Representational capacity of a set of independent neurons

Inés Samengo and Alessandro Treves
Scuola Internazinale di Studi Superiori Avanzati, Programme in Neuroscience, Via Beirut 2 - 4, 34014 Trieste, Italy
(Received 25 August 2000; published 27 December 2000)

The capacity with which a system of independent neuron-like units represents a given set of stimuli is studied by calculating the mutual information between the stimuli and the neural responses. Both discrete noiseless and continuous noisy neurons are analyzed. In both cases, the information grows monotonically with the number of neurons considered. Under the assumption that neurons are independent, the mutual information rises linearly from zero, and approaches exponentially its maximum value. We find the dependence of the initial slope on the number of stimuli and on the sparseness of the representation.

DOI: 10.1103/PhysRevE.63.011910 PACS number(s): 87.19.La, 87.18.Sn, 87.19.Bb

I. INTRODUCTION

Neural systems have the capacity, among others, to represent stimuli, objects and events in the outside world. Here, we use the word *representation* to refer to an association between a certain pattern of neural activity and some external correlate. Irrespective of the identity or the properties of the items to be represented, information theory provides a framework where the capacity of a specific coding scheme can be quantified. How much information can be extracted from the activity of a population of neurons about the identity of the item that is being represented at any one moment? Such a problem, in fact, has already been studied experimentally [1–11]. Typically a discrete set of p stimuli is presented to a subject, while the activity of a population of N neurons is recorded. At its simplest, this activity can be described as an N dimensional vector $\mathbf{r}$, whose components are the firing rates of individual neurons computed over a predefined time window. The measured response is expected to be selective, at least to some degree, to each one of the stimuli. This degree of selectivity can be quantified by the mutual information between the set of stimuli and the responses [12]

$$I=\sum_{s=1}^{p} P(s)\sum_{\mathbf{r}} P(\mathbf{r}|s)\log_2\left[\frac{P(\mathbf{r}|s)}{P(\mathbf{r})}\right], \tag{1}$$

where $P(s)$ is the probability of showing stimulus s, $P(\mathbf{r}|s)$ is the conditional probability of observing response $\mathbf{r}$ when the stimulus s is presented and

$$P(\mathbf{r})=\sum_{s=1}^{p} P(s)P(\mathbf{r}|s). \tag{2}$$

The mutual information I characterizes the mapping between the p stimuli and the response space, and represents the amount of information conveyed by $\mathbf{r}$ about which of the p stimuli was shown. If each stimulus evokes a unique set of responses, i.e., no two different stimuli induce the same response, then Eq. (1) reduces to the entropy of the stimulus set, and is, therefore, $\log_2 p$. On the other hand, if a response $\mathbf{r}$ may be evoked by more than one stimulus, the mutual information is less than the entropy of the stimuli. In the extreme case where the responses are independent of the stimulus shown, $I=0$.

In Fig. 1 we show the mutual information extracted from neural responses from the inferior temporal cortex of a macaque when exposed to p visual stimuli [8]. Diamonds correspond to $p=20$, squares to $p=9$ and triangles to $p=4$. The graph is plotted as a function of the number of neurons considered. Initially, the information rises linearly. As N grows, the increase of $I(N)$ slows down, apparently saturating at some asymptotic value compatible with the theoretical maximum $\log_2 p$.

The behavior shown in Fig. 1 is quite a common observation also in other experiments of the same type [6,7,10,11]. From the theoretical point of view, different conclusions

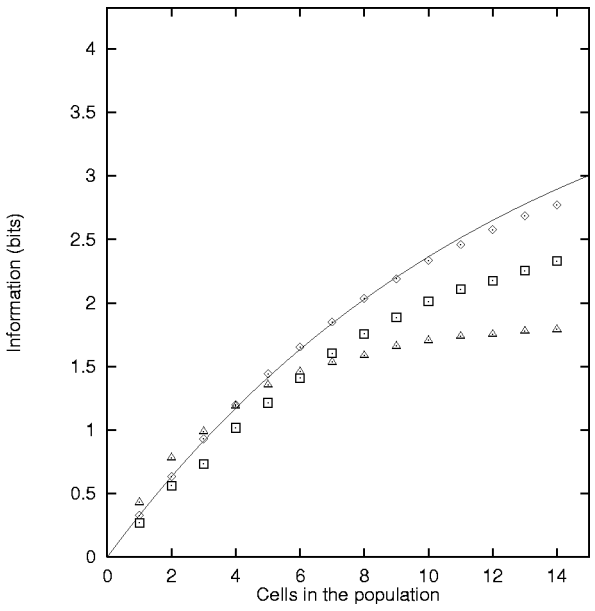

FIG. 1. Mutual information extracted from the activity of inferior temporal cortical neurons of a macaque when exposed to p visual stimuli. Diamonds correspond to $p=20$, squares to $p=9$ and triangles to $p=4$. The graph is plotted as a function of the number of neurons considered, once an average upon all the possible permutations of neurons has been carried out. The theoretical maximum is, in each case, $\log_2 p=4.32$ bits, 3.16 bits, and 2 bits, respectively. The full line shows a fit of Eq. (3) to the case of $p=20$.

have been drawn, over the years, from these curves. Obviously, the saturation in itself implies that, after a while, adding more and more neurons provides no more than redundant information. Gawne and Richmond [13] have considered a simple model which yields an analytical expression for $I(N)$ under the assumption that each neuron provides a fixed amount of information, $I(1)$, and that a fixed fraction of such an amount, y, is redundant with the information conveyed by any other neuron. The model yields $I(\infty)=I(1)/y$. Rolls *et al.* [8] have considered a more constrained model that, in addition, assumes that $y=I(1)/\log_2 p$. Later it was shown that this is, in fact, the mean pairwise redundancy if the information provided by different cells has a random overlap [14]. In this kind of phenomenological description, the information provided by a population of N cells reads

$$I(N)=\log_2(p)[1-(1-y)^N]. \tag{3}$$

The full line in Fig. 1 shows a fit of Eq. (3) to the data, in the case of $p=20$.

It has also been suggested [8] that monitoring the linear rise for small N may tell whether the representation of the stimuli is distributed or local. In a distributed scheme many neurons participate in coding for each stimulus. On the contrary, in a local representation—sometimes called grandmother cell encoding—each stimulus is represented by the activation of just one or a very small number of equivalent neurons.

Here we present a theoretical analysis of the dependence of I on N for independent units. In contrast to the previous phenomenological description, we model the response of each neuron to every stimulus. In Secs. II and III we derive $I(N)$ for several choices of the single unit response probability. In Sec. IV we discuss the relation of the mutual information defined in Eq. (1) to an informational measure of retrieval accuracy. We end in Sec. V with some concluding remarks.

II. DISCRETE, NOISELESS UNITS

In what follows, the issue of quantifying the mean amount of information provided by N units is addressed. To do so, the response of each unit to every stimulus is specified. From such responses, the mutual information is calculated using Eq. (1). Two types of models are considered. In this section we deal with discrete noiseless units, while in Sec. III we turn to continuous noisy ones.

We consider N units responding to a set of stimuli. The response r_i of unit i is taken to vary in a discrete set of f possible values. The states of the whole assembly of N units are written as $\mathbf{r}\in\mathcal{R}$, where $\mathbf{r}=(r_1,\ldots,r_N)$. Throughout the paper, letters in bold stand for vectors in an N-dimensional space. The total number of states in $\mathcal{R}$ is therefore f^N.

The stimuli $\{s\}$ to be discriminated constitute a discrete set $\mathcal{S}$ of p elements. For simplicity, we assume that they are all presented to the neural system with the same frequency, namely

$$P(s)=\frac{1}{p}. \tag{4}$$

In order to calculate the mutual information between $\mathcal{S}$ and $\mathcal{R}$ we assume that each stimulus has a representation in $\mathcal{R}$. In other words, for each stimulus s there is a fixed N-dimensional vector $\mathbf{r}^s$. Superscripts label stimuli, while subscripts stand for units.

The fact that the neurons are noiseless means that the mapping between stimuli and responses is deterministic. That is to say, for every stimulus there is a unique response $\mathbf{r}^s$. Mathematically,

$$P(\mathbf{r}|s)=\begin{cases}1 & \text{if} \quad \mathbf{r}=\mathbf{r}^s,\\ 0 & \text{if} \quad \mathbf{r}\neq\mathbf{r}^s.\end{cases} \tag{5}$$

Therefore, for every $s\in\mathcal{S}$ there is one and only one $\mathbf{r}\in\mathcal{R}$. The reciprocal, however, is in general not true. If several stimuli happen to have the same representation—which may well be the case if too few units are considered—then a given $\mathbf{r}$ may come as a response to more than one stimulus. In order to provide a detailed description of the way the stimuli are associated to the responses, we define $S_{\mathbf{r}}$ as the number of stimuli whose representation is state $\mathbf{r}$. Clearly,

$$\sum_{\mathbf{r}} S_{\mathbf{r}}=p, \tag{6}$$

and

$$P(\mathbf{r})=\frac{S_{\mathbf{r}}}{p}. \tag{7}$$

When the conditional probability (5) is inserted in Eq. (1), the sum on the responses can be carried out, since only a single vector $\mathbf{r}=\mathbf{r}^s$ gives a contribution. The mutual information reads

$$I=\sum_{\mathbf{r}}\frac{S_{\mathbf{r}}}{p}\log_2\left(\frac{p}{S_{\mathbf{r}}}\right). \tag{8}$$

Thus, I is entirely determined by the way the stimuli are clustered in the response space. For example:

- Consider the case where all stimuli evoke the same response. This means that all the $\mathbf{r}^s$ coincide. Accordingly, $S_{\mathbf{r}^s}=p$ while all the other $S_{\mathbf{r}}$ vanish. There is no way the responses can give information about the identity of the patterns, and $I=0$.
- If every stimulus evokes its distinctive response there are no two equal $\mathbf{r}^s$. This means that a number p of the $S_{\mathbf{r}}$ are equal to one, while the remaining vanish. The responses fully characterize the stimuli, and $I=\log_2 p$.
- Consider the case of even clustering, where the representations are evenly distributed among all the states of the system. This, or something close to it, may in fact happen when the number of patterns is much larger than the number of states $p\gg f^N$. Thus, $S_{\mathbf{r}}=p/f^N$, for all $\mathbf{r}$, and $I=\log_2(f^N)$. This is the maximum amount of information that can be extracted when the set of stimuli has been partitioned in f^N

subsets, and the responses are only capable of identifying the subsets, but not individual stimuli.

A. A local coding scheme

We now consider another example, namely that of a local coding scheme, sometimes called a system of *grandmother cells*. In 1972 Barlow proposed a single neuron doctrine for perceptual psychology [15]. If a system is organized in order to achieve as complete a representation as possible with the minimum number of active neurons, at progressively higher levels of sensory processing fewer and fewer cells should be active. However the firing of each one of these high level units should code for a very complex stimulus (as for example, one's grandmother). The encoding of information of such a scheme is described as local.

Local coding schemes have been shown to have several drawbacks [11], as their extreme fragility to the damage of the participating units. Nevertheless, there are some examples in the brain of rather local strategies such as, for example, retinal ganglion cells (only activated by spots of light in a particular position of the visual field [16]) or the rodent's hippocampal place cells (only responding when the animal is in a specific location in its environment [17]).

We now evaluate the mutual information in such grandmother-cell scheme, making use of Eq. (8). For simplicity, we take the units to be binary ($f=2$). We assume that each unit j responds to a single stimulus $s(j)$. Let us take that response to be 1, and the response to any other stimulus to be 0. All units are taken to respond to one single stimulus and, at first, we take at most one responsive unit per stimulus. Thus, for the time being, $N \leqslant p$.

This particular choice for the representations means that out of the 2^N states of the response space, only a subset of $N+1$ vectors is ever used. Actually, $S_{\mathbf{0}}=p-N$, while for all one-active-unit states $\mathbf{e}$, $S_{\mathbf{e}}=1$. For the remaining responses, $S_{\mathbf{r}}=0$. Therefore, the mutual information reads

$$I=\frac{N}{p}\log_2(p)+\frac{p-N}{p}\log_2\left(\frac{p}{p-N}\right). \tag{9}$$

In Fig. 2 we show the dependence of I on the number of cells, for several values of p. It can be readily seen that for $N \ll p$

$$I \approx \frac{N}{p}\,\frac{1+\ln p}{\ln 2}+\mathcal{O}(N/p)^2. \tag{10}$$

In the limit of large p Eq. (10) coincides with the intuitive approximation

$$I(N)=NI(1)=N\left[\frac{1}{p}\log_2 p+\frac{p-1}{p}\log_2\left(\frac{p}{p-1}\right)\right]. \tag{11}$$

A linear rise in $I(N)$ means that different neurons provide different information, or, in other words, that there is no redundancy in the responses of the different cells. As seen in Fig. 2, this is, in fact, the case when N is small and p is large enough. When a cell does not respond, it is still providing some information, namely, that it is not recognizing its specific stimulus. When two cells are considered, a part of this non-specific information overlaps with the information conveyed by the second cell, when responding. In other words, if two cells respond to different stimuli then, when one of them is in state 1, the other is, for sure, in state 0. Therefore, strictly speaking, the information provided by different neurons in a grandmother-like encoding is not independent. However, in the limit of $N/p \to 0$ the number of stimuli not evoking responses in any single cell is large enough as to make the information approximately additive.

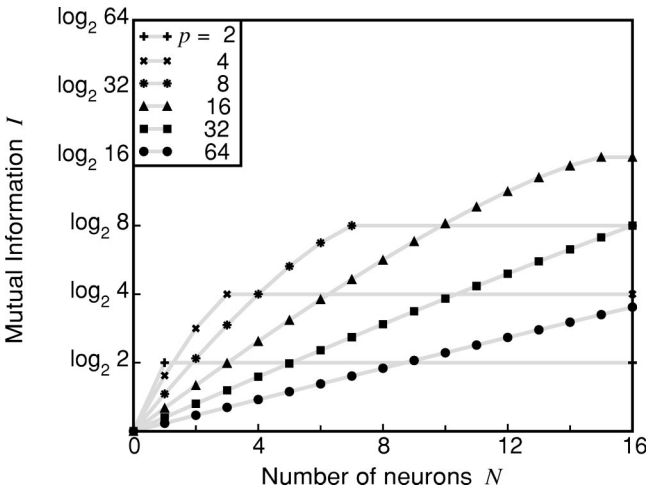

FIG. 2. Mutual information I as a function of the number of cells N, for different sizes of the set of stimuli, in the case of localized encoding. For small N, the information rises linearly with a slope proportional to $1/p$. When $N=p-1$, I saturates at $\log_2 p$.

As N approaches p, such an independence no longer holds, so the growth of $I(N)$ decelerates, and the curve approaches $\log_2 p$. For $N=p-1$, the mutual information is exactly equal to $\log_2 p$, and remains constant when more units are added. In fact, $p-1$ noiseless units are enough to accurately identify p stimuli. If all $p-1$ are silent, then the stimulus shown is the one represented by the missing unit.

In a slightly more sophisticated approach, each unit can have any number of responses f. But as long as the conditional probability $P(r_j|s)$ is the same for all those s that are not $s(j)$, Eq. (9) still holds.

It should be kept in mind that up to now we have considered the optimal situation, in that different units always respond to different stimuli. If several cells respond to the same stimulus, a probabilistic approach is needed since otherwise, the growth of $I(N)$ depends on the order in which the units are taken. Averaging over all possible selections of N cells from a pool of M units (the whole set is such that there are M/p cells allocated to each stimulus) the result shown in Fig. 3 is obtained. We have taken $p=32$, and different curves correspond to various values of M. The probabilistic approach smooths the sharp behavior observed in Fig. 2. Actually, the asymptote $\log_2 p$ can only be reached when there is certainty that there are $p-1$ units responding to different stimuli, that is, for $N=1+(p-2)M/p$. However, it is readily seen that with M/p as large as 5, the curves are already very near to the limit case of $M/p \to \infty$.

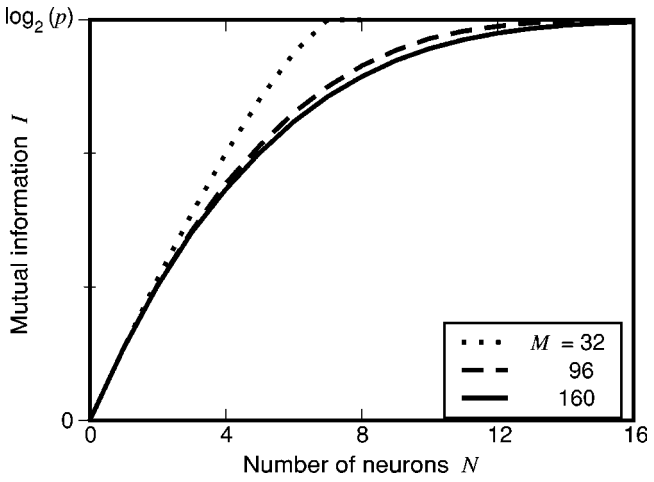

FIG. 3. Mutual information I as a function of the number of recorded units N, once averaged over all the possible selections of N cells picked up from a pool of M (the latter constituted of M/p units responding to each stimulus). Different curves correspond to various values of M, and $p=32$.

B. Distributed coding schemes

As an alternative to the local coding scheme described above, we now treat the case of distributed encoding, ranging from sparsely to fully distributed. However, in doing so, we employ a different approach, namely, we average the information upon the details of the representation.

Equation (8) implies that the amount of information that can be extracted from the responses depends on the specific representations of the p stimuli. Since it is desirable to have a somewhat more general result, we define an averaged mutual information $\langle I\rangle$:

$$\langle I\rangle=\sum_{\mathbf{r}^1,\ldots,\mathbf{r}^p} P_0(\mathbf{r}^1,\ldots,\mathbf{r}^p)I, \tag{12}$$

where the mean is taken over a probability distribution $P_0(\mathbf{r}^1,\ldots,\mathbf{r}^p)$ of having the representation in positions $\mathbf{r}^1,\ldots,\mathbf{r}^p$. This distribution, of course, is determined by the coding scheme used by the system. By averaging the information we depart from the experimental situation, where the recorded responses strongly depend on the very specific set of stimuli chosen. But, in return, the resulting information characterizes, more generally, the way neurons encode a certain type of stimuli, rather than the exact stimuli that have actually been employed.

We write P_0 as a product of single distributions for each representation,

$$P_0(\mathbf{r}^1,\ldots,\mathbf{r}^p)=\prod_{s=1}^{p} P_1(\mathbf{r}^s). \tag{13}$$

This implies that the representation of one item does not bias the probability distribution of the representation of any other. In this sense, we can say that Eq. (13) assumes that representations are independent from one another.

If in one particular experiment the set of stimuli is large enough to effectively sample $P_1(\mathbf{r}^s)$ the averaged information will be close to the experimental result.

We further assume that there is a probability distribution $\rho(r_j)$ that determines the frequency at which unit j goes into state r_j (or fires at rate r_j). If ρ is strongly peaked at a particular state—which can always be taken as zero—the code is said to be sparse. On the contrary, a flat ρ gives rise to a fully distributed coding scheme.

Finally, we assume that different units are independent. In other words, we factorize the probability that a given stimulus is represented by the state $\mathbf{r}$ as

$$P_1(\mathbf{r})=\prod_{j=1}^{N}\rho(r_j). \tag{14}$$

In order to average the information (8) we need to derive the probability that stimuli are clustered into any possible set of $\{S_\mathbf{r}\}$. Such a probability reads

$$P(\{S\})=\binom{p}{\{S\}}\Pi_\mathbf{r}[P_1(\mathbf{r})]^{S_\mathbf{r}}, \tag{15}$$

where

$$\binom{p}{\{S\}}=\frac{p!}{\Pi_\mathbf{r}S_\mathbf{r}!}. \tag{16}$$

Therefore, the average mutual information may be written as

$$\langle I\rangle=\sum_{\{S\}} P(\{S\})I. \tag{17}$$

The summation runs over all sets $\{S\}$ such that $\Sigma_\mathbf{r}S_\mathbf{r}=p$. Replacing Eq. (8) in Eq. (17), we obtain

$$\langle I\rangle=\sum_{\{S\}}\binom{p}{\{S\}}\Pi_\mathbf{r}[P_1(\mathbf{r})]^{S_\mathbf{r}}\sum_{\mathbf{r}'}\frac{S_{\mathbf{r}'}}{p}\log_2\left(\frac{p}{S_{\mathbf{r}'}}\right). \tag{18}$$

Rearranging the summation so as to explicitly separate out a single $S_{\mathbf{r}_j}$ one may write

$$\langle I\rangle=\sum_{\mathbf{r}}\sum_{S_\mathbf{r}=1}^{p}\frac{p!}{S_\mathbf{r}!}[P(S_\mathbf{r})]^{S_\mathbf{r}}\frac{S_\mathbf{r}}{p}\log_2\left(\frac{p}{S_\mathbf{r}}\right)\frac{1}{(p-S_\mathbf{r})!}A, \tag{19}$$

where A is the sum over all other S, namely

$$A=\sum_{\{S_{\mathbf{r}'\neq\mathbf{r}}\}}\binom{p-S_\mathbf{r}}{\{S_{\mathbf{r}'\neq\mathbf{r}}\}}\prod_{\mathbf{r}''\neq\mathbf{r}}[P_1(\mathbf{r}'')]^{S_{\mathbf{r}''}}=[1-P_1(\mathbf{r})]^{p-S_\mathbf{r}}. \tag{20}$$

Thus,

$$\langle I\rangle=\sum_{\mathbf{r}}\sum_{S_\mathbf{r}=1}^{p-1}\frac{(p-1)!}{(S_\mathbf{r}-1)![(p-1)-(S_\mathbf{r}-1)]!}\log_2\left(\frac{p}{S_\mathbf{r}}\right)\times[P_1(\mathbf{r})]^{S_\mathbf{r}}[1-P_1(\mathbf{r})]^{p-S_\mathbf{r}}. \tag{21}$$

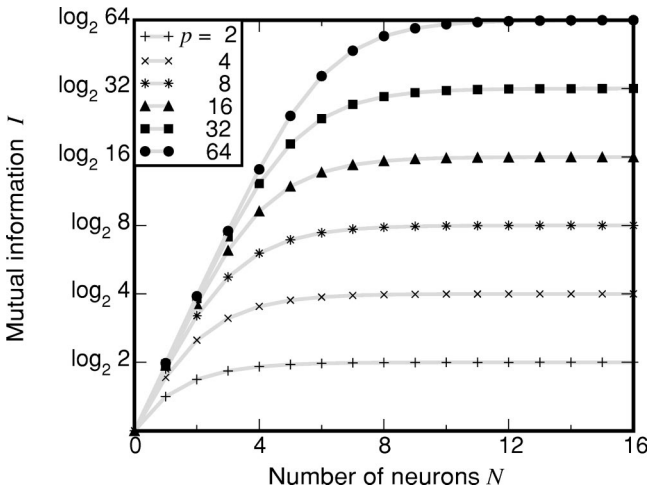

FIG. 4. Mean mutual information $\langle I\rangle_{\rm dis}$ as a function of the number of neurons N for several values of p. Initially the information rises linearly with a slope only slightly depending on p. As N increases, $\langle I\rangle_{\rm dis}$ eventually saturates at $\log_2 p$.

We now discuss two particular cases of Eq. (21). First, we take the encoding to be fully distributed, namely $\rho(r_j)=1/f$. Therefore, $P_1(\mathbf{r}_j)=1/f^N$. If this is replaced in the previous expression, we obtain

$$\langle I\rangle_{\rm dis}=(1-f^{-N})^{p-1}\sum_{S=0}^{p-2}\frac{(p-1)!}{S!(p-1-S)!}$$

$$\times(f^N-1)^{-S}\log_2\left(\frac{p}{S+1}\right). \tag{22}$$

It may be seen that the dependence of the information on f and N always involves the combination f^N. This means that neither the number of units, nor how many distinctive firing rates each unit has are relevant in themselves. Only the total number of states matters.

In Fig. 4 we plot the relation between $\langle I\rangle_{\rm dis}$ and N for several values of p. Initially the information rises linearly with a slope only slightly dependent on p. As N increases, $\langle I\rangle_{\rm dis}$ eventually saturates at $\log_2 p$. The limit cases are easily derived:

$$\lim_{N\ln f\to 0}\langle I\rangle_{\rm dis}=N(p-1)\ln f\log_2\left(\frac{p}{p-1}\right), \tag{23}$$

$$\lim_{f^N/p\to\infty}\langle I\rangle_{\rm dis}=\log_2 p-(p-1)f^{-N}. \tag{24}$$

If the number of stimuli is large, Eq. (23) becomes

$$\lim_{N\ln f\to 0}\lim_{p\to\infty}\langle I\rangle_{\rm dis}=N\frac{p-1}{p}\log_2 f. \tag{25}$$

Notice that in contrast to the local coding scheme Eq. (9), the initial slope of $I(N)$ hardly depends on p (actually, it increases slightly with p). This makes the distributed encoding a highly efficient way to read out information about a large set of stimuli by the activity of just a few units.

As opposed to the fully distributed case, a sparse distributed encoding is now considered, with $f=2$, $\rho(1)=q$, $\rho(0)=1-q$ and $q\ll 1$. This choice is again a binary case, but with one response much more probable than the other. As a consequence, the most likely representations in $\mathcal{R}$ space are those with either zero or at most one active neuron. In fact, $P_1(\mathbf{r}^s=\mathbf{0})=(1-q)^N$, whereas if the representation is a one-active-unit state $\mathbf{e}$, $P_1(\mathbf{e})=q(1-q)^{N-1}$. The probability of all other representations is higher order in q.

Accordingly, to first order in q, we only consider the combinations of p representations with at least $p-1$ of them in state $\mathbf{r}^s=\mathbf{0}$. These are the only responses with a probability P_0 at most linear in q. More precisely, the probability of representing all p stimuli with the same state $\mathbf{r}=\mathbf{0}$ is $P_0(\mathbf{0},\mathbf{0},\ldots,\mathbf{0})=[P_1(\mathbf{0})]^p\approx 1-Npq$. In the same way, the probability of having $N-1$ stimuli in $\mathbf{0}$ and a single one-active-unit state is q. There are N different possible one-active-unit states, and any one of the p stimuli can be such a state. Taking all this into account, we find that up to the first order in Npq,

$$\langle I\rangle_{\rm spa}=Npq\left[\frac{p-1}{p}\log_2\left(\frac{p}{p-1}\right)+\frac{1}{p}\log_2 p\right]. \tag{26}$$

Expanding this expression for large p, we obtain

$$\lim_{p\to\infty}\langle I\rangle=Nq\frac{1+\ln p}{\ln 2}. \tag{27}$$

This means that from the experimental measurement of the slope of $\langle I(N)\rangle$ it is possible to extract the sparseness of an equivalent binary model, which can be compared with a direct measurement of the sparseness. If the number of stimuli cannot be considered large, the whole of Eq. (26) can be used to derive a value for q.

It should be noticed that if $q=1/p$ Eq. (26) coincides with the expression (11) for a grandmother-like encoding. This makes sense, since $q=1/p$ implies that, on average, any one unit is activated by a single pattern. In short, it corresponds to a probabilistic description of the localized encoding. Notice, though, that $q=1/p$ is outside the range of validity of our limit $Npq\ll 1$.

III. CONTINUOUS, NOISY NEURONS

In this section we turn to a more realistic description of the single neuron responses. Specifically, we allow the states r_j to take any real value. Therefore, the response space $\mathcal{R}$ is now Re^N. In addition, we depart from the deterministic relationship between stimuli and responses. This means that upon presentation of stimulus s, there is no longer a unique response. Instead, the response vector $\mathbf{r}$ is most likely centered at a particular $\mathbf{r}^s$, and shows some dispersion to nearby vectors. The aim is to calculate the mutual information between the responses and the stimuli requiring as little as possible from the conditional probability $P(\mathbf{r}|s)$. A single parameter σ is introduced as a measure of the noise in the representation. Thus,

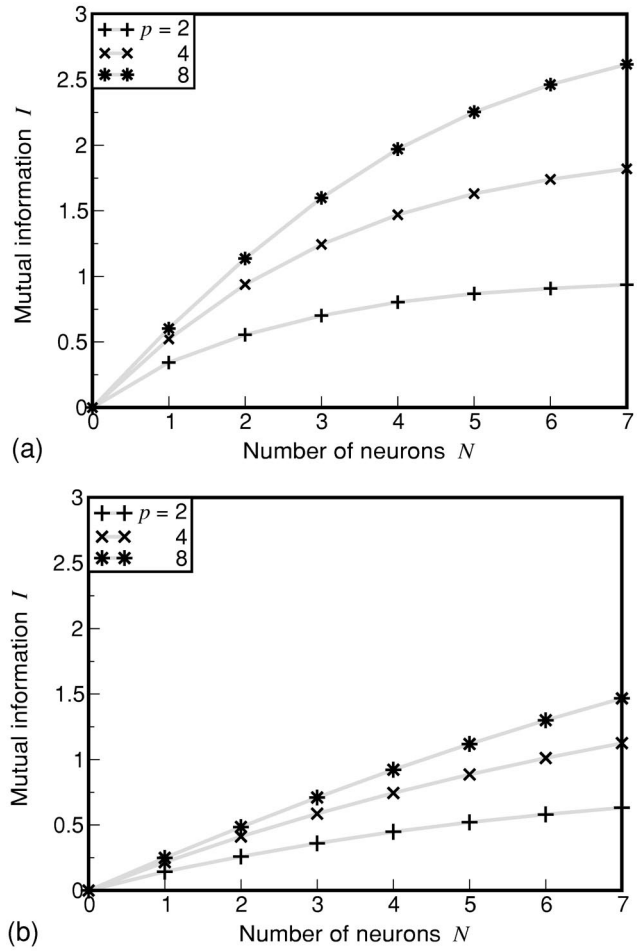

FIG. 5. Results of the numerical evaluation of the mutual information for continuous noisy neurons, where p is the number of stimuli in the set. In (a) $\sigma=\lambda/2$, and in (b) $\sigma=\lambda$.

$$P(\mathbf{r}|s)=\prod_{j=1}^{N}\frac{e^{-(r_j-r_j^s)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}, \tag{28}$$

where the index s takes values from 1 to p. The conditional probability depends on the distance between the actual response $\mathbf{r}$ and a fixed vector $\mathbf{r}^s\in\mathcal{R}$, which is the mean response of the system to stimulus s. There is one such $\mathbf{r}^s$ for every element in $\mathcal{S}$. The choice of Gaussian functions is only to keep the description simple and analytically tractable. By factorizing $P(\mathbf{r}|s)$ in a product of one component probabilities an explicit assumption about the independence of the neurons is being made.

Figure 5 shows a numerical evaluation of the information (1), when the probability $P(\mathbf{r}|s)$ is as in Eq. (28). The information, just as in the previous section, has been averaged upon many selections of the representations $\mathbf{r}^s$. The curve is a function of the number of neurons considered N. Different lines correspond to different sizes of the set of stimuli, while in (a) $\sigma=\lambda/2$, and in (b) $\sigma=\lambda$, where λ is a parameter quantifying the mean discriminability among patterns, to be defined precisely later. Just as in the discrete distributed case, we observe an initial linear rise and a saturation at $\log_2 p$. Moreover, and pretty much as in the experimental situation of Fig. 1, the initial slope does not seem to depend strongly on the number of stimuli, at least for large values of the noise σ. In what follows, an analytical study of these numerical results is carried out. In particular, the relevant parameters determining the shape of $I(N)$ are identified.

We write the mutual information as

$$I=H_1-H_2, \tag{29}$$

where

$$H_1=-\frac{1}{p}\sum_{s=1}^{p}\int d\mathbf{r}\,P(\mathbf{r}|s)\log_2\left[\frac{1}{p}\sum_{s'=1}^{p}P(\mathbf{r}|s')\right]$$

$$=-\int d\mathbf{r}\,P(\mathbf{r})\log_2[P(\mathbf{r})], \tag{30}$$

is the total entropy of the responses, and

$$H_2=-\frac{1}{p}\sum_{s=1}^{p}\int d\mathbf{r}\,P(\mathbf{r}|s)\log_2[P(\mathbf{r}|s)] \tag{31}$$

is the conditional entropy of $P(\mathbf{r}|s)$, averaged over s.

H_2 can be easily calculated. It reads

$$H_2=\frac{N}{2\ln 2}[1+\ln(2\pi\sigma^2)]. \tag{32}$$

It is therefore linear in N. This stems from the independence of the units, since the entropy of the response space increases linearly with its dimension. It does not depend on the location of the representations $\mathbf{r}^s$, and it is a growing function of the noise σ.

In Appendix A we solve the integral in $\mathbf{r}$ of H_1 using the replica method. We obtain

$$H_1=\frac{-1}{\ln 2}\lim_{n\to 0}\frac{1}{n}\left(\left\{\frac{1}{p^{n+1}(2\pi\sigma^2)^{Nn/2}(n+1)^{N/2}}\sum_{\{K\}}\binom{n+1}{\{K\}}\right.\right.$$
$$\times\prod_{j=1}^{N}\exp\left[\frac{-1}{4\sigma^2(n+1)}\sum_{\ell=1}^{p}\sum_{m=1}^{p}K_\ell K_m\right.$$
$$\left.\left.\left.\times(r_j^\ell-r_j^m)^2\right]\right\}-1\right), \tag{33}$$

where $\{K\}$ now stands for the set $\{K_1,K_2,\ldots,K_p\}$ specifying how many replicas are representing each pattern. The summation in $\{K\}$ runs over all sets of K such that $\Sigma_{s=1}^{p}K_s=n+1$. The symbol in brackets is defined in Eq. (16). Equation (33) shows that the information depends explicitly on the ratio between all the possible differences $|\mathbf{r}^\ell-\mathbf{r}^m|$ and the noise σ. In other words, the capacity to determine which stimulus is being shown is given by a signal-to-noise ratio, characterizing the discriminability of the responses.

The mutual information I characterizes the selectivity of the correspondence between stimuli and responses. If the distance between any two vectors $|\mathbf{r}^\ell-\mathbf{r}^m|$ is much greater than

the noise σ, then the mapping is (almost) injective. Thus, in this limit the mutual information approaches its maximal value, $\log_2 p$.

If, on the other hand, the noise level in $P(\mathbf{r}|s)$ is enough to allow for some vectors $\mathbf{r}$ to be evoked with appreciable probability by more than one stimulus, the mutual information decreases. In this sense, I can be interpreted as a comparison between the noise in $P(\mathbf{r}|s)$ and the distance between any two mean responses. For a specific choice of the representations, the distance between any two of them is a non-linear function of their components. Therefore, in general, even though Eq. (28) implies that different units are independent, it is not possible to write I as a sum over units of single-units information.

Just as before, we now average the mutual information (1) over a probability distribution $P_0(\mathbf{r}^1, \ldots, \mathbf{r}^p)$ of the representations $\mathbf{r}^1, \ldots, \mathbf{r}^p$, namely

$$\langle I \rangle = \int \prod_{j=1}^{p} d\mathbf{r}^j P_0(\mathbf{r}^1, \ldots, \mathbf{r}^p) I. \tag{34}$$

Under the assumption that the responses to different stimuli are independent, P_0 reads

$$P_0(\mathbf{r}^1, \ldots, \mathbf{r}^p) = \prod_{s=1}^{p} P_1(\mathbf{r}^s). \tag{35}$$

Adding the requirement of independent units,

$$P_1(\mathbf{r}^s) = \prod_{j=1}^{N} \rho(r_j^s). \tag{36}$$

By replacing the average (12) in the separation (29) we write

$$\langle I \rangle = \langle H_1 \rangle - H_2, \tag{37}$$

since H_2 does not depend on the vectors $\mathbf{r}^s$.

So we now turn to the calculation of $\langle H_1 \rangle$, namely,

$$\langle H_1 \rangle = -\frac{1}{\ln 2} \lim_{n\to 0} \frac{1}{n} \left[\frac{1}{(n+1)^{N/2} (2\pi\sigma^2)^{Nn/2} p^{n+1}} \times \sum_{\{K\}} \binom{n+1}{\{K\}} \langle A_{\{K\}} \rangle^N - 1 \right], \tag{38}$$

where

$$\langle A_{\{K\}} \rangle = \int \prod_{s=1}^{p} dr^s \rho(r^s) \times \exp\left[-\frac{1}{4(n+1)\sigma^2} \sum_{m=1}^{p} \sum_{\ell=1}^{p} K_m K_\ell (r^m - r^\ell)^2 \right]. \tag{39}$$

The main step forward introduced by the average in Eq. (12) is that now, $\langle H_1 \rangle$ is symmetric under the exchange of any two responses, or any two neurons. In contrast, before the averaging process, the location of every single response by every single unit was relevant.

The limit in Eq. (38) can be calculated in some particular cases. In the first place, we analyze the large N limit. From Eq. (39) it is clear that $\langle A_{\{K\}} \rangle \leqslant 1$. The equality holds, in fact, only when there is a single K different from zero. In the calculation of $\langle H_1 \rangle$, as stated in Eq. (38), $A_{\{K\}}$ appears to the N-th power. Therefore, when $N \to \infty$ only the terms with $A_{\{K\}} = 1$ give a non-vanishing contribution. There are p of such terms. When the sum in (38) is replaced by p, it may be shown that once more, $\langle I \rangle = \log_2 p$.

In the following two subsections we compute $\langle I(N) \rangle$ for both large and small values of the noise σ.

A. Information in the large noise limit

We now make the assumption that the noise σ is much larger than some average width of $\rho(r)$. In other words, we suppose $\sigma^2 \gg (r^\ell - r^m)^2$, for all r^ℓ and r^m with non-vanishing probability. In this case, the exponential in Eq. (39) may be expanded in Taylor series. Up to the second order,

$$\exp\left[-\frac{1}{4(n+1)\sigma^2} \sum_{m=1}^{p} \sum_{\ell=1}^{p} K_m K_\ell (r^m - r^\ell)^2 \right] \approx 1 - \frac{1}{4(n+1)\sigma^2} \sum_{m=1}^{p} \sum_{\ell=1}^{p} K_m K_\ell (r^m - r^\ell)^2 + \frac{1}{2} \left[\frac{1}{4(n+1)\sigma^2} \sum_{m=1}^{p} \sum_{\ell=1}^{p} K_m K_\ell (r^m - r^\ell)^2 \right]^2. \tag{40}$$

If only the constant term is considered, the integral in Eq. (39) becomes the normalization condition for P_0. Thus, the sums in Eq. (38) give p^{n+1}, and it is readily seen that $\langle H_1 \rangle$ exactly cancels H_2. As expected, in the limit $\sigma^2 \to \infty$ the mutual information vanishes.

The next order of approximation is to consider the expansion (40) up to the linear term. Thus, the integral in Eq. (39) becomes

$$\langle A_{\{K\}} \rangle = 1 - \frac{\lambda^2}{4(n+1)\sigma^2} \sum_{m=1}^{p} \sum_{\ell=1, \ell\neq m}^{p} K_m K_\ell, \tag{41}$$

where

$$\lambda^2 = \int dr^1 \, dr^2 \, \rho(r^1) \rho(r^2) (r^1 - r^2)^2 \tag{42}$$

is the parameter quantifying the discriminability among patterns, and appearing in Fig. 5. We have now gained a more precise insight of the large σ limit. It stands for taking $\sigma \gg \lambda$.

Since in Eq. (38) $A_{\{K\}}$ appears to the N-th power, in order to proceed further we have to estimate the size of $N\lambda^2/\sigma^2$. We first consider the small N limit and assume, to start with, that $N\lambda^2/\sigma^2 \ll 1$. Thus, we may expand

$$(A_{\{K\}})^N \approx 1 - \frac{N\lambda^2}{4(n+1)\sigma^2} \sum_{m=1}^{p} \sum_{\ell=1,\ell\neq m}^{p} K_m K_\ell . \quad (43)$$

In Appendix B we calculate the sums in Eq. (43), thus obtaining $\langle H_1 \rangle$. When the result is replaced in Eq. (38) we get

$$\langle I \rangle = \frac{N}{\ln 2} \frac{(p-1)}{p} \left(\frac{\lambda}{2\sigma} \right)^2 . \quad (44)$$

For a large amount of noise, the information rises linearly with the number of neurons. This dependence should be compared with Eq. (25), in the discrete distributed case. The two expressions coincide, if the number of discrete states f is associated to $\exp(\lambda^2/4\sigma^2)$. Therefore, as regards to the mutual information, a dispersion σ in the representation is equivalent to having a number $\exp(\lambda^2/4\sigma^2)$ of distinguishable discrete responses. Notice that both the noisy, continuous and the discrete, deterministic approach show the same dependence on the number of patterns.

Regarding the dependence on σ, it is readily seen that as the noise decreases, the slope of I increases. In other words, every single neuron provides a larger amount of information. Since the mutual information saturates at $\log_2 p$ for $N \to \infty$, a small value of σ implies that the ceiling is quickly reached. As a consequence, the assumption $N \ll \sigma^2/\lambda^2$ can now be more precisely stated as $N \ll (\sigma^2/\lambda^2)\log_2 p$. In this regime, linearity holds.

As N increases, saturation effects become evident, and the mutual information is no longer linear. The first hint of the presence of an asymptote at $\log_2 p$ is given by the quadratic contribution to $I(N)$. In order to describe it, the whole of expansion (40) must be replaced in Eq. (39). Carrying out the integral in $r^1, \ldots, r^p$,

$$\langle A_{\{K\}} \rangle = 1 - \frac{\lambda^2}{4\sigma^2(n+1)} \sum_{\ell=1}^{p} \sum_{m=1,m\neq\ell}^{p} K_\ell K_m + \frac{\eta^4}{32\sigma^4(n+1)^2}, \quad (45)$$

where

$$\eta^4 = \int \left[\sum_{\ell=1}^{p} \sum_{m=1}^{p} (r^\ell - r^m)^2 K_\ell K_m \right]^2 \prod_{s=1}^{p} \rho(r^s) dr^s . \quad (46)$$

Extracting the sums from the integral, the limit in Eq. (38) can be solved, and

$$\langle I \rangle = \frac{N}{\ln 2} \frac{p-1}{p} \left[\frac{\lambda^2}{4\sigma^2} + \frac{1}{2(4\sigma^2)^2} C \right] - \frac{N^2}{\ln 2} \left(\frac{\lambda^2}{4\sigma^2} \right)^2 \frac{p-1}{p^2} . \quad (47)$$

Here,

$$C = \frac{2\lambda^4}{p} - 2\Lambda_1 \left(1 - \frac{2}{p} + \frac{2}{p^2} \right) - 4\Lambda_2 \frac{(p-2)}{p} \left(\frac{2}{p} - 1 \right) - 2\Lambda_3 \frac{(p-2)(p-3)}{p^2} \quad (48)$$

with

$$\Lambda_1 = \int dr^1 dr^2 \rho(r^1)\rho(r^2)(r^1 - r^2)^4,$$

$$\Lambda_2 = \int dr^1 dr^2 dr^3 \rho(r^1)\rho(r^2)\rho(r^3)(r^1 - r^2)^2 (r^1 - r^3)^2, \quad (49)$$

$$\Lambda_3 = \int dr^1 dr^2 dr^3 dr^4 \rho(r^1)\rho(r^2)\rho(r^3)\rho(r^4)(r^1 - r^2)^2 \times (r^3 - r^4)^2 .$$

Our numerical simulations corroborate that if a quadratic function is fit to the initial rise of $I(N)$, the coefficients accompanying N and N^2 depend on p and σ just as predicted by Eq. (47).

B. The limit of vanishing noise

In the first place, we take $\sigma \to 0$. If the conditional probability (28) is replaced by a δ-function, it is readily seen that $I = \log_2 p$.

In Appendix C we show that for small—but not vanishing—values of the noise σ, the mutual information is expected to grow as

$$\langle I \rangle = \log_2(p) \left[1 - \frac{p-1}{\log_2 p} (4\sqrt{\pi}\sigma B_2)^N \right], \quad (50)$$

where

$$B_2 = \int \rho^2(r) dr . \quad (51)$$

In order to corroborate this result, we have fit a function of the form $\log_2(p)[1 - a\exp(bN)]$ to the numerical evaluation of Eq. (37). In Fig. 6 we show the dependence of a and b with σ and p. We observe that coefficient a shows a dependence with the noise σ, in contrast to what is predicted by Eq. (50). It is also in contrast to the prediction of the phenomenological model leading to Eq. (3), where $a = 1$. In addition, b shows a variation with the number of stimuli p. Thus, although it is very easy to calculate the mutual information when σ is exactly equal to 0, we have not been able to derive analytically the approach to the $\log_2 p$ limit, as $\sigma \to 0$.

IV. A RELATED INFORMATIONAL MEASURE OF ACCURACY

Up until now, we have considered the mutual information of Eq. (1), a quantifier of the capacity with which a given

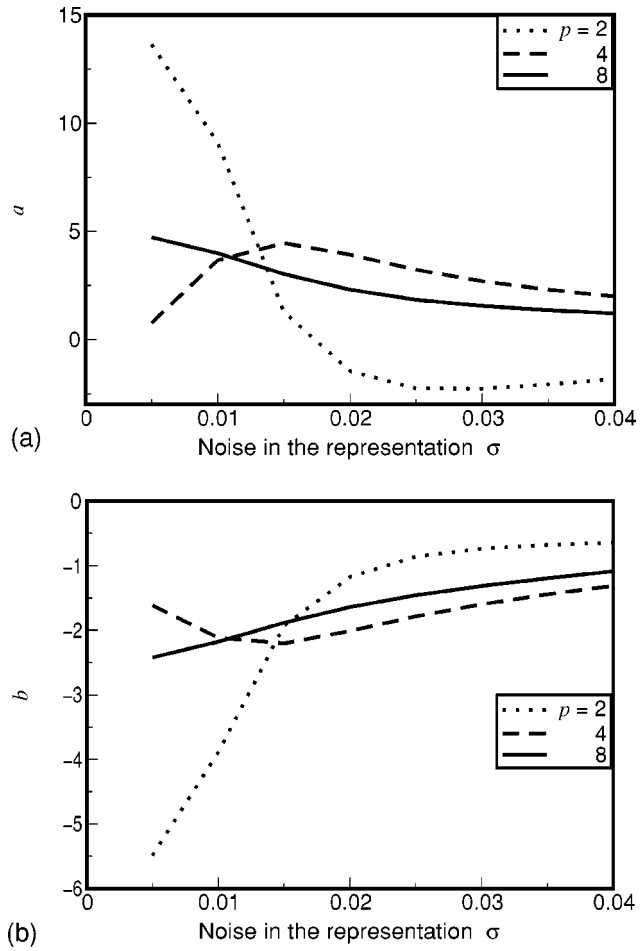

FIG. 6. Dependence of the coefficients a and b extracted from numerical evaluations of the mutual information, with the parameters p and σ.

group of units can represent a fixed set of p stimuli. This is a measure of direct relevance to neuronal recording experiments. A somewhat different information measure has been used in analyzing mathematical network models, in particular models of memory storage and retrieval. We would like to clarify the relationship between the two measures.

Consider the variability with which a typical stimulus is represented, which in a mathematical model might be described by a formula as simple as Eq. (28). There, $\mathbf{r}$ is the response during a trial, while $\mathbf{r}^s$ is the average response across trials with the same stimulus. The average variability may be quantified by the mutual information between $\mathbf{r}$ and $\mathbf{r}^s$,

$$\tilde{I}=\int d\mathbf{r}^s P(\mathbf{r}^s)\int d\mathbf{r} P(\mathbf{r}|\mathbf{r}^s)\log_2\left[\frac{P(\mathbf{r}|\mathbf{r}^s)}{P(\mathbf{r})}\right], \qquad (52)$$

where $\mathbf{r}^s$ is taken to span the space of average responses, described by the probability distribution $P(\mathbf{r}^s)$. In a different model, $\mathbf{r}^s$ might be the first response produced, and $\mathbf{r}$ the second, or any successive response; in yet other models [18–22], $\mathbf{r}^s$ might be the stored representation of a memory item, and $\mathbf{r}$ the representation emerging when the item is being retrieved. In all such cases, one need not refer to a discrete set of p stimuli, but only to a probability distribution $P(\mathbf{r}^s)$ [and, of course, to a conditional probability distribution $P(\mathbf{r}|\mathbf{r}^s)$]. This measure of accuracy is simply related to the mutual information we have considered in this paper: it is given by its $p\to\infty$ limit. In particular, the initial linear rise of $\tilde{I}$ with N is the only regime relevant to the accuracy measure, which for independent units is always purely linear in N.

Let us see this in formulas. Just as before, we assume that $P(\mathbf{r}^s)$ factorizes as

$$P(\mathbf{r}^s)=\prod_{j=1}^{N} P(r_j^s). \qquad (53)$$

The equivalent of Eq. (2) is now

$$P(\mathbf{r})=\int d\mathbf{r}^s P(\mathbf{r}^s)P(\mathbf{r}|\mathbf{r}^s). \qquad (54)$$

In Appendix D we show that

$$\tilde{I}=\frac{N}{\ln 2}\frac{\lambda^2}{4\sigma^2}. \qquad (55)$$

In the derivation of Eq. (55) no assumption of small N has been made. By comparison with Eq. (44) we see that, indeed, the information measure (52) introduced in this section coincides with the initial rise of the information about which stimulus is being shown (Sec. III), when the latter is calculated for a large number of stimuli.

V. SUMMARY AND DISCUSSION

The capacity with which a system of N independent units can code for a set of p stimuli has been studied. More precisely, the growth of the mutual information I between stimuli and responses has been calculated, for different models of the neural responses. In all these models, the units were supposed to operate independently. That is to say, the conditional probability of response $\mathbf{r}$ given stimulus s is always a product of single-unit conditional probabilities. Of course, the fact that neurons operate independently does not mean that they provide independent information. As stated in Eq. (29), the mutual information can always be separated into the difference between the entropy of the responses (H_1) and the averaged stimulus specific entropy (H_2), sometimes called noise entropy. For independent units, H_2 is always linear in N. However, the factorization of the conditional probabilities does not imply the factorization of $P(\mathbf{r})$, meaning that H_1 need not be linear in the number of units. In other words, even independent units may produce correlated responses, and indeed strongly correlated, simply because every unit is driven by the same set of stimuli. Imagine that each unit provides a very precise representation of the stimuli. If stimulus 1 is shown, the responses of the N units will show almost no trial to trial variability. When the stimulus is changed, another set of N responses is obtained. But the first responses always come together (driven by stimulus 1), and so do the second ones. Even after averaging over all stimuli, this coherent behavior implies strong correlations

between the responses. In this example, $H_2 \approx 0$ and $H_1 \approx \log_2 p$.

In other situations, when the number of stimuli is very large, or the representation of each one of them is noisier, the correlations in the responses are weaker. We have seen that, in these cases, H_1 tends to become linear in N.

Throughout the work, the responses of the units was described by a vector $\mathbf{r}$. Nothing was said, however, about what the components of the vector really are. In the experiment of Fig. 1, r_j was the firing rate of neuron j in a pre-defined time window. One might however consider a slightly more complex description in which a subset of M components is associated to the response of unit j, for example, the first M principal components of its time course [1]. Our analysis would still apply, replacing N units by MN components.

In the Introduction, reference was made to the phenomenological models where the growth of $I(N)$ as given by Eq. (3) is entirely explained by ceiling effects. In such models, the information provided by different neurons is supposed to be independent, inasmuch this is compatible with the fact that the total amount of information must be $\log_2 p$. The models presented in this paper are not in principle opposed to the phenomenological ones; rather they are at a more detailed level of description. Instead of a direct assumption on how different units share the available information, we specify conditional probabilities for the responses. As a result, we find global trends that closely resemble those of Eq. (3), that is to say, an initial linear rise and an exponential saturation at $\log_2 p$. The detailed shape of $I(N)$ is, however, different for each model.

It should be kept in mind that whatever the detailed shape of the curve, the approach to $\log_2 p$ is no more than a consequence of the fact that the number of stimuli is limited. The maximum information that can be extracted from the neural responses is $\log_2 p$. It is clear that if we have a set of neurons that already provides information very near to this maximum, by adding one more neuron we will gain no more than redundant information. In other words, we have reached a regime where the neural responses correctly distinguish the identity of each stimulus. But we cannot deduce from this that the representational capacity of the responses remains unchanged when the number of neurons increases. One should rather realize that the task itself is no longer appropriate to test the way additional neurons contribute in the encoding of stimuli. In contrast, the slope of the initial linear rise is an accurate quantification of the capacity of the system to represent items.

We have found that distributed coding schemes result in an initial slope that is roughly independent of the number of stimuli. This means that the number of units needed to reach a given fraction of the maximum information scales as $\log_2 p$—at least, for large p. In contrast, when a grandmother-cell encoding is used, the initial slope is proportional to $1/p$, and hence, one should have $N \propto p$. This makes distributed encoding much more efficient than localized schemes. In the example of the experiment of Fig. 1, the information measure supports the conclusion, already evident from the responses themselves, that the representation of faces in the inferior temporal cortex of the macaque is distributed.

ACKNOWLEDGMENTS

We thank Damian Zanette for a critical reading of the manuscript. This work has been supported by Human Frontier Science Programm, Grant No. RG 01101998B.

APPENDIX A: CALCULATION OF H_1 USING THE REPLICA METHOD

Replacing the identity

$$\ln \alpha = \lim_{n \to 0} \frac{1}{n} (\alpha^n - 1) \tag{A1}$$

in Eq. (30) the integral in $\mathbf{r}$ can be evaluated. This we show in the present appendix.

$$H_1 = -\sum_{s=1}^{p} \frac{1}{p} \int d\mathbf{r}\, P(\mathbf{r}|s) \frac{1}{\ln 2} \lim_{n \to 0} \frac{1}{n} \left\{ \left[\sum_{s'=1}^{p} \frac{1}{p} P(\mathbf{r}|s') \right]^n - 1 \right\}$$

$$= \frac{-1}{\ln 2} \lim_{n \to 0} \frac{1}{n} \left\langle \frac{H_{1a}}{p^{n+1}} - 1 \right\rangle, \tag{A2}$$

where

$$H_{1a} = \sum_{s=1}^{p} \int d\mathbf{r}\, P(\mathbf{r}|s) \times \left[\sum_{s'=1}^{p} P(\mathbf{r}|s') \right]^n \sum_{s_1=1}^{p} \cdots \sum_{s_{n+1}=1}^{p} \prod_{k=1}^{N} H_{1b}(j), \tag{A3}$$

and

$$H_{1b}(j) = \frac{1}{(2\pi\sigma^2)^{(n+1)/2}} \int dr_j \exp\left[-\frac{1}{2\sigma^2} \sum_{k=1}^{n+1} (r_j - r_j^{s_k})^2 \right] \tag{A4}$$

is a factor that depends on the j-th component of one particular way of distributing p stimuli among the $n+1$ replicas. To calculate it we observe that

$$\sum_{k=1}^{n+1} (r_j - r_j^{s_k})^2 = (n+1) \left[r_j - \frac{1}{n+1} \sum_{\ell=1}^{n+1} r_j^{s_\ell} \right]^2 + \vec{\xi}_j^{\dagger} A \vec{\xi}_j, \tag{A5}$$

where $\vec{\xi}_j$ is a vector of $n+1$ components such that $\xi_j^k = r_j^{s_k}$. The vector notation is used for arrays of $n+1$ components. The matrix A has dimensions $(n+1) \times (n+1)$, and reads

$$A = I_d - \frac{1}{n+1} U, \tag{A6}$$

where U is an $(n+1) \times (n+1)$ matrix, with all its coefficients equal to unity.

Thus, the quadratic factor in Eq. (A5) can be extracted outside the integral in Eq. (A4), and

$$H_{1b}(j)=(2\pi\sigma^2)^{-(n+1)/2}\sqrt{\frac{2\pi\sigma^2}{n+1}}\exp\left(\frac{-1}{2\sigma^2}\vec{\xi}_j^{\dagger}A\vec{\xi}_j\right). \tag{A7}$$

Replacing this expression in Eq. (A3)

$$H_{1a}=[\sqrt{n+1}(2\pi\sigma^2)^{n/2}]^{-N}\times\sum_{s_1=1}^{p}\cdots\sum_{s_{n+1}=1}^{p}\exp\left(\frac{-1}{2\sigma^2}\sum_{j=1}^{N}\vec{\xi}_j^{\dagger}A\vec{\xi}_j\right). \tag{A8}$$

We now re-arrange the summation in Eq. (A8), according to the number d of *different* stimuli appearing in the $n+1$ replicas. For each realization of $s_1,s_2,\ldots,s_{n+1}$, the replicas can be divided in d classes, such that all the replicas belonging to the same class are associated to the same stimulus, and replicas of different classes correspond to different stimuli. The number of replicas adscribed to stimulus j is K_j. Clearly, the sum of all the K_j is $n+1$, and only d of the K_i are different from zero. Therefore,

$$\sum_{s_1=1}^{p}\sum_{s_2=1}^{p}\cdots\sum_{s_{n+1}=1}^{p}=\sum_{\{K\}}\binom{n+1}{\{K\}}. \tag{A9}$$

where the term in brackets is defined in Eq. (16), and the $(n+1)$-fold summation involves all possible sets of $K_1,\ldots,K_p$ ranging from 0 to $n+1$, and whose total sum is $n+1$.

The advantage of this rearrangement is that the exponent in Eq. (A8) can be written as a function of only the differences between representations, namely

$$\vec{\xi}_j^{\dagger}A\vec{\xi}_j=\frac{1}{2(n+1)}\sum_{m=1}^{p}\sum_{\ell=1}^{p}K_mK_\ell(r_j^m-r_j^\ell)^2. \tag{A10}$$

Therefore, replacing Eqs. (A9) and (A10) in Eq. (A3) we arrive at Eq. (33).

APPENDIX B: INITIAL RISE OF $\langle I(N)\rangle$ IN THE LARGE NOISE LIMIT

Replacing Eqs. (40) in (38), we get

$$\langle H_1\rangle=-\frac{1}{\ln 2}\lim_{n\to 0}\frac{1}{n}\left\{\frac{1}{(n+1)^{M/2}(2\pi\sigma^2)^{Nn/2}p^{n+1}}\times\left[p^{n+1}-\frac{\lambda^2N}{4(n+1)\sigma^2}S\right]-1\right\}, \tag{B1}$$

with

$$S=\sum_{\{K\}}\binom{n+1}{\{K\}}\sum_{m=1}^{p}\sum_{\ell=1,\ell\neq m}^{p}K_mK_\ell. \tag{B2}$$

In order to compute S we interchange the order of summation

$$S=\sum_{m=1}^{p}\sum_{\ell=1,\ell\neq m}^{p}\sum_{\{K\}}\binom{n+1}{\{K\}}K_mK_\ell. \tag{B3}$$

The terms with K_ℓ or K_m equal to zero do not contribute to S. Therefore, we can restrict the sum in Eq. (B3) to $K_m\neq 0\neq K_\ell$. Thus, the addition over all K's ranging from 0 to $n+1$ whose total sum is $n+1$ can be replaced by another addition, where all K's different from K_ℓ and K_m range from 0 to $n-1$, K_m and K_ℓ go from 1 to n, and the sum of all the K's is $n+1$. Since there are $p(p-1)$ choices for K_ℓ and K_m,

$$S=p(p-1)(n+1)np^{n-1}. \tag{B4}$$

Replacing Eq. (B4) in Eq. (B1) we get

$$\langle H_1\rangle=\frac{1}{\ln 2}\left[\frac{N}{2}+\frac{N}{2}\ln(2\pi\sigma^2)\right]+\frac{N}{\ln 2}\frac{p-1}{p}\left(\frac{\lambda}{2\sigma}\right)^2. \tag{B5}$$

When H_2 is summed to $\langle H_1\rangle$, Eq. (44) is obtained.

APPENDIX C: THE SMALL σ LIMIT

We go back to Eq. (38). We re-write Eq. (39) as

$$\langle A_{\{K\}}\rangle=\int\prod_{s=1}^{p}dr^s\rho(r^s)\exp\left[-\frac{1}{2(n+1)\sigma^2}\overline{\chi}^{\dagger}M\overline{\chi}\right], \tag{C1}$$

where $\overline{\chi}$ is a vector of p components, such that $\chi_s=r^s$, and

$$M=(n+1)\begin{pmatrix}K_1 & 0 & \ldots & 0\\ 0 & K_2 & \ldots & 0\\ & & \ldots & \\ 0 & 0 & \ldots & K_p\end{pmatrix}-\begin{pmatrix}K_1K_1 & K_1K_2 & \ldots & K_1K_p\\ K_2K_1 & K_2K_2 & \ldots & K_2K_p\\ & & \ldots & \\ K_pK_1 & K_pK_2 & \ldots & K_pK_p\end{pmatrix}. \tag{C2}$$

The integrand in Eq. (C1) is 1 in the origin, and also along the eigenvectors of M corresponding to a zero eigenvalue. The number of such eigenvalues is equal or larger than the number of K that are zero. We therefore re-arrange the numbering of the patterns in such a way as to put all those with K different from zero in the first d places. Thus, $K_{d+1}=K_{d+2}=\cdots=K_p=0$. With this ordering, matrix M is filled with zeros in all those positions with a row or a column greater than d. Integrating in $r^{d+1},r^{d+2},\ldots,r^p$ we get

$$\langle A_{\{K\}}\rangle = \int \prod_{s=1}^{d} dr^s \rho(r^s) \exp\left[-\frac{1}{2(n+1)\sigma^2} \overline{\chi'}^{\dagger} M' \overline{\chi'} \right], \tag{C3}$$

where $\overline{\chi'}$ and M' are defined as $\overline{\chi}$ and M, but live in a space of d dimensions (and not p).

In order to integrate Eq. (C3) we observe that M' has a single eigenvalue λ_1 equal to zero, with eigenvector

$$w_1 = \frac{1}{\sqrt{d}} \begin{pmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{pmatrix}. \tag{C4}$$

We call $w_2, \ldots, w_d$ all the other eigenvectors corresponding to non-vanishing eigenvalues $\lambda_2, \ldots, \lambda_d$. We choose the eigenvectors normalized, and orthogonal to each other and to w_1 (the symmetry of M allows us to do so). With this set of vectors we construct a new basis, and call $\overline{w}$ the collection of coordinates in this new system. We define a matrix C as the change of basis

$$\overline{\chi} = C\overline{w}, \tag{C5}$$

where

$$C = \begin{pmatrix} 1/\sqrt{d} & c_{12} & \ldots & c_{1d} \\ 1/\sqrt{d} & c_{22} & \ldots & c_{2d} \\ & & \cdots & \\ 1/\sqrt{d} & c_{d2} & \ldots & c_{dd} \end{pmatrix} \tag{C6}$$

and $\det(C) = 1$. In this new basis,

$$\langle A_{\{K\}}\rangle = \int \prod_{j=1}^{d} dw_j \rho(w_1/\sqrt{d} + c_{j2}w_2 + \cdots + c_{jd}w_d) \times \exp\left[-\frac{1}{2} \sum_{\ell=1}^{d} \frac{\lambda_\ell}{\sigma^2(n+1)} w_\ell^2 \right]. \tag{C7}$$

Multiplying and dividing by the product of all $2\pi(n+1)\sigma^2/\lambda_\ell$, for $\ell \in [2,d]$, we get

$$\langle A_{\{K\}}\rangle = \left(\prod_{j=2}^{d} \sqrt{\frac{2\pi\sigma^2(n+1)}{\lambda_j}} \right) \int \prod_{j=1}^{d} [dw_j \rho(w_1/\sqrt{d} + c_{j2}w_2 + \cdots + c_{jd}w_d)] \times \frac{\exp\left[-\frac{1}{2} \sum_{k=1}^{d} \frac{\lambda_k}{\sigma^2(n+1)} w_k^2 \right]}{\prod_{j=2}^{d} \sqrt{\frac{2\pi\sigma^2(n+1)}{\lambda_j}}}. \tag{C8}$$

In the limit $\sigma \to 0$, the integrand in Eq. (C8) includes $d-1$ delta functions. Once integrated,

$$\lim_{\sigma\to 0} \langle A_{\{K\}}\rangle = \prod_{j=2}^{d} \sqrt{\frac{2\pi\sigma^2(n+1)}{\lambda_j}} \int dw_1 [\rho(w_1/\sqrt{d})]^d. \tag{C9}$$

It may be shown that

$$\prod_{j=2}^{d} \lambda_j = d(n+1)^{d-2} \prod_{\ell=1}^{d} K_\ell. \tag{C10}$$

Thus,

$$\lim_{\sigma\to 0} \langle A_{\{K\}}\rangle = (2\pi\sigma^2)^{(d-1)/2} \frac{(n+1)^{1/2}}{\sqrt{d \prod_{\ell=1}^{d} K_\ell}} B_d, \tag{C11}$$

where

$$B_d = \int dx [\rho(x)]^d. \tag{C12}$$

We now turn to the calculation of

$$S = \sum_{\{K\}} \binom{n+1}{\{K\}} \langle A_{\{K\}}\rangle^N, \tag{C13}$$

where, as before, the summation runs over all sets of $\{K\}$ that add up to $n+1$. Equation (C11) states that $\langle A_{\{K\}}\rangle$ depends on d, that is, on the number of K that are different from zero. Therefore, we write the sum in Eq. (C13) as

$$S = \sum_{d=1}^{n+1} \frac{1}{d!} \frac{p!}{(p-d)!} [(2\pi\sigma^2)^{(d-1)/2} \sqrt{n+1} B_d]^N S_1, \tag{C14}$$

where

$$S_1 = \sum_{\{K\}'} \binom{n+1}{\{K\}} \left[\frac{1}{\prod_{j=1}^{d} K_j} \right]^{N/2}. \tag{C15}$$

The sum in S_1 involves only the d values of K that are different from zero. We now make the approximation

$$S_1 \approx \left(\frac{d}{n+1} \right)^{Nd/2} \sum_{\{K\}'} \binom{n+1}{\{K\}}. \tag{C16}$$

But

$$\sum_{\{K\}'} \binom{n+1}{\{K\}} = \sum_{j=0}^{d} (-1)^j \frac{d!}{d!(d-j)!} (d-j)^{n+1}. \tag{C17}$$

And, taking the limit

$$\lim_{n\to 0} \sum_{\{K\}'} \binom{n+1}{\{K\}} = \delta_{d,1} + n \sum_{j=0}^{d} \frac{d!}{(d-j)!j!} j \ln j (-1)^{d-j}. \tag{C18}$$

Moreover, if N is large, as d grows $(B_d)^N \to 0$. Therefore, keeping just $d=1$ and $d=2$ we may approximate

$$S \approx p + p(p-1)\ln 2 (2\pi\sigma^2)^{N/2}(n+1)^{N/2}(B_2)^N n. \tag{C19}$$

Replacing in Eqs. (38) and (29) we arrive at Eq. (50).

APPENDIX D: INFORMATION BETWEEN THE ACTUAL RESPONSE AND THE STORED REPRESENTATION

The aim is to calculate Eq. (52) under the assumption (28). Replacing Eq. (28) in Eq. (54) the probability $P(\mathbf{r})$ can be written as

$$P(\mathbf{r}) = \prod_{j=1}^{N} \zeta(r_j), \tag{D1}$$

where

$$\zeta(r_j) = \int dr_j^0 P(r_j^0) \frac{e^{-(r_j - r_j^0)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}. \tag{D2}$$

Just as before, we separate

$$\tilde{I} = H_1 - H_2, \tag{D3}$$

where

$$\begin{aligned} H_2 &= -\int d\mathbf{r}^0 \int d\mathbf{r}\, P(\mathbf{r}|\mathbf{r}^0) P(\mathbf{r}^0) \log_2[P(\mathbf{r}|\mathbf{r}^0)] \\ &= \frac{N}{2\ln 2}[1 + \ln(2\pi\sigma^2)], \end{aligned}$$

$$\begin{aligned} H_1 &= -\int d\mathbf{r}^0 \int d\mathbf{r}\, P(\mathbf{r}|\mathbf{r}^0) P(\mathbf{r}^0) \log_2[P(\mathbf{r})] \\ &= -N \int dt\, \zeta(t) \log_2[\zeta(t)]. \end{aligned} \tag{D4}$$

Inserting the definition (D2) of $\zeta(t)$, and using the expression (A1) for the logarithm we get

$$\begin{aligned} H_1 = &-\frac{N}{\ln 2} \lim_{n\to 0} \frac{1}{n} \int dt \prod_{j=1}^{n+1} \frac{e^{-(t-x_j)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} \\ &- \int dx P(x) \int dt \frac{e^{-(x-t)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}. \end{aligned} \tag{D5}$$

The last term in Eq. (D5) is nothing but the integral of $\zeta(x)$ over all x, which can be shown to give 1. To carry out the integral in t in the first line of Eq. (D5) we observe that

$$\begin{aligned} \prod_{j=1}^{n+1} \frac{e^{-(t-x_j)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} &= (2\pi\sigma^2)^{-(n+1)/2} \\ &\times \exp\left\{ \frac{-(n+1)}{2\sigma^2} \left[t - \frac{1}{n+1} \sum_{j=1}^{n+1} x_j \right]^2 \right. \\ &\left. - \frac{1}{2\sigma^2} \left[\sum_{j=0}^{n+1} x_j^2 - \frac{1}{n+1} \left(\sum_{k=1}^{n+1} x_j \right)^2 \right] \right\}. \end{aligned}$$

When replacing this expression in Eq. (D5), the integration in t can be done right away. The result is $[2\pi\sigma^2/(n+1)]^{1/2}$. Therefore,

$$\begin{aligned} H_1 = &-\frac{N}{\ln 2} \lim_{n\to 0} \frac{1}{n} \left[(2\pi\sigma^2)^{-n/2} \ (n+1)^{-1/2} \right. \\ &\times \int \prod_{\ell=1}^{n+1} dx_\ell P(x_\ell) \\ &\left. \times \exp\left\{ \frac{-1}{2\sigma^2} \left[\sum_{j=1}^{n+1} x_j^2 - \frac{1}{n+1} \left(\sum_{k=1}^{n+1} x_k \right)^2 \right] \right\} - 1 \right]. \end{aligned} \tag{D6}$$

In the same way as in Eq. (A10), we write

$$\sum_{\ell=1}^{n+1} x_\ell^2 - \frac{1}{n+1} \left(\sum_{j=1}^{n+1} x_j \right)^2 = \frac{1}{2(n+1)} \sum_{\ell=1}^{n+1} \sum_{m=1}^{n+1} (x_\ell - x_m)^2. \tag{D7}$$

Thus, replacing Eq. (D7) in Eq. (D6), and making the expansion

$$\begin{aligned} &\exp\left[-\frac{1}{4\sigma^2(n+1)} \sum_{\ell=1}^{n+1} \sum_{m=1}^{n+1} (x_\ell - x_m)^2 \right] \\ &\approx 1 - \frac{1}{4\sigma^2(n+1)} \sum_{\ell=1}^{n+1} \sum_{m=1}^{n+1} (x_\ell - x_m)^2, \end{aligned} \tag{D8}$$

H_1 can be calculated. The result is

$$H_1 = \frac{N}{\ln 2} \left\{ \tfrac{1}{2}[1 + \ln(2\pi\sigma^2)] + \frac{\lambda^2}{4\sigma^2} \right\}. \tag{D9}$$

When H_2 is substracted, Eq. (55) is obtained. It should be noticed that Eq. (D8) is not an approximation. The j-th order in the Taylor expansion of the exponential grows as $[n(n+1)]^j$. Therefore, only the linear term gives a contribution for $n \to 0$.

[1] L. M. Optican and B. J. Richmond, J. Neurophysiol. **57**, 162 (1987).
[2] E. N. Eskandar, B. J. Richmond, and L. M. Optican, J. Neurophysiol. **68**, 1277 (1992).
[3] T. W. Kjaer, J. A. Hertz, and B. J. Richmond, J. Comput. Neurosci. **1**, 109 (1994).
[4] J. Heller, J. A. Hertz, T. W. Kjaer, and B. J. Richmond, Comput. Neurosci. **2**, 175 (1995).
[5] M. J. Tovée, E. T. Rolls, A. Treves, and R. J. Bellis, J. Neurophysiol. **70**, 640 (1993).
[6] E. T. Rolls, A. Treves, R. G. Robertson P. Georges-Francois, and S. Panzeri, J. Neurophysiol. **79**, 1797 (1998).
[7] E. T. Rolls, H. D. Critchley, and A. Treves, J. Neurophysiol. **75**, 1982 (1996).
[8] E. T. Rolls, A. Treves, and M. J. Tovee, Exp. Brain Res. **114**, 149 (1997).
[9] A. Treves, W. E. Skaggs, and C. A. Barnes, Hippocampus **6**, 666 (1996).
[10] A. Treves, BioSystems **40**, 189 (1997).
[11] E. T. Rolls and A. Treves, *Neural Networks and Brain Function* (Oxford University Press, Oxford, 1998).
[12] C. E. Shannon, AT&T Tech. J. **27**, 379 (1948).
[13] T. J. Gawne and B. J. Richmond, J. Neurosci. **7**, 2758 (1993).
[14] I. Samengo, Network (to be published).
[15] H. B. Barlow, Perception **1**, 371 (1972).
[16] S. W. Kuffler, J. Neurophysiol. **16**, 37 (1953).
[17] J. O'Keefe and J. Dostrovsky, Brain Res. **34**, 171 (1971).
[18] A. Treves, Phys. Rev. A **42**, 2418 (1990).
[19] A. Treves and E. T. Rolls, Network **2**, 371 (1991).
[20] A. Treves, J. Comput. Neurosci. **2**, 259 (1995).
[21] S. R. Schulz and E. T. Rolls, Hippocampus **9**, 582 (1999).