Information-Theoretic Analysis of Neural Coding

Don H. Johnson[*†], Charlotte M. Gruner[*], Keith Baggerly[†], Chandran Seshagiri[*]
Computer and Information Technology Institute
Department of Electrical and Computer Engineering[*]
Department of Statistics[†]
Rice University, MS 366, Houston, Texas 77251–1892
dhj@rice.edu, cmgruner@rice.edu, kabagg@rice.edu, cseshag@rice.edu

Revised

July 27, 2000

Abstract

We describe an approach to analyzing single- and multi-unit (ensemble) discharge patterns based on information-theoretic distance measures and on empirical theories derived from work in universal signal processing. In this approach, we quantify the difference between response patterns, be they time-varying or not, using information-theoretic distance measures. We apply these techniques to single and multiple unit processing of sound amplitude and sound location. These examples illustrate that neurons can simultaneously represent at least two kinds of information with different levels of fidelity. The fidelity can persist through a transient and a subsequent steady-state response, indicating that it is possible for an evolving neural code to represent information with constant fidelity.

1 Introduction

Neural coding has been classified into two broadly defined types: rate codes—the average rate of spike discharge—and timing codes—the timing pattern of discharges. Debates rage over the effectiveness of one code over another and over whether this categorization even applies. Complicating the debate are recent results that suggest single-unit responses inadequately explain information coding. For example, theoretical considerations indicate that single-neuron discharge patterns in the mammalian auditory pathway are too random to effectively represent sound [19]. Coordinated responses of neurons within sensory and motor nuclei have been found, with discharge timing relations among neural outputs having significance [1, 3, 8, 22, 24, 27, 29]. Consequently, much current work has focused on population activity, using the fundamental assumption that *coordinated sequences of action potential occurrence times produced by groups of neurons collectively represent the stimulus-response relationship*. Thus, today the "neural code" is taken to mean how groups of neurons, responding individually and collectively, represent sensory information with their discharge patterns [5]. Knowing the code would unlock the secrets of how neurons, working in concert, process and represent information.

In developing data analysis techniques that seek to elucidate the neural code, some kind of averaging is required because of the stochastic nature of the neural response. The simplest techniques require *stationary* responses: The probability law governing the response must not change with time. Among these are several interspike interval statistics [19], and auto- and cross-correlation techniques [2]. To quantify time-varying responses, responses are averaged over several stimulus presentations that are spaced sufficiently far apart in time to prevent adaptation and sequential stimulus effects. Such responses are said to be *cyclostationary* [12]: The probability law varies with time but does so periodically, the period equaling the inter-stimulus interval. The well-known PST histogram measures, under a Poisson point process assumption, how the discharge rate changes after each stimulus presentation [19]. All of these measures spring from a point process view of neural discharge patterns. Using point-process-based measures and elaborations of them, we could in principle estimate the point process model that accurately describes a neuron's response to each stimulus. In our experience, having such a point process model does not reveal what stimulus features are being coded and when, and how effective the code is in representing the stimulus. For example, we developed a point process model for the tone-burst responses of single units located in the lateral superior olive (LSO) [31, 32] and an underlying computational biophysical model [33]. While these models provide a notion of the response's structure, they do *not* help us determine the typical LSO unit's information processing role and what processing function the variety of LSO response types may engender. What has been left out is quantifying both the response's significance and its effectiveness in representing sensory information.

New techniques are emerging that take a broader view. One developed by Victor and Purpura (1997) measures the distance between two responses by determining the number of steps it takes to systematically transform one discharge pattern into another. Mutual information calculations [11, 23] measure how effectively a spike train encodes the stimulus under Poisson assumptions. Neural network models have been trained on recorded neural responses to determine how well these responses can be used to determine stimulus parameters [21]. These methods yield response metrics that are difficult to relate to how well systems can extract information from single- or multiple-neuron responses. When inputs to a neural population have been measured and characterized, how do we judge how well the population extracts sensory information? In fact, how would we know if a population did extract information or simply served as a relay, passing the information along (possibly in a different neural code) to its projections? We must be able to quantify the *neural code*: What aspect of a neural ensemble's collective output represents information and what is the fidelity of this representation? This paper describes a new information-theoretic technique for measuring when and to what degree responses differ in a way that can be related to how optimal systems perform in decoding neural responses.

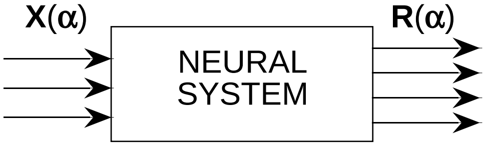

Figure 1: A neural system has as inputs the vector quantity $\mathbf{X}$ that depend on a collection of stimulus parameters denoted by the vector $\boldsymbol{\theta}$. The output $\mathbf{R}$ thus also depends on the stimulus parameters. Both input and output implicitly depend on time. Note that the precise nature of the input is deliberately unclear. It can represent the stimulus itself or a population's collective input.

Consider the simple system shown in Figure 1. Conceptually, this system accepts inputs $\mathbf{X}$ that represent a stimulus or a neural population conveying information (parameterized by $\boldsymbol{\theta}$) and produces outputs $\mathbf{R}$ that collectively code some or all of stimulus attributes. The boldfaced symbols represent vectors, and are intended to convey the notion that our system—a neural ensemble—has multiple inputs and multiple outputs. Presumably, input stimulus features preserved in the output are those extracted by the system; those de-emphasized in the output are discarded by the system and define its feature extraction properties. To probe the system and its representation of sensory information, we experimentally measure the system's output and its inputs as we vary stimulus conditions.

Relating input and output for each condition amounts to measuring the intensity of an underlying point process model, which does not help quantify the effectiveness of neural coding. Instead, what we look for is how the inputs and the outputs *change* as the stimulus undergoes a controlled change. No change means no coding of the perturbed aspect of the stimulus; the bigger the change, the more the system accentuates that sensory aspect. To quantify change, we need a measure that quantifies its degree. In short, what we seek is a *distance measure*: Given two sets of stimulus conditions $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$, we need to measure how different the corresponding responses $\mathbf{R}(\boldsymbol{\theta}_1)$, $\mathbf{R}(\boldsymbol{\theta}_2)$ are—how far apart they are—with some distance metric $d\big(\mathbf{R}(\boldsymbol{\theta}_1), \mathbf{R}(\boldsymbol{\theta}_2)\big)$. Assuming that population codes are subtle, this metric needs to apply to ensemble responses, to dynamic as well as steady-state responses, to changes in transneural correlations, and to changes in temporal correlation structure.

Anticipating our results, Figure 2 demonstrates quantifying how two responses differ. The distance computation does *not* have any *a priori* notion of the neural code and is based on the data, *not* the PST histogram. What we want the distance measure to uncover is (1) what portion of the response most represents the stimulus change, (2) whether the later constant-rate response reveal more or less than the early time-varying response, and (3) how well the stimulus change is represented by the response change? By plotting how distance accumulates with time since stimulus presentation, we can answer these questions. If the distance measure plateaus, the responses don't differ over that time interval; if it increases, the responses differ in their statistical structure and the sharper the increase, the more the responses differ during that time segment than in other segments. First of all, we see that the response begins at about a 3 ms latency, but the distance measure suggests that they don't differ significantly until after 5 ms. Between 5 and 10 ms, the distance accumulates about 2 bits, and from 10 to 25 ms, the distance increases by another 2 bits.[1] Thus, the response during the 5–10 ms intervals reveals much about the amplitude change while the constant rate portion reveals just as much but over a time interval three times longer. From this example, we see that the distance calculation applies to both time-varying and constant rate responses. The distance measurement applies to single and multineuron responses equally as well. What the distance measure does not reveal is the precise nature of amplitude coding by this neuron; it only assesses the code's quality and when the coding occurs. However, just because some component of the neural response more accurately represents some stimulus component that others does not mean that destination populations make more effective use of that response component than others. As we shall see, results from information theory can be used to

[1] Our distance measure has units of bits only because we use base-2 logarithms in its computation and does not imply an information rate. We describe in succeeding sections how to interpret distance values.

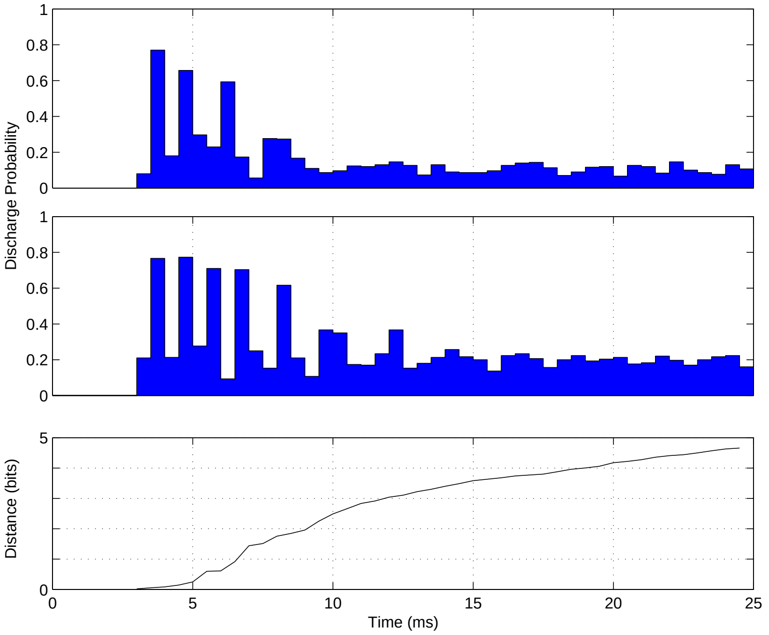

Figure 2: The PST histograms shown in the upper two panels show the dramatically time-varying nature of a simulated auditory neuron's to brief tone pulses that had different amplitudes. Shown separately in the bottom panel is the distance between these cyclostationary responses. Distance between responses accumulates with post-stimulus time, meaning that the distance a post-stimulus time t is the distance between response measured up to time t. Detailed analysis results are shown in Figure 5 for the same simulations.

determine the limits to which systems can extract information no matter what response component(s) they may use.

2 Information-theoretic distance measures

While the merits of one distance measure versus another can be debated [4], we describe here a collection of information-theoretic distances that have a clear, intuitive mathematical foundation. The underlying theory is not rooted in the classic results of Shannon, but in modern classification theory, the key results from which are detailed in the appendix. In this theory, we try to assign a response to one of a set of pre-assigned response categories. For example, discerning whether the stimulus is on or off is a two-category classification problem. The ease of classification depends on how different the categories are; it is through this aspect of the classification problem that distance measures arise. We use this classification theoretic approach because recent results from universal signal processing[2] provide distance measures and classification techniques that assume little about the data yet yield (in a certain sense) optimal classification results. In addition to information-theoretic distance measures having a strong mathematical foundation, direct empirical results have been derived. For example, we can determine how complex a data analysis we can perform given a certain amount of data.

Error probabilities in optimal classifiers decrease exponentially with the distance between the categories. Using $\Pr[\text{error}]$ to denote a generic classification error probability, $\Pr[\text{error}] \sim 2^{-d(C_1,C_2)}$, where $d(C_1, C_2)$ denotes an information-theoretic distance between two categories. The distance measure of particular interest here is the *Kullback-Leibler* distance defined to be

$$\mathcal{D}(p_1 \| p_2) = \int p_1(\mathbf{R}) \log \frac{p_1(\mathbf{R})}{p_2(\mathbf{R})} \, d\mathbf{R} \ , \tag{1}$$

[2] The theory surrounding how to process information universally without much regard to the underlying distribution of the data.

where p_1, p_2 are probability distributions that characterize the two categories and $\mathbf{R}$ symbolically represents a neural response, be it from one or several neurons. Following the convention of information theory, we use the base-two logarithm, which means that distance has units of bits. In the Gaussian case (categories defined to have different means but the same variance), the Kullback-Leibler distance equals ${d'}^2/(2\ln 2)$ bits, with $d' = |m_1 - m_2|/\sigma$. The quantity d' is frequently used in psychophysics to assess how easily stimuli can be distinguished. When applied to non-Gaussian problems, the Kullback-Leibler distance represents a generalization of d' to all binary classification problems: The larger this distance, the easier the classification problem. It measures how different two probability distributions are, and it has several important properties.

1. $\mathcal{D}(p_1\|p_2) \geq 0$ and $\mathcal{D}(p\|p) = 0$.
   The Kullback-Leibler distance is always non-negative, with zero distance occurring only when the probability distributions are the same.

2. $\mathcal{D}(p_1\|p_2) = \infty$ whenever, for some R domain, $p_2(R) = 0$ and $p_1(R) \neq 0$. If $p_1(R) = 0$, the value of $p_1(R)\log\frac{p_1(R)}{p_2(R)}$ is defined to be zero.

3. When the underlying stochastic quantities are random vectors having statistically independent components with respect to both p_1 and p_2, the Kullback-Leibler distance equals the sum of the component distances. Stated mathematically, if $p_1(\mathbf{R}) = \prod_i p_1(R_i)$ and $p_2(\mathbf{R}) = \prod_i p_2(R_i)$,

   $$\mathcal{D}(p_1(\mathbf{R})\|p_2(\mathbf{R})) = \sum_i \mathcal{D}(p_1(R_i)\|p_2(R_i)) \ . \tag{2}$$

   Furthermore, if p_1, p_2 describe Markovian data,[3] the Kullback-Leibler distance has a similar summation property. Taking the first-order Markovian case as an example, wherein $p_1(\mathbf{R}) = p_1(R_1)\prod_i p_1(R_{i+1}|R_i)$ and $p_2(\cdot)$ has a similar structure,

   $$\mathcal{D}(p_1(\mathbf{R})\|p_2(\mathbf{R})) = \mathcal{D}(p_1(R_1)\|p_2(R_1)) + \sum_i \mathcal{D}(p_1(R_{i+1}|R_i)\|p_2(R_{i+1}|R_i)) \ . \tag{3}$$

   where

   $$\mathcal{D}(p_1(R_{i+1}|R_i)\|p_2(R_{i+1}|R_i)) = \int p_1(R_i, R_{i+1}) \log\frac{p_1(R_{i+1}|R_i)}{p_2(R_{i+1}|R_i)}\, dR_i\, dR_{i+1} \tag{4}$$

4. $\mathcal{D}(p_1\|p_2) \neq \mathcal{D}(p_2\|p_1)$.
   The Kullback-Leibler distance is usually not a symmetric quantity. In some special cases, it can be symmetric (like the just described Gaussian example), but symmetry cannot, and should not, be expected.

5. $\mathcal{D}(p(x_1, x_2)\|p(x_1)p(x_2)) = I(x_1; x_2)$.
   The Kullback-Leibler distance between a joint probability density and the product of the marginal distributions equals what is known in information theory as the *mutual information* between the random variables x_1, x_2. From the properties of the Kullback-Leibler distance, we see that the mutual information equals zero only when the random variables are statistically independent.

The word "distance" should appear in quotes because $\mathcal{D}(\cdot\|\cdot)$ violates some of the fundamental properties a distance metric must have: A distance *must* be symmetric in its arguments. As explained in the appendix, classification error probabilities need not have the same exponential decay rate and this results

[3] A sequence of random variables is D^{th}*-order Markov* if the conditional probability of any element of the sequence given values for the previous ones depends *only* on the D most previous: $p(R_i|R_{i-1}, R_{i-2}, \dots) = p(R_i|R_{i-1}, R_{i-2}, \dots, R_{i-D})$.

in the Kullback-Leibler distance's asymmetry. This asymmetry property does not hinder theoretical developments but does affect measuring the distance between recorded neural responses. Consequently, we later propose a symmetric distance measure directly related to the Kullback-Leibler distance. The Gaussian example also indicates that the Kullback-Leibler distance has the form of a *squared-distance*: These distances are proportional to $\sum_i (m_1^{(i)} - m_2^{(i)})^2$, which corresponds to the *square* of the Euclidean distance. Thus, we have a second reason to put distance in quotes.

In addition to quantifying the exponential decay rate of the error probabilities in optimal classifiers, information-theoretic distances determine the ease of estimating parameters represented by the data. Consider the situation where two categories differ slightly according to the values of a scalar parameter θ: symbolically, $p_1(R) = p_\theta(R)$ and $p_2(R) = p_{\theta+\delta\theta}(R)$. Intuitively, if we can easily distinguish between two such categories (small error probabilities), we should also be able to estimate the parameter accurately (smaller estimation error). For sufficiently small values of the difference $\delta\theta$, the Kullback-Leibler distance is proportional to the reciprocal of the smallest mean-squared estimation error that can be achieved. The mathematical results are[4]

$$\mathcal{D}(p_{\theta+\delta\theta} \| p_\theta) \approx \frac{1}{2\ln 2} F(\theta)(\delta\theta)^2 \tag{5}$$

Here, $F(\theta)$ denotes the *Fisher information*.

$$F(\theta) = \mathcal{E}\left[\left(\frac{\partial \ln p_\theta(R)}{\partial\theta}\right)^2\right] = \int \left(\frac{\partial \ln p_\theta(R)}{\partial\theta}\right)^2 p_\theta(R)\, dR$$

with $\mathcal{E}[\cdot]$ denoting expected value. The significance of these formulas rests in the *Cramér-Rao bound*, which states that the mean-squared error for *any* unbiased estimator $\widehat{\theta}$ of θ cannot be smaller than $1/F(\theta)$ [15: §6.2.4].

$$\mathcal{E}\left[(\widehat{\theta} - \theta)^2\right] \geq \frac{1}{F(\theta)} \tag{6}$$

When two or more parameters change, Fisher information becomes a matrix, and the distance formulas become what are known as quadratic forms.

$$\mathcal{D}(p_{\boldsymbol{\theta}+\boldsymbol{\delta\theta}} \| p_{\boldsymbol{\theta}}) \approx \frac{1}{2\ln 2} \boldsymbol{\delta\theta}' \mathbf{F}(\boldsymbol{\theta}) \boldsymbol{\delta\theta} \tag{7}$$

with $\mathbf{F}(\boldsymbol{\theta}) = \mathcal{E}\left[(\nabla_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(R))\,(\nabla_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(R))'\right]$. Here, $(\cdot)'$ means transpose and $\nabla_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(R)$ means the gradient of the log probability density function: $\nabla_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\theta}}(R) = \mathrm{col}\left[\frac{\partial}{\partial\theta_1} \ln p_{\boldsymbol{\theta}}(R), \ldots, \frac{\partial}{\partial\theta_N} \ln p_{\boldsymbol{\theta}}(R)\right]$. The Cramér-Rao bound still holds, but in a more complicated form.

$$\mathcal{E}[(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})'] \geq \mathbf{F}^{-1}(\boldsymbol{\theta})$$

This result means that the mean-squared estimation error for any one parameter must be greater than the corresponding diagonal entry in the *inverse* of the Fisher information matrix: $\mathcal{E}[(\widehat{\theta}_i - \theta_i)^2] \geq \left(\mathbf{F}^{-1}\right)_{ii}(\boldsymbol{\theta})$. Thus for any given stimulus parameter perturbation $\boldsymbol{\delta\theta}$, the larger the Kullback-Leibler distance becomes (the further apart the distributions become), the larger the Fisher information (equation 5), and the least possible mean-squared error in estimating the parameter becomes proportionally smaller. In short, *larger distances mean smaller estimation errors*. This relationship not only reinforces the notion that our distance measures do indeed measure how distinct two classification categories are, but also allows us to determine how well parametric information can be gleaned from data. Information-theoretic distance measures assess the limits of information processing.

[4] The term $\ln 2$ arises because the definition of Kullback-Leibler distance (1) uses $\log_2$ and the definition of Fisher information uses natural logarithms.

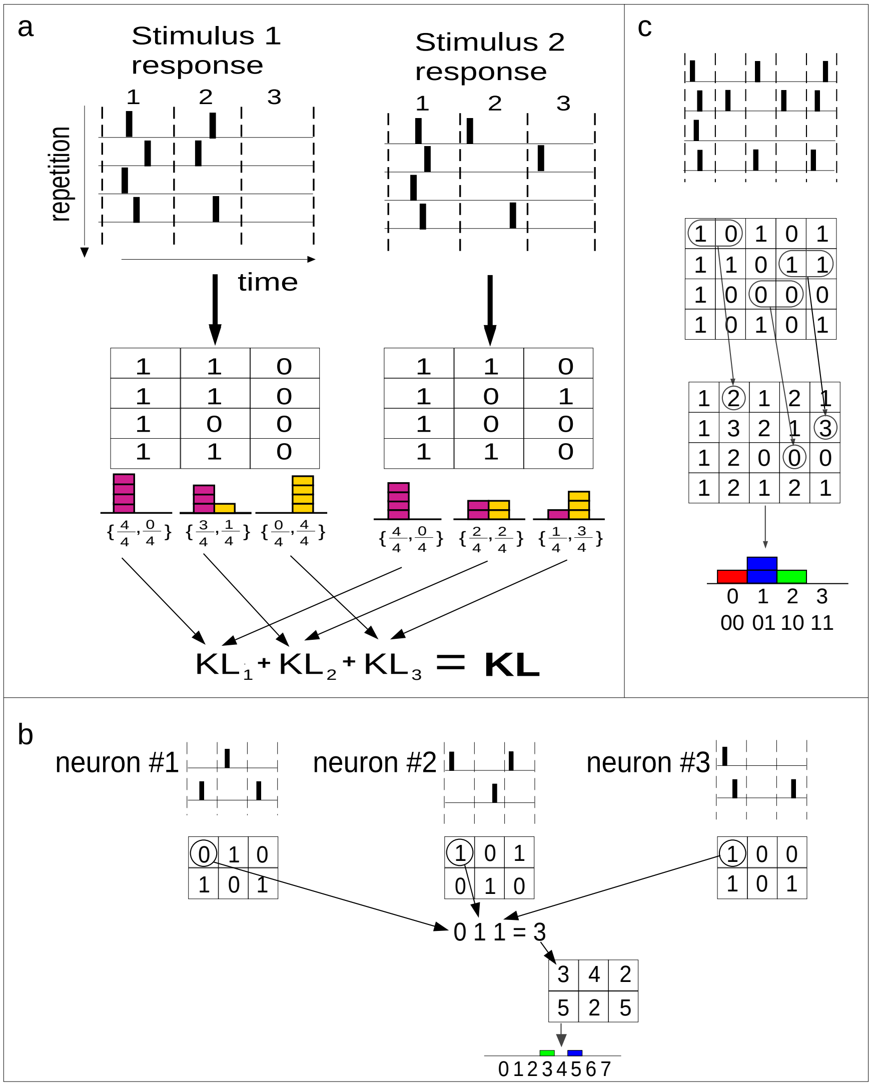

Figure 3: Panel (a) portrays how we estimate the Kullback-Leibler distance between single neuron responses to different stimuli. The response to each stimulus repetition is time aligned as in PST histogram computation, and a table is formed from the spike occurrence times (denoted by an "×") quantized to the binwidth Δ. For each bin, a "1" indicates that a spike occurred in a bin and a "0" indicates that a spike did not occur. We accumulate the type for each bin, forming a histogram of spike occurrences and nonoccurrences separately for each bin from the M stimulus presentations (four are shown in the figure). A similar set of types is computed from the responses to a second stimulus. When we assume the Markov order D to be zero, we compute the Kullback-Leibler distance between corresponding bins and sum the results. Panel (b) generalizes the computations of panel (a) to the multi-neuron case. In the depicted ensemble of three neurons, the spike pattern at any bin could be one of eight ($2^3 = 8$) possible patterns. Each possible pattern is represented by an integer between 0 and 7 in the table to the lower right. Types are formed from these quantities and distances are again computed separately between corresponding bins and summed when $D = 0$. Panel (c) illustrates how first-order distance analysis is computed. For each neuron's (or ensemble's) responses, the response pattern for two bins at a time is represented by an integer between 0 and 3 in the bottom table. Note that the first bin is special as no bin precedes it. This edge effect corresponds to the first term in equation (3). We compute the zeroth order distance for it and the first-order distances for the others, then sum the result to form the total distance.

3 Digital representation of neural responses

To develop a measure of the population's response, we first convert the population's discharge pattern into a convenient representation for computational analysis (figure 3b). Here, a neural population's response during the b^{th} bin is summarized by a single number R_b that equals a binary coding for which neurons, if any, discharged during the bin. This procedure generalizes the approach taken for the single neuron case, wherein the occurrence of a spike in a bin was represented by a zero or a one. Note that this digitization process for a neural ensemble is reversible (up to the temporal precision of the binwidth): The sequence $\mathbf{R} = \{R_1, \ldots, R_b, \ldots, R_B\}$ completely characterizes the population response, and the entire discharge pattern can be recreated from it. *In developing techniques to analyze neural coding, we need only consider the statistical structure of this sequence*. When we present a stimulus periodically M times, we form the dataset $\{\mathbf{R}_1, \ldots, \mathbf{R}_M\}$ from the component responses. Here, the response to the m^{th} stimulus is $\mathbf{R}_m = \{R_{m,1}, \ldots, R_{m,B}\}$. Because we repeat the same stimulus and take the usual precautions to mitigate adaptation, the resulting response is cyclostationary [12] (each $\mathbf{R}_m$ obeys the same probability law) and, in addition, they are statistically independent of each other.

In information theory terminology, a discrete-valued random variable, such as the response in a bin, takes on values that are *letters* r drawn from an *alphabet* $\mathcal{A}$. When we have a N-neuron population and a sufficiently small binwidth, R_b takes on values from the alphabet $\{r_0, \ldots, r_{K-1}\} = \{0, \ldots, 2^N - 1\}$, For the three-neuron population exemplified in Figure 3, the collection $\{0, 1, 2, 3, 4, 5, 6, 7\}$ forms the alphabet. The letter $r = 3 = 011_2$ means that a discharge occurred in both neurons #2 and #3 and not in neuron #1 during a particular bin. We could form a *population PST histogram* of the population's response at a particular bin to estimate the probability that $\Pr[R_b = r_k]$. Information theorists term such histograms estimates of probabilities *types* [7: Chap. 12].

$$\widehat{P}[R_b = r_k] = \frac{(\#\text{times } r_k \text{ occurs in } \{R_{1,b}, \ldots, R_{M,b}\})}{M}$$

By accumulating this multineuron PST histogram in this way, we obtain the distribution of neural discharge occurrence across the entire population within each bin. This histogram generalizes the PST histogram used to analyze single-neuron responses [19]. In the single-neuron case, the alphabet consists of $\{0, 1\}$, and only the probability of one letter need be calculated. The PST histogram consists of a type at each bin that estimates the probability of a discharge (the letter 1) occurring. The only other remaining value of the type — $\widehat{P}_{R_b}(R_b = 0)$ — is found by subtracting $\widehat{P}_{R_b}(R_b = 1)$ from one.

Just as in the usual PST histogram, this multiunit PST histogram does not faithfully represent temporal dependence that may be present in the ensemble response [17]. The multiunit PST histogram essentially assumes responses occur independently from bin to bin — what amounts to a Poisson assumption — because no record is kept of what preceded a particular population discharge pattern in each bin when the type is calculated. This assumption is more serious here than in the single-unit case: While departures from Poisson behavior may not be significant in the single-unit case, a discharge in one neuron may well affect another's discharge occurring several bins later. We want our analysis techniques to be sensitive to this possibility, and go beyond the PST histogram in providing insight into the neural code. In the most general case, we should estimate the joint probability of population response across all bins: $\Pr[R_1 = r_{k_1}, R_2 = r_{k_2}, \ldots, R_B = r_{k_B}]$. Because we have 2^N possible letters in each bin and B bins, we need to estimate 2^{NB} probabilities. Most of these will be zero — certain discharge patterns will not occur — but knowing this does not alleviate the overwhelming demand achieving an accurate probability estimate places on data collection.

To *approximate* the temporal dependence structure of the population response, we assume that it has a Markovian structure: The probability of a particular population response in a bin depends only on what responses occurred in the previous D bins. This assumption means that we approximate the joint probability of the neural response by

$$P(\mathbf{R}) = P(R_1, \ldots, R_D) \prod_{b=D+1}^{B} P(R_b | R_{b-1}, \ldots, R_{b-D}) \ .$$

The demands placed on data collection are greatly reduced when we employ this approximation. Recent results in information theory [30] prescribe how much data are needed by *any* bin-based technique to analyze data to a given degree of dependence:

$$D \leq \frac{\log(L+1)}{\log(2^N+1)} \ , \tag{8}$$

where L is the amount of averaging used in analyzing the data and N the number of neurons. For nonstationary responses L equals the number of stimulus repetitions M while for stationary (constant rate) responses it equals the number of bins in the measured response. This result makes the point that the amount of data needed grows exponentially in the dependence order and in the number of neurons in the ensemble: $L > 2^{D \cdot N}$. Computational experiments indicate that this bound is not particularly tight, and we do not analyze data to as high an order as the bound permits.

As shown in figure 3c, temporal dependence is easily incorporated into type-based distance calculations. For each bin, a joint type estimates the joint probability that a given ensemble response pattern occurs in it and the preceding D bins. Extending our example, rather than just counting the number of times R_b assumes various values, we need to know the joint distribution of (R_{b-1}, R_b) to assess first-order Markovian dependence. Because the PST histogram is equivalent to zeroth-order analysis, employing joint types in measuring response differences can reveal response changes *not* revealed by the PST histogram, be it a single- or multi-unit histogram. Temporal dependence in discharge probabilities can arise in a variety of ways: among them are dependence on discharge history [18, 31, 32], non-exponential interval distributions[5], and syn-fire response patterns in ensembles [1, 22]. The Markov dependence D corresponds to a temporal analysis window $D\Delta$ s long. The bound on D determines the longest interval over which interneuron and intraneuron response dependencies can be included in the analysis. For a given amount of data, the only way to extend the analysis interval is to use a larger binwidth, which means that temporal precision is reduced.

4 Calculating distance between responses

Let $\mathbf{R}(\boldsymbol{\theta}_1)$ and $\mathbf{R}(\boldsymbol{\theta}_2)$ represent the responses of a neural population to two stimulus conditions parameterized by $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$. What we want to measure is the distance between the *joint* probability distributions corresponding to these responses. Using the Kullback-Leibler distance as an example, we would want to find $\mathcal{D}(P_{\boldsymbol{\theta}_1}(\mathbf{R}) \| P_{\boldsymbol{\theta}_2}(\mathbf{R}))$. To manage this statistical complexity, we must assume that the response in a given bin depends (in the statistical and practical sense) *only* on the responses that occur in the immediately preceding D bins. Once this *analysis dependence order* is chosen, the distance calculation generalizes equation (3).

$$\begin{aligned}\mathcal{D}(P_{\boldsymbol{\theta}_1}(\mathbf{R}) \| P_{\boldsymbol{\theta}_2}(\mathbf{R})) \cong\ & \mathcal{D}(P_{\boldsymbol{\theta}_1}(R_1, \ldots, R_D) \| P_{\boldsymbol{\theta}_2}(R_1, \ldots, R_D)) \\ & + \sum_{b=D+1}^{B} \mathcal{D}(P_{\boldsymbol{\theta}_1}(R_b | R_{b-1}, \ldots, R_{b-D}) \| P_{\boldsymbol{\theta}_2}(R_b | R_{b-1}, \ldots, R_{b-D}))\end{aligned} \tag{9}$$

The "$\cong$" relation means that this relation is only true according to assumption, and the data's actual dependence structure may differ. If D equals or exceeds the memory present in the responses, this equation holds: Picking D too large does not hurt. The problem arises when D is chosen too small; in this case the two sides of equation (9) are not equal. Mathematical analysis suggests that Kullback-Leibler distances calculated using a smaller-than-required dependence order could be smaller or larger than the actual value. Thus, to measure accurately the distance between two responses, distances must be computed using increasingly larger values of D until the calculated values stabilize (don't change with increasing D) or the upper limit

[5] This situation is particularly subtle. Even when the response can be well modeled as a renewal process (interspike intervals are statistically independent from each other), the probability of a discharge in a bin depends on how long ago the previous discharge occurred.

of (8) is reached. We decide when the distance reaches a constant value by employing statistical tests; in the following sections we describe the statistical properties of distances and how to estimate their confidence intervals. On the other hand, if we have insufficient data to reach a stabilizing value of D, we cannot say whether the computed value is a lower or an upper bound. Because temporal dependence in discharge patterns usually spans some time interval, we need to choose a larger binwidth (at the sacrifice of temporal resolution) to span this interval with the same number of bins to obtain accurate distance measurements. We address the binwidth selection problem in section 5.3.

As succeeding examples will show, the Kullback-Leibler distance's asymmetry property—$\mathcal{D}(p\|q) \neq \mathcal{D}(q\|p)$—is indeed real, with the distance between two responses depending on the order the responses appear in the formula. In some applications, a reference stimulus condition occurs naturally, and we want to measure how responses differ from a reference; the Kullback-Leibler distance can be used in such situations (the second argument denotes the reference distribution). Otherwise, the asymmetry becomes a nuisance and we need a symmetric distance measure, like the Chernoff distance defined in the appendix. Calculating the Chernoff distance can be so daunting that we need to consider alternative symmetric distances. One possibility is known as the J-divergence, which equals the average of the two Kullback-Leibler distances that can be defined for two probability distributions.

$$\mathcal{J}(p_1, p_2) = \frac{\mathcal{D}(p_1\|p_2) + \mathcal{D}(p_2\|p_1)}{2} \tag{10}$$

This distance is not as powerful as the Kullback-Leibler or the Chernoff distances in that it only *bounds* the average error probability of an optimal classifier.

$$\lim_{M\to\infty} \frac{\log P_{\mathrm{e}}}{M} \geq -\mathcal{J}(p_1, p_2)$$

Calculations show that this bound can be quite generous (not very tight). Though it may be easy to find (it's the sum of easily calculated Kullback-Leibler distances), but we can only approximately relate it to classification error rates: The J-divergence is overly optimistic.

A more accurate approximation is the so-called *resistor average* of the two Kullback-Leibler distances.

$$\mathcal{R}(p_1, p_2) = \frac{\mathcal{D}(p_1\|p_2)\mathcal{D}(p_2\|p_1)}{\mathcal{D}(p_1\|p_2) + \mathcal{D}(p_2\|p_1)} \tag{11}$$

The origin of the name "resistor average" arises because a simple rewriting of this definition yields a formula that resembles the parallel resistor formula.

$$\frac{1}{\mathcal{R}(p_1, p_2)} = \frac{1}{\mathcal{D}(p_1\|p_2)} + \frac{1}{\mathcal{D}(p_2\|p_1)}$$

This quantity is not arbitrary. Rather it is derived in a way analogous to the Chernoff distance, and half of it approximates the Chernoff distance well: $\mathcal{C}(p_1, p_2) \approx \mathcal{R}(p_1, p_2)/2$. In our Gaussian and Poisson examples of Figure 9, the Chernoff distances are 0.03125 and 0.01796, respectively. The corresponding J-divergences are 0.125 and 0.07192, and half the resistor average values are 0.03125 (exact equality) and 0.01794. Thus, when we want to contrast two response patterns, rather than computing the correct distance measure, the Chernoff distance, we compute the much simpler quantity, the resistor average, instead.

One note on these "distances:" None of the Kullback-Leibler, Chernoff, and resistor-average "distances" can be distances in the mathematical sense. To qualify, a proposed distance $d(a, b)$ must be symmetric ($d(a, b) = d(b, a)$), be stricly positive unless the two arguments to the distance are equal ($d(a, b) > 0$ for $a \neq b$, $d(a, a) = 0$), and obey the triangle inequality ($d(a, b) \leq d(a, c) + d(c, b)$). The Kullback-Leibler distance is not symmetric, and the Chernoff and resistor-average distances don't satisfy the triangle inequality. This mathematical issue does not affect their utility in judging how different two responses are in a meaningful way: All of these measures can be related to the performance of optimal signal processing systems (see the discussion following equation 5 and the appendix).

5 Statistical properties

5.1 Estimation of distance measures

The most direct approach to estimating distance measures is to use types in their definitions.

$$\widehat{\mathcal{D}}(P_{\boldsymbol{\theta}_1}(\mathbf{R})\|P_{\boldsymbol{\theta}_2}(\mathbf{R})) = \mathcal{D}(\widehat{P}_{\boldsymbol{\theta}_1}(\mathbf{R})\|\widehat{P}_{\boldsymbol{\theta}_2}(\mathbf{R}))$$
$$\widehat{\mathcal{R}}(P_{\boldsymbol{\theta}_1}(\mathbf{R}), P_{\boldsymbol{\theta}_2}(\mathbf{R})) = \mathcal{R}(\widehat{P}_{\boldsymbol{\theta}_1}(\mathbf{R}), \widehat{P}_{\boldsymbol{\theta}_2}(\mathbf{R}))$$

Here, the Kullback-Leibler distances are computed assuming some Markov order as described by equation (9).

However, this direct approach does have problems. When the type for the reference distribution has a zero-valued probability estimate for some letter at which the other type is nonzero, we obtain an infinite answer, which may not be accurate (the true reference distribution has a nonzero probability for the offending letter). To alleviate this problem, the so-called *K-T estimate* [20] is employed. Each type is modified by adding one half to the histogram estimate *before* it is normalized to yield a type. Thus, for the k^{th} letter, the K-T estimate at bin b is

$$\widehat{P}^{\text{KT}}_{R_b}(r_k) = \frac{(\#\text{times } r_k \text{ occurs in } \{R_{1,b},\ldots,R_{M,b}\}) + \frac{1}{2}}{M + \frac{K}{2}}$$

Now, no letter will be assigned a zero estimate of its probability of occurrence *and* the estimate remains asymptotically unbiased with increasing number of observations. When applied to joint types, we add $1/2$ to each bin and normalize according to the total number of letters in the joint type. For first-order Markovian dependence analysis, we need the joint type defined over two successive bins, and we apply the K-T procedure according to

$$\widehat{P}^{\text{KT}}_{R_{b-1},R_b}(r_{k_1}, r_{k_2}) = \frac{(\#\text{ times } (r_{k_1}, r_{k_2}) \text{ occurs in } \{(R_{1,b-1}, R_{1,b}),\ldots,(R_{M,b-1}, R_{M,b})\}) + \frac{1}{2}}{M + \frac{K^2}{2}}$$

This estimation procedure is not arbitrary: It is based on theoretical considerations of what *a priori* distribution for the probabilities estimated by a type sways the estimate the least.

5.2 Bootstrap removal of bias

All distance measures presented here have the property that they can only attain non-negative values. Any quantity having this property cannot be estimated without bias. For example, if the true distributions are identical, distance measures are zero, but types calculated from two datasets drawn from the same distribution are unlikely to themselves be equal and the resulting distance estimate will be positively biased. While the estimates are asymptotically unbiased, experience shows that the bias is significant even for large datasets, and can lead to analysis difficulties. Analytic expressions for the bias of a related quantity—entropy—are known [6], and they indicate that bias expressions will depend on the underlying distribution in complicated ways.

Fortunately, recent work in statistics provides a way of estimating the bias and removing it from *any* estimator without requiring additional data. The essence of this procedure, known as the *bootstrap*, is to employ computation as a substitute for a larger dataset. The bootstrap procedure is one of several *resampling* techniques that attempt to provide auxiliary information—variance, bias, and confidence intervals—about a statistical estimate. Another method in this family is the so-called jackknife method, and it has been used for removal of bias in entropy calculations [10]. The book by Efron and Tibshirani [9] provides excellent descriptions of the bootstrap procedure and its theoretical properties.

In a general setting, let $\mathbf{R} = \{R_1,\ldots,R_M\}$ denote a dataset from which we estimate the quantity $\theta(\mathbf{R})$. Our quantities of interest here are the Kullback-Leibler and resistor-average distance measures. We create a sequence of bootstrap datasets $\mathbf{R}^*_j = \{R^*_{1,j}, \ldots, R^*_{M,j}\}$, $j = 1, \ldots, M_B$. Each bootstrap dataset has the same number of elements as the original, and is created by selecting elements from the original randomly and with replacement. Thus, elements in the original dataset may or may not appear in a given bootstrap dataset, and each can appear more than once. For example, suppose we had a dataset having four data elements $\{R_1, R_2, R_3, R_4\}$; a possible bootstrap dataset might be $\mathbf{R}^* = \{R_2, R_3, R_1, R_1\}$. The parameter estimated from the m^{th} bootstrap dataset is denoted by $\theta^*_m = \theta(\mathbf{R}^*_m)$. From the M_B bootstrap datasets, we estimate the quantity of interest, obtaining the collection of estimates $\{\theta^*_1, \ldots, \theta^*_{M_B}\}$. The suggested number of bootstrap datasets and estimates is a few hundred [9].

The bootstrap estimates cannot be used improve the precision of the original estimate, but they can provide estimates of $\theta(\mathbf{R})$'s auxiliary statistics, such as variance, bias, and confidence intervals. The *bootstrap estimate of bias* is found by averaging the bootstrap estimates, and subtracting from this average the original estimate: $\text{bias} = \frac{1}{M_B}\sum_m \theta^*_m - \theta(\mathbf{R})$. The bootstrap-debiased estimate is, therefore, $2\theta(\mathbf{R}) - \frac{1}{M_B}\sum_m \theta^*_m$. Calculation of bootstrap-debiased distances can result in negative distances when the actual distance is small.

Confidence intervals of level β can be estimated from the bootstrap estimates by sorting them, and determining which values correspond to the $\beta/2$ and $1-\beta/2$ quantiles. Let $\{\theta^*_{(1)}, \ldots, \theta^*_{(M_B)}\}$ denote the sorted (from smallest to largest) estimates. A raw confidence interval estimate corresponds to $[\theta^*_{(\lfloor M_B - \beta M_B/2 \rfloor)}, \theta^*_{(\lceil \beta M_B/2 \rceil)}]$. Thus, for the 90% confidence interval, $\beta = 0.9$, and the raw confidence interval corresponds to the 5^{th} and 95^{th} percentiles. Because we want confidence intervals on the bootstrap-debiased estimate rather than the original, we reverse the interval and center it around the debiased estimate: $[2\theta(\mathbf{R}) - \theta^*_{(\lceil \beta M_B/2 \rceil)}, 2\theta(\mathbf{R}) - \theta^*_{(\lfloor M_B - \beta M_B/2 \rfloor)}]$.

5.3 Dependence on binwidth

Ideally, the calculation of distance measures between two responses would not depend on the binwidth Δ used in the digitization process. However, discharge probability at any specific time varies as binwidth varies. Since distances measure how different two probability distributions are, we expect that distance calculations could depend on binwidth. To analyze this situation, let's assume a single neuron population, with the probability of an event equaling some rate times the binwidth: $\Pr[R_b = 1] = \lambda\Delta$ and $\Pr[R_b = 0] = 1 - \lambda\Delta$. The Kullback-Leibler distance between two such random variables (having rates λ_1 and λ_2) is given by

$$\mathcal{D}(\lambda_2 \| \lambda_1) = \lambda_2\Delta \log \frac{\lambda_2\Delta}{\lambda_1\Delta} + (1 - \lambda_2\Delta) \log \frac{1 - \lambda_2\Delta}{1 - \lambda_1\Delta} \ .$$

The first term is clearly proportional to binwidth; if we assume that the discharge probability is small ($\lambda\Delta \ll 1$), then the total expression is proportional to the binwidth.

$$\mathcal{D}(\lambda_2 \| \lambda_1) \approx \left(\lambda_2 \log \frac{\lambda_2}{\lambda_1} + \lambda_1 - \lambda_2 \right) \Delta$$

All the other distances are also proportional to binwidth when $\lambda\Delta \ll 1$. Once the binwidth is chosen small enough, we have found the temporal resolution necessary to maximally distinguish the two responses. Distance calculations can also be deliberately made with *larger* binwidths to assess the role temporal resolution has on distinguishing two responses.

When we accumulate the distance across bins that span a given time interval having duration T, as suggested in property 3 and equation (9), the number of bins equals T/Δ. If the discharge rates are such that discharge probabilities are small, the accumulation *over a given time interval* cancels the binwidth dependence, which leaves the accumulated distance independent of the binwidth. Let's be more concrete about this point. Assuming for the moment that the data are statistically independent from bin to bin (Markov order $D = 0$), the computation of the Kullback-Leibler distance between two responses equals

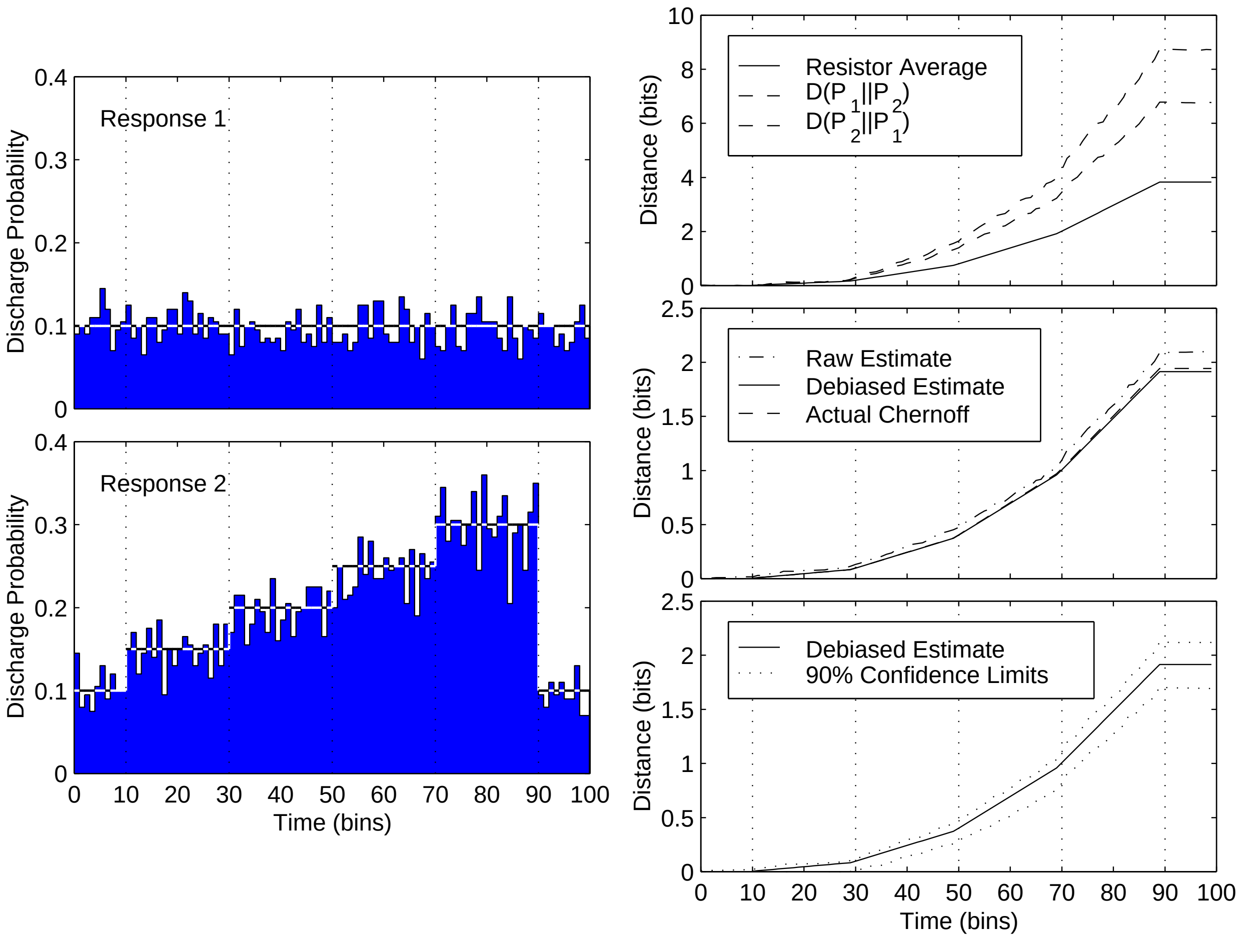

Figure 4: Single-neuron responses were simulated based on a Poisson discharge model. The first response had a constant rate, and the second response was a staircase; these constitute an example chosen to illustrate type-based analysis. These two responses equaled each other during the initial and final ten bins. The discharge probabilities controlled the occurrence of $M = 200$ simulated responses. The resulting PST histograms are shown, with the actual discharge probability shown by dashed lines and the dotted vertical lines indicating when rate changes occurred. The right column displays the various information-theoretic distance measures calculated from these responses. The top panel shows the accumulated Kullback-Leibler distances estimated with the K-T modification using each response as the reference (dashed lines), along with the resistor-average of these two shown (solid line). All of these were debiased using the bootstrap. In the middle panel, the resistor-average (scaled by two) before (dot-dashed) and after (solid) applying the bootstrap is compared with the theoretical Chernoff distance (dashed). The bottom panel again shows the debiased resistor-average (again scaled by two) along with the 90% confidence limits (dotted) estimated via the bootstrap. In all cases, two hundred bootstrap samples were used.

the sum of the distances between the responses occurring within a bin. This distance will be proportional to binwidth if the bins are small enough. However, when we add them up to form the total inter-response distance, the value we get will not depend on the binwidth. For this reason, we prefer plotting accumulated distance (as expressed in equations (2) and (9) in the independent and Markov cases respectively) across the response.

5.4 Example

Figure 4 illustrates a simple application of this procedure for simulated (Poisson) single-neuron discharges. The two ways of computing the Kullback-Leibler distance from the simulated responses differ substantially. We find that this difference is statistically significant, and occurs frequently in simulations and in actual recordings. Because we have no clear reference stimulus in this example, we use the Chernoff distance or its resistor-average approximation to compare two responses. The resistor average depicted in the top right panel consists of a series of straight lines, which correspond to time segments of constant rate differences between the responses. The greater slopes correspond to greater rate differences. Note that when the rates are equal, the distances do not change, indicating no response difference. The middle plot shows that the bias in the initial estimate of the resistor average is quite large. We have found the bootstrap bias compensation procedure described in section 5.2 to be necessary for obtaining accurate distance estimates. To employ bootstrap in the cyclostationary case, we consider our dataset to consist of the responses to individual stimulus presentations for a given parameter setting: $\{\mathbf{R}_1(\boldsymbol{\theta}), \ldots, \mathbf{R}_M(\boldsymbol{\theta})\}$, and our bootstrap datasets contain M responses selected randomly from this original. We independently perform the bootstrap on each response resulting from each stimulus condition, compute types from each bootstrap sample, and calculate the distance between these samples. As illustrated in figure 4, the bootstrap substantially removes the inherent positive bias. We also see that half the resistor-average distance quite closely approximates the actual Chernoff distance between the responses. Examining the bottom right panel of figure 4 shows that the actual Chernoff distance would lie well within the 90% confidence interval. Hence, we use the computationally simpler resistor-average distance measure. Note how the confidence interval widens as we progress across the response. This effect occurs naturally because we are adding more and more statistical quantities as we accumulate the total distance. These intervals would be substantially smaller if we considered accumulated distances over portions of the response.

To interpret this distance calculation, we refer to modern classification theory reviewed in the appendix. Because Chernoff distance is related through equation (A3) to the classification error rate, it reveals how easily the two responses can be distinguished: the bigger the distance, the smaller the probability of an error in distinguishing the two. Note that this error probability is known only up to a constant: We cannot compute it precisely. Asymptotic error probability changes with time roughly according to $2^{-d(t)}$, where $d(t)$ is the accumulated distance, be it the Kullback-Leibler or Chernoff distance, and t is post-stimulus time. Thus, each unit (one bit) increase in distance corresponds to a factor of two smaller error probability. The accumulation of distance with time is not an arbitrary choice. This procedure corresponds to the Kullback-Leibler distance's property 3, which states that the distance between the joint probability distributions characterizing a response over a given number of bins equals the sum of the component distances.

As the two responses are identical over the first ten bins, no distance is accumulated. As the rates differ more in each twenty-bin section, we see that the distance accumulated in each section increases. In this example, the accumulated (approximate) Chernoff distance increases from the beginning to the end of each section are 0.1, 0.3, 0.55, and 0.95 bits. These quantities were calculated by subtracting the accumulated distance at section beginning from its value at the end. Finally, the responses have identical rates during the last ten bins, and we see distance does not increase further. When we analyze responses, we concentrate on those portions of the response that contribute most to accumulated distance since they provide the most effective coding (in terms of classification errors). In our simple example, the response during bins 70–90 contributes most because the rate difference is greater there. As we consider more complicated examples of coding, it becomes increasingly important that we can use type-based analysis to determine important sections of the response *without* assuming the nature of the code.

To relate these distance calculations to estimation error, let's assume that the staircase response corresponds to increasing some stimulus parameter by equal amounts. The smallest increment yielded a difference of 0.1 bits over a 20-bin interval. Using the perturbational results of equation (5), we find that the Fisher information equals $0.1 \times 8 \ln 2/(\delta\theta)^2 = 0.555/(\delta\theta)^2$. This calculation means that this parameter is encoded by a rate code in such a way that the mean-squared estimation error incurred in determining the parameter from this response must be at least $(\delta\theta)^2/0.555 = 1.8(\delta\theta)^2$, where we need to know how the amount of perturbation to produce a numeric value. If N statistically independent neurons represented the stimulus the same way, the mean-squared error would decrease inversely with N.

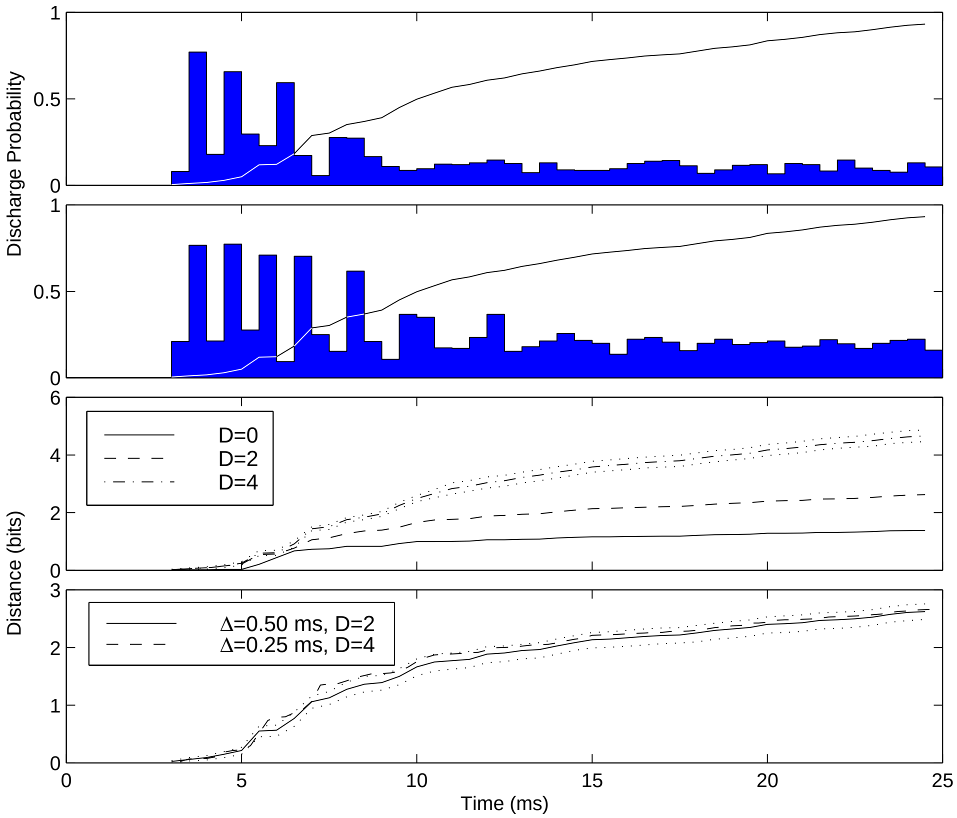

Figure 5: The upper panels show the PST histograms of a simulated lateral superior olive neuron's response to two choices of stimulus level (binwidth equals 0.5 ms). The simulations modeled the neuron's biophysics [33]. The bottom panels show the resistor-average distance between these two responses; the computations were performed under several conditions. The first of these shows the resistor-average distance (divided by two) between these responses computed for $D = 0, 2, 4$ bins (corresponding to 0, 1, and 2 ms of temporal dependence, respectively). The dotted lines straddling the $D = 4$ curve portray the 90% confidence interval. The curve superimposed upon the PST histograms is the $D = 4$ curve. Finally, the bottom plot displays the resistor-average distance (divided by two) between the responses for two choices of binwidth, but with the dependence parameter D chosen so that the assumed temporal dependence for each spans the same time interval. The 90% confidence interval for the $\Delta = 0.5$ ms is displayed with dotted lines.

Figure 5 portrays how choice of analysis order can affect distance calculations. Recall that exploring nonzero analysis orders amounts to seeking response differences *not* conveyed by the PST histogram. During the first few milliseconds, no significant response differences are evident. After about 5 ms, significant differences occur, with the various choices of analysis orders yielding about the same result. These distances then depart at about 7 ms, with the $D = 4$ curve being significantly larger. This result indicates significant temporal dependence in the responses as it differs greatly from the $D = 0$ curve, which always corresponds to assuming the data are statistically independent from bin to bin. The value of dependence parameter D is one of the few assumptions our information-theoretic approach must make. Ideally, all values that can be computed based on the amount of available data (equation 8) should be explored. As D increases, the distance calculations will eventually not change, and the best value for the dependence parameter is the smallest of these. In the example portrayed in figure 5, the resistor-average distance kept increasing, leaving us no choice but to use the largest possible value. What the actual distance might be, even whether it is larger or smaller than the $D = 4$ result, cannot be determined without more data. Using the $D = 4$ result, the distance between the responses increases most sharply during the second portion of the transient response.

If one sacrifices temporal resolution by using larger bins, the distance computation can span longer time intervals. The bottom panel of figure 5 shows two distance calculations that span the same amount of temporal dependence, one using twice the binwidth of the other and half the dependence order. The restrictions placed on the dependence parameter by (8) apply to the Markov order, not to the amount of time spanned by a discharge pattern's dependence structure. Thus, we could analyze the data with the same maximal dependence order allowed by the bound, but over longer dependence time intervals by manipulating the binwidth. It should be re-emphasized that the restrictions of (8) apply to *any* data analysis technique and the idea of choosing different binwidths applies to other methods as well.

This way of displaying distance—accumulated as post-stimulus time increases—also illustrates our general finding that the distance measures smooth the response variations found in PST histograms. Although the displayed responses came from simulations, actual recordings also demonstrate rapid rate oscillations found during the first 10 ms. One of our analysis technique's most powerful features is that it can assess response differences without regard to whether response rates and/or interspike dependencies are time varying or not. Note that during the latter portion of the response the distance measures increase roughly linearly. This effect usually indicates a difference in sustained rates, which can be discerned from the PST histograms. Furthermore, about half the total distance accumulated over 25 ms (4.65 bits) is garnered in the first 10 ms. We conclude that the initial transient of the response allows equal discriminability in the first 10 ms (actually 7 ms as there is about a 3 ms latency) as does the response obtained during the last 13 ms. Thus, the initial portion of the response conveys as much as the stimulus does during the latter portion in less time.

Binwidth effects are also demonstrated in figure 5. From the example shown there, we conclude that the larger binwidth of 0.5 ms would suffice as joint types computed over the same time span but with different binwidths yield nearly the same results. The time epochs over which the distance calculations disagree most occurs during the high-probability-of-discharge segments of both responses, a result consistent with the analysis of section 5.3.

6 Applications

We have shown how information-theoretic distances can assess how two responses differ in a meaningful way: Using them, we can infer the performance limits of information processing systems. We can also probe interdependencies in population responses. We describe this and other application of our approach for understanding the neural code.

6.1 Assessing neural codes

The simplest application of distance analysis is assessing which part of the response changes significantly as with stimulus changes. Perhaps the most powerful aspect of type-based analysis is that it makes no *a priori* assumption about the nature of neural encoding. It and other techniques that make no *a priori* assumptions about the neural code are limited to Markov dependence orders that (8) allows. Calculating response distance quantifies how well the code expresses stimulus changes regardless of its form, whether it be a timing code, a rate code, or some combination of these. "Significant change" has two meanings here. The first is whether the distance measure is significantly different from zero during some portion of the response. Inferring this statistical significance is the role for confidence intervals, which we compute using the bootstrap. The second type of significance is which portion of the response contributes most to accumulated distance. We judge this by computing how much distance changes over a given time interval. One consequence of making this kind of calculation is that we can *directly* evaluate one response component's importance relative to another's. For well-defined portions of the response, like the initial transient and later sustained response that typifies auditory neuron responses to tone bursts, we can directly compare how different portions are. Furthermore, the cumulative distance reveals how long it takes to yield a certain level of discrimination. We can then begin to answer questions such as how long it takes to determine from the population's response a just noticeable stimulus change.

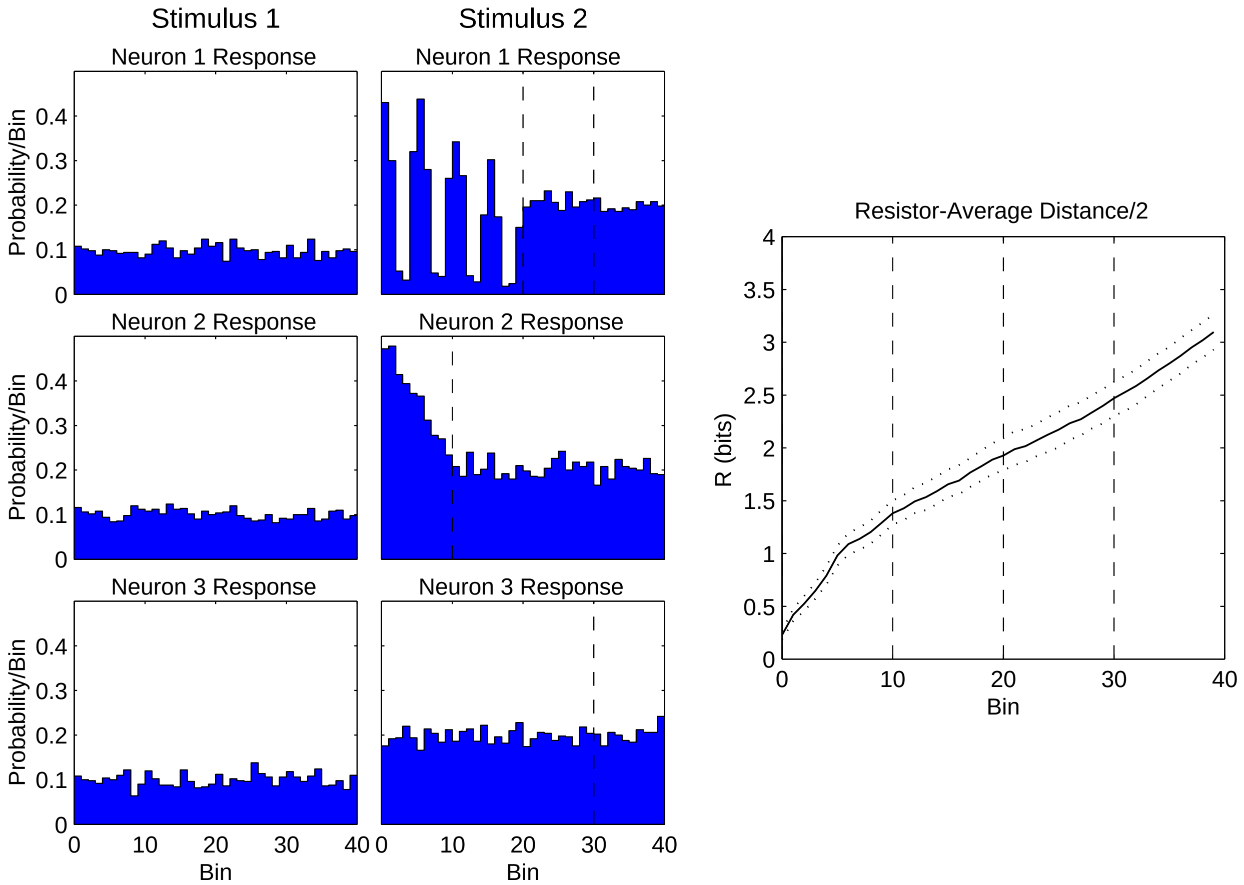

Figure 6: We simulated a three-neuron ensemble responding to two stimulus conditions. The left portion of the display shows PST histograms of each neuron. As far as can be discerned from these histograms, the first stimulus yielded a constant-rate response in each simulated neuron. The second stimulus produced different responses in each neuron: The first had an oscillatory response lasting 20 bins, the second a transient rate increase for 10 bins, and the third a rate change. The dashed vertical lines in the PST histograms indicate the boundaries of these various response portions. During the first stimulus, and until the last ten bins of the second stimulus, the neurons produced discharges statistically independent of the others. In the last ten bins, the first and third neurons' discharges became correlated (coefficient = 0.6). Throughout all responses, the responses were produced by a first-order Markov model having a correlation coefficient of -0.1. The right panel shows the result of computing the resistor-average distance between the two responses. The solid line shows half the resistor-average distance, with its 90% confidence interval shown with a dotted line. Dashed vertical lines correspond to stimulus 2 response components.

An example of this analysis for the single neuron case is displayed in figure 5. Figure 6 illustrates applying this approach to a simple population of three neurons. Both a stimulus-induced rate response and a transneural correlation can be detected, and the relative contribution of each response component to sensory discrimination quantified. Clearly, the initial portion of the response produced the greatest distance change. During the next ten bins, when the latter portion of neuron #1's oscillatory response and the rate responses of the other two are present, about 0.5 bits of distance were gained. This increase means that the probability of not being able to discriminate between the two stimulus conditions decreased by a factor of about $2^{0.5} = 1.4$. A much larger change (1.4 bits) occurred during the first ten bins. Consequently, the first portion of the response contributes much more to stimulus discrimination than the second. The third portion of the response contains only constant discharge rates. The distance accumulated during this time (bins 20–29) roughly equals the distance accumulated during the previous ten bins, when neuron #1's response contained an oscillatory component. This equality of accumulated distance means that the oscillatory response and

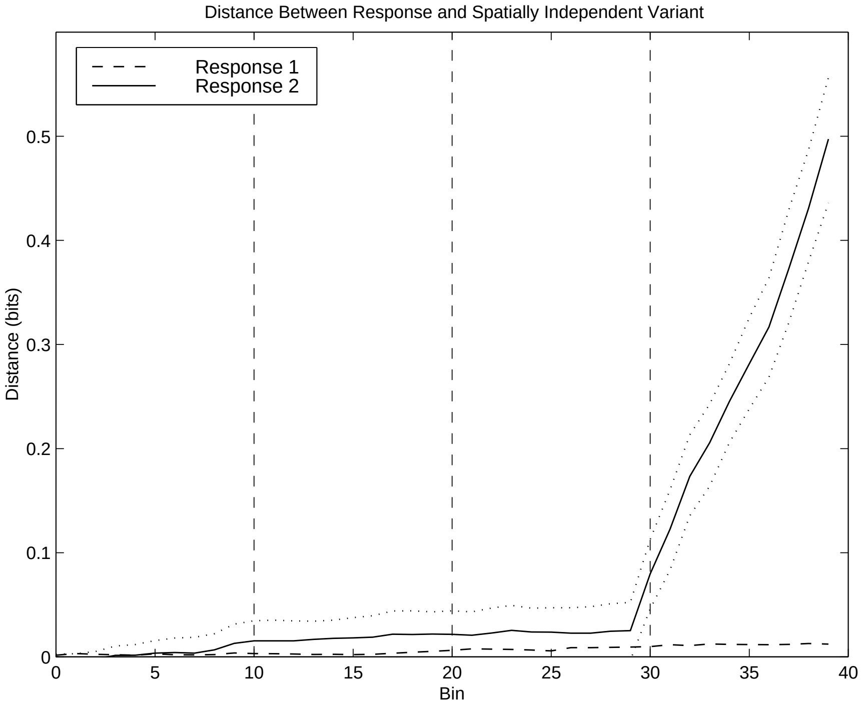

Figure 7: The resistor average (divided by two) between the type computed from response and the type computed derived from it that forces a spatially independent ensemble response structure is shown for the two stimulus conditions used in figure 6. The dashed line shows the result for the first response (histograms shown in the left column of figure 6), the solid line for the second (center column). As was simulated, the responses to stimulus 1 demonstrated no transneural correlation. The second stimulus did induce a correlation in the latter portion of the response, and the distance clearly indicates the presence of such correlation. The 90% confidence interval for the second response is indicated by the dotted lines. Note that the confidence interval's lower edge was less than zero for the first 30 bins.

the constant-rate response are equally effective in representing the stimulus difference. Interestingly, the introduction of spatial correlation (found in the last ten bins) increased only slightly the accumulated distance beyond what the rate response by itself would have.

6.2 Uncovering neural codes

The calculation of distances between responses quantifies neural coding without revealing what the code is. Distance calculations can offer some insights as well into what aspects of the response contribute to the code. For example, we can determine the presence of correlation in an ensemble's response, be it stimulus- or connectivity-induced. In the former case, spike trains can be correlated merely because neurons are responding to the same stimulus. In the latter, the neurons receive common inputs or are interconnected. We compute the type of the measured ensemble response and derive from it the type that would have been produced by the ensemble if it had statistically independent members (spatial dependence) and/or had no temporal dependence. Referring to figure 3 for an example, the probability of each neuron discharging in each bin can be calculated from the joint probability of various response patterns occurring in a bin. For example, $\Pr[\text{discharge in neuron \#1}] = \Pr[R_n = 4] + \Pr[R_n = 5] + \Pr[R_n = 6] + \Pr[R_n = 7]$ because the leading bit of the binary representations of these symbols, which corresponds to neuron #1, equals 1: $4 = 100_2$, $5 = 101_2$, etc. From these component probabilities, we estimate the probability of all possible ensemble response patterns by multiplying according to the ensemble response the probabilities of each neuron discharging or not ($\Pr[R_n = 3] = \Pr[\text{no discharge in neuron \#1}] \cdot \Pr[\text{discharge in neuron \#2}] \cdot \Pr[\text{discharge in neuron \#3}]$ because $3 = 011_2$). By calculating the distance between these two types, we can infer when correlated responses are present; figure 7 illustrates an example.

The presence of interneuron correlation in the fourth response segment shown in figure 6 is not discernable when compared to the distance accumulated in the third segment, when only rate differences are present. One might infer from the analysis shown in figure 7 that the amount accumulated in the fourth segment should exceed that of the third by about 0.5 bit. The fact that this difference does not occur when analyzing the data demonstrates a subtlety in using distance measures: The distance between responses does represent how easily an optimal classifier can distinguish them, but the various factors that contribute to this distance are *not* necessarily additive. Just because correlation analysis reveals 0.5 bit of difference does not mean that interneuron correlation increases the distance contributed by average-rate differences by the same amount.

In this analysis, we can use the Kullback-Leibler distance directly. It equals the mutual information between the component discharge patterns of the population (property 5). Zero mutual information corresponds to statistically independent responses and it increases as the discharge patterns become more interdependent. When applying this analysis to populations of three or more neurons, we are extending the definition of mutual information (property 5) to N random variables in a novel way.

$$I(x_1; \dots ; x_N) = \mathcal{D}(p(\mathbf{x}) \| p(x_1)p(x_2)\cdots p(x_N))$$

Here, $p(\mathbf{x})$ denotes the joint distribution of the N random variables. We note that from an information transfer viewpoint, statistically independent responses do not always correspond to the best situation [13].

6.3 Uncovering feature extraction

As part of developing these new techniques, we re-examined how the signal processing function of any system should be assessed. Consider a nonlinear, adaptive system—a neural ensemble—that accepts inputs and produces outputs (as shown in figure 1), about which we have only general insight into the system's function (for example, it processes acoustic information). Assume that the inputs depend on a collection of stimulus parameters represented by the vector $\boldsymbol{\theta}$. Curiously, knowing the system's input-output relation may not be helpful in understanding its signal processing function: Nonlinear systems are just too complicated. Our approach examines response sensitivity to stimulus changes and derives from it the ability of an optimal signal processing system to estimate the stimulus parameters. The key idea underlying this approach is the perturbational result of equation (5), which relates distance measure changes to the Fisher information matrix.

In our approach, we measure responses recorded in response to a reference stimulus parameterized by $\boldsymbol{\theta}_0$ and a family of responses parameterized by $\boldsymbol{\theta}_0 + \delta\boldsymbol{\theta}$, with $\delta\boldsymbol{\theta}$ a perturbation. We compute types from ensemble responses to both stimuli and quantify the "distance" between them. We use the Kullback-Leibler distance in this application since we have a natural choice for a reference response. Figure 8 shows the surfaces generated by perturbing two stimulus parameters—sound amplitude and azimuthal location of the sound—about a reference stimulus. Interestingly, our responses, obtained from accurate biophysical simulations of binaurally sensitive lateral superior olive (LSO) neurons [33], indicate that during different portions of the response, the two stimulus features are coded with differing fidelity. We measure fidelity as the ability (standard deviation of error) of an optimal system to estimate the stimulus parameters from the response. Early on, the transient response encodes both stimulus features well. Twenty milliseconds later, the fidelity of angle encoding remains about the same, although the form of the response has changed from a transient to a gradual rate change. During this period, the amplitude encoding has greatly worsened, with the standard deviation increasing by over a factor of five. During the constant-rate portion of the response starting 20 ms later, the amplitude estimate has worsened more with the angle estimate's quality remaining about the same.

What these results indicate is that this LSO neuron is processing its inputs (which greatly resemble the primary neural outputs of the two ears) in such a way that stimulus amplitude and angle are encoded in its response. In short, the neuron's discharge pattern multiplexes stimulus information. The fidelity of this representation changes rapidly with time after stimulus onset, with the azimuth being the primary stimulus attribute encoded in the response. Thus, the information coding provided by this neuron's discharges is multidimensional and time varying. The initial portion of the response could be used along with other neural responses in the auditory pathway to estimate stimulus amplitude, but later portions are less useful. Because azimuth is consistently represented by these neural discharges, we conclude the primary, but not only, role for the lateral superior olive, is sound localization. However, downstream neural structures could use the amplitude information conveyed by these responses. Parallel neural systems present in the auditory pathway clearly represent amplitude more effectively; presumably they have greater impact on amplitude processing.

7 Conclusions and discussion

The goal of information-theoretic distance analysis is to compute the distance between responses. We explored several ideas on how these distance calculations can be used to measure and assess the neural code. In all of these, the basic procedure is as follows.

1. Given sets of individual or simultaneous recordings, the analysis of the population's response begins with the digitization process described in section 3. The important consequence of this procedure is that single and multi-unit recordings have a common data representation.

2. Compute the joint type of user-specified order D, employing the K-T modification if the Kullback-Leibler distance is needed.
$$\widehat{P}_{R_b,\dots,R_{b-D}}(r_0,\dots,r_D) = \frac{(\#\text{times } R_b = r_0,\dots,R_{b-D}=r_D)+\frac{1}{2}}{M+2^{D\cdot N-1}}$$

3. Compute the Kullback-Leibler distance or resistor-average approximation to the Chernoff distance using the Markov decomposition expressed in equation (9). The conditional distribution needed in this computation is found from the joint type by a formula that mimics the definition of conditional probabilities.
$$\widehat{P}_{R_b|R_{b-1},\dots,b-D}(r_0|r_1,\dots r_D) = \frac{\widehat{P}_{R_b,\dots,R_{b-D}}(r_0,r_1,\dots,r_D)}{\sum_r \widehat{P}_{R_b,R_{b-1},\dots,R_{b-D}}(r,r_1\dots,r_D)}$$

4. Use the bootstrap debiasing and confidence interval procedure on the distance thus calculated. When analyzing cyclostationary responses, we consider the responses to individual stimulus presentations as the fundamental "data quantum." Our bootstrap samples are drawn from this collection of M datasets.

5. Our examples plot the cumulative debiased distance as each term is accumulated (using an expression similar to that of equation (9)).

We have presented information-theoretic distance measures that can be used to quantify neural coding, and described techniques that exploit them. These distances depend on the probabilistic descriptions of the neural discharges, about which we want to assume as little as possible. The theory of types suggests that empirical estimates of these distributions can be used to accurately compute these distance measures, with the sole modeling assumption being the Markov dependence parameter. Given sufficient data, this parameter can also be determined solely by the data's statistical structure. If this statistical structure spans a long time interval, our technique and any other will not fully reveal the neural code unless temporal resolution is compromised or more data are acquired. The examples we have presented here, particularly the feature extraction one, demonstrate that neural information coding can be quite complex, being both time

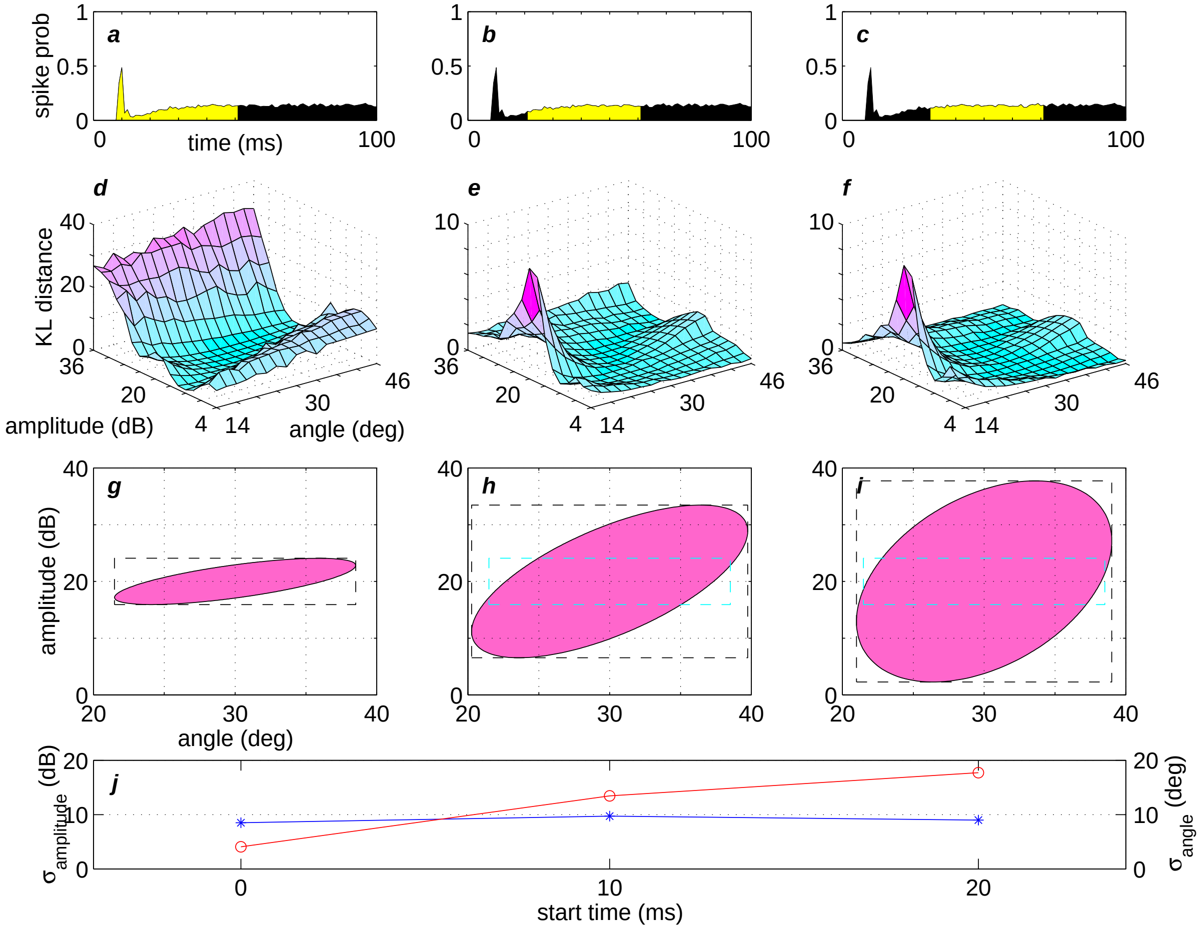

Figure 8: We simulated [33] the response of a single lateral superior olive neuron [25, 26] to high-frequency tone bursts presented at various amplitudes and azimuthal locations. The top row (panels a-c) shows the PST histogram of the simulated response at the reference condition (20 dB, 30°). The light areas in each indicate the 40 ms portion of the response subjected to type-based analysis. The next row (panels d-f) shows three-dimensional surfaces of the corresponding values of Kullback-Leibler distance between the reference response and the responses resulting from varying stimulus amplitude and angle. From these surfaces, we fit a two-dimensional third-order polynomial, and used its parabolic terms to estimate the elements of the Fisher information matrix according to equation (5). The inverse of this matrix provides lower bounds on estimates of angle and amplitude derived from the analyzed portion of the response. Panels g-i show sensitivity ellipses that trace one standard deviation that an optimal system would yield if it estimated amplitude and angle simultaneously. The horizontal and vertical extents of these ellipses correspond to the standard deviations of angle and amplitude estimates, respectively, and these determined the rectangles shown in each panel. Panel g's rectangle is repeated in the other panels for comparison. The bottom panel shows how these standard deviations changed during the response. The circles indicate the standard deviation of the amplitude estimate (left vertical scale) and the asterisks the standard deviation of the angle estimate (right scale).

varying and expressed by both discharge timing and rate. Thus, any technique for assessing information-coding fidelity must make as few assumptions as possible; the type-based analysis described here fulfills that criterion.

A second powerful aspect of our approach is its ability to cope with ensemble responses. As shown in figure 3, the analysis technique can conceptually be applied to any sized population. The information bound (equation 8) suggests that the amount of data required for a given level of analysis grows *exponentially* in the number of neurons and in the Markov parameter D. In practical terms, our technique can be used only for small populations. However, since the bound applies to any bin-based technique, without additional assumptions about the neural code, all such techniques are similarly data-limited. Whether a similar limitation applies to other techniques, such as those based on interspike intervals, is not known.

For judging coding quality, we prefer the Chernoff distance. Because of its computational complexity, we use the resistor-average distance to approximate it. The Kullback-Leibler distance, despite its theoretical importance, is difficult to use empirically because it is asymmetric with respect to the two component responses. We used it in the stimulus perturbation analysis because a natural reference response emerges, and it is the simplest computationally to estimate.

The procedures we have described here can assess neural coding, but they do not directly reveal what the code is. We showed one approach to assessing transneural correlation in figure 7. In general, coding mechanisms can be inferred from the component types; precisely how we have not yet determined. Be that as it may, the information-theoretic procedures developed here offer flexible but computationally intense analysis techniques that can meaningfully quantify the nature of neural coding within populations.

Appendix: Classification theory

Classification theory concerns how observations can be optimally classified into predefined categories. Stating the problem formally in the paper's context of neural signal analysis, a set of measured responses $\{\mathbf{R}_1, \dots, \mathbf{R}_M\}$ is to be classified as belonging to one of J categories. Here, the parameter M represents the number of stimulus presentations, and each $\mathbf{R}_m$ represents the population response to each presentation. Each category is defined according to a known probabilistic description (which may have unknown parameters or other controlled uncertainties). The most frequently studied variant of the classification problem is the binary classification problem: which of two categories C_1, C_2 best match the observations. Given the probabilistic descriptions of the categories, the optimal rule for classifying observations, the likelihood ratio test, is well known [14]. Interestingly, what is very difficult to calculate is how well this optimal rule works. Developing approximations for calculating performance leads to important notions that directly apply to the neural response analysis problem. Classifier performance is usually expressed in terms of error probabilities. Terminology for the error probabilities originated in radar engineering, where C_1 meant that no target was present and C_2 that one was. The *false-alarm probability* $P_{\text{fa}} = \Pr[\text{say } C_2 \mid C_1]$ is the probability that the classification rule announces C_2 when the data actually were produced according to C_1 (the classifier incorrectly declared a target was present), and the *miss probability* is $P_{\text{miss}} = \Pr[\text{say } C_1 \mid C_2]$ (the classifier incorrectly announced that no target was present when one was). The *average error probability* P_{e} is the average of these individual error probabilities: $P_{\text{e}} = \pi_1 P_{\text{fa}} + \pi_2 P_{\text{miss}}$, where π_1, π_2 are the *a priori* probabilities that data conform to the categories. Note that in the two-category problem $\pi_1 = 1 - \pi_2$.

The likelihood ratio test results when we seek the classifier that minimizes either the average probability of error or the false-alarm probability [15].[6] Picking which error probability to minimize would seem not to matter (the same classifier results) if it were not for the fact that the optimized error probabilities usually differ, easily by orders of magnitude in many cases. Furthermore, in many situations the optimal classifier's false-alarm probability can differ significantly from its miss probability. Because error probabilities provide the portal through which we develop information-theoretic analysis techniques, appreciating which error probability best suits a given experimental paradigm leads to better interpretation of measured distances. In neuroscience applications, we want to present two different stimulus conditions and to quantify how easily these stimuli can be distinguished based on the responses of some neural population. Average error probability summarizes how well an optimal classifier can distinguish data that could arise from either of two stimulus conditions. False-alarm (or miss) probability better summarizes performance when one of our stimuli can be considered a reference, and we want to know how well an optimal classifier can distinguish some neural response from a nominal.

No general formulae for any error probabilities are known for any optimal classifiers except in special cases, such as the classic Gaussian problem. We can answer the question "How does performance change as the amount of data becomes large?" When the observations are statistically independent and identically distributed under both categories, $p_{C_j}(\{\mathbf{R}_1, \dots, \mathbf{R}_M\}) = \prod_m p_{C_j}(\mathbf{R}_m)$, results generically known as Stein's Lemma [7: §12.8,12.9] state that error probabilities decay exponentially in the amount of data available,

[6] Note that in optimizing false-alarm probability (making it as small as possible), we must constrain the miss-probability to not exceed a pre-specified value. If not, we can make the false-alarm probability zero by having our classifier announce "class C_2 models the data." That way we are never wrong when the category C_2 is true, but we'll always be wrong if category C_1 is true (the miss probability will be one). The likelihood ratio test emerges when we minimize P_{fa} subject to $P_{\text{miss}} \leq \alpha$.

with the so-called *exponential rate* being an information-theoretic distance measure.

$$\lim_{M\to\infty} \frac{\log P_{\text{fa}}}{M} = -\mathcal{D}\left(p_{C_2}(\mathbf{R}) \| p_{C_1}(\mathbf{R})\right) \quad \text{for fixed } P_{\text{miss}} \tag{A1a}$$

$$\lim_{M\to\infty} \frac{\log P_{\text{miss}}}{M} = -\mathcal{D}\left(p_{C_1}(\mathbf{R}) \| p_{C_2}(\mathbf{R})\right) \quad \text{for fixed } P_{\text{fa}} \tag{A1b}$$

$$\lim_{M\to\infty} \frac{\log P_{\text{e}}}{M} = -\mathcal{C}\left(p_{C_1}(\mathbf{R}), p_{C_2}(\mathbf{R})\right) \tag{A1c}$$

$\mathcal{D}(\cdot\|\cdot)$ is known as the *Kullback-Leibler distance* between two probability distributions. It applies to both probability densities p, q or probability mass functions P, Q.

$$\mathcal{D}(p\|q) = \int p(R) \log \frac{p(R)}{q(R)}\, dR \tag{A2}$$

$\mathcal{C}(\cdot,\cdot)$ is the *Chernoff distance*, defined as

$$\mathcal{C}(p,q) = -\min_{0\le u\le 1} \log \int [p(R)]^{1-u}[q(R)]^u\, dR \tag{A3}$$

Note these definitions apply to both univariate and multivariate distributions. When the observations are not statistically independent, all these results apply to the multivariate distribution of the observations [16]. For example,

$$\lim_{M\to\infty} \frac{\log P_{\text{fa}}}{M} = -\lim_{M\to\infty} \frac{\mathcal{D}\left(p_{C_2}(\{\mathbf{R}_1,\dots,\mathbf{R}_M\}) \| p_{C_1}(\{\mathbf{R}_1,\dots,\mathbf{R}_M\})\right)}{M} \quad \text{for fixed } P_{\text{miss}}\ .$$

Stein's Lemma (A1) is not stated directly in term of error probabilities because of subtle, but important, technical details. Focusing on the false-alarm probability, Stein's Lemma for the case of independent observations can be stated more directly as

$$P_{\text{fa}} \to f(M) 2^{-M\mathcal{D}(p_{C_1}(\mathbf{R})\|p_{C_2}(\mathbf{R}))} \quad \text{for fixed } P_{\text{miss}}\ ,$$

with $\lim_{M\to\infty}[\log f(M)]/M = 0$. The term $f(\cdot)$ changes more slowly in comparison to the exponential, and it depends on the problem at hand. What this formula means is that if we plot any of the error probabilities logarithmically against M linearly, we will always obtain a straight line for large values of M (see figure 9). Stein's Lemma says that the false-alarm probability's slope on semi-logarithmic axes—its exponential rate—equals the negative of the Kullback-Leibler distance between the probability distributions defining our classification problem. Because of the presence of the problem-dependent quantity $f(\cdot)$, we cannot determine in general the vertical origin for the error probability or how large M must be for straight-line behavior to take over. Consequently, we cannot use asymptotic formulas to compute error probabilities, but we do know that error probabilities ultimately decay exponentially for *any* classification problem solved with the optimal classifier, and we know the rate of this decay. We can also say that if further observations increase any of these distances by one unit (a bit), the corresponding error probability decreases by a factor of two. Furthermore, having exponentially decreasing error probabilities defines a set of "good" classifiers. Optimal classifiers produce error probabilities that decay exponentially with the quantity multiplying M equal to the Kullback-Leibler distance or the Chernoff distance. Suboptimal but "good" ones will have a smaller slope, with poor ones not yielding exponentially decaying error probabilities. The exponential rate cannot be steeper than the Chernoff distance for the classifier that optimizes average error probability and Kullback-Leibler distance for the Neyman-Pearson classifier. Thus, these distances define *any* classification problem's difficulty. The greater the distance, the more quickly error probabilities decrease (the exponential rate is larger) and the "easier" the classification problem. *Whether we use an optimal classifier or not, the*

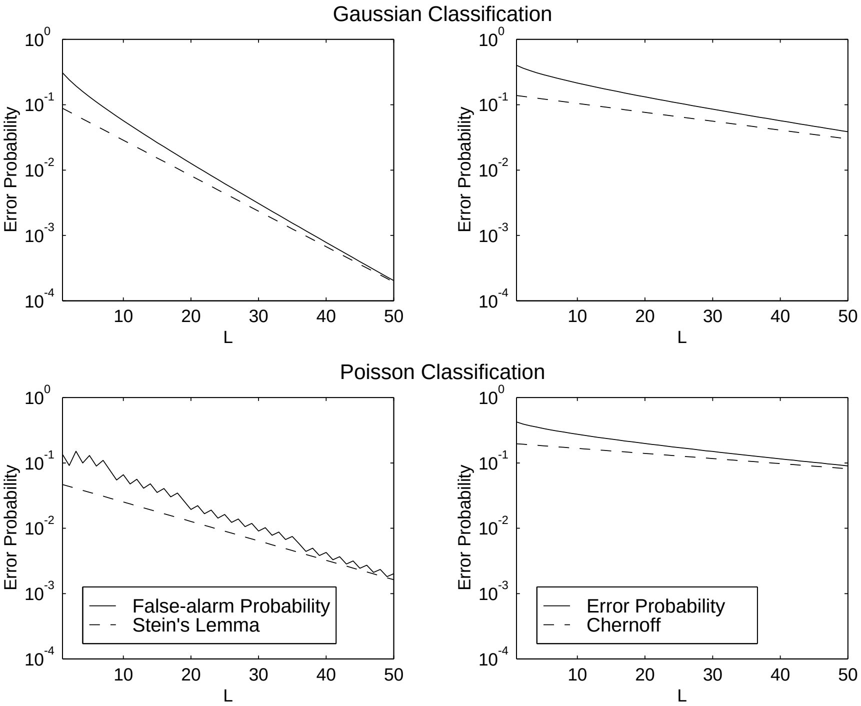

Figure 9: Using the Gaussian and Poisson classification problems as examples, we plot the false-alarm probability (left panels) and average probability of error (right panels) for each as a function of the amount of statistically independent data used by the optimal classifier. The miss-probability criterion was that it be less than or equal to 0.1. The *a priori* probabilities are $1/2$ in the right-column examples. As shown here, the average error probability produced by the minimum P_e classifier typically decays more slowly than the false-alarm probability for the classifier that fixes the miss probability. The dashed lines depict the behavior of the error probabilities as predicted by asymptotic theory (A1). In each case, these theoretical lines have been shifted vertically for ease of comparison.

Chernoff and Kullback-Leibler distances quantify the ultimate performance any classifier can achieve, and therefore measure intrinsic classification problem difficulty. We therefore want to estimate these distances from measured responses to quantify what response differences make distinguishing them easier.

Note that the Kullback-Leibler distance (equation A2) is asymmetric with respect to the two distributions defining the classification problem. The false-alarm probability achieved under a fixed miss-probability constraint and the miss probability achieved under a fixed false-alarm-probability constraint not only have different values, they may have different exponential rates. The Chernoff distance (equation A3) is symmetric with respect to the two probability distributions, and should be used to assess the classification problem. What prevents using it in applications is the required minimization process. Technically, this minimization is simple: The function to be minimized is bowl-shaped, leaving it with a unique minimum. Computationally, finding the quantity to be minimized can be quite complicated. In typical problems, many more calculations are needed to find the Chernoff distance than required to compute the Kullback-Leibler distance.

References

1. M. Abeles. *Corticonics: Neural Circuits of the Cerebral Cortex*. Cambridge University Press, New York, 1991.
2. M. Abeles and M. H. Goldstein, Jr. Multispike train analysis. *Proc. IEEE*, 65:762–773, 1977.
3. J.-M. Alonso, W. M. Usrey, and R. C. Reid. Precisely correlated firing in cells of the lateral geniculate nucleus. *Nature*, 383:815–818, 1996.
4. M. Basseville. Distance measures for signal processing and pattern recognition. *Signal Processing*, 18:349–369, 1989.

5. W. Bialek, F. Rieke, R. R. de Ruyter van Steveninck, and D. Warland. Reading a neural code. *Science*, 252:1852–1856, 1991.
6. A. G. Carlton. On the bias of information estimates. *Psychological Bulletin*, 71:108–109, 1969.
7. T. M. Cover and J .A. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.
8. R.C. deCharms and M.M. Merzenich. Primary cortical representation of sounds by the coordination of action-potential timing. *Nature*, 381:610–613, 1996.
9. B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993.
10. R. M. Fagan. Information measures: Statistical confidence limits and inference. *J. Th. Biol.*, 73:61–79, 1978.
11. F. Gabbiani and C. Koch. Coding of time-varying signals in spike trains of integrate-and-fire neurons with random threshold. *Neural Computation*, 8:44–66, 1996.
12. W. A. Gardner. An introduction to cyclostationary signals. In *Cyclostationarity in Communications and Signal Processing*, chapter 1. IEEE Press, New York, 1994.
13. C. M. Gruner and D. H. Johnson. Correlation and neural information coding efficiency. In *Computational Neuroscience*, Santa Barbara, CA, 1998.
14. R. V. Hogg and A. T. Craig. *Introduction to Mathematical Statistics*. Prentice Hall, Englewood Cliffs, NJ, fifth edition, 1995.
15. D. H. Johnson and D. E. Dudgeon. *Array Signal Processing: Concepts and Techniques*. Prentice-Hall, 1993.
16. D. H. Johnson and G. C. Orsak. Performance of optimal non-Gaussian detectors. *IEEE Trans. Comm.*, 41:1319–1328, 1993.
17. D. H. Johnson and A. Swami. The transmission of signals by auditory-nerve fiber discharge patterns. *J. Acoust. Soc. Am.*, 74:493–501, 1983.
18. D. H. Johnson, C. Tsuchitani, D. A. Linebarger, and M. Johnson. The application of a point process model to the single unit responses of the cat lateral superior olive to ipsilaterally presented tones. *Hearing Res.*, 21:135–159, 1986.
19. D.H. Johnson. Point process models of single-neuron discharges. *J. Computational Neuroscience*, 3:275–299, 1996.
20. R. E. Krichevsky and V. K. Trofimov. The performance of universal encoding. *IEEE Trans. Info. Th.*, IT–27:199–207, 1981.
21. J. C. Middlebrooks, A. E. Clock, L. Xu, and D. M. Green. A panoramic code for sound location by cortical neurons. *Science*, 264:842–844, 6 May 1994.
22. A. Riehle, S. Grun, M. Diesmann, and A. Aertsen. Spike synchronization and rate modulation differentially involved in motor cortical function. *Science*, 278:1950–1953, 1997.
23. F. Rieke, D.A. Bodnar, and W. Bialek. Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory efferents. *Proc. R. Soc. Lond. B*, 262:259–265, 1995.
24. W. Singer and C. M. Gray. Visual feature integration and the temporal correlation hypothesis. *Ann. Rev. Neuroscience*, 18:555–586, 1995.
25. C. Tsuchitani. The inhibition of cat lateral superior olivary unit excitatory responses to binaural tone bursts: I. The transient chopper discharges. *J. Neurophysiol.*, 59:164–183, 1988.
26. C. Tsuchitani. The inhibition of cat lateral superior olivary unit excitatory responses to binaural tone bursts: II. The sustained discharges. *J. Neurophysiol.*, 59:184–211, 1988.
27. E. Vaadla, I. Haalman, M. Abeles, H. Bergman, Y. PRut, H. Solvin, and A.Aertsen. Dynamics of neuronal interactions in monkey cortex in relation to behavioural events. *Nature*, 373:515–518, 1995.
28. J. D. Victor and K. P. Purpura. Metric-space analysis of spike trains: theory, algorithms and applications. *Network: Comput. Neural Sys.*, 8:127–164, 1997.
29. M. Wehr and G. Laurent. Odour encoding by temporal sequences of firing in oscillating neural assemblies. *Nature*, 384:162–166, 1996.

30. M. J. Wienberger, J. J. Rissanen, and M. Feder. A universal finite memory source. *IEEE Trans. Info. Th.*, 41:643–652, 1995.
31. M. Zacksenhouse, D. H. Johnson, and C. Tsuchitani. Excitatory/inhibitory interaction in the LSO revealed by point process modeling. *Hearing Res.*, 62:105–123, 1992.
32. M. Zacksenhouse, D. H. Johnson, and C. Tsuchitani. Excitation effects on LSO unit sustained responses: Point process characterization. *Hearing Res.*, 68:202–216, 1993.
33. M. Zacksenhouse, D. H. Johnson, J. Williams, and C. Tsuchitani. Single-neuron modeling of LSO unit responses. *J. Neurophysiol.*, 79:3098–3110, 1998.