PUPT-1611

Entropy and Information in Neural Spike Trains

Steven P. Strong,[1] Roland Koberle,[1,2] Rob R. de Ruyter van Steveninck,[1] and William Bialek[1]
[1] *NEC Research Institute, 4 Independence Way, Princeton, New Jersey 08540*
[2] *Department of Physics, Princeton University, Princeton, New Jersey 08544*
(March 15, 1996)

Fermilab Library
0 1160 0042690 2

The nervous system represents time-dependent signals in sequences of discrete action potentials or spikes; all spikes are identical so that information is carried only in the spike arrival times. We show how to quantify this information, in bits, free from any assumptions about which features of the spike train or input waveform are most important. We apply this approach to the analysis of experiments on a motion-sensitive neuron in the fly visual system.

FERMILAB
MAY 21 1996
LIBRARY

As you read this text, optical signals reaching your retina are encoded into sequences of identical pulses, termed action potentials or spikes, that propagate along the $\sim 10^6$ fibers of the optic nerve from eye to brain. This spike encoding appears almost universal, occurring in animals as diverse as worms and man, and spanning all the sensory modalities [1]. The molecular mechanisms for the generation and propagation of action potentials are well understood [2], as are the mathematical reasons for the selection of stereotyped pulses by the dynamics of the nerve cell membrane [3]. Less well understood is the function of these spikes as a code [4]: How do the sequences of spikes represent the sensory world, and how much information is conveyed in this representation?

Experiments on spiking neurons are usually analyzed either by making assumptions about which features of the spike train are significant in the code, or by making assumptions about which features of the sensory input are being encoded. These assumptions become less compelling in deeper layers of the brain, so one would like a model independent approach.

Imagine that we present a long duration sensory stimulus drawn from some particular ensemble of stimuli, perhaps the natural one. During this long period the spike train varies, and we can quantify this variability by the entropy per unit time of the spike train. Because neurons have a minimum "refractory period" separating successive spikes, we can digitize the spike train in bins of size $\Delta\tau$ so that we never observe more than one spike per bin (Fig. 1a). The spikes then form an Ising chain where spin up (down) represents the presence (absence) of a spike in a given time bin. If there is a deterministic mapping from input signals to output spike trains, then the entropy of this Ising chain is exactly the information, as defined by Shannon [5], that the spike train provides about the sensory stimulus. In fact repeated presentations of the same stimulus lead to different spike trains (Fig. 1b). This variability can also be quantified by an entropy, which we call the conditional or noise entropy. We think of this as the entropy of an ensemble of Ising chains with quenched interactions determined by the stimulus. The information that the spike train provides about the stimulus is the difference between the total spike train entropy and this conditional entropy.

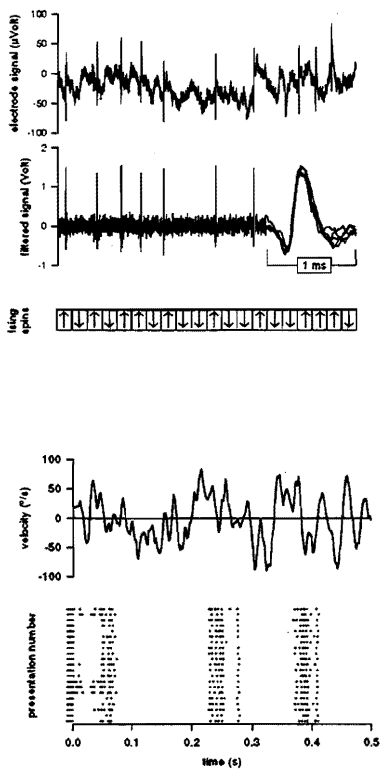

FIG. 1. (A) Raw voltage records from a tungsten microelectrode near the cell H1 are filtered and discretized to form (for physicists) a chain of Ising spins or (for computer scientists) a binary string. (B)Angular velocity of a pattern moving across the fly's visual field produces a sequence of spikes in H1, indicated by dots. Repeated presentations produce slightly different spike sequences. For experimental methods see Ref. [14].

The problem of measuring information transmission by spiking neurons is thus exactly the problem of estimating

the entropy of Ising chains from examples of their configurations. This analogy is model independent, making no assumptions about the structure of the input signal or the encoding process. These ideas allow us to quantify the response of neurons in a very general way, even when they are responding to complex, time dependent stimuli. The difficulty is that brute force application of these ideas requires large amounts of data.

The most obvious way to estimate the entropy from examples of the configuration is to accumulate a large number of these configurations into a histogram, assume that the (normalized) count in each bin of the histogram approximates the probability of that configuration, and evaluate the entropy directly from the definition $S = -\sum_i p_i \log_2 p_i$ bits. This procedure works only if each bin contains many counts. Imagine that we have a spike train with mean spike rate $\bar{r} \sim 40$ spikes/s and we sample with a time resolution $\Delta\tau = 3$ ms. The maximum entropy distribution consistent with this mean rate is, in the Ising language, a set of uncoupled spins with magnetic field chosen so that the probability of spin up is $\bar{r}\Delta\tau = 0.12$. Then the entropy of the configurations in a window of size $T = 100$ ms is

$$S_{\max} = \frac{T}{\Delta\tau}[-(\bar{r}\Delta\tau)\log_2(\bar{r}\Delta\tau) - (1-\bar{r}\Delta\tau)\log_2(1-\bar{r}\Delta\tau)] \sim 17.8\,\text{bits}. \quad (1)$$

Naively, then, we need to see more than $2^S \sim 2 \times 10^5$ examples of the 100 ms windows. If these windows have to be non-overlapping, then we need more than three hours of data, and one might think that we need *much* more data than this to insure that the probability of occupying each bin is estimated with reasonable accuracy. Such large quantities of data are generally inaccessible for experiments on real neurons.

Here we report that it is possible to make progress despite these pessimistic estimates. There are several ingredients: First, we examine explicitly the dependence of our entropy estimates on the size of the data set and find regular behaviors [6] that can be extrapolated to the infinite data limit. Second, generalizing the suggestion by Ma [7], we evaluate lower bounds to the entropy that are much less prone to sampling errors. Third, we are interested in the extensive component of the entropy in large time windows, and we find that a clean approach to extensivity is visible before sampling problems set in. Finally, we study a neuron—the motion-sensitive neuron H1 in the fly's visual system—where we can actually collect many hours of data. Lessons from this large data set provide some guidance for the design and analysis of experiments with more limited recording times.

H1 responds to motion across the entire visual field, producing more spikes for an inward horizontal motion and fewer spikes for an outward motion; vertical motions have no effect on this cell, but are coded by other neurons [8]. This circuitry provides visual feedback for flight control. In the experiments analyzed here [9], the fly is immobilized and views computer generated images on a display oscilloscope. For simplicity these images consist of a fixed pattern of vertical stripes with randomly chosen grey levels, and this pattern takes a random walk in the horizontal direction. The stimulus is the time dependent velocity of this walk.

We begin our analysis with time bins of size $\Delta\tau = 3$ ms. For a window of $T = 30$ ms—corresponding to the behavioral response time of the fly [10]—we can estimate the entropy rather accurately by the naive procedure described above. Figure 2 shows the resulting entropies as a function of the amount of data included in the analysis, and we see that there are very small finite data set corrections ($< 10^{-3}$), well fit by $S_{\text{naive}} = S_0 + S_1/\text{size} + S_2/\text{size}^2$ [6]. Under these conditions we feel confident that the extrapolated S is the correct entropy for this combination of T and $\Delta\tau$. If we extend the naive procedure to windows of $T = 200$ ms, finite size corrections become much larger, the contribution of the second correction is significant and the extrapolation to infinite size is unstable.

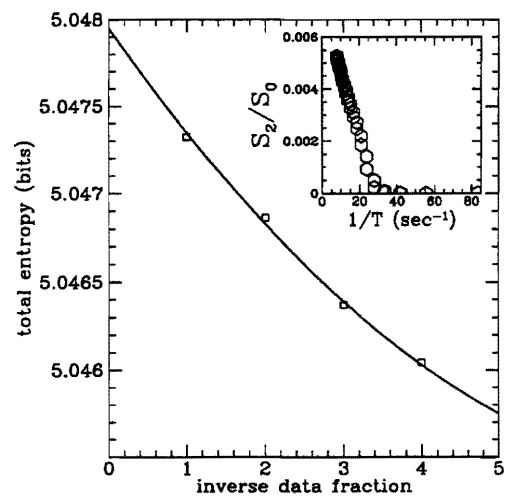

FIG. 2. The estimated entropy, measured in bits, plotted versus the inverse size of the data set used in the estimation. Results are for windows of length $T = 30$ ms and a time resolution of $\Delta\tau = 3$ ms; size = 1 corresponds to 2.5 hours of recording. Also plotted is a least squares fit to the form $S = S_0 + S_1/\text{size} + S_2/\text{size}^2$. The intercept S_0 is our extrapolation to the true value of the entropy with infinite data.

Ma [7] discussed the problem of entropy estimation in the undersampled limit. For probability distributions that are uniform on a set of N bins (as in the microcanonical ensemble), the entropy is $\ln N$ and the problem is to estimate N. Ma noted that this could be done by counting the number of times that two randomly chosen observations yield the same configuration, since the probability of such a coincidence is $1/N$. In the undersampled limit, Ma's procedure is actually the optimal estimator

of the entropy in the least-squares sense if all values of N are equally likely *a priori* [11]. More generally, for non-uniform distributions the probability of a coincidence is $P_c = \sum p_i^2$, and it is clear that

$$S = -\sum p_i \log_2 p_i = -\langle \log_2 p_i \rangle$$
$$\geq -\log_2 (\langle p_i \rangle) = -\log_2 P_c, \qquad (2)$$

so we can compute a lower bound the the entropy by counting coincidences. This is the minimum entropy consistent with a given P_c, and it is one of the Renyi entropies [12]. It is also at the heart of algorithms for the analysis of attractors in dynamical systems [13].

The bound in Eq. (2) is tightest for distributions that are close to uniform. The distributions of spike sequences cannot be uniform because the spikes are sparse. But the distribution of sequences with fixed spike count $N_{\rm sp}$ could be more nearly uniform, and so we apply the Ma bounding procedure independently in each $N_{\rm sp}$ sector. Thus we obtain a lower bound to the entropy,

$$S_{\rm Ma} = -\sum_{N_{\rm sp}} P(N_{\rm sp}) \times \log_2 \left[P(N_{\rm sp}) \frac{n_c(N_{\rm sp})}{N_{\rm obs}(N_{\rm sp})[N_{\rm obs}(N_{\rm sp}) - 1]} \right], \qquad (3)$$

where $n_c(N_{\rm sp})$ is the number of coincidences observed among the words with $N_{\rm sp}$ spikes, $N_{\rm obs}(N_{\rm sp})$ is the total number of occurrences of words with $N_{\rm sp}$ spikes, and $P(N_{\rm sp})$ is the fraction of words with $N_{\rm sp}$ spikes.

The bound $S_{\rm Ma}$ suffers from two types of sampling errors. First, there are those associated with the estimate of $P(N_{\rm sp})$. A slight generalization of the analysis of Ref. [6] shows that the leading error in the entropy computation for each $N_{\rm sp}$ is simply $-[N_{\rm obs}(N_{\rm sp}) \ln 2]^{-1}$. The second type of errors are associated with estimating $\sum p_i^2$ from the observed number of coincidences $n_c(N_{\rm sp})$. This produces an error in the contribution to the entropy at each $N_{\rm sp}$,

$$\Delta S_{\rm Ma} = \frac{2}{N_{\rm obs}(N_{\rm sp})} \frac{\sum' p_i^3 - (\sum' p_i^2)^2}{(\ln 2)(\sum' p_i^2)^2} + \ldots, \qquad (4)$$

where the sum is taken only over words with $N_{\rm sp}$ spikes. This is again $\sim 1/N_{\rm obs}$ and exponentially smaller than the sampling errors expected for the naive estimator. For the case of uniform probability distributions at fixed spike number, even this leading term vanishes.

In accord with these arguments, we found a very weak dependence of the Ma bound on the size of the data set. Furthermore, the error bars which we estimate by subdividing the data are also extremely small. These results give us confidence that the procedure in Eq. (3) gives a reliable lower bound to the entropy. In Fig. 3 we finally plot the entropy as a function of the window size T, with results from both the naive procedure and from the Ma bound. For sufficiently large windows the naive procedure gives an answer smaller than the Ma bound, and hence the naive answer must be wrong because it is more sensitive to sampling problems. Before this sampling disaster the lower bound and the naive estimate are never more than 10–15% apart. The point at which the naive estimate crashes into the Ma bound is also the point at which the second correction, S_2, becomes significant and we lose control over the extrapolation to the infinite data limit. This point occurs at a window size of roughly 100 ms. We can trust the Ma bound beyond this point, but it becomes steadily less powerful. The sudden transition from accurate estimation to a crude lower bound occurs because the number of possible spike trains is growing exponentially with window size, and hence we are either very well sampled or disastrously undersampled. Conversely, the critical window size is only logarithmically dependent on the size of the data set, so we do not sacrifice much if we have more limited duration recordings.

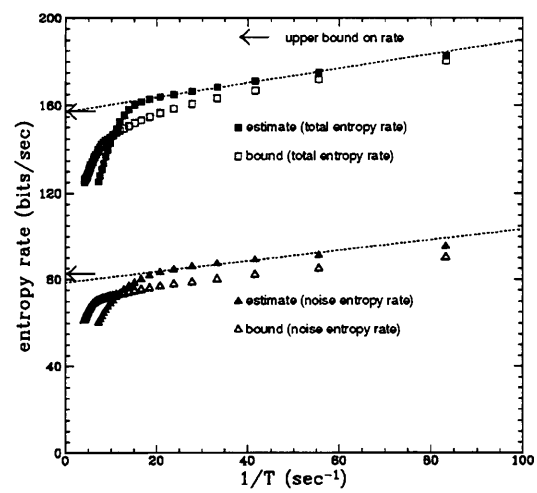

FIG. 3. The total and noise entropies per unit time (in bits per second) are plotted versus the reciprocal of the window size (in s^{-1}), with the time resolution held fixed at $\Delta\tau = 3$ ms. Results are given both for the direct estimate and for the bounding procedure described in the text, and for each data point we apply the extrapolation procedures of Fig. 2. As discussed in the text, the plot is linear for wordlengths that are not too long, where undersampling is severe, and not too short, a regime where correlations spoil extensivity. The linearity allows us to extrapolate to the entropy per unit time for words of infinite length. Arrows indicate upper bounds obtained by differentiating $S(T)$, as explained in the text. The information transmitted per unit time about the stimulus is the extrapolated total entropy per unit time minus the extrapolated noise entropy per unit time.

If the correlations in the spike train have finite range, then the leading sub-extensive contribution to the entropy will be a constant. Thus, if we plot the entropy per unit time as a function of inverse window size, we should see a linear behavior which can be extrapolated to the

infinite time limit. This is seen clearly in Fig. 3, and emerges before the sampling disaster. Indeed the break away from linear behavior is close to the point where the naive estimate crashes into the Ma bound and S_2 becomes large, confirming our identification of this point. Given the clean linear behavior in a well sampled region of the plot, we trust the extrapolation and arrive at an estimate of the entropy per unit time as seen with 3 ms resolution, 157 ± 3 bits/s.

In principle we can obtain an upper bound to the entropy per unit time simply by differentiating $S(T)$ [5]. This bound becomes progressively tighter at larger T, until sampling problems set in. In fact there is a broad plateau ($\pm 2.7\%$) in the range $18 < T < 60$ ms, leading to $S \leq 157 \pm 4$ bits/s, in excellent agreement with the extrapolation in Fig. 3.

The above discussion applies to the estimate of the total entropy. All the same issues arise in estimating the noise entropy, where the same stimulus has to be shown repeatedly, and Fig. 3 also shows the noise entropy results as a function of the window size. The difference between the two entropies is the information which the cell transmits, $R_{\text{info}} = 78 \pm 18$ bits/s, or 1.8 ± 0.4 bits/spike at a resolution $\Delta\tau = 3$ ms.

We repeated the analysis of entropy and information at several different values of the time resolution $\Delta\tau$, from 0.7 ms to 12 ms. The information rate R_{info} has a very linear dependence on the log of the time resolution throughout this range; loss of one bit of precision in the specification of spike times ($\Delta\tau \to 2\Delta\tau$) causes a ~ 10 bit/s drop in the information rate. This is not what we expect for a system with a single scale characterizing the intrinsic temporal precision of the code. Instead it suggests that different regions of the spike train have different intrinsic precisions [14].

The temporal sequence of spikes provides a large capacity for transmitting information, as emphasized by MacKay and McCulloch nearly 45 years ago [15]. One central question in studies of the brain is whether this large capacity is used, or whether the variations in spike timing represent noise which must be averaged away [16]. We suggest that the proper quantitative formulation of this question is the comparison of information transmission rates (in bits/s) with the total entropy of the spike train (also in bits/s) as a function of the time resolution $\Delta\tau$ [17]. For the sensory neurons which first convert continuous input signals into discrete spike trains, we understand enough about what the spike train represents that direct "decoding" of the spike train is possible; this leads to information rates which are more than half the total entropy with timing resolutions of roughly 1 ms [17,18]. The idea that sensory neurons provide a maximally efficient representation of the outside world has also been suggested as an optimization principle from which many features of these cells' responses can be derived [19]. But much of the current debate is focused on cells in the central nervous system [16], where assumptions about what is being encoded should be viewed with caution.

Although we understand a good deal about the signals represented in H1 [8,20], our present analysis does not hinge on this knowledge. Similarly, although it is possible to collect very large data sets from H1, it is clear from Fig.'s 2 and 3 that more limited data sets would compromise our conclusions only slightly. It therefore seems feasible to apply these same analysis techniques to cells in the mammalian brain. Like cells in the monkey or cat primary visual cortex, H1 is four layers 'back' from the array of photodetectors and receives its inputs from thousands of synapses. For this central neuron, we find that information transmission rates are half the total spike train entropy at a time resolution of 1.5 ms.

We thank K. Miller and P. Mitra for helpful discussions, and G. Lewen for his help with the experiments. R. K., on leave from the Institute of Physics, Universidade de São Paulo, São Carlos, Brasil, was supported in part by CNPq and FAPESP.

---

[1] E. D. Adrian, *The Basis of Sensation* (W. W. Norton, New York, 1928).

[2] D. J. Aidley, *The Physiology of Excitable Cells, 3d ed.* (Cambridge University Press, Cambridge, 1989)

[3] A. L. Hodgkin and A. F. Huxley, *J. Physiol.* **117**, 500 (1952); D. G. Aronson and H. F. Weinberger, *Adv. Math.* **30**, 33 (1978).

[4] For a review see F. Rieke, D. Warland, R. R. de Ruyter van Steveninck and W. Bialek, *Spikes: Exploring the Neural Code* (MIT Press, Cambridge MA, in press).

[5] C. E. Shannon and W. Weaver (1949). *The Mathematical Theory of Communication* (University of Illinois Press, Urbana).

[6] A. Treves and S. Panzeri, *Neural Comp.* **7**, 399 (1995).

[7] S. K. Ma, *J. Stat. Phys.* **26**, 221 (1981).

[8] For reviews on the motion sensitive neurons see the chapters by N. Franceschini, A. Riehle and A. le Nestour and by K. Hausen and M. Egelhaaf in *Facets of Vision,* D. G. Stavenga and R. C. Hardie, eds. (Springer-Verlag, Berlin, 1989). For an overview of the fly's anatomy, emphasizing the nearly crystalline structure of the neural circuits, see the contribution by N. J. Strausfeld to the same volume.

[9] For a discussion of experimental methods see Ref. [14].

[10] M. F. Land and T. S. Collett, *J. Comp. Physiol.* **89**, 331 (1974).

[11] S. P. Strong, unpublished.

[12] A. Renyi, *Probability Theory* (North Holland, Amsterdam, 1970).

[13] D. Ruelle, *Chaotic Evolution and Strange Attractors* (Cambridge University Press, Cambridge UK, 1989).

[14] One simple idea is that short intervals between spikes are determined with (at least) proportionately greater

precision than long intervals. A completely independent analysis led to this same conclusion, again with a range of time resolutions from $\sim 1 - 15$ ms: R. R. de Ruyter van Steveninck and W. Bialek, *Proc. R. Soc. Lond. Ser. B.* **234,** 379 (1988).

[15] D. MacKay and W. S. McCulloch, *Bull. Math. Biophys.* **14,** 127 (1952).

[16] One can trace the debate through the references and texts of recent commentaries. See, for example, the exchange between M. N. Shadlen and W. T. Newsome, *Curr. Opin. Neurobiol.* **5,** 248 (1995) and W. R. Softky, *ibid.* **5,** 239 (1995), as well as T. J. Sejnowksi, *Nature* **376,** 21 (1995); D. Ferster and N. Spruston, *Science* **270,** 756 (1995); C. F. Stevens and A. Zador, *Curr. Biol.* **5,** 1370 (1995).

[17] F. Rieke, D. Warland, and W. Bialek, *Europhys. Lett.* **22,** 151 (1993).

[18] In the frog auditory system, information rates reach 90% of the total entropy when signals are chosen from an ensemble that captures the spectral structure of naturally occurring frog calls. F. Rieke, D. A. Bodnar, and W. Bialek, *Proc. R. Soc. Lond. Ser. B* **262,** 259 (1995).

[19] J. J. Atick, in *Princeton Lectures on Biophysics,* W. Bialek, ed., pp. 223-289 (World Scientific, Singapore, 1992).

[20] W. Bialek, F. Rieke, R. R. de Ruyter van Steveninck, and D. Warland, *Science* **252,** 1757 (1991).