The Lie algebras* $(\mathfrak{o}(Q), +, \cdot, [\cdot,\cdot])$ *and* $(\mathbb{R}^3, +, \cdot, \times_Q)$ *are isomorphic, where the product* $\times_Q : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}^3$ *is defined by* $u \times_Q v := Q \cdot (u \times v)$ *for any* $u, v \in \mathbb{R}^3$.

Proof. One can easily verify that the map

$$\Phi : (m_1, m_2, m_3) \in \mathbb{R}^3 \longmapsto \begin{bmatrix} 0 & m_3 & -2m_2 \\ -m_3 & 0 & 2m_1 \\ m_2 & -m_1 & 0 \end{bmatrix} \in \mathfrak{o}(Q)$$

is a Lie algebra isomorphism. ∎

An immediate consequence of the above proposition is that $\mathbb{R}^3 \cong (\mathfrak{o}(Q))^*$ can be regarded as the dual of the Lie algebra $\mathfrak{o}(Q)$, and hence it has a natural Poisson structure, namely the so-called minus Lie–Poisson structure, a structure that is given in coordinates using matrix notation by

$$\Pi^-_{LP}(m_1, m_2, m_3) = \begin{bmatrix} 0 & 2m_3 & -m_2 \\ -2m_3 & 0 & m_1 \\ m_2 & -m_1 & 0 \end{bmatrix}.$$

The following theorem gives a realization of the Rikitake system as a Lie–Poisson mechanical system on the dual of the Lie algebra $\mathfrak{o}(Q)$.

Theorem 2.2. *The dynamics* (2.1) *has the following Hamilton–Poisson realization:*

$$(\mathbb{R}^3, \Pi^-_{LP}, H_\beta), \tag{2.2}$$

where

$$\Pi_{LP}^{-}(x,y,z)=\begin{bmatrix} 0 & 2z & -y \\ -2z & 0 & x \\ y & -x & 0 \end{bmatrix}$$

is the minus Lie–Poisson structure on $(\mathfrak{o}(Q))^* \cong \mathbb{R}^3$, *and the Hamiltonian* $H_\beta \in C^\infty(\mathbb{R}^3,\mathbb{R})$ *is given by* $H_\beta(x,y,z)=\frac{1}{4}(-x^2+y^2)-\beta z$.

Proof. Indeed, we have successively

$$\Pi_{LP}^{-}(x,y,z)\cdot\nabla H_\beta(x,y,z)=\begin{bmatrix} 0 & 2z & -y \\ -2z & 0 & x \\ y & -x & 0 \end{bmatrix}\cdot\begin{bmatrix} -\frac{x}{2} \\ \frac{y}{2} \\ -\beta \end{bmatrix}=\begin{bmatrix} yz+\beta y \\ xz-\beta x \\ -xy \end{bmatrix}=\begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \end{bmatrix},$$

as required. ∎

Remark 2.3. *It is not hard to see that the center of the Poisson algebra* $C^\infty(\mathbb{R}^3,\mathbb{R})$ *is generated by the Casimir invariant* $C\in C^\infty(\mathbb{R}^3,\mathbb{R})$, $C(x,y,z)=\frac{1}{2}(x^2+y^2)+z^2$.

Remark 2.4. *The Rikitake system can also be considered as a dynamical system resulting from the reduction by the symmetry group* $\mathcal{O}(Q)$ *of some symplectic dynamics on the cotangent bundle* $T^*\mathcal{O}(Q)$.

The next proposition gives other Hamilton–Poisson realizations of the Rikitake system.

Proposition 2.5. *The dynamics* (2.1) *admits a family of Hamilton–Poisson realizations parameterized by the group* $SL(2,\mathbb{R})$. *More exactly,* $(\mathbb{R}^3,\{\cdot,\cdot\}_{a,b}^{\beta},H_{c,d}^{\beta})$ *is a Hamilton–Poisson realization of the dynamics* (2.1), *where the bracket* $\{\cdot,\cdot\}_{a,b}^{\beta}$ *is defined by*

$$\{f,g\}_{a,b}^{\beta}:=\nabla C_{a,b}^{\beta}\cdot(\nabla f\times\nabla g)$$

for any $f,g\in C^\infty(\mathbb{R}^3,\mathbb{R})$, *and the functions* $C_{a,b}^{\beta}$ *and* $H_{c,d}^{\beta}$ *are given by*

$$C_{a,b}^{\beta}(x,y,z)=\frac{1}{2}\left(a-\frac{b}{2}\right)x^2+\frac{1}{2}\left(a+\frac{b}{2}\right)y^2+az^2-b\beta z,$$

$$H_{c,d}^{\beta}(x,y,z)=\frac{1}{2}\left(c-\frac{d}{2}\right)x^2+\frac{1}{2}\left(c+\frac{d}{2}\right)y^2+cz^2-d\beta z,$$

respectively; the matrix of coefficients a,b,c,d *is* $\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]\in SL(2,\mathbb{R})$.

Proof. The conclusion follows directly, taking into account that the matrix formulation of the Poisson bracket $\{\cdot,\cdot\}_{a,b}^{\beta}$ is given in coordinates by

$$\Pi_{a,b}^{\beta}(x,y,z)=\begin{bmatrix} 0 & 2az-b\beta & -\left(a+\frac{b}{2}\right)y \\ -2az+b\beta & 0 & \left(a-\frac{b}{2}\right)x \\ \left(a+\frac{b}{2}\right)y & -\left(a-\frac{b}{2}\right)x & 0 \end{bmatrix}.$$ ∎

3. Stability of equilibria.

In this section we analyze the stability properties of the equilibrium states of the Rikitake system (2.1), making use of the geometry of the problem.

Remark 3.1. *The system* (2.1) *can be written in the following equivalent form:*

$$\dot{u}=\nabla H_\beta(u)\times\nabla C(u),\qquad u=(x,y,z)\in\mathbb{R}^3,$$

where $H_\beta, C \in C^\infty(\mathbb{R}^3, \mathbb{R})$ are respectively the Hamiltonian of the system and the Casimir of the Poisson manifold $(\mathfrak{o}(Q))^*$, introduced in Theorem 2.2 and Remark 2.3.

Using this remark, one gets the following characterization for the equilibrium states of the Rikitake system.

Proposition 3.2. *A point $u \in \mathbb{R}^3$ is an equilibrium state of the Rikitake system if and only if $u \in Crit(\mathcal{EC}_\beta)$, the set of critical points of the energy-Casimir map $\mathcal{EC}_\beta \in C^\infty(\mathbb{R}^3, \mathbb{R}^2)$, and $\mathcal{EC}_\beta := (H_\beta, C)$.*

Proof. Let $u_0 = (x_0, y_0, z_0) \in \mathbb{R}^3$ be an arbitrary point. Then,

$$\mathbf{d}(H_\beta, C)(u_0) = \begin{bmatrix} \frac{\partial H_\beta}{\partial x}(u_0) & \frac{\partial H_\beta}{\partial y}(u_0) & \frac{\partial H_\beta}{\partial z}(u_0) \\ \frac{\partial C}{\partial x}(u_0) & \frac{\partial C}{\partial y}(u_0) & \frac{\partial C}{\partial z}(u_0) \end{bmatrix}$$

has rank strictly less than 2 if and only if $\nabla H_\beta(u_0) \times \nabla C(u_0) = 0$. ∎

Consequently, the equilibrium states of the Rikitake system are given as the union of the following three families:

$$\mathcal{E}_{1,\beta} = \{(M, 0, \beta) : M \in \mathbb{R}\},$$

$$\mathcal{E}_{2,\beta} = \{(0, M, -\beta) : M \in \mathbb{R}\},$$

$$\mathcal{E}_{3,\beta} = \{(0, 0, M) : M \in \mathbb{R}\}.$$

Let us begin the stability analysis by considering the equilibria from the first two families.

Theorem 3.3. *All the equilibrium states from the families $\mathcal{E}_{1,\beta}$ and $\mathcal{E}_{2,\beta}$ are nonlinearly stable except for the equilibria $(0, 0, \beta) \in \mathcal{E}_{1,\beta}$ and $(0, 0, -\beta) \in \mathcal{E}_{2,\beta}$, which are unstable.*

Proof. To study the nonlinear stability of the equilibria from $\mathcal{E}_{1,\beta}$ and $\mathcal{E}_{2,\beta}$ for $\beta \neq 0$ we are using the Arnold stability test (see, e.g., [2]). To do that, let $F_\lambda \in C^\infty(\mathbb{R}^3, \mathbb{R})$ be defined by $F_\lambda = H_\beta - \lambda C$, where $\lambda \in \mathbb{R}$ is a real parameter. Then, we have successively the following:

(i) $\mathbf{d}F_\lambda(0, M, -\beta) = 0$ if and only if $\lambda = \frac{1}{2}$, $\mathbf{d}F_\lambda(M, 0, \beta) = 0$ if and only if $\lambda = -\frac{1}{2}$;

(ii) $W = \ker \mathbf{d}C(0, M, -\beta) = \operatorname{span}_{\mathbb{R}}\{(1, 0, 0), (0, 1, -\frac{M}{2\beta})\}$, and $V = \ker \mathbf{d}C(M, 0, \beta) = \operatorname{span}_{\mathbb{R}}\{(0, 1, 0), (1, 0, \frac{M}{2\beta})\}$;

(iii) $\mathbf{d}^2 F_{\frac{1}{2}}(0, M, -\beta)\big|_{W \times W} = \operatorname{diag}(-1, -\frac{M^2}{4\beta^2})$ is negative definite for $M \neq 0$, and $\mathbf{d}^2 F_{-\frac{1}{2}}(M, 0, \beta)\big|_{V \times V} = \operatorname{diag}(1, \frac{M^2}{4\beta^2})$ is positive definite for $M \neq 0$.

Hence, from the Arnold stability test we conclude that all the equilibria from $\mathcal{E}_{1,\beta}$ and $\mathcal{E}_{2,\beta}$ for $\beta \neq 0$ are nonlinearly stable, except for $(0, 0, \beta) \in \mathcal{E}_{1,\beta}$ and $(0, 0, -\beta) \in \mathcal{E}_{2,\beta}$.

To study the nonlinear stability of the equilibria from the families $\mathcal{E}_{1,0}$ and $\mathcal{E}_{2,0}$ we will construct in each case a Lyapunov function. Let $(M, 0, 0) \in \mathcal{E}_{1,0}$ be an arbitrary equilibrium state. It is easy to verify that the smooth function $F_{(M,0,0)} \in C^\infty(\mathbb{R}^3, \mathbb{R})$, $F_{(M,0,0)}(x, y, z) = (-x^2 + y^2 + M^2)^2 + y^2 + z^2$ is a Lyapunov function for the equilibrium $(M, 0, 0)$.

Let $(0, M, 0) \in \mathcal{E}_{2,0}$ be an arbitrary equilibrium state. It is easy to verify that the smooth function $F_{(0,M,0)} \in C^\infty(\mathbb{R}^3, \mathbb{R})$, $F_{(0,M,0)}(x, y, z) = (x^2 - y^2 + M^2)^2 + x^2 + z^2$ is a Lyapunov function for the equilibrium $(0, M, 0)$.

Note that the Lyapunov functions defined above require no restriction on M, so both of them for $M = 0$ can be considered as Lyapunov functions for the equilibrium $(0, 0, 0)$. Note that the Casimir C can be considered also as a Lyapunov function for the origin.

The equilibria $(0, 0, \beta)$ and $(0, 0, -\beta)$ are unstable because in any arbitrary small neighborhood around them there exist equilibria from the family $\mathcal{E}_{3,\beta}$ admitting homoclinic orbits. Details about these homoclinic orbits will be given later in a special section. ■

The following theorem describes the stability of the equilibria from the family $\mathcal{E}_{3,\beta}$.

Theorem 3.4. *Let $e_M = (0, 0, M) \in \mathcal{E}_{3,\beta}$ be an arbitrary equilibrium state. The equilibrium $e_M \in \mathcal{E}_{3,\beta}$ is nonlinearly stable for $|M| < |\beta|$ and unstable for $|M| \geq |\beta|$.*

Proof. The characteristic polynomial associated with the linear part of the system at the equilibrium e_M is given by

$$p_{e_M}(\lambda) = \lambda(\lambda^2 + \beta^2 - M^2).$$

For $|M| > |\beta|$ the roots of p_{e_M} are $\lambda_1 = 0$ and $\lambda_{2,3} = \pm\sqrt{M^2 - \beta^2} \in \mathbb{R}$; hence e_M is a saddle and consequently an unstable equilibrium state. For $|M| = |\beta|$ we recover the special cases, $(0, 0, -\beta)$ and $(0, 0, \beta)$, already discussed in Theorem 3.3. To study the stability of the equilibria e_M for $|M| < |\beta|$ and $M \neq 0$, we will use the Arnold stability test. Note that the case $M = 0$ gives the equilibrium $(0, 0, 0)$ already discussed in Theorem 3.3. We are applying the Arnold test, using the same function test as in Theorem 3.3, namely, $F_\lambda \in C^\infty(\mathbb{R}^3, \mathbb{R})$, $F_\lambda = H_\beta - \lambda C$, where $\lambda \in \mathbb{R}$ is a real parameter. Then, we have successively the following:

(i) $\mathbf{d}F_\lambda(0, 0, M) = 0$ if and only if $\lambda = \frac{-\beta}{2M}$,
(ii) $W = \ker \mathbf{d}C(0, 0, M) = \operatorname{span}_{\mathbb{R}}\{(1, 0, 0), (0, 1, 0)\}$,
(iii) $\mathbf{d}^2 F_{\frac{-\beta}{2M}}(0, 0, M)\big|_{W\times W} = \operatorname{diag}\left(\frac{\beta - M}{M}, \frac{\beta + M}{M}\right)$ is definite for $|M| < |\beta|$.

Hence, from the Arnold stability test we conclude that for $|M| < |\beta|$, $M \neq 0$, the equilibrium states $(0, 0, M) \in \mathcal{E}_{3,\beta}$ are nonlinearly stable. ■

We gather all the stability results obtained in Theorems 3.2 and 3.3 in Tables 1 and 2. The splittings we are considering will be used later on, in the study of the energy-Casimir mapping. The superscripts "u" and "s" stand respectively for "unstable" and "stable." The global behavior of the equilibrium families $\mathcal{E}_{1,\beta}, \mathcal{E}_{2,\beta}, \mathcal{E}_{3,\beta}$ as the parameter β passes the value $\beta = 0$ is described by the bifurcation diagrams in Figure 1.

4. The existence of the periodic solutions. In this section we are making use of the property that the dynamics described by a Hamilton–Poisson system takes place on the symplectic leaves of the Poisson configuration manifold, to prove the existence of periodic orbits by looking for periodic orbits of the symplectic Hamiltonian completely integrable system obtained by the restriction of the Rikitake system to the regular coadjoint orbits of $(\mathfrak{o}(Q))^*$. This procedure will be implemented around nonlinearly stable equilibrium states.

The procedure is the following: we consider the system restricted to a regular coadjoint orbit of $(\mathfrak{o}(Q))^*$ that contains a nonlinearly stable equilibrium and then apply the Lyapunov center theorem in order to get the existence of periodic solutions for the restricted system. Taking into account the discussion above, these periodic solutions are periodic solutions also for the unrestricted system.

Table 1
Splittings for $\beta \neq 0$.

$A \subseteq \mathbb{R}^3$	$\mathcal{E}_{1,\beta}$ $(M,0,\beta) : M \in \mathbb{R}$			$\mathcal{E}_{2,\beta}$ $(0,M,-\beta) : M \in \mathbb{R}$			$\mathcal{E}_{3,\beta}$ $(0,0,M) : M \in \mathbb{R}$			
$A = \bigcup_k A_k$	$\mathcal{E}^s_{1,\beta}$ $(\xi,0,\beta)$ $\xi \neq 0$		$\mathcal{E}^u_{1,\beta}$ $(0,0,\beta)$	$\mathcal{E}^s_{2,\beta}$ $(0,\xi,-\beta)$ $\xi \neq 0$		$\mathcal{E}^u_{2,\beta}$ $(0,0-\beta)$	$\mathcal{E}^s_{3,\beta}$ $(0,0,\xi)$ $\xi \in (-\lvert\beta\rvert, \lvert\beta\rvert)$	$\mathcal{E}^u_{3,\beta}$ $(0,0,\xi)$ $\xi \in \mathbb{R} \setminus (-\lvert\beta\rvert, \lvert\beta\rvert)$		
$A = \bigcup_k \left(\bigcup_j A^j_k \right)$	$\mathcal{E}^{s,-}_{1,\beta}$ $(\xi,0,\beta)$ $\xi < 0$	$\mathcal{E}^{s,+}_{1,\beta}$ $(\xi,0,\beta)$ $\xi > 0$	$\mathcal{E}^u_{1,\beta}$ $(0,0,\beta)$	$\mathcal{E}^{s,-}_{2,\beta}$ $(0,\xi,-\beta)$ $\xi < 0$	$\mathcal{E}^{s,+}_{2,\beta}$ $(0,\xi,-\beta)$ $\xi > 0$	$\mathcal{E}^u_{2,\beta}$ $(0,0,-\beta)$	$\mathcal{E}^s_{3,\beta}$ $(0,0,\xi)$ $\xi \in (-\lvert\beta\rvert, \lvert\beta\rvert)$	$\mathcal{E}^{u,-}_{3,\beta}$ $(0,0,\xi)$ $\xi < -\lvert\beta\rvert$	$\mathcal{E}^{u,0}_{3,\beta}$ $(0,0,\pm\beta)$	$\mathcal{E}^{u,+}_{3,\beta}$ $(0,0,\xi)$ $\xi > \lvert\beta\rvert$

Table 2
Splittings for $\beta = 0$.

$A \subseteq \mathbb{R}^3$	$\mathcal{E}_{1,0}$ $(M,0,0) : M \in \mathbb{R}$			$\mathcal{E}_{2,0}$ $(0,M,0) : M \in \mathbb{R}$			$\mathcal{E}_{3,0}$ $(0,0,M) : M \in \mathbb{R}$		
$A = \bigcup_k A_k$	$\mathcal{E}^s_{1,0}$ $(\xi,0,0)$ $\xi \in \mathbb{R}$		$\emptyset$	$\mathcal{E}^s_{2,0}$ $(0,\xi,0)$ $\xi \in \mathbb{R}$		$\emptyset$	$\mathcal{E}^s_{3,0}$ $(0,0,0)$	$\mathcal{E}^u_{3,0}$ $(0,0,\xi)$ $\xi \neq 0$	
$A = \bigcup_k \left(\bigcup_j A^j_k \right)$	$\mathcal{E}^{s,*}_{1,0}$ $(\xi,0,0)$ $\xi \neq 0$	$\mathcal{E}^{s,0}_{1,0}$ $(0,0,0)$	$\emptyset$	$\mathcal{E}^{s,*}_{2,0}$ $(0,\xi,0)$ $\xi \neq 0$	$\mathcal{E}^{s,0}_{2,0}$ $(0,0,0)$	$\emptyset$	$\mathcal{E}^s_{3,0}$ $(0,0,0)$	$\mathcal{E}^{u,-}_{3,0}$ $(0,0,\xi)$ $\xi < 0$	$\mathcal{E}^{u,+}_{3,0}$ $(0,0,\xi)$ $\xi > 0$

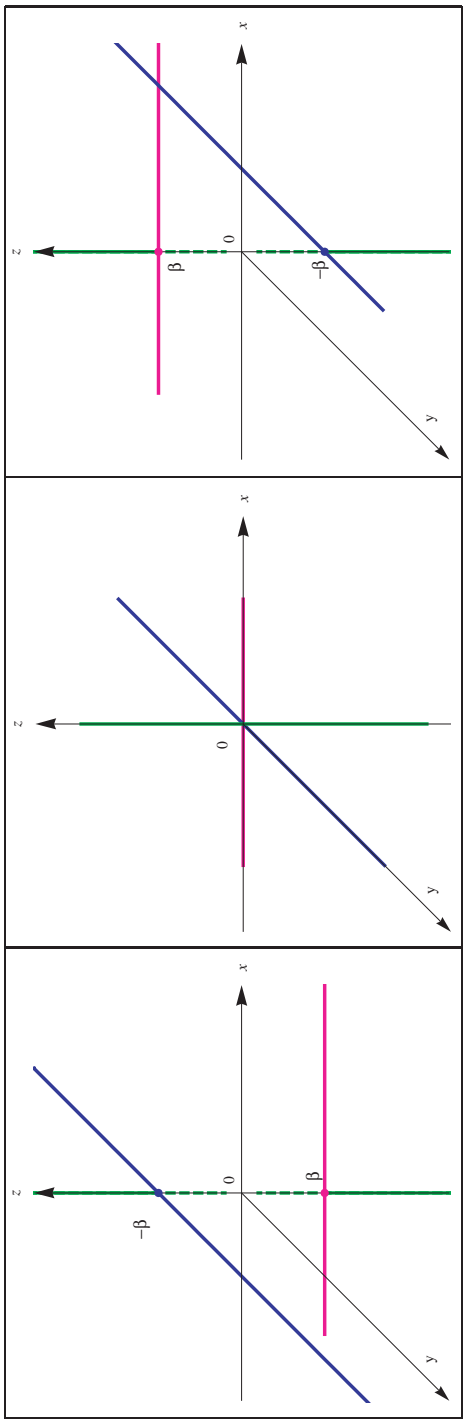


Figure 1. *Bifurcation diagram for equilibria.*

Let us begin our study by searching for periodic solutions around nonlinearly stable regular equilibria from the family $\mathcal{E}_{3,\beta}$. Recall from Theorem 3.4 that an equilibrium $(0,0,M) \in \mathcal{E}_{3,\beta}$ is nonlinearly stable if and only if $|M| < |\beta|$. Note that if $|M| < |\beta|$ and $M \neq 0$, the equilibrium state e_M is a regular one (meaning that it belongs to a regular coadjoint orbit of $(\mathfrak{o}(Q))^*$).

The Rikitake system (2.1) restricted to the regular coadjoint orbit

$$\mathcal{O}_{e_M} = \left\{ (x,y,z) \in \mathbb{R}^3 : \frac{1}{2}(x^2+y^2) + z^2 = M^2 \right\}, \quad M \neq 0,$$

has the characteristic polynomial associated with the linearization at $e_M \in \mathcal{E}_{3,\beta}$ given by

$$\tilde{p}_{e_M}(\lambda) = \lambda^2 + \beta^2 - M^2.$$

Because $e_M \in \mathcal{E}_{3,\beta}$ is a nonlinearly stable equilibrium state, we have that $|M| < |\beta|$, and hence the roots of $\tilde{p}_{e_M}$ are $\lambda_\pm = \pm i\sqrt{\beta^2 - M^2}$.

By applying the Lyapunov center theorem for the system (2.1) restricted to $\mathcal{O}_{e_M}$, we get that around e_M the coadjoint orbit $\mathcal{O}_{e_M}$ is locally foliated by periodic orbits of the restricted system, and hence as the dynamics takes place on the symplectic leaves, these periodic orbits are also periodic orbits for the system (2.1). Summarizing, we have proved the following theorem.

Theorem 4.1. *Let $e_M = (0,0,M) \in \mathcal{E}_{3,\beta}$ be such that $|M| < |\beta|$ and $M \neq 0$. Then there exist $\varepsilon_0 > 0$ and a one-parameter family $(\gamma_\varepsilon^{e_M})_{0<\varepsilon\leq\varepsilon_0}$ of periodic solutions of the Rikitake system (2.1) that approaches e_M as $\varepsilon \to 0$ with periods $T_\varepsilon^{e_M} \xrightarrow{\varepsilon\to 0} \frac{2\pi}{\sqrt{\beta^2-M^2}}$. The union*

$$\{e_M\} \cup \bigcup_{0<\varepsilon\leq\varepsilon_0} \gamma_\varepsilon^{e_M}$$

forms a smooth two dimensional manifold with boundary $\gamma_{\varepsilon_0}^{e_M}$, a manifold that is diffeomorphic to the closed disk in $\mathbb{R}^2$.

Using the same argument for the nonlinearly stable equilibria from the families $\mathcal{E}_{1,\beta}$ and $\mathcal{E}_{2,\beta}$, we get the following result.

Theorem 4.2. *Let $M \neq 0$. Then for each equilibrium state $(M,0,\beta) \in \mathcal{E}_{1,\beta}$ (respectively, $(0,M,-\beta) \in \mathcal{E}_{2,\beta}$) there exist $\varepsilon_0, \varepsilon_1 > 0$ and two one-parameter families $(\gamma_\varepsilon^{(M,0,\beta)})_{0<\varepsilon\leq\varepsilon_0}$ (respectively, $(\delta_\varepsilon^{(0,M,-\beta)})_{0<\varepsilon\leq\varepsilon_1}$) of periodic solutions of the Rikitake system (2.1) that approaches the equilibrium $(M,0,\beta)$ (respectively, $(0,M,-\beta)$) as $\varepsilon \to 0$ and have the periods $T_\varepsilon^{(M,0,\beta)} \xrightarrow{\varepsilon\to 0} \frac{2\pi}{|M|}$ (respectively, $T_\varepsilon^{(0,M,-\beta)} \xrightarrow{\varepsilon\to 0} \frac{2\pi}{|M|}$). The union*

$$\{(M,0,\beta)\} \cup \bigcup_{0<\varepsilon\leq\varepsilon_0} \gamma_\varepsilon^{(M,0,\beta)}$$

and respectively

$$\{(0,M,-\beta)\} \cup \bigcup_{0<\varepsilon\leq\varepsilon_1} \delta_\varepsilon^{(0,M,-\beta)},$$

forms each of them a two dimensional manifold with boundary $\gamma_{\varepsilon_0}^{(M,0,\beta)}$ (respectively, $\delta_{\varepsilon_1}^{(0,M,-\beta)}$), those manifolds being diffeomorphic with the closed disk in $\mathbb{R}^2$.

Remark 4.3. *Each regular coadjoint orbit of $(\mathfrak{o}(Q))^*$ containing a nonlinearly stable equilibrium state of the Rikitake system is locally foliated around the equilibrium by periodic solutions.*

5. The image of the energy-Casimir mapping. The aim of this section is to study the image of the Casimir mapping $Im(\mathcal{EC}_\beta)$ associated with the Hamilton–Poisson realization (2.2) of the Rikitake system (2.1).

We are considering convexity properties of the image of the map $\mathcal{EC}_\beta$ as well as semialgebraic partitions of it and the associated Whitney canonical stratifications. For details on Whitney stratifications of semialgebraic manifolds, see, e.g., [17].

All this will be used later on to obtain dynamical information about the system (2.1), for example, the explicit form of homoclinic and heteroclinic orbits, integrability via elliptic functions, and also a topological classification of the orbits of (2.1).

Recall first that the energy-Casimir mapping, $\mathcal{EC}_\beta \in C^\infty(\mathbb{R}^3, \mathbb{R}^2)$, is given by

$$\mathcal{EC}_\beta(x, y, z) = (H_\beta(x, y, z), C(x, y, z)), \qquad (x, y, z) \in \mathbb{R}^3,$$

where $H_\beta, C \in C^\infty(\mathbb{R}^3, \mathbb{R})$ are respectively the Hamiltonian of the system (2.1) and the Casimir of the dual of the Lie algebra $(\mathfrak{o}(Q))^*$, both of them as considered in Theorem 2.2.

Recall also that the equilibria of the system are exactly the critical points of the energy-Casimir mapping.

The next proposition gives a characterization of the image of the energy-Casimir map $\mathcal{EC}_\beta$.

Proposition 5.1. *The image of the energy-Casimir map is a convex subset of* $\mathbb{R}^2$ *given by the following:*

(i) $Im(\mathcal{EC}_\beta) = S_-^\beta \cup S_0^\beta \cup S_+^\beta$ *for* $\beta \neq 0$, *where the sets* $S_-^\beta, S_0^\beta, S_+^\beta \subset \mathbb{R}^2$ *are semialgebraic manifolds given by*

$$S_+^\beta = \left\{ (h, c) \in \mathbb{R}^2 : -\beta^2 + 2h \leq c \leq \frac{h^2}{\beta^2}; h \geq \beta^2 \right\},$$
$$S_0^\beta = \left\{ (h, c) \in \mathbb{R}^2 : c \geq \frac{h^2}{\beta^2} \right\},$$
$$S_-^\beta = \left\{ (h, c) \in \mathbb{R}^2 : -\beta^2 - 2h \leq c \leq \frac{h^2}{\beta^2}; h \leq \beta^2 \right\}.$$

(ii) $Im(\mathcal{EC}_\beta) = S_-^0 \cup S_+^0$ *for* $\beta = 0$, *where the sets* $S_-^0, S_+^0 \subset \mathbb{R}^2$ *are semialgebraic manifolds given by*

$$S_-^0 = \left\{ (h, c) \in \mathbb{R}^2 : c \geq -2h; h \leq 0 \right\},$$
$$S_+^0 = \left\{ (h, c) \in \mathbb{R}^2 : c \geq 2h; h \geq 0 \right\}.$$

In this case we also have that the whole image of the energy-Casimir map is itself a semialgebraic manifold; more exactly,

$$Im(\mathcal{EC}_0) = \left\{ (h, c) \in \mathbb{R}^2 : c \geq -2h; c \geq 2h \right\}.$$

Proof. The conclusions follow directly by simple algebraic computations using the definition of the energy-Casimir mapping. ∎

As any semialgebraic manifold has a canonical Whitney stratification, the next step is to identify this stratification and try to connect it with the image through the energy-Casimir map of subsets of the families of equilibria of the Rikitake system.

Remark 5.2. *The semialgebraic canonical Whitney stratifications of* $S^{\beta}_{-}, S^{\beta}_{0}, S^{\beta}_{+}, S^{0}_{-}, S^{0}_{+}$ *are given by the following:*

(i) $S^{\beta}_{-} = Im(\mathcal{EC}_\beta|_{\mathcal{E}^{u,-}_{3,\beta}}) \cup Im(\mathcal{EC}_\beta|_{\mathcal{E}^{u}_{1,\beta}}) \cup Im(\mathcal{EC}_\beta|_{\mathcal{E}^{s}_{1,\beta}}) \cup \Sigma^p(S^{\beta}_{-})$ *for* $\beta < 0$ *and* $S^{\beta}_{-} = Im(\mathcal{EC}_\beta|_{\mathcal{E}^{u,+}_{3,\beta}}) \cup Im(\mathcal{EC}_\beta|_{\mathcal{E}^{u}_{1,\beta}}) \cup Im(\mathcal{EC}_\beta|_{\mathcal{E}^{s}_{1,\beta}}) \cup \Sigma^p(S^{\beta}_{-})$ *for* $\beta > 0$, *where* $\Sigma^p(S^{\beta}_{-}) = \left\{(h,c) \in \mathbb{R}^2 : -\beta^2 - 2h < c < \frac{h^2}{\beta^2};\ h < -\beta^2\right\}$ *is the principal stratum.*

(ii) $S^{\beta}_{0} = Im(\mathcal{EC}_\beta|_{\mathcal{E}_{3,\beta}}) \cup \Sigma^p(S^{\beta}_{0})$, *where* $\Sigma^p(S^{\beta}_{0}) = \left\{(h,c) \in \mathbb{R}^2 : c > \frac{h^2}{\beta^2}\right\}$ *is the principal stratum.*

(iii) $S^{\beta}_{+} = Im(\mathcal{EC}_\beta|_{\mathcal{E}^{u,+}_{3,\beta}}) \cup Im(\mathcal{EC}_\beta|_{\mathcal{E}^{u}_{2,\beta}}) \cup Im(\mathcal{EC}_\beta|_{\mathcal{E}^{s}_{2,\beta}}) \cup \Sigma^p(S^{\beta}_{+})$ *for* $\beta < 0$, *and* $S^{\beta}_{+} = Im(\mathcal{EC}_\beta|_{\mathcal{E}^{u,-}_{3,\beta}}) \cup Im(\mathcal{EC}_\beta|_{\mathcal{E}^{u}_{2,\beta}}) \cup Im(\mathcal{EC}_\beta|_{\mathcal{E}^{s}_{2,\beta}}) \cup \Sigma^p(S^{\beta}_{+})$ *for* $\beta > 0$, *where* $\Sigma^p(S^{\beta}_{+}) = \left\{(h,c) \in \mathbb{R}^2 : -\beta^2 + 2h < c < \frac{h^2}{\beta^2};\ h > \beta^2\right\}$ *is the principal stratum.*

(iv) $S^{0}_{-} = Im(\mathcal{EC}_0|_{\mathcal{E}^{s,*}_{1,0}}) \cup Im(\mathcal{EC}_0|_{\mathcal{E}^{s}_{3,0}}) \cup Im(\mathcal{EC}_0|_{\mathcal{E}^{u}_{3,0}}) \cup \Sigma^p(S^{0}_{-})$, *where* $\Sigma^p(S^{0}_{-}) = \left\{(h,c) \in \mathbb{R}^2 : c > -2h;\ h < 0\right\}$ *is the principal stratum.*

(v) $S^{0}_{0} = Im(\mathcal{EC}_0|_{\mathcal{E}^{s}_{3,0}}) \cup \Sigma^p(S^{0}_{0})$, *where* $\Sigma^p(S^{0}_{0}) = Im(\mathcal{EC}_0|_{\mathcal{E}^{u}_{3,0}})$ *is the principal stratum of* S^{0}_{0}. *Note that the semialgebraic dimension of this principal stratum is* 1. *This is not a full measure stratum as in the previous cases.*

(vi) $S^{0}_{+} = Im(\mathcal{EC}_0|_{\mathcal{E}^{s,*}_{2,0}}) \cup Im(\mathcal{EC}_0|_{\mathcal{E}^{s}_{3,0}}) \cup Im(\mathcal{EC}_0|_{\mathcal{E}^{u}_{3,0}}) \cup \Sigma^p(S^{0}_{+})$, *where* $\Sigma^p(S^{0}_{+}) = \left\{(h,c) \in \mathbb{R}^2 : c > 2h;\ h > 0\right\}$ *is the principal stratum.*

All the startification results can be gathered as shown in the Figures 2, 3, 4, and 5, according to the sign of the parameter $\beta \in \mathbb{R}$. Note that for simplicity we have adopted the notation $Im(\mathcal{EC}_\beta|_{\mathcal{E}}) \overset{not}{=}: \Sigma$.

Remark 5.3. *As a convex set, the image of the energy-Casimir map is convexly generated by the images of the stable equilibria of the system, namely,*

(i) $Im(\mathcal{EC}_\beta) = \overline{\text{co}}\{Im(\mathcal{EC}_\beta|_{\mathcal{E}^{s}_{1,\beta}}), Im(\mathcal{EC}_\beta|_{\mathcal{E}^{s}_{2,\beta}}), Im(\mathcal{EC}_\beta|_{\mathcal{E}^{s}_{3,\beta}})\}$ *for* $\beta \neq 0$,
(ii) $Im(\mathcal{EC}_\beta) = \overline{\text{co}}\{Im(\mathcal{EC}_\beta|_{\mathcal{E}^{s}_{1,0}}), Im(\mathcal{EC}_\beta|_{\mathcal{E}^{s}_{2,0}})$ *for* $\beta = 0$.

6. The topology of the fibers of the energy-Casimir mapping.

In this section we describe the topology of the fibers of $\mathcal{EC}_\beta$, considering for our study fibers over regular values of $\mathcal{EC}_\beta$ as well as fibers over the singular values. It will remain an open question how these fibers fit all together in an more abstract fashion, such as bundle structures in the symplectic Arnold–Liouville integrable regular case.

Once we get the topological classification of the fibers, we will recover some of the dynamical objects already found (equilibrium points, periodic solutions) and also get some "new" dynamical objects (homoclinic orbits, heteroclinic orbits). Remember that the existence of homoclinic orbits was already used in order to obtain the instability of some equilibrium states $((0,0,-\beta),\ (0,0,\beta)$ for $\beta \neq 0)$.

Proposition 6.1. *According to the stratifications from the previous section, the topology of the fibers of* $\mathcal{EC}_\beta$ *can be described as follows:*

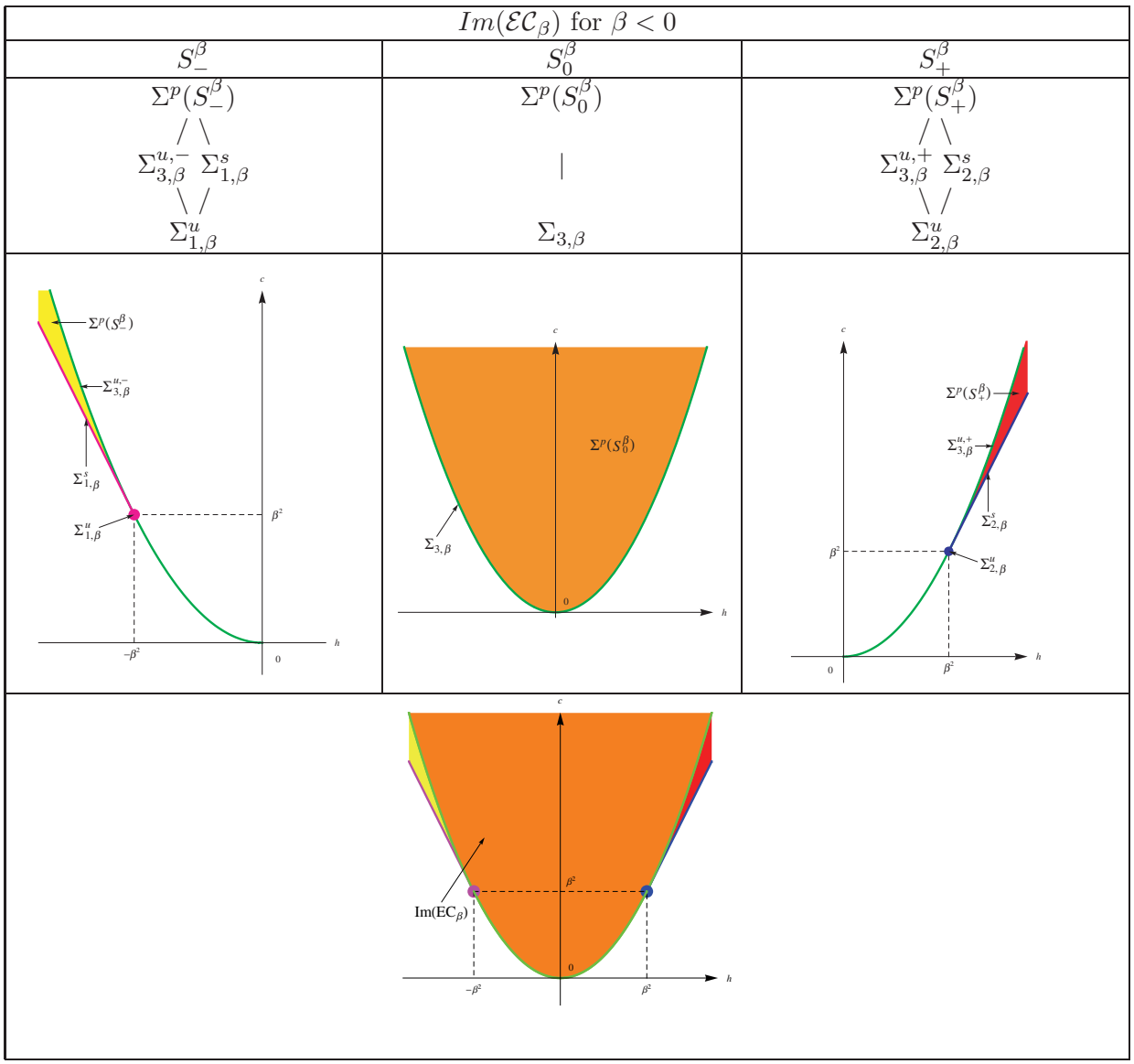


Figure 2. *Whitney canonical stratifications of the semialgebraic manifolds S_-^β, S_0^β, S_+^β for $\beta < 0$.*

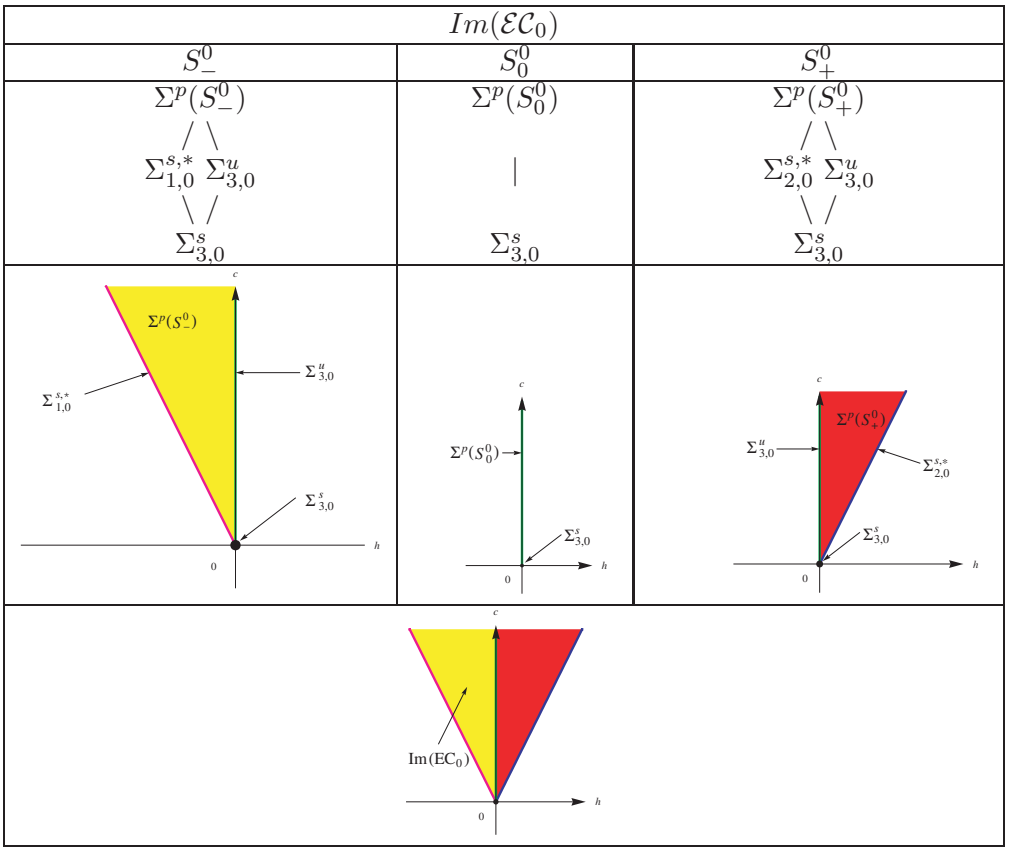


Figure 3. *Whitney canonical stratifications of the semialgebraic manifolds* S^0_-, S^0_0, S^0_+.

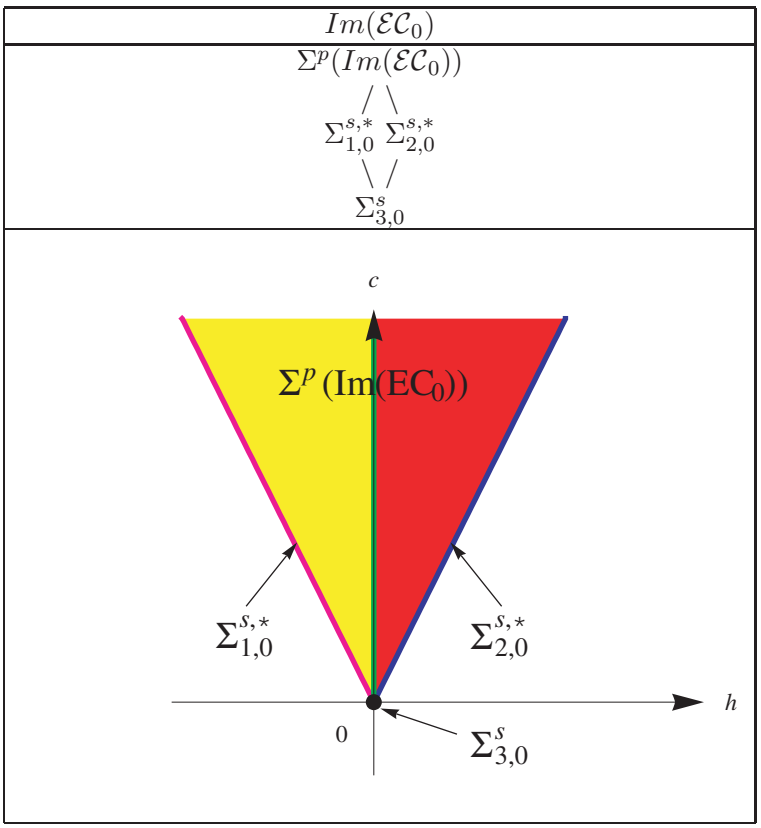


Figure 4. *Whitney canonical stratification of the semialgebraic manifold* $Im(\mathcal{EC}_0)$*, where* $\Sigma^p(Im(\mathcal{EC}_0)) = \{(h,c) \in \mathbb{R}^2 : c > -2h;\ c > 2h\}$ *is its principal stratum.*

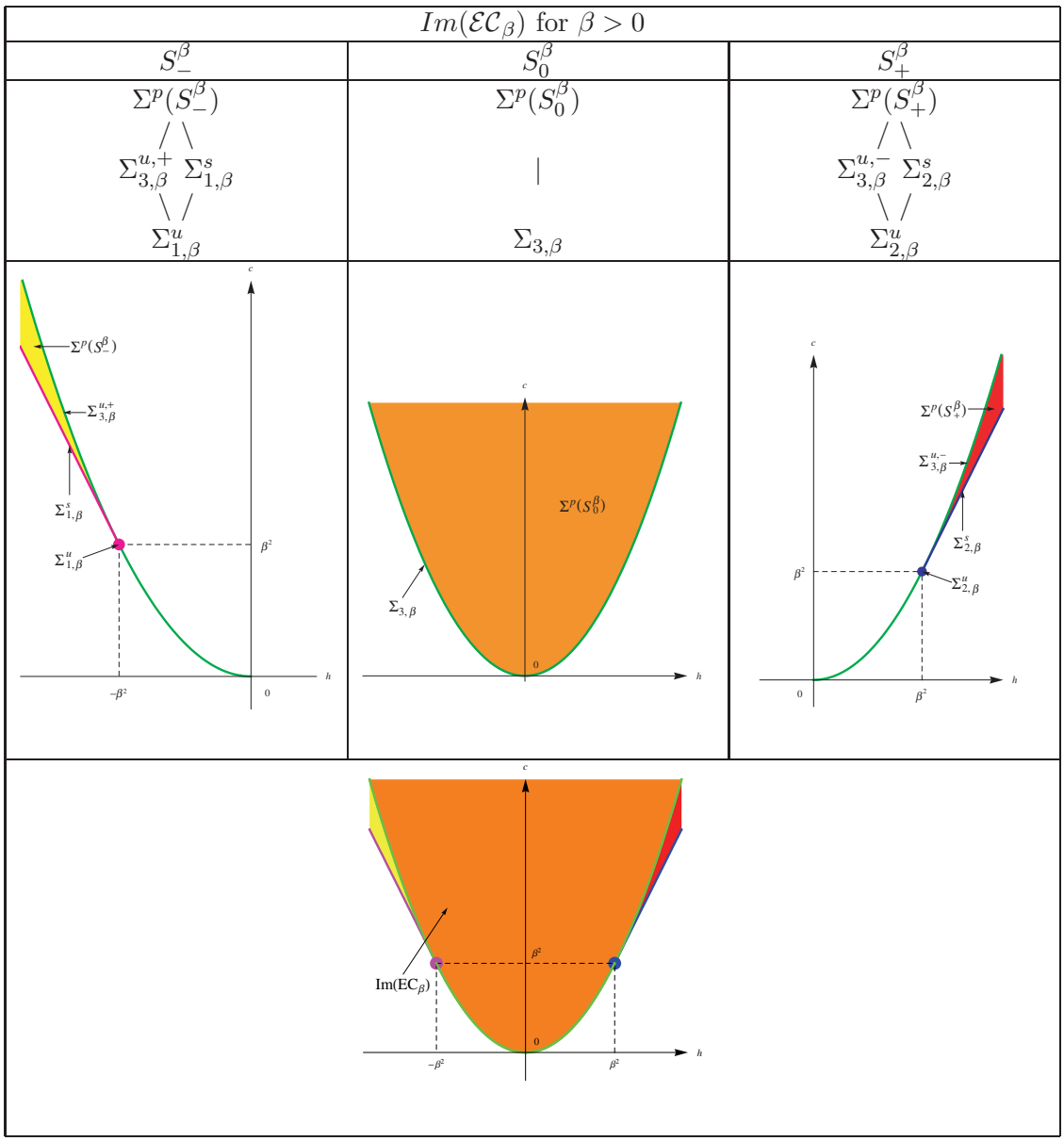
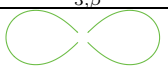


Figure 5. *Whitney canonical stratifications of the semialgebraic manifolds S_-^β, S_0^β, S_+^β for $\beta > 0$.*

(i) *The topological classification of the fibers* $\mathcal{F}_{(h,c)}$ *with* (h,c) *belonging to the strata of* $Im(\mathcal{EC}_\beta)$ *for* $\beta < 0$:

S^β_-				
$A \subseteq S^\beta_-$	$\Sigma^p(S^\beta_-)$	$\Sigma^{u,-}_{3,\beta}$	$\Sigma^s_{1,\beta}$	$\Sigma^u_{1,\beta}$
$\mathcal{F}_{(h,c)} \subseteq \mathbb{R}^3$ $(h,c) \in A$	$S^1 \amalg S^1$		$\{pt\} \times \{pt'\}$ $pt \neq pt'$	$(0,0,\beta)$
Dynamical description	*Pair of periodic orbits*	*Pair of homoclinic orbits*	*Pair of stable equilibrium points*	*Unstable equilibrium points*

S^β_0		
$A \subseteq S^\beta_0$	$\Sigma^p(S^\beta_0)$	$\Sigma_{3,\beta} \setminus (\Sigma^{u,-}_{3,\beta} \cup \Sigma^{u,+}_{3,\beta} \cup \Sigma^u_{1,\beta} \cup \Sigma^u_{2,\beta})$
$\mathcal{F}_{(h,c)} \subseteq \mathbb{R}^3, (h,c) \in A$	$S^1 \amalg S^1$	$\{pt\}$
Dynamical description	*Pair of periodic orbits*	*Stable equilibrium*

S^β_+				
$A \subseteq S^\beta_+$	$\Sigma^p(S^\beta_+)$	$\Sigma^{u,+}_{3,\beta}$	$\Sigma^s_{2,\beta}$	$\Sigma^u_{2,\beta}$
$\mathcal{F}_{(h,c)} \subseteq \mathbb{R}^3$ $(h,c) \in A$	$S^1 \amalg S^1$		$\{pt\} \times \{pt'\}$ $pt \neq pt'$	$(0,0,-\beta)$
Dynamical description	*Pair of periodic orbits*	*Pair of homoclinic orbits*	*Pair of stable equilibrium points*	*Unstable equilibrium points*

(ii) *The topological classification of the fibers* $\mathcal{F}_{(h,c)}$ *with* (h,c) *belonging to the strata of* $Im(\mathcal{EC}_0)$:

S^0_-				
$A \subseteq S^0_-$	$\Sigma^p(S^0_-)$	$\Sigma^u_{3,0}$	$\Sigma^{s,*}_{1,0}$	$\Sigma^s_{3,0}$
$\mathcal{F}_{(h,c)} \subseteq \mathbb{R}^3$ $(h,c) \in A$	$S^1 \amalg S^1$		$\{pt\} \times \{pt'\}$ $pt \neq pt'$	$(0,0,0)$
Dynamical description	*Pair of periodic orbits*	*Quadruplet of heteroclinic orbits*	*Pair of stable equilibrium points*	*Stable equilibrium point*

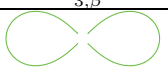
S^0_0		
$A \subseteq S^0_0$	$\Sigma^p(S^0_0)$	$\Sigma^s_{3,0}$
$\mathcal{F}_{(h,c)} \subseteq \mathbb{R}^3, (h,c) \in A$		$(0,0,0)$
Dynamical description	*Quadruplet of heteroclinic orbits*	*Stable equilibrium point*

S^0_+				
$A \subseteq S^0_+$	$\Sigma^p(S^0_+)$	$\Sigma^u_{3,0}$	$\Sigma^{s,*}_{2,0}$	$\Sigma^s_{3,0}$
$\mathcal{F}_{(h,c)} \subseteq \mathbb{R}^3$ $(h,c) \in A$	$S^1 \amalg S^1$		$\{pt\} \times \{pt'\}$ $pt \neq pt'$	$(0,0,0)$
Dynamical description	*Pair of periodic orbits*	*Quadruplet of heteroclinic orbits*	*Pair of stable equilibrium points*	*Stable equilibrium point*

(iii) *The topological classification of the fibers* $\mathcal{F}_{(h,c)}$ *with* (h,c) *belonging to the strata of* $Im(\mathcal{EC}_\beta)$ *for* $\beta > 0$:

S^β_-				
$A \subseteq S^\beta_-$	$\Sigma^p(S^\beta_-)$	$\Sigma^{u,+}_{3,\beta}$	$\Sigma^s_{1,\beta}$	$\Sigma^u_{1,\beta}$
$\mathcal{F}_{(h,c)} \subseteq \mathbb{R}^3$ $(h,c) \in A$	$S^1 \amalg S^1$		$\{pt\} \times \{pt'\}$ $pt \neq pt'$	$(0,0,\beta)$
Dynamical description	*Pair of periodic orbits*	*Pair of homoclinic orbits*	*Pair of stable equilibrium points*	*Unstable equilibrium points*

S^β_0		
$A \subseteq S^\beta_0$	$\Sigma^p(S^\beta_0)$	$\Sigma_{3,\beta} \setminus (\Sigma^{u,-}_{3,\beta} \cup \Sigma^{u,+}_{3,\beta} \cup \Sigma^u_{1,\beta} \cup \Sigma^u_{2,\beta})$
$\mathcal{F}_{(h,c)} \subseteq \mathbb{R}^3, (h,c) \in A$	$S^1 \amalg S^1$	$\{pt\}$
Dynamical description	*Pair of periodic orbits*	*Stable equilibrium*

S^β_+				
$A \subseteq S^\beta_+$	$\Sigma^p(S^\beta_+)$	$\Sigma^{u,-}_{3,\beta}$	$\Sigma^s_{2,\beta}$	$\Sigma^u_{2,\beta}$
$\mathcal{F}_{(h,c)} \subseteq \mathbb{R}^3$ $(h,c) \in A$	$S^1 \amalg S^1$		$\{pt\} \times \{pt'\}$ $pt \neq pt'$	$(0,0,-\beta)$
Dynamical description	*Pair of periodic orbits*	*Pair of homoclinic orbits*	*Pair of stable equilibrium points*	*Unstable equilibrium points*

Proof. The conclusion follows by simple computations according to the topology of the solution set of the system:

$$\begin{cases} H_\beta(x,y,z) = h, \\ C(x,y,z) = c, \end{cases}$$

where (h,c) belongs to the strata of the semialgebraic Whitney stratifications introduced in section 5. ∎

A whole picture of the above described behavior is given by Figure 6. Note that this can be considered as a visual "empirical" attempt to a singular Poisson Arnold–Liouville integrability result.

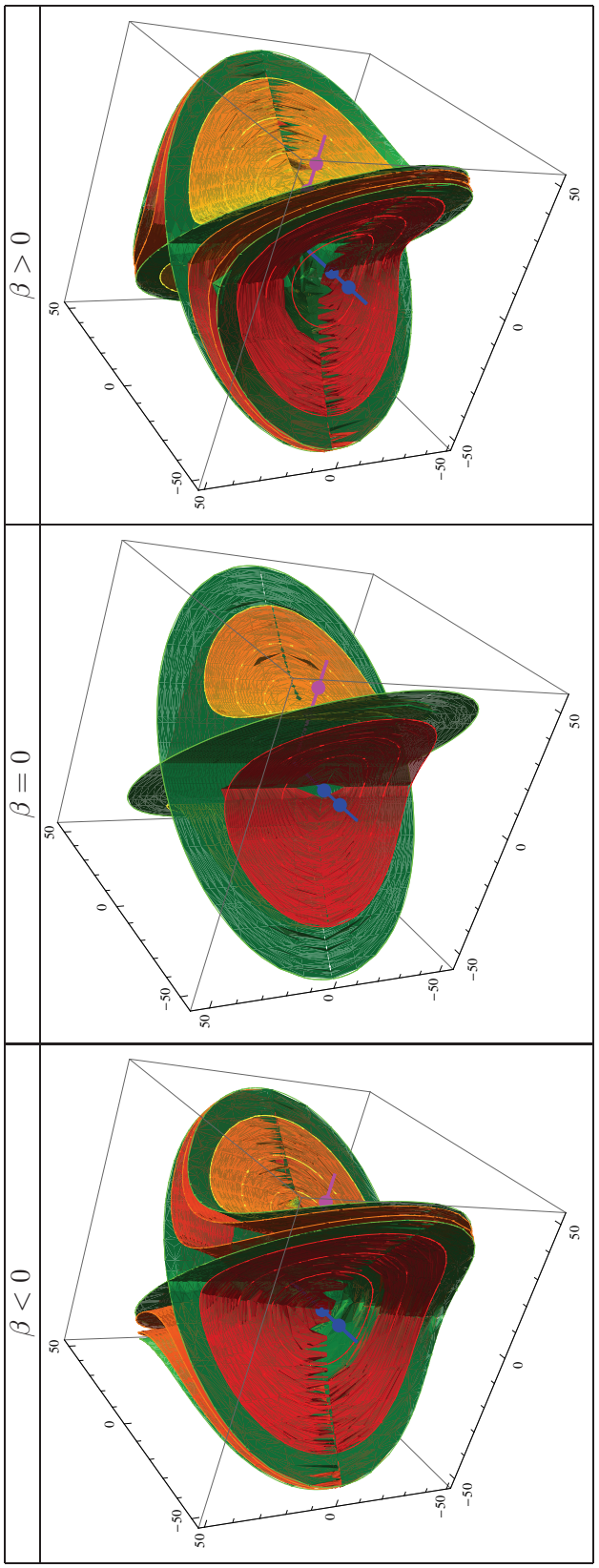


Figure 6. *Phase portraits.*

7. Homoclinic and heteroclinic orbits. In this section, using the topology of the fibers of $\mathcal{EC}_\beta$, we obtain explicit formulas for homoclinic (respectively, heteroclinic) orbits of the Rikitake system.

We also give a bifurcation diagram to show how these objects behave as the parameter $\beta \in \mathbb{R}$ passes the value $\beta = 0$.

Let us begin by discussing the existence of homoclinic orbits. These orbits are obtained in the case when $\beta \neq 0$, and they come in families of pairs connecting the unstable equilibria $(0, 0, M) \in \mathcal{E}_{3,\beta}$ for $|M| > |\beta|$.

Recall from the previous section that the existence of homoclinic orbits corresponds in the image of the energy-Casimir map to points on the strata $\Sigma^{u,-}_{3,\beta}$, or $\Sigma^{u,+}_{3,\beta}$. Using an implicit formulation, these orbits are described by the set of solutions of the following semialgebraic systems:

$$\begin{cases} \mathcal{EC}_\beta(x, y, z) = (h, c), \\ (h, c) \in \Sigma^{u,-}_{3,\beta}, \\ (h, c) = \mathcal{EC}_\beta(0, 0, M) \end{cases} \tag{7.1}$$

or

$$\begin{cases} \mathcal{EC}_\beta(x, y, z) = (h, c), \\ (h, c) \in \Sigma^{u,+}_{3,\beta}, \\ (h, c) = \mathcal{EC}_\beta(0, 0, M). \end{cases} \tag{7.2}$$

To obtain a parametric form of the homoclinic orbits, we first reduce (2.1) from three degrees of freedom to one degree of freedom by making use of one of the first integral conditions,

$$\begin{cases} H_\beta(x, y, z) = h, \\ C(x, y, z) = c, \\ (h, c) \in \Sigma^{u,-}_{3,\beta}, \\ (h, c) = \mathcal{EC}_\beta(0, 0, M) \end{cases} \tag{7.3}$$

and respectively

$$\begin{cases} H_\beta(x, y, z) = h, \\ C(x, y, z) = c, \\ (h, c) \in \Sigma^{u,+}_{3,\beta}, \\ (h, c) = \mathcal{EC}_\beta(0, 0, M), \end{cases} \tag{7.4}$$

and then integrate the resulting reduced differential equation.

Reconstructing the solution of the initial equation, one gets the following parametrization of homoclinic orbits around the unstable equilibrium state $(0, 0, M) \in \mathcal{E}_{3,\beta}$.

Let us first give the following notation:

$$\begin{aligned} x_c^{\pm}(t) = {} & \pm 2\sqrt{2}\frac{(a+\beta)(1+\beta\cosh p(t)+\beta\sinh p(t))}{1+2a\cosh p(t)+\beta^2\cosh(2p(t))+2a\sinh p(t)+\beta^2\sinh(2p(t))} \\ & \cdot\sqrt{(a-\beta)(\cosh p(t)+\sinh p(t))}, \\ y_c^{\pm}(t) = {} & \pm 2\sqrt{2}\frac{(a-\beta)(-1+\beta\cosh p(t)+\beta\sinh p(t))}{1+2a\cosh p(t)+\beta^2\cosh(2p(t))+2a\sinh p(t)+\beta^2\sinh(2p(t))} \\ & \cdot\sqrt{(a+\beta)(\cosh p(t)+\sinh p(t))}, \end{aligned}$$

$$z_c(t) = \frac{a - 2a^2\cosh p(t) + 4\beta^2\cosh p(t) + a\beta^2\cosh(2p(t))}{1 + 2a\cosh p(t) + \beta^2\cosh(2p(t)) + 2a\sinh p(t) + \beta^2\sinh(2p(t))} + \frac{-2a^2\sinh p(t) + 4\beta^2\sinh p(t) + a\beta^2\sinh(2p(t))}{1 + 2a\cosh p(t) + \beta^2\cosh(2p(t)) + 2a\sinh p(t) + \beta^2\sinh(2p(t))},$$

with $p(t) = 2\sqrt{a^2 - \beta^2}(-c + t)$, $a = M$, and $t_0 = c - \frac{\ln|\beta|}{2\sqrt{M^2-\beta^2}}$, where $c \in \mathbb{R}$ is an arbitrary real constant.

Using the notation above, the parametrization of the "figure eight-like" families of pairs of homoclinic orbits connecting the unstable equilibrium state $(0, 0, M) \in \mathcal{E}_{3,\beta}$ is given by

$$\mathcal{H}^{(0,0,M)}_{c,+} := (x^+_c, \tilde{y}_c, z_c) : \mathbb{R} \to \mathbb{R}^3 \tag{7.5}$$

and respectively

$$\mathcal{H}^{(0,0,M)}_{c,-} := (x^-_c, \tilde{y}_c, z_c) : \mathbb{R} \to \mathbb{R}^3, \tag{7.6}$$

where by $\tilde{y}_c : \mathbb{R} \to \mathbb{R}^3$ we denote the function

$$\tilde{y}_c(t) = \begin{cases} y^-_c(t), & t \le t_0, \\ y^+_c(t), & t > t_0. \end{cases}$$

Hence, around each unstable equilibrium state $(0, 0, M) \in \mathcal{E}_{3,\beta}$ we have a "figure eight-like" family of pairs of homoclinic orbits $(\mathcal{H}^{(0,0,M)}_{c,-}, \mathcal{H}^{(0,0,M)}_{c,+})_{c\in\mathbb{R}}$.

Using the same technique, in the case when $\beta = 0$, around each pair of unstable equilibrium points $(0, 0, -M)$ and $(0, 0, M)$ such that $M \neq 0$, we get four heteroclinic connections.

Recall from the above section that the existence of heteroclinic connection corresponds in the image of the energy-Casimir map to points on the stratus $\Sigma^u_{3,0}$.

Using an implicit formulation, these orbits are described by the set of solutions of the following semialgebraic system:

$$\begin{cases} \mathcal{EC}_0(x, y, z) = (h, c), \\ (h, c) \in \Sigma^u_{3,0}, \\ (h, c) = \mathcal{EC}_0(0, 0, M). \end{cases} \tag{7.7}$$

To obtain a parametric form of the heteroclinic solutions, we reduce (2.1) from three degrees of freedom to one degree of freedom by making use of the integral (7.7) condition, then integrate the resulting reduced differential equation. Reconstructing the solution of the initial equation, one gets the following parametrization of the four heteroclinic orbits connecting the unstable equilibria $(0, 0, -M)$ and $(0, 0, M)$ for $M \neq 0$.

Let us first give the following notation:

$$\begin{cases} x(t) = M\operatorname{sech}(Mt), \\ y(t) = M\operatorname{sech}(Mt), \\ z(t) = M\tanh(Mt), \end{cases} \qquad t \in \mathbb{R}.$$

Using the notation above, the four heteroclinic orbits connecting the equilibria $(0, 0, -M)$ and $(0, 0, M)$ are the following:

$$\mathcal{H}^{(0,0,\pm M)}_{(+,+,-)}(t) =: (x(t), y(t), -z(t)),$$
$$\mathcal{H}^{(0,0,\pm M)}_{(-,-,-)}(t) =: (-x(t), -y(t), -z(t)),$$
$$\mathcal{H}^{(0,0,\pm M)}_{(-,+,+)}(t) =: (-x(t), y(t), z(t)),$$

and respectively

$$\mathcal{H}^{(0,0,\pm M)}_{(+,-,+)}(t) =: (x(t), -y(t), z(t)),$$

where all of them are defined for all $t \in \mathbb{R}$.

Remark 7.1. *Note that the four heteroclinic orbits lie in the pair of orthogonal planes* $x = \pm y$.

The above results can be gathered into the bifurcation diagram shown in Figure 7, which presents the behavior of these special orbits as the real parameter $\beta \in \mathbb{R}$ passes the value $\beta = 0$.

8. Integration via Jacobi elliptic functions. In this section we present another consequence of the results from the section 6, namely, the integration via Jacobi elliptic functions of the Rikitake system (2.1) in the case $\beta = 0$.

Analogously to the method used to find homoclinic and heteroclinic orbits, we are reducing the system from three degrees of freedom to one degree of freedom, this time using the first integral condition:

$$\begin{cases} \mathcal{EC}_0(x, y, z) = (h, c), \\ (h, c) \in Im(\mathcal{EC}_0). \end{cases}$$

We then integrate the reduced differential equation via Jacobi elliptic functions and finally reconstruct the solutions of the initial equation, namely, the Rikitake system for $\beta = 0$. Using the above, we obtain that the solutions of the Rikitake system in the case when $\beta = 0$ are the following four types:

Type I:

$$\begin{cases} x(t) = \sqrt{c-2h}\ \mathrm{cn}(-\sqrt{c-2h}\ t + k), \\ y(t) = \sqrt{c+2h}\ \mathrm{dn}(-\sqrt{c-2h}\ t + k), \\ z(t) = \sqrt{c-2h}\ \mathrm{sn}(-\sqrt{c-2h}\ t + k), \end{cases} \qquad k \in \mathbb{R}.$$

Type II:

$$\begin{cases} x(t) = -\sqrt{c-2h}\ \mathrm{cn}(-\sqrt{c+2h}\ t + k), \\ y(t) = -\sqrt{c+2h}\ \mathrm{dn}(-\sqrt{c+2h}\ t + k), \\ z(t) = -\sqrt{c-2h}\ \mathrm{sn}(-\sqrt{c+2h}\ t + k), \end{cases} \qquad k \in \mathbb{R}.$$

Type III:

$$\begin{cases} x(t) = -\sqrt{c-2h}\ \mathrm{cn}(\sqrt{c+2h}\ t + k), \\ y(t) = \sqrt{c+2h}\ \mathrm{dn}(\sqrt{c+2h}\ t + k), \\ z(t) = \sqrt{c-2h}\ \mathrm{sn}(\sqrt{c+2h}\ t + k), \end{cases} \qquad k \in \mathbb{R}.$$

Type IV:

$$\begin{cases} x(t) = \sqrt{c-2h}\ \mathrm{cn}(\sqrt{c+2h}\ t + k), \\ y(t) = -\sqrt{c+2h}\ \mathrm{dn}(\sqrt{c+2h}\ t + k), \\ z(t) = \sqrt{c-2h}\ \mathrm{sn}(\sqrt{c+2h}\ t + k), \end{cases} \qquad k \in \mathbb{R}.$$

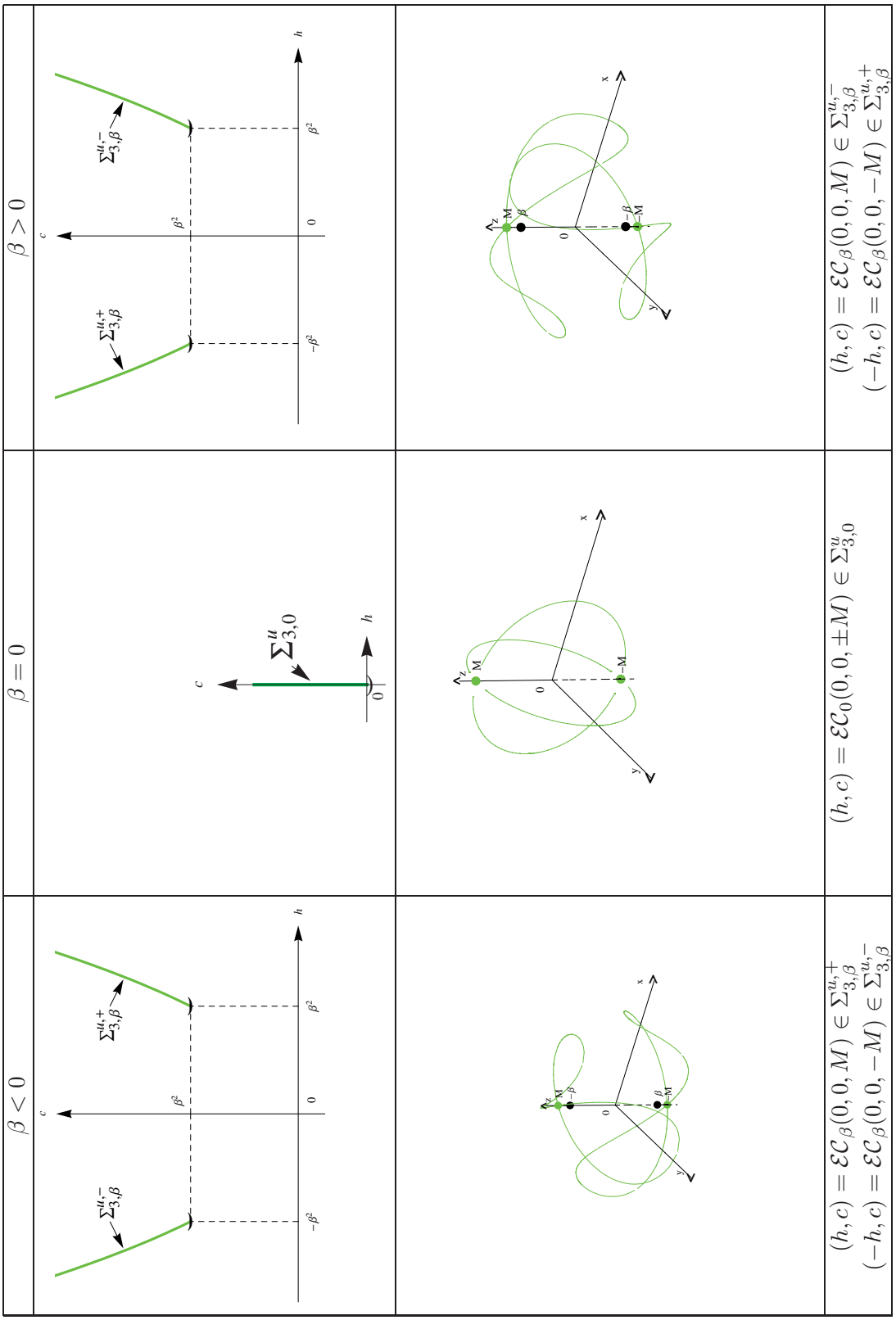


Figure 7. *Bifurcation diagram.*

Remark 8.1. *In the case when $c = \pm 2h$, we get the stable equilibrium points, points that correspond in the $Im(\mathcal{EC}_0)$ with the strata $\Sigma^s_{3,0}$, $\Sigma^{s,*}_{1,0}$ and respectively $\Sigma^{s,*}_{2,0}$. Recall that these are exactly the singular strata of the canonical Whitney stratification of $Im(\mathcal{EC}_0)$ viewed as a semialgebraic manifold.*

Remark 8.2. *For the general case when $\beta \neq 0$, using the same idea as before and the software Mathematica 6, we get the following integration result:*

$$\begin{aligned} x(t) &= \sqrt{c + 2h - 2\beta z(t) - z(t)^2}, \\ y(t) &= \sqrt{c + 2h + 2\beta z(t) - z(t)^2}, \\ z(t) &= f^{-1}(k - t), \end{aligned}$$

where

$$\begin{aligned} f(x) = \Bigg(& 2\left(2\beta + \sqrt{\beta^2 + c - 2h} + \sqrt{\beta^2 + c + 2h}\right) \\ & \text{EllipticF}\Bigg(\text{Arcsin}\Bigg(\sqrt{\frac{\left(2\beta - \sqrt{\beta^2 + c - 2h} + \sqrt{\beta^2 + c + 2h}\right)\left(\beta + \sqrt{\beta^2 + c - 2h} + x\right)}{\left(2\beta + \sqrt{\beta^2 + c - 2h} + \sqrt{\beta^2 + c + 2h}\right)\left(\beta - \sqrt{\beta^2 + c - 2h} + x\right)}}\Bigg), \\ & \frac{\beta^2 - c - \sqrt{\beta^2 + c - 2h}\sqrt{\beta^2 + c + 2h}}{\beta^2 - c + \sqrt{\beta^2 + c - 2h}\sqrt{\beta^2 + c + 2h}}\Bigg) \\ & \sqrt{\frac{\sqrt{\beta^2 + c - 2h}(\beta + \sqrt{\beta^2 + c + 2h} - x)}{\left(2\beta + \sqrt{\beta^2 + c - 2h} + \sqrt{\beta^2 + c + 2h}\right)\left(-\beta + \sqrt{\beta^2 + c - 2h} - x\right)}} \\ & \left(\beta - \sqrt{\beta^2 + c - 2h} + x\right)^2 \\ & \sqrt{\frac{\left(2\beta - \sqrt{\beta^2 + c - 2h} + \sqrt{\beta^2 + c + 2h}\right)\left(\beta + \sqrt{\beta^2 + c - 2h} + x\right)}{\left(2\beta + \sqrt{\beta^2 + c - 2h} + \sqrt{\beta^2 + 2c + 2h}\right)\left(\beta - \sqrt{\beta^2 + c - 2h} + x\right)}} \\ & \sqrt{\frac{\sqrt{\beta^2 + c - 2h}(-\beta + \sqrt{\beta^2 + c + 2h} + x)}{\left(2\beta + \sqrt{\beta^2 + c - 2h} - \sqrt{\beta^2 + c + 2h}\right)\left(\beta - \sqrt{\beta^2 + c - 2h} + x\right)}}\Bigg) \\ & \div \left(\sqrt{\beta^2 + c - 2h}\left(-2\beta + \sqrt{\beta^2 + c - 2h} - \sqrt{\beta^2 + c + 2h}\right)\right. \\ & \left.\sqrt{c + 2h + (2\beta - x)x}\sqrt{c - 2h - (2\beta + x)x}\right), \end{aligned}$$

where for $-\frac{\pi}{2} < \phi < \frac{\pi}{2}$, $\text{EllipticF}(\phi, m) = \int_0^\phi (1 - m\sin^2(\theta))^{-1/2} \mathrm{d}\theta$ and $k \in \mathbb{R}$ is an arbitrary real constant.

9. Lax formulation. In this section we present two types of Lax formulations of the Rikitake system, namely, a real three dimensional one in the case when $\beta = 0$ and a complex six dimensional one for $\beta \neq 0$.

Let us first note that as the system (2.1) is restricted to a regular coadjoint orbit, $\Sigma_c := C^{-1}(\{c\})$, $c \neq 0$, gives rise to a symplectic Hamiltonian system that is completely integrable in the sense of Liouville, and consequently it has a Lax formulation.

It is natural to ask whether the unrestricted system admits a Lax formulation. The answer is positive and is given by the following proposition.

Proposition 9.1. *The system* (2.1) *can be written in the Lax form* $\dot{L} = [L, B]$, *where the matrices* L *and* B *are given by*

$$L = \begin{bmatrix} 0 & -\beta - z & x & 0 & 0 & 0 \\ \beta + z & 0 & 0 & 0 & 0 & 0 \\ -x & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -i\beta & -i(x-y) \\ 0 & 0 & 0 & i\beta & 0 & x - y \\ 0 & 0 & 0 & i(x-y) & -x + y & 0 \end{bmatrix},$$

$$B = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -y & 0 & 0 & 0 \\ 0 & y & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -iz & -i(x+y) \\ 0 & 0 & 0 & iz & 0 & x + y \\ 0 & 0 & 0 & i(x+y) & -x - y & 0 \end{bmatrix}.$$

In the case when $\beta = 0$ we have another Lax formulation using real matrices of smaller dimension.

Proposition 9.2. *The system* (2.1) *for* $\beta = 0$ *can be written in the Lax form* $\dot{L} = [L, B]$, *where the matrices* L *and* B *are given by*

$$L = \begin{bmatrix} 0 & x - y & x + y \\ -x + y & 0 & 2z \\ -x - y & -2z & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 & -y & y \\ y & 0 & z \\ -y & -z & 0 \end{bmatrix}.$$

10. Numerical integration. We shall discuss now the numerical integration of the dynamics (2.1) via the Lie–Trotter integrator and also via the Kahan integrator.

Let us observe for the beginning that the Hamiltonian vector field X_{H_β} splits as follows:

$$X_{H_\beta} = X_{H_1} + X_{H_2} + X_{H_3^\beta},$$

where

$$H_1(x, y, z) = -\frac{x^2}{4}, \quad H_2(x, y, z) = \frac{y^2}{4}, \quad H_3^\beta(x, y, z) = -\beta z.$$

The integral curves of the vector fields X_{H_1}, X_{H_2}, and $X_{H_3^\beta}$ are given by

$$\begin{aligned}
c_1(t; x(0), y(0), z(0)) &= M_1(t) \begin{bmatrix} x(0) & y(0) & z(0) \end{bmatrix}^T, \\
c_2(t; x(0), y(0), z(0)) &= M_2(t) \begin{bmatrix} x(0) & y(0) & z(0) \end{bmatrix}^T, \\
c_3^\beta(t; x(0), y(0), z(0)) &= M_3^\beta(t) \begin{bmatrix} x(0) & y(0) & z(0) \end{bmatrix}^T,
\end{aligned}$$

where

$$M_1(t) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\frac{x(0)t}{\sqrt{2}} & \sqrt{2}\sin\frac{x(0)t}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}}\sin\frac{x(0)t}{\sqrt{2}} & \cos\frac{x(0)t}{\sqrt{2}}, \end{bmatrix},$$

$$M_2(t) = \begin{bmatrix} \cos\frac{y(0)t}{\sqrt{2}} & 0 & \sqrt{2}\sin\frac{y(0)t}{\sqrt{2}} \\ 0 & 1 & 0 \\ -\sin\frac{y(0)t}{\sqrt{2}} & 0 & \cos\frac{y(0)t}{\sqrt{2}}, \end{bmatrix}, \text{ and}$$

$$M_3^\beta(t) = \begin{bmatrix} \cos\beta t & \sin\beta t & 0 \\ -\sin\beta t & \cos\beta t & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Hence, the Lie–Trotter integrator can be written in the following form:

$$\begin{bmatrix} x^{n+1} & y^{n+1} & z^{n+1} \end{bmatrix}^T = M_1 M_2 M_3^\beta \begin{bmatrix} x^n & y^n & z^n \end{bmatrix}^T. \tag{10.1}$$

Some long but straightforward computations lead us to the following conclusion.

Proposition 10.1. *The numerical integrator* (10.1) *has the following properties:*

(i) *It is a Poisson integrator.*

(ii) *Its restriction to the regular coadjoint orbits*

$$\left\{(x, y, z) \in \mathbb{R}^3 : \frac{1}{2}(x^2 + y^2) + z^2 = \text{const.};\ \text{const.} \neq 0\right\}$$

gives rise to a symplectic integrator.

(iii) *It does not preserve the Hamiltonian H_β of the Rikitake system* (2.1).

Let us now point out some properties of the Kahan integrator related to this dynamical system. The Kahan integrator of the Rikitake system (2.1) is given by

$$\begin{cases} x^{n+1} - x^n = \dfrac{h}{2}(\beta(y^{n+1} + y^n) + y^{n+1}z^n + y^n z^{n+1}), \\ y^{n+1} - y^n = \dfrac{h}{2}(-\beta(x^{n+1} + x^n) + x^{n+1}z^n + x^n z^{n+1}), \\ z^{n+1} - z^n = \dfrac{h}{2}(-x^{n+1}y^n - x^n y^{n+1}). \end{cases}$$

After some long but straightforward computations, we get the following proposition, which shows the incompatibility of the Kahan integrator with the Poisson geometric structure of the Rikitake system.

Proposition 10.2. *The Kahan integrator does not preserve the minus Lie–Poisson structure, does not preserve the Hamiltonian H_β, and also does not preserve the Casimir C.*

Acknowledgment. The authors wish to acknowledge an anonymous referee for useful suggestions and comments.

REFERENCES

[1] V. I. Arnold, *Mathematical Methods of Classical Mechanics*, 2nd ed., Grad. Texts in Math. 60, Springer, New York, 1989.
[2] V. Arnold, *Conditions for nonlinear stability of stationary plane curvilinear flows on an ideal fluid*, Akad. Nauk. Doklady SSSR, 162 (1965), pp. 773–777.
[3] M. Barge, *Invariant manifolds and the onset of reversal in the Rikitake two-disk dynamo*, SIAM J. Math. Anal., 15 (1984), pp. 514–529.
[4] Y. X. Chang, X. J. Liu, and X. F. Li, *Chaos and chaos control of the Rikitake two-disk dynamo*, J. Liaoning Norm. Univ. Nat. Sci., 29 (2006), pp. 422–426.
[5] A. E. Cook, *Two-disc dynamo with viscous friction and time delay*, Proc. Camb. Phil. Soc., 71 (1972), pp. 135–153.
[6] A. E. Cook and P. H. Roberts, *The Rikitake two-disc dynamo system*, Proc. Camb. Phil. Soc., 68 (1970), pp. 547–569.
[7] R. H. Cushman and L. Bates, *Global Aspects of Classical Integrable Systems*, Birkhäuser, Basel, 1977.
[8] A. Figueiredo, T. M. Rocha Filho, and L. Brenig, *Algebraic structures and invariant manifolds of differential systems*, J. Math. Phys., 39 (1998), pp. 2929–2946.
[9] Y. Hardy and W.-H. Steeb, *The Rikitake two-disc dynamo system and domains with periodic orbits*, Internat. J. Theoret. Phys., 38 (1999), pp. 2413–2417.
[10] D. D. Holm and J. E. Marsden, *The rotor and the pendulum*, in Symplectic Geometry and Mathematical Physics, Prog. Math. 99, P. Donato, C. Duval, J. Elhadad, and G. M. Tuynman, eds., Birkhäuser Boston, Cambridge, MA, 1991, pp. 189–203.
[11] D. D. Holm, J. E. Marsden, T. Ratiu, and A. Weinstein, *Nonlinear stability of fluid and plasma equilibria*, Phys. Rep., 123 (1985), pp. 1–116.
[12] R. Hide, *Structural instability of the Rikitake disk dynamo*, Geophys. Res. Lett., 22 (1995), pp. 1057–1059.
[13] K. Ito, *Chaos in the Rikitake two-disc dynamo system*, Earth Planet. Sci. Lett., 51 (1980), pp. 451–456.
[14] J. Llibre and X. Zhang, *Invariant algebraic surfaces of the Rikitake system*, J. Phys. A, 33 (2000), pp. 7613–7635.
[15] J. E. Marsden, *Lectures on Mechanics*, London Math. Soc. Lecture Note Ser. 174, Cambridge University Press, Cambridge, UK, 1992.
[16] J. E. Marsden and T. S. Ratiu, *Introduction to Mechanics and Symmetry*, 2nd ed., Texts in Appl. Math. 17, Springer, Berlin, 1999.
[17] M. J. Pflaum, *Analytic and Geometric Study of Stratified Spaces*, Lecture Notes in Math. 510, Springer, Berlin, 2001.
[18] F. Plunian, Ph. Marty, and A. Alemany, *Chaotic behaviour of the Rikitake dynamo with symmetric mechanical friction and azimuthal currents*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 454 (1998), pp. 1835–1842.
[19] T. S. Ratiu, R. M. Tudoran, L. Sbano, E. Sousa Dias, and G. Terra, *A crash course in geometric mechanics*, in Geometric Mechanics and Symmetry: The Peyresq Lectures, London Math. Soc. Lecture Note Ser. 306, Cambridge University Press, Cambridge, UK, 2005, pp. 23–156.
[20] T. Rikitake, *Oscillations of a system of disk dynamos*, Proc. Cambridge Philos. Soc., 54 (1958), pp. 89–105.
[21] W.-H. Steeb, *Continuous symmetries of the Lorenz model and the Rikitake two-disc dynamo system*, J. Phys. A: Math. Gen., 15 (1982), pp. 389–390.
[22] D. L. Turcotte, *Fractals and Chaos in Geology and Geophysics*, 2nd ed., Cambridge University Press, Cambridge, UK, 1997.
[23] C. Valls, *Rikitake system: Analytic and Darbouxian integrals*, Proc. Roy. Soc. Edinburgh Sect. A, 135 (2005), pp. 1309–1326.
[24] T. Yajima and H. Nagahama, *KCC-theory and geometry of the Rikitake system*, J. Phys. A, 40 (2007), pp. 2755–2772.
[25] A. Weinstein, *Normal models for non-linear Hamiltonian systems*, Invent. Math., 20 (1973), pp. 47–57.

SIAM J. Applied Dynamical Systems
Vol. 8, No. 1, pp. 480–507
© 2009 Society for Industrial and Applied Mathematics

Computing Stability of Multidimensional Traveling Waves*

Veerle Ledoux†, Simon J. A. Malham‡, Jitse Niesen§, and Vera Thümmler¶

Abstract. We present a numerical method for computing the pure-point spectrum associated with the linear stability of multidimensional traveling fronts to parabolic nonlinear systems. Our method is based on the Evans function shooting approach. Transverse to the direction of propagation we project the spectral equations onto a finite Fourier basis. This generates a large, linear, one-dimensional system of equations for the longitudinal Fourier coefficients. We construct the stable and unstable solution subspaces associated with the longitudinal far-field zero boundary conditions, retaining only the information required for matching, by integrating the Riccati equations associated with the underlying Grassmannian manifolds. The Evans function is then the matching condition measuring the linear dependence of the stable and unstable subspaces and thus determines eigenvalues. As a model application, we study the stability of two-dimensional wrinkled front solutions to a cubic autocatalysis model system. We compare our shooting approach with the continuous orthogonalization method of Humpherys and Zumbrun. We then also compare these with standard projection methods that directly project the spectral problem onto a finite multidimensional basis satisfying the boundary conditions.

Key words. multidimensional stability, parabolic systems, Evans function

AMS subject classifications. 65L15, 65L10

DOI. 10.1137/080724009

1. Introduction.

Our goal in this paper is to present a practical extension of the Evans function shooting method to computing the stability of *multidimensional* traveling wave solutions to parabolic nonlinear systems. The problem is thus to solve the linear spectral equations for small perturbations of the traveling wave. Hence we need to determine the values of the spectral parameter for which solutions *match* longitudinal far-field zero boundary conditions and, in more than one spatial dimension, the transverse boundary conditions. These are the *eigenvalues*. There are two main approaches to solving such eigenvalue problems.

- *Projection*: Project the spectral equations onto a finite dimensional basis, which by construction satisfies the boundary conditions. Then solve the resulting matrix eigen-

*Received by the editors May 12, 2008; accepted for publication (in revised form) by C. E. Wayne December 21, 2008; published electronically March 6, 2009.
http://www.siam.org/journals/siads/8-1/72400.html
†Vakgroep Toegepaste Wiskunde en Informatica, Ghent University, Krijgslaan 281-S9, B-9000 Gent, Belgium (Veerle.Ledoux@ugent.be). This author is a postdoctoral fellow of the Fund of Scientific Research—Flanders (F.W.O.–Vlaanderen).
‡Maxwell Institute of Mathematical Sciences and School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh EH14 4AS, UK (S.J.Malham@ma.hw.ac.uk). This author was supported by EPSRC First Grant GR/S22134/01, Nuffield Foundation Newly Appointed Science Lecturers Grant SCI/180/97/129, and London Mathematical Society Scheme 2 Grant 2611.
§School of Mathematics, University of Leeds, Leeds, LS2 9JT, UK (jitse@maths.leeds.ac.uk). This author was supported by EPSRC First Grant GR/S22134/01 and Australian Research Council grant DP0559083.
¶Fakultät für Mathematik, Universität Bielefeld, 33501 Bielefeld, Germany (thuemmle@math.uni-bielefeld.de).

value problem.

- *Shooting*: Start with a specific value of the spectral parameter and the correct boundary conditions at one end. Shoot/integrate toward the far end. Examine how close this solution is to satisfying the boundary conditions at the far end. Readjust the spectral parameter accordingly. Repeat the shooting procedure until the solution *matches* the far boundary conditions.

The projection method can be applied to traveling waves of any dimension. Its main disadvantage is that to increase accuracy or include adaptation is costly and complicated as we have to reproject onto the finer or adapted basis. Further, from the set of eigenvalues produced by the resulting large algebraic eigenvalue problem, care must be taken to distinguish two subsets as follows. In the limit of the vanishing discretization scale, one subset corresponds to isolated eigenvalues of finite multiplicity, while the other subset converges to the absolute or essential spectrum, depending on the boundary conditions (see Sandstede and Scheel [27]). The latter subset can be detected by changing the truncated domain size and observing which eigenvalues drift into the complex spectral parameter plane. In contrast, for the shooting method, it is much easier to fine-tune accuracy and adaptivity, and zeros of the matching condition are isolated eigenvalues of finite multiplicity (we do not have to worry about spurious eigenvalues as described above). Its disadvantage, though, is that by its very description it is a one-dimensional method.

When implementing the shooting method, a Wronskian-like determinant measuring the discrepancy in the boundary *matching condition* in the spectral problem is typically used to locate eigenvalues in the complex spectral parameter plane. This is known as the Evans function or miss-distance function—see Alexander, Gardner, and Jones [1] or Greenberg and Marletta [14]. It is a function of the spectral parameter whose zeros correspond to eigenvalues, i.e., a *match*. In practice, we match at a point roughly centered at the front. Further, for parabolic systems it is typically analytic in the right-half complex parameter plane, and the argument principle allows us to globally determine the existence of zeros and therefore unstable eigenvalues by integrating the Evans function along suitably chosen contours.

The one-dimensional label of the Evans function has started to be overturned, and, in particular, Deng and Nii [10], Gesztesy, Latushkin, and Makarov [12], and Gesztesy, Latushkin, and Zumbrun [13] have extended the Evans function theoretically to multiple dimensions. Importantly, from a computational perspective, Humpherys and Zumbrun [20] outlined a computational method crucial to its numerical evaluation in the multidimensional scenario. Their method overcame a restriction on the order of the system that could be investigated numerically using shooting and, in particular, using exterior product spaces. It is based on continuous orthogonalization and is related to integrating along the underlying Grassmannian manifold while evolving the coordinate representation for the Grassmannian. More recently, Ledoux, Malham, and Thümmler [22] provided an alternative (though related) approach to that of Humpherys and Zumbrun. This approach is based on choosing coordinatizations of the Grassmannian manifold and integrating the resulting Riccati systems.

The main idea of this paper is to determine the stability of multidimensional traveling waves by essentially combining the two main approaches in the natural way: we project the spectral problem transversely and shoot longitudinally. The basic details are as follows.

1. Transverse to the direction of propagation of the traveling wave, we project onto a finite

Fourier basis. This generates a large, linear, one-dimensional system of equations for the longitudinal Fourier coefficients.
2. We construct the stable and unstable solution subspaces associated with the longitudinal far-field zero boundary conditions by integrating with given Grassmannian coordinatizations (chosen differently for each subspace).
3. The Evans function is the determinant of the matrix of vectors spanning both subspaces and measures their linear dependence. We evaluate it at an intermediate longitudinal point.

We compare the construction of the stable and unstable solution subspaces by integrating the spectral problem using the continuous orthogonalization method of Humpherys and Zumbrun. We also compare these with standard projection methods that project the spectral problem onto a finite multidimensional basis satisfying the boundary conditions and solve the resulting large algebraic eigenvalue problem.

The parabolic nonlinear systems that we will consider here are of the form

$$\partial_t U = B\,\Delta U + c\,\partial_x U + F(U), \tag{1.1}$$

on the cylindrical domain $\mathbb{R}\times\mathbb{T}$. Here U is a vector of component fields, and $F(U)$ represents a nonlinear coupling reaction term. The diagonal matrix B encodes the positive constant diffusion coefficients. We suppose we have Dirichlet boundary conditions in the infinite longitudinal direction $x\in\mathbb{R}$ and periodic boundary conditions in the transverse direction $y\in\mathbb{T}$. We assume also that we are in a reference frame traveling in the longitudinal direction with velocity c. Though we assume only two spatial dimensions, our subsequent analysis in the rest of this paper is straightforwardly extended to higher dimensional domains which are transversely periodic. We also suppose we have a traveling wave solution $U = U_c(x,y)$ of velocity c satisfying the boundary conditions, whose stability is the object of our scrutiny.

As a specific application we consider a cubic autocatalysis reaction-diffusion system that admits wrinkled cellular traveling fronts. They appear as solutions bifurcating from planar traveling wave solutions as the autocatalytic diffusion parameter is increased. The initiation of this first bifurcation is known as cellular instability and has been studied in some detail in the reaction kinetics and combustion literature; see, for example, Horváth et al. [19] or Terman [30]. However, our main interest here is the stability of these cellular fronts themselves. It has been shown by direct numerical simulation of the reaction-diffusion system that the cellular fronts become unstable as the autocatalyst diffusion parameter is increased further—see, for example, Horváth et al. [19] and Malevanets, Careta, and Kapral [24]. We apply our multidimensional Evans function shooting method to this problem, and we reveal the explicit form of the spectrum associated with such fronts. We compare our shooting methods with standard direct projection methods obtaining high accuracy concurrency. We confirm the literature on these instabilities and establish the shooting method as an effective and accurate alternative to standard direct projection methods.

Our paper is structured as follows. In section 2 we review Grassmannian and Stiefel manifolds. We show how to construct the stable and unstable solution subspaces of the spectral problem using a given Grassmannian coordinatization and define the Evans function using coordinatizations for each subspace. We show how to project the spectral problem onto a finite dimensional transverse basis in detail in section 3. A large portion of numerical effort

is devoted to accurately constructing the multidimensional fronts whose stability we wish to study, and in section 4 we detail the technique we use to do this. In section 5 we outline our main application—the cubic autocatalysis problem—and review the known traveling wave solutions and stability properties. We then bring to bear our full range of numerical techniques to the cubic autocatalysis problem in section 6, comparing their accuracy. In section 7 we provide some concluding remarks and outline future directions of investigation.

2. Review: Grassmann flows and spectral matching.

2.1. Grassmannian coordinate representations. An m-frame is an m-tuple of $m \leq n$ linearly independent vectors in $\mathbb{C}^n$. The *Stiefel manifold* $\mathbb{V}(n,m)$ of m-frames is the open subset of $\mathbb{C}^{n\times m}$ of all m-frames centered at the origin. Hence any element $Y \in \mathbb{V}(n,m)$ can be represented by an $n \times m$ matrix of rank m:

$$Y = \begin{pmatrix} Y_{11} & \cdots & Y_{1m} \\ \vdots & & \vdots \\ Y_{n1} & \cdots & Y_{nm} \end{pmatrix}. \tag{2.1}$$

The set of m-dimensional subspaces of $\mathbb{C}^n$ forms a complex manifold $\mathrm{Gr}(n,m)$ called the *Grassmann manifold* of m-planes in $\mathbb{C}^n$; it is compact and connected (see Steenrod [29, p. 35] or Griffiths and Harris [15, p. 193]). There is a quotient map from $\mathbb{V}(n,m)$ to $\mathrm{Gr}(n,m)$, sending each m-frame centered at the origin to the m-plane it spans; see Milnor and Stasheff [25, p. 56]. Any m-plane in $\mathbb{C}^n$ can be represented by an $n\times m$ matrix of rank m, like Y above. However, any two such matrices Y and Y' related by a rank m transformation $u \in \mathrm{GL}(k)$, so that $Y' = Yu$, will represent the same m-plane. Hence we can cover the Grassmann manifold $\mathrm{Gr}(n,m)$ by coordinate patches $\mathbb{U}_\mathfrak{i}$, labeled with multi-index $\mathfrak{i} = \{i_1,\dots,i_m\} \subset \{1,\dots,n\}$, where $\mathbb{U}_\mathfrak{i}$ is the set of m-planes $Y \in \mathrm{Gr}(n,m)$ which have a matrix representation $y_{\mathfrak{i}^\circ} \in \mathbb{C}^{n\times m}$ whose $m \times m$ submatrix, designated by the rows $\mathfrak{i}$, is the identity matrix. Each such patch is an open dense subset of $\mathrm{Gr}(n,m)$ isomorphic to $\mathbb{C}^{(n-m)m}$. For example, if $\mathfrak{i} = \{1,\dots,m\}$, then $\mathbb{U}_{\{1,\dots,m\}}$ can be uniquely represented by a matrix of the form

$$y_{\mathfrak{i}^\circ} = \begin{pmatrix} I_m \\ \hat{y} \end{pmatrix} \tag{2.2}$$

with $\mathfrak{i}^\circ = \{m+1,\dots,n\}$ and $\hat{y} \in \mathbb{C}^{(n-m)m}$. For each $\mathfrak{i}$, there is a bijective map $\varphi_\mathfrak{i}\colon \mathbb{U}_\mathfrak{i} \to \mathbb{C}^{(n-m)m}$ given by $\varphi_\mathfrak{i}\colon y_{\mathfrak{i}^\circ} \mapsto \hat{y}$, representing the local coordinate chart for the coordinate patch $\mathbb{U}_\mathfrak{i}$. For more details see Griffiths and Harris [15, pp. 193–194].

2.2. Riccati flow. Consider a nonautonomous linear vector field defined on the Stiefel manifold of the form $V(x,Y) = A(x)\,Y$, where $Y \in \mathbb{V}(n,m)$ and $A\colon \mathbb{R} \to \mathfrak{gl}(n)$, the general linear algebra of rank n matrices. Following the exposition in Ledoux, Malham, and Thümmler [22], fix a coordinate patch $\mathbb{U}_\mathfrak{i}$ for $\mathrm{Gr}(n,m)$ for some $\mathfrak{i}$, and decompose $Y \in \mathbb{V}(n,m)$ into

$$Y = y_{\mathfrak{i}^\circ} u.$$

If we substitute this form into the ordinary differential system $Y' = V(x,Y)$, we obtain

$$y'_{\mathfrak{i}^\circ} u + y_{\mathfrak{i}^\circ} u' = (A_\mathfrak{i} + A_{\mathfrak{i}^\circ} y_{\mathfrak{i}^\circ})\, u,$$

where $A_{\mathfrak{i}}$ denotes the submatrix obtained by restricting the matrix $A(x)$ to its $\mathfrak{i}$th columns. If we project this last system onto its $\mathfrak{i}^\circ$th and $\mathfrak{i}$th rows, respectively, we generate the following pair of equations for the coordinate chart variables $\hat{y} = \varphi_{\mathfrak{i}} \circ y_{\mathfrak{i}^\circ}$ and transformations u, respectively:

$$\hat{y}' = c + d\hat{y} - \hat{y}a - \hat{y}b\hat{y} \tag{2.3a}$$

and

$$u' = (a + b\hat{y})\, u. \tag{2.3b}$$

Here a, b, c, and d denote the $\mathfrak{i} \times \mathfrak{i}$, $\mathfrak{i} \times \mathfrak{i}^\circ$, $\mathfrak{i}^\circ \times \mathfrak{i}$, and $\mathfrak{i}^\circ \times \mathfrak{i}^\circ$ submatrices of A, respectively.

We observe that the flow on the Stiefel manifold $\mathbb{V}(n, m)$ generated by the linear vector field $V(x, Y)$ can be decomposed into an *independent Riccati flow* in the coordinate chart variables of a fixed coordinate patch $\mathbb{U}_{\mathfrak{i}}$ of $\mathrm{Gr}(n, m)$, and a *slaved linear flow* of transformations in $\mathrm{GL}(m)$. Hence we can reconstruct the flow on the Stiefel manifold generated by the linear vector field $V(x, Y)$, by simply solving the Riccati equation above and then, if required, subsequently integrating the linear equation for the transformations. In general, solutions to the Riccati equation (2.3a) can blow up. However, this is simply the manifestation of a poorly chosen local representative coordinate patch; see, for example, Schiff and Shnider [28]. This can be gleaned from the fact that simultaneously the determinant of u becomes zero—after all, the linear Stiefel flow, with globally smooth coefficients, does not blow up. To resolve this, we simply change the patch and keep a record of the determinant of the patch swapping transformation. Further details and strategies can be found in Ledoux, Malham, and Thümmler [22].

2.3. Drury–Oja flow. The continuous orthogonalization method of Humpherys and Zumbrun [20] can be thought of as the flow on the Grassmann manifold corresponding to the linear vector field V, with the coordinatization evolving according to a given unitary flow (see Ledoux, Malham, and Thümmler [22]). Indeed, their method is specified by a Drury–Oja flow for the $n \times m$ orthogonal matrix Q, together with the flow for the determinant of the matrix R, in the QR-decomposition of Y:

$$Q' = (I_n - QQ^\dagger)A(x)Q, \tag{2.4a}$$

$$(\det R)' = \mathrm{Tr}\big(Q^\dagger A(x) Q\big)(\det R). \tag{2.4b}$$

Here $Q^\dagger$ denotes the Hermitian transpose of Q. In practice, the determinant of R is exponentially rescaled—see Humpherys and Zumbrun [20]. We also did not find it necessary to apply any of the stabilization techniques suggested therein.

2.4. Spectral problem. Consider the linear spectral problem on $\mathbb{R}$ with spectral parameter λ in standard first order form with coefficient matrix $A(x; \lambda) \in \mathbb{C}^{n \times n}$:

$$Y' = A(x; \lambda)\, Y. \tag{2.5}$$

We assume there exists a subdomain $\Omega \subset \mathbb{C}$ containing the right-half complex plane, which does not intersect the essential spectrum. For $\lambda \in \Omega$, we know that there exist exponential

dichotomies on $\mathbb{R}^-$ and $\mathbb{R}^+$ with the same Morse index m in each case (see Sandstede [26]). Hence, for $\lambda \in \Omega$, let $Y^-(x;\lambda) \in \mathbb{V}(n,m)$ denote the matrix whose columns are solutions to (2.5) which span the unstable subspace of solutions decaying exponentially to zero as $x \to -\infty$, and let $Y^+(x;\lambda) \in \mathbb{V}(n,n-m)$ denote the matrix whose columns are the solutions which span the stable subspace of solutions decaying exponentially to zero as $x \to +\infty$. The values of spectral parameter λ for which the columns of Y^- and columns of Y^+ are linearly dependent on $\mathbb{R}$ are pure-point eigenvalues.

2.5. Evans determinant. The Evans function $D(\lambda)$ is the measure of the linear dependence between the two basis sets Y^- and Y^+:

$$D(\lambda) \equiv \mathrm{e}^{-\int_0^x \operatorname{Tr} A(\xi;\lambda)\,\mathrm{d}\xi} \det\bigl(Y^-(x;\lambda)\ Y^+(x;\lambda)\bigr) \tag{2.6}$$

(see Evans [11], Alexander, Gardner, and Jones [1], or Sandstede [26] for more details). Note that for the decompositions $Y^- = y_{\mathfrak{i}_-^\circ}\, u_-$ and $Y^+ = y_{\mathfrak{i}_+^\circ}\, u_+$, assuming $\det u_- \neq 0$ and $\det u_+ \neq 0$, we have

$$\det\bigl(Y^-\ Y^+\bigr) = \det\begin{pmatrix} y_{\mathfrak{i}_-^\circ} & y_{\mathfrak{i}_+^\circ} \end{pmatrix} \cdot \det u_- \cdot \det u_+.$$

Let $Y_0^-(\lambda)$ denote the $n \times m$ matrix whose columns are the m eigenvectors of $A(-\infty;\lambda)$ corresponding to eigenvalues with a positive real part. Analogously let $Y_0^+(\lambda)$ denote the $n \times (n-m)$ matrix whose columns are the $(n-m)$ eigenvectors of $A(+\infty;\lambda)$ corresponding to eigenvalues with a negative real part.

Suppose we fix a coordinate patch for $\mathrm{Gr}(n,m)$ labeled by $\mathfrak{i}_-$ (which has cardinality m). We integrate the corresponding Riccati differential problem (2.3a) for $\hat{y}_- = \varphi_{\mathfrak{i}_-} \circ y_{\mathfrak{i}_-^\circ}$ from $x = -\infty$ to $x = x_*$. The initial data $\hat{y}_-(-\infty;\lambda)$ is generated by postmultiplying the $\mathfrak{i}_-^\circ \times \{1,\dots,m\}$ submatrix of $Y_0^-(\lambda)$ by the inverse of the $\mathfrak{i}_- \times \{1,\dots,m\}$ submatrix of $Y_0^-(\lambda)$ (i.e., perform the simple decomposition into $y_{\mathfrak{i}_-^\circ}$ and u_- for $Y_0^-(\lambda)$, and use the $\mathfrak{i}_-^\circ \times \{1,\dots,m\}$ block of $y_{\mathfrak{i}_-^\circ}$ as the initial data for $\hat{y}_-$). Similarly, we fix a patch for $\mathrm{Gr}(n,n-m)$ labeled by $\mathfrak{i}_+$ (with cardinality $(n-m)$) and integrate the Riccati differential problem (2.3a) for $\hat{y}_+ = \varphi_{\mathfrak{i}_+} \circ y_{\mathfrak{i}_+^\circ}$ from $x = +\infty$ to $x = x_*$. We perform the analogous decomposition on $Y_0^+(\lambda)$ with index $\mathfrak{i}_+$ to generate the initial data $\hat{y}_+(+\infty;\lambda)$. If the solutions to both Riccati problems do not blow up, so that $\det u_-(x;\lambda) \neq 0$ for all $x \in (-\infty, x_*]$ and $\det u_+(x;\lambda) \neq 0$ for all $x \in [x_*, \infty)$, then we define the modified Evans function, which is analytic in λ, by

$$D(\lambda; x_*) \equiv \det\begin{pmatrix} y_{\mathfrak{i}_-^\circ}(x_*;\lambda) & y_{\mathfrak{i}_+^\circ}(x_*;\lambda) \end{pmatrix}, \tag{2.7}$$

where $x_* \in \mathbb{R}$ is the matching point. In our practical implementation in section 5, we took the patches labeled by $\mathfrak{i}_- = \{1,\dots,m\}$ and $\mathfrak{i}_+ = \{m+1,\dots,n\}$ for the left and right problems, respectively. We used the generally accepted rule of thumb for Evans function calculations: to match at a point roughly centered at the front. In our example application we took $x_* = 50$ and did *not* observe singularities in $\hat{y}^-$ or $\hat{y}^+$. However, generally, varying x_* may introduce singularities for $\hat{y}^-$ or $\hat{y}^+$ in their respective domains $(-\infty, x_*]$ and $[x_*, \infty)$. Indeed, this was observed in several examples considered in Ledoux, Malham, and Thümmler [22]. There, a robust remedy is provided which on numerical evidence allows matching anywhere in the computational domain (and which in future work we intend to apply to our context here).

If we use the Humpherys and Zumbrun continuous orthogonalization approach, then we define the Evans function to be

$$D_{\mathrm{HZ}}(\lambda;x_*) \equiv \det\bigl(Q^-(x_*;\lambda)\ Q^+(x_*;\lambda)\bigr)\cdot\det R^-(x_*;\lambda)\cdot\det R^+(x_*;\lambda). \tag{2.8}$$

This is analytic in λ when we include the $\det R^{\pm}$ terms.

2.6. Angle between subspaces. We can also measure the angle between the unstable and stable linear subspaces $\mathbb{V}(n,m)$ and $\mathbb{V}(n,n-m)$; see Björck and Golub [8]. When the columns of the matrices $Y^-(x;\lambda)$ and $Y^+(x;\lambda)$ which span $\mathbb{V}(n,m)$ and $\mathbb{V}(n,n-m)$, respectively, are nearly linearly dependent, the angle will be small. It is defined as follows: without loss of generality we assume $m \geq n/2$. Let Q^- and Q^+ be unitary bases for $\mathbb{V}(n,m)$ and $\mathbb{V}(n,n-m)$. They are specified by the QR-decompositions $Y^- = Q^-R^-$ and $Y^+ = Q^+R^+$. The cosine of the smallest angle θ between $\mathbb{V}(n,m)$ and $\mathbb{V}(n,n-m)$ is given by the largest singular value of the matrix $(Q^-)^\dagger Q^+$. However, when θ is small, it is more accurate to compute $\sin\theta$, which is given by the smallest singular value of

$$\bigl(I_n - Q^-(Q^-)^\dagger\bigr)Q^+.$$

Hence, if the singular values of this matrix are σ_i, $i=1,\ldots,n-m$, define

$$\theta(\lambda;x_*) \equiv \arcsin \min_{i=1,\ldots,n-k} \sigma_i,$$

where, again, x_* is the matching point. Eigenvalues correspond to zeros of $\theta(\lambda;x_*)$.

3. Projection onto a finite transverse Fourier basis. Consider the parabolic nonlinear system (1.1) posed on the cylindrical domain $\mathbb{R}\times\mathbb{T}$. Recall that we assume Dirichlet boundary conditions in the infinite longitudinal direction $x\in\mathbb{R}$ and periodic boundary conditions in the transverse direction $y\in\mathbb{T}$ (here $\mathbb{T}$ is the one-torus with period/circumference L). Suppose we have a traveling wave solution $U = U_c(x,y)$ of velocity c satisfying the boundary conditions. The stability of this solution is determined by the spectrum of the linear differential operator

$$\mathcal{L} = B\Delta + c\,\partial_x + \mathrm{D}F(U_c), \tag{3.1}$$

where $\mathrm{D}F$ is the Jacobian of the interaction term. Hence the eigenvalue problem is

$$B\Delta U + c\,\partial_x U + \mathrm{D}F(U_c)U = \lambda U. \tag{3.2}$$

We wish to define and compute the Evans function for this spectral problem. The idea is to perform a Fourier decomposition in the transverse direction. This transforms the problem into a one-dimensional problem, for which we can apply the Evans function shooting approach. Hence suppose that we can write

$$U(x,y) = \sum_{k=-\infty}^{\infty} \hat{U}_k(x)\,\mathrm{e}^{iky/\tilde{L}}, \tag{3.3}$$

where $\hat{U}_k \in \mathbb{C}^N$ are the Fourier coefficients (with N denoting the number of components of U) and $\tilde{L} = 2\pi L$; this implicitly restricts our attention to perturbations U with period L. Similarly, we expand $\mathrm{D}F(U_c)$ as

$$\mathrm{D}F(U_c(x,y)) = \sum_{k=-\infty}^{\infty} \hat{D}_k(x)\, \mathrm{e}^{iky/\tilde{L}}, \tag{3.4}$$

where the Fourier coefficients $\hat{D}_k \in \mathbb{C}^{N\times N}$. Substituting our Fourier expansions for U and $\mathrm{D}F(U_c)$ in (3.3) and (3.4) into the eigenvalue problem (3.2) yields

$$B\,(\partial_{xx}+\partial_{yy}) \sum_{k=-\infty}^{\infty} \hat{U}_k\, \mathrm{e}^{iky/\tilde{L}} + c\,\partial_x \sum_{k=-\infty}^{\infty} \hat{U}_k\, \mathrm{e}^{iky/\tilde{L}} + \sum_{\ell=-\infty}^{\infty} \hat{D}_\ell\, \mathrm{e}^{i\ell y/\tilde{L}} \sum_{k=-\infty}^{\infty} \hat{U}_k\, \mathrm{e}^{iky/\tilde{L}} = \lambda \sum_{k=-\infty}^{\infty} \hat{U}_k\, \mathrm{e}^{iky/\tilde{L}}.$$

Reordering the double sum

$$\sum_{\ell=-\infty}^{\infty} \hat{D}_\ell\, \mathrm{e}^{i\ell y/\tilde{L}} \sum_{k=-\infty}^{\infty} \hat{U}_k\, \mathrm{e}^{iky/\tilde{L}} = \sum_{k=-\infty}^{\infty} \sum_{\nu=-\infty}^{\infty} \hat{D}_{k-\nu}\hat{U}_\nu\, \mathrm{e}^{iky/\tilde{L}}$$

yields that for all $k \in \mathbb{Z}$ we must have

$$\partial_{xx}\hat{U}_k - (k/\tilde{L})^2\hat{U}_k + c\,B^{-1}\partial_x\hat{U}_k + \sum_{\nu=-\infty}^{\infty} B^{-1}\hat{D}_{k-\nu}\hat{U}_\nu = \lambda B^{-1}\hat{U}_k.$$

In practical implementation, we consider only the first K Fourier modes. Hence we replace the above infinite system of equations by the following finite system of equations which we now also express in first order form:

$$\partial_x\hat{U}_k = \hat{P}_k, \tag{3.5a}$$

$$\partial_x\hat{P}_k = \lambda B^{-1}\hat{U}_k + (k/\tilde{L})^2\hat{U}_k - c\,B^{-1}\hat{P}_k - \sum_{\nu=-K}^{K} B^{-1}\hat{D}_{k-\nu}\hat{U}_\nu \tag{3.5b}$$

for $k = -K, -K+1, \ldots, K$. This is a large system of ordinary differential equations involving the spectral parameter λ, so we can apply the standard Evans function approach to it. In matrix form this is

$$\partial_x \begin{pmatrix} \hat{U} \\ \hat{P} \end{pmatrix} = \begin{pmatrix} O_{N(2K+1)} & I_{N(2K+1)} \\ A_3(x;\lambda) & A_4 \end{pmatrix} \begin{pmatrix} \hat{U} \\ \hat{P} \end{pmatrix}, \tag{3.6}$$

where $\hat{U}$ is the vector of $\mathbb{C}^N$-valued components $\hat{U}_{-K}, \hat{U}_{-K+1}, \ldots, \hat{U}_K$ and $\hat{P}$ is the vector of $\mathbb{C}^N$-valued components $\hat{P}_{-K}, \hat{P}_{-K+1}, \ldots, \hat{P}_K$. The lower right block matrix is $A_4 =$

$-c\,B^{-1} \otimes I_{2K+1}$. The lower left block matrix is given by $A_3 = \tilde{A}_3(\lambda) + \hat{A}_3(x)$, where, if we set $E_k(\lambda) \equiv \lambda B^{-1} + (k/\tilde{L})^2 I_N$, then

$$\tilde{A}_3(\lambda) = \begin{pmatrix} E_{-K}(\lambda) & O & \cdots & O \\ O & E_{-K+1}(\lambda) & \cdots & O \\ \vdots & \vdots & \ddots & \vdots \\ O & \cdots & O & E_{+K}(\lambda) \end{pmatrix}$$

and

$$\hat{A}_3(x) = -B^{-1} \otimes \begin{pmatrix} \hat{D}_0 & \hat{D}_{-1} & \cdots & \hat{D}_{-2K} \\ \hat{D}_1 & \hat{D}_0 & \cdots & \hat{D}_{-2K+1} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{D}_{2K} & \hat{D}_{2K-1} & \cdots & \hat{D}_0 \end{pmatrix}.$$

A similar Galerkin approximation for nonplanar fronts can be found in Gesztesy, Latushkin, and Zumbrun [13, pp. 28, 37].

4. The freezing method. To study the long-time behavior of the wrinkled front solution, we use the method of freezing traveling waves described in Beyn and Thümmler [7] and analyzed in Thümmler [31]. This method transports the problem into a moving frame whose velocity is computed during the computation. For the traveling waves whose stability is our concern here, the freezing method has two advantages over direct simulation: First, we can do long-time computations of time dependent (for example, oscillating) solutions. This is because the method picks out the correct speed for the moving frame in which the wave is roughly stationary (in particular, it does not drift out of the chosen domain). Second, the freezing method can be used to obtain a suitably accurate guess for the waveform, which can then be used to initiate a Newton solver that computes the traveling wave as a stationary solution in the comoving frame. For completeness we will briefly review the idea of the freezing method here. Consider a parabolic nonlinear system of the form (1.1) in a stationary frame of reference:

$$\partial_t U = B\,\Delta U + F(U). \tag{4.1}$$

If we substitute the ansatz $U(x, y, t) = V(x - \gamma(t), y, t)$ into this equation, where $\gamma(t)$ is the unknown reference position for a wave traveling in the longitudinal x-direction, we generate the partial differential equation for V and γ':

$$\partial_t V = B\,\Delta V + \gamma'(t)\partial_x V + F(V). \tag{4.2a}$$

In order to compensate for the additional degree of freedom which has been introduced by the unknown position γ, we specify the following additional algebraic constraint:

$$0 = \int_{\mathbb{R}\times\mathbb{T}} \big(\partial_x \hat{V}(x, y, t)\big)^{\mathrm{T}} \big(\hat{V}(x, y, t) - V(x, y, t)\big)\,\mathrm{d}x\,\mathrm{d}y. \tag{4.2b}$$

This constraint selects a unique solution from the one-parameter group of possible solutions generated by longitudinal translation. It is a phase condition that normalizes the system with

respect to a given template function; for example, a suitable choice is to set $\hat{V}(x,y,t)$ to be the initial data $V(x,y,0)$. Hence we have generated a partial differential algebraic system (4.2) for the unknowns $V(x,y,t)$ and $\zeta(t)=\gamma'(t)$, with initial data $V(x,y,0)$ and $\zeta(0)$ within the attractive set of the stable traveling wave and its velocity.

A stationary solution $(V,\zeta)=(V_c,c)$ of the partial differential algebraic system (4.2) represents a traveling wave solution $U_c(x,y,t)=V_c(x-ct,y)$ with fixed speed c to the parabolic nonlinear problem (4.1). If the traveling wave U_c is stable, then the stationary solution (V_c,c) is also stable, and the time evolution of partial differential algebraic system (4.2) will converge to (V_c,c). Thus the numerical method consists of solving the partial differential algebraic system (4.2) until the solution (V,ζ) has reached the equilibrium (V_c,c). Alternatively one can also solve the partial differential algebraic system (4.2) until some time $t=T$ when the solution is deemed sufficiently close to equilibrium and then use a direct method (e.g., Newton's method) to solve the stationary equation

$$0 = B\,\Delta v + \theta\partial_x v + F(v), \tag{4.3a}$$

$$0 = \int_{\mathbb{R}\times\mathbb{T}} \big(\partial_x \hat{V}(x,y)\big)^{\mathrm{T}} \big(\hat{V}(x,y)-v(x,y)\big)\,\mathrm{d}x\,\mathrm{d}y \tag{4.3b}$$

for (v,θ) starting with initial values $v_0(x,y)=V(x,y,T)$ and $\theta_0=\zeta(T)$.

5. Review: Autocatalytic system. As our model application we consider the cubic autocatalytic system

$$u_t = \Delta u + c\partial_x u - uv^2, \tag{5.1a}$$

$$v_t = \delta\Delta v + c\partial_x v + uv^2. \tag{5.1b}$$

Here $u(x,y,t)$ is the concentration of the reactant and $v(x,y,t)$ is the concentration of the autocatalyst in the infinitely extended medium $(x,y)\in\mathbb{R}\times\mathbb{T}$. In the far field, we suppose that (u,v) approaches the stable homogeneous steady state $(0,1)$ as $x\to-\infty$ and the unstable homogeneous steady state $(1,0)$ as $x\to+\infty$. The parameter δ is the ratio of the diffusivity of the autocatalyst to that of the reactant. This system is globally well-posed for smooth initial data and any finite $\delta>0$.

From Billingham and Needham [4, 5] we know that, in one spatial dimension, a unique heteroclinic connection between the unstable and stable homogeneous steady states exists for wavespeeds $c\geq c_{\min}$. The unique traveling wave for $c=c_{\min}$ converges exponentially to the homogeneous steady states. They are readily constructed by shooting (see, for example, Balmforth, Craster, and Malham [3]) and represent the planar front solution referred to hereafter.

It is well known that these planar traveling wave solutions become too unstable to transverse perturbations when δ is greater than a critical ratio $\delta_{\mathrm{cr}}\approx 2.3$ (from Malevanets, Careta, and Kapral [24, p. 4733]), and our transverse domain is sufficiently large. This is known as the cellular instability—see, for example, Terman [30], Horváth et al. [19], and Malevanets, Careta, and Kapral [24]. Calculating the spectrum that reveals the mechanism of the instability is also straightforward. If we consider small perturbations about the underlying planar traveling wave solution of the form $\hat{U}(x)\mathrm{e}^{\lambda t+\mathrm{i}(k/\tilde{L})y}$ for any $k\in\mathbb{Z}$, then the associated linear operator $\mathcal{L}$ whose spectrum we wish to determine has the form

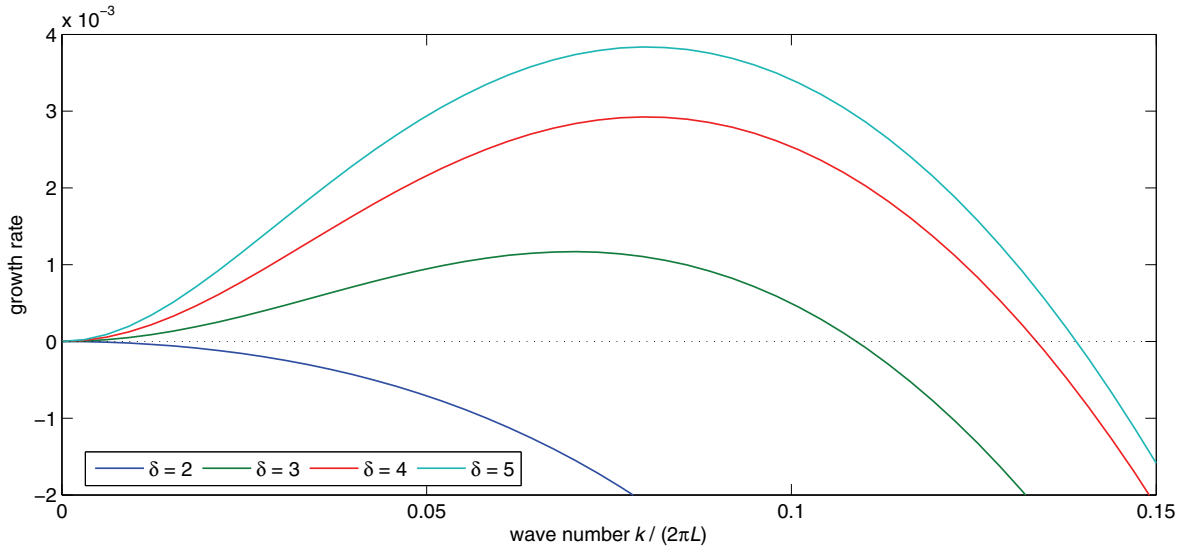


Figure 1. *Dispersion relations for the planar front for different values of δ. For $\delta = 2$ all perturbations decay and the planar front is linearly stable, but for the other values of δ there is a growing interval $(0, k_\delta)$ of unstable transverse wavenumbers.*

$$\mathcal{L} = B\big(\partial_{xx} - (k/\tilde{L})^2 I\big) + c\,\partial_x + \mathrm{D}F(U_c). \tag{5.2}$$

The spectrum of $\mathcal{L}$ consists of the pure-point spectrum, i.e., isolated eigenvalues of finite multiplicity, together with the essential spectrum. By a result of Henry [16], if we consider $\mathcal{L}$ as an operator on the space of square integrable functions, we know that the essential spectrum is contained within the parabolic curves of the continuous spectrum in the left-half complex plane. The continuous spectrum for this model problem is given for all $\mu \in \mathbb{R}$ and each fixed $k \in \mathbb{Z}$ by the locus in the complex λ-plane of the curves

$$\begin{aligned}
\lambda &= -\delta(k/\tilde{L})^2 + \mathrm{i}c\mu - \delta\mu^2,\\
\lambda &= -1 - \delta(k/\tilde{L})^2 + \mathrm{i}c\mu - \delta\mu^2,\\
\lambda &= -(k/\tilde{L})^2 + \mathrm{i}c\mu - \mu^2,
\end{aligned}$$

where the last curve appears with multiplicity two. Hence the continuous spectrum touches the imaginary axis at the origin when k is zero. The origin is a simple eigenvalue for $\mathcal{L}$ due to translational invariance of the underlying traveling wave $U_c(x)$ in the longitudinal direction. Thus, even though the essential spectrum does not affect linear stability, we cannot automatically deduce asymptotic stability unless we consider a gently weighted space of square integrable functions, which will shift the essential spectrum to the left.

We now notice that the spectral problem for the linear differential operator $\mathcal{L}$ is fourth order and one-dimensional; the transverse perturbation information is encoded through the transverse wavenumber parameter k. Hence we can employ the Evans function shooting approach; in particular, in Figure 1 we plot the dispersion relation for planar fronts corresponding to different values of the physical parameter δ. The dispersion relation gives the relationship

between the real part of the exponential growth factor of small perturbations, $\mathrm{Re}(\lambda)$, and transverse wavenumber k (in fact, to make later comparisons easier we have used $k/\tilde{L}$ on the abscissa). To construct the graph shown, we locate zeros of the Evans function on the real λ axis and follow them as the transverse wavenumber k is varied. We see in Figure 1 that the interval $(0, k_\delta)$ of unstable transverse wavenumbers k increases as δ increases (at least up to $\delta = 5$). The wavenumber where the maximum of $\mathrm{Re}(\lambda)$ occurs has the largest growth and is thus typically the transverse mode expressed in the dynamics.

However, our real concern here is the stability of the wrinkled cellular fronts themselves as we increase the diffusion ratio parameter δ beyond δ_{cr}. We know the planar waves are unstable in this regime and the question is whether the bifurcating wrinkled cellular fronts are stable for values of δ just beyond δ_{cr} but do become unstable at another critical value of δ. Using direct simulations of the reaction-diffusion system, Malevanets, Careta, and Kapral [24] demonstrated that when $\delta = 5$ the transverse structure of the propagating fronts was very complex with spatio-temporal dynamics of its own. Our goal is to elicit in practice the initial mechanisms that precipitate the behavior they observe through a careful study of the spectra of the underlying multidimensional traveling fronts.

6. Numerics.

6.1. The method. We will compute the spectrum of the two-dimensional traveling front by computing the Evans function associated with the linear operator $\mathcal{L}$ defined in (3.1). First, we compute the front with the freezing method, as explained in section 4. Then, as in section 3, we project the problem onto a finite number of transverse Fourier modes $k = -K, \ldots, K$. This generates the large linear system (3.6) which is of the standard linear form (2.5) where the coefficient matrix is given by

$$A(x;\lambda) = \begin{pmatrix} O_{N(2K+1)} & I_{N(2K+1)} \\ A_3(x;\lambda) & A_4 \end{pmatrix},$$

where the matrices A_3 and A_4 are explicitly given at the end of section 3. This coefficient matrix depends on both the traveling wave solution U_c and the spectral parameter λ. It also depends on the nonlinear reaction term. For our autocatalytic problem the Jacobian $\mathrm{D}F$ of the nonlinear reaction term and its Fourier transform $\hat{D}$ are given by

$$\mathrm{D}F(U) = \begin{pmatrix} -v^2 & -2uv \\ v^2 & 2uv \end{pmatrix} \qquad \text{and} \qquad \hat{D} = \begin{pmatrix} -\hat{v} * \hat{v} & -2\hat{u} * \hat{v} \\ \hat{v} * \hat{v} & 2\hat{u} * \hat{v} \end{pmatrix},$$

where $\hat{u}$ and $\hat{v}$ are the Fourier transforms of u and v, respectively, and $*$ denotes the convolution defined by $(\hat{u} * \hat{v})_k = \sum_{\ell=-\infty}^{\infty} \hat{u}_\ell \hat{v}_{k-\ell}$.

As described in section 2.5, we construct the unstable subspace $Y_0^-(\lambda)$ of (3.6) at $x = -\infty$ as the matrix whose columns are the eigenvectors of $A(-\infty;\lambda)$ corresponding to eigenvalues with a positive real part. Similarly, we construct the stable subspace $Y_0^+(\lambda)$ at $x = +\infty$ from the eigenvectors of $A(+\infty;\lambda)$ corresponding to eigenvalues with a negative real part. Then to construct the modified Evans function (2.7) at the matching point x_* we proceed as follows (for both the planar and wrinkled front cases for the autocatalytic problem below). The full system is of size $n = 4(2K+1)$. We use the coordinate patch identified by $\mathfrak{i}_- = \{1, \ldots, m\}$, where

$m = 2(2K+1)$ for the interval $(-\infty, x_*]$. On this interval we solve the Riccati equation (2.3a) for $\hat{y}_-$; with the chosen coordinate patch, the coefficient matrices in the Riccati equation are $a = O_m$, $b = I_m$, $c = A_3$, and $d = A_4$. The initial data for $\hat{y}_-$ is simply the lower $(n-m)\times m$ block of $Y_0^-(\lambda)$ postmultiplied by the inverse of the upper $m \times m$ block of $Y_0^-(\lambda)$. For the interval $[x_*, +\infty)$ we use the coordinate patch identified by $\mathtt{i}_+ = \{m+1, \dots, n\}$. We solve the Riccati equation (2.3a) for $\hat{y}_+$; the coefficient matrices in this case are $a = A_4$, $b = A_3$, $c = I_m$, and $d = O_m$. The initial data for $\hat{y}_+$ is the upper $m \times (n-m)$ block of $Y_0^+(\lambda)$ postmultiplied by the inverse of the lower $(n-m)\times(n-m)$ block of $Y_0^+(\lambda)$. Of course, in practical calculations, our intervals of integration are $[L_x^-, x_*]$ and $[x_*, L_x^+]$ for suitable choices of L_x^- and L_x^+. In particular, we assume the interval $[L_x^-, L_x^+]$ contains the traveling front and extends far enough in both directions that we can assume the front is sufficiently close to the far-field homogeneous steady states. Further, the respective coordinate patches for the left and right intervals, and matching point x_*, are chosen to avoid singularities in the Riccati flows in these intervals (in all our calculations for the autocatalytic problem we were always able to find such a matching point when using these two patches). Hence we evaluate the Evans function by computing the determinant (2.7). This information is used to find the spectrum of the traveling wave.

6.2. The planar front. We first consider planar traveling waves. This will not give us new information, but it is a good test for our procedure. The traveling waves can be found by the freezing method but also by the shooting method explained in Balmforth, Craster, and Malham [3] for the one-dimensional case. The next step is to find the unstable subspace of (3.5) at $x = -\infty$. The traveling wave U_c is planar and thus independent of y, so the Fourier coefficient $\hat{D}_k$ vanishes for $k \neq 0$. This implies that the differential equation (3.5) decouples in $2K+1$ systems, one for every $k = -K, \dots, K$, each of the form

$$\begin{aligned} \partial_x \hat{U}_k &= \hat{P}_k, \\ \partial_x \hat{P}_k &= \lambda B^{-1}\hat{U}_k + \left(\tfrac{k}{L}\right)^2 \hat{U}_k - B^{-1}\hat{D}_0(x)\,\hat{U}_k - c\,B^{-1}\hat{P}_k. \end{aligned}$$

For the autocatalytic system (5.1), this becomes

$$\partial_x \begin{pmatrix} \hat{U}_k \\ \hat{P}_k \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \frac{\lambda}{\delta} + \left(\frac{k}{L}\right)^2 + \frac{1}{\delta}u_2^2 & \frac{2}{\delta}u_1 u_2 & -\frac{c}{\delta} & 0 \\ -u_2^2 & \lambda + \left(\frac{k}{L}\right)^2 - 2u_1 u_2 & 0 & -c \end{pmatrix} \begin{pmatrix} \hat{U}_k \\ \hat{P}_k \end{pmatrix}. \tag{6.1}$$

We have $(u_1, u_2) \to (0, 1)$ in the limit $x \to -\infty$. The coefficient matrix in (6.1) has two unstable eigenvectors when λ is to the right of the continuous spectrum, and these eigenvectors are

$$\left(1, \tfrac{1}{\nu}, \mu, \tfrac{\mu}{\nu}\right)^{\mathrm{T}}, \qquad \text{where} \qquad \begin{cases} \mu = -\frac{c}{2\delta} + \sqrt{\frac{c^2}{4\delta^2} + \left(\frac{k}{L}\right)^2 + \frac{\lambda+1}{\delta}}, \\ \nu = \left(\frac{k}{L}\right)^2 + \lambda - c\mu - \mu^2, \end{cases} \tag{6.2a}$$

and

$$(0, 1, 0, \mu)^{\mathrm{T}}, \qquad \text{where} \qquad \mu = -\tfrac{1}{2}c + \sqrt{\tfrac{1}{4}c^2 + \left(\tfrac{k}{L}\right)^2 + \lambda}. \tag{6.2b}$$

These eigenvectors are all analytic functions of λ except at those points where either ν or one of the expressions under the square root sign vanishes. A small calculation shows that the latter can happen only when λ is on the negative real axis, so it will not interfere with stability computations, which concern eigenvalues with positive real part. In contrast, we may have $\nu = 0$ for positive λ, but in the situations considered here ν vanishes only for values of λ which are so large that they do not concern us here. Alternatively, the singularities caused by $\nu = 0$ can easily be avoided by using $(\nu, 1, \mu\nu, \mu)^{\mathrm{T}}$ as the eigenvector instead of $(1, 1/\nu, \mu, \mu/\nu)^{\mathrm{T}}$.

We thus have two unstable directions for every wavenumber k. Together these form the $2(2K + 1)$-dimensional unstable subspace $Y_0^-(\lambda)$ at $x = -\infty$. Using this, as described in section 6.1 above, we integrate the Riccati equation (2.3a) from $x = L_x^-$ to $x = x_*$, where L_x^- corresponds to a position sufficiently far behind the front. The integration is done with MATLAB function `ode45`, which is based on the explicit fifth order Runge–Kutta method due to Dormand and Prince. In all experiments, we use an absolute tolerance of 10^{-8} and a relative tolerance of 10^{-6}.

In the other limit, where $x \to +\infty$, we need the two stable eigenvectors of the coefficient matrix in (6.1), which are

$$(1, 0, \mu, 0)^{\mathrm{T}}, \qquad \text{where} \qquad \mu = -\tfrac{c}{2\delta} - \sqrt{\tfrac{c^2}{4\delta^2} + \left(\tfrac{k}{L}\right)^2 + \tfrac{\lambda}{\delta}}, \tag{6.3a}$$

and

$$(0, 1, 0, \mu)^{\mathrm{T}}, \qquad \text{where} \qquad \mu = -\tfrac{1}{2}c - \sqrt{\tfrac{1}{4}c^2 + \left(\tfrac{k}{L}\right)^2 + \lambda}. \tag{6.3b}$$

Again, we get a $2(2K + 1)$-dimensional subspace $Y_0^+(\lambda)$. We use this to generate the initial condition for the Riccati equation (2.3a). We integrate backward from $x = L_x^+$ to $x = x_*$ (note that the coefficients are different from those for the left interval); see section 6.1 above. Here L_x^+ is a longitudinal position sufficiently far ahead of the front. In our actual calculations we used $L_x^\pm = \pm 25$ and $x_* = 0$. Finally, the Evans function is computed using the modified form (2.7).

This concludes the explanation for the Riccati approach. The process for the Drury–Oja approach due to Humpherys and Zumbrun [20] is very similar. The difference is that it uses the Drury–Oja flow (2.4) to propagate the subspaces and (2.8) to evaluate the Evans function.

As an illustration, we show in Figure 2 a graph of the Evans function along the real axis for $\delta = 3$, with the Fourier cut-off at $K = 24$. We see that the Evans function has a zero at $\lambda = 0$ and double zeros around 0.0005, 0.0007, and 0.0011. The zero at $\lambda = 0$ corresponds to the eigenvalue at the origin caused by translational symmetry. The other zeros correspond to double eigenvalues. They have positive real part, and thus we can conclude that the planar wave is unstable for $\delta = 3$, coinciding with the long-established statements in section 5.

In fact, the zeros for the Evans function can be related to the dispersion curve in Figure 1. As mentioned above, the differential equation (3.5) decouples into $2K + 1$ subsystems of the form (6.1), each corresponding to the linear operator $\mathcal{L}$ given in (5.2) for a particular wave number. The basis vectors we choose for the unstable subspace are zero for all but one of these subsystems. Since the subsystems are decoupled, the solution at $x = x_*$ is also zero for all but one of these subsystems. Thus, the matrix $\big(Y^-(x_*;\lambda)\ Y^+(x_*;\lambda)\big)$ at the matching

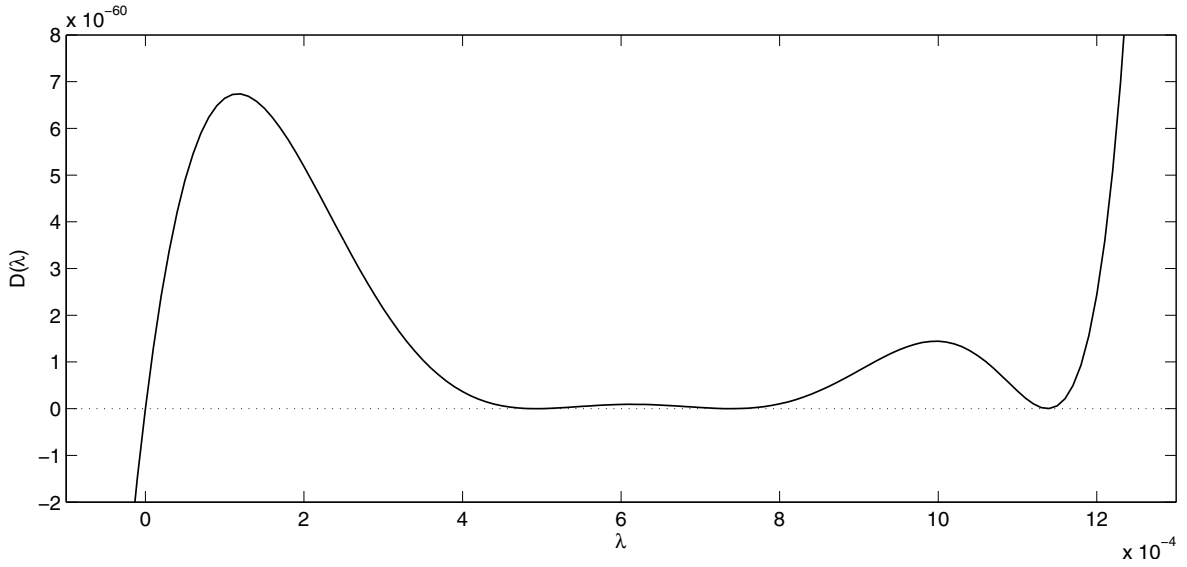


Figure 2. *The Evans function along the real axis for the planar front at $\delta = 3$. The Fourier cut-off is $K = 24$, and the transversal length of the domain is $L = 200$. The Evans function is computed using the Drury–Oja approach. A very similar plot can be produced using the Riccati approach.*

point is (after reordering) a block-diagonal matrix consisting of $2K + 1$ blocks of size 4×4, and its determinant is the product of the determinants of the 4×4 blocks. However, every determinant of a subblock is the Evans function D_k of the operator (5.2) for a particular wavenumber k. Thus, the Evans function for the two-dimensional planar wave is the product of the Evans functions for the one-dimensional wave with respect to transverse perturbations:

$$D(\lambda) = \prod_{k=-K}^{K} D_k(\lambda). \tag{6.4}$$

We choose $L = 200$, and so $k = 1$ corresponds to a wavenumber of $\frac{2\pi}{200} \approx 0.0314$ in Figure 1. The plot shows that the corresponding growth rate is approximately 0.0005; thus $D_1(\lambda)$ has a zero around $\lambda = 0.0005$. The dispersion relation is symmetric, and so $D_{-1}(\lambda)$ has a zero at the same value. This explains why the two-dimensional Evans function has a double zero around $\lambda = 0.0005$. Similarly, the double zeros around $\lambda = 0.0011$ and $\lambda = 0.0007$ correspond to $k = 2$ and $k = 3$, respectively.

As Figure 2 shows, the Evans function is of the order 10^{-60} in the interval of interest. It is important to emphasize that we are primarily interested in the zeros of the Evans function, which correspond to eigenvalues of the operator $\mathcal{L}$ defined in (3.1). The scale of the Evans function between the zeros, which depends on the normalization of the eigenvectors (6.2) and (6.3), is of lesser relevance in this context. However, for completeness, we address this issue in section 6.5.

6.3. The wrinkled front. Having established that our method works for planar fronts, we now turn to the wrinkled front. The most immediate problem is that of computing reliable steady wrinkled fronts. The procedure we use to do this is outlined as follows:

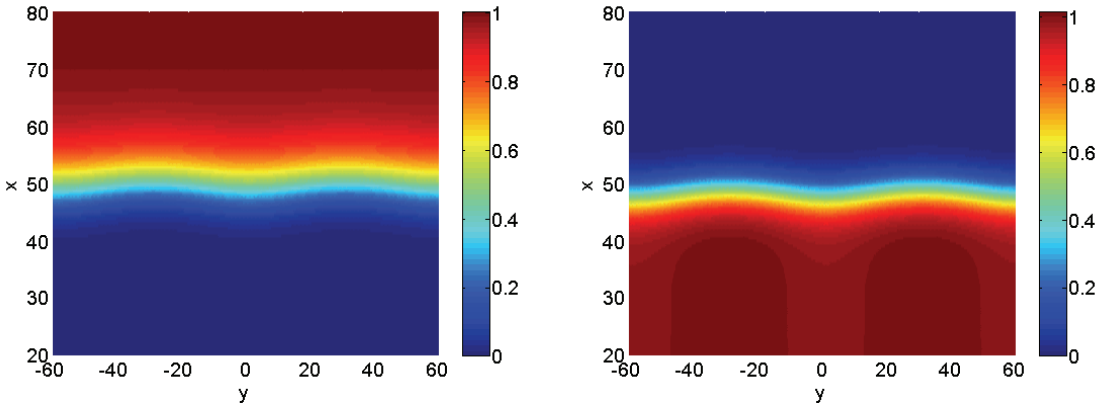


Figure 3. *The wrinkled front for $\delta = 3$. The left panel shows the u component, and the right panel shows the v component.*

- The partial differential algebraic system (4.2) is solved on the computational domain $[-150, 150] \times [-60, 60]$ with a second order finite element method on a grid of right triangles formed by dividing a grid of 300×240 rectangles along their diagonals. We use the finite element package COMSOL Multiphysics$^{\text{TM}}$ [9], which includes a version of the DAE solver `daspk`.
- To obtain reasonable starting waveform profiles for the partial differential algebraic system (4.2), we construct the one-dimensional waveform, which we then sweep out uniformly in the transverse direction.
- For traveling waves that appeared to be stable, we use the freezing method described in section 4 to obtain good starting waveform profiles to initiate the use of Newton's method in COMSOL.
- For traveling waves that appeared to be unstable, we use simple parameter continuation (also using COMSOL) starting with a (parametrically) nearby stable wave.

Figure 3 shows the front which results for $\delta = 3$.

The wrinkled front varies along the transverse y-direction. Therefore, the Fourier coefficients $\hat{D}_k$ do not vanish for $k \neq 0$, and the differential equation (3.6) does not decouple. In the limits $x \to \pm\infty$, we have restricted ourselves to wrinkled front profiles that approach a y independent state, which is the same as for the planar front. Hence, the computation of the stable and unstable subspaces for the planar front, which we reported above, remains valid for the wrinkled front.

The computation now proceeds as with the planar front. We solve the Riccati differential equation (2.3a) in the left and right intervals (as described in section 6.1 above) for the Riccati approach and (2.4) for the Drury–Oja approach. The front U_c is determined as explained above, with interpolation being used for those points that fall between the grid points of the finite element solution. We use cubic interpolation because linear interpolation is not sufficiently accurate. Finally, the Evans function is computed using either the modified form (2.7) or the Humpherys and Zumbrun form (2.8).

Figure 4 shows a plot of the Evans function along the real axis for $\delta = 3$, computed

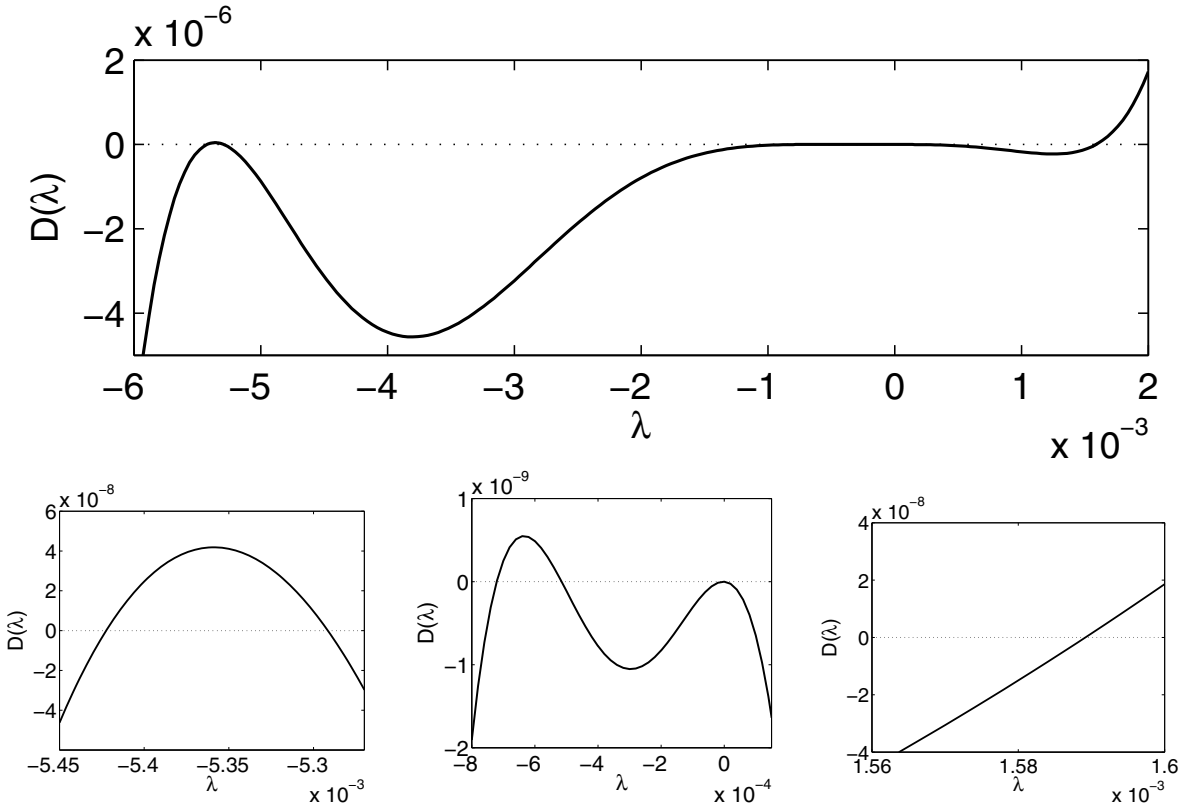


Figure 4. *The Evans function along the real axis for the wrinkled front at $\delta = 3$, computed using the Riccati approach. The three plots in the bottom row zoom in on the zeros of the Evans function.*

using the Riccati approach. The result of the Drury–Oja approach is shown in Figure 5. For both approaches, we used projection on $K = 9$ modes and we solved the Riccati differential equations (2.3a) (with appropriate coefficients for the left and right intervals), and the Drury–Oja equations (2.4), by the MATLAB ODE-solver `ode45` with absolute and relative tolerances of 10^{-8} and 10^{-6}. The domain is $[-150, 150] \times [-60, 60]$, and as the matching point we have chosen $x_* = 50$, which is roughly the front location (see Figure 3). Both approaches produce Evans functions with zeros at (approximately) the same λ-values: there is one eigenvalue with a positive real part around $\lambda = 0.0016$, a double zero around 0 (due to translation invariance), and pairs of eigenvalues around -0.0006 and -0.0053. In particular, the eigenvalue around $\lambda = 0.0016$ shows that the wrinkled front is unstable for $\delta = 3$.

However, the Riccati and Drury–Oja approaches do differ in some respects. The Riccati approach involves integrating a system of half the dimension as compared to the Drury–Oja approach, so we can expect it to be faster, especially for large values of K (in comparable experiments for large K, it was typically two to three times faster). See Ledoux, Malham, and Thümmler [22], where some detailed accuracy versus cost comparisons can be found. We also note that the Drury–Oja approach again produced very small values for the Evans function, which, as explained in section 6.5, is not as alarming as it first seems. The Riccati

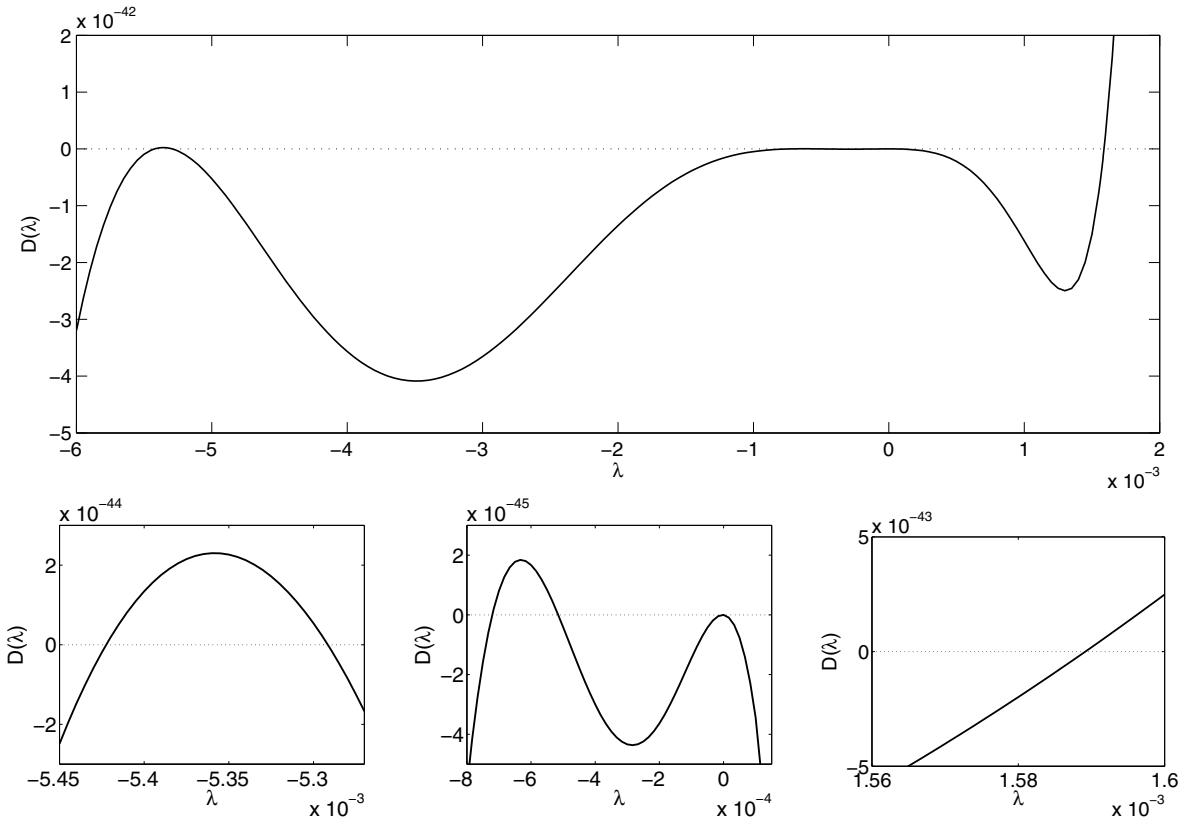


Figure 5. *Same as Figure 4, but with the Drury–Oja approach.*

approach, which uses a different scaling of the Evans function, yields only moderately small values, though again see section 6.5.

We also used the angle between the unstable and stable subspaces $\theta(\lambda; x_*)$ to determine the location of the eigenvalues. Though this scales better than the Evans determinant, it is nonnegative by definition and zeros occur as touch-downs in the complex spectral parameter plane. Actual zeros of the function are thus harder to establish compared to sign changes in the real and imaginary parts of the Evans determinant. Furthermore, $\theta(\lambda; x_*)$ is not analytic in λ.

More accurate values for the zeros of the Evans function can be found by using a standard root-finding method. This yields the values listed in Table 1. The results of the Riccati and Drury–Oja approaches agree (to the precision shown) in all cases. The table also shows results for different values of the wavenumber cut-off K. The eigenvalues converge quickly when K increases, and projection on $K = 7$ modes is sufficient to get six digits of accuracy, at least in this example. The largest computation we performed is with $K = 24$, in which case the one-dimensional equation has order 196.

We also computed the eigenvalues by a projection method to check these results. The COMSOL package, which we used to compute the traveling front, has an eigenvalue solver based on the Arnoldi method as implemented in ARPACK [23]. This solver uses a shift σ

Table 1

Zeros of the Evans function for the wrinkled front at $\delta = 3$ for different values of the wavenumber cut-off K, as computed with the Riccati and Drury–Oja approaches. The last row shows the eigenvalues computed by ARPACK.

K	Eigenvalues (Evans function)						
3	0.001609	−0.000026	−0.000781	−0.001296	−0.000670	−0.006078	−0.006177
4	0.001609	0.000002	−0.000001	−0.000519	−0.000670	−0.006078	−0.006177
5	0.001589	0.000002	−0.000001	−0.000519	−0.000720	−0.005295	−0.005426
6	0.001589	−0.000002	−0.000003	−0.000515	−0.000720	−0.005295	−0.005426
7	0.001589	−0.000002	−0.000003	−0.000515	−0.000721	−0.005292	−0.005422
8	0.001589	−0.000002	−0.000003	−0.000515	−0.000721	−0.005292	−0.005422
9	0.001589	−0.000002	−0.000003	−0.000515	−0.000721	−0.005292	−0.005422
⋮				⋮			
24	0.001589	−0.000002	−0.000003	−0.000515	−0.000721	−0.005292	−0.005422
	Eigenvalues (ARPACK)						
	0.001592	0.000000	0.000000	−0.000514	−0.000719	−0.005289	−0.005420

to allow the user to choose on which part of the spectrum to concentrate. We found that the choice of this shift is not straightforward. It is unwise to use a shift which is exactly an eigenvalue, so, in particular, the default value of $\sigma = 0$ should not be used. We used a shift of $\sigma = 1$ or $\sigma = 3$, which proved to be safe. The resulting eigenvalues are also listed in Table 1. They are in close agreement with the eigenvalues found using the Evans function, giving additional confidence in our computation.

Figure 6 shows the contours where the real and imaginary parts are zero; eigenvalues correspond with the intersections of these contours. All eigenvalues are real. Interestingly, the contours are well separated on the boundary of the depicted region of the complex plane even though the eigenvalues are packed closely together. This suggests that it may be better to study the Evans function in a contour surrounding the eigenvalues instead of concentrating on the eigenvalues themselves; see section 6.4.

Figures 7 and 8 show the Evans function for the case $\delta = 2.5$. The Evans function no longer has a zero with positive real part, meaning that there are no unstable eigenvalues. Hence for this value of δ the wrinkled front is in fact stable. We have thus established the following overall stability scenario for traveling wave solutions of the cubic autocatalysis problem. When δ is less than $\delta_{\mathrm{cr}} \approx 2.3$, planar traveling waves are stable to transverse perturbations. For δ beyond this instability there exist wrinkled fronts that are stable for $\delta = 2.5$. However, for $\delta = 3$ these wrinkled fronts themselves become unstable. These stability results may depend on the transverse domain size (which we fixed at $L = 120$); see Horváth et al. [19].

6.4. Global eigenvalue search methods. If one is interested only in the stability of the underlying front, then several methods exist to detect eigenvalues in the right-half complex spectral parameter plane—which also require relatively few Evans function evaluations. One important method uses the argument principle. We have the following sectorial estimate for the spectrum of $\mathcal{L}$: if $\lambda \in \sigma(\mathcal{L})$, then

$$\mathrm{Re}(\lambda) \le \|\mathrm{D}F(U_c)\|_{L^\infty(\mathbb{R}\times\mathbb{T})}, \tag{6.5a}$$

$$\mathrm{Re}(\lambda) + |\mathrm{Im}(\lambda)| \le \frac{c^2}{4\kappa} + 2\,\|\mathrm{D}F(U_c)\|_{L^\infty(\mathbb{R}\times\mathbb{T})}, \tag{6.5b}$$

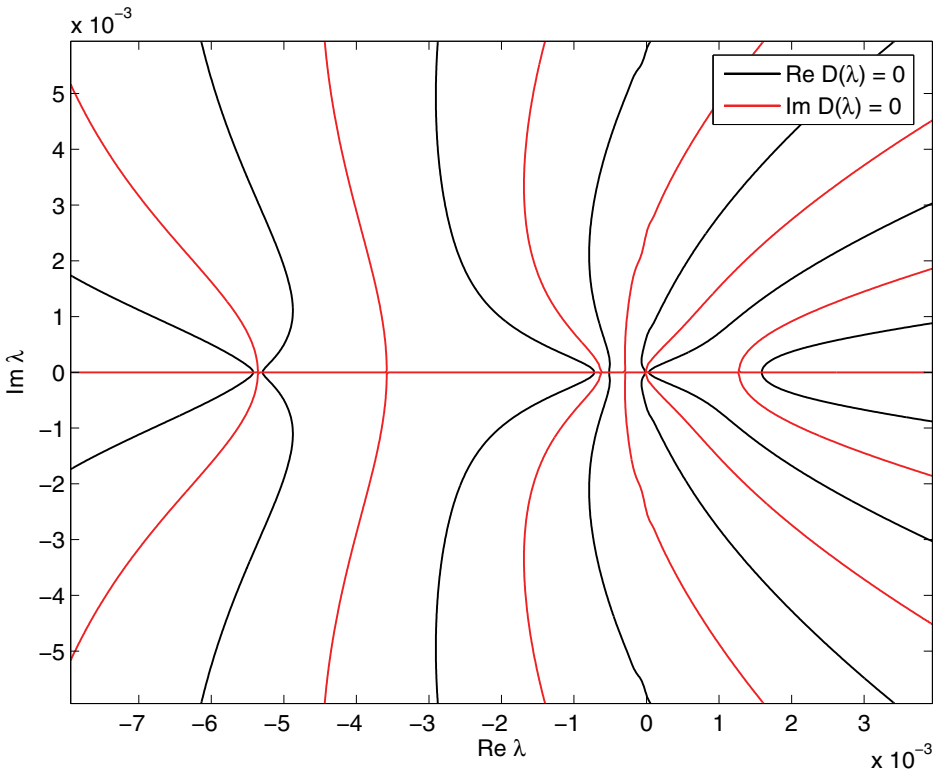


Figure 6. *The locus where the real and imaginary parts of the Evans function vanish for the wrinkled front at $\delta = 3$. This plot uses a 100×100 grid in the depicted region of the complex plane. The Evans function was computed on the 5000 grid points in the top half and extended by symmetry to the bottom half.*

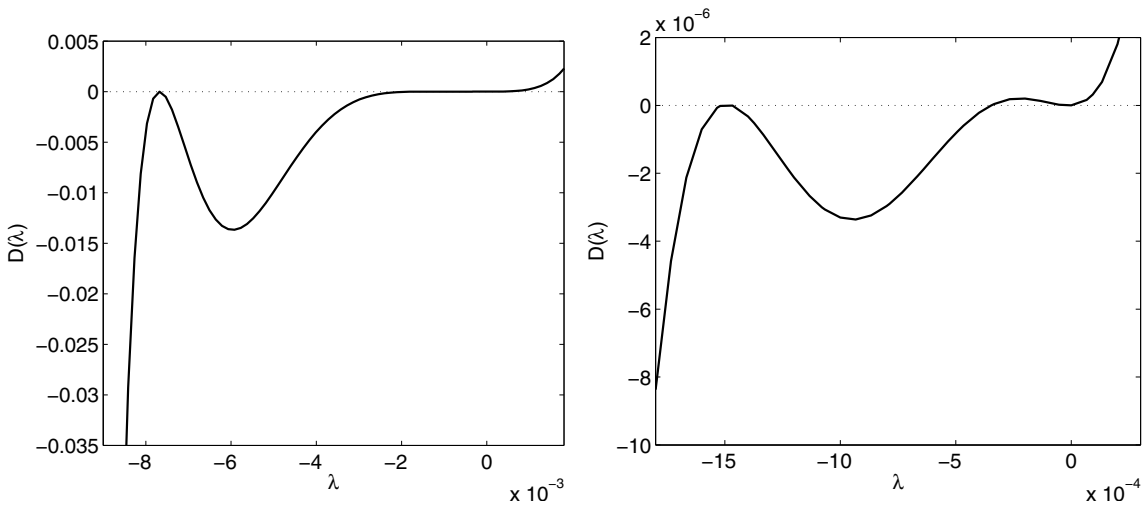


Figure 7. *The Evans function along the real axis for the wrinkled front at $\delta = 2.5$, computed using the Riccati approach. The right panel zooms in on the flat plateau around the origin in the left panel.*

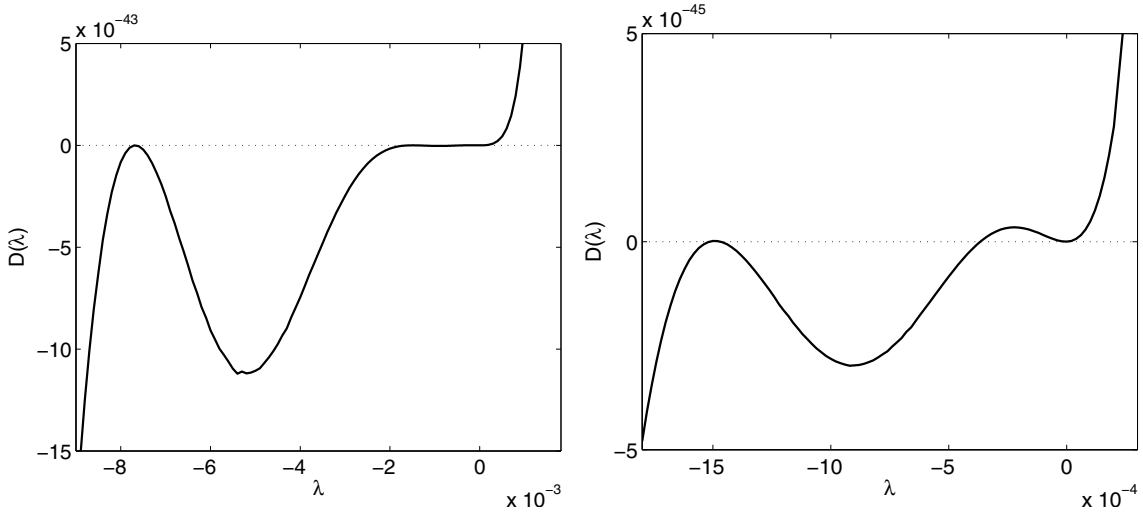


Figure 8. *Same as Figure 7, but with the Drury–Oja approach.*

where $\kappa \equiv \min\{B_{11}, B_{22}\}$; we provide a proof in Appendix A. Hence we can restrict our search for unstable eigenvalues to the region of the complex λ-plane bounded by these inequalities.

In our case, we have $c \approx 0.577$ and $\|\mathrm{D}F(U_c)\|_{L^\infty} \approx 1.422$ for $\delta = 2.5$, and $c \approx 0.548$ and $\|\mathrm{D}F(U_c)\|_{L^\infty} \approx 1.450$ for $\delta = 3$. So, the above estimate implies that any eigenvalue λ satisfies $\mathrm{Re}(\lambda) < 1.5$ and $\mathrm{Re}(\lambda) + |\mathrm{Im}(\lambda)| < 3$. Any unstable eigenvalue is thus inside the contour bounded by these lines and the imaginary axis; this contour is depicted in the left part of Figure 9 (the small semicircle around the origin is to exclude the double eigenvalue at the origin, as explained below). The Evans function is analytic inside this contour, and thus we can count the number of zeros inside the contour by computing the change of the argument of the Evans function as λ goes around the contour.

This suggests the following method. We divide the contour $\mathcal{C}$ into small segments, compute the Evans functions at the end points of the segments, find the change in the argument over each section, and sum these to find the total change over the whole contour. The number of zeros is then the change of the argument divided by 2π. The segments in which the contour is divided have to be chosen small enough that the change of the argument over each segment is smaller than π. Adaptive procedures to achieve this have been proposed by, among others, Ying and Katz [32], but for simplicity we determined the partition of the contour by trial and error.

The Evans function, with our choice of initial conditions, satisfies $D(\overline{\lambda}) = \overline{D(\lambda)}$. Hence, it suffices to transverse half the contour $\mathcal{C}$. The plot at the right of Figure 9 shows how the argument of $D(\lambda)$ changes as λ transverses the top half of the contour. We see that the change over the whole contour is -2π, and thus there is one zero inside the contour. This is the zero around $\lambda = 0.0016$ shown in Figure 5; there are no other unstable eigenvalues.

The operator $\mathcal{L}$ is invariant under translations, and hence it has two eigenvalues at $\lambda = 0$. Thus, the Evans function is zero at $\lambda = 0$ and its argument is undefined. For this reason, the contour $\mathcal{C}$ includes a small semicircle of radius 10^{-4} to avoid the origin. It is of course possible that we miss some eigenvalues because of this deformation. To guard against this, we

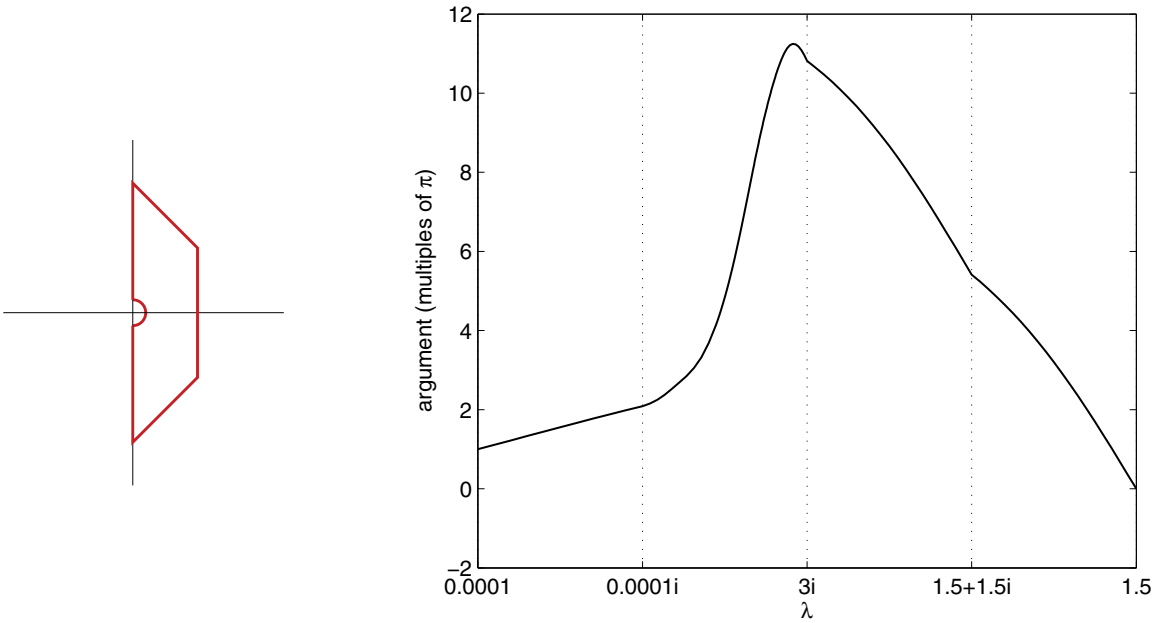


Figure 9. *The left figure shows the contour $\mathcal{C}$ in the complex plane. The right figure shows the argument of $D(\lambda)$ when λ transverses the top half of this contour (for the wrinkled front with $\delta = 3$). The plot is split up into four parts, corresponding to the quarter circle, the segment along the imaginary axis, the diagonal segment, and the segment going down. The scale on the horizontal axis is linear on the first, third, and fourth parts and logarithmic on the second part.*

also compute the change in the argument of $D(\lambda)$ as λ transverses a circle around the origin with radius 10^{-4}. We find that the argument changes by 4π, and hence the Evans function has two zeros inside the circle. This shows that we have not missed any eigenvalues by the deformation of $\mathcal{C}$ around the origin.

Another method is the parity-index argument (which we have not investigated here, but we include a description for completeness). To determine if there are eigenvalues on the positive real axis, we could use that

$$D(\lambda) \sim 2\delta^{-K-1/2}$$

as $\lambda \to +\infty$, as we prove in Appendix B. A change in sign of the Evans function along the real axis indicates an unstable eigenvalue. This can be exposed by comparing the sign of the Evans function in the asymptotic limit $\lambda \to +\infty$ to its sign elsewhere on the positive real axis, or the derivative or second derivative of the Evans function along the real axis evaluated at the origin. Note that to implement this technique we need to ensure that the Evans function evaluated in the neighborhood of the origin and the one generating the asymptotic limit have been normalized, as far as their signs are concerned, in the same way. This may require keeping careful track of the the nonzero analytic multiples dropped in our definition of the modified Evans function, i.e., the normalization of the initial data, the scalar exponential factor, and the determinants of u_- and u_+ (which can be computed analytically in the asymptotic limit also).

6.5. Convergence of the Evans function. The results reported in Table 1 show that the values of λ for which the Evans function $D(\lambda)$ is zero converge as $K \to \infty$. However, they do not show how the Evans function scales for other values of λ (not in the spectrum) as K becomes large. We see in Figure 2 for the planar front that, in the domain of interest with λ of order 10^{-4}, the Evans function computed using the Drury–Oja approach with $K = 24$ is of order 10^{-60}. Similarly, for the wrinkled front, we see in Figures 5 and 8, also computed using the Drury–Oja approach with $K = 9$, that the scale of the Evans function is of order 10^{-45} for λ of order 10^{-4}. These small results we find for the Evans function—far smaller than machine precision—suggest that the computation might be inaccurate because of round-off error. This is not the case here. The solution of the Drury–Oja equation (2.4) does not contain such small numbers. They appear only when the determinant is computed at the matching point, which is a stable computation; see, for example, Higham [18, sect. 13.5]. The determinant is computed by performing an LU-decomposition with partial pivoting and then multiplying the diagonal entries. Both steps are numerically stable even though the final result may be very small (as long as it does not underflow). The factorization (6.4) for the planar front provides another viewpoint: the Evans function for $K = 24$ is the product of $2K + 1 = 49$ one-dimensional Evans functions, and if each of them is 0.06—not a very small number—then their product is of the order 10^{-60}, which is a very small number.

For the modified Evans function that we construct using the Riccati approach, we see in Figure 4 that, in the domain of interest, it is of order 10^{-6} when $K = 9$. If we reproduce Figure 4 for increasing values of K, we find that the modified Evans function actually grows. For large K, it increases roughly by a factor of 10 per unit increase in K. However, it is important to emphasize that the solutions to the Riccati systems, which are used to construct the modified Evans function, remain bounded. It is only when we evaluate the determinant that, for example, when $K = 21$, the scale of the modified Evans function in Figure 4 is of order 10^4. Hence the question of the scaling of the Evans function comes into play in the multidimensional context.

In the one-dimensional case, the standard definition of the Evans function is (2.6), given in Alexander, Gardner, and Jones [1]. Of course this definition stands modulo any analytic nonzero multiplicative factor—after all, the zeros of the Evans function are what we are typically after. A different appropriate normalization of the initial data $Y_0^{\pm}(\lambda)$ rescales the Evans function by such a factor. Further, it is common to drop the scalar exponential factor. Hence, for example, the difference between the Humpherys and Zumbrun Evans function (2.8) and the standard Evans function is that the scalar exponential factor is dropped and the initial data might be normalized differently when it is QR-factorized. For the modified Evans function in (2.7), we dropped the exponential factor and also the determinants of some full rank submatrices in the determinant of the standard Evans function.

However, such analytic nonzero factors impact the scale of the Evans function in the multidimensional context, as they depend on K. This can lead to growth or decay of the Evans function employed away from the eigenvalues as K is increased, depending on which factors are kept and which are dropped. This is precisely what we have observed for the Drury–Oja and Riccati approaches, which decay and grow, respectively, as K increases.

Gesztesy, Latushkin, and Zumbrun [13, sect. 4.1.4] show that the Evans function, when suitably normalized, coincides with the 2-modified Fredholm determinant of a certain operator

related to $\mathcal{L}$. Importantly, the 2-modified Fredholm determinant converges as $K \to \infty$. They treat the self-adjoint case; the non-self-adjoint operator $\mathcal{L}$ we consider can, using suitable exponential weight functions, be transformed into a self-adjoint one (this is not numerically necessary, though, and could introduce further complications due to the introduction of exponential factors in the potential term). Gesztesy, Latushkin, and Zumbrun define the Evans function much like we have, through a Galerkin approximation, using the matching condition (2.6). Their normalization, natural to their context, involves ensuring the Evans function does not depend on the coordinate system chosen and is invariant to similarity transformations (see Gesztesy, Latushkin, and Makarov [12]). Crucially, Gesztesy, Latushkin, and Zumbrun demonstrate in [13, Theorem 4.15] that their 2-modified Fredholm determinant equals the Evans function multiplied by an exponential factor (nonzero and analytic), whose exponent diverges as $K \to \infty$ (it can diverge to positive or negative infinity depending on the precise form of the coefficient matrix $A(x;\lambda)$). Hence their Evans function diverges as well. Consequently, the 2-modified Fredholm determinant is a natural candidate for the canonical Evans function.

With this in mind, our observation regarding the measured divergence of the modified and Humpherys and Zumbrun Evans functions in our experiments is not too surprising. Their divergence, the modified Evans function to large values and the Humpherys and Zumbrun Evans function to small values, is simply a reflection of the over/underestimation of the crucial, correct multiplicative exponential factor that equates these Evans functions with the 2-modified Fredholm determinant of Gesztesy, Latushkin, and Zumbrun, which guarantees a finite limit as $K \to \infty$. It would be interesting to see whether one can explicitly compute the factor relating the Evans function of Gesztesy, Latushkin, and Makarov and the Evans function as defined in this paper.

7. Concluding remarks. Our goal in this paper was to show, for the first time, that spectral shooting methods can be extended to study the stability of genuinely multidimensional traveling fronts. It was already obvious to the people working in this field that this can be done by a Fourier decomposition in the transverse dimension, thus reducing the problem to a large one-dimensional problem; however, no one had yet managed to tackle the resulting large one-dimensional problems. This issue was overcome in theory in the work of Humpherys and Zumbrun [20] and more recently in that of Ledoux, Malham, and Thümmler [22], which showed that enough information for matching can be retained by integrating along the natural underlying Grassmannian manifolds (these ideas have been well known in control theory and quantum chemistry for a long while, though; see, for example, Hermann and Martin [17] and Johnson [21]). The dimension of the Grassmannian manifolds scales with the square of the order of the original problem. These methods brought the possibility of multidimensional shooting within our grasp, but there were still challenges to overcome. We have addressed the most immediate of these in this paper:

1. How to project the problem onto the finite transverse Fourier basis and ensure a numerically well-posed large system of linear ordinary differential spectral equations.
2. How to construct the genuinely multidimensional fronts; this was a particular challenge which we managed to overcome by using the combined techniques of freezing and parameter continuation (the latter for the unstable multidimensional fronts).
3. Choose a nontrivial but well-known example problem which exhibits the development

of planar front instabilities (an important initial testing ground for us) but also two-dimensional wrinkled fronts that themselves became unstable as a physical parameter was varied.

4. Apply the Humpherys and Zumbrun continuous orthogonalization as well as the Riccati methods to study the stability of the wrinkled fronts, in particular integrating a large system of order 196 using $K = 24$ transverse modes; hitherto the largest system considered in Evans function calculations was of order 6 (see Allen and Bridges [2]).
5. Ultimately demonstrating the *feasibility*, *accuracy*, *numerical convergence*, and *robustness* of multidimensional shooting to determine eigenvalues.

However, there is still more to do. Though multidimensional shooting is competitive in terms of accuracy, how does it compare in terms of computational cost? There are multiple sources of error including approximation of the underlying wrinkled front, domain truncation, transverse Fourier subspace finite projection, approximation of the solution to the projected one-dimensional spectral equations, and the matching position. Our experience is that the accuracy of the underlying wrinkled front is especially important. Although we have endeavored to keep all these errors small, not only for additional confidence in our numerical results, but also for optimal efficiency, we would like to know the relative influence of each one of these sources of error in order to target our computational effort more effectively. Finally, we would also like to see how the method transfers to hyperbolic traveling waves and more complicated mixed partial differential systems.

Appendix A. Sectorial estimate on the spectrum. We follow the standard arguments that prove that the parabolic linear operator $\mathcal{L}$ is sectorial; see, for example, Brin [6, p. 26]. First consider the L^2 complex inner product with the spectral problem

$$B\,\Delta U + c\partial_x U + \mathrm{D}F(U_c)\,U = \lambda U.$$

Premultiplying this spectral equation with $U^\dagger$ and integrating over the domain $\mathbb{R}\times\mathbb{T}$, we get

$$-\int_{\mathbb{R}\times\mathbb{T}} \nabla U^\dagger B \nabla U \,\mathrm{d}x\,\mathrm{d}y + c\int_{\mathbb{R}\times\mathbb{T}} U^\dagger\,\partial_x U + U^\dagger\,\mathrm{D}F(U_c)\,U\,\mathrm{d}x\,\mathrm{d}y = \lambda\|U\|_{L^2}^2.$$

In this last equation, if we set $B = P^{\mathrm{T}}P$, where $P \equiv \mathrm{diag}\{B_{11}^{1/2}, B_{22}^{1/2}\}$, we get

$$\lambda\|U\|_{L^2}^2 + \|P\nabla U\|_{L^2}^2 = c\int_{\mathbb{R}\times\mathbb{T}} U^\dagger\,\partial_x U + U^\dagger\,\mathrm{D}F(U_c)\,U\,\mathrm{d}x\,\mathrm{d}y. \tag{A.1}$$

If we consider the complex conjugate transpose of the spectral equation, postmultiply with U, and integrate over the domain $\mathbb{R}\times\mathbb{T}$, we get

$$\bar\lambda\|U\|_{L^2}^2 + \|P\nabla U\|_{L^2}^2 = c\int_{\mathbb{R}\times\mathbb{T}} \partial_x U^\dagger\,U + U^\dagger\left(\mathrm{D}F(U_c)\right)^{\mathrm{T}} U\,\mathrm{d}x\,\mathrm{d}y.$$

If we now add these last two equations, then we get

$$\begin{aligned}\mathrm{Re}(\lambda)\|U\|_{L^2}^2 + \|P\nabla U\|_{L^2}^2 &= \int_{\mathbb{R}\times\mathbb{T}} U^\dagger\,\tfrac{1}{2}\big(\mathrm{D}F + (\mathrm{D}F)^{\mathrm{T}}\big)(U_c)\,U\,\mathrm{d}x\,\mathrm{d}y \\ &\le \|\mathrm{D}F(U_c)\|_{L^\infty}\|U\|_{L^2}^2.\end{aligned} \tag{A.2}$$

Now dividing through by $\|U\|_{L^2}^2$, we get (6.5a).

Next we consider taking the imaginary part of (A.1), followed by the absolute value on both sides, and using Hölder's inequality to get

$$|\mathrm{Im}(\lambda)|\,\|U\|_{L^2}^2 \le c\|U\|_{L^2}\|\nabla U\|_{L^2} + \|\mathrm{D}F(U_c)\|_{L^\infty}\|U\|_{L^2}^2. \tag{A.3}$$

Adding inequalities (A.2) and (A.3) and using that $\|P\nabla U\|_{L^2}^2 \ge \kappa\|\nabla U\|_{L^2}^2$, where $\kappa \equiv \min\{B_{11}, B_{22}\}$, and the arithmetic-geometric mean inequality

$$c\|U\|_{L^2}\|\nabla U\|_{L^2} \le \frac{c^2}{4\kappa}\|U\|_{L^2}^2 + \kappa\|\nabla U\|_{L^2}^2,$$

we get (6.5b) after dividing through by $\|U\|_{L^2}^2$.

Appendix B. Evans function when spectral parameter is large. We follow the arguments for such results given in Alexander, Gardner, and Jones [1, p. 186] and Sandstede [26, p. 24]. We begin by rescaling $\tilde{x} = \sqrt{|\lambda|}\,x$ in (3.5) for each $k = -K, -K+1, \ldots, K$ and taking the limit $|\lambda| \to \infty$ to get

$$\partial_{\tilde{x}\tilde{x}}\hat{U}_k - \mathrm{e}^{\mathrm{i}\arg\lambda}B^{-1}\hat{U}_k = 0.$$

Hence we have a decoupled system of $2K+1$ equations of the form

$$\partial_{\tilde{x}}\begin{pmatrix}\hat{U}\\ \hat{P}\end{pmatrix} = \begin{pmatrix}O_{2(2K+1)} & I_2\otimes I_{2K+1}\\ \tilde{B}\otimes I_{2K+1} & O_{2(2K+1)}\end{pmatrix}\begin{pmatrix}\hat{U}\\ \hat{P}\end{pmatrix},$$

where $\tilde{B} = \mathrm{e}^{\mathrm{i}\arg\lambda}B^{-1}$. This is a constant coefficient linear problem and the so-called spatial eigenvalues μ are the roots of the characteristic polynomial

$$\det\bigl((\mu^2 I - \tilde{B})\otimes I_{2K+1}\bigr) = \bigl(\det(\mu^2 I - \tilde{B})\bigr)^{2K+1},$$

where for any matrix C we have used that $\det(C\otimes I_k) \equiv \det(C^k) \equiv (\det C)^k$. Hence the spatial eigenvalues are $\pm\sqrt{B_{11}}$, $\pm\sqrt{B_{22}}$, each with algebraic multiplicity $2K+1$. Associated with each of these four basic spatial eigenvalue forms we have the $(2K+1)$-dimensional eigenspaces:

$$\begin{aligned}
\mu = +\sqrt{B_{11}}\colon &\quad \mathrm{span}\{e_{2k}^{+}(+\sqrt{B_{11}})\colon k = -K, -K+1, \ldots, +K\},\\
\mu = -\sqrt{B_{11}}\colon &\quad \mathrm{span}\{e_{2k}^{-}(-\sqrt{B_{11}})\colon k = -K, -K+1, \ldots, +K\},\\
\mu = +\sqrt{B_{22}}\colon &\quad \mathrm{span}\{e_{2k+1}^{+}(+\sqrt{B_{22}})\colon k = -K, -K+1, \ldots, +K\},\\
\mu = -\sqrt{B_{22}}\colon &\quad \mathrm{span}\{e_{2k+1}^{-}(-\sqrt{B_{22}})\colon k = -K, -K+1, \ldots, +K\},
\end{aligned}$$

where $e_i(\beta)$ is the vector with 1 in position i, with β in position $2(2K+1)+i$, and with zeros elsewhere. Hence, as $|\lambda| \to \infty$, the Evans function has the asymptotic form

$$\begin{aligned}
D(\lambda) &\sim \det\begin{pmatrix}I_2\otimes I_{2K+1} & -I_2\otimes I_{2K+1}\\ \tilde{B}^{1/2}\otimes I_{2K+1} & \tilde{B}^{1/2}\otimes I_{2K+1}\end{pmatrix}\\
&= 2\det\bigl(\tilde{B}^{1/2}\otimes I_{2K+1}\bigr)\\
&= 2\left(\delta^{-1/2}\mathrm{e}^{\mathrm{i}\arg\lambda}\right)^{2K+1}.
\end{aligned}$$

Acknowledgments. Simon J. A. Malham attended a workshop at AIM in Palo Alto in May 2005 on *Stability Criteria for Multidimensional Waves and Patterns* organized by Chris Jones, Yuri Latushkin, Bob Pego, Björn Sandstede, and Arnd Scheel that instigated this work. The authors would like to thank Marcel Oliver for his invaluable advice and help at the beginning of this project and also Gabriel Lord, Tom Bridges, Björn Sandstede, and Jacques Vanneste for useful discussions. A large part of this work was conceived and computed while all four authors were visiting the Isaac Newton Institute in the Spring of 2007. We would like to thank Arieh Iserles and Ernst Hairer for inviting us and providing so much support and enthusiasm. Last, we would like to thank the anonymous referees whose helpful comments and suggestions markedly improved the overall presentation of the paper.

REFERENCES

[1] J. C. Alexander, R. Gardner, and C. K. R. T. Jones, *A topological invariant arising in the stability analysis of traveling waves*, J. Reine Angew. Math., 410 (1990), pp. 167–212.
[2] L. Allen and T. J. Bridges, *Numerical exterior algebra and the compound matrix method*, Numer. Math., 92 (2002), pp. 131–149.
[3] N. J. Balmforth, R. V. Craster, and S. J. A. Malham, *Unsteady fronts in an autocatalytic system*, Proc. R. Soc. Lond. A, 455 (1999), pp. 1401–1433.
[4] J. Billingham and D. Needham, *The development of travelling waves in quadratic and cubic autocatalysis with unequal diffusion rates,* I, Phil. Trans. R. Soc. Lond. A, 334 (1991), pp. 1–24.
[5] J. Billingham and D. Needham, *The development of travelling waves in quadratic and cubic autocatalysis with unequal diffusion rates,* II, Phil. Trans. R. Soc. Lond. A, 336 (1991), pp. 497–539.
[6] L. Q. Brin, *Numerical Testing of the Stability of Viscous Shock Waves*, Ph.D. thesis, Indiana University, Bloomington, IN, 1998.
[7] W.-J. Beyn and V. Thümmler, *Freezing solutions of equivariant evolution equations*, SIAM J. Appl. Dyn. Syst., 3 (2004), pp. 85–116.
[8] Å. Björck and G. H. Golub, *Numerical methods for computing angles between linear subspaces*, Math. Comp., 27 (1973), pp. 579–594.
[9] Comsolab, *COMSOL Multiphysics* 3.3, COMSOL, Inc., Burlington, MA, 2007, http://www.comsol.com.
[10] J. Deng and S. Nii, *Infinite-dimensional Evans function theory for elliptic eigenvalue problems in a channel*, J. Differential Equations, 225 (2006), pp. 57–89.
[11] J. W. Evans, *Nerve axon equations,* IV: *The stable and unstable impulse*, Indiana Univ. Math. J., 24 (1975), pp. 1169–1190.
[12] F. Gesztesy, Y. Latushkin, and K. A. Makarov, *Evans functions, Jost functions, and Fredholm determinants*, Arch. Rational Mech. Anal., 186 (2007), pp. 361–421.
[13] F. Gesztesy, Y. Latushkin, and K. Zumbrun, *Derivatives of (modified) Fredholm determinants and stability of standing and travelling waves*, J. Math. Pures Appl., 90 (2008), pp. 160–200.
[14] L. Greenberg and M. Marletta, *Numerical solution of non–self-adjoint Sturm–Liouville problems and related systems*, SIAM J. Numer. Anal., 38 (2001), pp. 1800–1845.
[15] P. Griffiths and J. Harris, *Principles of Algebraic Geometry*, Wiley Classics Library Edition, Wiley, New York, 1994.
[16] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer–Verlag, New York, 1981.
[17] R. Hermann and C. Martin, *Lie and Morse theory for periodic orbits of vector fields and matrix Riccati equations,* I: *General Lie-theoretic methods*, Math. Systems Theory, 15 (1982), pp. 277–284.
[18] N. Higham, *Accuracy and Stability of Numerical Algorithms*, 2nd ed., SIAM, Philadelphia, 2002.
[19] D. Horváth, V. Petrov, S. K. Scott, and K. Showalter, *Instabilities in propagating reaction-diffusion fronts*, J. Chem. Phys., 98 (1993), pp. 6332–6343.
[20] J. Humpherys and K. Zumbrun, *An efficient shooting algorithm for Evans function calculations in large systems*, Phys. D, 220 (2006), pp. 116–126.

[21] B. R. Johnson, *New numerical methods applied to solving the one-dimensional eigenvalue problem*, J. Chem. Phys., 67 (1977), pp. 4086–4093.
[22] V. Ledoux, S. J. A. Malham, and V. Thümmler, *Grassmannian spectral shooting*, Math. Comp., submitted; available online at http://arxiv.org/abs/0710.1037v2, 2008.
[23] R. B. Lehoucq, D. C. Sorensen, and C. Yang, *ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*, Software Environ. Tools 6, SIAM, Philadelphia, 1998.
[24] A. Malevanets, A. Careta, and R. Kapral, *Biscale chaos in propagating fronts*, Phys. Rev. E, 52 (1995), pp. 4724–4735.
[25] J. W. Milnor and J. D. Stasheff, *Characteristic Classes*, Ann. of Math. Stud. 76, Princeton University Press, Princeton, NJ, 1974.
[26] B. Sandstede, *Stability of travelling waves*, in Handbook of Dynamical Systems II, B. Fiedler, ed., North–Holland, Amsterdam, 2002, pp. 983–1055.
[27] B. Sandstede and A. Scheel, *Absolute and convective instabilities of waves on unbounded and large bounded domains*, Phys. D, 145 (2000), pp. 233–277.
[28] J. Schiff and S. Shnider, *A natural approach to the numerical integration of Riccati differential equations*, SIAM J. Numer. Anal., 36 (1999), pp. 1392–1413.
[29] N. Steenrod, *The Topology of Fibre Bundles*, Princeton University Press, Princeton, NJ, 1951.
[30] D. Terman, *Stability of planar wave solutions to a combustion model*, SIAM J. Math. Anal., 21 (1990), pp. 1139–1171.
[31] V. Thümmler, *Numerical Analysis of the Method of Freezing Traveling Waves*, Ph.D. thesis, Bielefeld University, Bielefeld, Germany, 2005.
[32] X. Ying and I. Katz, *A reliable argument principle algorithm to find the number of zeros of an analytic function in a bounded domain*, Numer. Math., 53 (1988), pp. 143–163.

SIAM J. Applied Dynamical Systems
Vol. 8, No. 1, pp. 508–526
© 2009 Society for Industrial and Applied Mathematics

Study of Singularities in Nonsmooth Dynamical Systems via Singular Perturbation*

Jaume Llibre†, Paulo R. da Silva‡, and Marco A. Teixeira§

Abstract. In this article we describe some qualitative and geometric aspects of nonsmooth dynamical systems theory around typical singularities. We also establish an interaction between nonsmooth systems and geometric singular perturbation theory. Such systems are represented by discontinuous vector fields on $\mathbb{R}^\ell$, $\ell \geq 2$, where their discontinuity set is a codimension one algebraic variety. By means of a regularization process proceeded by a blow-up technique we are able to bring about some results that bridge the space between discontinuous systems and singularly perturbed smooth systems. We also present an analysis of a subclass of discontinuous vector fields that present transient behavior in the 2-*dimensional* case, and we dedicate a section to providing sufficient conditions in order for our systems to have local asymptotic stability.

Key words. regularization, vector fields, singular perturbation, discontinuous vector fields, sliding vector fields

AMS subject classifications. 34C20, 34C26, 34D15, 34H05

DOI. 10.1137/080722886

1. Introduction.

The purpose of this article is to present some aspects of the geometric theory of a class of nonsmooth systems. Our main concern is understanding the dynamics of such systems by means of tools in the geometric singular perturbation theory. Many similarities between such fields were observed, and a comparative study of the two categories is called for. The main task for the future is to bring the theory of nonsmooth dynamical systems to a maturity similar to that of smooth systems. Needless to say, geometric singular perturbation theory is an important tool in the field of continuous dynamical systems, of which very good surveys are available (see [6, 7, 9]). The techniques of geometric singular perturbation theory can be used to obtain information on the dynamics of the nonsmooth system, mainly in searching minimal sets.

The study of nonsmooth dynamical systems has in recent years established an important bridge between Mathematics, Physics, and Engineering. The book [5] presents motivating models that arise in the occurrence of impact motion in impact systems as well as switchings in electronic systems and hybrid dynamics in control systems. For a survey on qualitative

*Received by the editors April 30, 2008; accepted for publication (in revised form) by H. Kokubu December 17, 2008; published electronically March 18, 2009.
http://www.siam.org/journals/siads/8-1/72288.html
†Departament de Matemàtiques, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Catalonia, Spain (jllibre@mat.uab.cat). This author was partially supported by MTM2005-06098-C02-01 and by CICYT grant 2005SGR 00550.
‡Departamento de Matemática, IBILCE–UNESP, Rua C. Colombo, 2265, CEP 15054–000 S. J. Rio Preto, São Paulo, Brazil (prs@ibilce.unesp.br). This author was partially supported by FAPESP, CNPq, CAPES, and FAPESP–BRAZIL grant 2007/06896-5.
§IMECC–UNICAMP, CEP 13081–970, Campinas, São Paulo, Brazil (teixeira@ime.unicamp.br). This author was partially supported by FAPESP–BRAZIL grant 2007/06896-5.

aspects of such systems we refer the reader to [14] and references therein. In many applications examples of nonsmooth systems where the discontinuities are located on algebraic varieties are available (see, for example, [1, 2]). See, for instance, the system represented by $\ddot{x} + x\, sign(x) + sign(\dot{x}) = 0$. Concerning theoretical results on this subject we refer the reader to [3, 12]. This paper is mainly motivated by such issues. It is worthwhile to point out that in [4] and [10] preceding results were established in two dimensions and three dimensions, respectively, when the discontinuity set is a codimension one submanifold.

Let X be a nonsmooth vector field in $\mathbb{R}^l$, and denote by $\mathcal{S}$ its discontinuity set. First we focus our attention on the flow (determined by the orbits of X tending to $\mathcal{S}$) that is tangent to $\mathcal{S}$, which is denominated the *sliding mode*. It appears when the flow of X across $\mathcal{S}$ and points "inward" cannot leave such a surface.

Now we address the discussion to a more general context. Let $\mathcal{U} \subseteq \mathbb{R}^\ell$, $\ell \geq 2$, be an open set with $0 \in \mathcal{U}$. Let $S = \mathcal{S}_1 \bigcup \mathcal{S}_2$, where $\mathcal{S}_1$ and $\mathcal{S}_2$ are codimension one submanifolds of $\mathcal{U}$ with $0 \in \mathcal{S}_1 \cap \mathcal{S}_2$ that are in general position. Around $0 \in \mathcal{U}$ we have that $\mathcal{S}_1 \cup \mathcal{S}_2$ separates $\mathcal{U}$ into four open quadrants: $M_1, \ldots, M_4$. In our approach $\mathcal{S}$ will be the discontinuity boundary of our systems, also named the *switching variety*.

Let X_i, $i = 1, \ldots, 4$, be C^κ vector fields, with $\kappa \geqslant 1$, $\kappa = \infty$, or $\kappa = \omega$, defined on $\mathcal{U}$. We are concerned with the behavior of the sliding mode generated from a discontinuous differential system expressed by

$$\dot{z} = X(z) = X_i(z), \quad z \in M_i, \quad i = 1, \ldots, 4. \tag{1.1}$$

We will denote these systems by $X = (X_1, \ldots, X_4) \in \Omega^\kappa_{1234}(\mathcal{U})$ and the intersection $\overline{M_i} \cap \overline{M_j}$ by $\mathcal{S}_{ij}$.

Throughout the paper we consider local coordinates $(x_1, x_2, \ldots, x_\ell)$ such that $\mathcal{S}_1 = \mathcal{S}_{14} \cup \mathcal{S}_{23} = \{x_2 = 0\}$, $\mathcal{S}_2 = \mathcal{S}_{12} \cup \mathcal{S}_{34} = \{x_1 = 0\}$, and

$$M_1 = \{x_1 > 0,\, x_2 > 0\}, \quad M_2 = \{x_1 < 0,\, x_2 > 0\},$$
$$M_3 = \{x_1 < 0,\, x_2 < 0\}, \quad M_4 = \{x_1 > 0,\, x_2 < 0\}.$$

Denote by $\mathcal{S}^* = \mathcal{S} \setminus (\mathcal{S}_1 \cap \mathcal{S}_2)$ the regular part of $\mathcal{S}$. In $\mathcal{S}^*$ the definition of an orbit-solution obeys, whenever possible, the Filippov convention (see [8]). Consider M_i and M_j, $i \neq j$, having a common boundary. According to this convention there may exist generically a *sliding region* $\mathcal{S}^{sl} \subset \mathcal{S}^*$ such that any orbit which meets $\mathcal{S}^{sl}$ remains tangent to $\mathcal{S}^*$ for positive time. This region is the part of $\mathcal{S}^*$ on which X_i and X_j point inward to $\mathcal{S}^*$. Analogously there may exist generically an *escaping region* $\mathcal{S}^{es} \subset \mathcal{S}^*$ such that any orbit which meets $\mathcal{S}^{es}$ remains tangent to $\mathcal{S}^*$ for negative time.

On $\mathcal{S}^{sl} \bigcup \mathcal{S}^{es}$ the flow slides on $\mathcal{S}^*$; the flow follows a well-defined vector field $X^{\mathcal{S}}$ called the *sliding vector field*. See Figure 1. The sewing region $\mathcal{S}^{sw} \subset \mathcal{S}^*$ is the part of $\mathcal{S}^*$ where the flow crosses $\mathcal{S}^*$. The boundary between the three regions is the locus of points where the vector field is tangent to $\mathcal{S}^*$ and the flow grazes the switching surface. In our context such points together with the critical points of $X^{\mathcal{S}}$ and $\mathcal{S}_1 \cap \mathcal{S}_2$ constitute the set of singularities of X.

We remark that outside the three above-mentioned open regions (grazing regions), non-uniqueness of solutions is allowed, and the setting of local qualitative analysis has successfully

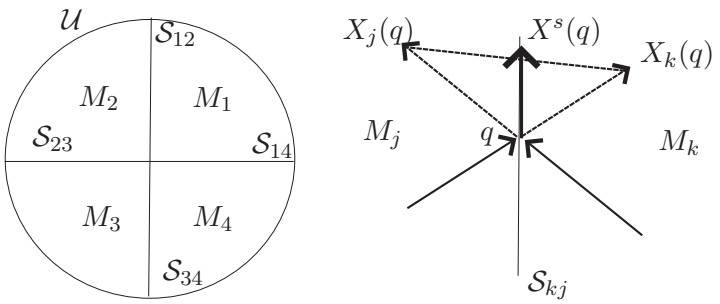


Figure 1. *Boundary of four regions and the sliding vector field.*

been the object of many studies. In the specific topic addressed in this paper the situation may be even more complicated since we want to study the dynamics of the vector field X around the origin. Alexander and Seidman in [1, 2] established a discussion of such a situation by using two mechanisms termed *blending* and *hysteresis*. Our approach is closest in spirit to the works [4] and [10], and one of our concerns is to know when a sliding flow in S^* can be continued until the intersection $\mathcal{S}_1 \cap \mathcal{S}_2$. In addition, we also present some applications of the techniques developed in [4] and [10] for analyzing the behavior of the so-called transient flow around 0.

We begin by explaining the main concepts with a concrete example before discussing the general result of this setting.

Example. Consider $\mathcal{U} \subseteq \mathbb{R}^3$, an open set with $0 \in \mathcal{U}$ and $X = (X_1, \dots, X_4) \in \Omega^{\kappa}_{1234}(\mathcal{U})$. We denote $X_i = (f_i, g_i, h_i)$, $i = 1, \dots, 4$. Suppose that $f_1 < 0$, $g_1 < 0$; $f_2 > 0$, $g_2 < 0$; $f_3 > 0$, $g_3 > 0$; and $f_4 < 0$, $g_4 > 0$. This means that all vector fields point inward to the intersecting switching curve $\Sigma = \mathcal{S}_1 \cap \mathcal{S}_2 = \{(0, 0, x_3)\}$. It is natural to extend the definition of the sliding vector field to this set. A possible sliding vector field is a convex combination of X_1, X_2, X_3, and X_4. We denote this combination by

$$X^{\Sigma}_{\lambda_1,\dots,\lambda_4} = \lambda_1 X_1 + \cdots + \lambda_4 X_4, \quad \sum_{i=1}^{i=4} \lambda_i = 1.$$

If

$$\operatorname{rank} \begin{bmatrix} f_1(0) & f_2(0) & f_3(0) & f_4(0) \\ g_1(0) & g_2(0) & g_3(0) & g_4(0) \\ 1 & 1 & 1 & 1 \end{bmatrix} = 3,$$

then there exists an open set $0 \in \mathcal{U} \subseteq \mathbb{R}^3$ such that for any $(0, 0, x_3) \in \mathcal{U}$ the set

$$\left\{ \lambda \in \mathbb{R}^4; \sum_{i=1}^{i=4} \lambda_i = 1,\ X^{\Sigma}_{\lambda_1,\dots,\lambda_4}(0, 0, x_3) \in \Sigma \right\}$$

is a nonempty set with dimension one in $\mathbb{R}^4$. Clearly we face here an ambiguity situation, and naturally some questions arise. For example, *What is required to avoid such ambiguity?*, or *How about the dynamics of smooth systems nearby X?*

Let $X = (X_1, \dots, X_4) \in \Omega^{\kappa}_{1234}(\mathcal{U})$. Summarizing, in what follows we give a rough overall description of the main results of this paper.

(a) There exist four curves

$$\{\gamma_i(\psi, x_1, \rho) = 0\}, \quad \{\alpha_i(\theta, x_2, \rho) = 0\}, \quad i = 1, 2,$$

with $\theta, \psi \in [0, \pi]$, $x_1, x_2 \in \mathbb{R}$, $\rho \in \mathbb{R}^{\ell-2}$, and such that the sliding region $\mathcal{S}^{sl} \bigcup \mathcal{S}^{es}$ on the regular part $\mathcal{S}^*$ is homeomorphic to one of them. Moreover, the sliding vector field $X^{\mathcal{S}}$ is topologically equivalent to one of the following reducing problems:

$$\gamma_i(\psi, x_1, \rho) = 0, \quad \dot{x_1} = \delta_i(\psi, x_1, \rho), \quad \dot{\rho} = \nu_i(\psi, x_1, \rho)$$

or

$$\alpha_i(\theta, x_2, \rho) = 0, \quad \dot{x_2} = \beta_i(\theta, x_2, \rho), \quad \dot{\rho} = \sigma_i(\theta, x_2, \rho).$$

See Theorem 2.2(a), (b).

(b) For $\ell \geq 3$ there exists a differential system

$$\zeta(\theta, \psi, \rho) = 0, \quad \xi(\theta, \psi, \rho) = 0, \quad \rho' = \phi(\theta, \psi, \rho),$$

which extends the concept of the Filippov system for the case $\mathcal{S}_1 \cap \mathcal{S}_2$. See Theorem 2.2(c).

(c) For $\ell = 2$, $X_1 = X_3$, and $X_2 = X_4$, the flow of the regular vector fields which approach $X = (X_1, X_2, X_1, X_2)$ in the transient case is described. See Proposition 5.4.

2. Preliminaries and statement of results. In this section basic concepts and the main result of the paper are presented.

The sliding vector field $X^{\mathcal{S}}$ is defined at $q \in \mathcal{S}_{kj} \cap (\mathcal{S}^{sl} \bigcup \mathcal{S}^{es})$ by $X^{\mathcal{S}}(q) = m - q$, with m being the point where the segment joining $q + X_k(q)$ and $q + X_j(q)$ cuts $\mathcal{S}_{kj}$. In previous works (see [4, 10, 11]) it was shown that if $p = (x_1, \dots, x_\ell) \in \mathcal{S}_1 \cup \mathcal{S}_2$ and $x_1^2 + x_2^2 \neq 0$, then the sliding vector field on a neighborhood $\mathcal{V}$ of p can be studied via singular perturbation theory.

Suppose that $p = (0, x_2, \dots, x_\ell) \in \mathcal{S}_{12}$, $X = X_1 = (f_1, g_1) \in \mathbb{R} \times \mathbb{R}^{\ell-1}$ if $x_1 > 0$, $X = X_2 = (f_2, g_2) \in \mathbb{R} \times \mathbb{R}^{\ell-1}$ if $x_1 < 0$, and $f_1(p) \cdot f_2(p) \neq 0$.

Definition 2.1. *A C^∞-function $\varphi : \mathbb{R} \longrightarrow \mathbb{R}$ is a transition function if $\varphi(s) = -1$ for $s \leqslant -1$, $\varphi(s) = 1$ if $s \geqslant 1$, and $\varphi'(s) > 0$ if $s \in (-1, 1)$.*

The *φ_{x_2}-regularization* of X is the one-parameter family given by

$$X^{12}_{\varepsilon} = (1/2)\left[(1 + \varphi(x_1/\varepsilon))\, X_1 + (1 - \varphi(x_1/\varepsilon))\, X_2\right] \tag{2.1}$$

for $x_2 > 0$, $\varepsilon > 0$.

Consider, around the point p, the surface composed by the union of $M_1 \cap \mathcal{V}$, $M_2 \cap \mathcal{V}$, and $\widehat{\mathcal{S}_{12}} = \{(\theta, \rho) : \theta \in (0, \pi), \rho = (x_2, \dots, x_\ell) \in \mathcal{S}_{12} \cap \mathcal{V}\}$. We denote $\mathcal{M} = (M_1 \cap \mathcal{V}) \cup \widehat{\mathcal{S}_{12}} \cup (M_2 \cap \mathcal{V})$ and remark that the set $\{(0, \rho) : \rho \in S_{12} \cap \mathcal{V}\}$ has two distinct copies: $\partial(M_1 \cap \mathcal{V})$ and $\partial(M_2 \cap \mathcal{V})$.

The blow-up process is the change of coordinates $x_1 = r \cos\theta$, $\varepsilon = r \sin\theta$. It induces a smooth vector field on $\mathcal{M}$ whose trajectories coincide with those of X_1 on M_1, of X_2 on M_2, and of a singular perturbation problem on $\widehat{\mathcal{S}_{12}}$ described by

$$\theta' = \alpha(r, \theta, \rho), \quad \rho' = r\beta(r, \theta, \rho), \tag{2.2}$$

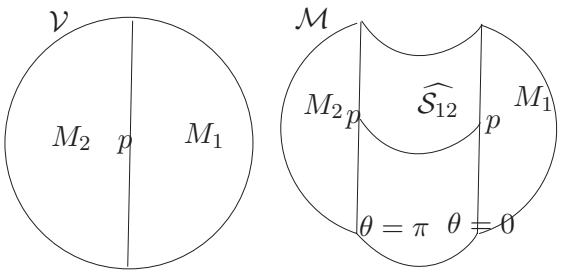


Figure 2. *The open set $\mathcal{V}$ and the set $\mathcal{M}$.*

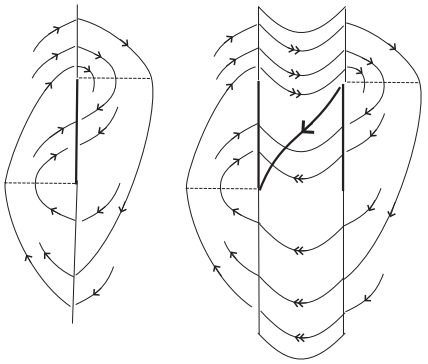


Figure 3. *Fast and slow dynamics with the slow manifold connecting two folds.*

with $r \geq 0$, $\theta \in (0, \pi)$, $\rho \in \mathcal{S}_{12} \cap \mathcal{V}$, and α and β of class C^κ. We remark that the functions $\alpha(r, \theta, \rho)$ and $\beta(r, \theta, \rho)$ are given by

$$\alpha = -\sin\theta \left[\frac{f_1 + f_2}{2} + \varphi(\cot\theta) \frac{f_1 - f_2}{2} \right], \quad \beta = \frac{g_1 + g_2}{2} + \varphi(\cot\theta) \frac{g_1 - g_2}{2},$$

with the functions f_i, g_i, $i = 1, 2$, depending on $(r\cos\theta, \rho)$. See Figures 2 and 3.

Clearly in a neighborhood of $0 \in \mathbb{R}^\ell$ we have two singular perturbation problems, SP1 and SP2, defined on the sets $r \geq 0$, $\theta \in (0, \pi)$ for $\rho \in \mathcal{S}_{12}$, and for $\rho \in \mathcal{S}_{34}$, respectively. Moreover, another two singular perturbation problems, SP3 and SP4, defined on the sets $u \geq 0$, $\psi \in (0, \pi)$ for $\eta \in \mathcal{S}_{14}$, and for $\eta \in \mathcal{S}_{23}$, respectively, still occur. Our main result is the following.

Theorem 2.2. *Consider $X = (X_1, \ldots, X_4) \in \Omega^\kappa_{1234}(\mathcal{U})$. There exist a neighborhood $\mathcal{V} \subseteq \mathcal{U}$ of 0 and the following singular perturbation problems defined at $0 \in \mathbb{R}^\ell$:*

$$\theta' = \alpha_i(r, \theta, x_2, \rho), \quad {x_2}' = r\beta_i(r, \theta, x_2, \rho), \quad \rho' = r\sigma_i(r, \theta, x_2, \rho), \quad i = 1, 2; \tag{2.3}$$

$$\psi' = \gamma_i(u, \psi, x_1, \rho), \quad {x_1}' = u\delta_i(u, \psi, x_1, \rho), \quad \rho' = u\nu_i(u, \psi, x_1, \rho), \quad i = 1, 2; \tag{2.4}$$

and

$$u\psi' = r\zeta(r, u, \theta, \psi, \rho), \quad \theta' = \xi(r, u, \theta, \psi, \rho), \quad \rho' = r\phi(r, u, \theta, \psi, \rho) \tag{2.5}$$

with $r, u \geq 0$, $\theta, \psi \in [0, \pi]$, $(x_1, x_2, \rho) \in \mathcal{V}$, $\rho = (x_3, \ldots, x_\ell)$. *Moreover,* α_i, β_i, σ_i, γ_i, δ_i, ν_i, ζ, ξ, *and* ϕ *are of class* C^κ *for* $i = 1, 2$ *and satisfying the following:*

(a) *The sliding region* $(\mathcal{S}^{sl} \bigcup \mathcal{S}^{es}) \cap \mathcal{S}_1$ *is homeomorphic to the slow manifold* $\{\gamma_i(0, \psi, x_1, \rho) = 0\} \setminus \mathcal{Z}_\psi = \{\psi = 0, \pi\}$ *of* (2.4), *with* $i = 1$ *at* $\widehat{\mathcal{S}_{14}}$ *and with* $i = 2$ *at* $\widehat{\mathcal{S}_{23}}$. *The sliding vector field* $X^{\mathcal{S}}$ *is topologically equivalent to the reduced problem* $\gamma_i(0, \psi, x_1, \rho) = 0$, $\dot{x_1} = \delta_i(0, \psi, x_1, \rho)$, *and* $\dot{\rho} = \nu_i(0, \psi, x_1, \rho)$.

(b) *The sliding region* $(\mathcal{S}^{sl} \bigcup \mathcal{S}^{es}) \cap \mathcal{S}_2$ *is homeomorphic to the slow manifold* $\{\alpha_i(0, \theta, x_2, \rho) = 0\} \setminus \mathcal{Z}_\theta = \{\theta = 0, \pi\}$ *of* (2.3), *with* $i = 1$ *at* $\widehat{\mathcal{S}_{12}}$ *and with* $i = 2$ *at* $\widehat{\mathcal{S}_{34}}$. *The sliding vector field* $X^{\mathcal{S}}$ *is topologically equivalent to the reduced problem* $\alpha_i(0, \theta, x_2, \rho) = 0$, $\dot{x_2} = \beta_i(0, \theta, x_2, \rho)$, *and* $\dot{\rho} = \sigma_i(0, \theta, x_2, \rho)$.

(c) *The singular perturbation* (2.5) *is the blowing up of the regularization of the systems* (2.3) *for* $i = 1, 2$. *The slow manifold is given by* $SM^1 = \{\zeta(0, 0, \theta, \psi, \rho) = 0\}$. *Furthermore, for* $\ell \geq 3$, *the slow flow is the limit, for* $r, u \downarrow 0$, *of the trajectories of another singular perturbation expressed by*

$$r\dot{\theta} = \xi(r, u, \theta, \psi, \rho), \quad \dot{\rho} = \phi(r, u, \theta, \psi, \rho). \tag{2.6}$$

The slow manifold of (2.6) *is the set on* $\mathbb{R}^\ell$ *given by*

$$SM^2 = \{\xi(0, 0, \theta, \psi, \rho) = 0, \zeta(0, 0, \theta, \psi, \rho) = 0\} \subseteq SM^1.$$

In section 3 we focus on the concept of regularization discussed above, and we prove Theorem 2.2. The sets given by $\alpha_i(0, \theta, x_2, \rho) = 0$, $\gamma_i(0, \psi, x_1, \rho) = 0$, and $\zeta(0, 0, \theta, \psi, \rho) = 0$ are called *slow manifolds*, and they will be denoted by SM. We say that a slow manifold is *nontrivial* if it is not contained in $\mathcal{Z} = \{\theta = 0, \pi\} \cup \{\psi = 0, \pi\}$.

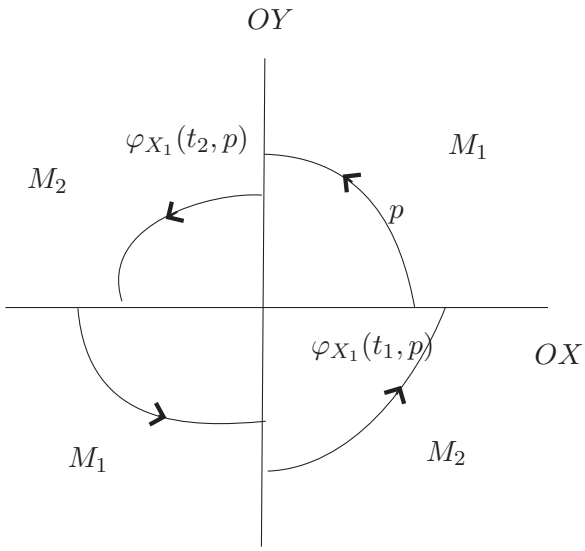


Figure 4. *Transient vector fields around* 0.

Let $X = (X_1, X_2) \in \Omega^k_{1234}(U)$ be a vector field belonging to the class $\Omega^k_{12}(U)$ of discontinuous vector fields satisfying $X_3 = X_1$ and $X_4 = X_2$. In section 5 we provide sufficient

conditions for the local asymptotic stability. Finally, in section 5 we study systems in this class that present a transient behavior. See section 5 for a precise definition. For an illustration see Figure 4.

3. Proof of Theorem 2.2. First we introduce some basic definitions. Consider a transition function φ.

(a) The φ_{x_2}*-regularization* of X is the one-parameter family given by

$$X_\varepsilon^{12} = (1/2)\left[(1+\varphi(x_1/\varepsilon))\,X_1 + (1-\varphi(x_1/\varepsilon))\,X_2\right] \tag{3.1}$$

for $x_2 > 0$, $\varepsilon > 0$, and

$$X_\varepsilon^{43} = (1/2)\left[(1+\varphi(x_1/\varepsilon))\,X_4 + (1-\varphi(x_1/\varepsilon))\,X_3\right] \tag{3.2}$$

for $x_2 < 0$, $\varepsilon > 0$.

(b) The φ_{x_1}*-regularization* of X is the one-parameter family given by

$$X_a^{14} = (1/2)\left[(1+\varphi(x_2/a))\,X_1 + (1-\varphi(x_2/a))\,X_4\right] \tag{3.3}$$

for $x_1 > 0$, $a > 0$, and

$$X_a^{23} = (1/2)\left[(1+\varphi(x_2/a))\,X_2 + (1-\varphi(x_2/a))\,X_3\right] \tag{3.4}$$

for $x_1 < 0$, $a > 0$.

Denote

$$\vartheta_{ij}^{+} = \frac{\vartheta_i + \vartheta_j}{2}, \quad \vartheta_{ij}^{-} = \frac{\vartheta_i - \vartheta_j}{2}, \quad \lambda(z) = \varphi(\cot(z)), \quad S_\alpha = -\sin\alpha,$$

and

$$\vartheta_{abcd}^{**\circ} = \frac{\vartheta_a * \vartheta_b \star \vartheta_c \circ \vartheta_d}{4},$$

where $*, \star, \circ \in \{+,-\}$.

Let $\varphi : \mathbb{R} \longrightarrow \mathbb{R}$ be a transition function. We remark that the mapping $\lambda : [0,\pi] \longrightarrow \mathbb{R}$ given by $\lambda(\theta) = \varphi(\cot\theta)$ for $0 < \theta < \pi$, $\lambda(0) = 1$, and $\lambda(\pi) = -1$ is a nonincreasing smooth function which is strictly decreasing for $\frac{\pi}{4} < \theta < \frac{3\pi}{4}$.

Proof of Theorem 2.2. For simplicity we suppose that $\ell = 3$, and we write $X_i = (f_i, g_i, h_i)$, $i = 1,\dots,4$. Denote by X_r^1, X_r^2 the equations of (3.1) and (3.2), respectively, after the coordinate change $\varepsilon = r\sin\theta$, $x_1 = r\cos\theta$. Let X_1^r and X_2^r be these equations after the time rescaling $t = r\tau$. Let X_r^3, X_r^4 be the systems given by (3.3) and (3.4) now written in the coordinates $a = u\sin\psi$, $x_2 = u\cos\psi$. Represent by X_3^r, X_4^r those systems after the time rescaling $t = u\tau$. Thus X_i^r, $i = 1,2$, are the vector fields corresponding to the differential systems (2.3), given in Theorem 2.2, on $\widehat{\mathcal{S}_{12}}$ and on $\widehat{\mathcal{S}_{34}}$, respectively. Analogously, X_i^r, $i = 3,4$, are the vector fields corresponding to the differential systems (2.4), given in Theorem 2.2, on $\widehat{\mathcal{S}_{14}}$ and on $\widehat{\mathcal{S}_{23}}$, respectively. See the equations in Table 1 and a diagram with the local vector fields in the blowing-up locus in Figure 5.

Table 1
The equations of the regularizations.

$X_r^1 \begin{cases} r\dot{\theta} = S_\theta\left(f_{12}^+ + \lambda(\theta) f_{12}^-\right) \\ \dot{x_2} = g_{12}^+ + \lambda(\theta) g_{12}^- \\ \dot{x_3} = h_{12}^+ + \lambda(\theta) h_{12}^- \end{cases}$	$X_r^2 \begin{cases} r\dot{\theta} = S_\theta\left(f_{43}^+ + \lambda(\theta) f_{43}^-\right) \\ \dot{x_2} = g_{43}^+ + \lambda(\theta) g_{43}^- \\ \dot{x_3} = h_{43}^+ + \lambda(\theta) h_{43}^- \end{cases}$
$X_r^3 \begin{cases} u\dot{\psi} = S_\psi\left(g_{14}^+ + \lambda(\psi) g_{14}^-\right) \\ \dot{x_1} = f_{14}^+ + \lambda(\psi) f_{14}^- \\ \dot{x_3} = h_{14}^+ + \lambda(\psi) h_{14}^- \end{cases}$	$X_r^4 \begin{cases} u\dot{\psi} = S_\psi\left(g_{23}^+ + \lambda(\psi) g_{23}^-\right) \\ \dot{x_1} = f_{23}^+ + \lambda(\psi) f_{23}^- \\ \dot{x_3} = h_{23}^+ + \lambda(\psi) h_{23}^- \end{cases}$
$X_1^r \begin{cases} \theta' = S_\theta\left(f_{12}^+ + \lambda(\theta) f_{12}^-\right) \\ x_2' = r\left(g_{12}^+ + \lambda(\theta) g_{12}^-\right) \\ x_3' = r\left(h_{12}^+ + \lambda(\theta) h_{12}^-\right) \end{cases}$	$X_2^r \begin{cases} \theta' = S_\theta\left(f_{43}^+ + \lambda(\theta) f_{43}^-\right) \\ x_2' = r\left(g_{43}^+ + \lambda(\theta) g_{43}^-\right) \\ x_3' = r\left(h_{43}^+ + \lambda(\theta) h_{43}^-\right) \end{cases}$
$X_3^r \begin{cases} \psi' = S_\psi\left(g_{14}^+ + \lambda(\psi) g_{14}^-\right) \\ x_1' = u\left(f_{14}^+ + \lambda(\psi) f_{14}^-\right) \\ x_3' = u\left(h_{14}^+ + \lambda(\psi) h_{14}^-\right) \end{cases}$	$X_4^r \begin{cases} \psi' = S_\psi\left(g_{23}^+ + \lambda(\psi) g_{23}^-\right) \\ x_1' = u\left(f_{23}^+ + \lambda(\psi) f_{23}^-\right) \\ x_3' = u\left(h_{23}^+ + \lambda(\psi) h_{23}^-\right) \end{cases}$

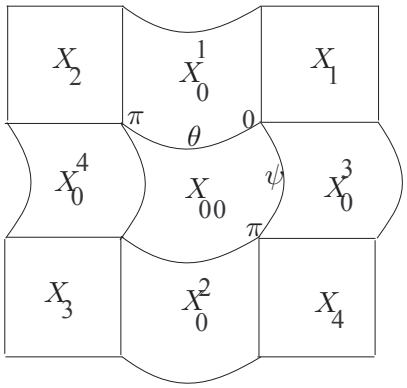


Figure 5. *Local vector fields in the blowing-up locus.*

Since the proofs of the cases (a) and (b) are similar, we will prove only (b) for $x_2 > 0$. The slow manifold for $\theta \in (0, \pi)$ is implicitly determined by the equation $f_{12}^+ + \lambda(\theta) f_{12}^- = 0$. We have that $f_{12}^- = 0$ if and only if $f_1(0, x_2, x_3) = f_2(0, x_2, x_3)$.

We denote $f(x_1, x_2, x_3)$ and $f(r, \theta, x_2, x_3)$ before and after the blow-up, respectively. If $(0, \theta, x_2, x_3) \in (\mathcal{S}^{sl} \bigcup \mathcal{S}^{es})$, then the vector fields point inward or outward. This implies that

$$f_1(0, \theta, x_2, x_3).f_2(0, \theta, x_2, x_3) < 0.$$

Thus we have that $f_{12}^-(0, \theta, x_2, x_3) \neq 0$ for any $\theta \in (0, \pi)$, $(0, x_2, x_3) \in (\mathcal{S}^{sl} \bigcup \mathcal{S}^{es})$. Since

$$\frac{f_{12}^+}{f_{12}^-} = \pm 1 \Rightarrow f_1.f_2 = 0,$$

we have

$$-1 < -\frac{f_{12}^+(0, \theta, x_2, x_3)}{f_{12}^-(0, \theta, x_2, x_3)} < 1$$

for all $(0,\theta,x_2,x_3)\in(\mathcal{S}^{sl}\bigcup\mathcal{S}^{es})$. Moreover, the inverse of the restriction $\lambda|_{(\frac{\pi}{4},\frac{3\pi}{4})}$ is increasing on $(-1,1)$, and the equation $f_{12}^{+}+\lambda(\theta)f_{12}^{-}=0$ defines a continuous graphic contained in

$$\left\{(\theta,x_2,x_3)\mid x_2\in\mathbb{R},\, x_3\in\mathbb{R},\, \theta\in\left(\frac{\pi}{4},\frac{3\pi}{4}\right)\right\}.$$

In accordance with the definition of $X^{\mathcal{S}}$, we have that $X^{\mathcal{S}}=X_1+k(X_2-X_1)$ with $k\in\mathbb{R}$ such that $X_1(x_1,x_2,x_3)+k(X_2-X_1)(x_1,x_2,x_3)=(0,y_2,y_3)$ for some y_2,y_3. Thus it is easy to see that $X^{\mathcal{S}}$ is given by $X^{\mathcal{S}}=\left(0,\frac{f_1g_2-f_2g_1}{f_1-f_2},\frac{f_1h_2-f_2h_1}{f_1-f_2}\right)$. The reduced problem is then represented by $\dot{x}_2=g_{12}^{+}+\lambda(\theta)g_{12}^{-}$, $\dot{x}_3=h_{12}^{+}+\lambda(\theta)h_{12}^{-}$ under the restriction $\lambda(\theta)=-\frac{f_{12}^{+}}{f_{12}^{-}}$. Then we must have $\dot{x}_2=\frac{f_1g_2-f_2g_1}{f_1-f_2}$, $\dot{x}_3=\frac{f_1h_2-f_2h_1}{f_1-f_2}$. It follows immediately that the flows of $X^{\mathcal{S}}$ and the reduced problem are equivalent.

To prove case (c) we consider the φ_{x_1}-regularization of X_1^r and X_2^r:

$$X_a^{1^r,2^r}\left\{\begin{array}{rcl}\theta' &=& S_\theta\left(f_{1234}^{+++}+\lambda(\theta)f_{1243}^{-+-}+\varphi(x_2/a)f_{1234}^{+--}+\varphi(x_2/a)\lambda(\theta)f_{1243}^{--+}\right),\\ x_2' &=& r\left(g_{1234}^{+++}+\lambda(\theta)g_{1243}^{-+-}+\varphi(x_2/a)g_{1234}^{+--}+\varphi(x_2/a)\lambda(\theta)g_{1243}^{--+}\right),\\ x_3' &=& r\left(h_{1234}^{+++}+\lambda(\theta)h_{1243}^{-+-}+\varphi(x_2/a)h_{1234}^{+--}+\varphi(x_2/a)\lambda(\theta)h_{1243}^{--+}\right).\end{array}\right. \tag{3.5}$$

The equations of the equivalent system with $a=u\sin\psi$, $x_2=u\cos\psi$ are

$$\left\{\begin{array}{rcl}\theta' &=& S_\theta\left(f_{1234}^{+++}+\lambda(\theta)f_{1243}^{-+-}+\lambda(\psi)f_{1234}^{+--}+\lambda(\psi)\lambda(\theta)f_{1243}^{--+}\right),\\ u\psi' &=& rS_\psi\left(g_{1234}^{+++}+\lambda(\theta)g_{1243}^{-+-}+\lambda(\psi)g_{1234}^{+--}+\lambda(\psi)\lambda(\theta)g_{1243}^{--+}\right),\\ x_3' &=& r\left(h_{1234}^{+++}+\lambda(\theta)h_{1243}^{-+-}+\lambda(\psi)h_{1234}^{+--}+\lambda(\psi)\lambda(\theta)h_{1243}^{--+}\right).\end{array}\right. \tag{3.6}$$

By means of the time rescaling $\tau=us$ we get (keeping the notation θ',ψ',x_3' for simplicity)

$$\left\{\begin{array}{rcl}\theta' &=& uS_\theta\left(f_{1234}^{+++}+\lambda(\theta)f_{1243}^{-+-}+\lambda(\psi)f_{1234}^{+--}+\lambda(\psi)\lambda(\theta)f_{1243}^{--+}\right),\\ \psi' &=& rS_\psi\left(g_{1234}^{+++}+\lambda(\theta)g_{1243}^{-+-}+\lambda(\psi)g_{1234}^{+--}+\lambda(\psi)\lambda(\theta)g_{1243}^{--+}\right),\\ x_3' &=& r\left(h_{1234}^{+++}+\lambda(\theta)h_{1243}^{-+-}+\lambda(\psi)h_{1234}^{+--}+\lambda(\psi)\lambda(\theta)h_{1243}^{--+}\right).\end{array}\right. \tag{3.7}$$

Thus we have

$$\xi(r,u,\theta,\psi,x_3)=S_\theta\left(f_{1234}^{+++}+\lambda(\theta)f_{1243}^{-+-}+\lambda(\psi)f_{1234}^{+--}+\lambda(\psi)\lambda(\theta)f_{1243}^{--+}\right),$$

$$\zeta(r,u,\theta,\psi,x_3)=S_\psi\left(g_{1234}^{+++}+\lambda(\theta)g_{1243}^{-+-}+\lambda(\psi)g_{1234}^{+--}+\lambda(\psi)\lambda(\theta)g_{1243}^{--+}\right),$$

and

$$\phi(r,u,\theta,\psi,x_3)=h_{1234}^{+++}+\lambda(\theta)h_{1243}^{-+-}+\lambda(\psi)h_{1234}^{+--}+\lambda(\psi)\lambda(\theta)h_{1243}^{--+}.$$

We define the slow system ($u=0$ at (3.6)) and the fast system (divide by r and $u=0$ at (3.7)) at $(x_1,x_2,x_3)=(0,0,x_3)$, respectively, by

$$X_{00}^{S}\left\{\begin{array}{rcl}0 &=& S_\psi\left(g_{1234}^{+++}+\lambda(\theta)g_{1243}^{-+-}+\lambda(\psi)g_{1234}^{+--}+\lambda(\psi)\lambda(\theta)g_{1243}^{--+}\right),\\ \theta' &=& S_\theta\left(f_{1234}^{+++}+\lambda(\theta)f_{1243}^{-+-}+\lambda(\psi)f_{1234}^{+--}+\lambda(\psi)\lambda(\theta)f_{1243}^{--+}\right),\\ x_3' &=& 0\end{array}\right. \tag{3.8}$$

and

$$X_{00}^F \left\{ \begin{array}{rcl} \psi' &=& S_\psi \left(g_{1234}^{+++} + \lambda(\theta) g_{1243}^{-+-} + \lambda(\psi) g_{1234}^{+--} + \lambda(\psi)\lambda(\theta) g_{1243}^{--+} \right), \\ \theta' &=& 0, \\ x_3' &=& 0. \end{array} \right. \tag{3.9}$$

This finishes the proof. ∎

Next the results of the theorem are illustrated.

Example 1. Consider $X_1(x_1, x_2) = (-1, -1)$, $X_2(x_1, x_2) = (2, -1)$, $X_3(x_1, x_2) = (4, 5)$, and $X_4(x_1, x_2) = (-2, 3)$. The trajectories of X_0^1 are the solutions of the reduced problem $S_\theta \left(\frac{1}{2} - \frac{3}{2}\lambda(\theta)\right) = 0$, $\dot{x_2} = -1$, where $S_\theta = -\sin\theta$. The slow manifold is the curve $\theta = \theta_0$ with $\lambda(\theta_0) = 1/3$, and the slow flow points to region $x_2 < 0$. The fast vector field is $(\theta', 0)$ with $\theta' < 0$ for $\theta > \theta_0$ and $\theta' > 0$ for $\theta < \theta_0$. The trajectories of X_0^2 are the solutions of the reduced problem $S_\theta (1 - 3\lambda(\theta)) = 0$, $\dot{x_2} = 4 - \lambda(\theta)$. The slow manifold is the curve $\theta = \theta_0$ with $\lambda(\theta_0) = 1/3$, and since $4 - \lambda(\theta) > 0$ for $\theta \in (0, \pi)$, the slow flow points to region $x_2 > 0$. The fast vector field is $(\theta', 0)$ with $\theta' < 0$ for $\theta > \theta_0$ and $\theta' > 0$ for $\theta < \theta_0$. The trajectories of X_0^3 are the solutions of the reduced problem $S_\psi (1 - 2\lambda(\psi)) = 0$, $\dot{x_1} = -\frac{3}{2} + \frac{1}{2}\lambda(\psi)$. The slow manifold is the curve $\psi = \psi_0$ with $\lambda(\psi_0) = 1/2$, and since $\dot{x_1} < 0$ for $\psi \in (0, \pi)$, the slow flow points to region $x_1 < 0$. The fast vector field is $(\psi', 0)$ with $\psi' > 0$ for $\psi < \psi_0$ and $\psi' < 0$ for $\psi > \psi_0$. The trajectories of X_0^4 are the solutions of the reduced problem $S_\psi (2 - 3\lambda(\psi)) = 0$, $\dot{x_1} = 3 - \lambda(\psi)$. The slow manifold is the curve $\psi = \psi_1$ with $\lambda(\psi_1) = 2/3$, and since $\dot{x_1} > 0$ for $\psi \in (0, \pi)$, the slow flow points to region $x_1 > 0$. The fast vector field is $(\psi', 0)$ with $\psi' > 0$ for $\psi < \psi_1$ and $\psi' < 0$ for $\psi > \psi_1$. The trajectories of X_{00} are the solutions of the reduced problem $S_\psi \left(\frac{3}{2} - \frac{1}{2}\lambda(\theta) - \frac{5}{2}\lambda(\psi) + \frac{1}{2}\lambda(\psi)\lambda(\theta)\right) = 0$, $\theta' = S_\theta \left(\frac{3}{4} - \frac{9}{2}\lambda(\theta) - \frac{1}{4}\lambda(\psi) + \frac{3}{4}\lambda(\psi)\lambda(\theta)\right)$. The slow manifold is the curve $\frac{3}{2} - \frac{1}{2}\lambda(\theta) - \frac{5}{2}\lambda(\psi) + \frac{1}{2}\lambda(\psi)\lambda(\theta) = 0$, which connects the points $(\theta, \psi) = (0, \psi_0)$ and $(\theta, \psi) = (\pi, \psi_1)$ with $\lambda(\psi_0) = 1/2$ and $\lambda(\psi_1) = 2/3$. In fact, if $\theta = 0$, then $\lambda(\theta) = 1$, and $\frac{3}{2} - \frac{1}{2}\lambda(\theta) - \frac{5}{2}\lambda(\psi) + \frac{1}{2}\lambda(\psi)\lambda(\theta) = 0$ implies $\lambda(\psi) = \frac{1}{2}$, and if $\theta = \pi$, then $\lambda(\theta) = -1$, and $\frac{3}{2} - \frac{1}{2}\lambda(\theta) - \frac{5}{2}\lambda(\psi) + \frac{1}{2}\lambda(\psi)\lambda(\theta) = 0$ implies $\lambda(\psi) = \frac{2}{3}$. Observe that $\frac{3}{2} - \frac{1}{2}\lambda(\theta) - \frac{5}{2}\lambda(\psi) + \frac{1}{2}\lambda(\psi)\lambda(\theta) = 0$ implies $\theta' = \frac{S_\theta}{4(\lambda(\theta)-5)}(-15\lambda^2(\theta) + 83\lambda(\theta) - 12)$. There exists a unique $\alpha = \frac{83}{30} - \frac{1}{30}\sqrt{6169} \in (-1, 1)$ such that $-15\alpha^2 + 83\alpha - 12 = 0$. Furthermore, if $\lambda < \alpha$, then $-15\lambda^2 + 83\lambda - 12 < 0$, and if $\lambda > \alpha$, then $-15\lambda^2 + 83\lambda - 12 > 0$. So there exists a unique $\theta_* \in (\frac{\pi}{4}, \frac{3\pi}{4})$ such that $\lambda(\theta_*) = \alpha$ and

$$0 < \theta < \theta_* \Rightarrow \lambda(\theta) > \lambda(\theta_*) \Rightarrow \theta' > 0$$

and

$$\theta > \theta_* \Rightarrow \lambda(\theta) < \lambda(\theta_*) \Rightarrow \theta' < 0.$$

Moreover, $\lambda(\theta_0) = 1/3$ and $\theta' > 0$ imply that $\theta_0 < \theta_*$. See Figure 6.

Example 2. Consider $X_1(x, y) = (-y, x)$ and $X_2(x, y) = (1, -1)$. The trajectories of X_0^1 are the solutions of the reduced problem

$$\frac{-y+1}{2} + \lambda(\theta)\frac{-y-1}{2} = 0, \quad \dot{y} = -\frac{1}{2}(1 - \lambda(\theta)).$$

The slow manifold is the curve $y = y(\theta)$ given by $y = \frac{1-\lambda(\theta)}{1+\lambda(\theta)}$. This function is increasing:

$$\lim_{\theta \longrightarrow \frac{\pi}{4}} y(\theta) = 0, \quad \lim_{\theta \longrightarrow \frac{3\pi}{4}} y(\theta) = +\infty.$$

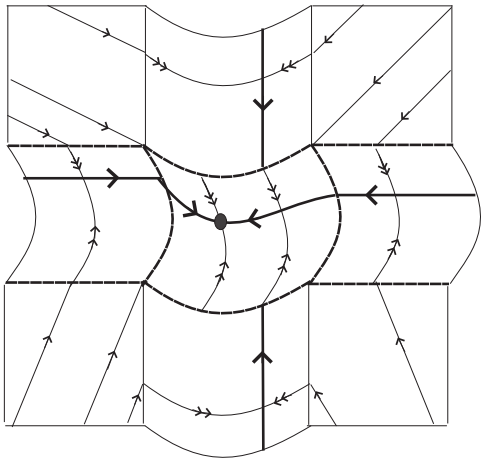


Figure 6. *Local vector fields in the blowing-up locus of Example 1.*

Since $1-\lambda(\theta)\geq 0$, we have that $\dot{y}\leq 0$. So the slow flow points to region $y<0$. Denote $\theta(y)$ the inverse function of $y(\theta)$. The fast vector field is $(\theta',0)$ with $\theta'<0$ for $\theta(y)<\theta<\pi$ and $\theta'>0$ for $0<\theta<\theta(y)$. The trajectories of X_0^2 are the solutions of the reduced problem

$$\frac{-y+1}{2}+\lambda(\theta)\frac{y+1}{2}=0,\quad \dot{y}=-\frac{1}{2}\left(1+\lambda(\theta)\right).$$

The slow manifold is the empty set. In fact, the right-hand side of $y=\frac{1+\lambda(\theta)}{-\lambda(\theta)+1}$ is positive for $\theta\in(\frac{\pi}{4},\frac{3\pi}{4})$, and here we have $y<0$. The fast vector field is $(\theta',0)$ with $\theta'<0$ for $0<\theta<\pi$.

The trajectories of X_0^3 are the solutions of the reduced problem

$$\frac{x-1}{2}+\lambda(\psi)\frac{x+1}{2}=0,\quad \dot{x}=\frac{1}{2}\left(1-\lambda(\psi)\right).$$

The slow manifold is the curve $x=x(\psi)$ given by $x=\frac{1-\lambda(\psi)}{1+\lambda(\psi)}$. This function is increasing:

$$\lim_{\psi\longrightarrow\frac{\pi}{4}}x(\psi)=0,\qquad \lim_{\psi\longrightarrow\frac{3\pi}{4}}x(\psi)=+\infty.$$

We have $\dot{x}\geq 0$. So the slow flow points to region $x>0$. Denote by $\psi(x)$ the inverse function of $x(\psi)$. The fast vector field is $(\psi',0)$ with $\psi'<0$ for $0<\psi<\psi(x)$ and $\psi'>0$ for $\psi(x)<\psi<\pi$. The trajectories of X_0^4 are the solutions of the reduced problem

$$\frac{x-1}{2}+\lambda(\psi)\frac{-x-1}{2}=0,\quad \dot{x}=\frac{1}{2}\left(1+\lambda(\psi)\right).$$

The slow manifold is the empty set. In fact, the right-hand side of $x=\frac{1+\lambda(\psi)}{-\lambda(\psi)+1}$ is positive, and here we have $x<0$. The fast vector field is $(\psi',0)$ with $\psi'<0$ for $0<\psi<\pi$. The trajectories of X_{00} are the solutions of the reduced problem

$$\frac{-1}{2}+\lambda(\psi)\lambda(\theta)\frac{1}{2}=0,\quad \theta'=\frac{S_\theta}{2}\left(1-\lambda(\psi)\lambda(\theta)\right).$$

The slow manifold is given by the equation $\lambda(\psi)\lambda(\theta) = 1$, and the slow flow is $\theta' = 0$, which is highly degenerated. The fast vector field is $(\psi', 0)$ with $\psi' \leq 0$ for $0 < \psi < \pi$. Note that $\lambda(\psi)\lambda(\theta) = 1$ for all $(\theta, \psi) \in ([0, \pi/4] \times [0, \pi/4]) \cup ([3\pi/4, \pi] \times [3\pi/4, \pi])$. See Figure 7.

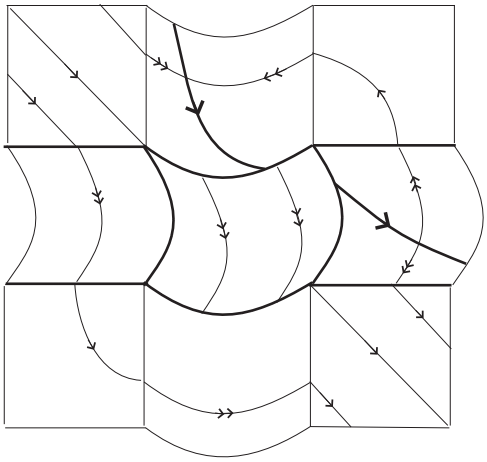


Figure 7. *Local vector fields in the blowing-up locus of Example 2.*

4. Intersecting switching surfaces on $\mathbb{R}^3$. In this section we consider the case $\ell = 3$. We provide sufficient conditions in order for an equilibrium point of the reduced problem (2.6) to be a local attractor. See previous results in this direction in [12, 13].

In what follows we use the notation

$$\frac{\partial(\Phi, \Psi)}{\partial(a, b)} = \begin{vmatrix} \Phi_{aa} & \Psi_{ab} \\ \Phi_{ba} & \Psi_{bb} \end{vmatrix}.$$

Proposition 4.1. *Consider $\mathcal{U} \subseteq \mathbb{R}^3$, an open set with $0 \in \mathcal{U}$ and $X = (X_1, \ldots, X_4) \in \Omega^{\kappa}_{1234}(\mathcal{U})$. Denote $X_i = (f_i, g_i, h_i)$, $i = 1, \ldots, 4$. Let $\overline{\zeta}$, $\overline{\psi}$ be the functions which satisfy $\zeta = S_\theta \overline{\zeta}$ and $\xi = S_\psi \overline{\xi}$ with ζ, ξ, and ϕ given by (2.5). Let $D \subseteq (0, \pi) \times (0, \pi) \times \mathbb{R}$ be an open neighborhood of $(\theta_0, \psi_0, 0)$. Suppose that $\overline{\zeta}(0, 0, \theta_0, \psi_0, 0) = \overline{\xi}(0, 0, \theta_0, \psi_0, 0) = 0$ and that for $(\theta, \psi, x_3) \in D$ we have*

$$\frac{\partial(\overline{\zeta}, \overline{\xi})}{\partial(\theta, \psi)} > 0, \quad \frac{\partial(\overline{\zeta}, \overline{\xi})}{\partial(\theta, x_3)} < 0, \quad \frac{\partial(\overline{\zeta}, \overline{\xi})}{\partial(\psi, x_3)} < 0.$$

Then the slow manifold SM_2 is a curve parameterized by $(\theta(x_3), \psi(x_3), x_3)$. Moreover, if

$$\phi(\theta_0, \psi_0, 0) = 0, \quad \frac{\partial}{\partial x_3}\phi(\theta_0, \psi_0, 0) < 0,$$

then $(\theta_0, \psi_0, 0)$ is an attracting singular point of the reduced system (2.6).

Proof. It is enough to apply the implicit function theorem. The assertions about the singular point follow the fact that the reduced problem is $x_3' = \alpha x_3 + \cdots$ with $\alpha < 0$. ∎

Example. Consider $X = (X_1, \dots, X_4) \in \Omega^{\kappa}_{1234}(\mathcal{U})$ given by $X_1 = (-1, -1, -x_3)$, $X_2 = (2, -3, -x_3)$, $X_3 = (1, 2, -x_3)$, and $X_4 = (-3, 1, -x_3)$. System (2.5) takes the form

$$\begin{aligned} u\psi' &= rS_\psi(-\tfrac{1}{4} + \tfrac{1}{4}\lambda(\theta) - \tfrac{7}{4}\lambda(\psi) + \tfrac{3}{4}\lambda(\psi)\lambda(\theta)), \\ \theta' &= S_\theta(-\tfrac{1}{4} - \tfrac{7}{4}\lambda(\theta) + \tfrac{3}{4}\lambda(\psi) + \tfrac{1}{4}\lambda(\theta)\lambda(\psi)), \\ x_3' &= -rx_3. \end{aligned}$$

The slow manifold SM^1 is given by $-1 + \lambda(\theta) - 7\lambda(\psi) + 3\lambda(\theta)\lambda(\psi) = 0$. This is a surface parameterized by $(\theta, \psi(\theta), x_3)$ with $\theta \in [0, \pi]$. Moreover, the curve $(\theta, \psi(\theta))$ is such that it connects $(\theta, \psi) = (0, \pi/2)$ to (π, ψ_2) with $\lambda(\psi_2) = -1/5$. The fast vector field is such that $\psi' > 0$ on the upper side and $\psi' < 0$ otherwise. The slow manifold SM_2 is represented by

$$\begin{cases} -1 + \lambda(\theta) - 7\lambda(\psi) + 3\lambda(\theta)\lambda(\psi) &= 0, \\ -1 - 7\lambda(\theta) + 3\lambda(\psi) + \lambda(\theta)\lambda(\psi) &= 0. \end{cases}$$

This curve is parameterized by (θ_0, ψ_0, x_3), $x_3 \in \mathbb{R}$, with $\lambda(\psi_0) = \frac{11-4\sqrt{11}}{2}$ and $\lambda(\theta_0) = \frac{3-\sqrt{11}}{2}$. Since $x_3' = -x_3$, the slow flow on SM_2 has an attracting singular point at $(\theta_0, \psi_0, 0)$. See Figure 8.

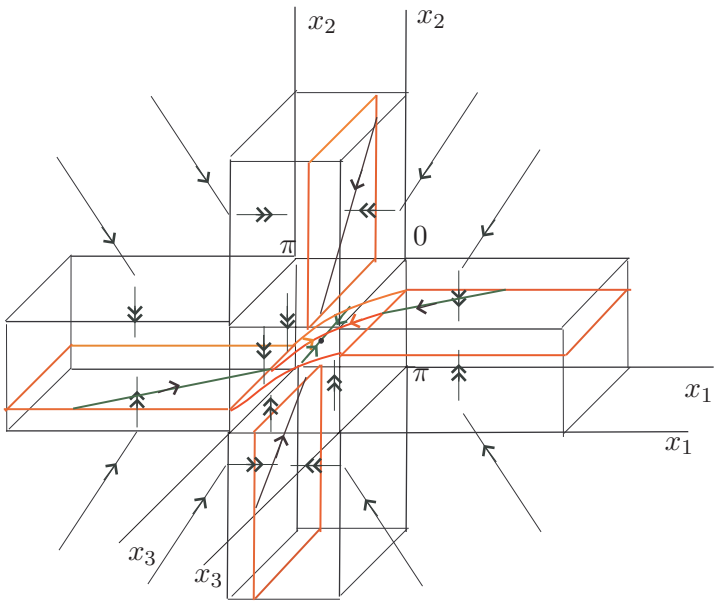


Figure 8. *A sequence of slow manifolds $SM_2 \subseteq SM_1$.*

5. Transient vector fields on $\mathbb{R}^2$. Now we give special attention to the class of discontinuous vector fields that present a transient behavior around the singularity. The following study program can be considered. Let $F : U \subseteq \mathbb{R}^2, 0 \to \mathbb{R}, 0$ be a generic Morse C^∞-function. This means that we may take a coordinate system (x_1, x_2) around $(0, 0)$ such that it takes either the form (a) $F(x_1, x_2) = x_2$ or the form (b) $F(x, y) = x^2 + \varepsilon y^2$ with $\varepsilon = \pm 1$. We denote by $\mathcal{C}^\kappa(U, \mathbb{R}^2)$ the set of all vector fields of class $\mathcal{C}^\kappa$ defined on U, with $\kappa \geqslant 1$, endowed with

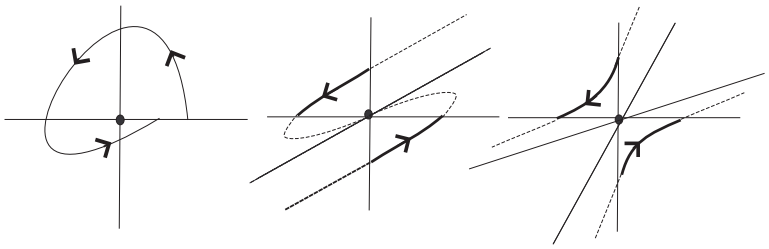


Figure 9. *T-type singular points.*

the C^κ-topology. Consider $X_1, X_2 \in \mathcal{C}^\kappa(U, \mathbb{R}^2)$. The idea is to study the following nonsmooth system in U:

$$X(q) = \begin{cases} X_1(q) = (f_1(q), g_1(q)) & \text{if} \quad F(q) > 0, \\ X_2(q) = (f_2(q), g_2(q)) & \text{if} \quad F(q) < 0. \end{cases} \tag{5.1}$$

The case (a) was treated in [11]. The case (b) can be seen as a particular case of the system studied above when $\varepsilon = -1$. In this case we will denote $X = (X_1, X_2) \in \Omega_{12}^\kappa(U)$. The case when $\varepsilon = 1$ does not deserve to be analyzed. We choose local coordinates such that $M = F^{-1}(0) = \{xy = 0\}$. Thus $M_1 = \{x.y > 0\}$ and $M_2 = \{x.y < 0\}$. In what follows we denote by $\mathcal{O}_{X_i}(p) = \{\varphi_{X_i}(t, p),\ t \in \mathbb{R}\}$ the orbit of the vector field X^i, $i = 1, 2$, through the point p.

Definition 5.1. *We say that $X = (X_1, X_2) \in \Omega_{12}^\kappa(U)$, $U \subseteq \mathbb{R}^2$, is a transient vector field around 0 if, for all $p = (x, y) \in M_i$ for some $i = 1, 2$, there exist $t_1, t_2 \in \mathbb{R}$ such that $\mathcal{O}_{X_i}(p)$ ($i = 1$ if $xy > 0$ or $i = 2$ if $xy < 0$) cuts $0X \setminus \{0\}$ and $0Y \setminus \{0\}$ transversally for $t = t_1$ and $t = t_2$, respectively. Moreover, $t_1 t_2 < 0$, and for any t between t_1 and t_2 we have that $\varphi_{X_i}(t, p) \in M_i$.*

Definition 5.2. *Let $A \in M(2)$ be a 2×2 real matrix. We say that $(0, 0)$ is a T-type singular point of A if the Jordan form of A, $J(A)$, is one of the following.*

- *(focus-type) $J(A) = \left[\begin{smallmatrix} \alpha & -\beta \\ \beta & \alpha \end{smallmatrix}\right]$, with $\alpha \cdot \beta \neq 0$;*
- *(improper node–type) $J(A) = \left[\begin{smallmatrix} \alpha & 1 \\ 0 & \alpha \end{smallmatrix}\right]$, with $\alpha \neq 0$ and with the eigenvector $\omega \in M_i$;*
- *(saddle- or node-type) $J(A) = \left[\begin{smallmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{smallmatrix}\right]$, with $\lambda_1 \neq \lambda_2 \in \mathbb{R}$ and with both eigenvectors in the same M_i.*

See Figure 9.

Proposition 5.3. *Let $X = (X_1, X_2) \in \Omega_{12}^\kappa(U)$, $U \subseteq \mathbb{R}^2$, be a transient vector field. If $(0, 0)$ is a hyperbolic singular point of X_i, $i = 1, 2$, then $(0, 0)$ is a singular point of the linear vector field $DX_i(0, 0)$ of T type.*

Proof. We have that $(0, 0)$ is a saddle or a node or a focus. The transient behavior of X_i implies that we cannot have orbits on $M_i \cup \{(0, 0)\}$ with α-limit or ω-limit equal to $\{(0, 0)\}$. Moreover, if $(0, 0)$ is a saddle or a node, then the stable and unstable manifolds are on $M_i \cup \{(0, 0)\}$. ∎

Now we introduce some notation. $Q1 = \{(x, y);\ x > 0,\ y > 0\}$.

- We say that X_i for $i = 1, 2$ is of F_{ij} type in a neighborhood of $(0, 0)$ if $(0, 0)$ is a focus or an improper node of X_i. Moreover, the trajectories are positively oriented if $j = 1$ and negatively oriented if $j = 2$.

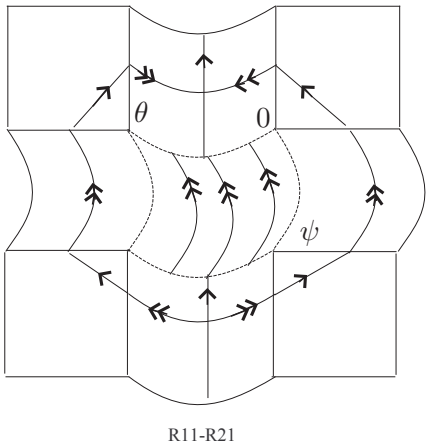


Figure 10. *Transient vector field of* (R_{11}, R_{21}) *type.*

- We say that X_i for $i = 1, 2$ is of S_{ij} type in a neighborhood of $(0, 0)$ if $(0, 0)$ is a saddle or a node of X_i and the trajectories on M_i are positively oriented if $j = 1$ and negatively oriented if $j = 2$.
- We say that X_1 is of R_{1j} type in a neighborhood of $(0, 0)$ if $(0, 0)$ is a regular point of X_1 and $X_1(0, 0) \in Q2$ if $j = 1$ and $X_1(0, 0) \in Q4$ if $j = 2$.
- We say that X_2 is of R_{2j} type in a neighborhood of $(0, 0)$ if $(0, 0)$ is a regular point of X_2 and $X_2(0, 0) \in Q1$ if $j = 1$ and $X_1(0, 0) \in Q3$ if $j = 2$.
- We say that (X_1, X_2) is of (A, B) type if X_1 is of A type and X_2 is of B type.

Remark. We have 36 possible combinations for X_1 and X_2. However, we can reduce this number, observing that some cases can be identified by means of a rotation or a change of sign. For example, roughly speaking, (F_{12}, F_{21}) can be identified with $-(F_{11}, F_{22})$, and (S_{11}, F_{21}) can be identified with $R_{\pi/2}(F_{11}, S_{21})$. Thus we have the following:

(a) If $X_1(0, 0) \in M_2$, $X_2(0, 0) \in M_1$, then, by means of a rotation, $X = (X_1, X_2)$ is such that its phase portrait is of (R_{11}, R_{21}) type. See Figure 10.

(b) If $|X_1(0, 0)|.|X_2(0, 0)| = 0$ and $|X_1(0, 0)|^2 + |X_2(0, 0)|^2 \neq 0$, then $X = (X_1, X_2)$ is such that its phase portrait is of (F_{11}, R_{21}), (F_{11}, R_{22}), (S_{11}, R_{21}), or (S_{11}, R_{22}) type. See Figures 11 and 12.

(c) If $|X_1(0, 0)|^2 + |X_2(0, 0)|^2 = 0$, then $X = (X_1, X_2)$ is such that its phase portrait is of (F_{11}, F_{21}), (F_{11}, F_{22}), (F_{11}, S_{21}), (F_{11}, S_{22}), (S_{11}, S_{21}), or (S_{11}, S_{22}) type. See Figures 13, 14, and 15.

Example. Let $X = (X_1, X_2) \in \Omega^k_{12}(U)$ be a transient vector field defined on a neighborhood of $0 \in U \subseteq \mathbb{R}^2$. Suppose that $X_1(0, 0) = (a, b)$, $ab < 0$, and $X_2(0, 0) = (c, d)$, with $cd > 0$. The slow manifold on the blowing-up locus corresponding to $x_1 > 0$, $x_2 = 0$ is $S_\psi(\frac{b+d}{2} + \frac{b-d}{2}\lambda(\psi)) = 0$ and on the blowing-up locus corresponding to $x_1 < 0$, $x_2 = 0$ is $S_\psi(\frac{b+d}{2} + \frac{d-b}{2}\lambda(\psi)) = 0$. If we consider $bd < 0$, we have that $-\frac{b+d}{b-d} \in (-1, 1)$, which implies that there exist $\psi_1, \psi_2 \in (\frac{\pi}{4}, \frac{3\pi}{4})$ such that the slow manifolds are $\psi = \psi_1$ and $\psi = \psi_2$, respectively. Finally, the equation of the slow manifold in the intersection is $S_\psi(\frac{b+d}{2} + \frac{b-d}{2}\lambda(\theta)\lambda(\psi)) = 0$. We have that this manifold connects the points $(\theta, \psi) = (0, \psi_1)$ and $(\theta, \psi) = (\pi, \psi_2)$.

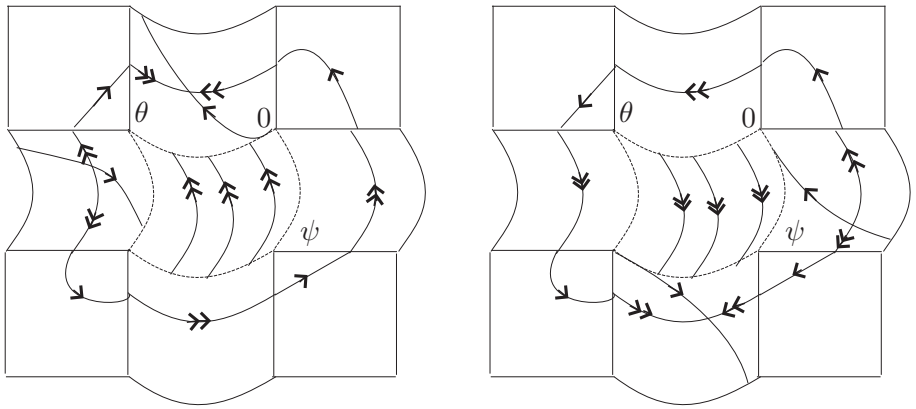


Figure 11. *Transient vector fields of* (F_{11}, R_{21}) *type and* (F_{11}, R_{22}) *type.*

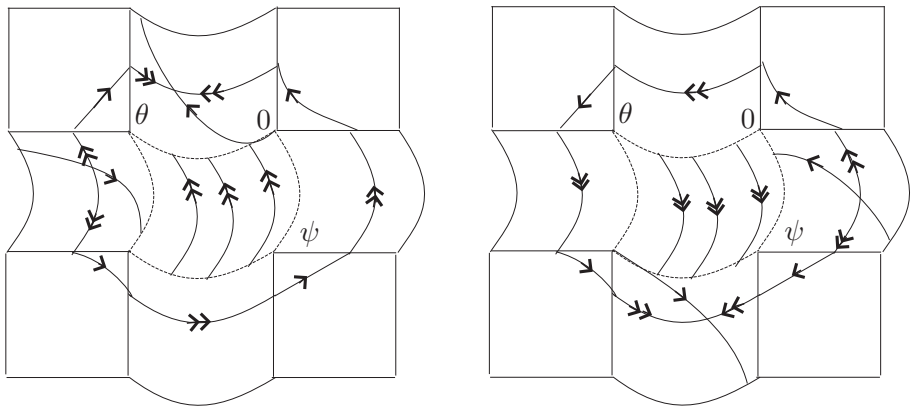


Figure 12. *Transient vector fields of* (S_{11}, R_{21}) *type and* (S_{11}, R_{22}) *type.*

Proposition 5.4. *Let* $X = (X_1, X_2) \in \Omega_{12}^k(U)$ *be a transient vector field defined on a neighborhood of* $0 \in U \subseteq \mathbb{R}^2$. *Denote*

$$A_i = \begin{bmatrix} a_i & -b_i \\ b_i & a_i \end{bmatrix}, \quad i = 1, 2,$$

$$B_\lambda = \begin{bmatrix} \sin(\lambda_1 + \lambda_2) & -2\cos\lambda_1 \cos\lambda_2 \\ 2\sin\lambda_1 \sin\lambda_2 & -\sin(\lambda_1 + \lambda_2) \end{bmatrix},$$

and

$$C_\lambda = \begin{bmatrix} -2\sin\lambda_1 \sin\lambda_2 & \sin(\lambda_1 + \lambda_2) \\ \sin(\lambda_1 + \lambda_2) & -2\cos\lambda_1 \cos\lambda_2 \end{bmatrix}.$$

(a) *If* $X_1(0,0) = X_2(0,0) = (0,0)$, $DX_1(0,0) = A_1$, *and* $DX_2(0,0) = A_2$, *then for* $b_1 > 0$, $b_2 > 0$ *the slow manifold is nontrivial and the reduced flow of system* (2.5) *is singular;*

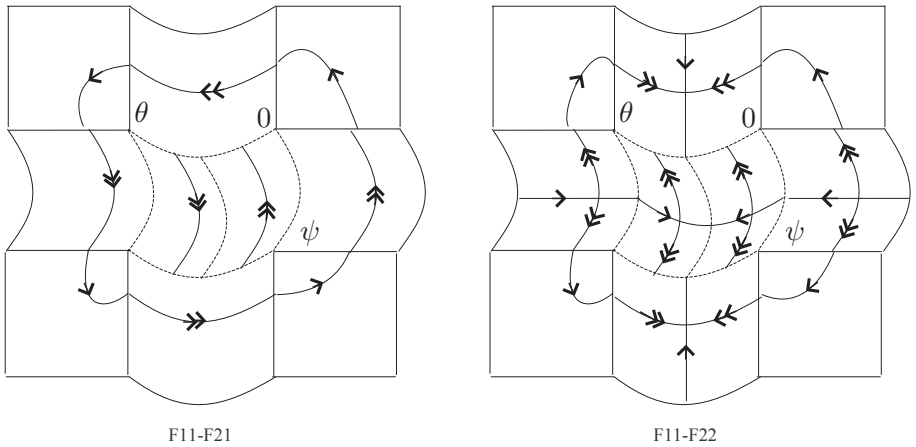


Figure 13. *Transient vector fields of* (F_{11}, F_{21}) *type and* (F_{11}, F_{22}) *type.*

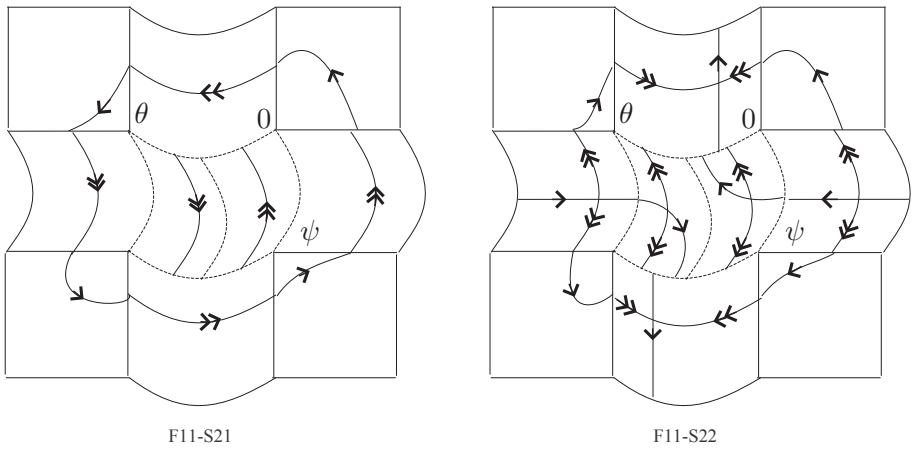


Figure 14. *Transient vector fields of* (F_{11}, S_{21}) *type and* (F_{11}, S_{22}) *type.*

and for $b_1 > 0$, $b_2 < 0$, *and* $a_1 \neq a_2$ *the slow manifold is nontrivial and the slow flow has a singular point.*

(b) *If* $X_1(0,0) = X_2(0,0) = (0,0)$, $DX_1(0,0) = A_1$, *and* $b_1 > 0$, $0 < \lambda_1 < \lambda_2 < \pi/2$, *then for* $DX_2(0,0) = B_\lambda$ *the slow manifold is nontrivial and the reduced flow of system* (2.5) *is singular; and for* $DX_2(0,0) = -B_\lambda$ *the slow manifold is nontrivial and the slow flow has not only singular points. Moreover, if* $2a_1 \sin(\lambda_1)\sin(\lambda_2) > b_1 \sin(\lambda_1 + \lambda_2)$, *then the slow flow has a singular point.*

(c) *If* $X_1(0,0) = X_2(0,0) = (0,0)$, $DX_1(0,0) = C_\lambda$, $0 < \lambda_1 < \lambda_2 < \pi/2$, *and* $0 < \mu_1 < \mu_2 < \pi/2$, *then for* $DX_2(0,0) = B_\mu$ *the slow manifold is nontrivial and the reduced flow of system* (2.5) *is singular; and for* $DX_2(0,0) = -B_\mu$ *the slow manifold is nontrivial and the slow flow has a part composed only of singular points and another part where there exists a singular point.*

We remark that A_i is the matrix of a linear vector field which has a singular point of

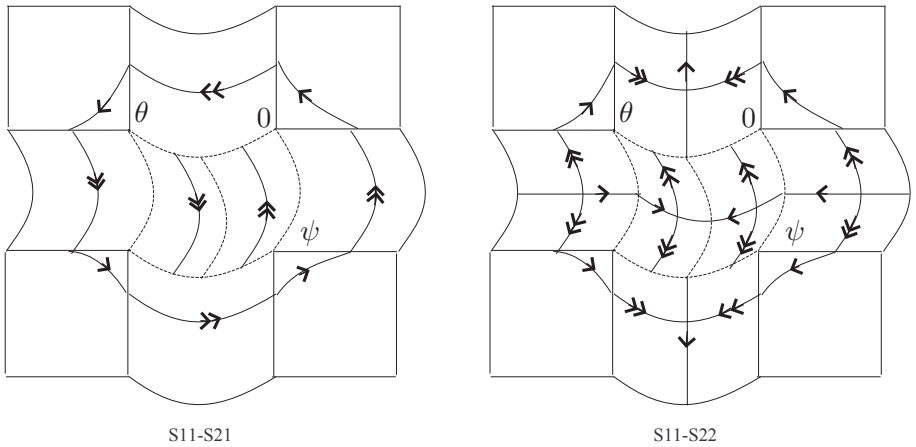


Figure 15. *Transient vector fields of* (S_{11}, S_{21}) *type and* (S_{11}, S_{22}) *type.*

focus type at $(0,0)$, and B_λ and C_λ are the matrices of linear vector fields which have singular points of saddle type at $(0,0)$ with eigenvectors $(\cos\lambda_1, \sin\lambda_1)$ and $(\cos\lambda_2, \sin\lambda_2)$ in the first and second quadrants, respectively.

Proof of Proposition 5.4. Suppose that X_1 and X_2 satisfy the hypothesis of (a). The equation of the slow manifold is $\frac{S_\psi}{2}\cos(\theta)((b_1+b_2)+(b_1-b_2)\lambda(\theta)\lambda(\psi))=0$. If $b_1>0$ and $b_2>0$, then $|\frac{b_1+b_2}{b_1-b_2}|>1$. This implies that the nontrivial part of the slow manifold is given by $\theta=\pi/2$. Moreover, the slow flow is determined by θ', which is zero for $\theta=\pi/2$. If $b_1>0$ and $b_2<0$, then $|\frac{b_1+b_2}{b_1-b_2}|<1$, and thus $(b_1+b_2)+(b_1-b_2)\lambda(\theta)\lambda(\psi)=0$ defines two smooth curves. The slow flow has a singular point because $\theta'=S_\theta\cos(\theta)(a_2-a_1)$ has a sign change at $\theta=\frac{\pi}{2}$. So (a) is proved.

Suppose now that both X_1 and X_2 satisfy the hypothesis of (b). The equation of the slow manifold is $\frac{S_\psi}{2}\cos(\theta)[(b_1+2S_{\lambda_1}S_{\lambda_2})+(b_1-2S_{\lambda_1}S_{\lambda_2})\lambda(\theta)\lambda(\psi)]=0$. If $b_1>0$ and $0<\lambda_1<\lambda_2<\pi/2$, then $S_{\lambda_1}S_{\lambda_2}>0$. This implies that the nontrivial part of the slow manifold is given by $\theta=\pi/2$. Moreover, the slow flow is determined by θ', which is zero for $\theta=\pi/2$. If $DX_2(0,0)=-B_\lambda$, then the slow manifold is $\frac{S_\psi}{2}\cos(\theta)[(b_1-2S_{\lambda_1}S_{\lambda_2})+(b_1+2S_{\lambda_1}S_{\lambda_2})\lambda(\theta)\lambda(\psi)]=0$. Since $\frac{b_1-2S_{\lambda_1}S_{\lambda_2}}{b_1+2S_{\lambda_1}S_{\lambda_2}}\in(-1,1)$, the equation defines at least two curves. The slow flow has a singular point because $\theta'=S_\theta\cos(\theta)\left(\frac{2a_1S_{\lambda_1}S_{\lambda_2}+b_1S_{\lambda_1+\lambda_2}}{b_1+2S_{\lambda_1}S_{\lambda_2}}\right)$ has a sign change at $\theta=\frac{\pi}{2}$. So (b) is proved.

Suppose now that X_1 and X_2 satisfy the hypothesis of (c). The equation of the slow manifold is $\frac{S_\psi}{2}\cos(\theta)[(-S_{\lambda_1+\lambda_2}+2S_{\mu_1}S_{\mu_2})+(-S_{\lambda_1+\lambda_2}-2S_{\mu_1}S_{\mu_2})\lambda(\theta)\lambda(\psi)]=0$. If $0<\lambda_1<\lambda_2<\pi/2$ and $0<\mu_1<\mu_2<\pi/2$, then $-S_{\lambda_1+\lambda_2}>0$ and $S_{\mu_1}S_{\mu_2}>0$. This implies that the nontrivial part of the slow manifold is given by $\theta=\pi/2$. Moreover, the slow flow is determined by θ', which is zero for $\theta=\pi/2$. If $DX_2(0,0)=-B_\mu$, then the slow manifold is $\frac{S_\psi}{2}\cos(\theta)[(-S_{\lambda_1+\lambda_2}-2S_{\mu_1}S_{\mu_2})+(-S_{\lambda_1+\lambda_2}+2S_{\mu_1}S_{\mu_2})\lambda(\theta)\lambda(\psi)]=0$. Since $\frac{-S_{\lambda_1+\lambda_2}-2S_{\mu_1}S_{\mu_2}}{-S_{\lambda_1+\lambda_2}+2S_{\mu_1}S_{\mu_2}}\in(-1,1)$, the equation defines at least two curves. The slow flow has a singular point because $\theta'=S_\theta\cos(\theta)\left(\frac{-8S_{\lambda_1}S_{\lambda_2}S_{\mu_1}S_{\mu_2}-2S_{\lambda_1+\lambda_2}S_{\mu_1+\mu_2}}{-S_{\lambda_1+\lambda_2}+2S_{\mu_1}S_{\mu_2}}\right)$ has a sign change at $\theta=\frac{\pi}{2}$. In fact, observe

that if

$$\delta = \frac{-8S_{\lambda_1}S_{\lambda_2}S_{\mu_1}S_{\mu_2} - 2S_{\lambda_1+\lambda_2}S_{\mu_1+\mu_2}}{-S_{\lambda_1+\lambda_2} + 2S_{\mu_1}S_{\mu_2}} \Rightarrow sgn(\delta) = -1,$$

then (c) is proved. ■

6. Conclusions. In this paper we propose a method of studying the dynamics around the intersection of codimension one discontinuity submanifolds. Using a regularization process proceeded by a blow-up, we get a singular perturbation problem. We also consider questions like transient behavior and asymptotic stability.

Acknowledgment. The authors express their gratitude to the referees for their helpful comments.

REFERENCES

[1] J. C. Alexander and T. I. Seidman, *Sliding modes in intersecting switching surfaces.* I. *Blending*, Houston J. Math., 24 (1998), pp. 545–569.
[2] J. C. Alexander and T. I. Seidman, *Sliding modes in intersecting switching surfaces.* II. *Hysteresis*, Houston J. Math., 25 (1999), pp. 185–211.
[3] M. Broucke, C. Pugh, and S. Simić, *Structural stability of piecewise smooth systems*, Comput. Appl. Math., 20 (2001), pp. 51–89.
[4] C. Buzzi, P. R. Silva, and M. A. Teixeira, *A singular approach to discontinuous vector fields on the plane*, J. Differential Equations, 231 (2006), pp. 633–655.
[5] M. di Bernardo, C. Budd, A. Champneys, and P. Kowalczyk, *Piecewise-Smooth Dynamical Systems. Theory and Applications*, Appl. Math. Sci. 163, Springer-Verlag, London, 2008.
[6] F. Dumortier and R. Roussarie, *Canard cycles and center manifolds*, Mem. Amer. Math. Soc., 121 (1996), 132708.
[7] N. Fenichel, *Geometric singular perturbation theory for ordinary differential equations*, J. Differential Equations, 31 (1979), pp. 53–98.
[8] A. F. Filippov, *Differential Equations with Discontinuous Righthand Sides*, Math. Appl. (Soviet Ser.), Kluwer Academic Publishers, Dordrecht, The Netherlands, 1988.
[9] C. Jones, *Geometric singular perturbation theory*, in Dynamical Systems, C.I.M.E. Lectures (Montecatini Terme, 1994), Lecture Notes in Math. 1609, Springer-Verlag, Heidelberg, 1995, pp. 44–118.
[10] J. Llibre, P. R. Silva, and M. A. Teixeira, *Regularization of discontinuous vector fields via singular perturbation*, J. Dynam. Differential Equations, 19 (2007), pp. 309–331.
[11] J. Llibre, P. R. Silva, and M. A. Teixeira, *Sliding vector fields via slow–fast systems*, Bull. Belg. Math. Soc. Simon Stevin, 15 (2008), pp. 851–869.
[12] J. Llibre and M. A. Teixeira, *Global asymptotic stability for a class of discontinuous vector fields in* $\mathbb{R}^2$, Dyn. Syst., 22 (2007), 133–146.
[13] M. A. Teixeira, *Stability conditions for discontinuous vector fields*, J. Differential Equations, 88 (1990), pp. 15–29.
[14] M. A. Teixeira, *Perturbation theory for non-smooth dynamical systems*, in Encyclopedia of Complexity and Systems Science, G. Gaeta, ed., Springer-Verlag, New York, to appear.