The minimum information principle and its application to neural code analysis

Amir Globerson [*], **Eran Stark** [†], **Eilon Vaadia** [† ‡], **and Naftali Tishby** [‡ §]

[*]Computer Science and Artificial Intelligence Laboratory,Massachusetts Institute of Technology, Cambridge, MA 02139, USA,[†]Department of Physiology, Hadassah Medical School, Hebrew University, Jerusalem 91120, Israel,[‡]The Interdisciplinary Center for Neural Computation, The Hebrew University, Jerusalem 91904, Israel, and [§] School of Computer Science and Engineering, The Hebrew University, Jerusalem 91904, Israel

Submitted to Proceedings of the National Academy of Sciences of the United States of America

The study of complex information processing systems requires appropriate theoretical tools to help unravel their underlying design principles. Information theory is one such tool, and has been utilized extensively in the study of the neural code. Although much progress has been made in information theoretic methodology, there is still no satisfying answer for the question: "what is the information that a given property of the neural response (e.g., the responses of single cells) carry about a set of stimuli". Here we answer this question via a novel approach, the Minimum Mutual Information (MinMI) principle. We quantify the information in any statistical property of the neural response by considering all hypothetical neuronal populations which have the given property, and finding the one that contains the minimum information about the stimuli. All systems with higher information values necessarily contain additional information processing mechanisms, and thus the minimum captures the information related to the given property alone. MinMI may be used to measure information in properties of the neural response, such as that conveyed by responses of small subsets of cells (e.g., singles or pairs) in a large population, and cooperative effects between subunits in networks. We show how the new framework can be used to study neural coding in large populations and to reveal properties that are not discovered by other information theoretic methods.

Neural Coding | Information Theory | Maximum Entropy | Population Coding

Some of the greatest challenges to science today involve complex systems, such as the brain and gene regulatory networks. Such systems are characterized by a very large number of interacting units that potentially cooperate in complex ways to produce ordered behavior. Some of the more interesting systems may be viewed, to a certain degree, as input-output systems. The brain, for example, receives multiple inputs from the environment and processes them to generate behavior. In order to obtain insight into such systems, this transmission of information needs to be quantified. An attractive mathematical tool in this context is Information Theory (IT), introduced by Claude Shannon in his mathematical theory of communication [1]. IT has been used in neuroscience since the 1950s [2], yielding insights into design principles in neural coding [3, 4], and offering new methods for analyzing data obtained in neurophysiological experiments [5]. The main information theoretic measure used in the literature is the mutual information (MI), which quantifies the level of dependence between two variables. Experimental works typically employ IT by studying the MI between aspects of the external world (e.g. motor activity [6] or visual stimuli [7]) and aspects of the neural response (e.g. spike counts [8] or precise spike times [9] among others).

Empirical studies of complex systems in general and information theoretic analyses in particular are fundamentally limited by the fact that the space of possible system states is extremely large. Thus any measurement of the system is bound to be limited and reveal only a subset of its possible states. For example, it is not practical to fully characterize the statistics of a 100 ms spike train of even a single neuron, because of its high dimensionality (2^{100} for $1ms$ precision) and the limited number of experimental trials. The problem of limited measurement is more acute for multiple neurons due to two reasons. First, the dimension of the response space grows exponentially with the number of neurons. Second, neurons are often not recorded simultaneously but rather over several recording sessions, so their joint statistics are not accessible. However, it is possible to reliably estimate partial measurements, or statistics, of the system, such as firing rates of single neurons or correlations between pairs of neurons.

Here, we present a new framework for extending information theoretic analysis to handle partial measurements of complex systems. At the basis of our approach is the assumption that the partial measurements hold for the true underlying system, whose complete characterization cannot be accessed. We next consider all hypothetical systems that are consistent with the observed partial measurements. Clearly, there is a large set of such systems, each with its own value of mutual information between input and output. Our goal is to find the value of information that can be attributed only to the given measurements. Intuitively, the systems with relatively high mutual information in the hypothetical set have some additional structure which cannot be inferred based on the given partial measurements. However, the system with minimum information in this set cannot be simplified further (in the MI sense) and its information can thus be taken to reflect the information available in the given measurements alone. Our minimum information (MinMI) principle thus states that given a set of measurements of a system, the MI available in these measurements is the minimum mutual information between input and output in any system consistent with the given measurements. An immediate implication of the above construction is that this minimum information is a lower bound on the information in the true underlying system, since the true system is also in the set we minimized over.

A conceptual tool which has previously been used to tackle partial measurements is the Maximum Entropy (MaxEnt) principle [10, 11, 12]. While MaxEnt may be quite successful in modeling response distributions [12], the MinMI principle is more appropriate for handling input-output systems where MI rather than entropy is the measure of interest. Furthermore, MinMI offers a bound on the information in the true system, whereas MaxEnt does not.

The minimum information bound

Consider a system with discrete input stimulus $S \in \{1, \ldots, n_s\}$ and output response $R \in \{1, \ldots, n_r\}$ (our formalism applies to continu-

Abbreviations: CI, conditionally independent, IT, information theory, MI, mutual information, MinMI, minimum mutual information, MaxEnt, maximum entropy, RDT, rate distortion theory

ous variables as well; we focus on discrete ones for presentation purposes). The mutual information (MI) between S and R is a measure of the dependence between these two variables. Denote by $p(s,r)$ the joint distribution of S and R. This distribution fully characterizes the input-output relations in the system. The mutual information between S and R is defined as

$$I_p(S;R) \equiv \sum_{s,r} p(s,r) \log \frac{p(s,r)}{p(s)p(r)}, \qquad [1]$$

where $p(s), p(r)$ are the marginal distributions of S, R [13]. The MI is zero if and only if the variables are independent. High MI indicates that the response R encodes properties of the stimulus S.

We focus on the case where $p(s,r)$ is not known. Rather, we have access to partial measurements of it, given by the expected value of some function of R given S. Formally, we consider a set of d functions $\vec{\phi} : \{1,\ldots,n_r\} \rightarrow \mathbb{R}^d$, and assume we know their expected values given S. We denote these expected values by $\vec{a}(s)$, so that

$$\vec{a}(s) \equiv \langle \vec{\phi}(r) \rangle_{p(r|s)}, \qquad [2]$$

where the expectation operator $\langle\rangle$ is defined by $\langle f(x) \rangle_{p(x)} = \sum_x f(x)p(x)$. For example, such expected values may be the firing rates of individual neurons given a stimulus s. The expected values are typically estimated from experimental data. Denote the experimental data by the set of pairs $(s_1,r_1),\ldots,(s_n,r_n)$. Then $\vec{a}(s)$ is estimated by $\frac{1}{m_s}\sum_{i:s_i=s} \vec{\phi}(r_i)$, where m_s is the number of data pairs where $s_i = s$. Due to finite sample effects, this empirical estimate will typically not equal the true expected value. In what follows, we shall assume that these values are identical. However, our approach may be extended to the cases where the expected values are known up to some confidence interval. We further assume that the prior probabilities of the stimulus variable are known, and denote these by $p(s)$. This assumption is reasonable since these probabilities are usually determined by the experimentalist, and can be estimated by $\frac{m_s}{\sum_{\hat{s}} m_{\hat{s}}}$. The above formalism encompasses a wide range of response characteristics, from the response of single neurons, through that of neurons over time, to joint statistics of any order.

To bound the information in our system from below, we consider all hypothetical systems (joint distributions) which could yield the given partial measurements. These systems are defined by the set of distributions which yield expected values of $\vec{\phi}(r)$ that are equal to the measured ones

$$\mathcal{P}(\vec{a}(s),p(s)) \equiv \left\{ \hat{p}(r,s) : \begin{array}{cc} \langle \vec{\phi}(r) \rangle_{\hat{p}(r|s)} = \vec{a}(s) & \forall s \\ \hat{p}(s) = p(s) & \forall s \end{array} \right\}. \qquad [3]$$

The true underlying distribution $p(s,r)$ is clearly in $\mathcal{P}(\vec{a}(s),p(s))$. Thus, the true underlying information $I_p(R;S)$ is lower bounded by the minimum information attainable in $\mathcal{P}(\vec{a}(s),p(s))$

$$I_p(R;S) \geq I_{min}\left[\vec{\phi}(r),\vec{a}(s)\right] \equiv \min_{\hat{p}\in\mathcal{P}(\vec{a}(s),p(s))} I_{\hat{p}}(R;S), \qquad [4]$$

where we have used $I_{min}\left[\vec{\phi}(r),\vec{a}(s)\right]$ to denote the minimum information value. The constrained optimization problem of Equation 4 is solved by introducing a set Lagrange multipliers $\vec{\psi}(s) \in \mathbb{R}^d$ (one vector per stimulus value), yielding a solution of the form

$$\hat{p}_{MI}(r|s) = \hat{p}_{MI}(r) e^{\vec{\phi}(r)\cdot\vec{\psi}(s)+\gamma(s)}, \qquad [5]$$

where $\gamma(s) = -\log\sum_r \hat{p}_{MI}(r) e^{\vec{\phi}(r)\cdot\vec{\psi}(s)}$ is a normalization factor. Note that the distribution $\hat{p}_{MI}(r)$ depends on $\hat{p}_{MI}(r|s)$ through marginalization: $\hat{p}_{MI}(r) = \sum_s p(s)\hat{p}_{MI}(r|s)$. Thus, the above characterization is not a closed form solution, but rather a set of equations involving $\hat{p}_{MI}(r|s)$. The parameters $\vec{\psi}(s)$ should be chosen to satisfy the constraints $\mathcal{P}(\vec{a}(s),p(s))$. Although there is no closed form solution for $\vec{\psi}(s)$, it can be calculated using iterative algorithms as shown in the Methods section.

In the above formulation, $\vec{\phi}(r)$ may be any function of the response space. In analyzing neural codes, we shall be specifically interested in the case where the expected values are k^{th} order marginals of the true distribution $p(r|s)$. We denote the minimum information given the set of all k^{th} order marginals by $I^{(k)}$ (see Supporting Inf., Section 1). In what follows, we shall specifically demonstrate how $I^{(1)}$ and $I^{(2)}$ may be used to study aspects of neural coding.

It is interesting to contrast the MinMI solution with that of MaxEnt (see Supporting Inf., Section 2). The key difference between the methods may be studied using a simple example. Consider a set of N binary neurons with the same first order responses $p(r_i|s)$ to two stimuli: $p(r_i = 1|s = 1) = \alpha$ and $p(r_i = 1|s = 2) = \beta$, and assume $p(s) = 0.5$. Clearly the information minimizing distribution is one where neurons are completely correlated (i.e., all fire or don't fire simultaneously). Thus the MinMI information will equal the information in a single neuron. On the other hand, the MaxEnt distribution in this case will correspond to neurons being conditionally independent (CI) given the stimulus. As $N \rightarrow \infty$, the information in the MaxEnt distribution will approach one (as long as $\alpha \neq \beta$), since an observer of the response R will be able to perfectly predict the identity of the stimulus S by averaging over the N neurons to obtain (with probability 1) the values α, β for $s = 1, 2$. Thus the MaxEnt approach becomes inadequate for measuring information in large populations, whereas the MinMI approach does not have this limitation. This difference will be illustrated in the experiments reported below.

Synergy and Redundancy Measures

A key issue in neural coding is the importance of high order statistics, and their contribution with respect to lower order statistics. One approach to quantifying this contribution is to compare the MI in a model based on higher order statistics, to one based on lower order statistics. A positive difference indicates synergy: information in higher order interactions, while a negative difference indicates redundancy. Several such measures have been suggested in the literature [14, 8, 15, 16]. One measure, previously studied in [8], is defined as:

$$SynSum(R;S) \equiv I_p(R;S) - \sum I_p(R_i;S). \qquad [6]$$

It measures the difference between the full information and the sum of individual (first order) informations. One shortcoming of the above measure is that the second term becomes dominant as N grows (the first is always bounded by $H(S)$). Thus, large populations will always appear redundant. Another possible measure compares the full information to the information in the case where neurons are conditionally independent (CI) given the stimulus

$$SynCI(R;S) \equiv I_p(R;S) - I_{p_{CI}}(R;S), \qquad [7]$$

where $p_{CI}(r|s) = \prod_{i=1}^n p(r_i|s)$ (SynCI was denoted by ΔI_{noise} in [15]). Note that this measure does not grow with N and will equal zero when the neurons are CI. Another related measure based on the CI case, but not directly using information, was introduced in [16].

Both $SynSum$ and $SynCI$ compare the full information to that in first order statistics. Moreover, the typical implementation of these measures is for the two neuron case, where the only statistics less than full order are first order. The generalization of synergy/redundancy measures to higher order statistics, and to $N > 2$ populations poses an important challenge. The $SynSum$ measure has been generalized

to this scenario in [17], where it was decomposed into elements measuring synergy in k^{th} order correlations. MinMI offers an elegant approach for generalizing the $SynCI$ measure to higher orders. At first sight, it seems like a reasonable approach is to take difference between the informations of the MaxEnt distributions for orders k and $k-1$. However, these two numbers will saturate as $N \to \infty$ (see Supporting Inf., Section 2), and thus this measure will be zero at the limit. MinMI offers a way around this problem, as we now illustrate for second order statistics. The $I^{(2)}$ measure quantifies the information available in a population given only its (first and) second order statistics. To turn it into a synergy/redundancy measure, we need to subtract the second order information in the CI model. If the neurons are CI, the pairwise statistics are expected to be $p(r_i, r_j|s) = p(r_i|s)p(r_j|s)$. We denote the minimum information in these pairwise statistics by $I^{(2)}_{CI}$. A natural measure of synergy is then the difference

$$SynI^{(2)}(R_1, \ldots, R_N, S) = I^{(2)} - I^{(2)}_{CI}. \qquad [8]$$

When the true population is CI, we have $SynI^{(2)} = 0$, as expected. Furthermore, when $N = 2$, we have that $SynI^{(2)} = SynCI$. Thus MinMI generalizes $SynCI$ to the study of pairwise interactions in large populations. Furthermore, the MinMI information does not saturate as the MaxEnt ones, and thus this measure is meaningful even as $N \to \infty$. The $SynI^{(2)}$ measure may be extended to the k^{th} order case by replacing $I^{(2)}_{CI}$ with the minimum information subject to k^{th} order statistics given by a MaxEnt model of order $k-1$.

Results

Neural population codes may be studied at several levels, corresponding to different coding strategies. The basic level is the single neuron code. Next is the relation between the codes of different single neurons. Higher order interactions between neurons constitute yet another level, along with the relation between multiple higher-order interactions. Finally, temporal structure may also be used to enhance coding efficiency. In the applications below, we show how the MinMI principle may be applied to the study of various neural coding schemes and quantify the level to which different populations use these schemes.

Two binary neurons and a binary stimulus. We begin with an illustration of MinMI calculation for the case of two toy binary neurons R_1, R_2 where each neuron has two possible responses: 0 or 1. The stimulus S is also taken to be binary. We assume that only first order statistics $p(r_1|s), p(r_2|s), p(s)$ are known and that $p(r_1, r_2|s)$ is unknown. We are interested in the minimum information $I^{(1)}$, i.e., the information available in a distribution $\hat{p}(r_1, r_2, s)$ satisfying the first order constraints $\hat{p}(r_i|s) = p(r_i|s)$, $i = 1, 2$. Note that any such distribution is completely defined by two numbers $\hat{p}(r_1 = 1, r_2 = 1|s)$ (for $s = 1, 2$), since for each S value $\hat{p}(r_1, r_2|s)$ has four free parameters and has to satisfy three constraints (two first order constraints and one normalization constraint). In this specific case, the space of possible distributions $\hat{p}(s, r)$ can be visualized in two dimensions, as in Figure 1. The figure shows the value of the MI for each possible distribution in $\hat{p}(s, r)$ satisfying the constraints above. This is done for two different pairs of neurons, with different first order responses. The figure shows (in yellow circles) the location of the MinMI distribution $\hat{p}_{MI}(r_1, r_2|s)$. Also shown (in white squares) is the distribution under which the neurons are CI given the stimulus: $\hat{p}(r_1, r_2|s) = \hat{p}(r_1|s)\hat{p}(r_2|s)$. By definition, this distribution has higher mutual information than $I^{(1)}$. In the first example (Figure 1A-C) the two neurons have the same response distributions $p(r_1|s) = p(r_2|s)$. The MinMI distribution shown in the figure is

A

s	$p(r_1 = 1)$	$p(r_2 = 1)$
1	0.2	0.2
2	0.6	0.6

B

s	$[0, 0]$	$[0, 1]$	$[1, 0]$	$[1, 1]$
1	0.8	0	0	0.2
2	0.4	0	0	0.6

D

s	$p(r_1 = 1)$	$p(r_2 = 1)$
1	0.4	0.6
2	0.8	0.7

E

s	$[0, 0]$	$[0, 1]$	$[1, 0]$	$[1, 1]$
1	0.3	0.3	0.1	0.3
2	0.1	0.1	0.2	0.6

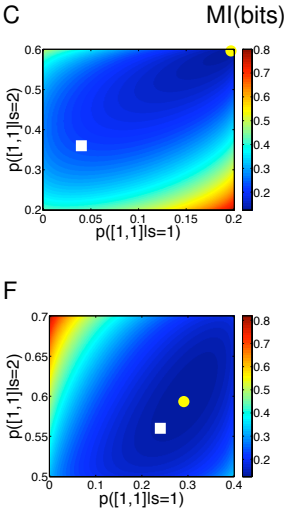

Fig. 1. Illustration of $I^{(1)}$ for two binary neurons and a binary stimulus ($p(s) = 0.5$). Only the first order statistics of each neuron are assumed to be known. The results for two different first order statistics are shown. Panels A,D give the first order statistics in the two cases. Panels B,E show the minimum information distribution $\hat{p}_{MI}(r|s)$ for the statistics in panels A,D respectively. Panels C,F show the information in all distributions satisfying with the given first order statistics in panels A,D respectively. The yellow dot shows the location of the MinMI distribution in this information plane, and the white square shows the CI distribution. The X and Y axes measure the probability of both neurons firing for stimuli $s = 1, 2$. Note that these two parameters specify the entire response distribution, given the constraints on single neuron responses.

then the one in which the neurons are completely correlated, and thus lies on the boundary of the space of possible distributions. It is intuitively clear why this is the minimum: the two neurons are, in the worst case, equivalent to a single neuron. In this case the CI information is higher since when the two neurons are CI, one can average over the noise to obtain more information about the stimulus.

In contrast, when the two neurons differ in their response distri-

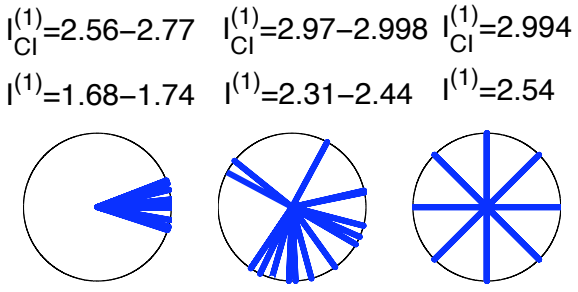

Fig. 2. The information $I^{(1)}$ for different population coding schemes. We consider three populations of 16 neurons responding to eight stimuli. The stimuli correspond to eight equally spaced directions on the circle ($s = \{0°, 45°, \ldots, 315°\}$). All neurons are cosine tuned with PDs given in the polar plots ($p(r_i|s) = Poiss\left(r_i|5 + 5\cos\left(s - \theta_i\right)\right)$, where $Poiss(r|\lambda)$ is probability of count r under a Poisson distribution with rate λ, and θ_i is the PD of neuron i; responses where $r_i \geq 25$ spikes are clipped $r_i = 25$). Left panel shows a setup where all neurons have similar PDs (directions were drawn uniformly in the range $\pm 22.5°$). Middle panel shows tuning to random directions. In the right panel, neurons are tuned to equally spaced directions, so that two neurons are assigned to each direction. $I^{(1)}$ and $I^{(1)}_{CI}$ values are given for each scenario (values for the overlapping and random tunings were obtained by drawing PDs 1000 times and calculating a 99% confidence interval).

A

	p_1	p_2
$[0,0]$	0.25	0.4
$[0,1]$	0.25	0.1
$[1,0]$	0.25	0.1
$[1,1]$	0.25	0.4

B $\mathbf{SynI^{(2)} = 0.07}$

s	r_1, r_2	r_3, r_4
1	p_1	p_1
2	p_2	p_2
3	p_1	p_1
4	p_2	p_2

C $\mathbf{SynI^{(2)} = 0.14}$

s	r_1, r_2	r_3, r_4
1	p_1	p_1
2	p_2	p_1
3	p_1	p_2
4	p_2	p_2

Fig. 3. Information in populations from pairwise statistics. We consider the responses of four toy neurons $r_1, \ldots, r_4$ to four stimuli $s = \{1, \ldots, 4\}$ ($p(s) = 0.25$). Neurons r_1, r_2 are conditionally independent from r_3, r_4. Panel A defines the response distributions of two single neurons. Note that the information in any single neuron is zero. Panels B and C give the response of the four neurons under two different scenarios by specifying the response of each pair. In both scenarios, pairwise synergy values ($SynSum$ and $SynCI$, which are equal in this case) are 0.07 for pairs (r_1, r_2) and (r_3, r_4) and zero for the other four pairs. However, the $SynI^{(2)}$ values for each distribution are different, as shown in the heading of panels B and C.

butions (Figure 1D-F), they cannot be completely correlated. Thus, the information minimizing distribution will not lie on the boundary as in the previous example (compare Figure 1C with Figure 1F), but will still be lower than the CI information (compare circle with square in Figure 1F).

Coding redundancy in single neurons. We next illustrate the use of MinMI in the study of single neuron codes and their combination in a population. As an example, consider a population of neurons where each neuron is tuned to some preferred direction of movement (PD) in the stimulus (e.g., the direction of hand movement in motor neurons, or stimulus motion in visual neurons). A population of neurons may have different distributions of such PDs. In one extreme, all neurons have the same PD, while in the other extreme PDs are uniformly distributed among neurons. It is intuitively clear that the second scenario is advantageous in terms of coding. However, it is not clear how to quantify this intuition in terms of information, especially when the joint distribution of the population cannot be estimated.

The MinMI principle provides a natural framework for tackling the above problem. Ideally, in studying information in populations we are interested in the quantity $I(R_1, \ldots, R_N; S)$. More specifically, we are interested in the contribution of single neuron codes to this information. Our $I^{(1)}$ measure provides precisely that. To illustrate how $I^{(1)}$ differentiates between different single neurons coding schemes, we simulate data from three hypothetical neuronal populations, with different degrees of overlap between single neuron codes. Figure 2 shows the code structure for these populations and the respective $I^{(1)}$ values. The results correspond to the intuition mentioned above: low $I^{(1)}$ values correspond to populations with high overlap between single neuron codes, and high values correspond to low overlap. Note that the MinMI calculation is model-free, and thus does not use the concept of directional tuning or preferred direction. It can thus detect differences in population coding in considerably more complex scenarios, which may be very hard to visualize. In Figure 2, we also compare the $I^{(1)}$ to the information in a distribution where neurons are CI given the stimulus (this is also the MaxEnt distribution subject to first order statistics). We denote this information by $I^{(1)}_{CI}$ (see Supporting Inf., Section 4, for $I^{(1)}_{CI}$ calculation method). The shortcoming of the CI assumption is that for large populations, the information will asymptotically reach its maximum value, since one can average over the responses of independent neurons and recover the exact stimulus as the population size grows. This behavior is apparent in the results in Figure 2, where the $I^{(1)}_{CI}$ measure approaches the maximum value of 3 bits for randomly distributed PDs, and a similarly high value for the uniformly spaced PDs. Thus, unlike $I^{(1)}$, the measure $I^{(1)}_{CI}$ does not discriminate between the different neuronal populations.

Pairwise coding in populations. Second order statistics between neurons have been shown to play a part in neural coding in the sense that their joint activity provides more information about a stimulus than their individual responses [6, 18]. Most research has been devoted to studying pairs of neurons, mostly due to sample size limitations (but see [19, 12]). It is intuitively clear, however, that if the second order statistics between all pairs in a population provide information, but about the same property of the stimulus, this should result in less information than if different pairs encoded different stimulus properties. This situation is the pairwise equivalent of the single neuron coding issue discussed in the previous section.

The information available from grouped pairwise responses in a population can be quantified using the $SynI^{(2)}$ measure. Figure 3 shows two toy populations, four neurons each, with identical pairwise statistics and therefore identical pairwise synergy values: the set of $SynSum(R_i, R_j; S)$ $(i, j \in \{1, \ldots, 4\})$ values is identical in both populations. Furthermore, in this case $SynSum = SynCI$ since $I(R_i; S) = 0$ for all neurons. In one population (Figure 3B), all synergistic coding provides information about the same property of the stimulus, whereas in the other (Figure 3C) the pairwise codes are designed to provide disparate information. The difference between these two populations is clearly seen in their $SynI^{(2)}$ values. Thus the MinMI principle can be used to differentiate between populations with different pairwise code designs. Although MaxEnt models [12] can also be applied to this case, they suffer from the same asymptotic behavior that we encountered for the $I^{(1)}_{CI}$ case, and will not be able to discriminate between different populations for large N.

Temporal Coding. Temporal response profiles of single neurons may transmit information about behaviorally relevant variables [20, 21, 22]. Intuitively, one could argue that if different behavioral parameters induce different response profiles, as measured by a peri-stimulus time histogram (PSTH), then the temporal response carries information about the variable. Our MinMI formalism allows us to make this statement explicit and to calculate the resulting information.

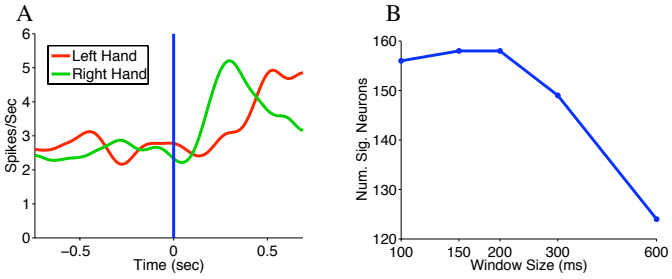

Fig. 4. Analysis of temporal coding using MinMI. Panel A shows the PSTHs of the response to the laterality signal (left hand - red, right hand - green), for a neuron recorded in the primary motor cortex. Time zero indicates the stimulus onset. The $I^{(1)}$ measure was significant for window size 200 ms and below but not for 300 ms or 600 ms. Panel B shows the number of neurons with significant $I^{(1)}$ ($p < 0.01$) as a function of the window size.

The response function of a neuron can be given by its response in a series of time bins $p(r_t|s)$, $t = 1 \dots T$. A PSTH is an example of such a profile where r_t is a binary variable, and one plots the rate function $p(r_t = 1|s)$. The responses $p(r_t|s)$ are merely a set of first order statistics and thus we can calculate $I^{(1)}$ for these statistics, so as to obtain a measure of information in a PSTH.

Figure 4 illustrates the application of MinMI to temporal coding in recordings from the primary motor cortex of behaving monkeys (see Supporting Inf., Section 5, for experimental methods and data analysis). We consider the response to a binary laterality signal (a visual stimulus), which instructs the monkey which hand to move. Figure 4 shows a PSTH of a neuron, where the total spike count over a period of 600 ms post-stimulus is similar for both conditions. However, the temporal profiles differ between the two conditions. To analyze this coding using $I^{(1)}$ we partitioned the 600 ms period into time windows of $600, 300, 200, 150, 100$ ms, and calculated $p(r_t|s)$ and the corresponding $I^{(1)}$ for each partition. We then shuffled the trials between laterality signals and compared the shuffled values to the raw $I^{(1)}$ in order to test if the raw information was significantly different from zero. For the neuron in Figure 4A, we found that it was not significant for window sizes of 300 ms and above, but was significant for all lower sized windows. This indicates that MinMI may be used to detect information related to temporal structure. We repeated the above procedure for the entire population of 827 neurons, and counted the number of significant neurons for each window size. Figure 4B shows this number as a function of window size. A large increase can be seen when moving from 600 ms to 200 ms, indicating relevant temporal structure at these time constants. The number then flattens for lower window sizes, suggesting that no information about the stimulus is added at these time scales.

Discussion

We have presented a framework for estimating MI in systems given partial measurements. Our MinMI principle has two attractive properties. The first is the ability to obtain a bound on information from a wide array of statistical observations and experimental scenarios. The second is the extension of standard information measures (such as information in single neurons or pairs of neurons) to large populations, allowing the detection of complex coding schemes not evident when analyzing neurons individually. Also, unlike previous decomposition methods (e.g., [23]) MinMI does not require knowledge of the complete joint distribution. These advantages improve on current IT-based methods in neuroscience and provide a comprehensive framework for tackling fundamental issues in neural coding.

Unlike other IT-based methods used in neuroscience, MinMI does not aim to estimate the underlying distribution directly, but rather uses the distribution as a variable to be optimized over. This in fact is the mathematical structure of the central coding theorems in information theory [13], where information is either minimized (as in the Rate Distortion theorem) or maximized (as in the Channel Capacity theorem) under some constraints on the joint distribution $p(s, r)$. Our approach is most closely related to Rate Distortion Theory (RDT), which sets the achievable limits on "lossy" compression, i.e. compression which results in some distortion of the original signal. The RDT compression bound is obtained by minimizing information with respect to a fixed $p(s)$, and a constraint on the expected distortion. In MinMI, we also fix $p(s)$ but introduce additional constraints on $p(r|s)$ via its expected values. This can be understood as searching for a distribution $p(r|s)$ as in RDT, but with the single distortion constraint replaced by multiple constraints on expected values of several functions.

Mutual information can be interpreted as a measure of the predictive power between two variables. Another measure is the optimal bayes error e^*, defined as the minimum probability of error that may be incurred in predicting the value of S from R. While e^* is a valid measure of predictability, it is not sensitive to changes in $p(s, r)$ which are intuitively related to predictive power. This is due to the fact that e^* only relies on the value of $p(s|r)$ at the maximum, and will thus not change as this maximum becomes less "sharp". MI does quantify such changes, and can thus track small, but relevant changes in the response distribution. Interestingly, there is a connection between these two measures. The MI can be used to obtain an upper bound on e^* given by $e^* \leq 0.5(H_p(S) - I_p(R; S))$, where $H_p(S)$ is the entropy of the stimulus [24]. This implies that in the MinMI case, although the distribution $p(s, r)$ is not known, we are guaranteed that $e^* \leq 0.5(H_p(S) - I_{min})$. MinMI thus yields an upper bound on the optimal prediction error in the true underlying distribution. The MinMI distribution $\hat{p}_{MI}(s|r)$ may also be used directly to predict S from R, and in fact it can be shown that the resulting error is bounded from above by $H_p(S) - I_{min}$ ([25], Section 6). Furthermore, it can also be shown that prediction via the MinMI distribution $\hat{p}_{MI}(s|r)$ minimizes a worst case bound on prediction error ([25], Section 3).

A common approach to calculating information in complex responses, is to apply some quantization to R via a function $f(r)$ (e.g., the total spike count in a spike train r), such that the quantized variable $f(R)$ has relatively few values and thus $p(s, f(r))$ may be estimated from small samples. The Data Processing Inequality [26] then states that $I_p(S; f(R)) \leq I_p(S; R)$, and thus the quantized information always provides a lower bound on the true information. It can be shown (see Supporting Inf., Section 6) that $I_p(S; f(R))$ is in fact the outcome of the following MinMI problem: what is the minimum information in a distribution $p(s, r)$ whose quantized version is $p(s, f(r))$. Thus MinMI may be viewed as generalizing the data processing inequality approach.

In presenting the method, we made the assumption that partial measurements are exact. Since these measurements are commonly estimated from finite samples, their exact values are usually not known, but rather lie in some range of values (with high probability). This range can be determined from the size of the sample via Chernoff like bounds [27]. To account for noisy measurements, the expectation constraints (Equation 3) merely need to be limited to this range. The solution in this case still has the general form of Equation 5, and corresponding algorithms may be derived.

While the results presented here were applied to neural coding, the MinMI principle is general and may be used for studying a wide array of complex systems. For instance, it may be used to estimate the information in a set of gene expression profiles about external conditions [28], and thus help in analyzing their functional role, and in comparing different gene regulatory networks.

Methods

In this section we describe algorithms for calculating the minimum information. The constrained optimization problem posed in Equation 4 is convex, since the target function $I_{\hat{p}}(S; R)$ is convex [13] and since the constraints are linear in $\hat{p}(s, r)$. It thus has no local minima and can be solved using convex optimization machinery [29]. Here we present two specialized iterative algorithms to solve this problem. The first algorithm is exact, but requires $O(n_r)$ resources. Since n_r may be large in some applications, we also present a second, approximate, algorithm to handle such cases. The basic building block of our first, exact, algorithm is the I-projection [30]. The I-projection of a distribution $q(r)$ on a set of distributions $\mathcal{F}$ is defined as the distribution $p^* \in \mathcal{F}$ which minimizes the Kullback-Leibler (KL) divergence to

the distribution $q(r)$: $p^* = \arg\min_{p\in\mathcal{F}} D_{KL}[p|q]$, where $D_{KL}[p|q]$ is defined as $\sum_r p(r)\log\frac{p(r)}{q(r)}$. The I-projection has a particularly simple form when $\mathcal{F}$ is determined by expectation constraints

$$\mathcal{F}(\vec{\phi}(r),\vec{a}) = \left\{\hat{p}(r) : \langle\vec{\phi}(r)\rangle_{\hat{p}(r)} = \vec{a}\right\}. \quad [9]$$

The I-projection is then given by

$$p^*(r) = q(r)e^{\vec{\phi}(r)\cdot\vec{\lambda}^*+\gamma^*}, \quad [10]$$

where $\vec{\lambda}^*$ are a set of Lagrange multipliers, chosen to fit the desired expected values, and $\gamma^*$is a normalization factor. The values of $\vec{\lambda}^*$ can be found using several optimization techniques. All involve the computation of the expected value of $\vec{\phi}(r)$ under distributions of the form $q(r)e^{\vec{\phi}(r)\cdot\vec{\lambda}}$. Here we use an L-BFGS based algorithm as in [31].

The structural similarity between the form of Equation 10 and the characterization of $\hat{p}_{MI}(r|s)$ in Equation 5 suggests that $\hat{p}_{MI}(r|s)$ is an I-projection of $\hat{p}_{MI}(r)$ on the set $\mathcal{F}(\vec{\phi}(r),\vec{a}(s))$. The fact that $\hat{p}_{MI}(r)$ depends on $\hat{p}_{MI}(r|s)$ through marginalization suggests that the minimization problem may be solved using an iterative algorithm where marginalization and projection are performed at each step. The iteration consists of the following steps [25]:

- For all s, set $\hat{p}_{t+1}(r|s)$ to be the I-projection of $\hat{p}_t(r)$ on $\mathcal{F}(\vec{\phi}(r),\vec{a}(s))$.
- Set $\hat{p}_{t+1}(r) = \sum_s \hat{p}_{t+1}(r|s)p(s)$.
- Calculate $I_{\hat{p}_{t+1}}(S;R)$. Stop iterating when $I_{\hat{p}_t}(S;R) - I_{\hat{p}_{t+1}}(S;R) \le \epsilon$.

The above procedure can be shown to converge to the minimum information (see convergence proof in Supporting Inf., Section 3).

The exact algorithm presented above is feasible when the size of the input space n_r is small enough to allow $O(n_r)$ memory and computational resources. For systems containing many elements, this is often not the case. For instance, when R is a response of 100 binary neurons, $n_r = 2^{100}$. To derive an approximate algorithm for the large system case, we first note that after t iterations of the exact iterative algorithm, the distribution $p_t(r)$ is a mixture of the form

$$p_t(r) = \sum_{k=1}^{(n_s)^t} c_k e^{\vec{\phi}(r)\cdot\vec{\psi}_k+\gamma_k}, \quad [11]$$

where every iteration increases the number of components by a factor of n_s. For the approximate algorithm, we limit the number of elements in this mixture to some constant K by clustering its components after each iteration using a K-means algorithm with K centroids (See [32], Chapter 10). The resulting mixture is represented using its mixing probabilities c_k and parameters $\vec{\psi}_k$ (resulting in $O(K)$ parameters). We denote the resulting approximate distribution by $\hat{p}'_t(r)$. The algorithm then proceeds as in the exact method, only with $\hat{p}'_t(r)$ instead of the exact $\hat{p}_t(r)$.

For the $I^{(1)}$ case, the k^{th} element in the mixture has the form $c_k e^{\sum_{i=1}^n \psi_k(r_i)+\gamma_k}$. Recall that in order to perform the I-projection one needs to calculate expected values for distributions of the form

$$\hat{p}'_t(r)e^{\vec{\lambda}\cdot\vec{\phi}(r)} = \sum_k c_k e^{\sum_{i=1}^n \psi_k(r_i)+\gamma_k+\lambda(r_i)}. \quad [12]$$

Because of the factorized form of each element in the sum, the first order marginals are straightforward to calculate. Thus the I-projection of $\hat{p}'_t(r)$ on the relevant constraints can be calculated, and the algorithm in the previous section may be implemented.

For the higher order cases, such as $I^{(2)}$, the marginals of the mixture do not have a closed form solution, and require approximate methods such as Markov chain Monte Carlo [33] or belief propagation [34]. For the applications presented here, we used the approximate algorithm only for the $I^{(1)}$ case. We have found empirically that the above approximation scheme works well for cases where we could compare it to the exact algorithm (up to $n_r = 50,000$). In the applications reported here, we used the exact algorithm for $n_r \le 50,000$.

1. Shannon, C. (1948) Bell System Technical Journal **27**, 379–423, 632–656.
2. Miller, G. (1956) The Psychological Review **63**, 81–97.
3. Barlow, H. (1960) in Current Problems in Animal Behaviour, eds. Thorpe, W & Zangwill, O. L. (Cambridge University Press), pp. 331–360.
4. Linsker, R. (1988) IEEE Computer **21**, 105–117.
5. Rieke, F, Warland, D, de Ruyter van Steveninck, R, & Bialek, W. (1997) Spikes. (MIT Press).
6. Hatsopoulos, N, Ojakangas, C, Paninski, L, & Donoghue, J. (1998) Proc. Natl. Acad. Sci. USA **95**, 15706–15711.
7. Bialek, W, Rieke, F, de Ruyter van Steveninck, R, & Warland, D. (1991) Science **252**, 1854–1857.
8. Gawne, T & Richmond, B. (1993) J. Neuroscience **13**, 2758–2771.
9. Dan, Y, Alonso, J, Usrey, W, & Reid, R. (1998) Nat Neurosci. **1**, 501–507.
10. Jaynes, E. (1957) Physical Review **106**, 620–630.
11. Martignon, L, Deco, G, Laskey, K, Diamond, M, Freiwald, W, & Vaadia, E. (2000) Neural Computation. **12**, 2621–2653.
12. Schneidman, E, Berry, M, Segev, R, & Bialek, W. (2006) Nature **440**, 1007–1012.
13. Cover, T & Thomas, J. (1991) Elements of information theory. (Wiley-Interscience).
14. Tononi, G, Sporns, O, & Edelman, G. (1999) Proc. Natl. Acad. Sci. **96**, 3257–3262.
15. Schneidman, E, Bialek, W, & Berry, M. (2003) J. Neuroscience **23**, 11539–11553.
16. Nirenberg, S & Latham, P. (2003) Proc. Natl. Acad. Sci. **100**, 7348–7353.
17. Schneidman, E, Still, S, Berry, M, & Bialek, W. (2003) Physical Review Letters **91**, 238701.
18. Vaadia, E, Haalman, I, Abeles, M, Bergman, H, Prut, Y, Slovin, H, & Aertsen, A. (1995) Nature **373**, 515–518.
19. Narayanan, N, Kimchi, E, & Laubach, M. (2005) J. Neuroscience **25**, 4207–4216.
20. Optican, L & Richmond, B. (1987) J. Neurophysiology **57**, 162–177.
21. Osborne, L, Bialek, W, & Lisberger, S. (2004) J. Neuroscience **24**, 3210–3222.
22. Victor, J & Purpura, K. (1996) J. Neurophysiology **76**, 1310–1326.
23. Pola, G, Thiele, A, Hoffmann, K, & Panzeri, S. (2003) Network: Comput. Neural Syst. **14**, 35–60.
24. Hellman, M & Raviv, J. (1970) IEEE Transactions on Information Theory **16**, 368–372.
25. Globerson, A & Tishby, N. (2004) in Proceedings of the 20th conference on Uncertainty in artificial intelligence, eds. Chickering, M & Halpern, J. (AUAI Press, Arlington, VA), pp. 193–200.
26. Borst, A & Theunissen, F. (1999) Nature Neuroscience **2**, 947–957.
27. Dudík, M, Phillips, S, & Schapire, R. E. (2004) in 17th Annual Conference on Computational Learning Theory, eds. Shawe-Taylor, J & Singer, Y. (Springer), pp. 472–486.
28. Gasch, A, Spellman, P, Kao, C, Carmel-Harel, O, Eisen, M, Storz, G, D, D. B, & Brown, P. (2000) Mol Biol Cell. **11**, 4241–4257.
29. Boyd, S & Vandenberghe, L. (2004) Convex Optimization. (Cambridge Univ. Press).
30. Csiszar, I. (1975) Annals of Probability **3**, 146–158.
31. Sha, F & Pereira, F. (2003) in Proceedings of the conference on Human Language Technology - NAACL. (ACL), pp. 134–141.
32. Duda, R. O, Hart, P. E, & Stork, D. G. (2000) Pattern Classification. (Wiley-Interscience).
33. Jordan, M, ed. (1998) Learning in graphical models. (MIT press, Cambridge, MA).
34. Yedidia, J, Freeman, W, & Weiss, Y. (2005) IEEE Trans. on Information Theory **51**, 2282–2312.