INFORMATION-THEORETIC ANALYSIS OF NEURAL CODING

Don H. Johnson and Charlotte M. Gruner

Computer and Information Technology Institute
Department of Electrical and Computer Engineering
Rice University
Houston, Texas 77005–1892
dhj@rice.edu, cmgruner@rice.edu

ABSTRACT

We describe a family of new techniques for analyzing single- and multi-unit discharge patterns. These techniques are based on information theoretic distance measures and on empirical theories derived from work on universal signal processing. They are capable of determining transneuron statistical dependencies even when time-varying responses occur. The response portion contributing most to information coding can be identified and the coding fidelity can be quantified regardless of the neural coding mechanisms—be it timing, rate or transneural correlations.

1. INTRODUCTION

For about half a century, the information-bearing aspect of individual neuron's discharge patterns has been thought to be the *times* at which discharges occur. If a neural population encodes information about a sensory stimulus, then discharge timing in each component neuron should somehow vary with stimulus changes. In neural systems, both simple stimulus-evoked variations, such as an average discharge rate change, and complicated stimulus-evoked variations, such as a change in multi-neuron discharge correlation structure are observed. Data analysis techniques for single-neuron discharges such as the PST histogram, the interval histograms, and several joint interval statistical measures—were inspired by the mathematical model for single neuron discharges, the point process [11]. These measures do not quantify the response to reveal what stimulus aspects are being represented, when these representations occur, what the representations are, and the quality of these representations. The limitation to single unit activity also means that population coding is not directly probed. Consequently, more recent work has focused on population activity, using the fundamental assumption that coordinated sequences of action potentials produced by groups of neurons collectively represent their response. Neural ensembles process their inputs to produce *joint* discharge patterns that encode those aspects of the stimulus enhanced by the ensemble. Thus, today the "neural code" is taken to mean how groups of neurons, responding individually and collectively, represent sensory information with their discharge patterns [3]. Knowing the code would unlock the secrets of how neurons, working in concert, process and represent information. From the viewpoint of point process theory, the code is equivalent to the intensity of an accurate vector-channel, point-process model [14] for the data. Unfortunately, traditional optimal estimation techniques depend heavily on the intensity's intrinsic structure (how one event depends on the timing of others) [3], which is part of the neural code we seek. Furthermore, stimulus changes induce time-varying responses, which confound many techniques for quantifying the population codes: Mutual information calculations [7, 15], cross- and autocorrelation techniques [2], and artificial neural networks [13] apply to stationary single-neuron response patterns and don't generalize easily to neural ensembles.

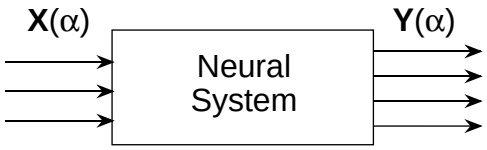

Figure 1: A neural system has as inputs the vector quantity $\mathbf{X}$ that depend on a collection of stimulus parameters denoted by the vector $\boldsymbol{\alpha}$. The output $\mathbf{Y}$ thus also depends on the stimulus parameters. Both input and output implicitly depend on time. Note that the 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 sensory information (parameterized by $\boldsymbol{\alpha}$) and produces outputs $\mathbf{Y}$ that codes some or all of the stimulus. The boldfaced symbols represent vectors, and are intended to convey the notion that our system is a neural ensemble and has multiple inputs and multiple outputs. Presumably, stimulus features preserved in the output are those extracted by the system; those deemphasized in the output are discarded by the system. To probe the system and its representation of sensory information, we experimentally measure the system's output and its inputs as we vary stimulus parameters. No change in the response 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 *distance measure*: Given two sets of stimulus conditions $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2$, we need to measure how different the responses are—how far apart they are—with a distance $d(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$. This metric needs to apply ensemble responses, to nonstationary as well as stationary response changes, to changes in transneural correlations, and to changes in discharge statistics.

While the merits of one measure versus another can be debated, we describe here how to use information theoretic distances that have a clear mathematical and intuitive foundation. Roots of the underlying theory are not in the classic results of Shannon, but in modern classification theory. In this theory, we try to assign a re-

sponse to one of a set of preassigned response categories. 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—the theory of how to process information universally without much regard to the underlying distribution of the data—provide a technique of measuring distance and demonstrate the technique's data-processing optimality.

2. CLASSIFICATION THEORY

Classification theory concerns how one can optimally classify empirical observations into predefined categories. Stating the problem formally, a set of observations $\mathbf{R} = \{R_1, \ldots, R_L\}$ is to be classified as belonging to one of J categories. The most frequently studied variant of this problem is the binary classification problem: which of two categories C_1, C_2 best match the observations. No general formulae for the miss and false-alarm error probabilities (P_M and P_F respectively) are known. What has been found are asymptotic expressions. When the observations $\mathbf{R}$ are statistically independent and identically distributed under both categories, a result known as Stein's Lemma [5: §12.8] states

$$\lim_{L\to\infty} \frac{\log P_F}{L} = -\mathcal{D}(p_{C_2}(R)\|p_{C_1}(R)) \quad \text{for fixed } P_M \qquad (1)$$

where $\mathcal{D}(p\|q)$ is known as the *Kullback-Leibler distance* between the probability densities p, q.

$$\mathcal{D}(p\|q) = \int p(x) \log \frac{p(x)}{q(x)}\, dx$$

Note that the definition of Kullback-Leibler distance applies to both univariate and multivariate distributions.[1] What Stein's Lemma means is that error probabilities decay exponentially in the amount of data, with a slope equal to the Kullback-Leibler distance between the probability distributions defining the classification problem. This slope, known the *exponential rate*, cannot be steeper than the Kullback-Leibler distance for any classifier. Whether we use an optimal classifier or not, the Kullback-Leibler distance quantifies the ultimate performance any classifier can achieve, and therefore measures *any* classification problem's intrinsic difficulty. The word "distance" should appear in quotes because $\mathcal{D}(\cdot\|\cdot)$ violates one of the fundamental properties a distance metric must have: It is not a symmetric function of its arguments. Be that as it may, geometric theories of the classification problem show that *no* distance metric exists for it, and that the Kullback-Leibler distance is the distance-like quantity that should be used to assess how different two categories are [6].

The Kullback-Leibler distance is also related to the ease of estimating parameters that define the classic classification problem. In the situation where two categories differ slightly according to the values of a parameter vector $\boldsymbol{\alpha}$—symbolically, $p_{C_1} = p(\boldsymbol{\alpha})$ and $p_{C_2} = p(\boldsymbol{\alpha} + \boldsymbol{\delta\alpha})$—for sufficiently small values of the difference $\boldsymbol{\delta\alpha}$, the Kullback-Leibler distance has the form

$$\begin{aligned} \mathcal{D}(p(\boldsymbol{\alpha} + \boldsymbol{\delta\alpha})\|p(\boldsymbol{\alpha})) &\approx \frac{1}{2}\overline{\boldsymbol{\delta\alpha}}'\mathbf{F}(\boldsymbol{\alpha})\overline{\boldsymbol{\delta\alpha}} \\ \mathbf{F}(\boldsymbol{\alpha}) &= \mathcal{E}\left[(\nabla_{\boldsymbol{\alpha}} \log p(\boldsymbol{\alpha}))\,(\nabla_{\boldsymbol{\alpha}} \log p(\boldsymbol{\alpha}))'\right] \end{aligned} \qquad (2)$$

[1] In these definitions, we use the base-two logarithm, which means that distance has units of bits.

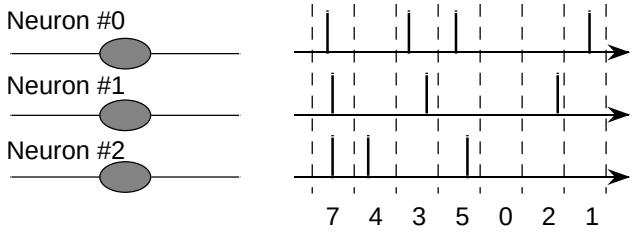

Figure 2: Individual neurons are given an arbitrary identification number. The discharge pattern for each is measured, and individual discharges placed in the b^{th} bin (each bin has width Δ). Using the neuron identification number and the presence of discharges in a binary code, a number denoted by R_b is assigned to each bin to represent which neurons fired during that bin. Because the intensity corresponding to each neuron typically varies with time, we estimate types for each bin separately using multiple stimulus presentations.

Here, $\mathbf{F}(\boldsymbol{\alpha})$ denotes the Fisher information, $(\cdot)'$ means transpose, $\nabla_{\boldsymbol{\alpha}} \log p(\boldsymbol{\alpha})$ means the gradient of the log probability density function, and $\mathcal{E}[\cdot]$ denotes expected value. The significance of these formulas rests in the Cramér-Rao bound, which states that the mean-squared error covariance matrix $\boldsymbol{\Sigma}_\epsilon$ for *any* unbiased estimator $\widehat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}$ cannot be "smaller" than $\mathbf{F}^{-1}(\boldsymbol{\alpha})$ in the sense that $\boldsymbol{\Sigma}_\epsilon - \mathbf{F}^{-1}(\boldsymbol{\alpha})$ is non-negative definite. In particular, this result means that the mean-squared error of an individual parameter is lowered bounded by the appropriate diagonal entry of the inverse of the Fisher information matrix. Thus for any given stimulus parameter perturbation $\overline{\delta\alpha}$, the larger the Kullback-Leibler distance, the larger the Fisher information, and hence the smaller the smallest possible mean-squared error. This relationship reinforces the notion that the Kullback-Leibler distance does indeed measure how distinct two classification categories are.

Instead of having a probabilistic description of the categories as in the classic classification problem, in the empirical classification problem we have only data. Gutman [8] found a classifier that is not only optimal (yielding maximal exponential rate), but will also, given enough training and observational data, produce error probabilities having the *same* exponential rate as the likelihood ratio classifier that clairvoyantly knows the underlying statistical model for the training data. His approach requires the computation of *types*—the histogram estimate of the probability mass function [5: Chap. 12]. We have demonstrated how type-based classifiers can be used in communication problems [9]. Our approach is to estimate the Kullback-Leibler distance between types computed from the neural recordings. From these distance calculations, we can directly infer when and how well sensory information is represented in neural responses.

3. QUANTIFYING 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 2). 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. In developing techniques to analyze neural coding, we need only consider the statistical structure of this sequence. Let $\mathbf{R}^{(1)}$ and $\mathbf{R}^{(2)}$ represent the responses of a neural population to

two stimulus conditions. What we want to measure is the distance between the *joint* probability distributions corresponding to these responses. Using the Kullback-Leibler as an example, we would want to find $\mathcal{D}\left(p(\mathbf{R}^{(2)})\|p(\mathbf{R}^{(1)})\right)$.

The most direct approach to estimating distance measures is to use types in their definitions. This approach has two difficulties bias and poorly formed probability estimates. When the type for the reference distribution has a zero-valued probability estimate for some letter at which the other type is nonzero, we would 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 [12] 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 is

$$\widehat{P}_{\mathbf{R}}^{\text{KT}}(a_k) = \frac{(\#\text{times } a_k \text{ occurs in } \mathbf{R}) + \frac{1}{2}}{L_R + \frac{K}{2}}$$

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. 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.

Because the Kullback-Leibler distance is non-negative, it cannot be estimated without bias. While the estimates are asymptotically unbiased, in our experience 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 [4], 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. In a general setting, let $\mathbf{R} = \{R_1, \ldots, R_L\}$ denote a dataset from which we estimate the quantity $\theta(\mathbf{R})$. We create a sequence of bootstrap datasets $\mathbf{R}_l^* = \{R_{1,l}^*, \ldots, R_{L,l}^*\}$, $l = 1, \ldots, L_B$. From each dataset, we estimate the parameter $\hat{\theta}_l(R_{1,l}^*)$. 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.

4. RESULTS

The simplest application of distance analysis is assessing which part of the response changes significantly as with stimulus changes. By calculating information theoretic distances from types, we measure response differences *no matter how they arise*. Figure 3 illustrates the application of 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. Because Kullback-Leibler distance is related through Stein's Lemma to 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. 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.

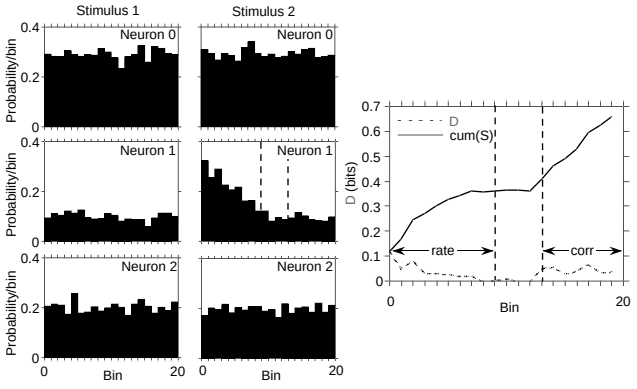

Figure 3: We simulated a three-neuron ensemble responding to two stimulus conditions. The left portion of the display shows PST histograms of each neuron, and these indicate that neuron 1 had a rate response to stimulus 2 (ending at the first vertical dashed line). The right panel shows the result of computing the Kullback-Leibler distance to measure the difference between the two responses. The dashed line shows the statistic computed in each bin and the solid the cumulative value of these component values. Not only does the rate response create a difference between the responses, but also a later response difference not evident in the PST histograms. This difference occurred because of a stimulus-induced correlation between neuron 1 and 2 in the last six bins. Interestingly, the correlation response is nearly as significant for distinguishing the stimuli as the rate response: The contribution of each to the total Gutman statistic is about the same.

We can use the Kullback-Leibler distance computed over portions of the neural response to detail the effectiveness of neural coding and what stimulus aspects are being coded. We used relation (2) between perturbations in detectability and Fisher information to infer responsiveness. For a given operating point, defined by parameter $\boldsymbol{\alpha}_0$, we estimated the change in the response of single lateral superior olive (LSO) neurons to systematic stimulus perturbations about this nominal stimulus condition. These neurons had been thought to be processors of sound location and which disregard other stimulus changes [10]. To test this theory, we chose azimuth and amplitude as our stimulus parameters. We then determined a least-squares fit of equation (2) to the measured Kullback-Leibler distances (incorporating the K-T modification to the types and bootstrapping) to estimate the Fisher information matrix $\widehat{\mathbf{F}}(\boldsymbol{\alpha}_0)$. By considering the concentration ellipse [16] defined by $\overline{\delta\boldsymbol{\alpha}}'\widehat{\mathbf{F}}(\boldsymbol{\alpha}_0)\overline{\delta\boldsymbol{\alpha}} = 1$, we can visually determine the standard deviation of the maximum likelihood estimator for each parameter (found by the extent of the ellipse along coordinate axes) and the general quality of coding from the ellipse's area. From these ellipses (figure 4), we found that the initial transient response of LSO neurons coded sound amplitude more effectively than azimuthal location, while the later portions of the response coded only azimuth. We have thus demonstrated the first known occurrence of a neuron time-multiplexing what it is coding in its discharge pattern.

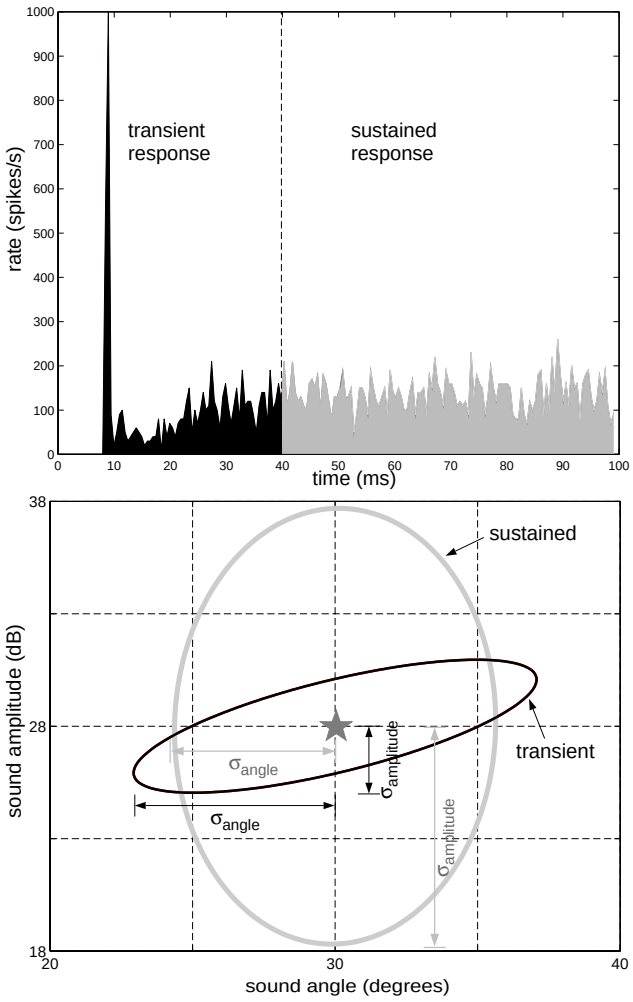

Figure 4: (top) A post stimulus time (PST) histogram showing a typical lateral superior olive response to binaural stimulus. The transient response is characterized by high discharge rates and non stationary behavior. The sustained response is characterized by a constant discharge rate. (bottom) The concentration ellipse from analysis of the transient and sustained responses at the particular operating point marked by the star. The extent of the ellipse along the angle axis shows the standard deviation of the expected estimation error of the optimal unbiased estimator of the angle parameter at this operating point. Similarly, the extent of the the ellipse along the level axis shows the standard deviation of the expected estimation error of the amplitude parameter. Amplitude could be estimated much more accurately based on observation of the transient response. Angle could be estimated slightly more accurately based on observation of the sustained response.

5. CONCLUSIONS

Type-based analysis is the only known technique that can measure the responsiveness of an ensemble, quantify the various contributions to this responsiveness, and provide some insight into the nature of the response. The technique is mathematically well-grounded and uses the amount of available data efficiently. We can also quantify the degree of response detail warranted by the amount of data without making the usual implicit assumptions other variability measures make that the response measure is Gaussian or that the response is stationary. Our results for neural response patterns can be generalized to other situations.

6. 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. W. Bialek, F. Rieke, R. R. de Ruyter van Steveninck, and D. Warland. Reading a neural code. *Science*, 252:1852–1856, 1991.
4. A. G. Carlton. On the bias of information estimates. *Psychological Bulletin*, 71:108–109, 1969.
5. T. M. Cover and J .A. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.
6. A.G. Dabak. *A Geometry for Detection Theory*. PhD thesis, Rice University, Houston, TX, 1992.
7. 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.
8. M. Gutman. Asymptotically optimal classification for multiple tests with empirically observed statistics. *IEEE Trans. Info. Th.*, 35:401–408, 1989.
9. D. H. Johnson, Y. K. Lee, O. E. Kelly, and J. L. Pistole. Type-based detection for unknown channels. In *ICASSP Proc.*, Atlanta, GA, 1996.
10. 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.
11. D.H. Johnson. Point process models of single-neuron discharges. *J. Computational Neuroscience*, 3:275–299, 1996.
12. R. E. Krichevsky and V. K. Trofimov. The performance of universal encoding. *IEEE Trans. Info. Th.*, IT–27:199–207, 1981.
13. 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.
14. M. I. Miller and D. L. Snyder. *Random Point Processes in Space and Time*. Springer-Verlag, New York, second edition, 1991.
15. 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.
16. L. L. Scharf. *Statistical Signal Processing: Detection, Estimation and Time Series Analysis*. Addison-Wesley, Reading, MA, 1991.