ISRAEL JOURNAL OF MATHEMATICS 149 (2005), 345–360

SOME REMARKS ON FILTERING AND PREDICTION OF STATIONARY PROCESSES

BY

BENJAMIN WEISS

Institute of Mathematics, The Hebrew University of Jerusalem
Givat Ram, Jerusalem 91904, Israel
e-mail: weiss@math.huji.ac.il

ABSTRACT

This note contains a some remarks concerning filtering and prediction theory. One of them is a solution to an old question of H. Furstenberg which indicates an unexpected phenomenon arising from the lack of integrability. Another gives some general results on the possibility of constructing two valued universal guessing schemes for distinguishing between classes of stochastic processes.

1. Introduction

The Ph.D. Thesis of Hillel Furstenberg was entitled *Stationary Processes and Prediction Theory* [F-1960] and contained many seminal ideas which were elaborated by him during the course of many years of research in probability theory and dynamical systems. His fundamental paper [F-1967] introduced more new ideas that enabled him to prove some striking general results in filtering, which is of course closely related to prediction theory. My purpose in this work is to present several remarks on these theories, mainly with the purpose of motivating the reader to peruse for himself these marvellous works.

The first remark will be an answer to one of the problems that Hillel posed in [F-1967]. That problem is a purely probabilistic question and asks the following:

1.1. PROBLEM: *Given the random variables U_1, U_2, V_1, V_2 such that U_1 has the same distribution as U_2 and V_1 has the same distribution as V_2, is it true that*

$$U_1 + V_1 \leq U_2 + V_2 \tag{1.1}$$

Received January 13, 2004 and in revised form September 19, 2004

implies

$$U_1 + V_1 = U_2 + V_2? \tag{1.2}$$

In §2, I will explain the motivation behind the question and then present a rather simple example showing that in general the answer is negative.

The second remark will summarize an exchange between Hillel and me last year when he taught a course which included his filtering result. This goes roughly as follows. Suppose we are given a stationary real valued stochastic process $\{X_n\}$ which represents the signal, and a second process $\{Y_n\}$ which represents some kind of additive noise. We observe their sum $Z_n = X_n + Y_n$. Under a basic assumption about the **disjointness** (see below) of the signal and noise processes and the assumption of integrability of these processes, he showed that one can almost surely filter out the noise, in the sense that the signal process is a function of the observed process $\{Z_n\}$. During his lecture on the filtering theorem, the question was raised as to the possibility of carrying this out in a unilateral way, i.e. can one determine X_0 from the knowledge of $\{Z_n : n \leq 0\}$ alone? Some discussions with Hillel resulted in a positive answer, which I will present in §3.

The third remark concerns more recent questions motivated by prediction theory, where one asks what can one say about an unknown stationary process $\{X_n\}$ after observing the first n outputs. Some questions can be answered in an "online" fashion. For example: what is the entropy of the process? For this, there are universal functions $g_n(X_1, X_2, \ldots, X_n)$ which have the property that for any finite valued stationary stochastic ergodic process, g_n converges almost surely to the entropy of the process (cf. [B-1976], [Z-1978], [OW-1990], [KASW-1998]).

Another example: is the process $\{X_n\}$ a discrete valued iid process or not? Here too, there are functions f_n which will almost surely stabilize on the correct answer (since there are only two values, convergence here means eventual agreement). Note that in the second question we are asking about a property that is not an isomorphism invariant. The isomorphism class of the iid processes are the Bernoulli processes and in [OW-1990] we showed that there is no such guessing scheme which will eventually stabilize and tell us whether a process is Bernoulli or not. I will present here some general results of this type, both positive and negative.

The positive result, given in §4, concerns a family $\mathcal{F}$ of ergodic stationary stochastic processes that is σ-compact and closed in the relative topology on ergodic processes with a fixed number of states (induced from the topology of convergence in finite distributions). For an example of this kind of family, take all ergodic Markov chains with a fixed bound on the number of states, or the processes that are obtained by collapsing states of such Markov chains. I will construct a sequence of two valued functions $g_n(X_1, X_2, \ldots, X_n) \in \{YES, NO\}$, which will converge almost surely to **YES** for any process in $\mathcal{F}$ and to **NO** for any process not in $\mathcal{F}$. For Markov chains such results are known (see, for example, [CS-2000] and the references there).

One of the negative results may be formulated as follows. Let $\mathcal{P}_0, \mathcal{P}_1$ be two distinct isomorphism classes of ergodic processes with zero entropy, and suppose that g_n are functions which will converge in probability to i for any process in $\mathcal{P}_i$. Then there is some ergodic process for which the g_n will fail to converge in probability. In particular, it is impossible to decide whether or not -1 is in the spectrum. This question was raised by Eli Glasner, and motivated me to prove this general result. This and another related result are in §5. I would like to thank Hillel Furstenberg for allowing me to present the results of our discussions.

Added during revision.

- I am indebted to an anonymous referee for pointing out to me a recent preprint of Andrew Nobel entitled *Hypothesis testing for families of ergodic processes* which has results that overlap §4.
- Right after the Ornsteinfest, held at Stanford University on the occasion of the 70th birthday and retirement of Donald S. Ornstein, he and I realized that the results of §5 could be extended to positive entropy and would lead to a remarkable characterization of the entropy as essentially the only **Finitely Observable**, isomorphism invariant of ergodic processes. A joint paper is being written with the details of this surprising result.

2. Disjointness and filtering

Since my first two remarks both refer to Hillel's basic filtering theorem, let me begin by describing this precisely. The basic definition is the following:

2.1. Definition: Two stationary stochastic processes $\{X_n\}$ and $\{Y_n\}$ are said to be disjoint if in any stationary coupling $\{(X_n, Y_n)\}$ of these processes they are independent.

With this definition Hillel proved the following:

2.2. THEOREM: *If $\{X_n\}$ and $\{Y_n\}$ are real valued integrable processes that are disjoint, then both are functions of the sum process $\{X_n + Y_n\}$.*

He also posed some problems which were designed to enable him to remove the assumption of integrability, which seemed to be merely a technical one. It is not hard to see how the first one relates to the filtering question:

2.3. PROBLEM: *Given the random variables X_1, X_2, Y_1, Y_2 such that X_1 has the same distribution as X_2 and Y_1 has the same distribution as Y_2 and, in addition, assume for $i, j = 1, 2$ that X_i is independent of Y_j, is it true that*

$$X_1 + Y_1 = X_2 + Y_2 \tag{2.1}$$

implies

$$X_1 = X_2 \quad \text{and} \quad Y_1 = Y_2? \tag{2.2}$$

In an attempt to answer this affirmatively, he formulated:

2.4. PROBLEM: *Given the random variables U_1, U_2, V_1, V_2 such that U_1 has the same distribution as U_2 and V_1 has the same distribution as V_2, is it true that*

$$U_1 + V_1 \leq U_2 + V_2 \tag{2.3}$$

implies

$$U_1 + V_1 = U_2 + V_2? \tag{2.4}$$

Let us see why an affirmative answer to Problem 2.4 implies an affirmative answer to Problem 2.3. Rewrite (2.1) as follows:

$$X_1 - X_2 = Y_2 - Y_1 \tag{2.5}$$

and, multiplying the two sides of the equation, we obtain

$$0 \leq X_1Y_2 - X_1Y_1 - X_2Y_2 + X_2Y_1 \tag{2.6}$$

or

$$X_1Y_1 + X_2Y_2 \leq X_1Y_2 + X_2Y_1. \tag{2.7}$$

Now by our assumptions about the X_i and Y_j, it follows that all variables that appear in the last equation have the same distribution, and so from a positive solution to Problem 2.4 we would conclude that equality holds, which gives the desired equality of the X_i and Y_j.

Furthermore, it is also easy to see that in case the variables are integrable, then Problem 2.4 has an affirmative answer. Indeed, taking expectations of both sides we would see that they are equal, and this would immediately imply that equality holds almost surely. Also, for independent variables, if the X_i and Y_j have finite integrals, then the same is true for the products X_iY_j, and it follows that for integrable variables Problem 2.3 also has an affirmative answer.

For an example in which (2.3) holds without (2.4), I will make use of the positive stable distributions of index p strictly less than 1. Recall that if U and V are independent identically distributed random variables with a stable distribution of index p, then $U + V$ has the same distribution as cU where the constant c is given by $(1^p + 1^p)^{1/p} = 2^{1/p}$. Now it is well known that for $0 < p < 1$ there are stable distributions that are concentrated on $[0, \infty)$. Their Laplace transform is given by $e^{-a\lambda^p}$. It is not very hard to verify that indeed this is the Laplace transform of a probability measure concentrated on $[0, \infty)$ (cf. [Feller-II] p. 448), while the stability property is evident from the form.

Take now for U_2 and V_2 independent identically distributed copies of a positive stable variable with index p. Then set $U_1 = V_1 = 2^{-1/p}(U_2 + V_2)$, and we see that (2.3) holds with a strict inequality everywhere. Note that since we can take p to be arbitrarily close to 1, no moment condition can replace integrability.

3. Unilateral filtering

Recall that in the filtering theorem in the previous section the signal process $\{X_n\}$ can be recovered with zero probability of error from the noisy version $Z_n = X_n + Y_n$ if the noise $\{Y_n\}$ is disjoint from the signal. The problem that we are now trying to resolve is: Can X_0 be recovered from $\{Z_n\}_{n \leq 0}$ alone? One cannot apply the proof that is in [F-1967] directly to this situation. The proof there is by contradiction. Assuming that the signal cannot be recovered perfectly leads to a coupling of $\{(X_n, Y_n)\}$ and a different version of this pair process $\{(\tilde{X}_n, \tilde{Y}_n)\}$ in which $X_n + Y_n = \tilde{X}_n + \tilde{Y}_n$. Using the disjointness and the integrability, one sees that in fact these two different versions must be the same, which yields the contradiction. To see if this proof works in the unilateral setting, we have to review how the coupling is achieved.

For this I switch to the language of systems rather than processes. Suppose $(X, \mathcal{X}, \mu, T)$ is an ergodic measure preserving transformation and π is a mapping to another such system $(Z, \mathcal{Z}, \eta, T)$ that takes μ to η and intertwines the actions, i.e. $T\pi = \pi T$. The Rokhlin theory of such factor spaces enables us to represent $(X, \mathcal{X}, \mu)$ as a product space with $(Z, \mathcal{Z}, \eta)$ as one factor and, say, $(Y, \mathcal{Y}, \nu)$ as the other, and the transformation T on X is expressed now as a skew product over $(Z, \mathcal{Z}, \eta, T)$ with $(Y, \mathcal{Y}, \nu)$ as fiber. For the relative independent coupling of X over Z we simply take the product measure space $(Y \times Y, \mathcal{Y} \times \mathcal{Y}, \nu \times \nu)$ as the new fiber over Z with the diagonal skew product.

To compare with the processes above, the space X corresponds to the pair process $\{(X_n, Y_n)\}$, Z corresponds to the $\{Z_n\}$ process and Y to the abstract fiber that would arise if perfect filtering weren't possible.

To carry this out unilaterally would require defining a relatively independent product in the setting where T is not invertible. Now the difficulty is that the canonical definition of a relatively independent product in general doesn't lead to a measure preserving system. For a simple example demonstrating this, consider the Markov chain $\{U_n\}$ on three states $\{0, 1, 2\}$ where each i moves to i or $i + 1 (\mathrm{mod}\ 3)$ with equal probabilities. We map this onto the 2-shift on the symbols $\{a, b\}$ by mapping pairs according to the rule that the pairs $\{00, 12, 20\}$ map to a and the pairs $\{01, 11, 22\}$ map to b. For the one-sided chain $\{U_n, n \geq 0\}$ this mapping is definitely two to one, and yet one cannot get from this a nontrivial coupling over the 2-shift since, as soon as the string aa occurs in the 2-shift, the two paths that map to it must merge.

In spite of this difficulty the unilateral result is valid. To see this we recall one of the basic results of [F-1967], namely that processes with positive entropy cannot be disjoint. It follows that if the two processes $\{X_n\}, \{Y_n\}$ are disjoint then at least one of them has zero entropy. Let us say that $\{X_n\}$ has zero entropy. According to the filtering theorem we know that $\{X_n\}$ is measurable with respect to the entire process $\{Z_n, n \in \mathbb{Z}\}$. Now we appeal to a classical result that basically goes back to Pinsker, namely that any zero entropy process which is measurable with respect to $\{Z_n, n \in \mathbb{Z}\}$ is also measurable with respect to the remote past of $\{Z_n, n \in \mathbb{Z}\}$. But the remote past of $\{Z_n, n \in \mathbb{Z}\}$ is clearly measurable with respect to $\{Z_n, n \leq 0\}$; this implies that we have the filtering result unilaterally.

For the sake of completeness I will sketch a proof of Pinsker's result, especially since we need it here for processes $\{Z_n\}$ that might not be finite valued and most of the proofs in the literature deal with finite valued processes. In the following I will use $\alpha, \beta, \gamma, \ldots$ to denote finite partitions of a probability space $(X, \mathcal{X}, \mu)$ and denote by ρ the Rokhlin metric on these partitions, given by $\rho(\alpha, \beta) = H(\alpha|\beta) + H(\beta|\alpha)$. With respect to this metric, for any measure preserving transformation T, the entropy $h(T,\alpha)$ is a Lipshitz function with constant one. Suppose now that we have a process, $\{Z_n\}$, that is defined over the probability space $(X,\mathcal{X},\mu)$ by means of a function f. This means that Z_n is given by $T^n f$ for all $n \in \mathbb{Z}$. If α is measurable with respect to the σ-algebra generated by the Z_n and $h(T,\alpha) = 0$, we want to show that α is measurable with respect to the remote past of the process $\{Z_n\}$.

Fix a small $\epsilon > 0$, and use the first assumption on α to find a finite partition β that is measurable with respect to the field $\mathcal{B}_N$ which is generated by the variables $\{Z_n, |n| \leq N\}$, and satisfies

$$\rho(\alpha,\beta) < \epsilon. \tag{3.1}$$

Since $h(T,\alpha) = 0$ one also has $h(T^{2N},\alpha) = 0$, and from the above that gives us $h(T^{2N},\beta) < \epsilon$. From the definition of h we have

$$H\left(\beta| \bigvee_{j=1}^{\infty} T^{-2jN}\beta\right) < \epsilon. \tag{3.2}$$

Recalling that β was measurable with respect to $\mathcal{B}_N$, this yields in turn

$$H\left(\beta| \bigvee_{j=1}^{\infty} T^{-2jN}\mathcal{B}_N\right) < \epsilon \tag{3.3}$$

or, using again the property of the Rokhlin metric,

$$H\left(\alpha| \bigvee_{j=1}^{\infty} T^{-2jN}\mathcal{B}_N\right) < 2\epsilon. \tag{3.4}$$

Denoting by $\mathcal{C}_{-N}$ the σ-algebra generated by the variables $\{Z_n, n \leq -N\}$, we deduce that

$$H(\alpha|\mathcal{C}_{-N}) < 2\epsilon. \tag{3.5}$$

By the martingale convergence theorem, the left-hand side converges as N tends to infinity to the conditional entropy of α with respect to the remote past of the $\{Z_n\}$ process, and since ϵ was arbitrary we conclude that indeed α is measurable with respect to this remote past. Thus the full Pinsker algebra, corresponding to the largest zero entropy factor, is measurable with respect to the remote past.

4. On classifying processes

One of the main classes of processes that are studied in detail in [F-1960] is the class of continuous functions of finite state Markov chains. These are very popular today in the mathematical biology literature under the name "Hidden Markov Models" (HMM). In [F-1960] one can find a very nice characterization of these processes as those which can be defined by a finite number of finite dimensional stochastic matrices. Essentially the same characterization was rediscovered several years later by A. Heller in [H-1965]. There has been much work in finding methods for finding the best HMM to fit some given data. In light of this it is natural to ask: Can one determine membership in this class or not by successive observations of $\{X_1, X_2, \ldots, X_n\}$? D. Bailey showed in his thesis [B-1976] that this is not even possible for the class of all k-step Markov chains (k arbitrary, fixed number of states). In [MW] we give a similar negative result for another extension of the class of all Markov chains — the finitary Markov processes.

On the other hand, if one restricts the order and the size of the state space, then there are guessing schemes g_n which will converge almost surely and test for membership; see, for example, [Ki-1994], [CS-2000]. (In these papers there are integer valued schemes which are shown to converge to the least k such that the process is a k-step Markov chain, and with an a priori bound on the value of k this can be used to produce a two-valued scheme which tests for membership in the class.) In [DP-1994], a more general problem of this nature is studied and some sufficient conditions are given for the possibility of discriminating by tests of this type between two families of processes.

I shall now explain how to devise such schemes for any family of ergodic processes with uniform rates in the ergodic theorem. Then I will show how to use a variant of this for the class of all ergodic HMM where there is an a priori bound on the number of states in the Markov chain. Note that since we are now dealing with functions of Markov chains, there is no loss of generality in restricting to classical 1-step chains and functions that depend on a single state.

Let $\mathcal{F}$ denote some family of ergodic stochastic processes on a fixed state space S with a finite number of symbols. We will identify these processes with the shift invariant measures on the compact space, S, of bi-infinite sequences of elements from S. On this space of measures we put the weak* topology to obtain a compact space. Convergence in this topology coincides exactly with convergence of all finite distributions. We will be concerned mainly with ergodic measures, since by the ergodic decomposition almost every sequence produced by any stationary process is a typical sequence for some ergodic process. On the ergodic processes we take the induced topology. Thus when we speak of a closed family of ergodic processes, we mean closed in this relative topology.

Our guessing functions will be based on properties of the empirical distribution of k-blocks in n-strings from S. Let us introduce the following notation for this empirical distribution. Let $b \in S^k$ be a fixed k-block and $u \in S^n$ an n-string; then define

$$\mathbf{D}(b|u) = |\{1 \leq i \leq n-k+1 : u[i,\ldots,i+k-1] = b\}|/(n-k+1). \tag{4.1}$$

4.1. Definition: A closed family of ergodic stochastic processes $\mathcal{F}$ has **uniform rates**, if for every $k \in \mathbb{N}$ and every $\epsilon > 0$ there is some $n = n(k,\epsilon)$ such that for all $P \in \mathcal{F}$ we have

$$P\{u \in S^n : |P(b) - \mathbf{D}(b|u)| < \epsilon, \text{ for all } b \in S^k\} > 1 - \epsilon. \tag{4.2}$$

With this definition, for any closed family with uniform rates, we will construct a guessing scheme with two values,{YES,NO}, which will almost surely stabilize on YES if the process belongs to $\mathcal{F}$ and to NO in the contrary case. To this end let $\mathcal{F}$ be a family with uniform rates, and fix a sequence ϵ_k such that

$$\sum_k \epsilon_k < \infty. \tag{4.3}$$

Let $n_k = n(k,\epsilon_k)$ be the sequence which the definition supplies for us, and define g_n as follows:

For n in the range $[n_k, n_{k+1} - 1]$, if for some $P \in \mathcal{F}$ we have that

$$|P(b) - \mathbf{D}(b|x_1, x_2, \ldots, x_{n_k})| < \epsilon_k \quad \text{for all } b \in S^k, \tag{4.4}$$

then set

$$g_n(x_1, x_2, \ldots x_n) = YES, \tag{4.5}$$

and if not set

$$g_n(x_1, x_2, \ldots, x_n) = NO. \tag{4.6}$$

With this definition we can prove the following proposition:

4.2. Proposition: *If the closed family of ergodic processes $\mathcal{F}$ has uniform rates and the g_n are defined by (3.4)–(3.6), then for almost every realization of a process P from the family $\mathcal{F}$ we have that eventually $g_n(x_1, x_2, \ldots, x_n) =$*

YES, while for almost every realization of an ergodic process that is not in $\mathcal{F}$ eventually $g_n(x_1, x_2, \ldots, x_n) =$ NO.

Proof: The first part of the theorem follows immediately by our assumption (4.3) and the Borel–Cantelli lemma. For the second part, arguing by contradiction, it suffices to show that if g_n returns YES infinitely often, then the process belongs to $\mathcal{F}$. As already remarked, we may assume that our observation is typical for some ergodic process. The fact that $g_n =$ YES along a sequence of n's tending to infinity means that this ergodic process is in the closure of $\mathcal{F}$, and since we have assumed that $\mathcal{F}$ is closed this means that the process indeed belongs to $\mathcal{F}$. ∎

The reader may well ask how one can find closed ergodic families with uniform rates. Perhaps the simplest method is given by the following:

4.3. PROPOSITION: *If $\mathcal{K}$ is a compact set of ergodic distributions then $\mathcal{K}$ has uniform rates.*

Proof: Suppose that $\epsilon > 0$ is given, and we are looking for an N such that for any process P in $\mathcal{K}$ we will have

$$P\{u \in S^N : |P(b) - \mathbf{D}(b|u)| < \epsilon, \text{ for all } b \in S\} > 1 - \epsilon. \tag{4.7}$$

First note that this condition can be expressed without the explicit use of $P(b)$, by noting that equation (4.7) would follow from the existence of a set $U \subset S^N$ with $P(U) > 1 - 1/10\epsilon$ so that for all $u, v \in U$ we have

$$|\mathbf{D}(b|u) - \mathbf{D}(b|v)| < 1/10\epsilon \quad \text{for all } b \in S. \tag{4.8}$$

Next, use the fact that every P in $\mathcal{K}$ is ergodic to find for each P an $N(P)$ and a U_P for which equation (4.8) holds for all $u, v \in U_P$. This is an open condition on the process measures and thus it holds for all Q in a neighborhood of P; and then by the compactness we can find a finite number of these neighborhoods that will cover all of $\mathcal{K}$. We still need to find a single N that will be good for all the P's.

For this, it suffices to remark that as soon as (4.7) holds for N, then it will automatically continue to hold for all multiples of N with ϵ replaced by $2\sqrt{\epsilon}$. Thus we can take for N the product of the finite number of $N(P)$'s to find N; naturally, to achieve our original goal we begin this procedure with a smaller ϵ.

The argument for k-blocks is identical, and this establishes the proposition. ∎

For example, all Markov processes defined by transition matrices of a fixed size and a uniform positive lower bound on their entries have uniform rates, since the set is clearly compact and consists of ergodic processes only. We can now formulate a theorem which is sufficiently general and whose assumptions are purely toplogical.

4.4. THEOREM: *If the family of ergodic processes $\mathcal{E}$ is closed (in the set of all ergodic processes) and is also σ-compact, then there are g_n such that for almost every realization of a process P from the family $\mathcal{E}$ we have that eventually $g_n(x_1, x_2, \dots, x_n) =$ YES, while for almost every realization of an ergodic process that is not in $\mathcal{E}$ eventually $g_n(x_1, x_2, \dots, x_n) =$ NO.*

Proof: Write $\mathcal{E} = \bigcup_{k=1}^{\infty} \mathcal{K}_k$, where the $\mathcal{K}_k$ are an increasing sequence of compact subsets of $\mathcal{E}$. By Proposition 4.3 each of these $\mathcal{K}_k$ has uniform rates, and thus if ϵ_k is a fixed convergent series of positive numbers, there are integers N_k that increase such that for all processes P in $\mathcal{K}_k$ we have

$$P\{u \in S^{N_k} : |P(b) - \mathbf{D}(b|u)| < \epsilon_k, \text{ for all } b \in S^k\} > 1 - \epsilon_k. \tag{4.9}$$

To define the g_n we proceed as before. For n in the range $[N_k, N_{k+1} - 1]$, if for some $P \in \mathcal{K}_k$ we have

$$|P(b) - \mathbf{D}(b|x_1, x_2, \dots x_{n_k})| < \epsilon_k \quad \text{for all } b \in S^k, \tag{1.1}$$

then set

$$g_n(x_1, x_2, \dots, x_n) = YES, \tag{4.11}$$

and if not set

$$g_n(x_1, x_2, \dots, x_n) = NO. \tag{4.12}$$

If a process P belongs to $\mathcal{E}$, then for some k_0 it belongs to $\mathcal{K}_{k_0}$. Applying the Borel–Cantelli lemma to the complements of the sets

$$\{u \in S^{N_k} : |P(b) - \mathbf{D}(b|u)| < \epsilon_k \text{ for all } b \in S^k\}$$

we see that eventually equation (4.10) will be satisfied and thus eventually g_n will be returning the answer YES.

On the other hand, if the process P is ergodic, then almost surely the sequence of outcomes that we observe will be typical for the process. Receiving a YES infinitely many times would imply that P is in the closure of $\mathcal{E}$, which is assumed to be relatively closed in the ergodic processes. Thus P would have to belong to $\mathcal{E}$. It follows that if P is ergodic and not in $\mathcal{E}$, then almost surely, eventually we must be receiving a NO answer. This completes the proof of the theorem. ∎

As examples of this theorem one can take all ergodic Markov processes with a fixed number of states. The σ-compactness can be seen by taking for the $\mathcal{K}_k$ all those ergodic Markov processes defined by transition matrices where, if an entry is non-zero, it is at least $1/k$. In a similar fashion one sees that all ergodic hidden Markov models with a fixed number of states and a bound on the window size of the function satisfy the hypotheses of the theorem.

5. Isomorphism classes cannot be distinguished

Let $\mathcal{P}_0, \mathcal{P}_1$ be two distinct isomorphism classes of ergodic processes. If the entropy of these processes is different, then using any one of the known universal entropy estimators (cf. [B-1976], [Z-1978], [OW-1993], [KASW-1998]) one can easily construct a sequence of two-valued functions g_n which will almost surely converge to i for any ergodic process in $\mathcal{P}_i$ for $i = 0, 1$. If they have the same entropy, and if we only want a test which will eventually serve to distinguish between the two given classes, then it is also possible to design such a test. For example, two irrational rotations by $\rho_0 \neq \rho_1$ can be distinguished by considering ergodic sums of the type

$$\left| 1/n \sum_{k=1}^{n} e^{2\pi i k \rho_j} X_k \right|$$

for $j = 0, 1$ and seeing which of the two tends to zero. However, if one wants a **universal** test which is applicable to all processes, then things change dramatically. For zero entropy systems, I will show that if there are functions g_n which will converge in probability to i for any process in $\mathcal{P}_i$, then there is some ergodic process for which the g_n will not converge in probability. In other words, one can never distinguish between zero entropy isomorphism classes of processes by universal two-valued functions.

For example, we can take for $\mathcal{P}_0$ the isomorphism class of any zero entropy weakly mixing process, and for $\mathcal{P}_1$ the isomorphism class of the same process product with the flip map on $\{-1, +1\}$. This latter process has -1 in the point spectrum. It follows that it is impossible to decide whether or not -1 is in the spectrum.

5.1. THEOREM: *Let $\mathcal{P}_0, \mathcal{P}_1$ be two distinct isomorphism classes of zero entropy ergodic processes with values in a fixed finite set S. Suppose g_n is a sequence of functions with domain S^n and range $\{0,1\}$ such that for any ergodic process $\{Z_n\}$ in $\mathcal{P}_i$,*

$$\lim_{n\to\infty} \mathbf{Prob}\{g_n(Z_1, Z_2, \ldots, Z_n) = i\} = 1. \tag{5.1}$$

Then there is an ergodic process $\{Y_n\}$ such that for both $i = 0$ and $i = 1$,

$$\limsup_{n\to\infty} \mathbf{Prob}\{g_n(Y_1, Y_2, \ldots, Y_n) = i\} = 1. \tag{5.2}$$

To see the main idea of the proof we will first formulate another version which is easier to prove. In this version we take two ergodic systems $(X_i, \mathcal{X}_i, \mu_i, T_i)$, $i = 0, 1$, and consider the class $\mathcal{Q}_i$ of **all** S-valued processes that are defined by partitions of X_i into $|S|$ sets. The difference between these classes and the $\mathcal{P}_i$ of the theorem is that for them we should be restricting attention to processes that are defined by **generating** partitions of the system $(X_i, \mathcal{X}_i, \mu_i, T_i)$. If there is to be a sequence of functions g_n that distinguishes between these classes, the classes themselves should be disjoint, that is to say, they should have no common factor. This would follow from the disjointness of the ergodic systems $(X_i, \mathcal{X}_i, \mu_i, T_i)$, but is a strictly weaker property.

5.2. THEOREM: *Let $\mathcal{Q}_0, \mathcal{Q}_1$ consist of all ergodic processes with values in a fixed finite set S that are defined by partitions of $(X_i, \mathcal{X}_i, \mu_i, T_i)$, two fixed ergodic systems that are not isomorphic. Suppose g_n is a sequence of functions with domain S^n and range $\{0,1\}$ such that for any ergodic process $\{Z_n\}$ in $\mathcal{Q}_i$,*

$$\lim_{n\to\infty} \mathbf{Prob}\{g_n(Z_1, Z_2, \ldots, Z_n) = i\} = 1. \tag{5.3}$$

Then there is an ergodic process $\{W_n\}$ such that for both $i = 0$ and $i = 1$,

$$\limsup_{n\to\infty} \mathbf{Prob}\{g_n(W_1, W_2, \ldots, W_n) = i\} = 1. \tag{5.4}$$

For the proof we will need the following well-known property of ergodic systems which we formulate as a lemma. It is a simple version of the copying lemmas that play a key role in the isomorphism theory of Bernoulli shifts. It can be easily established using the Rokhlin lemma, and the fact that any finite stationary distribution can be approximated arbitrarily well using empirical distributions of the form $\mathbf{D}(b|u)$ for u fixed and b ranging over all N blocks.

5.3. LEMMA: *Let $(X, \mathcal{X}, \mu, T)$ be an ergodic, nonatomic measure preserving system, and $\{U_n\}$ any finite-valued stationary stochastic process. Then for any N and positive ϵ there is a finite partition α such that the distribution of $\bigvee_{n=0}^{N} T^{-n}\alpha$ differs from the joint distribution of $\{U_n, 0 \leq n \leq N\}$ by less than ϵ.*

We turn now to the proof of Theorem 5.2:

Proof: In the proof we will shift back and forth between partitions of the spaces X_i and the corresponding stochastic processes. Begin with any nontrivial partition α_1 of X_0, and let $Y^{(1)}$ be the corresponding stochastic process. We fix some sequence tending rapidly to zero, for example $c_m = 10^{-m}$, and choose N_1 so that

$$\mathbf{Prob}(g_{N_1}(Y_1^{(1)}, Y_2^{(1)}, \ldots, Y_{N_1}^{(1)}) = 0) > 1 - c_1. \tag{5.5}$$

Choose an ϵ_1 so that any process whose N_1 distribution is within ϵ_1 of that of $Y^{(1)}$ will also satisfy equation (5.5). Apply Lemma 5.3 with these parameters to the space X_1, to find a partition β_2 of X_1 so that the corresponding process $Z^{(2)}$ satisfies

$$\mathbf{Prob}(g_{N_1}(Z_1^{(2)}, Z_2^{(2)}, \ldots, Z_{N_1}^{(2)}) = 0) > 1 - c_1. \tag{5.6}$$

However, this process is in $\mathcal{Q}_1$, and thus there will be an $N_2 > N_1$ so that

$$\mathbf{Prob}(g_{N_2}(Z_1^{(2)}, Z_2^{(2)}, \ldots, Z_{N_2}^{(2)}) = 1) > 1 - c_2. \tag{5.8}$$

Choose an ϵ_2 so that any process whose N_2 distribution is within ϵ_2 of that of $Z^{(2)}$ will continue to satisfy both equations (5.6) and (5.7). Apply the lemma again with these parameters to the space X_0, to find a partition α_3 of X_0 so that the corresponding process $Y^{(3)}$ satisfies

$$\mathbf{Prob}(g_{N_2}(Y_1^{(3)}, Y_2^{(3)}, \ldots, Y_{N_2}^{(3)}) = 1) > 1 - c_2 \tag{5.9}$$

as well as equation (5.5) (with $Y^{(1)}$ replaced by $Y^{(3)}$). It should now be clear how to continue and produce in the limit some stationary process for which equation (5.4) holds for both $i = 0$ and 1. This would complete the proof if only we would know that the resulting process, which we may denote by $\{W_n\}$, is ergodic. I will now explain how to choose the N_j's so that indeed $\{W_n\}$ is ergodic.

The point is that one may express the ergodicity of a process as a series of conditions on the empirical distributions of k-blocks within n-strings being approximately constant for n sufficiently large. This property, which we explored in the previous section when we talked about uniform rates, is stable under small perturbations of the finite distributions. Thus what is required of N_j in addition is that it be large enough so that the empirical distribution of j-blocks in the N_j-strings is within ϵ_j of the true distribution. Naturally, we also require the same to continue to hold when we apply the lemma to switch over to the other space. This can easily be done since the systems X_i on which all our processes are defined are ergodic. ∎

For the proof of Theorem 5.1 we will need another version of the copying lemma.

5.4. Lemma: *Let $(X, \mathcal{X}, \mu, T)$ be a zero entropy ergodic, nonatomic measure preserving system, and $\{U_n\}$ any finite-valued zero entropy stationary stochastic process. Then for any N and positive ϵ there is a finite partition α such that the distribution of $\bigvee_{n=0}^{N} T^{-n}\alpha$ differs from the joint distribution of $\{U_n, 0 \le n \le N\}$ by less than ϵ and, in addition, α is a generator, i.e. $\bigvee_{n=-\infty}^{\infty} T^{n}\alpha$ is the full σ-algebra $\mathcal{X}$.*

A proof may be found in ([DGS-1976], §28). With this, the proof of Theorem 5.1 is carried out just like the proof of Theorem 5.2 with Lemma 5.4 replacing the use of Lemma 5.3. There is nothing new in the execution of this idea and so we leave it to the reader.

References

[B-1976] D. Bailey, *Sequential Schemes for Classifying and Predicting Ergodic Processes*, Ph.D. thesis, Stanford University, 1976.

[CS-2000] I. Csiszár and P. Shields, *The consistency of the BIC Markov order estimator*, Annals of Statistics **28** (2000), 1601–1619.

[DGS-1976] M. Denker, C. Grillenberger and K. Sigmund, *Ergodic Theory on Compact Spaces*, Lecture Notes in Mathematics **527**, Springer, Berlin, 1976.

[DP-1994] A. Dembo and Y. Peres, *A topological criterion for hypothesis testing*, Annals of Statistics **22** (1994), 106–117.

[Feller-II] W. Feller, *An Introduction to Probability Theory and its Applications*, Volume 2, Wiley, New York, 1971.

[F-1960] H. Furstenberg, *Stationary Processes and Prediction Theory*, Princeton University Press, 1960.

[F-1967] H. Furstenberg, *Disjointness in ergodic theory, minimal sets, and a problem in Diophantine approximation*, Mathematical Systems Theory **1** (1967), 1–49.

[H-1965] A. Heller, *On stochastic processes derived from Markov chains*, Annals of Mathematical Statistics **36** (1965), 1286–1291.

[KASW-1998] I. Kontoyiannis, P. Algoet, Yu. M. Suhov and A. J. Wyner, *Nonparametric entropy estimation for stationary processes and random fields, with application to English text*, IEEE Transactions on Information Theory **44** (1998), 1319–1327.

[Ki-1994] J. Kieffer, *Strongly consistent code-based identification and order estimation for constrained finite-state model classes*, IEEE Transactions on Information Theory **39** (1993), 893–902.

[MW] G. Morvai and B. Weiss, *On classifying processes*, Bernoulli, to appear.

[OW-1990] D. Ornstein and B. Weiss, *How sampling reveals a process*, Annals of Probability **18** (1990), 905–930.

[OW-1993] D. S. Ornstein and B. Weiss, *Entropy and data compression schemes*, IEEE Transactions on Information Theory **39** (1993), 78–83.

[Z-1978] J. Ziv, *Coding theorems for individual sequences*, IEEE Transactions on Information Theory **24** (1978), 405–412.