Quantifying Neurotransmission Reliability Through Metrics-Based Information Analysis

Romain Brasselet
romain.brasselet@upmc.fr
Centre National de la Recherche Scientifique, Université Pierre et Marie Curie,
UMR 7102, F75005 Paris, France

Roland S. Johansson
Roland.S.Johansson@physiol.umj.se
Umeå University, Department of Integrative Medical Biology,
SE-901 87 Umeå, Sweden

Angelo Arleo
angelo.arleo@upmc.fr
Centre National de la Recherche Scientifique, Université Pierre et Marie Curie,
UMR 7102, F75005 Paris, France

We set forth an information-theoretical measure to quantify neurotransmission reliability while taking into full account the metrical properties of the spike train space. This parametric information analysis relies on similarity measures induced by the metrical relations between neural responses as spikes flow in. Thus, in order to assess the entropy, the conditional entropy, and the overall information transfer, this method does not require any a priori decoding algorithm to partition the space into equivalence classes. It therefore allows the optimal parameters of a class of distances to be determined with respect to information transmission. To validate the proposed information-theoretical approach, we study precise temporal decoding of human somatosensory signals recorded using microneurography experiments. For this analysis, we employ a similarity measure based on the Victor-Purpura spike train metrics. We show that with appropriate parameters of this distance, the relative spike times of the mechanoreceptors' responses convey enough information to perform optimal discrimination—defined as maximum metrical information and zero conditional entropy—of 81 distinct stimuli within 40 ms of the first afferent spike. The proposed information-theoretical measure proves to be a suitable generalization of Shannon mutual information in order to consider the metrics of temporal codes explicitly. It allows neurotransmission reliability to be assessed in the presence of large spike train spaces (e.g., neural population codes) with high temporal precision.

1 Introduction

Shannon information theory provides a mathematical framework to characterize the input-output relationship of probabilistic communication systems (Shannon, 1948; Cover & Thomas, 1991). Neural information processing involves multistage transmission mechanisms in which neurons, as well as neural populations, act as stochastic communication channels. Information-theoretical tools can then be used to quantify the knowledge encoded in neural responses and the reliability of neural encoding and decoding mechanisms (MacKay & McCulloch, 1952; Bialek, Rieke, de Ruyter van Steveninck, & Warland, 1991; Deco & Obradovic, 1997; Rieke, Warland, de Ruyter van Steveninck, & Bialek, 1997; Borst & Theunissen, 1999).

Here we propose an extension of Shannon information theory that accounts for the metrical properties of spike time patterns to assess neurotransmission. In contrast to Shannon mutual information (MI), which provides an upper bound of the knowledge that an ideal observer can extract from neural responses (Borst & Theunissen, 1999; Quian Quiroga & Panzeri, 2009), the metrical mutual information defined here can reflect the properties of an actual neural decoder (e.g., its temporal selectivity capability). It provides a parametric information analysis of the statistical dependence between stimulus and neural response. The parametric nature of the analysis comes from its dependence on a similarity measure that quantifies, as spikes flow in, the metrical relations represented by the temporal code.

A large body of work has been done to assess the capacity of neurons to process afferent signals and transmit meaningful accounts of their inputs to downstream neural stages (see Quian Quiroga & Panzeri, 2009, for a recent review). Henceforth, we review existing information-theoretical approaches to estimate neurotransmission reliability (see section 2). We stress the importance of reducing the dimensionality of the event space (formed by spike train patterns) by preserving the information content of spatiotemporal codes. In section 3, we set forth a metrical information analysis based on a novel definition of entropy that embeds an explicit measure of the metrical relations between input-output events. In section 4, we validate the metrical information analysis on a data set of human microneurography recordings and perform a temporal decoding analysis of the responses of fingertip mechanoreceptors to tactile stimulation (Johansson & Birznieks, 2004). Finally, in section 5, we discuss the approach with respect to existing methods.

2 Information-Theoretical Analysis of Neural Codes

Information-theoretical approaches have characterized neurotransmission at different organization levels, from single synapses (e.g., Zador, 1998; Manwani & Koch, 2001; Tiesinga, 2001; Fuhrmann, Segev, Markram, & Tsodyks, 2002; London, Schreibman, Häusser, Larkum, & Segev, 2002;

Manwani, Steinmetz, & Koch, 2002; Goldman, 2004), to single-cell responses
(e.g., Theunissen & Miller, 1991; de Ruyter van Steveninck, Lewen, Strong,
Koberle, & Bialek, 1997; Borst & Theunissen, 1999; Tiesinga, Fellous, José, &
Sejnowski, 2002; Butts & Goldman, 2006; Sharpee et al., 2006; Arleo, Nieus,
Bezzi, D'Errico, D'Angelo, & Coenen, 2010), to neural population activity
(e.g., Brenner, Strong, Koberle, Bialek, & de Ruyter van Steveninck, 2000; Re-
ich, Mechler, & Victor, 2001; Lu & Wang, 2004; Smith & Lewicki, 2006; Saal,
Vijayakumar, & Johansson, 2009; Quian Quiroga & Panzeri, 2009). Also, nu-
merous studies on sensory neural processing (e.g., Richmond & Optican,
1990; Buracas, Zador, DeWeese, & Albright, 1998; Panzeri, Petersen, Schultz,
Lebedev, & Diamond, 2001; Paz & Vaadia, 2004; Rainer, Lee, & Logothetis,
2004; Einhauser, Mundhenk, Baldi, Koch, & Itti, 2007) have demonstrated
the relevance of stimulus-specific information transmission measures (see,
e.g., DeWeese & Meister, 1999, for a review). Finally, recent work (Thomson
& Kristan, 2005; Victor & Nirenberg, 2008; Quian Quiroga & Panzeri, 2009)
has begun to elucidate the relation between information-theoretical analysis
and more explicit decoding schemes to infer the most likely stimulus that
elicited an observed response (e.g., Bayesian algorithms, k-nearest neighbor
decoders, and population vector approaches; see reviews by Rieke et al.,
1997; Brunel & Nadal, 1998; Borst & Theunissen, 1999; Pouget, Dayan, &
Zemel, 2000; Dayan & Abbott, 2001).

Shannon mutual information (MI) estimates the average amount of
knowledge that can be extracted from the neural responses $r \in R$ to the
inputs $s \in S$ (both considered as probabilistic variables) as

$$I(R; S) = H(R) - H(R|S), \tag{2.1}$$

where $H(R)$ is the marginal entropy in the output events and quantifies the
intrinsic variability of the response space:

$$H(R) = -\sum_{r \in R} p(r) \log_2 p(r), \tag{2.2}$$

and $H(R|S)$ is the conditional entropy in the response space given the input

$$H(R|S) = -\sum_{s \in S} p(s) \sum_{r \in R} p(r|s) \log_2 p(r|s). \tag{2.3}$$

$H(R|S)$ is also called neuronal noise (Borst & Theunissen, 1999) because
it estimates the average uncertainty in the neural response after stimulus
presentation. Shannon mutual information can then be rewritten as

$$I(R; S) = \sum_{r,s} p(r, s) \log_2 \left(\frac{p(r, s)}{p(r)p(s)} \right). \tag{2.4}$$

$I(R; S)$ quantifies the difference between signal and noise entropy by mea-
suring how much one can learn about the stimulus by observing the neural
responses (or vice versa). $I(R; S)$ is zero if stimuli s and responses r are

completely uncorrelated, $I(R; S) > 0$ otherwise. For a given set of stimuli, the mutual information is maximal (i.e., $I(R; S) = H(R)$) if and only if no response r is elicited by two different stimuli s, that is, if there is no ambiguity when reconstructing the input by observing the output ($H(R|S) = 0$).

In the following, we review some methods for estimating the entropy and the mutual information of spiking signals, with a focus on state-space partitioning and the use of metrics to analyze spike trains.

2.1 Event Space Quantization

2.1.1 Binary Word Coding. If the readout system considers the event state space as discrete, probability distributions (i.e., input a priori probability $p(s)$, output marginal probability $p(r)$, conditional and joint probabilities $p(r|s)$ and $p(r, s)$, respectively) can be estimated empirically on the basis of natural equivalence classes (e.g., spike counts). In the case of continuous event spaces (e.g., spike time), the equivalence classes will depend on the precision of the discretization process used to partition the continuous state space. For instance, spike trains can be processed by means of a binning procedure that maps them into binary words (see Figure 1a; Strong, Koberle, de Ruyter van Steveninck, & Bialek, 1997; Dimitrov & Miller, 2000; London et al., 2002), which preserves the information about spike timing up to a certain precision (e.g., a few milliseconds). Therefore, binary word coding relaxes the temporal resolution constraint (see, Panzeri, Senatore, Montemurro, & Petersen, 2007, for a review) and allows equivalence classes (with arbitrary borders) to be defined by reducing the dimensionality of the input-output space. Once the equivalence classes have been defined, both Shannon entropy and MI can be computed (see equations 2.1–2.4).

When neural responses are processed according to spike count decoding, two spike trains r and r' are considered as equal if and only if they have the same number of spikes N. That is, the following similarity measure $\phi(r, r')$ is considered:

$$\phi(r, r') = \begin{cases} 1 & \text{if } N(r) = N(r') \\ 0 & \text{otherwise} \end{cases}. \tag{2.5}$$

When the neural decoder accounts for the timing t_i of the spikes, the similarity measure can be written as

$$\phi(r, r') = \begin{cases} 1 & \text{if } \forall i, \ t_i(r) \sim t_i(r') \\ 0 & \text{otherwise} \end{cases}. \tag{2.6}$$

This measure obviously depends on the temporal resolution taken to determine when $t_i(r)$ and $t_i(r')$ can be considered similar ($t_i(r) \sim t_i(r')$). The more severe the similarity measure (e.g., with a temporal resolution of 1 ms), the more precise the decoding process, and the smaller the information loss when processing the neural responses. Yet precise temporal decoding
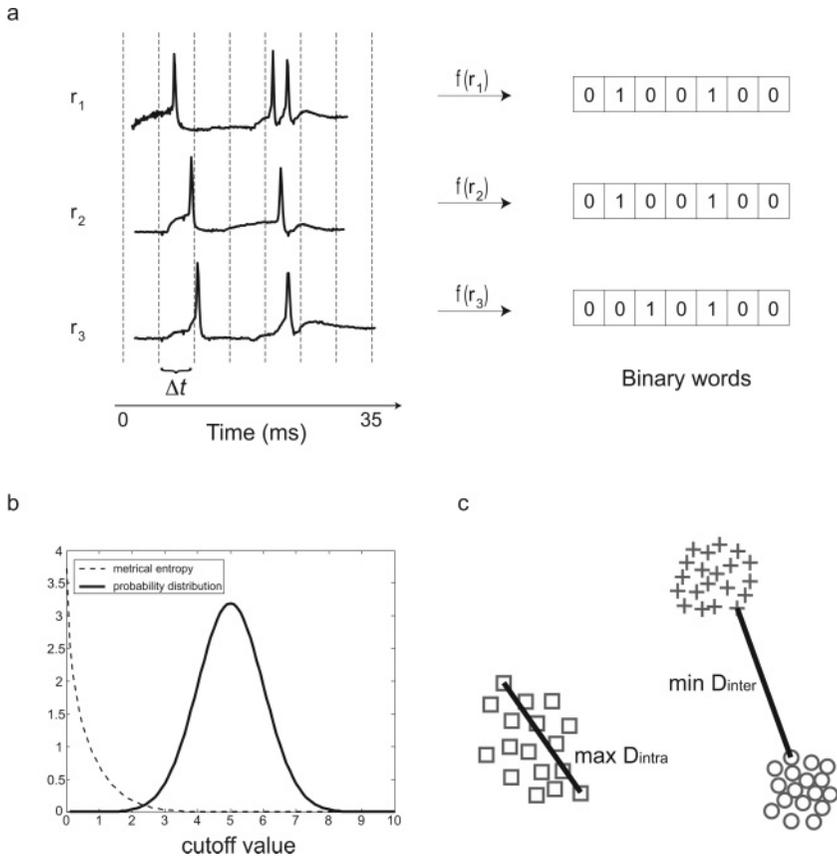
Figure 1: Methods: (a) Binary word decoding of neural responses. Time is partitioned according to a fixed temporal resolution Δt. If a spike occurs within a given time bin, the latter is assigned 1, otherwise 0. Besides potential disregard of action potentials (i.e., multiple spikes might occur within the same temporal bin), the binning procedure is prone to anisotropy effects: spike trains r_1 and r_2 are considered identical, whereas r_2 and r_3 are considered different (although the spike times of r_2 and r_3 are closer than those of r_1 and r_2). (b) Example of evolution of the metrical entropy (dashed curve) of a gaussian distribution of events (continuous curve) as a function of the cutoff when taking a Heaviside similarity measure (i.e., the similarity function is 1 for distances below the cutoff value and 0 otherwise). (c) In this toy example, the maximum intrastimulus distance $\max D_{intra}$ is smaller than the minimal interstimulus distance $\min D_{inter}$. Thus, we are in the perfect discrimination condition: $I^*(R; S)$ is maximal and $H^*(R|S)$ is nil.

makes it computationally intractable to estimate the Shannon entropy of spiking signals as the number of events increases, or their jitter is large, or in the presence of large neural populations (the curse of dimensionality; Borst & Theunissen, 1999; Quian Quiroga & Panzeri, 2009) because of a dramatic undersampling of the response space.

Note that binary word coding guarantees the transitivity of identity—$r \sim r'$ and $r' \sim r'' \Rightarrow r \sim r''$—but the binning procedure is prone to local anisotropy effects on the similarity measurements (see Figure 1a). For example, given a partitioning based on time bins of $\Delta t = 5$ ms, two spikes separated by 4 ms may be considered identical, whereas two spikes separated by 1 ms may be considered different.

2.1.2 State-Space Discretization Based on Rate Distortion Theory. Rate distortion theory (RDT) (Cover & Thomas, 1991) has been extensively used to study neural encoding and decoding through an information-theoretical approach. The goal is to find a compression (discretization) of the response space such that the information about the stimuli is preserved. The compression keeps the features of the code that are relevant for information transmission and discards the irrelevant ones. More technically, it amounts to finding an optimal quantization (partitioning into equivalence classes) of the response space with respect to a distortion (or cost) function.

In Tishby, Pereira, and Bialek (1999) and Schneidman, Slonim, Tishby, de Ruyter van Steveninck, and Bialek (2002), the cost function to be minimized is $\mathcal{L}(p(\tilde{r}|r)) = I(\tilde{R}; R) - \beta I(\tilde{R}; S)$, where $\tilde{R}$ is a compressed version of R. The goal is to find the quantization $\tilde{R}$ capturing the most information about the input S while discarding the unnecessary variability of the responses. The factor β determines the trade-off between quality of transmission and compression level. In Dimitrov, Miller, Aldworth, and Gedeon (2001) and Dimitrov, Miller, Gedeon, Aldworth, and Parker (2003), the state-space discretization $\tilde{R}$ of R is optimized with respect to a distortion function $D_I(R, \tilde{R}) = I(S; R) - I(S; \tilde{R})$ (with the same notation as above). This function estimates the loss of information when a quantization of the response space is applied. The optimal quantization is the one that maximizes $H(\tilde{R}|R)$ (i.e., it makes no further assumption on the code) while keeping the distortion function below some fixed threshold D. Both approaches aim at determining the codebook (or dictionary) between input and output and, thus, finding natural timescales for encoding information (Dimitrov & Miller, 2000).

2.2 Spike Train Metrical Analysis. A solution to avoid the discretization of the event space for the calculation of $I(R; S)$ is to account for spike train metrics (Victor & Purpura, 1996; Van Rossum, 2001; Quian Quiroga, Kreuz, & Grassberger, 2002; Aronov, Reich, Mechler, & Victor, 2003; Schreiber, Fellous, Whitmer, Tiesinga, & Sejnowski, 2003; Kreuz, Haas, Morelli, Abarbanel, & Politi, 2007; Houghton & Sen, 2008). For instance, a

well-established approach consists of estimating the Victor-Purpura (VP) distances between the neural responses (Victor & Purpura, 1996) and classifying them according to a clustering method such as the k-nearest neighbor algorithm (Duda, Hart, & Stork, 2001). That is, the response is assigned the class that constitutes the majority among its k-nearest neighbor responses (the positive-definite integer k being a free parameter of the method). This classification allows the so-called confusion matrix C_{ij} to be computed, whose terms are the probability of classifying a response to the ith stimulus as a response to the jth stimulus. In the presence of perfect decoding, the main diagonal of the confusion matrix C_{ij} should be equal to 1, whereas all the other entries should be zero. The use of VP spike train metrics coupled to clustering methods (e.g., k-nearest neighbors) to generate confusion matrices has proved to be suitable to estimate the lower bounds of Shannon mutual information $I(R; S)$ in sensory information processing, for example, auditory (Huetz, Del Negro, Lebas, Tarroux, & Edeline, 2006) and somatosensory (Saal et al., 2009).

Although this approach takes into account the metrical properties of the event space to compute Shannon mutual information, once equivalence classes have been defined, the metrics are forgotten. Whether the clusters are close to each other or far apart no longer influences the estimation of $I(R; S)$. In other words, spike train metrics is not explicitly embedded into the calculation of mutual information, which does not depend on either the shape of the clusters or the distance between them.

3 Metrical Information Theory: From Equivalence Classes to Distances

3.1 Extension of Shannon Entropy. Shannon entropy calls on a similarity measure $\phi(r, r')$ that can be thought of as a Kronecker function of two responses $r, r' \in R$:

$$\phi(r, r') = 1 \Leftrightarrow r = r'. \tag{3.1}$$

As a consequence, two events can be only identical or different (i.e., two slightly different or two very different events are treated the same way). This makes Shannon entropy diffeomorphism-invariant (or topological) because whether the mapping between S and R is random or isometric does not influence Shannon mutual information.

We propose an extension of Shannon entropy on the response space that is defined as follows,

$$H^*(R) = -\sum_{r \in R} p(r) \log_2 \left(\sum_{r' \in R} p(r')\phi(r, r') \right), \tag{3.2}$$

where R is the set of events and $\phi(r, r')$ is a yet unspecified similarity measure between the events r and r'. In the general framework, the similarity measure can be any real function with values in $[0, 1]$. The contrast

with Shannon entropy is the logarithm argument: instead of $p(r)$, we have $\sum_{r' \in R} p(r')\phi(r, r')$. It can thus be noted that:

- The logarithm argument is always higher than the probability of r because

$$\sum_{r' \in R} p(r')\phi(r, r') = p(r) + \sum_{r' \in R, r' \neq r} p(r')\phi(r, r'),$$

  the second term being positive. Therefore, this entropy is always lower than the Shannon entropy.
- If the similarity measure $\phi(r, r')$ is taken to be a Kronecker delta, then the logarithm argument of equation 3.2 is equal to $p(r)$, and thus $H^*(R)$ reduces to the Shannon entropy.
- Although the similarity measure is not specified yet, it can be seen that the broader $\phi(r, r')$, the lower the entropy $H^*(R)$.
- If $\forall r, r' \in R$, $\phi(r, r') = 1$, the entropy is nil.

An extension of the conditional entropy on the response given a specific stimulus $s \in S$ can then be defined as

$$H^*(R|s) = -\sum_{r \in R} p(r|s) \log_2 \left(\sum_{r' \in R} p(r'|s)\phi(r, r') \right), \tag{3.3}$$

and its average over the entire set of stimuli S gives

$$H^*(R|S) = \sum_{s \in S} p(s)H(R|s) = -\sum_{s \in S}\sum_{r \in R} p(r, s) \log_2 \left(\sum_{r' \in R} p(r'|s)\phi(r, r') \right). \tag{3.4}$$

Finally, similar to Shannon MI, the mutual information $I^*(R; S)$ is taken as a difference between marginal and conditional entropies:

$$I^*(R; S) = H^*(R) - H^*(R|S) \tag{3.5}$$

$$= \sum_{s \in S}\sum_{r \in R} p(r, s) \log_2 \left(\frac{\sum_{r' \in R} p(r'|s)\phi(r, r')}{\sum_{r' \in R} p(r')\phi(r, r')} \right). \tag{3.6}$$

This is the general framework of the theory. It encompasses as a specific case the unbinned Shannon entropy but also the binning procedure, where all the events belonging to the same bin have similarity equal to 1, while two events in different bins have similarity equal to 0. The codebook method can also be restated within this framework.

3.2 Metrical Entropy. In order to account for the metrical relations between events, we propose to define the similarity measure as a decreasing function of the distance between two events, following the intuitive idea that a distance provides a measure of dissimilarity. We thus call H^* a metrical entropy. Figure 1b shows an example of the metrical entropy (dashed

curve) for a one-dimensional gaussian distribution of events (solid curve) when defining the similarity measure as a Heaviside function of the distance between two events: $\phi(r, r') = \mathcal{H}(D_{\text{critic}} - d(r, r'))$. Henceforth, we will focus on this form for ϕ. When the cutoff D_{critic} (or critical distance) of the Heaviside function is zero, we get back to Shannon entropy, and when it goes to infinity, the metrical entropy becomes zero.

When applying the metrical information analysis to neurotransmission, we first need to select a distance function on the event space and define a class of similarity measures (see section 3.2.1), for example, a Heaviside function of the distance. Then we define an optimality condition for efficient neurotransmission in section 3.2.2. Finally, we determine the optimal values of the parameters with respect to the defined optimality condition in section 3.2.3.

3.2.1 Definition of the Similarity Measure. When the event space consists of spiking signals, a method to quantify the distance between neural responses is offered by the Victor-Purpura (VP) spike train metrics (Victor & Purpura, 1996), though other metrics may be suitable as well. We define the similarity measure $\phi(r, r')$ as a function of the VP distance $D_{VP}(r, r')$ between two responses r and r'. The distance $D_{VP}(r, r')$ depends on the VP cost parameter C_{VP} (Victor & Purpura, 1996), which determines the timescale of the temporal coding analysis. The cost parameter C_{VP} regulates the influence of spike timing versus spike count when calculating the VP distance between r and r'. As stated above, we wish to define the similarity as a decreasing function of the distance. A simple way to do that is to take a Heaviside step function of the distance $D_{VP}(r, r')$,

$$\phi(r, r') = \mathcal{H}(D_{\text{critic}} - D_{VP}(r, r')), \tag{3.7}$$

where the critical distance D_{critic} is the cutoff parameter: as long as $D_{VP}(r, r') < D_{\text{critic}}$ the responses r, r' are considered to be identical; otherwise they are classified as different. Note that if $D_{\text{critic}} = 0$, one recovers the Shannon entropy from equation 3.2.

It is worth mentioning that by taking a similarity measure different from the Kronecker delta, we introduce a bias in the computation of the entropy $H^*(R)$ (because any ϕ other than the Kronecker delta can only reduce the estimate). This bias depends on the characteristics of the readout system, for example, the properties of a downstream neural population. A consequence of this bias is that the metrical information I^* is not diffeomorphism-invariant (in contrast to Shannon mutual information) but rather depends on the metrical organization of events. Note also that since we take a similarity measure that depends on the $D_{VP}(r, r')$ distance only, the metrical quantities are invariant under the isometry group and under

isotropic homotheties (i.e., affine transformations preserving the ratios between event distances).

3.2.2 Optimality Condition for Efficient Neurotransmission. In the zero-noise limit, optimal feedforward information processing would require maximizing the metrical information $I^*(R; S)$ according to an infomax-like principle (Linsker, 1988; Bell & Sejnowski, 1995; Nadal, Brunel, & Parga, 1998). However, since neural information processing is not noise free and occurs through multiple encoding and decoding stages, the issue of minimizing the variability on the output representation at each stage constitutes a major requirement. Therefore, akin to the principle of redundancy reduction proposed for biological sensory processing by Barlow (1961), the metrical conditional entropy $H^*(R|S)$ constitutes a fundamental quantity for neurotransmission, and neural information processing should both maximize the metrical information $I^*(R; S)$ and (at the same time) minimize the conditional entropy $H^*(R|S)$. This perfect discrimination condition is met when all the responses elicited by the same stimulus are strictly identical ($\phi = 1$), whereas all the pairs of outputs elicited by two distinct stimuli are strictly different ($\phi = 0$). In the sequel, we will also refer to this situation as the optimal discrimination condition.

According to this optimality principle, a neural encoder should provide that any response is not elicited by two different stimuli and that the responses elicited by any stimulus remain as close as possible. If a neural encoder were not behaving this way, then a single stimulus would possibly elicit many different responses. How could this be effectively decoded by a downstream neural population? In other words, neural information processing requires encoding mechanisms capable of producing as few responses as possible to a given stimulus while keeping these responses different between stimuli (i.e., sparse coding; see, e.g., Földiák, 1990; Nadal & Toulouse, 1998; Olshausen & Field, 1996; Willmore & Tolhurst, 2001). It is worth mentioning that the idea of minimizing the conditional entropy can be traced back at least to work by MacKay and McCulloch (1952), in which the time resolution taken to study a spike train was chosen so that spike timing was not affected by the jitter.

3.2.3 Determining the Parameters of the Similarity Measure. The similarity measure $\phi(r, r')$ is the lever on which we can act to set the compromise between the overall information $I^*(R; S)$ and the mean variability of the response to a stimulus $H^*(R|S)$. There are two ways of understanding the similarity measure. On the one hand, provided that a comprehensive knowledge of the properties of the readout system (e.g., a neural decoder) is available, the selectivity level (i.e., the cutoff distance D_{critic}) of the similarity measure should account for the specific properties of the reader (e.g., time constant of synaptic integration). On the other hand, if no knowledge on the decoding system is available, the similarity measure should be set

to allow optimal information transmission (in terms of both $I^*(R;S)$ and $H^*(R|S)$) to be achieved in order to predict the values of the parameters that maximize information transmission.

A reasonable way to proceed to set the optimal parameters for $\phi(r, r')$ is to consider two sets of $D_{VP}(r, r')$ distances:

- The distances between the responses elicited by the same stimulus (henceforth referred to as intrastimulus distances)
- The distances between the responses elicited by different stimuli (the interstimulus distances).

In order to determine the optimal value for D_{critic}, we compute the minimum and maximum intrastimulus distances as well as the minimum and maximum interstimulus distances. The optimal coding condition, corresponding to maximum $I^*(R;S)$ and zero $H^*(R|S)$, occurs when the maximum intrastimulus distance (which provides the size of the largest cluster of responses) becomes smaller than the minimum interstimulus distance (which estimates the smallest distance between clusters of responses). We can thus set the cut-off distance between the maximum intrastimulus distance and the minimum interstimuli distance (see Figure 1c for a toy example).

In the case of neurotransmission, the relationship between intra- and interstimulus distance distributions tends to evolve over time as the input spike waves across multiple afferents flow in the readout system. Figure 2a shows an example of intra- and interstimulus distance distributions evolving over time. The two distributions separate from each other after about 110 ms. The critical parameter D_{critic} can then be taken as the distance at which the maximum intrastimulus distance becomes smaller than the minimum interstimulus distance (dashed line in Figure 2a). The time at which the critical distance D_{critic} can be determined indicates when perfect discrimination can be achieved. In other words, optimal discrimination occurs when the distributions of intra- and interstimulus distances stop overlapping, which means that (1) the conditional entropy $H^*(R|S)$ is nil, because all the responses elicited by any stimulus are identical, and (2) the information $I^*(R;S)$ is maximum (i.e., equal to $H^*(R)$), because any two responses elicited by two different stimuli are always correctly discriminated. Figure 2b illustrates an example of well-separated intra- and interstimulus distance distributions (bottom). It also shows the values for $I^*(R;S)$ and $H^*(R|S)$ as a function of the critical distance parameter (top). It appears clearly that taking $6.5 < D_{\text{critic}} < 8$ guarantees perfect discrimination (because within that range, $I^*(R;S)$ is maximum and $H^*(R|S)$ is nil).

The critical distance D_{critic} is interdependent on the VP cost parameter C_{VP} (Victor & Purpura, 1996). We define the optimal VP cost C_{VP}^* as the one that leads to earliest perfect discrimination (in the example in Figure 2a, a cost $C_{VP} = 0.15$ led to perfect discrimination after 110 ms).

Figure 2c presents a toy example comparing the estimates provided by Shannon mutual information (section 2), Shannon mutual information
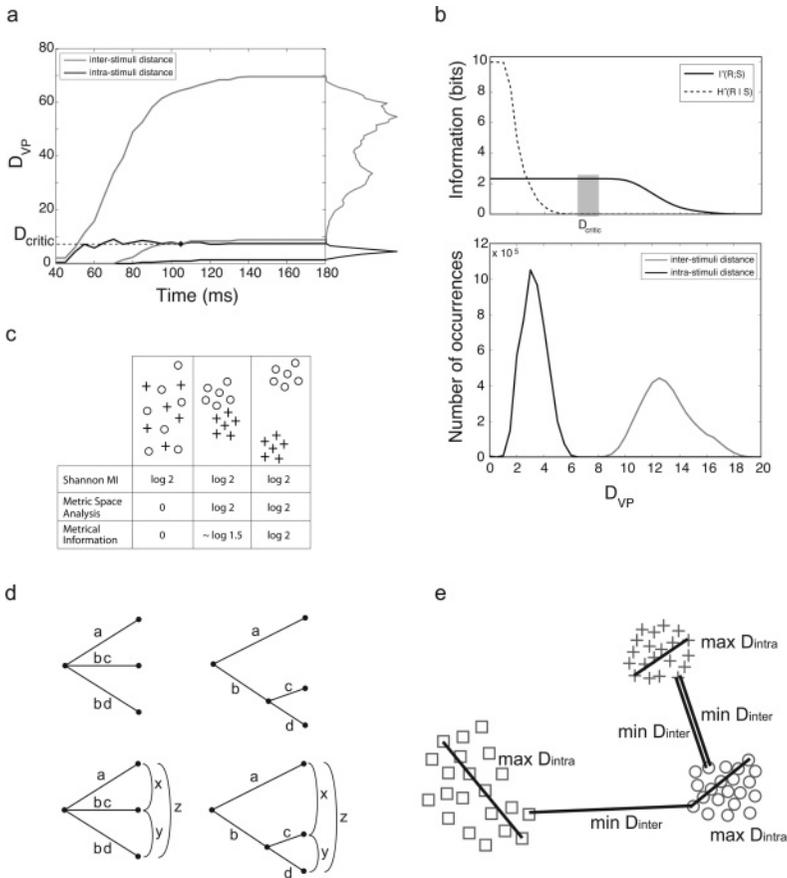
Figure 2:   Methods: (a) Example of intra- and interstimulus distance distributions (black and gray curves, respectively) over time for a VP cost parameter $C_{VP} = 0.15$. The optimal discrimination condition is met after about 110 ms, when the two distributions stop overlapping. (b) Example of how $I^*(R; S)$ and $H^*(R|S)$ vary as a function of the critical distance (top), given the distributions of intra- and interstimulus distances (bottom). (c) Simple comparison of Shannon MI, metrical space analysis, and metrical information analysis on a two-stimuli classification example. (d) Simple example showing the additivity property of Shannon entropy (top row) and metrical entropy (bottom row) with probabilities denoted a, b, c and similarities x, y, z. (e) Local versus global metrical information measures. Since the size of the cluster of square symbols is larger than the distance between the clusters of crosses and circles, the global similarity measure would make it impossible to maximize $I^*(R; S)$ while maintaining $H^*(R|S) = 0$. By contrast, a local version of the metrical entropy would consider different similarity measures for the clusters. Thus, the D_{critic} for the cluster of circles would be taken smaller than the one for the cluster of squares.

based on metrical space analysis (section 2.2), and metrical information analysis. It can be seen that the mutual information $I^*(R; S)$ captures the metrical property of the mapping from input to output spaces. For instance, if a stimulus s yields 10 different responses r randomly distributed on the output space, then $I^*(R; S)$ is nil, because the critical distance will be very large.

3.2.4 Additivity Property of Metrical Information. Shannon entropy satisfies the following properties:

1. It is continuous on the probabilities p.
2. If the equivalence classes are equiprobable, that is, $\forall p$, $p = \frac{1}{N}$ with N being the number of equivalence classes, then the entropy $H(R)$ is a monotonic increasing function of N.
3. If a probabilistic choice is further broken down into two successive choices, the overall entropy should be a weighted sum of the subchoice entropies. In Figure 2d (top), $H(a, bc, bd) = H(a, b) + bH(c, d)$, where a, b, c denote probabilities.

The metrical entropy H^* satisfies the first two requirements. Does it also satisfy the third property of additivity? Consider Figure 2d (bottom). The joint metrical entropy is

$$H^*(a, bc, bd) = -a \log(a) - bc \log(bc + bdy) - bd \log(bd + bcy).$$

$$(3.8)$$

Since

$$H^*(a, b) = -a \log(a) - b \log(b) \tag{3.9}$$

$$H^*(c, d) = -c \log(c + dy) - d \log(d + cy), \tag{3.10}$$

then,

$$bH^*(c, d) + H^*(a, b) = -a \log(a) - b \log(b) - bc \log(c + dy)$$

$$-bd \log(d + cy) \tag{3.11}$$

$$= -a \log(a) - bc \log(bc + bdy)$$

$$-bd \log(bd + bcy) \tag{3.12}$$

$$= H^*(a, bc, bd). \tag{3.13}$$

The metrical entropy thus satisfies the three previous fundamental requirements.

3.3 Metrical Entropy Based on a Local Similarity Function. A critical property of the metrical information $I^*(R; S)$ is the global nature of the similarity measure $\phi(r, r')$—the fact that the critical distance does not depend on

the position in the event space. In Figure 2e, $I^*(R; S)$ cannot be maximized while keeping $H^*(R|S) = 0$ because the size of one of the clusters is larger than the distance between two others. This is an undesired outcome of the analysis because all the responses within a cluster are closer to each other than the closest responses from other clusters.

Otherwise stated, the global version of the metrical information assumes a similarity measure that does not depend on the region of the event space. In order to circumvent this limitation, we also consider a locally specific discrimination selectivity function. We then derive a local version of the metrical entropy by taking a similarity measure that varies over the event space. Thus, we take a similarity measure at response r with respect to other's responses r' as $\phi_r(r') = \mathcal{H}(D_{\text{critic}}(r) - d(r, r'))$, with a critical distance that now depends on r. A collateral of this locality is that $\phi_r(r')$ is not necessarily equal to $\phi_{r'}(r)$ anymore. Note that we still constrain the similarity measure to be isotropic.

In the context of the optimality condition (see section 3.2.2), the critical distance for $\phi_r(r')$ becomes the size of the cluster of responses (to a given stimulus) that r belongs to. As a consequence, instead of the two parameters C_{VP} and D_{critic}, the local version of metrical information measure will depend on C_{VP} and a family of critical distances $\mathcal{D} = \{D_{\text{critic}}\}$ (see Figure 2e). When a local similarity measure is considered, the optimal discrimination condition, that is, maximum $I^*(R; S)$ and zero $H^*(R|S)$, will occur when the size of each cluster in the event space becomes smaller than the distance with its closest clusters.

4 Results: Temporal Decoding of Human Microneurography Haptic Signals

In order to validate the metrical information analysis we have presented, we studied the temporal decoding of the responses of fingertip mechanoreceptors to tactile stimuli. Mechanoreceptors innervate the epidermis and discharge as a function of the mechanical indentations and deformations of the skin. Recent microneurography studies in humans (Johansson & Birznieks, 2004) have suggested that the relative spike timing of mechanoreceptor responses can convey information about contact parameters faster than the fastest possible rate code, and fast enough to account for the use of tactile signals in natural manipulation (Johansson & Flanagan, 2009).

To investigate fast encoding and decoding of tactile signals, we concentrated on the responses of fast-adapting (FA-I) mechanoreceptors only (Johansson & Birznieks, 2004). The overall input state space consisted of the responses of 42 FA-I mechanoreceptors to 81 distinct stimuli obtained by varying four primary contact parameters:

- Curvature of the probe ($C = \{0, 100, 200\}$ m^{-1}, $|C| = 3$)
- Magnitude of the applied force ($F = \{1, 2, 4\}$ N, $|F| = 3$)

- Direction of the force (O = {ulnar, radial, distal, Proximal, Normal},
  $|O| = 5$)
- Angle of the force relative to the normal direction (A = {5, 10, 20}°,
  $|A| = 3$)

4.1 Metrical Information Analysis on a Limited Region of the Input-Output Space. First, we considered only the five force directions (ulnar, radial, distal, proximal, normal) as variable primary features (similar to Johansson & Birznieks, 2004, and Saal et al., 2009). We computed the VP distances D_{VP} across the population of 42 mechanoreceptor afferents, with each of the $|S| = 5$ stimuli presented 100 times.

According to the hypothesis that the variability in the first-spike latency domain with respect to stimulus feature is larger than the variability within repetitions of the same stimulus (Johansson & Birznieks, 2004), we focused on the first spike of each FA-I mechanoreceptor only. Thus, each tactile stimulus consisted of a single volley of spikes forming a spatiotemporal response pattern defined by the first-spike latencies across the afferent population (see Figure 3a for three sample recordings).

Figure 3b shows how the intra- and interstimulus distance distributions evolved over time. Within 25 ms of the the first mechanoreceptor discharge (and within 55 ms of stimulus onset), the critical cutoff $D_{critic} = 6.2$ of the Heaviside function could be set, which ensured that the perfect discrimination condition (maximum $I^*(R; S)$ and zero $H^*(R|S)$) was met. The black curves of Figure 3c confirm this result, showing that $I^*(R; S)$ (solid black curve), computed based on the global cutoff $D_{critic} = 6.2$, saturated at $\log_2(5)$ within 25 ms of the first spike arrival, while the condition entropy $H^*(R|S)$ (dashed black line) remained nil. The dark gray curves of Figure 3c show the time course of the local metrical information. Convergence to the perfect discrimination condition did not occur earlier than in the case of the global $I^*(R; S)$ (black curve). The reason is that the clusters happened to have similar sizes; therefore, all the local critical distances $\mathcal{D} = \{D_{critic}\} = \{5.3, 4.8, 4.8, 4.7, 6.2\}$ happened to be close to the global critical distance ($D_{critic} = 6.2$). The light gray curves of Figure 3c demonstrate that Shannon mutual information (light gray solid line), computed by considering a temporal binning of 1 ms resolution, increased faster than the metrical information. However, the conditional entropy (light gray dashed line) diverged.

Figure 3d displays two samples of distance matrices indicating how the decoding system clustered the input spike waves across the 42 mechanoreceptor afferents over time. Before the occurrence of the perfect discrimination condition (at 40 ms, left matrix) different stimuli could have relatively small distances, that is some interferences could impair the decoding process. After 60 ms (right matrix), all the initially overlapping contexts were separated, which removed all interferences across inputs and led to 100% accuracy in the discrimination process.
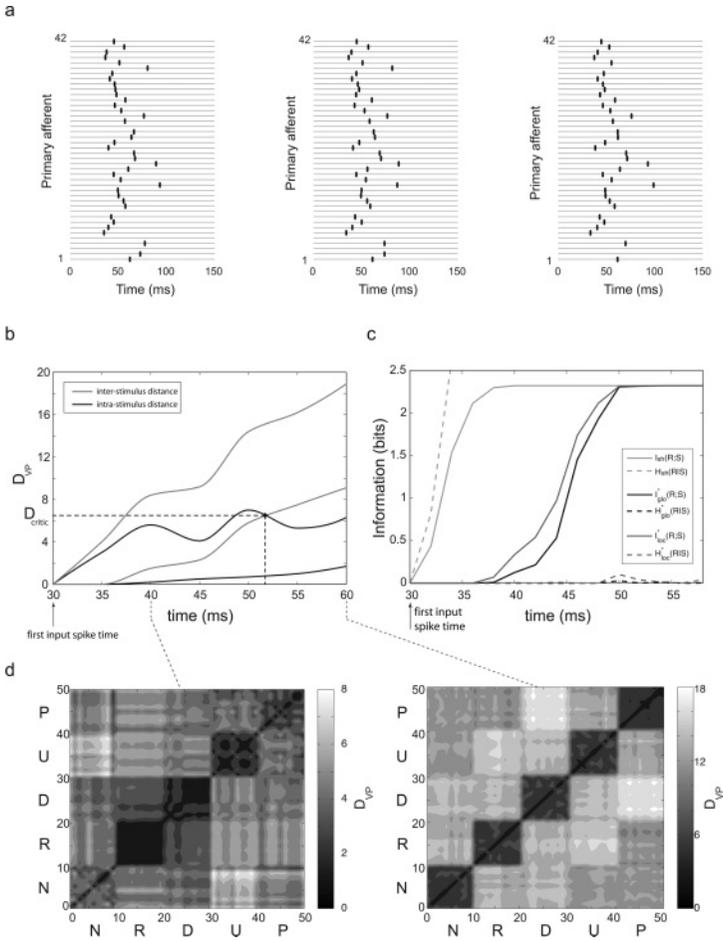
Figure 3:   Results: Information theoretical analysis on a limited stimulus set. The first spike latency patterns of $|S| = 5$ distinct stimuli, each of which presented 100 times, were considered for this analysis. (a) Three examples of human microneurography FA-I recordings showing the first spike waves across 42 afferents evoked by 3 distinct tactile stimuli (while varying only the orientation parameter). (b) Evolution of the intra- and interstimulus distances as spikes flowed in. The perfect discrimination condition was met about 55 ms after the stimulus onset, when the critical cutoff D_{critic} could be determined. (c) Shannon MI (light gray curves), global metrical information (black curve), and local metrical information (dark gray curves) over time. (d) Distance matrices before (left) and after (right) the occurrence of perfect discrimination. Only 10 presentations per stimulus were considered to generate these matrices. Whereas at 40 ms some interferences existed in the input-output mapping, all the events were well separated at 60 ms.

4.2 Metrical Information Analysis on an Extended Region of the Input-Output Space

4.2.1 First Spike Wave Analysis. We then scaled up the analysis to the entire set of microneurography recordings (i.e., 81 distinct stimuli encoded by 42 mechanoreceptor spike trains). Figure 4a compares Shannon mutual information against both the global and local versions of metrical information. In this case, only first spike latencies were considered by the analysis. Again, Shannon MI (the light gray solid line) reached its maximum faster than the metrical information, but the drawback is the large conditional entropy (the light gray dashed line). Global metrical information (the black line) shows that perfect discrimination ($I^*(R; S) = H^*(R)$) was reached within about 40 ms of the first afferent spike. The local version of metrical information (the dark gray line) converged faster than the global measure. This can be explained by the distribution of the local critical distances (i.e., the sizes of the event clusters) shown in Figure 4b. The global version of the I^* used a critical distance $D_{\text{critic}} = 6.8$, which was much too large for some stimuli. By contrast, the local version used the entire distribution of critical distances ($\mathcal{D} = \{1.7, \ldots, 6.8\}$) and it was then better adapted to the configuration of the input-output relationship.

It is worth noting that the time needed to achieve perfect discrimination was longer compared to the five stimulus case (40 versus 25 ms) not only because of the larger input set but also because some stimuli were rather difficult to separate in the D_{VP} distance space (e.g., two skin indentations with the same curvature, force, direction parameters, and with only a 5 degree difference on the angle). Figure 4c illustrates two examples of distance matrices largely and shortly before convergence (left and right matrices, respectively).

4.2.2 Full Spike Train Analysis. In a subsequent analysis, we took into account the complete spike train responses of the 42 recorded mechanoreceptors rather than the first spike waves only. Figure 5a shows an example of population spike latency pattern across 42 mechanoreceptors. When comparing again Shannon MI vsrsus global and local metrical information (see Figure 5b), we found that the time course of Shannon MI (the light gray line) did not change significantly compared to the first spike latency case.

Interestingly, the global metrical information measure (the black line) was impaired when taking into account the entire spike train. A plausible reason may be that the D_{VP} distance scaled up with the number of spikes: a stimulus with a large force amplitude (4 N) elicited many spikes per mechanoreceptor, making the size of the cluster corresponding to this stimulus large. Meanwhile, the distances between clusters corresponding to stimuli with low forces (1 N) remained small. As suggested by Figure 5b, it was more efficient to use the local version of I^* (the dark gray line), which behaved similar to the first spike wave case.
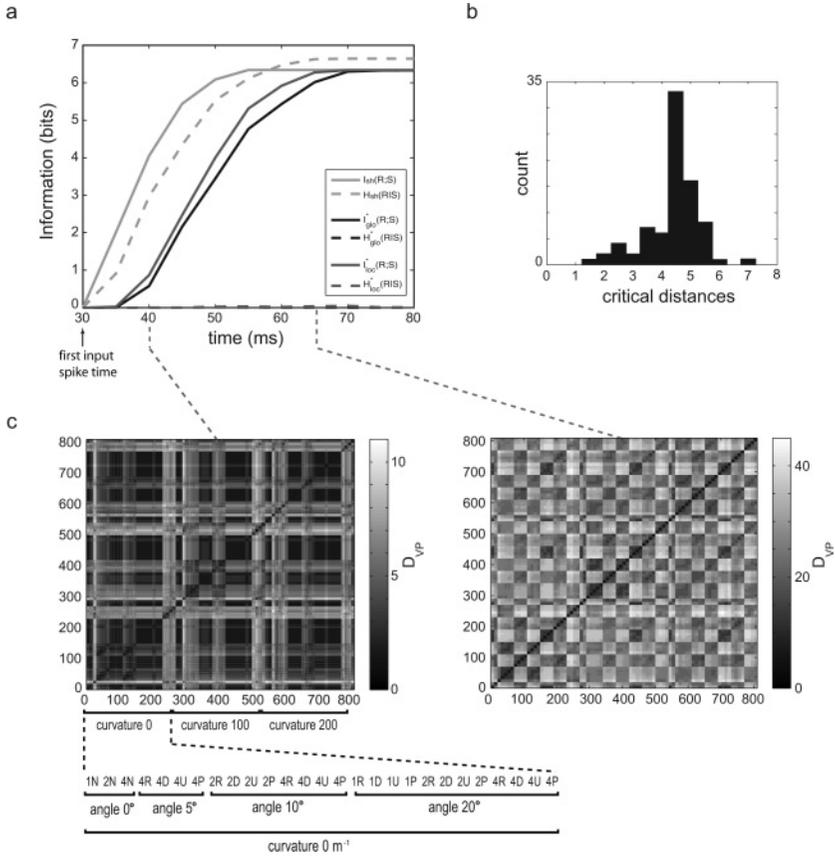
Figure 4: Results: Information-theoretical analysis on the entire stimulus set. The first spike latency patterns of $|S| = 81$ distinct stimuli, each presented 100 times, were considered by this analysis. (a) Shannon MI (light gray lines), global (black line), and local (dark gray line) metrical information measures as spikes flowed in. (b) Distribution of critical distances $\mathcal{D}$. (c) Distance matrices at 40 ms (left) and 65 ms (right), slightly before the occurrence of the perfect discrimination condition ($\sim$80 ms) (for clarity, these matrices were obtained on the basis of 10 presentations per stimulus). Bottom: Sequence of stimulus presentations according to the combinations of all contact parameters: curvature of the probe ($0\,\mathrm{m}^{-1}$, $100\,\mathrm{m}^{-1}$, and $200\,\mathrm{m}^{-1}$); angle of the force relative to the normal direction ($0°$, $5°$, $10°$, $20°$); magnitude of the force (1 N, 2 N, 4 N); direction of the force (normal, radial, distal, ulnar, proximal).
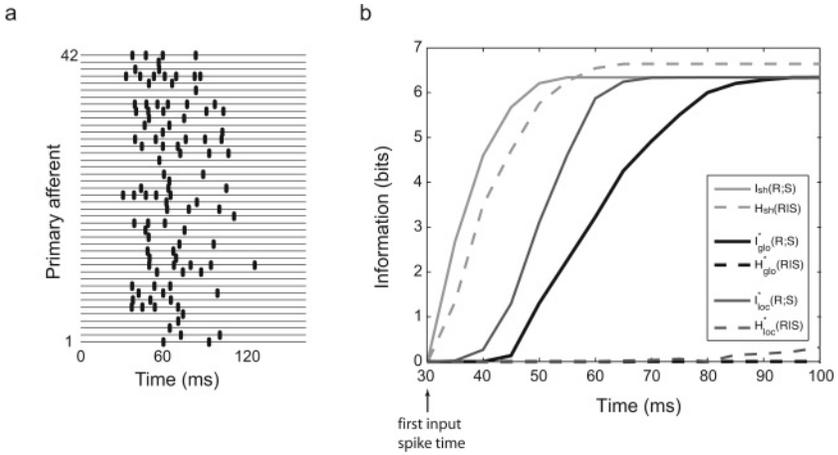
Figure 5: Results: Information-theoretical analysis on the complete stimulus set. The spike latencies of all spike trains, i.e. $|S| = 81$ distinct stimuli each presented 100 times, were considered by this analysis. (a) An example of human microneurography FA-I recordings showing the entire spike trains of 42 afferents. (b) Evolution of Shannon MI (light gray lines), global (black line), and local (dark gray lines) metrical information measures over time.

4.2.3 Detecting Regularities and Isometric Mapping. Since the time necessary to reach the optimal discrimination was similar when dealing with either first spike waves or whole spike trains, we investigated the possible contributions to information transmission of the second and following spike waves. The results shown in Figure 6 highlight a meaningful property of the entire spike trains.

We employed the Victor-Purpura distance by taking $C_{VP} = 0$ and measured the distances between the mechanoreceptor responses to stimuli with force amplitudes 1 and 2 N (i.e., $D_{VP}(1N, 2N)$), and then to stimuli with force 2 and 4 N (i.e., $D_{VP}(2N, 4N)$), and finally to stimuli with force 1 and 4 N ($D_{VP}(1N, 4N)$). The following relation was then verified for any set of the other free contact parameters: $D_{VP}(1N, 2N) + D_{VP}(2N, 4N) = D_{VP}(1N, 4N)$. This result held when only the first spike waves were considered (see Figure 6a) and the entire spike trains were analyzed (see Figure 6b). This means that the one-dimensional stimulus space was mapped onto a noncurved one-dimensional response space. This alignment property can be suitable for dissociating the problem of decoding the force of the stimulus from that of determining other features of the stimulus (i.e., pruning of the search state space).

Furthermore, Figure 6b shows that when entire spike trains are considered the distances between the outputs tended to reflect the distances

a
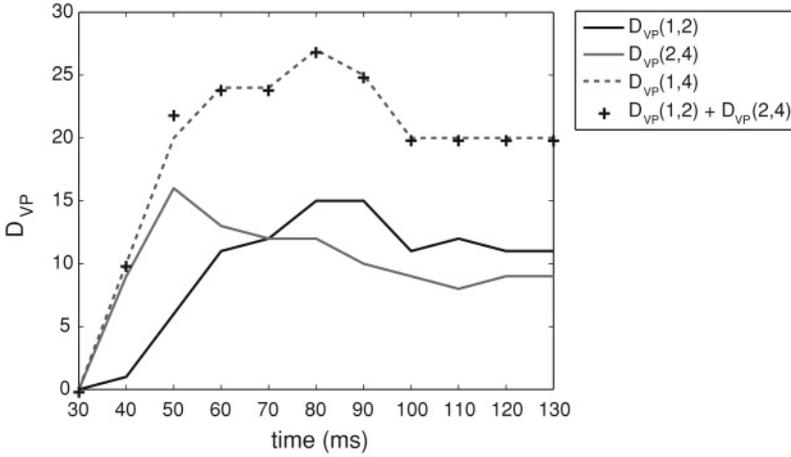
b
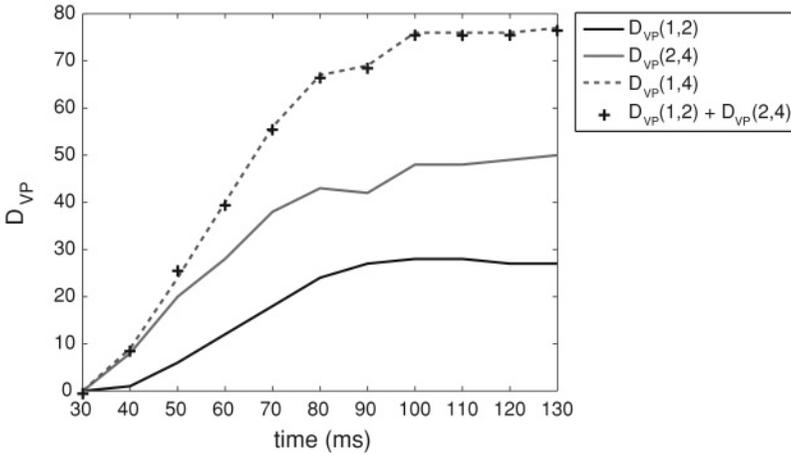
Figure 6: Results: Time course of the D_{VP} distances. Distances between the mechanoreceptor responses to stimuli with force amplitudes 1 and 2 N ($D_{VP}(1N, 2N)$), then to stimuli with force 2 and 4 N ($D_{VP}(2N, 4N)$), and finally to stimuli with force 1 and 4 N ($D_{VP}(1N, 4N)$). For each combination, all the other contact parameters were varied. (a) Results obtained when only the first spike of each mechanoreceptor was considered. (b) Results when the complete spike train was taken into account. In both cases, the equality $D_{VP}(1N, 2N) + D_{VP}(2N, 4N) = D_{VP}(1N, 4N)$ held through time. Furthermore, we observed an isometric input-output mapping when considering entire spike trains.

between the stimuli. Indeed, $D_{VP}(2N, 4N)$ was approximately two times $D_{VP}(1N, 2N)$, which suggests an isometry-like property of the decoding measure. This result can be interpreted as follows: temporal decoding based on first spike latencies allowed complete stimulus discrimination to be achieved very rapidly. Yet taking into account the spike timing of all the responses added to perfect discrimination the possibility of capturing some regularities of the input space (e.g., isometric mapping). Not only were the responses organized so that a quick discrimination was possible, but they were logically organized so that the mapping from the stimulus space to the response space was a simple transformation. This may be the basis for the ability to extrapolate or generalize the reconstruction of the stimulus when never-before-experienced stimuli are presented.

5 Discussion

This metrical information analysis proposes a complementary approach, compared to Shannon information theory, to study stochastic communication systems. Shannon mutual information estimates an upper bound on the quality of the coding, no matter the decoding system and the conditional entropy. As a consequence, Shannon information theory provides a benchmark to assess the fraction of the overall information actually captured by decoding algorithms (Borst & Theunissen, 1999; Quian Quiroga & Panzeri, 2009). Metrical information analysis incorporates the notion of decoding system (through the distance-based similarity function) and the importance of conditional entropy for the optimization of information transmission. This complementarity relation leads to a shift from an approach that tells how much information is there but not how to read it, to a constructive-like approach that considers the parameters of the decoding system that reads out the neural code.

This work attempts to contribute to existing methods, such as Fisher information analysis (Blahut, 1988; Clarke & Barron, 1990; Rissanen, 1996; Brunel & Nadal, 1998) and rate distortion theory (Bialek et al., 1991; Gabbiani & Koch, 1996; Dimitrov et al., 2001, 2003), that bridge the gap between pure decoding approaches (e.g., Bayesian algorithms, k-nearest neighbor decoders, population vector decoding; Rieke et al., 1997; Borst & Theunissen, 1999; Pouget et al., 2000; Dayan & Abbott, 2001) and pure information-theoretical analyses (e.g., London et al., 2002; Sharpee et al., 2006; Butts & Goldman, 2006). It also relates to recent works by Thomson and Kristan (2005), Victor and Nirenberg (2008), and Quian Quiroga and Panzeri (2009) that have studied the complementary properties of decoding and information theoretical algorithms to decipher the neural code. Our work is related to the approach by Treves (1997) that estimates the structure of the perceptual space by comparing results from information theory and maximum-likelihood estimations. Theoretical constraints may be derived between Shannon mutual information and the fraction of correct decodings.

For a given fraction of correct decodings, it is possible to infer lower and upper bounds to the MI depending on the structure of the perceptual space. When analyzing experimental data, knowing whether the results are close to the lower or upper bound allows the dimension and structure of the response space to be determined (Do they lie on a line or at the vertices of an n-simplex?).

Both rate distortion theory (RDT) (Cover & Thomas, 1991) and metrical information analysis look for a compression of the responses such that the information about the stimuli is preserved. The compression is to be realized by a neural decoder, that is, a layer of downstream neurons. The properties of the neural decoder impose some constraints on the possible compressions. RDT methods aim at identifying the features of the code that are most relevant to information transmission while discarding the irrelevant ones so as to compress the code. Metrical information analysis tries to find the optimal way in which neurons can look at the code (so as to minimize the noise entropy and maximize the information) given the constraints. These constraints are captured by the estimation of distances between spike trains. Indeed, the similarity measure is subordinated to distance estimation. It is then possible to infer the optimal parameters of the neurons (e.g., the membrane time constant) to transfer information while compressing the signal. In this framework, setting the critical distance so as to minimize the conditional entropy amounts to stressing the importance of the compression of the code besides the importance of input discriminability. The relaxation of this optimality constraint may lead to biologically plausible trade-offs between compression and information transfer (currently under examination; see the example in Figure 7).

From a machine learning perspective, the proposed method relates to statistical linear discrimination (e.g., for dimensionality reduction and classification) such as Fisher discriminant analysis (Fisher, 1936; Friedman, 1989; Mika, Rätsch, Weston, Schölkopf, & Müller, 1999; McLachlan, 2004). Indeed, there exists a qualitative link to the index of discriminability d', which quantifies the separability of two populations of events that are supposed to have identical gaussian conditional probability distributions. The d' measure is defined as the distance between the means of the two distributions, normalized by their standard deviation. The linear combination of parameters that maximizes the ratio d' leads to optimal discrimination (Fisher, 1936). For instance, d' can be used to assess the stimulus discriminability based on the spike count of neural responses for both single-cell and population coding (Petersen, Panzeri, & Diamond, 2002). As for the computation of d', estimating the metrical information $I^*(R; S)$ requires a comparison of the distances between the clusters of events with the distances within a cluster. Notice that instead of providing a continuous measure of discriminability without any upper bound (the ideal case would be reached when the two distributions are infinitely far apart), the measure presented
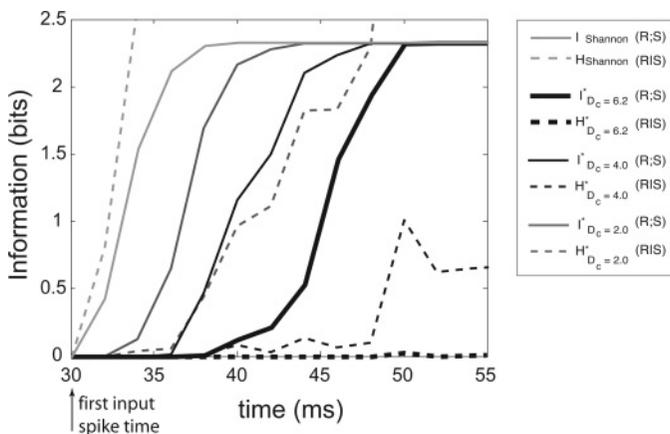
Figure 7: Results. The optimal discrimination constraint can be relaxed by reducing the critical distance, therefore giving relatively more importance to transmission than compression. The metrical information and noise entropy are plotted for three values of the critical distance (optimal $D_c = 6.2$, thick black curves; $D_c = 4$, thin black curves; and $D_c = 2$, dark gray curves), showing the trade-off between transmission and compression. In our analysis, Shannon information (light gray curves) constitutes a limit case of this trade-off, favoring transmission over compression.

here is bounded: two distributions are considered as fully separated as long as their size is smaller than their distance. Once this condition is met, no matter how far they are, their discriminability is estimated to be the same.

From a more general perspective, a link can be drawn between the proposed metrical information analysis and signal detection theory (SDT) (Dayan & Abbott, 2001). SDT offers a set of statistical tools to assess the capacity of neural systems to encode and transmit information (Tolhurst, Movshon, & Dean, 1983; Britten, Shadlen, Newsome, & Movshon, 1992; Guido, Lu, Vaughan, Godwin, & Sherman, 1995; Cheng & Wasserman, 1996; Petersen et al., 2002). In contrast to classical information theoretical principles, SDT makes an extensive use (either implicit or explicit) of metrics. As suggested in Figure 8, the metrical conditional entropy $H^*(R|S)$ can be related to the SDT "false-positive rate" concept, in the sense that when $H^*(R|S) = 0$, two identical stimuli are always correctly identified (0 false alarms). Likewise, $I^*(R; S)$ is akin to the "true positive rate" in the sense that when it is maximum, the hit rate is maximum (i.e., two different inputs are always correctly discriminated). Under the optimal discrimination condition (maximum $I^*(R; S)$ and nil $H^*(R|S)$), the receiver-operating
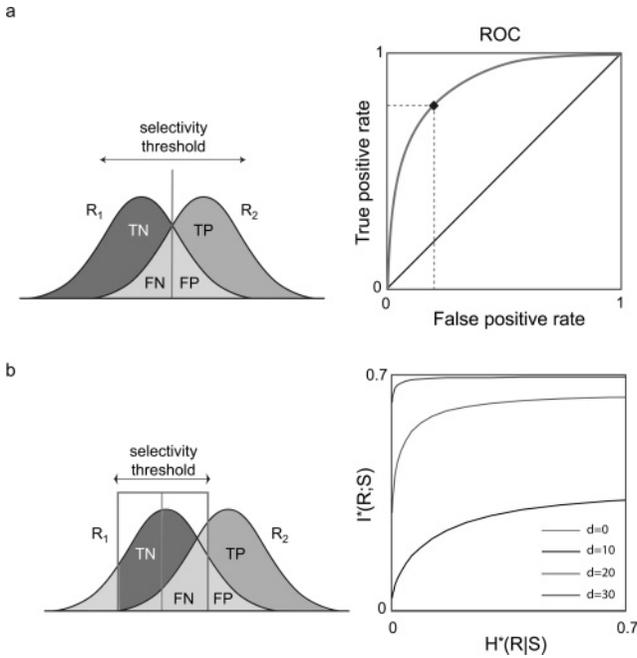
Figure 8: Signal detection theory (SDT) offers a set of statistical tools to evaluate the discrimination efficacy of decision-making processes (e.g., diagnostic tests). One of these tools is the receiver-operating characteristic (ROC) (Barlow & Levick, 1969; Cohn, Green, & Tanner, 1975; Geisler, Albrecht, Salvi, & Saunders, 1991), which can be used to determine the appropriate selectivity (or cutoff) threshold of a classification procedure. (a) Let R_1 and R_2 denote two populations of events characterized by overlapping gaussian distributions. The exploration of all possible cutoff values to discriminate R_1 from R_2 events can be quantified using an ROC analysis, which plots the true positive ("hit") rate (e.g., the fraction of correct classification of R_2 events) against the false-positive ("false alarm") rate (e.g., the fraction of R_1 events incorrectly classified as R_2 events) while varying the selectivity threshold. As a consequence, ROC curves permit trading all the relative importance of true positive and false-positive rates to enhance the accuracy of a discrimination test. Different points on the curve denote different choices of the discriminatory threshold. The diagonal line corresponds to a random guess (less accurate classification). The closer the curve is to the top-left quadrant, the more accurate the discrimination. TP: true positive; FP: false positive; TN: true negative; FN: false negative. (b) The metrical information with respect to the conditional metrical entropy as a function of the critical distance for two gaussians of variance 5 separated by d. The behavior highly resembles that of a classical ROC, as the metrical information (resp. conditional metrical entropy) roughly gives the true positive rate (resp. false-positive rate) of a discrimination task performed by an observer acting according to the similarity measure used.

characteristic (ROC) curve (Barlow & Levick, 1969; Cohn, Green, & Tanner, 1975; Geisler, Albrecht, Salvi, & Saunders, 1991) as a function of the critical distance follows the 0-miss and 0-false alarm axes. Therefore, the integral of the ROC curve is 1, and so it is the probability of discriminating correctly in a two-alternative forced-choice test (which holds here for any two stimuli) (Dayan & Abbott, 2001). Notice that in metrical information analysis, the emphasis is shifted from the values to the distances between the values. In SDT, the threshold is a certain value of the output parameter, whereas in metrical information analysis, the threshold is a certain distance between two outputs.

6 Conclusion

This letter highlights the importance of endowing an information-theoretical analysis with the capability of taking into full account the metrical relations (e.g., distances) between spike trains in order to quantify neurotransmission reliability. In contrast to Shannon mutual information, the proposed measure does not require any a priori partitioning of the event space into equivalence classes (e.g., temporal binning procedures or clustering of neural responses based on confusion matrices). Rather, the employed metrics (similarity measure) shapes the discretization of the state space over time. We show how the definition of metrical entropy incorporates the distances between events (computed, for instance, based on a Victor-Purpura spike train metrics; Victor & Purpura, 1996). The resultant information quantity, in contrast to Shannon information, is both a parametric and diffeomorphism-variant estimate of the statistical dependence between two variables (e.g., stimulus and neural response).

We put forth the hypothesis that the selectivity level represented by the similarity measure may be tuned to reflect some known properties of the readout system (e.g., actual temporal precision of a neural decoder). By contrast, Shannon information can be understood in terms of a Kronecker-like similarity measure—the selectivity capacity of an ideal observer. As a consequence, in certain situations, Shannon information measurements of the capacity of a neural code might happen to be not biologically relevant due to nonexistent neural decoders able to exploit the whole amount of estimated information (Quian Quiroga & Panzeri, 2009). Since the similarity measure is a projection of the properties of the reader onto the input space, metrical information analysis can help to solve this problem. Indeed, the metrical information $I^*(R; S)$ and conditional entropy $H^*(R|S)$ can be seen as quantitative answers to the question: Is it possible for a readout system that implements a decoding scheme based on the similarity measure $\phi(r, r')$ to perform perfect decoding, that is, maximize $I^*(R; S)$ and minimize $H^*(R|S)$?

References

Arleo, A., Nieus, T., Bezzi, M., D'Errico, A., D'Angelo, E., & Coenen, O. J. M. (2010). How synaptic release probability shapes neuronal transmission: Information theoretic analysis in a cerebellar granule cell. *Neural Comput., 22*(8), 2031–2058.

Aronov, D., Reich, D., Mechler, F., & Victor, J. (2003). Neural coding of spatial phase in v1 of the macaque monkey. *J. Neurophysiol. 89*, 3304–3327.

Barlow, H. B. (1961). The coding of sensory messages. In W. H. Thorpe & O. L. Zangwill (Eds.), *Current problems in animal behaviour* (pp. 331–360). Cambridge: Cambridge University Press.

Barlow, H. B., & Levick, W. R. (1969). Three factors limiting the reliable detection of light by retinal ganglion cells of the cat. *J. Physiol. (Lond.), 200*, 1–24.

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Comput., 7*, 1129–1159.

Bialek, W., Rieke, F., de Ruyter van Steveninck, R. R., & Warland, D. (1991). Reading a neural code. *Science, 252*, 1854–1857.

Blahut, R. S. (1988). *Principles and practice of information theory*. Reading, MA: Addison-Wesley.

Borst, A., & Theunissen, F. E. (1999). Information theory and neural coding. *Nat. Neurosci., 2*, 947–957.

Brenner, N., Strong, S. P., Koberle, R., Bialek, W. W., & de Ruyter van Steveninck, R. R. (2000). Synergy in a neural code. *Neural Comput., 12*, 1531–1552.

Britten, K. H., Shadlen, M. N., Newsome, W. T., & Movshon, J. A. (1992). The analysis of visual motion: A comparison of neuronal and psychophysical performance. *J. Neurosci., 12*, 4745–4765.

Brunel, N., & Nadal, J.-P. (1998). Mutual information, Fisher information, and population coding. *Neural Comput., 10*, 1731–1757.

Buracas, G. T., Zador, A. M., DeWeese, M. R., & Albright, T. D. (1998). Efficient discrimination of temporal patterns by motion-sensitive neurons in primate visual cortex. *Neuron, 20*, 959–969.

Butts, D., & Goldman, M. (2006). Tuning curves, neuronal variability, and sensory coding. *PLoS Biol, 4*, e92.

Cheng, Z., & Wasserman, G. S. (1996). Receiver operating characteristic (ROC) analysis of neural code efficacies. I. Graded photoreceptor potentials and data quality. *Biol. Cybern., 75*, 93–103.

Clarke, B., & Barron, A. (1990). Information-theoretic asymptotics of Bayes methods. *IEEE Trans. Inf. Theory, 36*, 453–471.

Cohn, T. E., Green, D. G., & Tanner, W. P. (1975). Receiver operating characteristic analysis: Application to the study of quantum fluctuation effects in optic nerve of *Rana pipiens*. *J. Gen. Physiol. 66*, 583–616.

Cover, T., & Thomas, J. (1991). *Elements of information theory*. New York: Wiley.

Dayan, P., & Abbott, L. (2001). *Theoretical neurosciences*. Cambridge, MA: MIT Press.

de Ruyter van Steveninck, R., Lewen, G., Strong, S., Koberle, R., & Bialek, W. (1997). Reproducibility and variability in neural spike trains. *Science*, *275*, 1805–1808.

Deco, G., & Obradovic, D. (1997). *An Information-theoretic approach to neural computing*. Berlin: Springer.

DeWeese, M., & Meister, M. (1999). How to measure the information gained from one symbol. *Network*, *10*, 325–340.

Dimitrov, A., & Miller, J. (2000). Natural time scales for neural encoding. *Neurocomputing*, *32*, 1027–1034.

Dimitrov, A., Miller, J., Aldworth, Z., & Gedeon, T. (2001). Non-uniform quantization of neural spike sequences through an information distortion measure. *Neurocomputing*, *38–40*, 175–181.

Dimitrov, A., Miller, J., Gedeon, T., Aldworth, Z., & Parker, A. (2003). Analysis of neural coding using quantization with an information-based distortion measure. *Network*, *14*(1), 151–176.

Duda, O. H., Hart, P. E., & Stork, D. G. (2001). *Pattern classification*. New York: Wiley.

Einhauser, W., Mundhenk, T. N., Baldi, P., Koch, C., & Itti, L. (2007). A bottom-up model of spatial attention predicts human error patterns in rapid scene recognition. *J. Vis.*, *7*, 1–13.

Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugen.*, *7*, 179–188.

Földiák, P. (1990). Forming sparse representations by local anti-Hebbian learning. *Biol. Cybern.*, *64*, 165–170.

Friedman, J. H. (1989). Regularized discriminant analysis. *J. Am. Stat. Assoc., 84*, 165–175.

Fuhrmann, G., Segev, I., Markram, H., & Tsodyks, M. (2002). Coding of temporal information by activity-dependent synapses. *J. Neurophysiol.*, *87*, 140–148.

Gabbiani, F., & Koch, C. (1996). Coding of time-varying signals in spike trains of integrate-and-fire neurons with random threshold. *Neural Comput.*, *8*(1), 44–66.

Geisler, W. S., Albrecht, D. G., Salvi, R. J., & Saunders, S. S. (1991). Discrimination performance of single neurons: Rate and temporal-pattern information. *J. Neurophysiol.*, *66*, 334–362.

Goldman, M. S. (2004). Enhancement of information transmission efficiency by synaptic failures. *Neural Comput.*, *16*, 1137–1162.

Guido, W., Lu, S. M., Vaughan, J. W., Godwin, D. W., & Sherman, S. M. (1995). Receiver operating characteristic (ROC) analysis of neurons in the cat's lateral geniculate nucleus during tonic and burst response mode. *Vis. Neurosci.*, *12*, 723–741.

Houghton, C., & Sen, K. (2008). A new multineuron spike train metric. *Neural Comput.*, *20*(6), 1495–1511.

Huetz, C., Del Negro, C., Lebas, N., Tarroux, P., & Edeline, J.-M. (2006). Contribution of spike timing to the information transmitted by HVC neurons. *Eur. J. Neurosci.*, *24*(4), 1091–1108.

Johansson, R., & Birznieks, I. (2004). First spikes in ensembles of human tactile afferents code complex spatial fingertip events. *Nat. Neurosci.*, *7*, 170–177.

Johansson, R. S., & Flanagan, J. R. (2009). Coding and use of tactile signals from the fingertips in object manipulation tasks. *Nat. Rev. Neurosci.*, *10*, 345–359.

Kreuz, T., Haas, J. S., Morelli, A., Abarbanel, H. D. I., & Politi, A. (2007). Measuring spike train synchrony. *J. Neurosci. Methods*, *165*, 151–161.

Linsker, R. (1988). An application of the principle of maximum information preservation to linear systems. In D. Touretzky (Ed.), *Advances in neural information processing systems, 1* (pp. 186–194). San Francisco: Morgan Kaufman.

London, M., Schreibman, A., äusser, M. H, Larkum, M. , & Segev, I. (2002). The information efficacy of a synapse. *Nat. Neurosci.*, *5*(4), 332–340.

Lu, T., & Wang, X. (2004). Information content of auditory cortical responses to time-varying acoustic stimuli. *J. Neurophysiol.*, *91*, 310–313.

MacKay, D., & McCulloch, W. (1952). The limiting information capacity of a neuronal link. *Bull. Math Biol.*, *14*(2), 127–135.

Manwani, A., & Koch, C. (2001). Detecting and estimating signals over noisy and unreliable synapses: Information-theoretic analysis. *Neural Comput.*, *13*, 1–33.

Manwani, A., Steinmetz, P. N., & Koch, C. (2002). The impact of spike timing variability on the signal-encoding performance of neural spiking models. *Neural Comput.*, *14*, 347–367.

McLachlan, G. J. (2004). *Discriminant analysis and statistical pattern recognition*. New York: Wiley Interscience.

Mika, S., Rätsch, G., Weston, J., Schölkopf, B., & Müller, K.-R. (1999). Fisher discriminant analysis with kernels. In Y.-H. Hu, J. Larsen, E. Wilson, & S. Douglas (Eds.), *Neural networks for signal processing IX* (pp. 41–48). Piscataway, NJ: IEEE.

Nadal, J. P., Brunel, N., & Parga, N. (1998). Nonlinear feedforward networks with stochastic outputs: Infomax implies redundancy reduction. *Network*, *9*, 207–217.

Nadal, J. P., & Toulouse, G. (1998). Information storage in sparsely coded memory nets. *Network*, *1*, 61–74.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*, 607–609.

Panzeri, S., Petersen, R., Schultz, S., Lebedev, M., & Diamond, M. (2001). The role of spike timing in the coding of stimulus location in rat somatosensory cortex. *Neuron*, *29*, 769–777.

Panzeri, S., Senatore, R., Montemurro, M. A., & Petersen, R. S. (2007). Correcting for the sampling bias problem in spike train information measures. *J. Neurophysiol.*, *98*, 1064–1072.

Paz, R., & Vaadia, E. (2004). Learning-induced improvement in encoding and decoding of specific movement directions by neurons in the primary motor cortex. *PLoS Biol 2*, E45.

Petersen, R., Panzeri, S., & Diamond, M. (2002). Population coding in somatosensory cortex. *Curr. Opin. Neurobiol.*, *12*, 441–447.

Pouget, A., Dayan, P., & Zemel, R. (2000). Information processing with population codes. *Nat. Rev. Neurosci.*, *1*, 125–132.

Quian Quiroga, R., Kreuz, T., & Grassberger, P. (2002). Event synchronization: A simple and fast method to measure synchronicity and time delay patterns. *Phys. Rev. E*, *66*, 041904.

Quian Quiroga, R., & Panzeri, S. (2009). Extracting information from neuronal populations: Information theory and decoding approaches. *Nat. Rev. Neurosci.*, *10*, 173–185.

Rainer, G., Lee, H., & Logothetis, N. K. (2004). The effect of learning on the function of monkey extrastriate visual cortex. *PLoS Biol*, *2*, E44.

Reich, D. S., Mechler, F. , & Victor, J. D. (2001). Independent and redundant information in nearby cortical neurons. *Science*, *294*, 2566–2568.

Richmond, B. J., & Optican, L. M. (1990). Temporal encoding of two-dimensional patterns by single units in primate primary visual cortex. II. Information transmission. *J. Neurophysiol.*, *64*, 370–380.

Rieke, F., Warland, D., de Ruyter van Steveninck, R. , & Bialek, W. (1997). *Spikes: Exploring the neural code*. Cambridge, MA: MIT Press.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Trans. Inf. Theory*, *42*, 40–47.

Saal, H., Vijayakumar, S., & Johansson, R. (2009). Information about complex fingertip parameters in individual human tactile afferent neurons. *J. Neurosci.*, *29*(25), 8022–8031.

Schneidman, E., Slonim, N., Tishby, N., de Ruyter van Steveninck, R. R., & Bialek, W. (2002). Analyzing neural codes using the information bottleneck method. (Tech. Rep.) Jerusalem: Hebrew University.

Schreiber, S., Fellous, J. M., Whitmer, S., Tiesinga, P., & Sejnowski, T. (2003). A new correlation-based measure of spike timing reliability. *Neurocomputing*, *52–54*, 925–931.

Shannon, C. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.*, *27*, 379–423.

Sharpee, T. O., Sugihara, H., Kurgansky, A. V., Rebrik, S. P., Stryker, M. P., & Miller, K. D. (2006). Adaptive filtering enhances information transmission in visual cortex. *Nature*, *439*, 936–942.

Smith, E., & Lewicki, M. (2006). Efficient auditory coding. *Nature*, *439*, 978–982.

Strong, S. P., Koberle, R., de Ruyter van Steveninck, R. R., & Bialek, W. (1997). Entropy and information in neural spike trains. *Phys. Rev. Lett.*, *80*(1), 197–200.

Theunissen, F., & Miller, I. (1991). Representation of sensory information in the cricket cercal sensory system. II. Information theoretic calculation of system accuracy and optimal tuning-curve widths of four primary interneurons. *J. Neurophysiol.*, *66*, 1690–1703.

Thomson, E. E., & Kristan, W. B. (2005). Quantifying stimulus discriminability: A comparison of information theory and ideal observer analysis. *Neural Comput.*, *17*, 741–778.

Tiesinga, P. H. (2001). Information transmission and recovery in neural communication channels revisited. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, *64*, 012901.

Tiesinga, P. H., Fellous, J. M., José, J. V., & Sejnowski, T. J. (2002). Information transfer in entrained cortical neurons. *Network*, *13*, 41–66.

Tishby, N., Pereira, F., & Bialek, W. (1999). The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing* (pp. 368–377). Urbana: University of Illinois.

Tolhurst, D. J., Movshon, J. A., & Dean, A. F. (1983). The statistical reliability of signals in single neurons in cat and monkey visual cortex. *Vision Res.*, *23*, 775–785.

Treves, A. (1997). On the perceptual structure of face space. *Biosystems*, *40*, 189–196.

Van Rossum, M. (2001). A novel spike distance. *Neural Comput.*, *13*, 751–763.

Victor, J. D., & Nirenberg, S. (2008). Indices for testing neural codes. *Neural Comput.*, *20*, 2895–2936.

Victor, J., & Purpura, K. (1996). Nature and precision of temporal coding in visual cortex: A metric-space analysis. *J. Neurophysiol.*, *76*, 1310–1326.

Willmore, B., & Tolhurst, D. J. (2001). Characterizing the sparseness of neural codes. *Network*, *12*, 255–270.

Zador, A. (1998). Impact of synaptic unreliability on the information transmitted by spiking neurons. *J. Neurophysiol.*, *79*, 1219–1229.