ARTICLE **Communicated by Jonathan Victor**

Estimating Information Rates with Confidence Intervals in Neural Spike Trains

Jonathon Shlens
shlens@salk.edu
Salk Institute, La Jolla, CA 92037, and Institute for Nonlinear Science, University of California, San Diego, La Jolla, CA 92093, U.S.A.

Matthew B. Kennel
mkennel@ucsd.edu
Institute for Nonlinear Science, University of California, San Diego, La Jolla, CA 92093, U.S.A.

Henry D. I. Abarbanel
habarbanel@ucsd.edu
Institute for Nonlinear Science, University of California, San Diego, La Jolla, CA 92093, U.S.A., and Department of Physics and Marine Physical Laboratory, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, CA 92093, U.S.A.

E. J. Chichilnisky
ej@salk.edu
Salk Institute, La Jolla, CA 92037, U.S.A.

Information theory provides a natural set of statistics to quantify the amount of knowledge a neuron conveys about a stimulus. A related work (Kennel, Shlens, Abarbanel, & Chichilnisky, 2005) demonstrated how to reliably estimate, with a Bayesian confidence interval, the entropy rate from a discrete, observed time series. We extend this method to measure the rate of novel information that a neural spike train encodes about a stimulus—the average and specific mutual information rates. Our estimator makes few assumptions about the underlying neural dynamics, shows excellent performance in experimentally relevant regimes, and uniquely provides confidence intervals bounding the range of information rates compatible with the observed spike train. We validate this estimator with simulations of spike trains and highlight how stimulus parameters affect its convergence in bias and variance. Finally, we apply these ideas to a recording from a guinea pig retinal ganglion cell and compare results to a simple linear decoder.

1 Introduction

How neurons encode and process information about the sensory environment is an essential focus of systems neuroscience. While much progress has been made (Rieke, Warland, de Ruyter, van Steveninck, & Bialek, 1997; Borst & Theunissen, 1999), many fundamental questions remain. Which features of the sensory world do neurons extract (de Ruyter van Steveninck, & Bialek, 1988; Bialek, Rieke, de Ruyter van Steveninck, & Warland, 1991; Dimitrov, Miller, Gedeon, Aldworth, & Parker, 2003; Paninski, 2003b; Aguera y Arcas & Fairhall, 2003; Sharpee, Rust, & Bialek, 2004)? How do correlations within a spike train (Mainen & Sejnowski, 1995; Berry, Warland, & Meister, 1997; Reinagel & Reid, 2002; Fellous, Tiesinga, Thomas, & Sejnowski, 2004; Abarbanel & Talathi, 2006) or between spike trains of different cells (Mastronarde, 1989; Meister, Lagnado, & Baylor, 1995; Usrey & Reid, 1999) convey information about the stimulus (Gawne & Richmond, 1993; Dimitrov et al., 2003; Rieke et al., 1997; Brenner, Strong, Koberle, Bialek, & de Ruyter van Steveninck, 2000; Warland, Reinagel, & Meister, 1997; Stanley, Li, & Dan, 1999; Gat & Tishby, 1999; Reich, Mechler, & Victor, 2001; Schnitzer & Meister, 2003; Nirenberg & Latham, 2003; Schneidman, Bialek, & Berry, 2003)? Are the tuning properties of sensory neurons optimized for their natural enviornment (Rieke, Bodnar, & Bialek, 1995; Lewen, Bialek, & de Ruyter van Steveninck, 2001; Simoncelli & Olhausen, 2001)? How does adaptation optimize information flow in sensory neurons (Brenner, Bialek, & de Ruyter van Steveninck, 2000; Fairhall, Lewen, Bialek, & de Ruyter van Steveninck, 2001; Meister & Berry, 1999)?

Addressing such questions requires a reliable measure of how much information about the sensory world is contained in a spike train. The average mutual information rate is a general measure of nonlinear correlation between dynamic sensory stimuli and neural responses (Shannon, 1948; Cover & Thomas, 1991). It does not assume a specific model for interpreting a spike train, yet it bounds the amount of information available to any method for reading or decoding a spike train. This quantity allows one to measure in a meaningful way how efficiently neural circuits encode the sensory stimulus (Strong, Koberle, de Ruyter van Steveninck, & Bialek, 1998) or evaluate on an absolute scale the performance of any decoding mechanism, whether biological or artificial (Buracas, Zador, DeWeese, & Albright, 1998; Warland et al., 1997).

A major obstacle to the broad application of mutual information measurements to neural coding problems is that estimating mutual information from experimental data sets is difficult. First, large data requirements of naive estimation techniques restrict the use of information-theoretic analyses to preparations with long-duration recordings. Judging the convergence and bias of conventional estimators of entropy and mutual information is difficult (Treves & Panzeri, 1995; Paninski, 2003a), making the application

of these tools particularly problematic in limited data sets. Second, common methods for estimating information rates require a heuristic judgement of the duration over which the stimulus and response histories affect the response (Theunissen & Miller, 1995). This subjective assessment can largely dominate the final estimate (Kennel, Shlens, Abarbanel, & Chichilnisky, 2005). Third, applying this heuristic precludes an objective estimate of a confidence interval for the information rate.

In this article, we address all of these shortcomings. We introduce a new estimator of information rates for the analysis of neural spike trains that (1) converges quickly to the true value in finite data sets, (2) requires no heuristic judgements, and (3) alone among all other methods provides accurate confidence intervals bounding the range of information rates from sources that most likely generated the observed data. The last issue is important because it permits one to compare information rates in a statistically meaningful manner.

We begin by discussing how to represent a spike train as a time series, and examine what the relevant quantities are for characterizing the information in this representation. We introduce an estimator of information rate, derived from Kennel et al. (2005), and briefly review major issues in estimating these quantities and their associated confidence intervals in neural data. We validate this estimator in simulation and discuss how stimulus parameters in an experiment affect bias and variance in the estimate. Finally, we test this approach on spike data from a guinea pig retinal ganglion cell, calculating the information rate and coding efficiency with confidence intervals, and comparing derived bounds to the performance of a simple linear decoder.

2 Quantifying Information in a Spike Train

We begin with several assumputions about how a spike train represents information. These assumptions ought to be as encompassing as possible, balanced by computational tractability, in order to handle two essential features of neural systems:

1. Spike trains are lists of stereotyped action potentials occurring in continuous time.
2. Spike trains are generated by a dynamical system with temporal dependencies.

In this setting, only the spike times and number of action potentials in a response can convey information about the stimulus.

We represent the spike train as the output of a symbolic dynamical system evolving in time (Lind & Marcus, 1996; Ott, 2002; Gilmore & Lefranc, 2002). By symbolic, we mean that the representation of the spike train is a sequence of integers with a finite alphabet. We select a short time

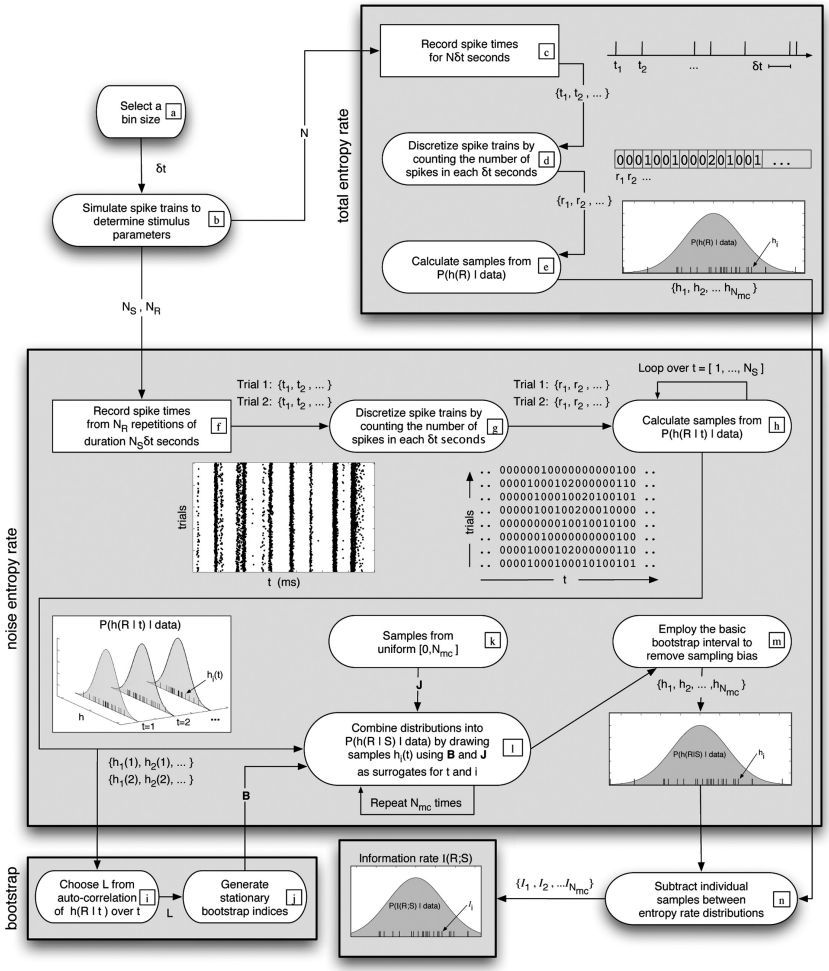
window δt termed the *bin size*, and count the number of spikes in each time bin (MacKay & McCulloch, 1952; Strong et al., 1998; see also Rapp et al., 1994). This is indicated in Figures 1d and 1g. The resulting time course of the observed response is represented by a symbol stream of integers $R = \{r_1, r_2, \ldots, r_N\}$, where each integer r_i is the number of spikes that occur between times $(i-1)\delta t$ and $i\delta t$. Typically, δt is chosen such that each r_i is binary.

The bin size is the effective temporal resolution of the spikes train (Theunissen & Miller, 1995; Reinagel & Reid, 2000). This coarse graining of the spike train assumes that the relevant timescale of neural dynamics is larger than the bin size. As a realization of the underlying symbol source, the symbol stream is now amenable to tools from symbolic dynamics and data compression (Willems, Shtarkov, & Tjalkens, 1995; Kennel et al. 2005).

Information theory provides a framework for quantifying the transmission of information in any communications channel (Shannon, 1948; Cover & Thomas, 1991; MacKay, 2003; Fano, 1961). In neural coding, the sensory stimulus S and neural response R are considered the respective input to and output of a lossy communications channel, represented probabilistically by

---

Figure 1: Block diagram of the entire estimation procedure divided into two larger tasks of estimating the total entropy rate (upper-right box) (see section 5.1). Simulate spike trains and bin at particular temporal resolution in order to determine appropriate stimulus parameters (a–b). Perform experiment and bin spike train at same temporal resolution (c–d). (e) Generate N_{MC} samples from the continuous distribution $P(h(R) \mid \mathsf{data})$ (see section 3.1). Estimate the noise entropy rate by recording multiple repetitions of a stimulus (f) and binning the resulting spike train (g) (see section 3.2). (h) Calculate $P(h(R|t) \mid \mathsf{data})$ using the adapted algorithm to also condition on phase time t. The result is N_S sampled distributions where t and i label the time and sample index in $h_i(t)$. (i) Selecting L to be several times the autocorrelation width suffices, although more sophisticated algorithms exist (see appendix A). (j) Generate a stationary bootstrap **B** to capture the autocorrelation in $h(R|t)$ by acting as a surrogate for the integer time indices t (see appendix A). (k) Draw multiple integers **J** from a uniform probability distribution $\in 1, \ldots, N_{\mathsf{MC}}$ (see section 3.3). (l) Select a set of samples $\{h_i(t)\}$ using **B** and **J** for the time and sample indices and average. Repeat this procedure to generate N_{MC} samples from $P(h(R|S) \mid \mathsf{data})$, each time generating new sets **B** and **J** of time and sample indices (see equation 3.4). (m) Subtract off bias discovered in the bootstrap resampling (see equation 3.5). (n) Subtract individual samples of entropy rates to generate samples of $P(I(S;R) \mid \mathsf{data})$. The mean and quantiles of $P(I(S;R) \mid \mathsf{data})$ provide the final estimate and confidence interval bounding the range of likely average mutual information rates (see section 3.5).



$P(R|S)$ (for reviews, see Rieke et al., 1997, and Borst & Theunissen, 1999).[1] In our framework, the stimulus could be a continuous or discrete signal, but the response will always be a stream of discrete integers. The goal is to quantify the nonlinear correlation or statistical dependence between the stimulus and spike train response.

[1] $P(X)$ denotes the probability distribution of the random variable X, while $P(X = x)$ denotes the probability that $X = x$. For convenience, we sometimes abbreviate the latter as $P(x)$, where X is implied.

The classical functional for measuring statistical dependence between two distributions is the mutual information $\log_2 \frac{P(s,r)}{P(s)P(r)}$ (Fano, 1961), where $P(s)$ and $P(r)$ are the probabilities of observing stimulus $s \in S$ and response $r \in R$. $P(s, r)$ is the joint probability for these events. One usually deals with the average of this statistic over the joint distribution (Cover & Thomas, 1991),

$$\mathcal{I}(S;\, R) = \sum_{s \in S, r \in R} P(s, r) \log_2 \frac{P(s, r)}{P(s)P(r)}.$$

$\mathcal{I}(S;\, R)$ is the average mutual information and answers the question, How much in bits, on average, do we learn about the observation r when we observe s?[2] The average mutual information can also be expressed as a difference in the entropy $H[\cdot]$ of the distributions,

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

where the entropy is defined,

$$H(S) = -\sum_{s \in S} P(s) \log_2 P(s),$$

and similarly for $H(R)$. The average conditional entropy is $H(S|R) = \sum_{r \in R} P(r) H(S|R = r)$, with

$$H(S|R = r) = -\sum_{s \in S} P(s, r) \log_2 \frac{P(s, r)}{P(r)}.$$

The benefit of this formulation is that it is easy to generalize this quantity to time series with temporal dependencies.

In a non-Poisson spike train R, successive symbols $r_{t-1}, r_t, \ldots$ are not statistically independent. The stationary distribution associated with a symbol stream is the conditional probability distribution $P(r_t|r_{t-1}, \ldots, r_{t-D})$, in which the symbol r_t observed at time t depends on D previous symbols. D is called the *conditioning depth* or the order of a Markov process. In a dynamical system, the conditioning depth is potentially infinite (Hilborn, 2000;

[2] A second interpretation is to recognize that $\mathcal{I}(S;\, R)$ measures the Kullback-Leibler divergence between the joint distribution and the product of the marginals and thus measures how inefficient the product of the marginal distribution is at encoding the joint distribution (Cover & Thomas, 1991).

Ott, 2002), though in practice, measurements depend on a finite number of previous symbols. In the limiting case, we define the entropy rate,

$$h(R) = \frac{1}{\delta t} \lim_{D \to \infty} H\left[P(r_t | r_{t-1}, \ldots, r_{t-D})\right], \tag{2.1}$$

with

$$H\left[P(r_t | r_{t-1}, \ldots, r_{t-D})\right] = -\sum_{r_t, \ldots, r_{t-D}} P(r_t, \ldots, r_{t-D}) \log_2 P(r_t | r_{t-1}, \ldots r_{t-D}).$$

$h(R)$ has units of bits per second. The entropy rate quantifies the uncertainty of the next symbol given knowledge of the infinite prior history of the underlying symbol source (Cover & Thomas, 1991; Kennel et al., 2005).

In a neuroscience context, the stimulus S and neural response R are two dynamical time series. The generalization of the average mutual information is the average mutual information rate expressed as the difference of entropy rates,

$$I(S; R) = h(R) - h(R|S) = h(S) - h(S|R). \tag{2.2}$$

This quantity expresses the average rate of novel information (bits per second) that newly observed responses provide about the stimulus time series. It is symmetric in the response and stimulus.

The average information rate can mask the contribution of specific stimuli or responses that are particularly informative because it is an average over all pairs of stimuli and responses. It is of interest then to quantify the information attributed to a single stimulus or neural response (DeWeese & Meister, 1999; Brenner, Strong et al., 2000; Butts, 2003; Bezzi, Samengo, Leutbeg, & Mizmori, 2002),

$$\begin{aligned} I_{sp}(s) &\equiv h(R) - h(R \mid S = s) \\ I_{sp}(r) &\equiv h(S) - h(S \mid R = r). \end{aligned} \tag{2.3}$$

These specific information rates quantify the amount of information in a single momentary stimulus or neural response. They are natural decompositions of the average mutual information rate,

$$I(S; R) = \sum_{r \in R} P(r) I_{sp}(r) = \sum_{s \in S} P(s) I_{sp}(s).$$

The specific information rates need not be positive, although the average mutual information rate is always nonnegative (Fano, 1961; DeWeese & Meister, 1999). We now discuss how to estimate these quantities from an

experimental data set by extending estimators of entropy rate from previous work.

3 Estimating Information in a Spike Train

The definition of average mutual information rate, equation 2.2, provides two alternatives for calculating the information rate in a neural spike train. One can estimate the information rate by examining the reduction in uncertainty by conditioning in stimulus space (Bialek et al., 1991) or response space (Strong et al., 1998). Stimulus space can be high dimensional and continuous, and thus difficult to sample experimentally. Therefore, we select the alternate formulation,

$$I(S;\, R) = h(R) - h(R|S),$$

because we can estimate the information rate by measuring solely the neural responses in a suitably designed experiment. In this framework $h(R)$ is termed the *total entropy rate*, h_{total}, and $h(R|S)$ the *noise entropy rate*, h_{noise} (Strong et al., 1998). We estimate h_{total} and h_{noise} from separate data sets recorded under different stimulus presentations and combine these results to estimate the information rate $I(S;\, R) = h_{\text{total}} - h_{\text{noise}}$. Each step of this entire procedure is outlined in Figure 1.

3.1 Estimating Total Entropy Rates. The total entropy rate h_{total} is the average rate of uncertainty observed in a neural spike train response over a (stationary) stimulus distribution. Experimentally, this means that we present a stimulus to a neuron to sample all possible spiking patterns over a long duration $N\delta t$. Typically, $N\delta t$ ranges from several hundred to several thousand seconds (Borst & Theunissen, 1999). The recorded spike train is discretized at a small bin size δt, producing a list of integer spike counts $R = \{r_1, r_2, \ldots, r_N\}$ where the subscript indexes time. These two steps are diagrammed in Figures 1c and 1d. We have discussed in detail how to estimate the entropy rate of a finite-length symbol stream in a related work (Kennel et al., 2005), but we review these ideas briefly here.

A paramount goal of entropy rate estimation is to determine the appropriate conditioning depth D of the underlying symbol source.[3] We must balance the selection of D as diagrammed in Figure 2 between two opposing biases on the estimate:

- Unresolved temporal dependencies due to finite D
- Undersampled probability distributions due to finite N

[3] The word length (Strong et al., 1998; Borst & Theunissen, 1999; Reinagel & Reid, 2000) is roughly equivalent to the conditioning depth D, but see Kennel et al. (2005).

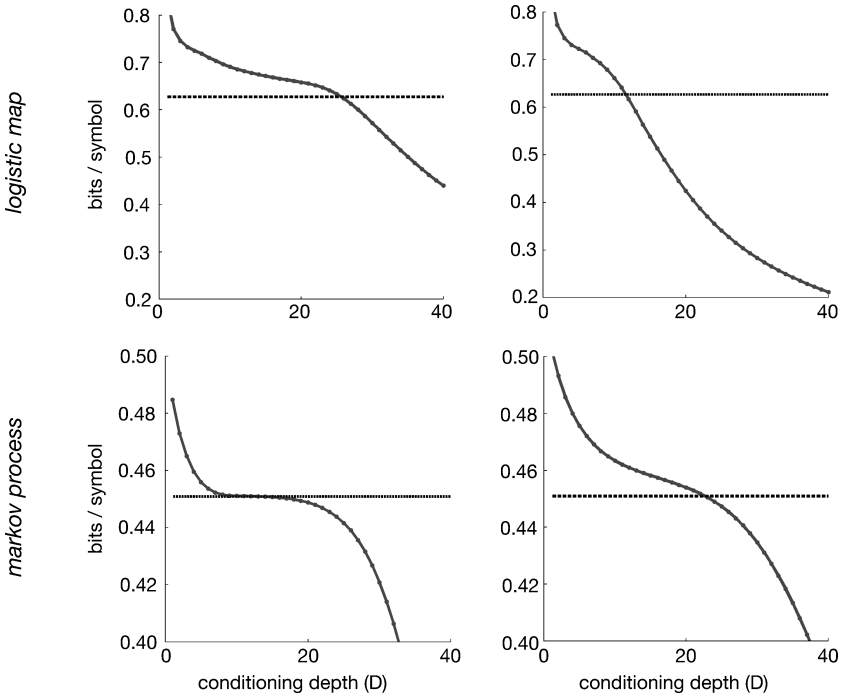


Figure 2: Selecting the conditioning depth D can dominate entropy rate estimation. This pedagogical figure reproduced from Kennel et al. (2005) demonstrates how temporal complexity and small data sizes affect entropy rate estimation in a discrete Markov process ($D = 10$) and a symbolized logistic map ($D = \infty$). Both symbolic systems use binary alphabets and have a maximum possible entropy rate of 1 bit per symbol. For $O(10^6)$ and $O(10^5)$ simulated symbols (left and right columns, respectively), the entropy rate was estimated as a function of conditioning depth $\hat{h}(D)$ using a naive entropy estimator (Miller & Madow, 1954). The dashed line is the true entropy rate calculated analytically. A common heuristic is to identify the final estimate of entropy rate as the plateau on this graph (Strong et al., 1998; Schurmann & Grassberger, 1996). This qualitative assessment is not reliable and can fail in the limit of small data sets (right column) and complex temporal dynamics (top row) (Kennel et al., 2005).

This requires selecting the appropriate probabilistic model complexity for a finite data set, a classic issue in statistical inference and machine learning (Duda, Hart, & Stork, 2001).

We resolve this problem by using a model weighting technique termed *context tree weighting*, originally developed for lossless data compression (Willems et al., 1995; Kennel & Mees, 2002; London, Schreibman, Hausser, Larkum, & Segev, 2002; Kennel et al., 2005), following the spirit of the

minimum description length principle (Rissanen, 1989). This weighting, along with local Bayesian entropy estimators (Wolpert & Wolf, 1995; Nemenman, Shalee, & Bialek, 2002), yields a direct estimator of the entropy rate of R, $\hat{h}_{\mathsf{total}}$.

We consider the range of plausible entropy rates compatible with the finite data set, not just a single point estimate. We take the Bayesian perspective and view the estimate $\hat{h}_{\mathsf{total}}$ as the expectation of a posterior distribution of entropy rates most consistent with the observed data,

$$\hat{h}_{\mathsf{total}} = \int h_{\mathsf{total}}\, P(h_{\mathsf{total}} \mid \mathsf{data})\, dh_{\mathsf{total}}.$$

The width of $P(h_{\mathsf{total}} \mid \mathsf{data})$ is a Bayesian confidence interval about its estimate. A wide (narrow) distribution implies large (small) uncertainty about the estimate.

The numerical algorithm presented in Kennel et al. (2005) yields N_{MC} numerical samples $\{h^*_{\mathsf{total}}\}$ of the entropy rate drawn from $P(h_{\mathsf{total}} \mid \mathsf{data})$ using Monte Carlo techniques (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953; Hastings, 1970; Gamerman, 1997). This step is diagrammed by Figure 1e. The mean of these N_{MC} samples is the estimate of $\hat{h}_{\mathsf{total}}$, and 5% and 95% quantiles of this empirical distribution give a 90% confidence interval.

3.2 Estimating Noise Entropy Rates. The noise entropy rate h_{noise} measures the reliability of a neuron in response to a stimulus. This term quantifies the amount of uncertainty in the response that is not attributable to the stimulus. This entropy rate can also be difficult to estimate in a finite data set. In this section we discuss how to extend the algorithm presented in Kennel et al. (2005) for estimating $h(R)$ to provide an estimator for $h(R|S)$.

The quantity h_{noise} is a conditional entropy rate and thus a weighted average over all conditional entropies,

$$\begin{aligned} h(R|S) &= \frac{1}{\delta t} \sum_{r_t, \mathsf{hist}} P(r_{\mathsf{hist}}, s_{\mathsf{hist}}) H\left[P(r_t | r_{\mathsf{hist}}, s_{\mathsf{hist}})\right] \\ &= \left\langle h(R|S = s_{\mathsf{hist}}) \right\rangle_S , \end{aligned} \tag{3.1}$$

where hist denotes all previous values up to but not including time t. Thus, r_{hist} and s_{hist} denote conditioning histories of response and stimulus space, respectively. To calculate this quantity, we must average over all possible stimuli (i.e., stimulus histories s_{hist}). This can be experimentally infeasible because stimulus space might be continuous and high dimensional.

A clever trick to circumvent this problem is to design an experiment where one replays the same time-varying stimulus segment over multiple

trials (Mainen & Sejnowski, 1995; Bair & Koch, 1996; de Ruyter van Steveninck, Lewen, Strong, Koberle, & Bialek, 1997; Strong et al., 1998). Typically, a stimulus of duration of 5 to 30 seconds is replayed dozens to hundreds of trials. The notion of phase time as used here refers to the time elapsed since the beginning of the trial. We index the phase time (measured in units of the time bin size δt) by an integer $t \in [1, \ldots, N_S]$. With N_R repeated stimulus presentations, this generates a raster plot (see Figure 1f or Figure 8), in which a neuron has received practically the same stimulus history s_{hist} at each phase time t. In other words, we equate each moment in phase time t with a particular stimulus history s_{hist}. We additionally assume that each repetition of the stimulus is sufficiently long so each trial is reasonably ergodic. This means

$$\begin{aligned} h(R|S) &= \langle h(R|S = s_{\text{hist}}) \rangle_S \\ &= \lim_{N_S \to \infty} \langle h(R|t) \rangle_{t=1}^{N_S}. \end{aligned} \tag{3.2}$$

We emphasize that this assumption might not hold when N_S is small or the stimulus contains temporal correlations as commonly observed in natural stimuli (Simoncelli & Olhausen, 2001). In practice, for each moment in phase time t we have observed N_R draws from the unknown conditional distribution $P(r_t \mid r_{\text{hist}}, s_{\text{hist}})$. In other words, each presentation has conditioned the response on the same (unspecified) stimulus history s_{hist}. The stimulus-conditioned response is the very distribution in equation 3.1 of which we need to calculate the entropy. The entropy of this distribution, averaged over stimulus space (or phase time), is the noise entropy.[4]

We use the same algorithm as previously used for $\hat{h}_{\text{total}}$ as a component, this time forming a new context-tree-based estimate at each phase time t from the N_R histories. The observations or draws of this conditional distribution are read off going "downward" in a raster plot as diagrammed in Figure 3, but the conditioning history is back in phase time as usual. Each of the N_S trees is processed in Figure 1h, following Kennel et al. (2005), providing N_{MC} samples from $P(h_{\text{noise}}(t) \mid \text{data})$, the Bayesian posterior for the instantaneous noise entropy rate. We denote the jth numerical sample

[4] We note that the definition of the noise entropy rate differs from the method commonly used in practice (Strong et al., 1998). The average noise entropy rate is by definition the average over stimulus space of all individual noise entropy rates $h(R|S = s)$. The direct method calculates the average entropy across stimulus space for varying conditioning depths; subsequently, it derives an entropy rate estimate from the average entropies. This out-of-order procedure in practice avoids many complicated extrapolations (Kennel et al., 2005).

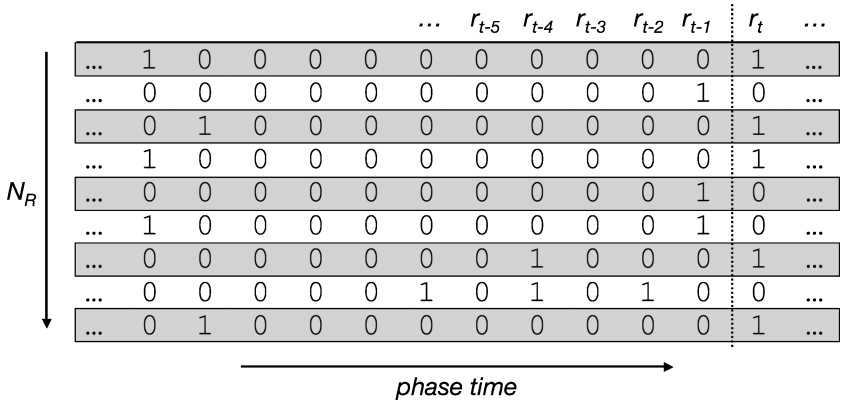


Figure 3: A diagram demonstrating how to estimate $P(r_t \mid r_{\text{hist}}, s_{\text{hist}})$ and, correspondingly, the noise entropy rate $\hat{h}_{\text{noise}}(t)$, conditioned on a particular stimulus. Each row is a discretized spike train response to a repeated stimulus. The set of N_R time series ending prior to some moment in phase time t shows the conditioning histories r_{hist}, explicitly, and s_{hist}, implicitly. With N_R samples of histories, we use the context tree to predict each r_t from its past history and estimate $\hat{h}_{\text{noise}}(t) = H\,[P(r_t|r_{\text{hist}}, s_{\text{hist}})]$, the instantaneous noise entropy rate.

of $P(h_{\text{noise}}(t) \mid \text{data})$ as $h^*_j(t)$. Averaging over those samples yields

$$\hat{h}_{\text{noise}}(t) = \frac{1}{N_{\text{MC}}} \sum_{j=1}^{N_{\text{MC}}} h^*_j(t).$$

The estimated average noise entropy rate is then the average over all phase times,

$$\hat{h}_{\text{noise}} = \frac{1}{N_S} \sum_{t=1}^{N_S} \hat{h}_{\text{noise}}(t). \tag{3.3}$$

3.3 Estimating the Distribution of Noise Entropy Rates. In addition to the estimate $\hat{h}_{\text{noise}}$, we also wish to estimate the distribution of likely values for the noise entropy, $P(h_{\text{noise}} \mid \text{data})$. Percentiles of this quantity will provide confidence intervals for $\hat{h}_{\text{noise}}$. Thus far, we have only computed samples from $P(h_{\text{noise}}(t) \mid \text{data})$ for every point in phase time. How do we appropriately combine these N_S sampled distributions into an overall distribution to provide for an error bar on $\hat{h}_{\text{noise}}$? At this point, one must make a philosophical decision about what types of variability are intended to be represented by the confidence interval. The proper procedure depends on

the intended use and interpretation of the estimated value. We identify two sources of variability, which conceivably ought to be represented in a useful confidence interval:

1. Variability due to estimation error point-wise in phase time, because only a finite number of repeats were recorded
2. Variability across the stimulus process itself, because only a finite duration stimulus was presented

The first source of variability quantifies uncertainty in the noise entropy rate pointwise in phase time due to finite samples (N_R). Within typical data sets (N_R ranges from 50–1000 repetitions), the width of this distribution, $P(\hat{h}_{\mathsf{noise}}(t) \mid \mathsf{data})$, tends to be narrow, and thus the variability attributable to point-wise estimation error is small (see, for instance, Figure 8, bottom).

The second source of variability is more subtle and often not explicitly recognized. Most often, we want to estimate $h(R|S)$ from the data, where S is the stimulus process, not the individual stimulus time series actually used in the experiment. In principle, to estimate this variability, we would repeat our experiment with many stimulus time series, each with many repeats; the variability across the ensemble of rasters (where each raster corresponds to a unique stimulus time series) would reflect the variability in the stimulus process. In practice, long-duration recordings are difficult, and thus only a single raster plot is measured in response to a single instance from the stimulus process. Thus, the goal set forth is to estimate the variability across an ensemble of raster plots from just a single raster plot. We employ a bootstrap procedure, detailed below, to achieve this goal.

Because the underlying stimulus process is (presumed to be) unobserved, the variability in the stimulus process is instead reflected in the range of spike patterns observed across phase time in a raster plot (see Figure 4a, top panel). In practice, we find this source of variability quite large because a neuron can fluctuate from nearly determinstic silence ($h_{\mathsf{noise}} \approx 0$) to bursting activity ($h_{\mathsf{noise}} \gg 0$) (see Figure 4a, top panel at $t = 200,\ 400$ ms, respectively) depending on the stimulus.

In order to estimate the effect of variation across the stimulus process, we use a generic time-series bootstrap of the observed effect of stimulus variation on $\hat{h}_{\mathsf{noise}}(t)$ to estimate the expected consequences on $\hat{h}_{\mathsf{noise}}$. We find ensembles of noise entropy rates based on resampled rasters, as if there were a new stimulus composed of randomized segments of the original stimulus. A bootstrapped sample is

$$h^*_{\mathsf{boot}} = \frac{1}{N_S} \sum_{t=1}^{N_S} h^*_{J_t}(B_t). \tag{3.4}$$

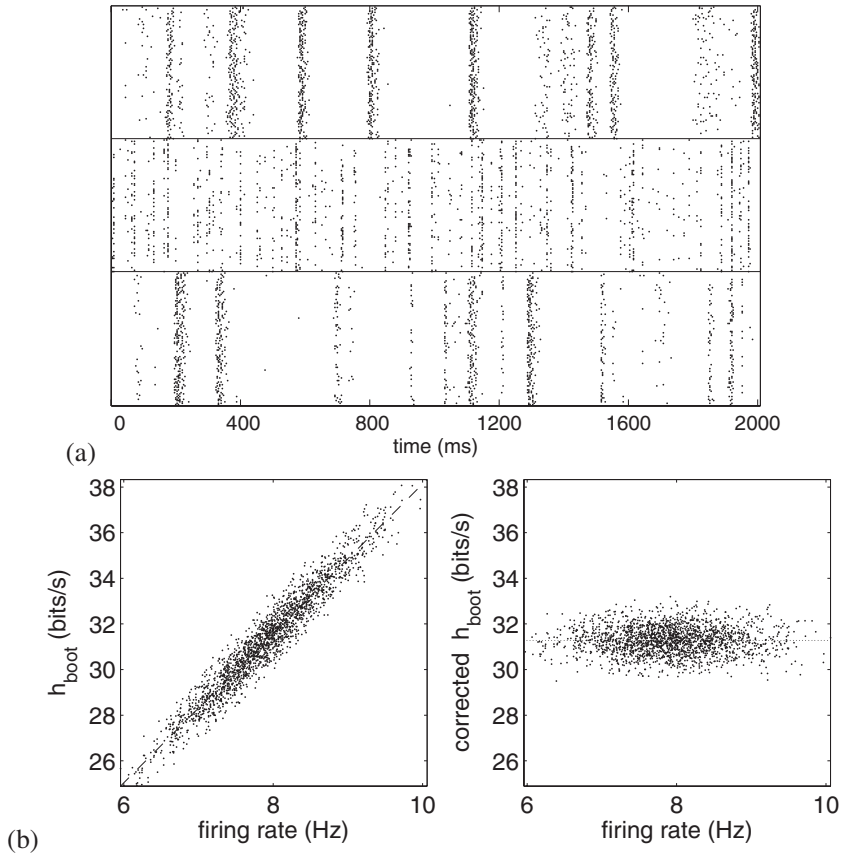


Figure 4: Effects of correlation and spike rate on bootstrap procedure on a short repeated stimulus segment from experimental data set ($\delta t = 4$ ms; see section 5). (a) Top: original raster plot. Middle: raster with phase time indices **B** resampled uniformly with replacement. Bottom: raster with phase times indices **B** resampled using stationary bootstrap procedure. (b) Conditioning on a constant spike rate. Left: noise entropy samples h^*_{boot} versus firing rate r^* over bootstrapped samples, demonstrating typical linear dependence (dashed line). Right: noise entropy samples corrected using linear Ansatz. The width of the distribution in h^*_{boot} (vertical dispersion) is substantially lowered using this procedure.

where, for each sample h^*_{boot}, B_t is a draw of time indices in $[1 \ldots N_S]$ (details discussed below) and J_t is a uniform random draw from the sample indices $[1,\ N_{\text{MC}}]$ with replacement. Procedurally, we draw a random bootstrap time series $\mathbf{B} = \{B_1, \ldots, B_{N_S}\}$ of time indices and, at each of those moments, draw

one of the N_{MC} samples from the point-wise estimated posterior and average them to give one bootstrap sample, h^*_{boot}.[5]

We employ the basic bootstrap interval (Davison et al., 1997) to remove bias discovered during the resampling procedure. The basic bootstrap interval presumes that the distribution of $\hat{h}_{noise} - h_{noise}$ (where h_{noise} is unknown) may be identified with the distribution of $h^*_{boot} - \hat{h}_{noise}$ (where h^*_{boot} has been sampled by the bootstrap). Samples of our estimate for the distribution of the noise entropy rate h^*_{noise} are computed using equations 3.3 and 3.4:

$$h^*_{noise} = \hat{h}_{noise} - (h^*_{boot} - \hat{h}_{noise}) = 2\hat{h}_{noise} - h^*_{boot}. \tag{3.5}$$

These samples are a hybrid Bayesian and bootstrap estimate of the distribution of h_{noise}, and quantiles of h^*_{noise} provide confidence intervals for the noise entropy.

The appropriate procedure for choosing **B** is subtle because the neural response contains significant temporal correlations across phase time. These correlations may be due to refractory neural dynamics as well as the effect of nearly overlapping stimuli within small shifts of phase time. These correlations in phase time create correlations between $\hat{h}_{noise}(t)$ and $\hat{h}_{noise}(t+1)$ and are reflected in the raster plot by the thickness of each burst of activity (or silence) in Figure 4a, top panel. A naive bootstrapping procedure for **B** is to uniformly draw random time indices (with replacement) from $[1, N_S]$. This standard procedure would result in creating the raster plot in Figure 4a, middle panel. Clearly, this procedure fails to capture the autocorrelation structure evident in Figure 4a, top panel. Correspondingly, when significant response autocorrelation exists, a naive bootstrap that presumes point-wise independence in phase time, this procedure will profoundly underestimate the true variance of $\hat{h}_{noise}$ across the underlying stimulus process.

To account for this significant correlation, we propose drawing bootstrapped phase times using a procedure developed by Politis and Romano for correlated time series (Politis & Romano, 1994; Politis & White, 2004). The Politis-Romano (PR) bootstrap procedure creates surrogate time indices **B** consisting of varying-length blocks of consecutive phase times (each block starting at random phase times), whose purpose is to capture most of the the autocorrelation structure in $\hat{h}_{noise}(t)$ (Figures 1i and 1j; see appendix A for details). This correlation structure can be quite large at small bin sizes or with stimuli exhibiting long temporal correlations (e.g., natural stimuli). Returning to Figure 4a, the phase time indices **B** are now resampled using the PR bootstrap. This bootstrap procedure captures the autocorrelation structure evident in the response (Figure 4a, bottom panel). With this bootstrap, the confidence interval calculated as quantiles on equation 3.5 better

[5] Note that if one did not want to include the stimulus variation in the confidence interval, one would select $B_t = t$.

capture the variability due to the underlying stimulus process. The consequences of this source of variability are discussed further in the application of this estimator.

3.4 Conditioning Noise Entropy Rate on a Fixed Spike Rate. Entropy rates are significantly correlated with the spike rate. For example, periods of silence in a raster plot correspond to $\hat{h}_{\text{noise}}(t) = 0$ reflecting practically deterministic activity (Kennel et al., 2005). Thus, sample h^*_{boot} will contain, by random chance, phase time indices **B** with fewer or greater periods of time with absolute silence significantly effecting its value. Figure 4b (left panel) demonstrates this empirically linear, systematic effect.

This source of variability is artificially introduced by the resampling procedure. Samples drawn with larger (smaller) proportions of phase times **B** with spiking activity exhibit higher (lower) entropy rates. We may null out this source of variability by adapting the resampling procedure. Conceptually, we are now estimating $h(R|S,\ \text{rate} = r_0)$, where r_0 is the empirically estimated average spike rate. For each moment in phase time t, we estimate the instantaneous spike rate $r(t)$. For each bootstrap sample h^*_{boot}, there is a corresponding set of phase time indices **B** and spike rate $r^* = \frac{1}{N_S}\sum_t r(B_t)$. We use least squares to fit a linear Ansatz between the sample spike rate and entropy rate, $h^*_{\text{boot}} \approx \alpha + \beta(r^* - r_0)$, where $\{\alpha, \beta\}$ are determined with an ordinary least squares fit (see Figure 4b, left panel, dashed line). Each entropy rate sample is corrected as if it had spike rate r_0,

$$h^{**}_{\text{boot}} = h^*_{\text{boot}} + \beta(r_0 - r^*). \tag{3.6}$$

These corrected samples can be plugged into equation 3.5, replacing h^*_{boot}. This procedure corrects for variability introduced by the bootstrap procedure, which is seen primarily through the altered spike rate and effectively shrinks the confidence interval of the noise entropy distribution when this effect is removed.

3.5 Estimating Mutual Information Rates. We now have all of the elements necessary to estimate the specific and average information rates. We remind the reader that the ergodic assumption in section 3.2 equated a specific stimulus history with a particular moment in phase time (i.e., $s_{\text{hist}} \sim t$). Because the total and noise entropy rate distributions are estimated from separately recorded experimental data sets, we can simulate samples of the likely information rates by drawing independent samples from the posterior distributions and taking their difference,

$$I^* = h^*_{\text{total}} - h^*_{\text{noise}} \tag{3.7}$$

$$I^*_{sp}(s) = h^*_{\text{total}} - h^*_{\text{noise}}(t). \tag{3.8}$$

This step is diagrammed in Figure 1n. The mean of these samples gives the estimated information rate, and the central portion gives a mixed Bayesian and bootstrap confidence interval on the range of information rates consistent with the data. We now test these ideas in simulation and highlight results on an experimental neural data set.

4 Testing the Estimator with Simulation

In previous work (Kennel et al., 2005), we validated and compared the performance of our estimator for the total entropy rate using simulations of nonlinear dynamical systems. In summary, we found that our estimate was asymptotically unbiased and converged reliably and rapidly compared to other estimators (Strong et al., 1998; Lempel & Ziv, 1976; Amigo, Szczepanski, Wajnryb, Sanchez-Vives, 2004; Kennel & Mees, 2002; London et al., 2002; Kontoyiannis, Algoet, Suhov, & Wyner, 1998). Furthermore, the size of the confidence interval estimate provided by the entropy rate estimator well matched the variation in entropy rate under sample fluctuations.

We now examine the convergence of our estimators and the confidence interval of $\hat{h}_{\text{total}}$ and $\hat{h}_{\text{noise}}$ using a simulation of an inhomogeneous Poisson neuron with an absolute and relative refractory period (Nemenman, Bialek, & de Ruyter van Steveninck, 2004; Victor, 2002). The parameters of this model have been chosen to approximate the statistics of a real neuron and binned at $\delta t = 4$ ms; see appendix B for details. The goals of these simulations are to (1) validate our estimates of $\hat{h}_{\text{total}}$, $\hat{h}_{\text{noise}}$ and their asssociated confidence intervals and (2) judge the necessary amount of data required for a good estimate in real neural data.

4.1 Testing Convergence. We examined the convergence of $\hat{h}_{\text{total}}$ using an inhomogeneous refractory Poisson simulation whose parameters are fit to model the temporal dynamics of a recorded neuron (see appendix B). In previous work we detemined that this estimator is asymptotically unbiased (Kennel et al., 2005). We judge the convergence and consistency of the estimator about an approximate truth ($N\delta t = 4000$ s, $N_R = 3000$ trials, $N_S\delta t = 32$ s; see also Nemenman et al., 2004) by examining the estimator bias and variance. For each duration $N\delta t$, we generated 100 independent data sets and computed $\hat{h}_{\text{total}}$ for each data set. In Figure 5a, the estimator bias is the difference between approximate truth and the mean of these 100 independent estimates. The estimator variance is the square of the accompanying error bar, and it measures the intrinsic fluctuations in this random process. We emphasize that this error bar is not the estimated confidence interval, but rather is the actual posterior distribution of entropy rates, which we will estimate later. Over increasing $N\delta t$, the estimator variance decreases, and within 125 s of data, the bias and variance (normalized by $\hat{h}_{\text{total}}$) are, respectively, less than 0.2% and 1.3%, suggesting a well-converged estimate.

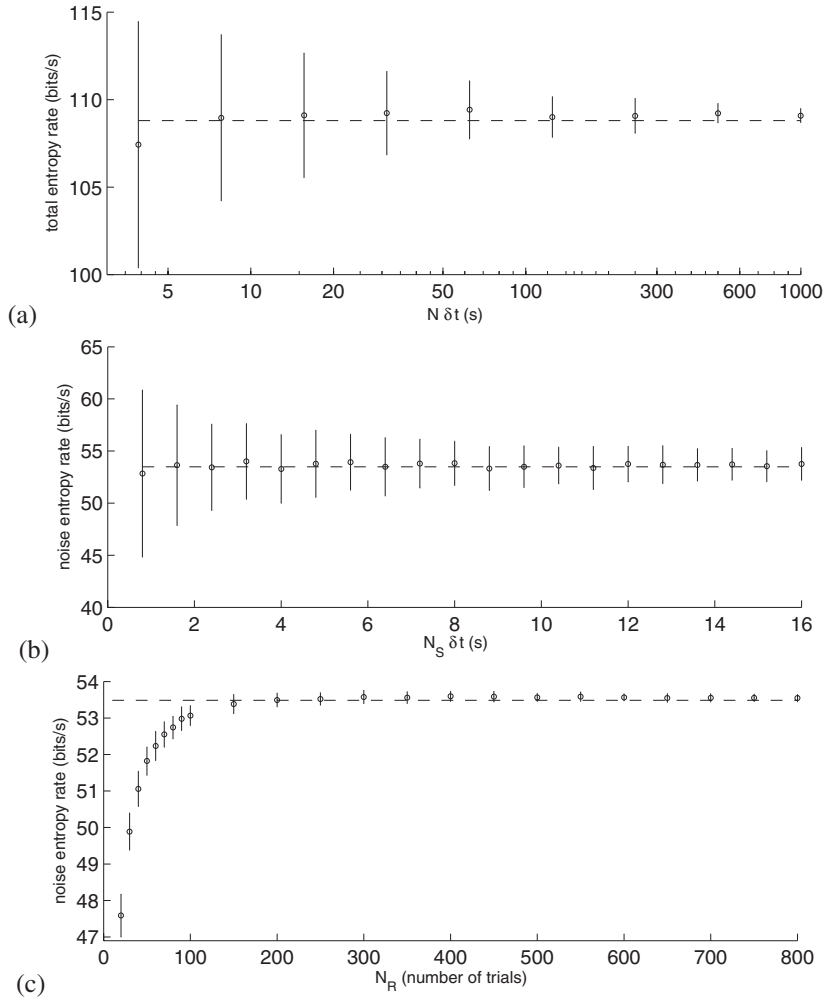


Figure 5: Convergence of entropy rate estimates on a simulated refractory Poisson neuron. Shown are ensembles over 100 different data sets; error bars give sample standard deviations over the data set ensembles. Dashed line is approximate truth ($N\delta t = 4000$ s, $N_R = 3000$ trials, $N_S\delta t = 32$ s). (a) Convergence of $\hat{h}_{\text{total}}$ in N. (b) Convergence of $\hat{h}_{\text{noise}}$ in N_S for a fixed number of trials $N_R = 400$. (c) Convergence of $\hat{h}_{\text{noise}}$ in N_R for a fixed trial duration $N_S\delta t = 8$ s.

Examining the convergence of $\hat{h}_{\text{noise}}$ is more subtle because the estimator contains two parameters, N_R and N_S. Because the convergence is difficult to visualize in multiple dimensions, we instead examine the convergence along each parameter while holding the other fixed at a reasonable value.

In Figure 5b, we again generated 100 independent data sets. This time each data set is an entire raster plot with N_S varied and $N_R = 400$ trials fixed. Each data set contains a new instantiation of the random process, as well as a new driving firing rate (λ_2) over $N_S \delta t$ seconds (see appendix B). The bias is the difference between approximate truth and the average of these 100 estimates. The estimator variance is again the width of the posterior distribution of entropy rates, although the posterior distribution of $\hat{h}_{\text{noise}}$ is calculated by applying the stationary bootstrap (see section 3.3) to account for correlations in time. The estimator bias remains small ($< 1\%$ of $\hat{h}_{\text{noise}}$), yet the estimator variance shrinks as N_S increases.[6]

Finally in Figure 5c, we judge the convergence of $\hat{h}_{\text{noise}}$ over N_R while holding $N_S \delta t = 8$ s fixed. In this panel, each data set is a new instantiation of the random process but retains the same firing rate (λ_1). Unlike Figure 5c, we do not select a new firing rate time series because we are interested in showing the significant effect of N_R on bias, and drawing new firing rate time series partially obscures the effect by submerging it in additional variance. The estimator bias is profound at small N_R, dominating the magnitude of the variance (Paninski, 2003a); however, by $N_R = 400$ trials, the bias is well contained within the variance of the estimate. Thus, we find that N_R dominates the bias in $\hat{h}_{\text{noise}}$ as compared to N_S. In other words, N_R dominates the bias, while N_S dominates the variance almost independently. Thus, variations across phase time (or stimulus histories s_{hist}), but not the error bar for individual $\hat{h}_{\text{noise}}(t)$, explain most of the variance in $\hat{h}_{\text{noise}}$. We return to this point in section 6 in discussing how to select experimental parameters for estimating these quantities.

4.2 Testing the Confidence Interval. We now examine how well the confidence interval, which can be estimated from a single data set, can replicate the actual variation of the underlying estimate seen in a large ensemble of new data sets. To test the confidence intervals for $\hat{h}_{\text{total}}$, we again generate 100 independent data sets with a fixed $N\delta t$. For each data set, we calculate $\hat{h}_{\text{total}}$ and the associated single trial confidence interval. We plot these estimates as circles and error bars sorted by increasing $\hat{h}_{\text{total}}$ in Figure 6a ($N\delta t = 125$ s). First, we note that approximate truth (dashed line) is contained within 90% of the individual confidence intervals. This indicates that the estimated confidence interval is well calibrated.

To summarize these results, we calculate the following. We consider the true distribution of this statistic (or posterior distribution of $\hat{h}_{\text{total}}$) to be the distribution of $\hat{h}_{\text{total}}$ (circles). The central 90% quantile of this posterior

[6] As an aside, notice that we are combining a Bayesian and frequentist bootstrap method for the estimator, and then testing it in a purely frequentist fashion (many draws from the source). The size and location of a Bayesian error bar need not match exactly the result from a frequentist-style experiment, but the general compatibility of the results here gives reassurance there are no peculiar anomalies.

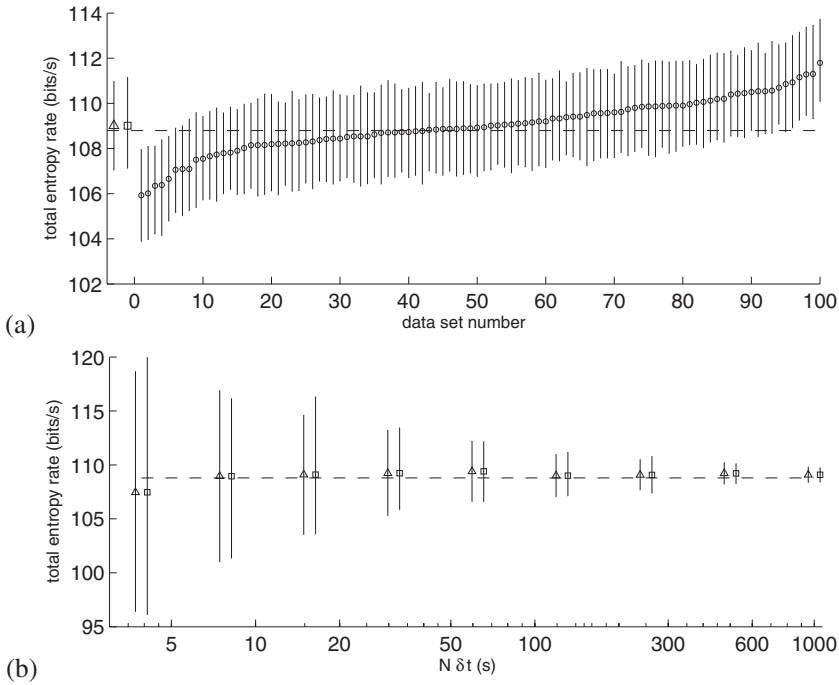


Figure 6: Testing the confidence interval of $\hat{h}_{\text{total}}$. (a) Individual estimates and associated 90% confidence intervals of 100 independent data sets of length $N\delta t = 125$ s (circles). Average of the individually estimated confidence intervals (triangle). 90% quantile on the distribution of $\hat{h}_{\text{total}}$ (circles), which measures actual posterior of h_{total} (square). The horizontal dashed line is approximate truth ($N\delta t = 4000$ s). (b) Comparison between average lower and upper confidence interval limits (triangle) and posterior distribution (square) across $N\delta t$. Across varying data set lengths, confidence intervals are well calibrated and engulf approximate truth.

distribution is denoted by the left-most square (see Figure 6a). We compare this error bar to the average lower and upper limits of the 100 individual confidence intervals (see Figure 6a, triangle). In Figure 6a, the average upper and lower limits contain approximate truth and match the estimated posterior distribution. We use these two statistics as a summary of the calibration and compare the two across varying $N\delta t$ (see Figure 6b). Across varying $N\delta t$, the average confidence interval matches the true posterior distribution and engulfs truth.

We repeat this same procedure for examining the confidence interval in $\hat{h}_{\text{noise}}$ across N_S and N_R (see Figure 7). In Figure 7a, we again draw a new instantiation of the random process and the firing rate (λ_2) for every sample,

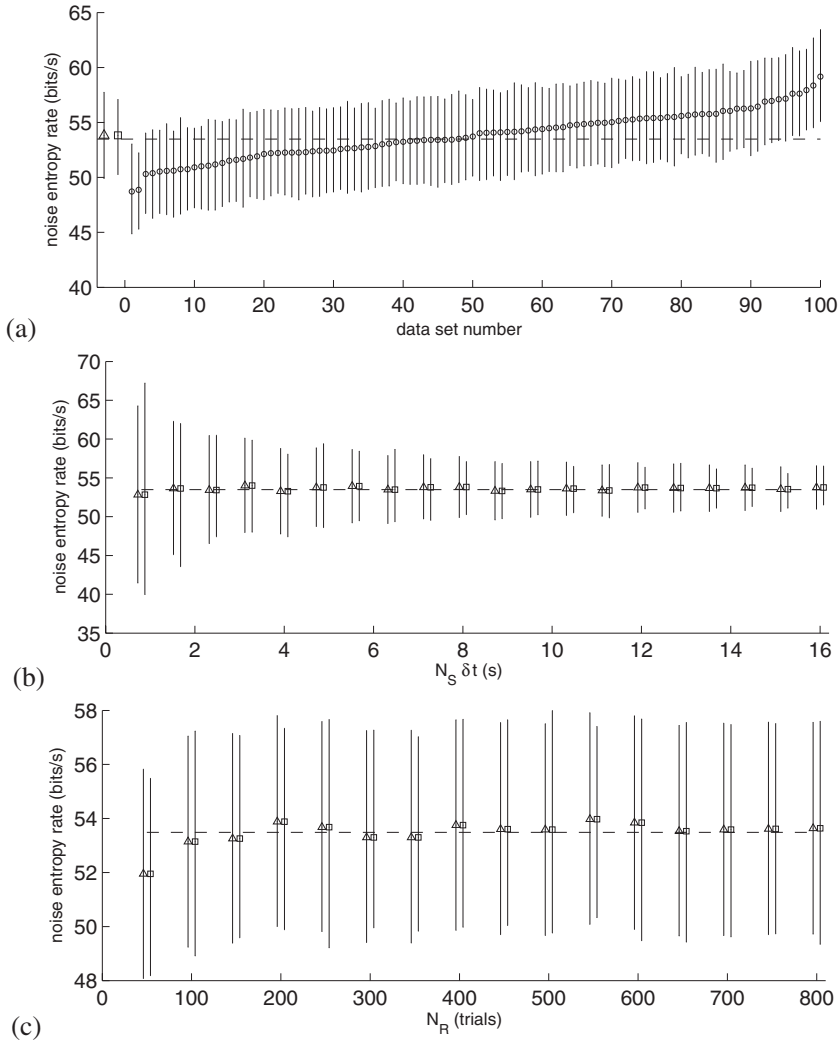


Figure 7: Testing the confidence interval of $\hat{h}_{\text{noise}}$ across variations in N_S and N_R. (a) This figure follows Figure 6. Each circle is an independent data set (raster) with $N_S\delta t = 8$ s and $N_R = 400$ trials. (b) Comparison between average lower and upper confidence interval limits (triangle) and posterior distribution (square) across $N_S\delta t$ with $N_R = 400$ fixed. (c) Comparison between average lower and upper confidence interval limits (triangle) and posterior distribution (square) across N_R with $N_S\delta t = 8$ s fixed. Across varying N_S and N_R, confidence intervals are well calibrated and engulf approximate truth (dashed line, $N_R = 3000$ trials, $N_S\delta t = 32$ s).

and we make use of the stationary bootstrap to account for correlations within $\hat{h}_{\text{noise}}(t)$. In Figure 7c, we repeat the same procedure, this time varying N_R while keeping $N_S \delta t = 8$ s. This is in contrast to Figure 5, where we did not draw new firing rate time series. The reason for this is that previously, the purpose was to demonstrate only the result that bias depends primarily on N_R; the additional variability induced by draws of new time series from the firing rate process would be irrelevant and would obscure the effect we intended to demonstrate. Now, however, the magnitude of the fluctuations and whether our estimation technology can account for them is the central object of interest, and so it is appropriate to account for this source of variability in the data-generating process. In summary, we find that the hybrid Bayesian-frequentist confidence interval is well calibrated for $\hat{h}_{\text{noise}}$ across a range of N_R and N_S in our example.

5 Results

We now bring these tools to bear on neural data from a guinea pig retinal ganglion cell (RGC). These data were recorded extracellularly with a multielectrode array (for experimental details, see Chichilnisky & Kalmar, 2003) and full-field binary monitor flicker at 120 Hz was presented for a long duration ($N\delta t = 1000$ s) and over repeated trials ($N_S \delta t = 9.5$ s, $N_R = 660$ trials) to sample the total and noise entropy rates, respectively. We isolated spike times of individual RGCs using a simple voltage threshold procedure and associated distinct clusters of spike times with individual neurons in a spike waveform feature spikes (Frechette et al., 2005). We present this analysis in full on a single RGC in Figures 8 to 11.

We selected the spike times of one well-isolated ON-type RGC (contamination rate < 0.02; Uzzell & Chichilnisky, 2004), binned at 4 ms, which appeared relatively stationary over the duration of the experiment (see Figure 8). We qualitatively judged stationarity by measuring fluctuations in the firing rate (7.9 ± 0.9 Hz) and the spike-triggered average (data not shown). In spite of this selective screening, it should be emphasized that nonstationarity exists in all recordings to varying degrees, as with most other experimental recordings. This nonstationarity could result from either changes in experimental conditions (e.g., temperature, cell death) or adaptation over longer timescales (Fairhall et al., 2001; Baccus & Meister, 2002). Regardless, all of these variations could result in either an increased noise entropy rate, a lack of convergence in our estimates, or an underestimation of the confidence interval. Although we attempted to mitigate these effects by careful control of experimental conditions, this issue is of potential concern because nonstationarity violates a central assumption in the definition of entropy rate.

5.1 Entropy Rates. Figure 9 examines the convergence of the entropy rate across increasing durations of data. Only a single data set is available;

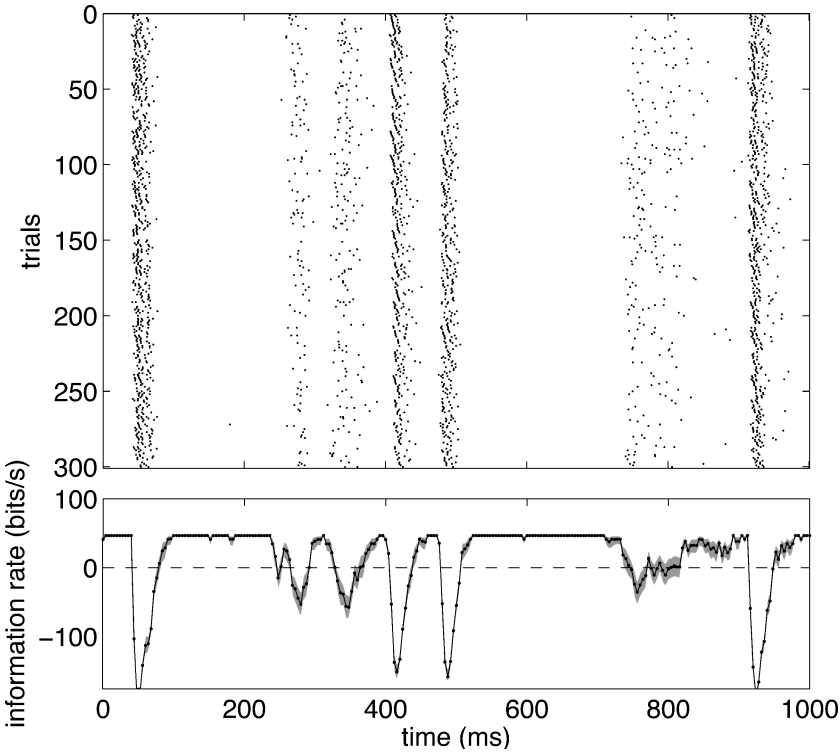


Figure 8: ON-type guinea pig retinal ganglion cell. A small portion of the raster from repeated trials of a stimulus. The conservation of spike times across trials suggests that the neuron responds reliably and precisely to repeated presentations of the stimulus. The lower plot shows the specific information rate of the retinal ganglion cell. The black dot is the estimate, and the gray shadow is the 95% confidence interval. The average mutual information rate is 15.2 bits per second.

thus, a complete examination of convergence and confidence intervals of the estimator across ensembles of data sets is not possible. As a rough approximation, we instead artificially divide our single long data set into K independent, smaller data sets. In Figure 9, this analysis is performed on the total and noise entropy rate across for $K \leq 8$ fractions of the original data set. The final estimate using the complete data set ($K = 1$) is displayed with the horizontal dashed line. Individual subsampled estimates (circles) converge to the final estimate as the number of data (N, N_R respectively) increases. Single-trial 95% confidence intervals engulf the final estimate a vast majority of the time, indicating that the confidence interval provides a

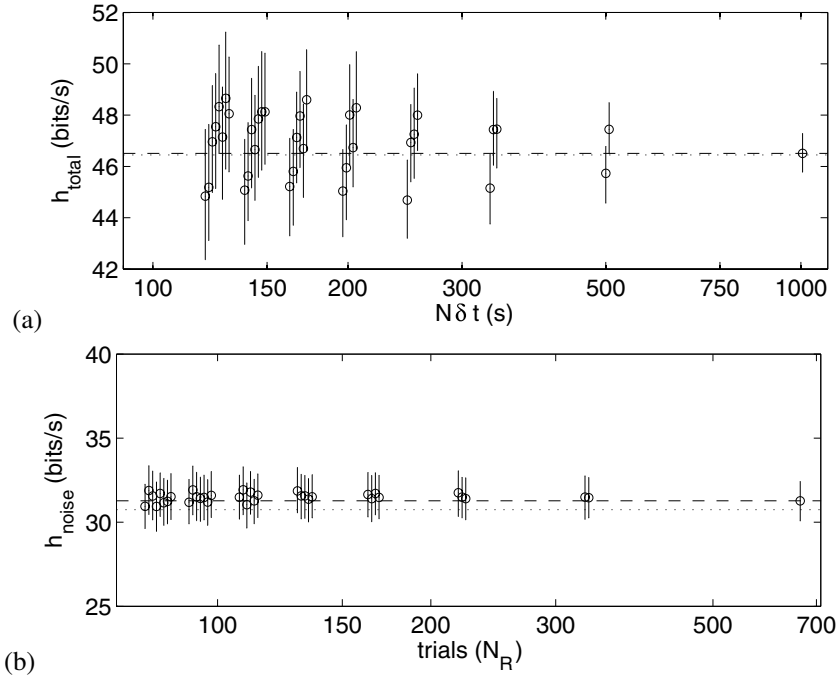


Figure 9: Convergence of entropy rate estimates (circles; error bar is 95% confidence interval). Each group of data points is an estimate from $K \leq 8$ independent data sets subsampled from the complete data set. Horizontal lines are entropy rate estimates from our estimator (dashed) and from the direct method (dotted) (Strong et al., 1998). (a) Total entropy rate estimates across subsamples of N. (b) Noise entropy rate estimates across subsamples of N_R. Note that single-trial confidence intervals from subsampled data sets engulf the final estimate using the complete data set.

reasonable bound on the range of potential estimates assuming more data were available.

For the noise entropy rate (see Figure 9b), the confidence interval is calculated using the stationary bootstrap procedure with the firing rate correction. The reason for selecting this type of confidence interval is explored in Figure 10 by examining estimates from data sets subsampled across N_S. The ergodicity assumption, equation 3.2, provides that each trial is long enough to explore a reasonable range of response space to provide a good estimate of $P(R|S)$ and consequently $\hat{h}_{\text{noise}}$. This assumption should be reflected not only in the estimate (circle) but also the confidence interval, such that different repeated presentations (i.e., subsampled N_S) do not substantially differ in their values. Indeed, for the standard bootstrap procedure (see

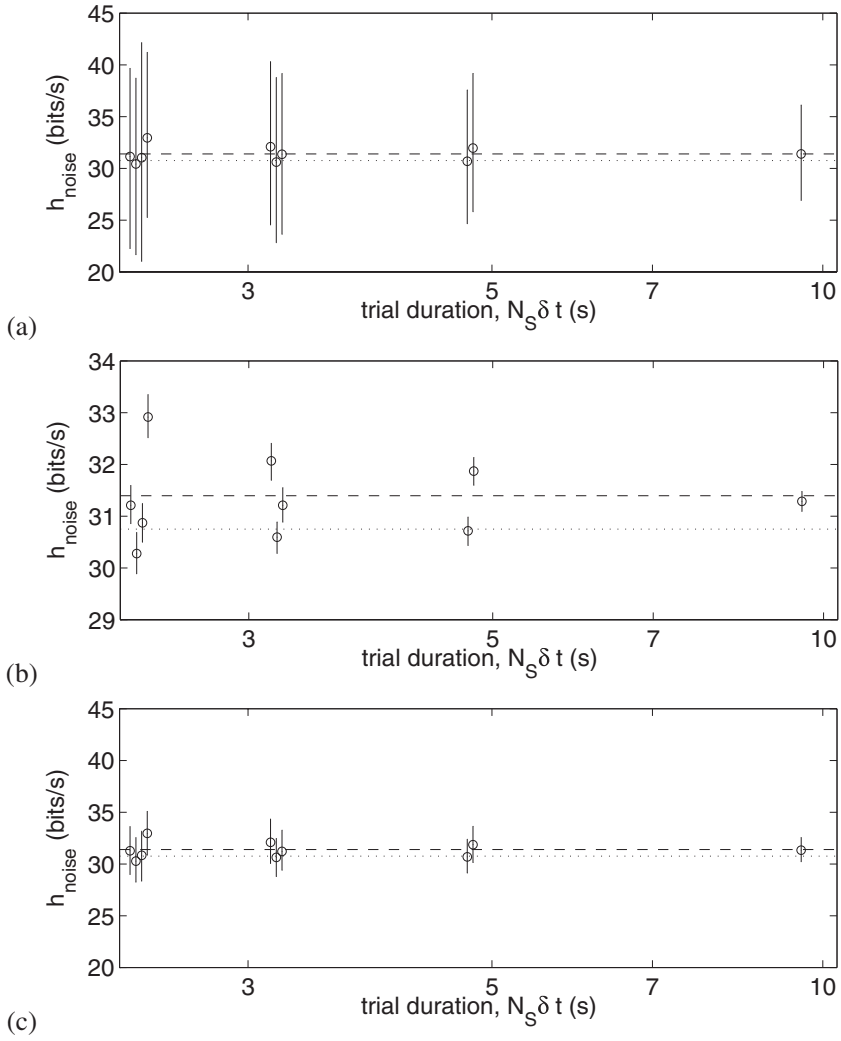


Figure 10: Confidence intervals of noise entropy rate estimation (symbols follow Figure 9). Estimates use three different methods for calculating 95% confidence intervals. Each panel uses varying fractions $K = 1, 2, 3, 4$ of the complete data set to generate independent estimates. (a) Stationary bootstrap procedure. (b) Standard bootstrap procedure for phase times **B**. (c) Stationary bootstrap procedure correcting for spike rate. Note expanded axes in b.

Figure 10b), the confidence interval would indicate otherwise. This suggests that for short stimulus presentations, the ergodicity assumption is incorrect. In contrast, the correlated bootstrap procedure (see Figure 10b) properly accounts for this uncertainty as the error bars indicate that these independent samples are not statistically different. Furthermore, this result holds even when removing systematic fluctuations in the firing rate introduced by the bootstrap procedure (see Figure 10c).

We suggest that several features give encouragement that the entropy rate estimates are fairly well converged. First, diagnostics, based on complexity measures introduced in Kennel et al. (2005) are well converged, suggesting a reasonable first check (data not shown). The convergence of these diagnostics is necessary but not sufficient for a well-converged entropy rate estimate. The simulations in the previous section are the strongest evidence for convergence. A model neuron with similar temporal dynamics, binned at the same temporal resolution, provides well-converged $\hat{h}_{\text{noise}}$ and $\hat{h}_{\text{total}}$ estimates within the total recording time of this neuron.

5.2 Information Rates of a Retinal Ganglion Cell. For this stimulus distribution, we estimate the average information rate $I(S; R)$ to be $15.2 \pm [1.1, 1.1]$ bits per second with 95% confidence.[7] The total entropy rate of the response $h(R)$ is $46.5 \pm [0.8, 0.8]$ bits per second, providing an average response efficiency $\frac{I(S;R)}{h(R)} = 32.8 \pm [2.8, 3.0]$%. The response efficiency characterizes how much of the response capacity is exploited by the neuron. Similar response efficiencies have been reported in neurons across varying bin sizes (Borst & Theunissen, 1999).

The stimulus distribution has an entropy rate $h(S) = 120$ bits per second, because there is an equal chance of a black or white screen (1 bit) selected randomly at 120 Hz. $H(S) > H(R)$ implies that even if the neuron were noise free (i.e., $H(R|S) = 0$), at a resolution of 4 ms, the neuron would not contain a large enough repertoire of spike patterns to losslessly transmit the stimulus.

The stimulus efficiency $\frac{I(S;R)}{h(S)} = 12.7 \pm [0.9, 0.9]$% characterizes how much of the stimulus the spike train transmits. The low stimulus efficiency suggests several possibilities: the neuron failed to transmit all of the information about the stimulus, or the bin size of the spike train is too large to extract all of the information about the stimulus. The absolute refractory period of the neuron sets a rough guide to the relevant timescale for the bin size. Recent work has suggested $I(S; R)$ must be calculated for bin sizes smaller than the refractory period in order to extract all information about a stimulus within a spike train (Strong et al., 1998; Reinagel & Reid, 2000; Liu, Tzonev, Rebrik, & Miller, 2001). However, estimating information rates

[7] Because our error bars can be asymmetric, we abbreviate lower and upper portions of the error bar as the first and second items within the brackets.

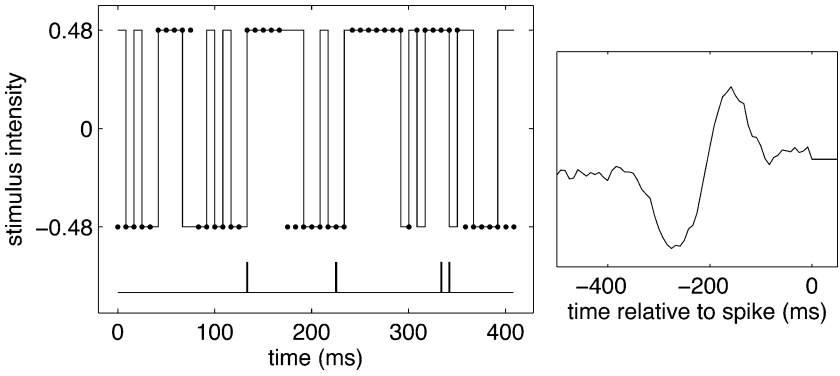


Figure 11: Reconstruction of a binary stimulus. In the left panel, a short segment of the displayed light intensity waveform of the full-field stimulus S (black line) and the corresponding RGC spike train (below). We calculate the reconstructed stimulus $\hat{S}$ (black dots) by convolving the discretized spike train ($\delta t = 4$ ms) with a filter (right panel) and thresholding the resulting value to a high or low stimulus intensity. The filter is the reverse correlation of the spike train with the stimulus divided by the autocorrelation of the response (Rieke et al., 1997). This simple decoding procedure recovers at least 3.3 bits per second, or $21.9 \pm [1.5, 1.7]$% of the available information.

at smaller bin sizes require-larger effective conditioning depths D and thus contains larger data requirements (Kennel et al., 2005). Most important, the response kinetics of ON-type guinea pig RGCs is slow compared to the refresh interval of the stimulus (see Figure 11). Thus, we hypothesize that the stimulus is probably flickering too rapidly for the RGC to respond accurately.

Average information rates mask the contributions of individual stimuli and responses. Recent work has investigated several extensions of information rates to measure how well particular stimuli are encoded (DeWeese & Meister, 1999; Brenner, Strong et al., 2000; Butts, 2003). A trivial extension of our estimator is to estimate the specific information rate $I_{sp}(s)$, or the contribution of individual stimuli. For a small segment of the raster in Figure 8, we plotted $I_{sp}(s)$ for comparison. The gray shadow is the 95% confidence interval and is largely dominated by the uncertainty in $\hat{h}_{\text{noise}}$. Stimuli eliciting silence provide the largest amount of information about the response because the neural response is reliably silent (Butts, 2003). During periods of silence, $h(R|t) = \hat{h}_{\text{noise}}(t) = 0$, implying that the specific information rate plateaus at

$$I_{sp}(s) = h(R) - 0 = h_{\text{total}}.$$

During periods of bursting activity, $\hat{h}_{\text{noise}} > \hat{h}_{\text{total}}$, giving a negative specific information $I_{sp}(s) < 0$. Negative specific information is not only possible but necessary to maintain additivity, a central assumption of information theory (Shannon, 1948; Brillouin, 1956; Fano, 1961; DeWeese & Meister, 1999; Abarbanel, Masuda, Rabinovich, & Tumer, 2001). The sign and distribution of specific information rates reflect the distribution of stimuli selected in the experiment.

5.3 Comparison to a Decoder. A principal feature of the average mutual information rate $I(S; R)$ is that it does not involve any reconstruction of the stimulus nor require any specific model for interpreting a spike train. However, $I(S; R)$ does bound the performance of any hypothesized decoding algorithm $\hat{S} = F(R)$, whether biological or artificial—meaning that it measures the performance an optimal decoder could achieve at a temporal resolution δt. We present a comparison of $I(S; R)$ to a simple decoding procedure based on linear reconstruction to illustrate this point. This decoding procedure reconstructs from the spike train R an estimate of the stimulus waveform

$$\hat{S} = \{\hat{s}_1, \hat{s}_2, \ldots\},$$

where the subscript indexes time. The performance of this artificial decoder can be judged on an absolute scale relative to the available information, $I(S; R)$ (Buracas et al., 1998; Warland et al., 1997; Abarbanel & Talathi, 2006).

Because the original stimulus waveform is binary, we constructed a simple decoder F by convolving the spike train with a filter and thresholding the resulting value to a high or low stimulus value. Figure 11 shows a segment of the resulting reconstruction. We calculated the filter using the first half of the data ($N\delta t = 500$ s); all subsequent calculations used the second half of the data set. The filter (Figure 11, right panel) is the optimal causal linear estimate of the stimulus given the spike train, calculated by reverse-correlating the spike train with the stimulus and dividing by the autocorrelation of the response (Rieke et al., 1997).

Preliminarily we can compare the ability of the decoder F to reconstruct the stimulus to the best possible predictive power dictated by the information rate. We define the error as a binary random variable $E \equiv \delta(S \neq \hat{S})$, where 1 and 0 indicate a correct and incorrect prediction, respectively. Fano's inequality (Cover & Thomas, 1991) guarantees that for a binary stimulus distribution, the probability of an error $P(E)$ implicitly is bounded below by the equation,

$$\begin{aligned} H(E) &\geq \delta t \; h(S|R) \\ &= \delta t \, [h(S) - I(S; R)]. \end{aligned} \tag{5.1}$$

The entropy rate of the stimulus, $h(S)$, is 120 bits per second by experimental design. The difference between the $h(S)$ and $I(S; R)$, the latter estimated from data, gives a lower bound on the entropy rate of the error. Solving equation 5.1 for $P(E)$ with our results gives a lower bound $P(E) \geq 37.8\%$. Empirically, we find our simple decoder failed to predict the stimulus 40.5% of the time.

We also characterize the performance of the decoder in absolute terms by examining the reduction in stimulus space to calculate a lower bound on the information attributed to the decoder,

$$I(S; F(R)) = h(S) - h(S|F(R)).$$

We can overestimate $h(S|F(R))$ by assuming the stimulus estimate $\hat{S}$ is identically and independently distributed (i.i.d.). This assumption ensures that our estimate of $I(S; F(R))$ is a conservative lower bound:

$$\begin{aligned} I(S; F(R)) &= h(S) - \frac{1}{\delta t} H(s_t \mid s_{\text{hist}}, \hat{s}_t, \hat{s}_{\text{hist}}) \\ &= h(S) - \frac{1}{\delta t} H(s_t \mid \hat{s}_t, \hat{s}_{\text{hist}}) \\ &\geq h(S) - \frac{1}{\delta t} H(s_t \mid \hat{s}_t) \\ &= h(S) - \frac{1}{\delta t} H(S|\hat{S}), \end{aligned} \tag{5.2}$$

where we have recognized that S is i.i.d., and hist denotes all previous values in time up to but not including time t. By the data processing inequality, we know that $I(S; F(R)) \leq I(S; R)$ with equality if and only if F captures all of the information contained in R (or is a sufficient statistic of R). For this simple decoding procedure, we recover at least 3.3 bits per second, or $21.9 \pm [1.5, 1.7]$% of the available information. Building a better decoding procedure through dimensional reduction of stimulus space or other biological priors could recover a greater percentage of this available information and decrease the probability of reconstruction error (Bialek et al., 1991; Aguera y Arcas & Fairhall, 2003; Simoncelli, Paninski, Pillow, & Schwartz, 2004; Abarbanel & Talathi, 2006).

6 Discussion

We have discussed how to estimate information rates in neural spike trains with a confidence interval by extending an estimator introduced in a related

work (Kennel et al., 2005).[8] We apply a time series bootstrap procedure to estimate the uncertainty in the noise entropy under the (usually unrecognized) assumption that the presented stimulus is but a single example from some underlying distribution. We combined these ideas to calculate a pointwise specific information rate $I_{sp}(s)$ and average information rate $I(S; R)$ in an experimental data set. We calculated the stimulus and response efficiencies to characterize how well the neuron conveys information about the stimulus and how well the neuron exploits its capacity to send information. Finally, we compared $I(S; R)$ to the performance of a simple decoder based on linear reconstruction.

Testing these ideas in simulation, we developed an empirical means of judging convergence in real spike trains. Importantly, we showed in simulation for noise entropy estimation that bias is dominated by the number of trials N_R, while variance is independently dominated by the duration of each trial N_S. This bias-variance trade-off suggests a systematic strategy for determining experimental parameters for estimating the noise entropy rate. First, determine the minimal number of trials for convergence in simulation for a selected bin size (see Figures 1a and 1b). Second, given a fixed experiment duration, maximize the duration of each stimulus presentation in order to minimize the estimator variance. Ideally, the duration of each presentation should be far larger than the correlation length of the neural respose to ensure ergodicity (see equation 3.2). Finally, on real spike trains with inherent nonstationarity, our estimator appears convergent, and the confidence interval captures the range of reasonable variation assuming the source were stationary. Mitigating the effects of nonstationarity in the estimate is thus the focus of our experiment and potential future analysis.

Our estimator is a methodological advance because it (1) empirically converges quickly to the entropy rate (Kennel et al., 2005), (2) removes heuristic judgements of conditioning depth D (Schurmann & Grassberger, 1996; Strong et al., 1998), and (3) provides confidence intervals about its estimate. The first point is imperative for any estimator because it minimizes data requirements, thereby increasing confidence in its accuracy. The second point is a special case of judging the appropriate model complexity for a neural spike train. This point removes heuristics and subjective assessments often relied on in practice in the application of the direct method (Reinagel & Reid, 2000; Lewen et al., 2001).

These heuristics either ignore the model selection problem or reject an estimate post hoc based on inappropriate convergence behavior. For example, in the application of the direct method (Strong et al., 1998), physical arguments suggest that a linear extrapolation within a selected scaling region should provide a good estimate of the entropy rate provided enough data

[8] Source code as well as a Matlab interface for both Bayesian entropy rate and information rate estimation is available online at http://www.snl.salk.edu/~shlens/infotheory.html.

samples exist. Thus, a common post hoc criterion is to reject this extrapolation if deviations from linearity become too large. One difficulty with this approach (aside from the subjective assessment of a scaling region and a linearity criterion) is that no estimate of the entropy rate is provided if this post hoc criterion is not met. In other words, if a linear extrapolation is poor, then no estimate is provided at all. A potential example of this failure can be observed in the upper-right panel of Figure 2, where deviations from a linear extrapolation would be large. We view this shortcoming as fundamental because it skirts the larger problem of selecting the appropriate model complexity.

Selecting the appropriate probabilistic model complexity for a spike train allows one to generalize or sample the range of plausible entropy rates associated with the spike train, thus fulfilling the third point above. The computation of confidence intervals for information rates is unique to our estimator. A confidence interval provides a means for comparing information rates in a statistically significant manner. The lack of a reliable, heuristic-free estimator with an associated error bar has complicated the application of information theory in empirical data analysis.

Immediate applications of this estimator include several avenues of research where the lack of a reliable estimator has made analysis difficult. This technique can be applied generally to estimate the information of any situation with a symbolic observable and the opportunity to repeat stimuli. We now discuss immediate applications in neuroscience. One important application is to measure the temporal precision of spike times by comparing information rates across high temporal resolutions (Strong et al., 1998; Reinagel & Reid, 2000; Liu et al., 2001). The selection of the optimal bin size (or any discretization; Rapp et al., 1994) to represent a spike train remains an open question, potentially resolved through model selection criteria (Rissanen, 1989). A second avenue is to explore the correlation structure between neurons by comparing information rates across multiple neurons (Puchalla, Schneidman, Harris, & Berry, 2005; Dan, Alonso, Usrey, & Reid, 1998; Gat & Tishby, 1999; Reich et al., 2001; Schneidman et al., 2003). Furthermore, the extension of this estimator to calculate specific information quantities (Bezzi et al., 2002; DeWeese & Meister, 1999; Brenner, Strong et al., 2000; Butts, 2003) suggests the possibility of actively searching through stimulus space for features that elicit high temporal precision in single neurons or unique correlational structure between multiple neurons.

The benefits of this new estimator arise from having a justified theory to select the appropriate model complexity for the estimated probabilistic spiking response $P(R|S)$. Although our model of spiking activity $P(R|S)$ is statistical and not biophysical, this model does highlight how dynamics restrict the production of uncertainty in any neuron and, with further calculation, provides an upper bound to judge the quality of any decoding mechanism—whether experimentally derived or downstream in sensory processing.

Appendix A: The Politis-Romano (PR) Bootstrap

The PR bootstrap is a numerical procedure used to generate correlated bootstrap samples from a time series (Politis & Romano, 1994; Davison et al., 1997). In this appendix $Q(t)$ is a time series discretely sampled at $t = [1, \ldots, t_{\max}]$. The bootstrap random variable $\mathbf{B} = (B_1, B_2, \ldots)$ is a list of time indexes that preserve some of the autocorrelation in $Q(t)$.

The first value B_1 is chosen at random from the integer time indexes. Given an already selected time index B_i, with probability $p = 1/L$ (or with $p = 1$ if $B_i = t_{\max}$), we choose B_{i+1} randomly and uniformly; otherwise, $B_{i+1} = B_i + 1$. This procedure generates a bootstrap draw $\mathbf{B}$ composed of varying-length blocks of successive time indexes. The bootstrap resample of the time series is thus $Q(B_1), Q(B_2), \ldots$ The lengths of these blocks are exponentially distributed with mean length L. L is a free parameter chosen to reflect the duration of the correlation structure in the time series $Q(t)$. In the statistical literature, the PR method is often called the stationary bootstrap because the bootstrap process generating $\mathbf{B}$ is statistically stationary, as opposed to the fixed-L block bootstrap methods previously used, which are only cyclically stationary.

Several algorithms select L automatically given an observed time series (Politis & White, 2004). In practice, a simple but common heuristic is to select L to be a few times the width of the autocorrelation peak of $Q(t)$. In the raster plots, L is calculated as twice the autocorrelation across the phase time of the spike rate $r(t)$ or $\hat{h}_{\text{noise}}(t)$. Empirically, our results do not seem to depend significantly on the specific choice of L within a wide range of reasonable values. We have also used the software in Politis and White (2004) to estimate L, which results in no appreciable change in the results presented here.

Appendix B: Simulation of a Refractory Poisson Neuron

In section 4 we use a single model of a neuron to judge the convergence and quality of our estimate of entropy rates. Our model neuron is an inhomogeneous Poisson process with an absolute and relative refractory period. The goal of this model is solely to provide an approximation of the temporal dynamics of a neuron (but see Berry & Meister, 1998; Keat, Reinagel, Reid, & Meister, 2001; Pillow & Simoncelli, 2003; Paninski, Pillow, & Simoncelli, 2004).

The parameters that must be specified are the relative (and absolute) refractory periods and the firing rate over time. We matched the absolute refractory period of a real neuron (4 ms) by identifying the smallest interspike interval in the autocorrelation of the spike train. We used a sigmoidal recovery function fit to the firing statistics of the spike train to match the relative refractory period of the neuron (Uzzell & Chichilnisky, 2004).

We approximated the firing rate using two separate methods. In the first method, we measured the probability of firing in the PSTH of a cell in response to 385 repetitions of the an identical stimulus 8 seconds in duration. We termed this firing rate time series λ_1. The firing rate λ_1 is effectively a probability of a spike within the bin size δt. As can be seen from a raster plot, this quantity varies significantly with the stimulus history. Because λ_1 has a limited duration, we generated a second, artificial firing rate λ_2 of arbitrary length. We generated λ_2 by using the correlated bootstrap method on λ_1 (see appendix A) to preserve the correlation structure of λ_1. We used λ_2 to test the estimator convergence over long durations of time.

Acknowledgments

This work was supported by NSF IGERT training grant DGE-0333451 and La Jolla Interfaces in the Sciences (J.S.) and a Sloan Research Fellowship (E.J.C.). We thank J. Victor and P. Reinagel for valuable discussions; our reviewers for substantive comments and feedback; and colleagues at the Institute for Nonlinear Science, Santa Cruz Institute for Particle Physics, and the Systems Neurobiology Laboratory (Salk Institute) for tremendous feedback and technical assistance.

References

Abarbanel, H. D. I., Masuda, N., Rabinovich, M. I., & Tumer, E. (2001). Distribution of mutual information. *Physics Letters A*, *281*, 368–373.

Abarbanel, H. D., & Talathi, S. S. (2006). Neural circuitry for recognizing interspike interval sequences. *Physical Review Letters*, *96*, 148104.

Aguera y Arcas, B., & Fairhall, A. (2003). Computation in a single neuron: Hodgkin and Huxley revisited. *Neural Computation*, *15*, 1715–1749.

Amigo, J. M., Szczepanski, J., Wajnryb, E., & Sanchez-Vives, M. (2004). Estimating the entropy rate of spike trains via Lempel-Ziv complexity. *Neural Computation*, *16*, 717–736.

Baccus, S. A., & Meister, M. (2002). Fast and slow contrast adaptation in retinal circuitry. *Neuron*, *36*, 909–919.

Bair, W., & Koch, C. (1996). Temporal precision of spike trains in extrastriate cortex of the behaving macaque monkey. *Neural Computation*, *8*, 1185–1202.

Berry, M. J., & Meister, M. (1998). Refractoriness and neural precision. *Journal of Neuroscience*, *18*, 2200–2211.

Berry, M. J., Warland, D. K., & Meister, M. (1997). The structure and precision of retinal spike trains. *Proceedings of the National Academies of Science*, *94*, 5411–5416.

Bezzi, M., Samengo, I., Leutbeg, S., & Mizmori, S. J. Y. (2002). Measuring information spatial densities. *Neural Computation*, *14*, 405–420.

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

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

Brenner, N., Bialek, W., & de Ruyter van Steveninck, R. (2000). Adaptive rescaling maximizes information transmission. *Neuron*, *26*, 695–702.

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

Brillouin, L. (1956). *Science and information theory*. Orlando, FL: Academic Press.

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

Butts, D. (2003). How much information is associated with a particular stimulus? *Network*, *14*, 177–187.

Chichilnisky, E. J., & Kalmar, R. (2003). Temporal resolution of ensemble visual motion signals in primate retina. *Journal of Neuroscience*, *23*, 6681–6689.

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

Dan, Y., Alonso, J. M., Usrey, W. M., & Reid, R. E. (1998). Coding of visual information by precisely correlated spikes in the lateral geniculate nucleus. *Nature Neuroscience*, *1*, 501–507.

Davison, A., Hinkley, D., Gill, R., Ripley, B., Ross, S., Stein, M., & Williams, D. (1997). *Bootstrap methods and their applications*. Cambridge: Cambridge University Press.

de Ruyter van Steveninck, R., & Bialek, W. (1988). Real-time performance of a movement sensitive neuron in the blowfly visual system: Coding and information transfer in short spike sequences. *Proceedings of the Royal Society of London, Series B*, *234*, 379–414.

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

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

Dimitrov, A. G., Miller, J. P., Gedeon, T., Aldworth, Z., & Parker, A. E. (2003). Analysis of neural coding through quantization with an information-based distortion measure. *Network: Computation in Neural Systems*, *14*, 151–176.

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

Fairhall, A., Lewen, G., Bialek, W., & de Ruyter van Steveninck, R. (2001). Efficiency and ambiguity in an adaptive neural code. *Nature*, *412*, 787–792.

Fano, R. M. (1961). *Transmission of information: A statistical theory of communications*. Cambridge, MA: MIT Press.

Fellous, J. M., Tiesinga, P. H. E., Thomas, P. J., & Sejnowski, T. J. (2004). Discovering spike patterns in neuronal responses. *Journal of Neuroscience*, *24*, 2989–3001.

Frechette, E. S., Sher, A., Grivich, M. I., Petrusca, D., Litke, A. M., & Chichilnisky, E. J. (2005). Fidelity of the ensemble code for visual motion in primate retina. *Journal of Neurophysiology*, *94*, 119–135.

Gamerman, D. (1997). *Markov chain Monte Carlo: Stochastic simulation of Bayesian inference*. New York: CRC Press.

Gat, I., & Tishby, N. (1999). Synergy and redundancy among brain cells of behaving monkeys. In H. S. Kearns, S. Solla, & D. Cohn (Eds.), *Advances in neural information processing systems*, *11* (pp. 465–471). Cambridge, MA: MIT Press.

Gawne, T. J., & Richmond, B. J. (1993). How independent are the messages carried by adjacent inferior temporal cortical neurons? *Journal of Neuroscience*, *13*, 2758–2771.

Gilmore, R., & Lefranc, M. (2002). *The topology of chaos: Alice in stretch and squeeze land*. New York: Wiley.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrica*, *57*, 97.

Hilborn, R. (2000). *Chaos and nonlinear dynamics: An introduction for scientists and engineers*. New York: Oxford University Press.

Keat, J., Reinagel, P., Reid, R. C., & Meister, M. (2001). Predicting every spike: A model for the responses of visual neurons. *Neuron*, *30*, 803–817.

Kennel, M. B., & Mees, A. I. (2002). Context-tree modeling of observed symbolic dynamics. *Physical Review E*, *66*, 056209.

Kennel, M., Shlens, J., Abarbanel, H. D. I., & Chichilnisky, E. J. (2005). Estimating entropy rates with Bayesian confidence intervals. *Neural Computation*, *7*, 1531–1576.

Kontoyiannis, I., Algoet, P., Suhov, Y., & Wyner, A. (1998). Nonparametric entropy estimation for stationary processes and random fields with applications to English text. *IEEE Transactions in Information Theory*, *44*, 1319–1327.

Lempel, A., & Ziv, J. (1976). On the complexity of finite sequences. *IEEE Transactions in Information Theory*, *22*, 75–81.

Lewen, G. D., Bialek, W., & de Ruyter van Steveninck, R. R. (2001). Neural coding of naturalistic motion stimuli. *Network: Computation in Neural Systems*, *12*, 317–329.

Lind, D., & Marcus, B. (1996). *Symbolic dynamics and coding*. Cambridge: Cambridge University Press.

Liu, R. C., Tzonev, S., Rebrik, S., & Miller, K. D. (2001). Variability and information in a neural code of the cat lateral geniculate nucleus. *Journal of Neurophysiology*, *86*, 2789–2806.

London, M., Schreibman, A., Hausser, M., Larkum, M., & Segev, I. (2002). The information efficacy of a synapse. *Nature Neuroscience*, *5*, 332–340.

MacKay, D. (2003). *Information theory, inference and learning algorithms*. Cambridge: Cambridge University Press.

MacKay, D., & McCulloch, W. (1952). The limiting information capacity of a neuronal link. *Bulletin of Mathematical Biophysics*, *14*, 127–135.

Mainen, Z., & Sejnowski, T. (1995). Reliability of spike timing in neocortical neurons. *Science*, *268*, 1503–1506.

Mastronarde, D. N. (1989). Correlated firing of retinal ganglion cells. *Trends in Neuroscience*, *12*, 75–80.

Meister, M., & Berry, M. J. (1999). The neural code of the retina. *Neuron*, *22*, 435–450.

Meister, M., Lagnado, L., & Baylor, D. A. (1995). Concerted signaling by retinal ganglion cells. *Science*, *270*, 1207–1210.

Metropolis, N., Rosenbluth, M., Rosenbluth, A., Teller, M., & Teller, A. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, *21*, 1087.

Miller, G., & Madow, W. (1954). On the maximum likelihood estimate of the Shannon-Wiener measure of information. *Air Force Cambridge Research Center Technical Report*, *75*, 54.

Nemenman, I., Bialek, W., & de Ruyter van Steveninck, R. (2004). Entropy and information in neural spike trains: Progress on the sampling problem. *Physical Review E*, 69.

Nemenman, I., Shafee, F., & Bialek, W. (2002). Entropy and inference, revisited. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processsing systems*, *14*. Cambridge, MA: Cambridge University Press.

Nirenberg, S., & Latham, P. (2003). Decoding neuronal spike trains: How important are correlations? *Proceedings of the National Academies of Science*, *100*, 7348–7353.

Ott, E. (2002). *Chaos in dynamical systems*. Cambridge: Cambridge University Press.

Paninski, L. (2003a). Estmation of entropy and mutual information. *Neural Computation*, *15*, 1191–1254.

Paninski, L. (2003b). Convergence properties of three spike-triggered analysis techinques. *Network: Computation in Neural Systems*, *14*, 437–464.

Paninski, L., Pillow, J., & Simoncelli, E. (2004). Maximum likelihood estimation of stochastic integrate-and-fire neural model. *Neural Computation, 16*, 2533–2561.

Pillow, J., & Simoncelli, E. (2003). Biases in white noise analysis due to non-Poisson spike generation. *Neurocomputing*, *52*, 109–115.

Politis, D. N., & Romano, J. P. (1994). The stationary bootstrap. *J. Amer. Stat. Assoc.*, *89*, 1303–1313.

Politis, D. N., & White, H. (2004). Automatic block-length selection for dependent bootstrap. *Econometric Review*, *23*, 53–70.

Puchalla, J., Schneidman, E., Harris, R., & Berry R. J. II (2005). Redundancy in the population code of the retina. *Neuron*, *46*, 493–504.

Rapp, P. E., Zimmerman, I. D., Vining, E. P., Cohen, N., Albano, A. M., & Jimhez-Montaiio, M. A. (1994). The algorithmic complexity of neural spike trains increases during focal seizures. *Journal of Neuroscience*, *14*, 4731–4739.

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

Reinagel, P., & Reid, R. C. (2000). Temporal coding of visual information in the thalamus. *Journal of Neuroscience*, *20*, 5392–5400.

Reinagel, P., & Reid, R. C. (2002). Precise firing events are conserved across neurons. *Journal of Neuroscience*, *22*, 6837–6841.

Rieke, F., Bodnar, D., & Bialek, W. (1995). Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. *Proceedings of the Royal Society of London B: Biological Sciences*, *262*, 259–265.

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

Rissanen, J. (1989). *Stochastic complexity in statistical inquiry*. Singapore: World Scientific.

Schneidman, E., Bialek, W., & Berry, M. J. (2003). Synergy, redundancy, and independence in population codes. *Journal of Neuroscience*, *23*, 11539–11553.

Schnitzer, M. J., & Meister, M. (2003). Multineuronal firing patterns in the signal from eye to brain. *Neuron*, *37*, 499–511.

Schurmann, T., & Grassberger, P. (1996). Entropy estimation of symbol sequences. *Chaos*, *6*, 414–427.

Shannon, C. (1948). A mathematical theory of communication. *Bell Systems Technology Journal*, *27*, 379–423.

Sharpee, T., Rust, N., & Bialek, W. (2004). Anaylzing neural responses to natural signals: Maximally informative dimensions. *Neural Computation*, *16*, 223–250.

Simoncelli, E. P., & Olhausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, *24*, 1193–1215.

Simoncelli, E. P., Paninski, L., Pillow, J., & Schwartz, O. (2004). Characterization of neural responses with stochastic stimuli. In M. Gazzaniga (Ed.), *The cognitive neurosciences* (3rd ed.). Cambridge, MA: MIT Press.

Stanley, G. B., Li, F. F., & Dan, Y. (1999). Reconstruction of natural scenes from ensemble responses in the lateral geniculate nucleus. *Journal of Neuroscience*, *19*, 8036–8042.

Strong, S. P., Koberle, R., de Ruyter van Steveninck, R., & Bialek, W. (1998). Entropy and information in neural spike trains. *Physical Review Letters*, *80*, 197–200.

Theunissen, F., & Miller, J. (1995). Temporal encoding in the nervous system: A rigorous definition. *Journal of Computational Neuroscience*, *2*, 149–162.

Treves, A., & Panzeri, S. (1995). The upward bias in measures of information derived from limited data samples. *Neural Computation*, *7*, 399–407.

Usrey, W. M., & Reid, R. C. (1999). Synchronous activity in the visual system. *Annual Reviews in Physiology*, *61*, 435–456.

Uzzell, V. J., & Chichilnisky, E. J. (2004). Precision of spike trains in primate retinal ganglion cells. *Journal of Neurophysiology*, *92*, 780–789.

Victor, J. D. (2002). Binless strategies for estimation of information from neural data. *Physical Review E*, *66*, 051903.

Warland, D., Reinagel, P., & Meister, M. (1997). Decoding visual information from a population of retinal ganglion cells. *Journal of Neurophysiology*, *78*, 2336–2350.

Willems, F. M. J., Shtarkov, Y. M., & Tjalkens, T. (1995). The context tree weighting method: Basic properties. *IEEE Transactions in Information Theory*, *IT-41*, 653–664.

Wolpert, D., & Wolf, D. (1995). Estimating functions of probability distributions from a finite set of samples. *Physical Review E*, *52*, 6841–6854.

Received April 12, 2006; accepted October 4, 2006.