Entropy and information in neural spike trains: Progress on the sampling problem

Ilya Nemenman,[1,*] William Bialek,[2,†] and Rob de Ruyter van Steveninck[3,‡]
[1]*Kavli Institute for Theoretical Physics, University of California, Santa Barbara, California 93106, USA*
[2]*Department of Physics and the Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey 08544, USA*
[3]*Department of Molecular Biology, Princeton University, Princeton, New Jersey 08544, USA*
(Received 11 November 2003; published 24 May 2004)

The major problem in information theoretic analysis of neural responses and other biological data is the reliable estimation of entropy-like quantities from small samples. We apply a recently introduced Bayesian entropy estimator to synthetic data inspired by experiments, and to real experimental spike trains. The estimator performs admirably even very deep in the undersampled regime, where other techniques fail. This opens new possibilities for the information theoretic analysis of experiments, and may be of general interest as an example of learning from limited data.

DOI: 10.1103/PhysRevE.69.056111 PACS number(s): 02.50.Tt, 89.70.+c, 87.19.La, 87.80.Tq

I. INTRODUCTION

There has been considerable progress in using information theoretic methods to sharpen and to answer many questions about the structure of the neural code [1–8]. Where classical experimental approaches have focused on mean responses of neurons to relatively simple stimuli, information theoretic methods have the power to quantify the responses to arbitrarily complex and even fully natural stimuli [9,10], taking account of both the mean response and its variability in a rigorous way, independent of detailed modeling assumptions. Measurements of entropy and information in spike trains also allow us to test directly the hypothesis that the neural code adapts to the distribution of sensory inputs, optimizing the rate or efficiency of information transmission [11–15].

A problem with such measurements is that entropy and information depend explicitly on the full distribution of neural responses, just a limited sample of which is provided by experiments. In particular, we need to know the distribution of responses to each stimulus in our ensemble, and the number of samples from this distribution is limited by the number of times the full set of stimuli can be repeated. For natural stimuli with long correlation times the time required to present a useful "full set of stimuli" is long, limiting the number of independent samples we can obtain from stable neural recordings. Furthermore, natural stimuli generate neural responses of high timing precision, and thus the space of meaningful responses itself is very large [3,10,16,17]. These factors make the sampling problem more serious as we move to more interesting and natural stimuli.

A natural response to this problem is to give up the generality of a completely model independent information theoretic approach. Some explicit help from models is required to regularize learning of the underlying probability distributions from the experiments. The question is if we can keep the generality of our analysis by introducing the gentlest of regularizations for the abstract learning problem, or if we need stronger assumptions about the structure of the neural code itself (for example, introducing a metric on the space of responses [18,19]).

A classical problem suggests that we may succeed even with very weak assumptions. Remember that one needs to have only $N\sim 23$ people in a room before any two of them are reasonably likely to share the same birthday. This is much less than $K=365$, the number of possible birthdays. Turning this around, we can estimate the number of possible birthdays by polling N people and counting how often we find coincidences. Once N is large enough to have observed a few of those, we can get a pretty good estimate of K. This will happen with a significant probability for $N\sim\sqrt{K}\ll K$.

The idea of estimating entropy by counting coincidences was proposed long ago by Ma [20] for physical systems in the microcanonical ensemble where distributions should be uniform at fixed energy. Clearly, if we could generalize the Ma idea to arbitrary distributions, then we would be able to explore a much wider variety of questions about information in the neural code. Here we argue that a simple and abstract Bayesian prior, introduced in Ref. [21], comes close to the objective.

It is well known that one needs $N\sim K$ to estimate entropy universally with small additive or multiplicative errors [22,23]. Thus the main question is: does a particular method work well only for (possibly irrelevant) abstract model problems, or can it also be trusted for natural data? Hence our goal is neither to search for potential theoretical limitations of the approach (these must exist and have been found), nor to analyze the neural code (this will be left for the future). Instead we aim at convincingly showing that the method of Ref. [21] can generate reliable estimates of entropy well into a classically undersampled regime for an experimentally relevant case of neurophysiological recordings.

II. AN ESTIMATION STRATEGY

Consider the problem of estimating the entropy S of a probability distribution $\{p_i\}, S=-\Sigma_{i=1}^{K} p_i \log_2 p_i$, where the in-

*Electronic address: nemenman@kitp.ucsb.edu
†Electronic address: wbialek@princeton.edu
‡Present address: Department of Physics, Indiana University, 727 E. Third St., Bloomington, Indiana 47405. Electronic address: deruyter@indiana.edu

dex i runs over K possibilities (e.g., K possible neural responses). In an experiment we observe that in N examples each possibility i occurred n_i times. If $N \gg K$, we approximate the probabilities by frequencies, $p_i \approx f_i \equiv n_i/N$, and construct a naive estimate of the entropy,

$$S_{\text{naive}} = -\sum_{i=1}^{K} f_i \log_2 f_i. \tag{1}$$

This is also a maximum likelihood estimator, since the maximum likelihood estimate of the probabilities is given by the frequencies. Thus we will replace S_{naive} by S^{ML} in what follows.

It is well know that S^{ML} underestimates the entropy (cf. Ref. [22]). With good sampling ($N \gg K$), classical arguments due to Miller [24] show that the ML estimate should be corrected by a universal term $(K-1)/2N$, and several groups have used this correction in the analysis of neural data. In practice, many bins may have truly zero probability (for example, as a result of refractoriness; see below), and the samples from the distribution might not be completely independent. Then S^{ML} still deviates from the correct answer by a term $\propto 1/N$, but the coefficient is no longer known *a priori*. Under these conditions one can heuristically verify and extrapolate the $1/N$ behavior from subsets of the available data [4]. Alternatively, still agreeing on the $1/N$ correction, one can calculate its coefficient (interpretable as an effective number of bins K^*) for some classes of distributions [25–27]. All of these approaches, however, work only when the sampling errors are in some sense a small perturbation.

If we want to make progress outside of the asymptotically large N regime we need an estimator that does not have a perturbative expansion in $1/N$ with S_{ML} as the zeroth order term. The estimator of Ref. [21] has just this property. Recall that S_{ML} is a limiting case of Bayesian estimation with Dirichlet priors. Formally, we consider that the probability distributions $\mathbf{p} \equiv \{p_i\}$ are themselves drawn from a distribution $\mathcal{P}_\beta(\mathbf{p})$ of the form

$$\mathcal{P}_\beta(\mathbf{p}) = \frac{1}{Z(\beta;K)} \left[\prod_{i=1}^{K} p_i^{(\beta-1)} \right] \delta\left(\sum_{i=1}^{K} p_i - 1 \right), \tag{2}$$

where the delta function enforces normalization of distributions $\mathbf{p}$ and the partition function $Z(\beta;K)$ normalizes the prior $\mathcal{P}_\beta(\mathbf{p})$. Maximum likelihood estimation is Bayesian estimation with this prior in the limit $\beta \to 0$, while the natural "uniform" prior is $\beta = 1$. The key observation of Ref. [21] is that while these priors are quite smooth on the space of $\mathbf{p}$, the distributions drawn at random from $\mathcal{P}_\beta$ all have very similar entropies, with a variance that vanishes as K becomes large. Fundamentally, this is the origin of the sample size dependent bias in entropy estimation, and one might thus hope to correct the bias at its source. The goal then is to construct a prior on the space of probability distributions which generates a nearly uniform distribution of entropies. Because the entropy of distributions chosen from $\mathcal{P}_\beta$ is sharply defined *and* monotonically dependent on the parameter β, we can come close to this goal by an average over β,

$$\mathcal{P}_{\text{NSB}}(\mathbf{p}) \propto \int d\beta \frac{d\bar{S}(\beta;K)}{d\beta} \mathcal{P}_\beta(\mathbf{p}). \tag{3}$$

Here $\bar{S}(\beta;K)$ is the average entropy of distributions chosen from $\mathcal{P}_\beta$ [21,28],

$$\bar{S}(\beta;K) \equiv \xi = \psi_0(K\beta + 1) - \psi_0(\beta + 1), \tag{4}$$

where $\psi_m(x) = (d/dx)^{m+1} \log_2 \Gamma(x)$ are the polygamma functions.

Given this prior, we proceed in standard Bayesian fashion. The probability of observing the data $\mathbf{n} \equiv \{n_i\}$ given the distribution $\mathbf{p}$ is

$$P(\mathbf{n}|\mathbf{p}) \propto \prod_{i=1}^{K} p_i^{n_i}, \tag{5}$$

and then

$$P(\mathbf{p}|\mathbf{n}) = P(\mathbf{n}|\mathbf{p}) \mathcal{P}_{\text{NSB}}(\mathbf{p}) \cdot \frac{1}{P(\mathbf{n})}, \tag{6}$$

$$P(\mathbf{n}) = \int d\mathbf{p}\, P(\mathbf{n}|\mathbf{p}) \mathcal{P}_{\text{NSB}}(\mathbf{p}), \tag{7}$$

$$\left(S^{\text{NSB}}\right)^m = \int d\mathbf{p} \left(-\sum_{i=1}^{K} p_i \log_2 p_i \right)^m P(\mathbf{p}|\mathbf{n}). \tag{8}$$

Here we need to calculate the first two posterior moments of the entropy, $m = 1, 2$, in order to have an access to the entropy estimate and to its variance as well.

The Dirichlet priors allow all the (K dimensional) integrals over $\mathbf{p}$ to be done analytically, so that the computation of S^{NSB} and of its posterior error reduces to just three numerical one-dimensional integrals:

$$\left(S^{\text{NSB}}\right)^m = \frac{\int d\xi\, \rho(\xi,\mathbf{n}) S_\beta^m(\mathbf{n})}{\int d\xi\, \rho(\xi,\mathbf{n})}, \tag{9}$$

where

$$\rho(\xi,\mathbf{n}) = \frac{\Gamma[K\beta(\xi)]}{\Gamma[N + K\beta(\xi)]} \prod_{i=1}^{K} \frac{\Gamma[n_i + \beta(\xi)]}{\Gamma[\beta(\xi)]}, \tag{10}$$

where the one-to-one relation between β and ξ is given by Eq. (4), and $S_\beta^m(\mathbf{n})$ is the expectation value of the mth entropy moment at fixed β; the exact expression for $m = 1, 2$ is given in Ref. [28].

Details of the NSB method can be found in Refs. [21,29], and the source code of the implementations in either Octave/C++ or plain C++ is available from the authors. We draw attention to several points.

First, since the analysis is Bayesian, we obtain not only S^{NSB} but also its *a posteriori* standard deviation, δS^{NSB}—an error bar on our estimate, see Eq. (9).

Second, for $N\to\infty$ and $N/K\to 0$ the estimator admits asymptotic analysis. The important parameter is the number of coincidences $\Delta=N-K_1$, where K_1 is the number of bins with nonzero counts. If $\Delta/N\to\text{const}<1$ (many coincidences), then the standard saddle point evaluation of the integrals in Eq. (4) is possible. Interestingly, the second derivative at the saddle is $(\ln^2 2)\Delta$ to the leading order in Δ/N. The second asymptotic can be obtained for $\Delta\sim O(N^0)$ (few coincidences). Then

$$S^{\mathrm{NSB}}\approx\frac{C_\gamma}{\ln 2}-1+2\log_2 N-\psi_0(\Delta), \tag{11}$$

$$\delta S^{\mathrm{NSB}}\approx\sqrt{\psi_1(\Delta)}, \tag{12}$$

where C_γ is the Euler's constant. This is particularly interesting since S^{NSB} happens to have a finite limit for $K\to\infty$, thus allowing entropy estimation even for infinite (or unknown) cardinalities.

Third, both of the above asymptotics show that the estimation procedure relies on Δ to make its estimates; this is in the spirit of Ref. [20].

Finally, S^{NSB} is unbiased if the distribution being learned is typical in $\mathcal{P}_\beta(\mathbf{p})$ for some β, that is, its rank ordered (Zipf) plot is of the form

$$q_i\approx 1-\left[\frac{\beta B(\beta,K\beta-\beta)(K-1)\,i}{K}\right]^{1/(K\beta-\beta)}, \tag{13}$$

$$q_i\approx\left[\frac{\beta B(\beta,K\beta-\beta)(K-i+1)}{K}\right]^{1/\beta}, \tag{14}$$

for $i/K\to 0$ and $i/K\to 1$, respectively. If the Zipf plot has tails that are too short (too long), then the estimator should over (under) estimate. While underestimation may be severe (though always strictly smaller than that for S^{ML}), overestimation is very mild, if present at all, in the most interesting regime $1\ll\Delta\ll N$. S^{NSB} is also unbiased for distributions that are typical in some weighted combinations of $\mathcal{P}_\beta$ for different β's, in particular in $\mathcal{P}_{\mathrm{NSB}}$ itself. However, the typical Zipf plots in this case are more complicated and will be detailed elsewhere.

Before proceeding it is worth asking what we hope to accomplish. Any reasonable estimator will converge to the right answer in the limit of large N. In particular, this is true for S^{NSB}, which is a *consistent* Bayesian estimator [35]. The central problem of entropy estimation is systematic bias, which will cause us to (perhaps significantly) under- or overestimate the information content of spike trains or the efficiency of the neural code. The bias, which vanishes for $N\to\infty$, will manifest itself as a systematic drift in plots of the estimated value versus the sample size. A successful estimator would remove this bias as much as possible. Ideally we thus hope to see an estimate which for all values of N is within its error bars from the correct answer. As N increases the error bars should narrow, with relatively little variation of the (mean) estimate itself. When data are such that no reliable estimation is possible, the estimator should remain uncertain, that is, the posterior variance should be large. The main purpose of this paper is to show that the NSB procedure applied to natural and nature–inspired synthetic signals comes close to this ideal over a wide range of $N\ll K$, and even $N\ll 2^S$. The procedure thus is a viable tool for experimental analysis.

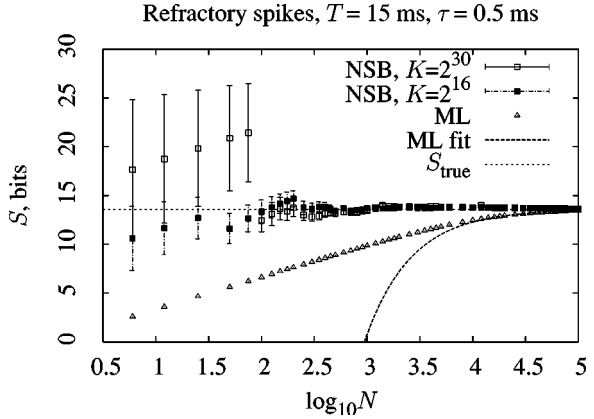

FIG. 1. Entropy estimation for a model problem. Notice that the estimator reaches the true value within the error bars as soon as $N^2\sim 2^S$, at which point coincidences start to occur with high probability. Slight overestimation for $N>10^3$ is expected (see text) since this distribution is atypical in $\mathcal{P}_{\mathrm{NSB}}$.

III. A MODEL PROBLEM

It is important to test our techniques on a problem which captures some aspects of real world data yet is sufficiently well defined that we know the correct answer. We constructed synthetic spike trains where intervals between successive spikes were independent and chosen from an exponential distribution with a dead time or refractory period of $g=1.8$ ms; the mean spike rate was $r=0.26$ spikes/ms. This corresponds to a rate of $r_0=r/(1-rg)=0.49$ spikes/ms for the part of the signal where spiking is not prohibited by refractoriness. These parameters are typical of the high spike rate, noisy regions of the experiment discussed below, which provide the greatest challenge for entropy estimation.

Following the scheme outlined in Ref. [4], we examine the spike train in windows of duration $T=15$ ms and discretize the response with a time resolution $\tau=0.5$ ms. Because of the refractory period each bin of size τ can contain at most one spike, and hence the neural response is a binary word with $T/\tau=30$ letters. The space of responses has $K=2^{30}\approx 10^9$ possibilities. Of course, most of these have probability exactly zero because of refractoriness, and the number of possible responses consistent with this constraint is bounded by $\sim 2^{16}\approx 10^5$. An approximation to the entropy of this distribution is given by an appropriate correction to Eq. (3.21) of Ref. [9], the entropy of a nonrefractory Poisson process:

$$S=\frac{rT}{\ln 2}\left[-\ln(1-e^{-r_0\tau})+\frac{r_0\tau\,e^{-r_0\tau}}{1-e^{-r_0\tau}}\right]=13.57\ \text{bits}. \tag{15}$$

In Fig. 1 we show the results of entropy estimation for this model problem. As expected, the naive estimate S^{ML}

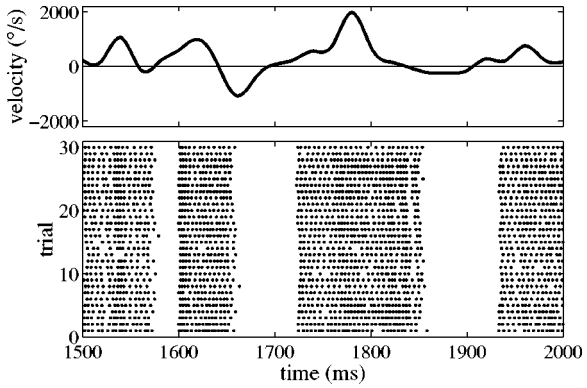

FIG. 2. Data from a fly motion sensitive neuron in a natural stimulus setting. Top: a 500 ms section of a 10 s angular velocity trace that was repeated 196 times. Bottom: raster plot showing the response to 30 consecutive trials; each dot marks the occurrence of a spike.

reaches its asymptotic behavior only when $N > 2^S$, thus the $1/N$ extrapolation becomes successful at $N \sim 10^4$ (the "ML fit" line on the plot). In contrast, we see that S^{NSB} gives the right answer within errors at $N \sim 100$. We can improve convergence by providing the estimator with the "hint" that the number of possible responses K is much smaller than the upper limit of 2^{30}, but even without this hint we have excellent entropy estimates already at $N \sim (2^S)^{1/2}$. This is in accord with expectations from Ma's analysis of (microcanonical) entropy estimation [20]. However, here we achieve these results for a nonuniform distribution.

IV. ANALYZING REAL DATA

For a test on real neurophysiological data, we use recordings from a wide field motion sensitive neuron (H1) in the visual system of the blowfly *Calliphora vicina*. While action potentials from H1 were recorded, the fly rotated on a stepper motor outside among the bushes, with time dependent angular velocity representative of natural flight. Figure 2 presents a sample of raw data from such an experiment (see Ref. [10] for details).

Following Ref. [4], the information content of a spike train is the difference between its total entropy and the entropy of neural responses to repeated presentations of the same stimulus [36]. The latter is substantially more difficult to estimate. It is called the noise entropy S_n since it measures response variations that are uncorrelated with the sensory input. The noise in neurons depends on the stimulus itself—there are, for example, stimuli which generate with certainty zero spikes in a given window of time—and so we write $S_{n|t}$ to mark the dependence on the time t at which we take a slice through the raster of responses. In this experiment the full stimulus was repeated 196 times, which actually is a relatively large number by the standards of neurophysiology. The fly makes behavioral decisions based on $\sim 10-30$ ms windows of its visual input [30], and under natural conditions the time resolution of the neural responses is of order 1 ms or even less [10], so that a meaningful analysis of neural responses must deal with binary words of length 10–30 or more. Refractoriness limits the number of these words which can occur with nonzero probability (as in our model problem), but nonetheless we easily reach the limit where the number of samples is substantially smaller than the number of possible responses.

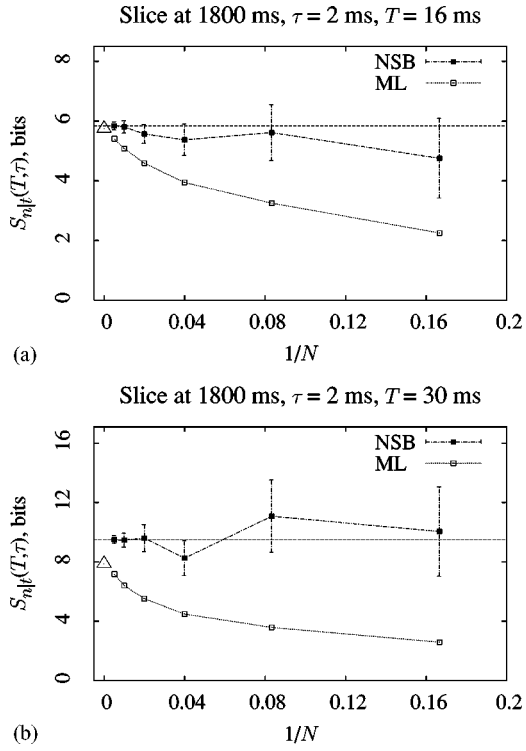

FIG. 3. Slice entropy vs sample size. Dashed line on both plots is drawn at the value of $S^{\mathrm{NSB}}|_{N=N_{\max}}$ to show that the estimator is stable within its error bars even for very low N. Triangle corresponds to the value of S^{ML} extrapolated to $N \to \infty$ from the four largest values of N. First and second panels show examples of word lengths for which S_{ML} can or cannot be reliably extrapolated. S^{NSB} is stable in both cases, shows no N dependent drift, and agrees with S^{ML} where the latter is reliable.

Let us start by looking at a single moment in time, $t = 1800$ ms from the start of the repeated stimulus, as in Fig. 2. If we consider a window of duration $T = 16$ ms at time resolution $\tau = 2$ ms [37], we obtain the entropy estimates shown in the first panel of Fig. 3. Notice that in this case we actually have a total number of samples which is comparable to or larger than $2^{S_{n|t}}$, and so the maximum likelihood estimate of the entropy is converging with the expected $1/N$ behavior. The NSB estimate agrees with this extrapolation. The crucial result is that the NSB estimate is correct within error bars across the whole range of N; there is a slight variation in the mean estimate, but the main effect as we add samples is that the error bars narrow around the correct answer. In this case our estimation procedure has removed essentially all of the sample size dependent bias.

As we open our window to $T = 30$ ms, the number of possible responses (even considering refractoriness) is vastly larger than the number of samples. As we see in the second panel of Fig. 3, any attempt to extrapolate the ML estimate of entropy now requires some wishful thinking. Nonetheless, in parallel with our results for the model problem, we find that

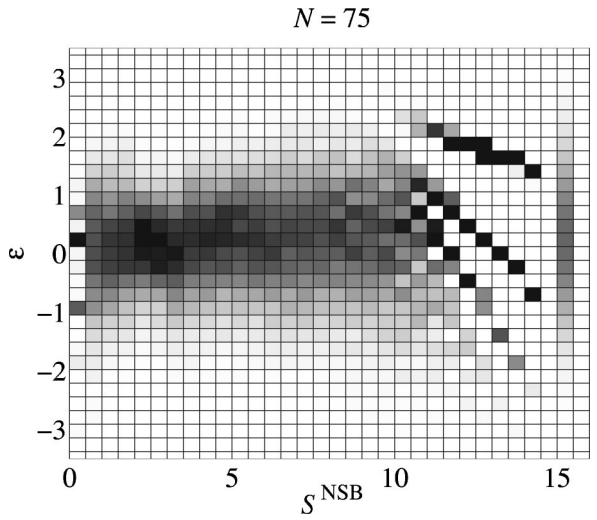

FIG. 4. Distribution of the normalized entropy error conditional on $S^{\mathrm{NSB}}(N_{\mathrm{max}})$ for $N=75$ and $\tau=0.75$ ms. Darker patches correspond to higher probability. The band in the right part of the plot is the normal distribution around zero with the standard deviation of 1 (the standard deviation of plotted conditional distributions averaged over S^{NSB} is about 0.7, which indicates a non-Gaussian form of the posterior for small number of coincidences [29]). For values of S^{NSB} up to about 12 bits the estimator performs remarkably well. For yet larger entropies, where the number of coincidence is just a few, the discrete nature of the estimated values is evident, and this puts a bound on reliability of S^{NSB}.

the NSB estimate is stable within error bars across the full range of available N.

For small T we can compare the results of our Bayesian estimation with an extrapolation of the ML estimate; each moment in time relative to the repeated stimulus provides an example. We have found that the results in the first panel of Fig. 3 are typical: in the regime where extrapolation of the ML estimator is reliable, our estimator agrees within error bars over a broad range of sample sizes. More precisely, if we take the extrapolated ML estimate as the correct answer, and measure the deviation of S^{NSB} from this answer in units of the predicted error bar, we find that the mean square value of this normalized error is of order 1. This is as expected if our estimation errors are random rather than systematic.

For larger T we do not have a calibration against the (extrapolated) S^{ML}, but we can still ask if the estimator is stable, within error bars, over a wide range of N. To check this stability we treat the value of S^{NSB} at $N=N_{\mathrm{max}}=196$ as our best guess for the entropy and compute the normalized deviation of the estimates at smaller values of N from this guess, $\varepsilon=[S^{\mathrm{NSB}}(N)-S^{\mathrm{NSB}}(N_{\mathrm{max}})]/\delta S^{\mathrm{NSB}}(N)$. Again, each moment in time is an example. Figure 4 shows the distribution of these normalized deviations conditional on the entropy estimate with $N=75$; this analysis is done for $\tau=0.75$ ms, with T in the range between 1.5 and 22.5 ms. Since the different time slices span a range of entropies, over some range we have $N>2^S$, and in this regime the entropy estimate must be accurate (as in the analysis of small T above). Throughout this range, the normalized deviations fall in a narrow band with mean close to zero and a variance of order 1, as expected if the only variations with the sample size were random. Remarkably this pattern continues for larger entropies, $S>\log_2 N=6.2$ bits, demonstrating that our estimator is stable even deep into the undersampled regime. This is consistent with the results obtained in our model problem, but here we find the same answer for the real data.

Note that Fig. 4 illustrates results with N less than one-half the total number of samples, so we really are testing for stability over a large range in N. This emphasizes that our estimation procedure moves smoothly from the well sampled into the undersampled regime without accumulating any clear signs of systematic error. The procedure collapses only when the entropy is so large that the probability of observing the same response more than once (a coincidence) becomes negligible.

V. DISCUSSION

The estimator we have explored here is constructed from a prior that has a nearly uniform distribution of entropies. It is plausible that such a uniform prior would largely remove the sample size dependent bias in entropy estimation, but it is crucial to test this experimentally. In particular, there are infinitely many priors which are approximately (and even exactly) uniform in entropy, and it is not clear which of them will allow successful estimation in real world problems. We have found that the NSB prior almost completely removed the bias in the model problem (Fig. 1). Further, for real data in a regime where undersampling can be beaten down by data the bias is removed to yield agreement with the extrapolated ML estimator even at very small sample sizes (Fig. 3, first panel). Finally and most crucially, the NSB estimation procedure continues to perform smoothly and stably past the nominal sampling limit of $N\sim 2^S$, all the way to the Ma cutoff $N^2\sim 2^S$ (Fig. 4). This opens the opportunity for rigorous analysis of entropy and information in spike trains under a much wider set of experimental conditions.

ACKNOWLEDGMENTS

We thank J. Miller for important discussions, G. D. Lewen for his help with the experiments, which were supported by the NEC Research Institute, and the organizers of the NIC'03 workshop for providing a venue for a preliminary presentation of this work. I.N. was supported by NSF Grant No. PHY99-07949 to the Kavli Institute for Theoretical Physics. I.N. is also very thankful to the developers of the following Open Source software: GNU Emacs, GNU Octave, GNUplot, and teTEX.

[1] W. Bialek, F. Rieke, R. R. de Ruyter van Steveninck, and D. Warland, Science **252**, 1854 (1991).
[2] E. Theunissen and J. P. Miller, J. Neurophysiol. **66**, 1690 (1991).
[3] M. J. Berry, D. K. Warland, and M. Meister, Proc. Natl. Acad. Sci. U.S.A. **94**, 5411 (1997).
[4] S. P. Strong, R. Koberle, R. R. de Ruyter van Steveninck, and W. Bialek, Phys. Rev. Lett. **80**, 197 (1998).
[5] A. Borst and F. E. Theunissen, Nat. Neurosci. **2**, 947 (1999).
[6] N. Brenner, S. P. Strong, R. Koberle, W. Bialek, and R. R. de Ruyter van Steveninck, Neural Comput. **12**, 1531 (2000).
[7] P. Reinagel and R. C. Reid, J. Neurosci. **20**, 5392 (2000).
[8] D. S. Reich, F. Mechler, and J. D. Victor, J. Neurophysiol. **85**, 1039 (2001).
[9] F. Rieke, D. Warland, R. de Ruyter van Steveninck, and W. Bialek, *Spikes: Exploring the Neural Code* (MIT Press, Cambridge, MA, 1997).
[10] G. D. Lewen, W. Bialek, and R. R. de Ruyter van Steveninck, Network Comput. Neural Syst. **12**, 312 (2001).
[11] H. B. Barlow, in *Proceedings of the Symposium on the Mechanization of Thought Processes*, edited by D. V. Blake and A. M. Uttley (H. M. Stationery Office, London, 1959), Vol. 2, pp. 537–574.
[12] H. B. Barlow, in *Sensory Communication*, edited by W. Rosenblith (MIT Press, Cambridge, MA, 1961), pp. 217–234.
[13] S. B. Laughlin, Z. Naturforsch. C **36c**, 910 (1981).
[14] N. Brenner, W. Bialek, and R. R. de Ruyter van Steveninck, Neuron **26**, 695 (2000).
[15] A. L. Fairhall, G. D. Lewen, W. Bialek, and R. R. de Ruyter van Steveninck, Nature (London) **412**, 787 (2001).
[16] Z. F. Mainen and T. J. Sejnowski, Science **268**, 1503 (1995).
[17] R. R. de Ruyter van Steveninck, G. D. Lewen, S. P. Strong, R. Koberle, and W. Bialek, Science **275**, 1805 (1997).
[18] J. D. Victor, Phys. Rev. E **66**, 051903 (2002).
[19] J. D. Victor and K. Purpura, Network Comput. Neural Syst. **8**, 127 (1997).
[20] S. Ma, J. Stat. Phys. **26**, 221 (1981).
[21] I. Nemenman, F. Shafee, and W. Bialek, in *Advances in Neural Information Processing Systems 14*, edited by T. G. Dietterich, S. Becker, and Z. Ghahramani (MIT Press, Cambridge, MA, 2002).
[22] L. Paninski, Neural Comput. **15**, 1191 (2003).
[23] T. Batu, S. Dasgupta, R. Kumar, and R. Rubinfeld, in *Proceedings of the 34th Symposium on Theory Computing*, ACM, 2002.
[24] G. A. Miller, in *Information Theory in Psychology; Problems and Methods II-B*, edited by H. Quastler (Free Press, Glencoe, IL, 1955), pp. 95–100.
[25] S. Panzeri and A. Treves, Network Comput. Neural Syst. **7**, 87 (1996).
[26] P. Grassberger, Phys. Lett. A **128**, 369 (1988).
[27] P. Grassberger, physics/0307138.
[28] D. Wolpert and D. Wolf, Phys. Rev. E **52**, 6841 (1995).
[29] I. Nemenman, e-print physics/0207009.
[30] M. F. Land and T. S. Collett, J. Comp. Physiol. **89**, 331 (1974).
[31] B. S. Clarke and A. R. Barron, IEEE Trans. Inf. Theory **36**, 453 (1990).
[32] I. Nemenman, Ph.D. thesis, Princeton University, 2000.
[33] W. Bialek, I. Nemenman, and N. Tishby, Neural Comput. **13**, 2409 (2001).
[34] R. de Ruyter van Steveninck and W. Bialek, in *Methods in Neural Networks IV*, edited by J. van Hemmen, J. D. Cowan, and E. Domany (Springer-Verlag, Heidelberg, New York, 2001), pp. 313–371 (see Fig. 17).
[35] In reference to Bayesian estimators, consistency usually means that, as N grows, the posterior probability concentrates around unknown parameters of the true model that generated the data. For finite parameter models, such as the one considered here, only technical assumptions like positivity of the prior for all parameter values, soundness (different parameters always correspond to different distributions) [31], and a few others are needed for consistency. For nonparametric models, the situation is more complicated. There one also needs ultraviolet convergence of the functional integrals defined by the prior [32,33].
[36] It may happen that information is a small difference between two large entropies. Then, due to statistical errors, methods that estimate information directly will have an advantage over NSB, which estimates entropies first. In our case, this is not a problem since the information is roughly a half of the total available entropy [4].
[37] For our and many other neural systems, the spike timing can be more accurate than the refractory period of roughly 2 ms [6,10,34]. For the current amount of data, discretization of $\tau \ll 1$ ms and large enough T will push the limits of all estimation methods, including ours, that do not make explicit assumptions about properties of the spike trains. Thus, to have enough statistics to convincingly show validity of the NSB approach, in this paper we choose $\tau = 0.75 \cdots 2$ ms, which is still much shorter than other methods can handle. We leave open the possibility that more information is contained in timing precision at finer scales.