



## Binless strategies for estimation of information from neural data

Jonathan D. Victor\*

*Department of Neurology and Neuroscience, Weill Medical College of Cornell University, 1300 York Avenue, New York, New York 10021*

(Received 5 November 2001; revised manuscript received 6 August 2002; published 11 November 2002)

We present an approach to estimate information carried by experimentally observed neural spike trains elicited by known stimuli. This approach makes use of an embedding of the observed spike trains into a set of vector spaces, and entropy estimates based on the nearest-neighbor Euclidean distances within these vector spaces [L. F. Kozachenko and N. N. Leonenko, *Probl. Peredachi Inf.* **23**, 9 (1987)]. Using numerical examples, we show that this approach can be dramatically more efficient than standard bin-based approaches such as the “direct” method [S. P. Strong, R. Koberle, R. R. de Ruyter van Steveninck, and W. Bialek, *Phys. Rev. Lett.* **80**, 197 (1998)] for amounts of data typically available from laboratory experiments.

DOI: 10.1103/PhysRevE.66.051903

PACS number(s): 87.10.+e, 02.70.-c, 02.50.-r, 87.19.La

### INTRODUCTION

How neurons represent, process, and transmit information is of fundamental interest in neuroscience [1,2]. It is accepted that neural information processing relies on the transmission of a series of stereotyped events. The basic biophysics that underlies the generation of these action potentials (spikes) is well established. However, the statistical features that convey information are not well understood. Possibilities include not only obvious features, such as the number of spikes fired by a population of neurons [3], but also more subtle ones, such as, their precise times of occurrence [4,5], the pattern of intervals [6], and various kinds of patterns of activity across a population [7,8]. A direct experimental assault on this question is difficult, since manipulations (such as chemical or electrical stimulation) that change one aspect of neural activity are likely to change others. Thus, an appropriate theoretical infrastructure is required to disentangle such potential confounds.

Shannon’s groundbreaking work in information and communication theory [9] is the natural basis for this theoretical infrastructure [1]. Quantifying the amount of information contained in neural activity, often in conjunction with appropriate simulations and models, makes it possible to determine the relevant statistical features of spike trains [10] and to examine overall biological strategies for information transfer [11].

However, as is becoming increasingly appreciated, estimation of information content from empirical data can be fraught with difficulties. Estimation of information in spike trains generally consists of several steps: (i) embedding spike trains into a space, (ii) clustering similar spike trains into groups, (iii) using a “plug-in” estimate for transmitted information based on how these groups relate to the stimuli, and (iv) estimating biases due to small sample size. Traditional approaches (e.g., the “direct” method of Strong *et al.* [12]) subdivide a spike train into narrow time bins (binning) as part of the embedding stage (i), with each bin corresponding to a separate dimension. Bins that are too wide lead to underestimates of information since temporal detail is lost,

while bins that are too narrow lead to biases associated with extreme undersampling. Standard [13–15] and jackknife estimators [16] of bias at stage (iv) can be helpful, but no bias correction is effective when the amount of data is very limited. Metric-space methods [17] avoid the binning problem, but still may underestimate information due to the clustering at stage (ii). The fundamental difficulty is that estimates of information that make few assumptions concerning the nature of the code suffer biases because of limited amounts of data, while methods that reduce the dimensionality of the problem by considering a parametric family of codes suffer biases if the neural code does not belong to one of the families.

This paper presents a strategy that bypasses the difficulties associated with binning and clustering, while making only weak assumptions concerning the nature of the code. Essentially, we assume that the neural code respects the continuity of time, but we make no assumptions as to the relationships between spike trains with different numbers of spikes. That is, we recognize the distinctive topology of the space of spike trains: there is a discrete component, corresponding to the number of spikes in a response, and there is a continuous component, corresponding to the timing of those spikes [18].

Implementation of the idea rests on a little-known asymptotically unbiased “binless” estimator of differential entropy [19]. We first show that this estimator has substantial computational advantages in a broader context: estimation of the entropy of a continuous distribution from a finite set of empirical samples in a Euclidean space. We then proceed to apply this estimator to spike trains. This requires grouping of the spike trains into strata according to the number of spikes that they contain, followed by separate analysis of each stratum. To preserve the advantages of the binless estimator within each stratum, we use linear, continuous embeddings of spike trains, rather than embeddings based on binning. Information is then estimated from the difference between the entropy of the set of all spike trains, and the entropies of the spike trains elicited by each stimulus. In simulations, the rapid convergence of the binless entropy estimator leads to marked improvements in information estimates in the regime of limited data.

\*FAX: 212 746 8984. Email address: jdvicto@med.cornell.edu

## RESULTS

### Statement of the problem and overview of the approach

We focus on estimating the amount of information transmitted by a neuron in a particular but common neurophysiologic laboratory situation. Each member of discrete collection of stimuli  $a_k$  ( $k=1,2,\dots,S$ ) is presented repeatedly to the preparation. The neural response elicited by each stimulus presentation is a “spike train,” namely, a sequence of stereotyped events (spikes) in a predefined observation period following the presentation of the stimulus. In a typical experiment, the observation period following each presentation is on the order of 0.1–1 s, the number of spikes may range from 0 to 20, the number of stimuli  $S$  is 2–12, and each stimulus is presented several dozen times. The investigator keeps track of which stimulus elicits which responses, but would like to determine the extent to which the responses themselves allow the stimuli to be distinguished.

As pointed out above, estimation of Shannon’s “transmitted information” [1,9] is natural for this purpose. In essence, the transmitted information is the difference between the entropy of all of the spike trains, and the entropy of the spike trains elicited by repeated presentations of the same stimulus. Thus, estimating transmitted information is closely related to estimating entropy.

Were a neural response fully characterized by the number of spikes it contained [20], it would suffice to describe an ensemble of spike trains in a discrete fashion. This description would be a tabulation of the list of the probabilities that, given a stimulus  $a_k$ , a response containing exactly  $n$  spikes is elicited. In this case, procedures for obtaining entropy estimates from discrete distributions could be applied. Such procedures are well known, and their behavior, including their biases, are well understood [13–15].

However, it does not suffice to characterize a spike train merely by the number of spikes that it contains, since the timing of the spikes in the response may also contribute to the ability to discriminate among the several stimuli [1,2,4,6–8,17,21,22]. The usual and currently standard approach is to break up the observation interval into a number of discrete time bins [12]. Once this is done, procedures for obtaining entropy estimates from a discrete distribution can again be applied. The fundamental difficulty with this approach is that the bins must be made small enough to capture the intrinsic precision of spike times, which may be as fine as 1 ms [22,23]. This requires estimation of a very large number of response probabilities (one for each possible way to distribute the  $n$  response spikes into these bins). This in turn incurs a large bias in the entropy estimates: bias is approximately proportional to the number of response probabilities to be estimated [13–15], and the latter increases exponentially with the time resolution used to analyze the responses. Since the estimated information is a difference between two estimated entropies, and the biases of these entropy estimates are large and unequal, large biases in estimated information can result.

The present approach avoids this exponential growth in bias with increasing time resolution. The transmitted information is broken into two parts: one ( $I_{\text{count}}$ ) that can be

gleaned from the number of spikes in each response and another ( $I_{\text{timing}}$ ) that can be gleaned from their timing.  $I_{\text{count}}$  is estimated via standard methods. While the estimate of  $I_{\text{count}}$  must be debiased, the bias is small and independent of time resolution.  $I_{\text{timing}}$  is further subdivided into contributions  $I_{\text{timing}}(n)$  from responses that contain exactly  $n$  spikes, a step that incurs no estimation error. To estimate  $I_{\text{timing}}(n)$ , we exploit the fact that spike trains that contain exactly  $n$  spikes are parametrized by  $n$  continuous parameters (namely, the times of the spikes). Thus, it is natural to consider estimation of their entropies within the context of estimation of entropies of continuous probability distributions on a Euclidean vector space. This additional structure provides a tool that avoids the difficulties associated with binning. As shown by Kozachenko and Leonenko [19], for a finite sample drawn from a continuous distribution in a Euclidean vector space, the statistics of the Euclidean distances between nearest neighbors provide for an asymptotically unbiased and consistent estimate of the entropy.

To implement this approach, and to show that it indeed has practical advantages, requires a number of logical steps. We begin by introducing the binless entropy estimator of Kozachenko and Leonenko [19] for continuous distributions in a Euclidean space, for one-dimensional distributions, and then for multidimensional ones. We next show (via numerical experiments) that this estimator indeed has practical advantages over the traditional estimators that rely on binning. We then describe how the binless estimator can be adapted to the estimation of information in neural data sets of the sort described above. This requires several steps: (a) stratification of spike trains into discrete sets based on how many spikes they contain, (b) estimation of the information  $I_{\text{count}}$  associated with step (a), (c) embedding the spike trains within each of these discrete sets into a separate Euclidean space, and (d) application of the binless estimator within each of the Euclidean spaces to obtain contributions to  $I_{\text{timing}}$ .

The results of Kozachenko and Leonenko [19], along with the chain rule property of information [24], guarantee that, in the limit of an infinite amount of data, the proposed procedure provides an unbiased estimate of information. With limited data acquired at finite resolution, there are practical problems that arise in the implementation of stages (b), (c), and (d). We make generic choices for how to solve them in the course of the development below. We make no claim that these choices are optimal, and we mention several variations in the Discussion and Appendix. Nevertheless, as a variety of numerical simulations demonstrate, these choices result in a procedure that has substantial advantages in comparison to traditional binned approaches, for data sets whose size and nature are typical of those obtained in the neurophysiology laboratory.

Finally, we note that this approach can also be applied to experimental data that consist of continuous responses (e.g., field potentials) rather than spike trains, and also to situations in which the stimulus set is continuous rather than discrete. The former situation is simpler than the one we consider in detail, since the partition of information into continuous and discrete components is not necessary.

**Binless entropy estimates of one-dimensional distributions**

Let  $p(x)$  be a continuous probability density on the real line  $[-\infty, \infty]$ . Our immediate goal is to estimate the differential entropy [25] of  $p(x)$ , defined as

$$H_{\text{diff}} = - \int_{-\infty}^{\infty} p(x) \log_2 p(x) dx, \tag{1}$$

from a finite sample of observations  $x_1, \dots, x_N$  drawn according to  $p(x)$ .

The differential entropy  $H_{\text{diff}}$  characterizes the behavior of the entropy of discretized versions of  $p(x)$  in the limit of small bin widths  $b$ , which we denote  $H_{\text{disc}}(b)$ . In the limit of small bin widths  $b$ , the probability that a sample  $x$  is between  $x_i$  and  $x_i + b$  is approximated by  $bp(x_i)$ . Using  $\sum_i bp(x_i) = 1$ , it follows (in the limit of small  $b$ ) that the differential entropy and the discretized entropy  $H_{\text{disc}}(b)$  are related by

$$\begin{aligned} H_{\text{disc}}(b) &\approx - \sum_i bp(x_i) \log_2 bp(x_i) \\ &\approx - \log_2 b - \sum_i bp(x_i) \log_2 p(x_i) \\ &\approx - \log_2 b - \int_{-\infty}^{\infty} p(x) \log_2 p(x) dx \\ &= H_{\text{diff}} - \log_2 b, \end{aligned}$$

where the  $x_i$  are the centers of equally sized bins of width  $b$ .

For information estimates obtained via discretization, only this limiting behavior is of interest, since it captures the greatest amount of detail (as formalized by the data processing theorem [24]). The final term  $-\log_2 b$  in the above equation is irrelevant to estimates of information, since the information estimates are differences of two entropy estimates, each obtained with the same bin size. The continuous approach bypasses this limiting process, and replaces the difference of discretized entropies with the equivalent difference of differential entropies.

We seek an estimate for the differential entropy [Eq. (1)] that depends continuously on the individual observations. We would like to exploit the (assumed) continuous nature of  $p$ , but to keep the estimation procedure local, so that sensitivity to the shape of  $p$  is preserved. The analysis below should be viewed as heuristic development of the binless estimator. For a rigorous proof, the reader is referred to Kozachenko and Leonenko [19].

The first step is to change the variable of integration in Eq. (1) to the cumulative probability density  $y$ , defined by

$$y = \int_{-\infty}^x p(t) dt.$$

Under this change of variables,  $dy = p(x) dx$ , and Eq. (1) transforms to

$$H_{\text{diff}} = - \int_0^1 \log_2 p(x) dy. \tag{2}$$

This equation states that the differential entropy is determined by the average of  $\log_2 p(x)$ , where the average is equally weighted with respect to the cumulative probability density  $y$ . However, the available data consist only of the  $N$  sample points  $x_j$ , and  $y$  is unknown. We estimate  $y$  by taking it to be the function that is *exactly* the cumulative probability distribution of the  $N$  observed samples. That is,  $y$  is estimated by a function that is 0 at  $x = -\infty$  and has an abrupt step of size  $1/N$  at each value  $x_j$ . Since  $y$  is a step function,  $dy$  is a formal sum of  $\delta$  functions of weight  $1/N$  at each value  $x_j$ . This provides an approximation of the integral of Eq. (2) by the sum

$$H_{\text{diff}} \approx - \sum_{j=1}^N \frac{1}{N} \log_2 p(x_j). \tag{3}$$

Note that the  $x_j$  are determined by random draws according to  $p$ , in contrast to the  $x_i$  above, which are determined by the positions of equally spaced bins.

We now estimate  $\log_2 p(x_j)$  from the Euclidean distance between  $x_j$  and its nearest neighbor. The rationale for an estimate of this sort is that in some sense, it is as local as possible given the available data. We proceed as follows. Let  $q(\lambda)$  be the probability that, after  $N - 1$  other samples have been drawn according to  $p$ , the nearest neighbor to a sample  $x_j$  is at a distance of least  $\lambda$ . The probability density for  $\lambda$  is thus  $-dq/d\lambda$ . As  $\lambda$  increases by  $\Delta\lambda$ ,  $q(\lambda)$  decreases according to the probability of encountering any of the  $N - 1$  samples in either of two intervals of length  $\Delta\lambda$  extending on either side of  $x_j$ . The continuity assumption for  $p$  means that within a sufficiently small neighborhood of  $x_j$ , we can approximate  $p$  by a locally uniform distribution of density  $p(x_j)$ . That is, we can approximate

$$\frac{dq}{d\lambda} \approx -2(N - 1)p(x_j)q(\lambda), \tag{4}$$

and consequently (since  $q(0) = 1$ ),

$$q(\lambda) \approx \exp[-2\lambda(N - 1)p(x_j)]. \tag{5}$$

Using Eqs. (4) and (5) and the substitution  $u = 2\lambda(N - 1)p(x_j)$ ,

$$\begin{aligned} \langle \log_2 \lambda \rangle &= \int_0^{\infty} \log_2(\lambda) \left( -\frac{dq}{d\lambda} \right) d\lambda \\ &\approx \int_0^{\infty} \log_2 \left[ \frac{u}{2(N - 1)p(x_j)} \right] e^{-u} du \\ &= -\log_2[2(N - 1)p(x_j)] - \frac{\gamma}{\ln(2)}, \end{aligned}$$

where  $\gamma = -\int_0^{\infty} e^{-v} \ln v dv$ , the Euler-Mascheroni constant ( $\approx 0.5772156649$ ). This can be rewritten as

$$-\log_2 p(x_j) \approx \langle \log_2 \lambda \rangle + \log_2[2(N - 1)] + \frac{\gamma}{\ln(2)}, \tag{6}$$

the desired relationship between  $\log_2 p(x_j)$  and the Euclidean distance to the nearest neighbor. This relationship, when substituted into Eq. (3), gives the estimate

$$H_{\text{diff}} \approx \frac{1}{N} \sum_{j=1}^N \log_2 \lambda_j + \log_2 [2(N-1)] + \frac{\gamma}{\ln(2)}, \quad (7)$$

where  $\lambda_j$  is the observed distance from  $x_j$  to its nearest neighbor.

**Binless entropy estimates of multidimensional distributions**

The above strategy readily extends to multidimensional distributions  $p(x)$ , where  $x$  is a point in an  $r$ -dimensional Euclidean space. The differential entropy [Eq. (1), interpreted as a multidimensional integral] and the discretized entropy calculated with respect to an  $r$ -dimensional bin of width  $b$  are related by

$$H_{\text{disc}}(b) \approx H_{\text{diff}} - r \log_2 b. \quad (8)$$

The finite sum approximation [Eq. (3)] remains valid, but the relationship between  $p(x_j)$  and the expected distribution  $q(\lambda)$  of Euclidean distances to the nearest neighbor of  $x_j$  must be modified [Eqs. (4) and (5)]. This is because the volume associated with a change in Euclidean distance from  $\lambda$  to  $\lambda + \Delta\lambda$  is the volume of an  $r$ -dimensional spherical shell of radius  $\lambda$  and thickness  $\Delta\lambda$ . That is,

$$\frac{dq}{d\lambda} \approx -S_r \lambda^{r-1} (N-1) p(x_j) q(\lambda), \quad (9)$$

where

$$S_r = \frac{r \pi^{r/2}}{\Gamma\left(\frac{r}{2} + 1\right)}$$

is the area of a unit  $r$ -dimensional spherical surface ( $S_1 = 2, S_2 = 2\pi, S_3 = 4\pi, \dots$ ). Following the same lines as the one-dimensional analysis above, we find

$$q(\lambda) \approx \exp\left[-\frac{S_r \lambda^r (N-1) p(x_j)}{r}\right]$$

and

$$\langle \log_2 \lambda \rangle \approx \frac{1}{r} \left( -\log_2 \left[ \frac{S_r (N-1) p(x_j)}{r} \right] - \frac{\gamma}{\ln(2)} \right),$$

which, when substituted into Eq. (3), provides the estimate

$$H_{\text{diff}} \approx \frac{r}{N} \sum_{j=1}^N \log_2(\lambda_j) + \log_2 \left[ \frac{S_r (N-1)}{r} \right] + \frac{\gamma}{\ln(2)}. \quad (10)$$

This is Eq. (2) of Kozachenko and Leonenko [19]. It was shown by these authors to be asymptotically unbiased and consistent, provided that  $p$  obeys certain conditions that control the convergence of integrals for the differential entropy.

The strategy of estimating differential entropy from nearest-neighbor Euclidean distances is related to the strategy of estimating fractal dimension from the statistics of nearest-neighbor distances, the third approach discussed by Grassberger [26]. Note, however, that the dimension corresponds to the slope of the dependence of  $\log$  (nearest-neighbor distance) on  $\log$  (number of samples), while differential entropy corresponds to the intercept. Thus, the debiasers developed in Ref. [26] for the dimension do not immediately extend to the present situation, in which the dimension is a known integer, and entropy is to be estimated.

**Numerical examples: Entropy estimates**

The asymptotically unbiased and consistent nature of the binless estimators suggests, but does not guarantee, its utility in practical application to finite data sets. We therefore illustrate the performance of the binless estimators of Eqs. (7) and (10) with some numerical examples, focusing on a comparison with standard estimates based on binning. Figure 1 considers a one-dimensional Gaussian of unit variance. The upper panels show that the binned estimates approach the correct value asymptotically, provided that the bin width is sufficiently small (i.e., 0.125 or 0.5). We show the behavior of two bias corrections for the binned estimates: (i) the bias correction of Miller [14] and Carlton [13] (which corresponds [27] to the bias correction for entropy proposed by Treves and Panzeri [15] and is henceforth referred to as the ‘‘classical’’ correction), and (ii) the jackknife [16]. The latter correction tends to result in a somewhat higher value and a smaller error, especially in the small-sample, small bin-width regime. The lower panels of Fig. 1 show that for a fixed number of samples, binned estimates (even if debiased) underestimate differential entropy when the bin width is small, and overestimate differential entropy when the bin width is large. The binless estimate (reproduced in all three upper panels) has essentially no bias, as expected from the analytical results of Kozachenko and Leonenko [19]. The trade-off for this lack of bias is that the binless estimates are considerably less precise than the binned estimates.

Numerical experiments (not shown) revealed similar behavior for other one-dimensional distributions, including: a uniform distribution, a one-sided exponential distribution, and a Lorentz distribution. The similarity is remarkable, considering that these distributions differ in whether their support is compact, semi-infinite, or infinite; whether the densities have discontinuities, and whether the distributions have finite variance.

Figure 2 compares these estimators for a three-dimensional Gaussian distribution. The above features remain, but their relative importance has changed. When bin size is small (0.5 or less), the underestimate of differential entropy associated with finite sample size has become so severe that hundreds of samples are required to achieve an acceptable estimate. The upward bias in differential entropy due to noninfinitesimal bins has also become severe. Thus, only a very narrow range of bin widths (ca. 1) will yield an acceptable estimate. On the other hand, the imprecision of the binless estimator has increased only slightly. Conse-

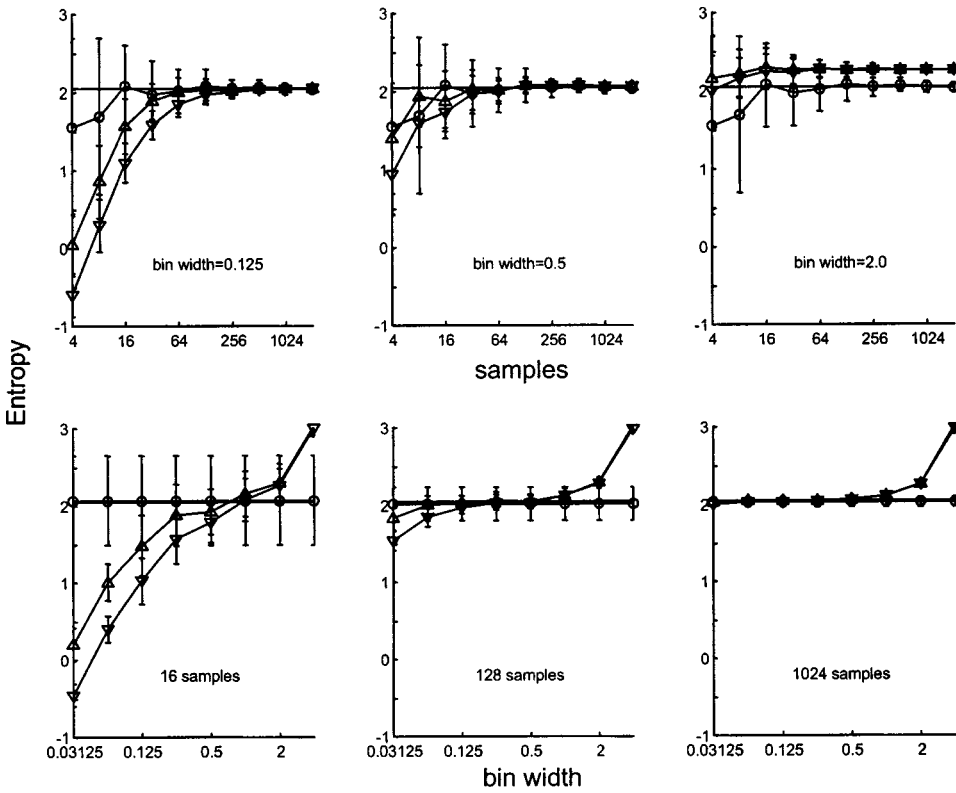


FIG. 1. Differential entropy of a one-dimensional Gaussian distribution (unit variance), as estimated from a finite sample of data via binned and binless approaches. Horizontal line: correct value, ca. 2.047. Triangles: mean of binned estimates, corrected via the classical method (down triangles) and the jackknife (up triangles). Open circles: binless estimates, calculated from Eq. (7). (These estimates do not depend on bin size, but are reproduced across bin sizes to facilitate comparison with the binned estimates.) The error bars represent the standard deviations of the estimates. Forty independent runs for each set of conditions. All calculations were carried out in MATLAB version 5.3.1 for Windows.

quently, an acceptable estimate of differential entropy can be obtained with 100 or fewer samples. These trends are even more apparent for a five-dimensional Gaussian distribution (Fig. 3).

Since Eq. 10 has a bias correction term that explicitly depends on the number of dimensions  $r$ , one might be con-

cerned that the performance of the binless estimator will be degraded when the dimensionality of a dataset is not clearcut. This is addressed in Fig. 4, which examines differential entropy estimates for a three-dimensional Gaussian whose variances along its three axes are in the ratio 1:10:100. The measurable bias associated with the binless

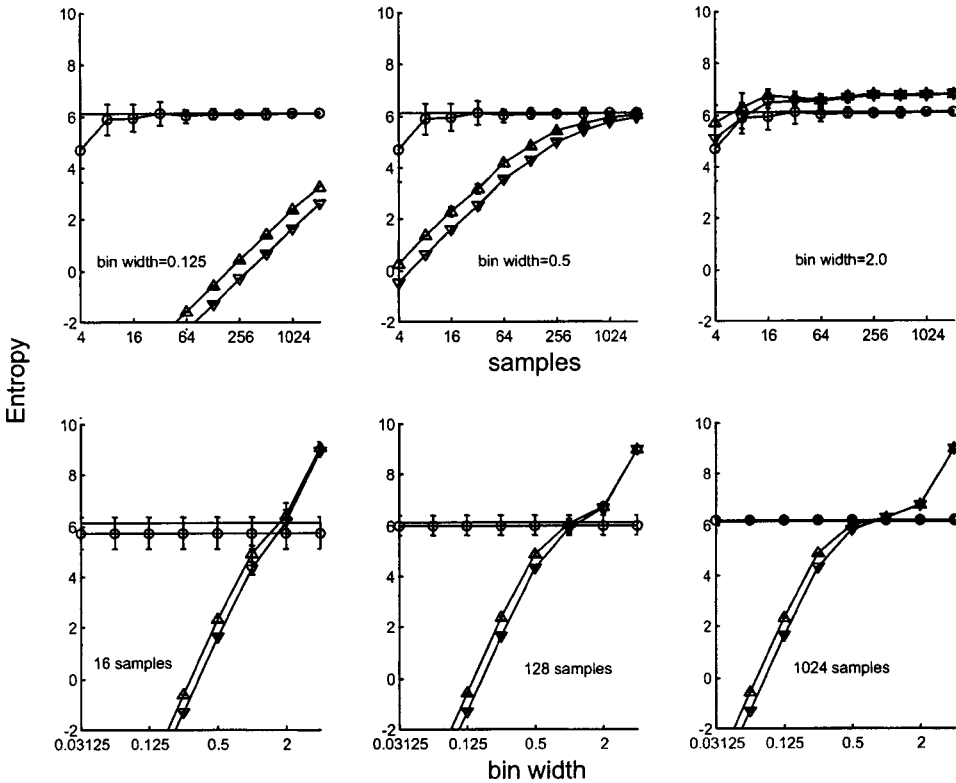


FIG. 2. Differential entropy of a three-dimensional Gaussian distribution (unit variance), estimated from a finite sample of data via binned and binless approaches. Correct value, ca. 6.141. Ten independent runs for each set of conditions. Display conventions as in Fig. 1.

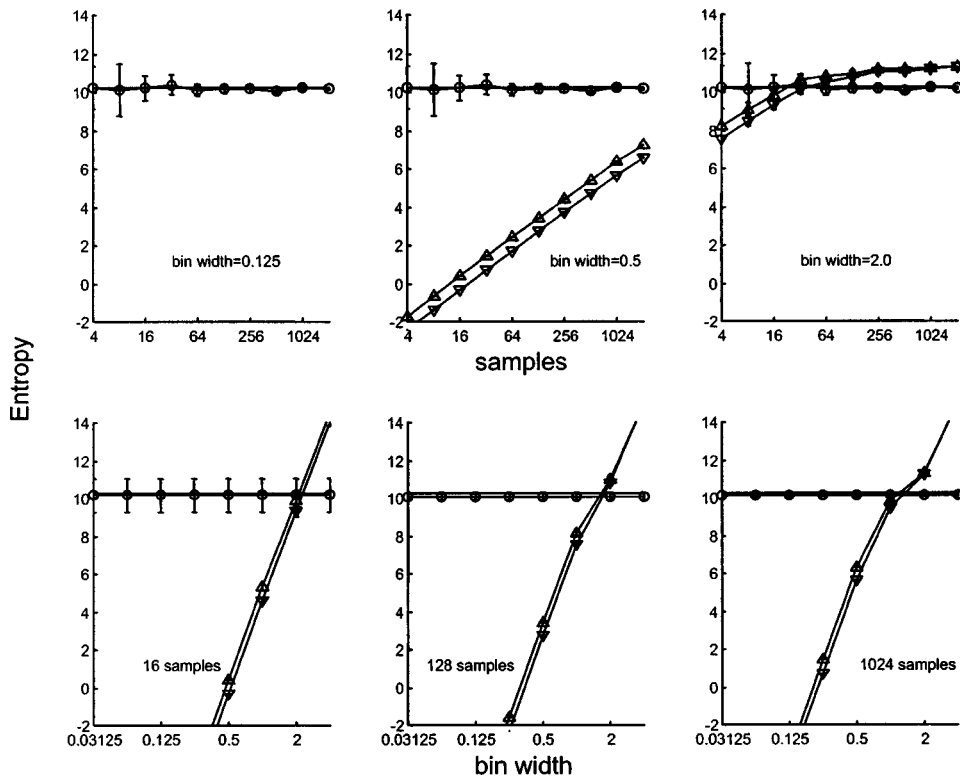


FIG. 3. Differential entropy of a five-dimensional Gaussian distribution (unit variance), estimated from a finite sample of data via binned and binless approaches. Correct value, ca. 10.236. Ten independent runs for each set of conditions. Display conventions as in Fig. 1.

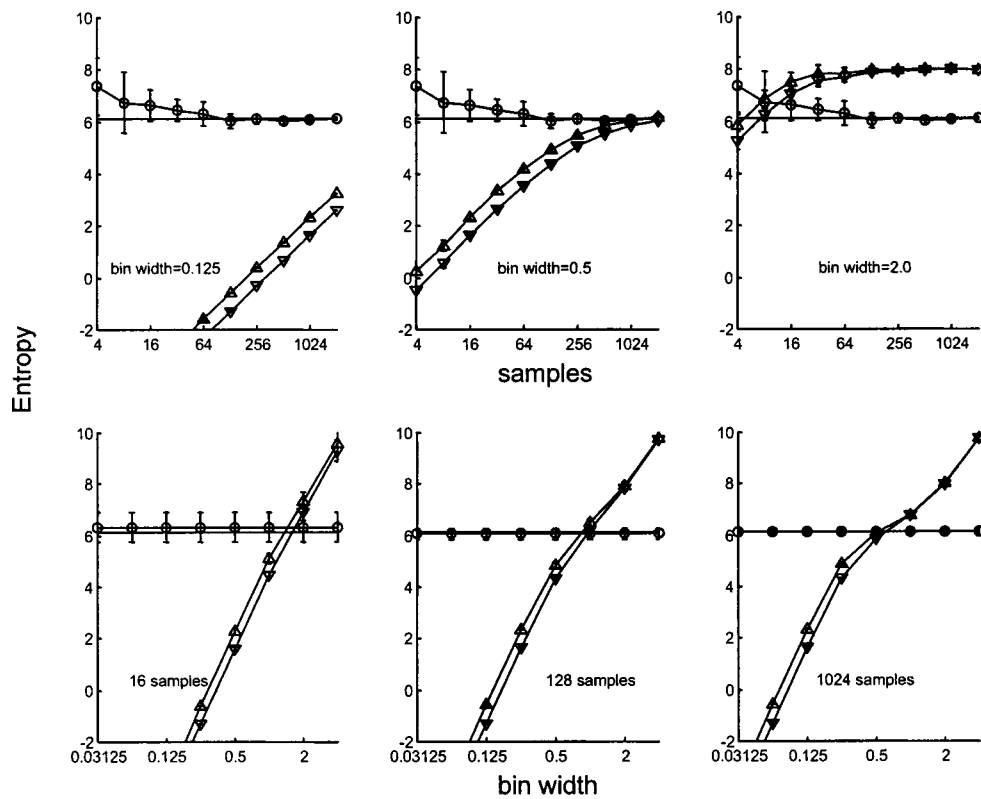


FIG. 4. Differential entropy of a three-dimensional Gaussian distribution whose variances along the orthogonal axes are in the ratio 1:10:100, estimated from a finite sample of data via binned and binless approaches. Correct value, ca. 6.141. Ten independent runs for each set of conditions. Display conventions as in Fig. 1.

estimators is restricted to small sample numbers ( $<16$ ), and is much less than the bias associated with the binned estimators, even after the latter have been “debiased.”

In sum, it appears that the binless estimates of a distribution’s differential entropy have significant advantages over binned estimates, particularly for high-dimensional distributions and when the size of the dataset is in the range 10–1000.

### Information estimates in a Euclidean space

We next consider estimation of information transmitted in the following setting: the input consists of a discrete set of  $S$  symbols  $a_1, \dots, a_s$ , presented with probabilities  $q_1, \dots, q_s$ . The resulting outputs  $x$  are characterized by conditional probability densities  $p_k(x) = p(x|a_k)$  in a Euclidean space of dimension  $r$ . In this context, the transmitted information is given by ([24], Sec. 2.4)

$$I = H_{\text{diff}} - \sum_{k=1}^S q_k H_{\text{diff}}(x|a_k), \quad (11)$$

where  $H_{\text{diff}}$  is the differential entropy for the (unconditional) density  $p(x)$ , and  $H_{\text{diff}}(x|a_k)$  is the differential entropy for the conditional density  $p_k(x) = p(x|a_k)$ . Substitution of Eq. (10) into Eq. (11) yields

$$I \approx \frac{r}{N} \sum_{j=1}^N \log_2 \left( \frac{\lambda_j}{\lambda_j^*} \right) - \sum_{k=1}^S \frac{N_k}{N} \log_2 \frac{N_k - 1}{N - 1}. \quad (12)$$

Here  $N_k$  is the number of presentations of the  $k$ th stimulus ( $N_k = q_k N$ ),  $\lambda_j$  is (as before) the minimum Euclidean distance between the observation  $x_j$  and any other observation, and  $\lambda_j^*$  is the minimum Euclidean distance between the observation  $x_j$  and any other observation elicited by the same stimulus. That is, Eq. (12) estimates information from the ratio between the minimum Euclidean distances between observations elicited by the same symbol, and the minimum Euclidean distances between observations elicited by all symbols.

### Information estimates for spike trains

We now consider how we can adapt this procedure to neural data, in which the outputs (responses) consists of spike trains. The main problem is that we cannot apply Eqs. (11) and (12) directly to neural data, since these equations assume that the spike trains are represented by quantities lying within a Euclidean space of a particular dimension  $r$ .  $n$  parameters are required to describe a spike train containing  $n$  spikes—effectively one for each spike time. Thus, the collection of all spike trains of finite duration can be considered to constitute a set of spaces, one of each dimension  $(0, 1, 2, \dots)$ , with the spike trains containing  $n$  spikes occupying the  $n$ -dimensional space. The binless approach outlined above can deal with the distribution of responses within each of these spaces, but it cannot deal with the overall distribution of responses across these spaces—since the latter is not characterized by any single dimension. This suggests that we

break the transmitted information into two kinds of contributions: one due to spike counts, and one due to spike times. We can then use binned estimates to determine the information carried by spike counts, and binless estimates to determine the information carried by spike times.

More formally, we write

$$I = I_{\text{count}} + \sum_{n=1}^{\infty} p(d(x)=n) I_{\text{timing}}(n), \quad (13)$$

where  $p(d(x)=n)$  is the probability that a response  $x$  contains exactly  $n$  spikes,  $I_{\text{count}}$  is the information carried by the number of spikes elicited by each stimulus, and  $I_{\text{timing}}(n)$  is the information carried by the distribution of spike times of all responses containing  $n$  spikes. The chain rule property of information [24] guarantees that the partitioning of information expressed by Eq. (13) is rigorously correct: information is unchanged by first considering how many spikes a response contains, and then, conditional on each particular number of spikes in a response, how those spikes are distributed in time. Note that this partitioning of information corresponds precisely to McFadden’s partitioning of the entropy of a point process into “numerical” and “locational” components [18].


Unfortunately, estimating  $I_{\text{timing}}(n)$  by embedding the spike trains containing  $n$  spikes into an  $n$ -dimensional space becomes impractical when  $n$  is large. Therefore, the above strategy must be modified in the following way. A maximal embedding dimension  $D$  is chosen. Each spike train of length  $n$  is then embedded (see below for details) as a point in a space of dimension  $r = \min(n, D)$ . This dimensional reduction for  $n > D$  may lead to a downward bias in the estimate of  $I_{\text{timing}}(n)$  (by the data processing theorem [24]), but as the numerical results will show (Fig. 6 and following), this downward bias is tolerable. Thus, given a choice of the maximal embedding dimension  $D$ , we estimate

$$I_{\text{timing}}(n) = \frac{r}{N(n)} \sum_{j=1}^{N(n)} \log_2 \left( \frac{\lambda_j}{\lambda_j^*} \right) - \sum_{k=1}^S \frac{N(n, a_k)}{N(n)} \log_2 \frac{N(n, a_k) - 1}{N(n) - 1}, \quad (14)$$

where  $r = \min(n, D)$  is the embedding dimension for  $n$ -element spike trains, the  $j$  summation is over all  $N(n)$  spike trains containing exactly  $n$  spikes, and  $N(n, a_k)$  is the number of trials in which a stimulus  $a_k$  elicits a response containing  $n$  spikes.  $\lambda_j$  and  $\lambda_j^*$  are the minimum Euclidean distances between  $x_j$  and all other responses that contain exactly  $n$  spikes ( $\lambda_j$ ) or those that contain exactly  $n$  spikes and are also elicited by the same stimulus as  $x_j$  ( $\lambda_j^*$ ). Note that the embedding process utilizes a single space for each dimension less than or equal to  $(D - 1)$ , but multiple spaces of dimension  $D$  (one for each value of  $n \geq D$ ). Each spike train is embedded into exactly one of these spaces, and the calculation of Eq. (14) is performed separately in each space. Other than the downward bias due to the dimensional reduction for  $n > D$ , it follows from the results of Kozachenko and

Leonenko [19] that the estimate of Eq. (14) is asymptotically unbiased and consistent. (As noted above, the results of Kozachenko and Leonenko [19] posit certain integrability conditions on the distributions of the embedded spike trains. These conditions are quite weak, and are guaranteed to hold if each multidimensional distribution of spike times is bounded and of finite support.)

Two quantities in Eq. (13) remain to be estimated.  $p(d(x)=n)$ , the probability that the number of spikes in a response is equal to  $n$ , can be estimated as  $N(n)/N$ .  $I_{\text{count}}$  can be estimated from a plug-in estimate based on  $N(n, a_k)$ , the number of trials in which a stimulus  $a_k$  elicits a response containing  $n$  spikes:



$$I_{\text{count}} \approx - \sum_{n=0}^{n_{\text{max}}} \sum_{k=1}^S \frac{N(n, a_k)}{N} \log_2 N(n, a_k) + \sum_{n=0}^{n_{\text{max}}} \frac{N(n)}{N} \log_2 N(n) + \sum_{k=1}^S q_k \log_2 q_k + I_{\text{bias}}, \quad (15)$$

where  $n_{\text{max}}$  is the maximum number of spikes in any response. The bias in the estimate of  $I_{\text{count}}$  can be estimated by standard methods for discrete entropy calculations. In the numerical examples, we will use two choices: the classical correction for entropy estimates [13–15]

$$I_{\text{bias}} \approx - \frac{(S-1)(n_{\text{max}}-1)}{2N \ln 2} \quad (16)$$

and the jackknife debiaser [16].

**The embedding**

To implement the above plan, we need to embed each  $n$ -element spike train as a point in a Euclidean space of dimension  $r = \min(n, D)$ . There are many reasonable choices for how to do this. However, the form of Eqs. (12) and (14) indicates that the estimated information will be insensitive to certain aspects of these choices. The information estimates depend only on the ratio of Euclidean distances to nearest neighbors. Thus, different embeddings related by a continuous distortion will lead to substantially the same estimate, provided that there are sufficiently many data points so that the distortion is relatively constant within the nearest-neighbor radius of each sample. This statement is nothing more than a reminder that information estimates are only likely to be valid if one has a sufficient amount of data to delineate the main features of the response probability distribution.

On the other hand, both common sense and the numerical examples of entropy calculations (Fig. 2 vs Fig. 4) suggest that the estimate is likely to be more efficient if each of the dimensions are relatively independent, and each contributes comparably to the overall scatter of the points. This motivates the following strategy, which we will use in the numerical examples that follow. First, the list of all spike times  $t_m$  (in all responses) is examined. A monotonic time-warping transformation [28,29]  $\tau(t)$  is applied so that the transformed

spike times  $\tau_m = \tau(t_m)$  are approximately equally spaced in the interval  $[-1, 1]$ . This can be accomplished by ordering all spike times serially, and assigning the  $j$ th spike to the time

$$\tau_j = -1 + 2 \frac{j - \frac{1}{2}}{M}, \quad (17)$$

where  $M$  is the total number of spikes. A set of spike times that are identical up to measurement precision are replaced by the mean of the values that would otherwise be assigned by the serial ordering and Eq. (17). The purpose of this transformation is to allow the creation of approximately independent coordinates via standard orthogonal polynomials.

The embedding coordinates are based on the Legendre polynomials  $P_h$ , which are orthogonal on  $[-1, 1]$ . In the usual normalization,

$$\frac{1}{2} \int_{-1}^1 P_h(\tau) P_k(\tau) d\tau = \frac{1}{2h+1} \delta_{h,k}. \quad (18)$$

The  $h$ th embedding coordinate  $c_h$  uses the  $h$ th Legendre polynomial to map a spike train  $x_j$  (containing  $n$  spikes at times  $t_{j_1}, \dots, t_{j_n}$ ) into the value

$$c_h(x_j) = \sqrt{2h+1} \sum_{k=1}^n P_h(\tau_k). \quad (19)$$

Together, the first  $r$  Legendre polynomials yield an embedding of a spike train  $x_j$  into a vector space of dimension  $r$ , namely, the point specified by the  $r$ -tuple  $c_1(x_1), \dots, c_r(x_j)$ . By virtue of the chosen normalization, if the spike times  $t_{j_1}, \dots, t_{j_n}$  within each spike train  $x_j$  were drawn at random from the pool of spike times, the mean-squared value of the  $h$ th coordinate of a spike train with  $n$  spikes would be  $n$ , because the transformed spike times  $\tau_m$  are approximately equally spaced in the interval  $[-1, 1]$ . Moreover (again assuming that spike times were drawn at random), coordinate values would be uncorrelated. This embedding thus fulfills the goal of creating approximately independent dimensions of approximately equal weight (if the spike times were independently drawn). However, we do not require the spike times or interspike intervals to be independent. If they are not independent, the above embedding nevertheless suffices to apply Eq. (14). The estimation procedure remains valid, but may suffer a loss of efficiency. Despite the possible loss of efficiency, the numerical examples of Figs. 8 and 9 show that the present approach retains its advantages when interspike intervals are strongly correlated.

We emphasize that the above embedding is based on the transformed spike times  $\tau_m$ . Thus, the resulting information estimates will not depend on the actual spike times, but only on their order (within and across spike trains). At first this might seem counterintuitive. However, the warping is, after all, an invertible topological transformation on the spike trains, which therefore cannot affect the information content. As a computational device, forcing the spike density to be uniform in the transformed time helps to make the embedding coordinates approximately independent, and thus makes the information estimate more robust.



As indicated above, the entire estimation procedure is parametric in a choice of a maximal embedding dimension  $D$ . Too small a choice for  $D$  will downwardly bias information estimates since it leads to a loss of detail about responses for  $n > D$ . On the other hand, too large a choice of  $D$  will also lead to a downward bias of the information estimate, but for a different reason: the asymptotic regime of the binless estimator is not reached. When the dimension of the embedding space is large, most of the embedded spikes are close to the boundary of the region of data. For points near the boundary, only a portion of the surrounding solid angle can possibly contain a nearest neighbor. The data-independent term  $S_r$  of Eq. (9) is too large to take into account this edge effect, and consequently, differential entropy estimates [such as Eq. (10)] are upwardly biased. This edge effect is larger for the conditional (within-stimulus) estimate of differential entropy  $H_{\text{diff}}(x|a_k)$  than for the unconditional across-stimulus estimate,  $H_{\text{diff}}$ , since there are fewer data points in the former estimate. Consequently, the ultimate effect on the information estimate [Eq. (11)] is a downward bias before the asymptotic regime is reached.

Thus, since we anticipate downward biases both for large  $D$  and for small  $D$ , the systematic way to proceed is to perform the above calculations parametrically in  $D$ , and to take the maximal value of the resulting information estimates is taken as the final estimate of  $I$ . The examples below (Figs. 6–9) show that this strategy is indeed practical, and that information estimates typically reach their maximal value for  $D=2, 3$ , or 4. We also note that terms in Eq. (14) may be undefined either because there are two spike trains that are embedded at precisely the same point (and thus, the nearest neighbor has a Euclidean distance of 0), or because there are no nearest neighbors. These eventualities can be handled as described in the Appendix.

#### Numerical examples: Information estimates

We illustrate the above approach with some calculations based on simulated data. The first simulation (Figs. 5 and 6) considers responses to five stimuli that produce Poisson spike trains differing in mean rate. Figure 5(a) illustrates information estimates obtained via the binless procedure [Eqs. (13)–(15)]. With increasing sample size, the debiased contribution from spike count alone,  $I_{\text{count}}$  [Eq. (15)], converges to the correct value. Prior to convergence, the classical bias estimate (down triangles) underestimates the correct value, while the jackknife bias estimate (up triangles) overestimates the correct value. Error bars represent one standard deviation of the range of values calculated in multiple independent simulations, and many error bars overlap.

In this simple simulation consisting of Poisson data, the contribution of spike timing (and thus,  $I_{\text{timing}}$ ) to the information is zero. Nevertheless it is useful to assess the bias and scatter of estimates of  $I_{\text{timing}}$  [Eq. (14)]. As seen in Figure 5(a), the estimates of  $I_{\text{timing}}$  indeed add considerable scatter. The classical debiaser (squares) typically underestimates the correct value, while the jackknife debiaser (diamonds) typically overestimates the correct value.

Figure 5(a) compares these information estimates to those obtained by estimating  $I_{\text{timing}}$  in a binned fashion from the embedding of Fig. 5(a). The estimates have greater precision

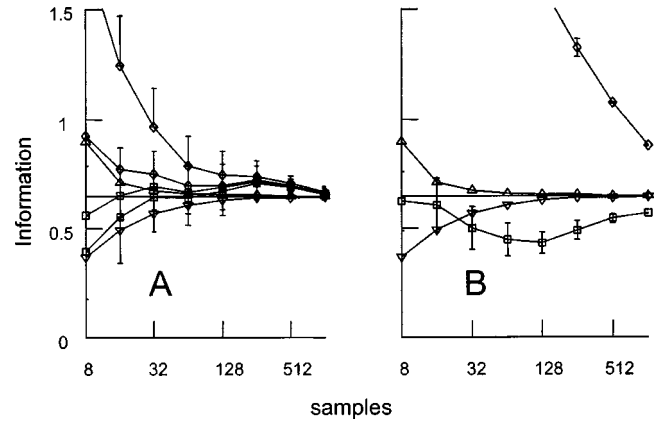


FIG. 5. (a) Information estimates from simulated Poisson data, using Eqs. (13)–(15), with a maximal embedding dimension  $D$  of 2. Simulated spike trains had duration 1 s and mean rates 2, 4, 6, 8, and 10 Hz. (This is identical to the simulation of Figs. 4A–D of Victor and Purpura [17].) 2048 such spike trains were generated in response to each stimulus, and we estimated information from datasets consisting of 8, 16, 32, . . . , 1024 examples of each response. Solid horizontal line: correct value. Triangles: contribution from spike count alone,  $I_{\text{count}}$  [Eq. (15)] as corrected by the classical bias estimate (down triangles) and the jackknife bias estimate (up triangles). Lines marked by squares and diamonds indicate total information estimate [Eq. (13)], adjusted by the classical and jackknife bias estimates, respectively. There are two traces marked by each set of symbols, corresponding to the strategy for handling terms in Eq. (14) that are undefined due to singletons (see the Appendix). The upper trace (with error bars extending only upwards) corresponds to estimates generated by considering singletons maximally informative. The lower trace (with error bars extending only downwards) corresponds to estimates generated by considering singletons maximally uninformative. Error bars represent one standard deviation of the range of values calculated in multiple independent simulations, and many error bars overlap. (b) Estimates of information derived from binning the embedded spike trains to obtain  $I_{\text{timing}}$ , rather than Eq. (14). Down and up triangles:  $I_{\text{count}}$  with the classical and jackknife bias corrections, as in Fig. 5(a). Squares and diamonds: total information estimates, corrected by the two kinds of bias estimates applied to the binned data. Since singletons are not treated as special cases, each kind of bias estimate leads to only one estimate (plotted with a double-sided error bar), rather than the upper and lower estimates of Fig. 5(a). The embedding dimension  $D$  is 2 [as in Fig. 5(a)]; the bin width is 1. Here and in subsequent figures, the numbers along each abscissa refer to the number of samples  $N$  that are used in the information estimates.

than those of the binless strategy. However, they have substantially less accuracy in the 32–256 sample range for the classical debiaser, and across the entire range for the jackknife debiaser. Over most of the range of sample number, the improved accuracy of the binless estimates [Fig. 5(a)] compared to the binned estimates [Fig. 5(b)] more than compensates for the inferior precision of the binless approach.

Figure 6 considers a wider set of information estimates for this simulation. The first three rows extend the comparisons of Fig. 5 to a range of maximal embedding dimensions  $D=1, 2$ , and 3. The choice of  $D$  has relatively little effect on the binless information estimates. However, for binned esti-

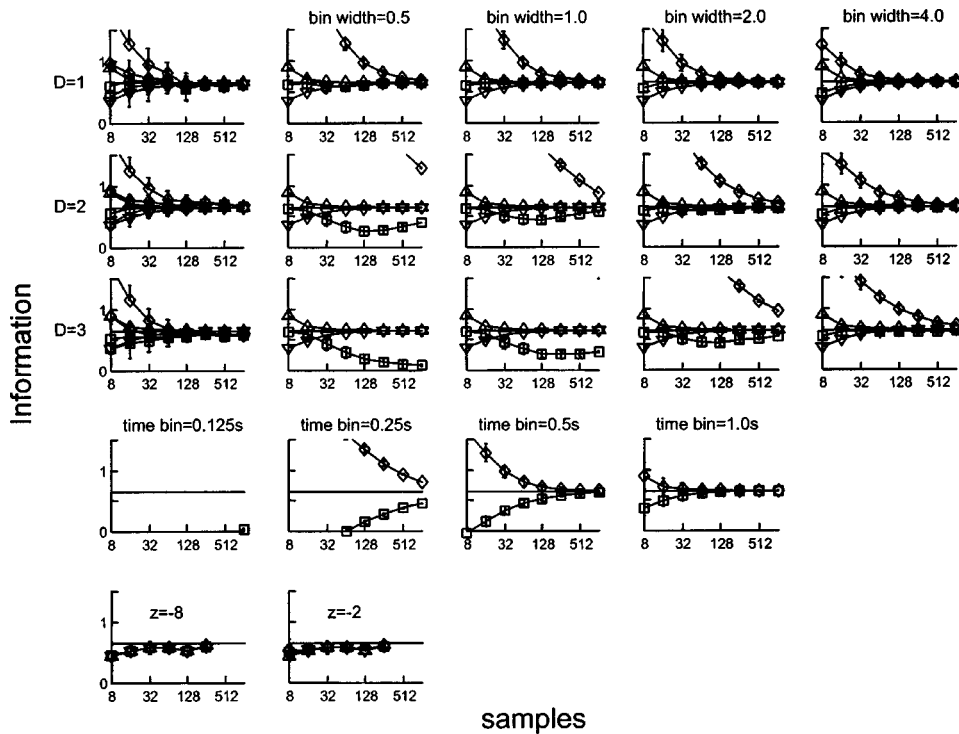


FIG. 6. A broader exploration of information estimates for the simulation of Fig. 5. First three rows: comparisons of information estimates obtained with the unbinned estimator and maximal embedding dimension  $D = 1, 2,$  and  $3$  [first column, displays conventions as in Fig. 5(a)] or with binned estimators [remaining columns, display conventions as in Fig. 5(b)]. Fourth row: information estimates obtained via standard time binning of the raw spike trains; bin widths  $0.125, 0.25, 0.5,$  and  $1$  s. The analysis based on  $1$  s bins reflects spike counts only, since the spike trains are  $1$  s in duration. Bias corrections and range of estimators displayed as in Fig. 5(b); note that estimates are largely off-scale for a bin width of  $0.125$  s. Fifth row: information estimates obtained via a binless approach based on metric-space embeddings [17]. The spike count contribution (up and down triangles, for the two kinds of bias corrections, here superimposed) is calculated from the “spike count” metric of [17]. The total information (diamonds and squares, for the two kinds of bias corrections, here superimposed) is calculated from the optimum “spike time” metric of [17]. The two graphs reflect two choices of the clustering exponent  $z$ , and calculations are limited to sample sizes of  $256$  or less because of computational constraints. For further details, see text and Victor and Purpura [17]. Bias corrections and range of estimators displayed as in Fig. 5(b).

mates,  $D$  and bin width have large effects. As bin width decreases, the upward bias of the jackknife estimate becomes large. This effect is magnified at higher maximal embedding dimensions  $D$ , so much so that the estimate is off-scale for much of the range of sample size.

The fourth row of Fig. 6 shows information estimates calculated by direct binning. That is, rather than embedding spike trains into a vector space by a procedure such as Eq. (19) and performing estimates (either binned or unbinned) on the embedded data, each response is represented as a sequence of integers corresponding to the number of spikes that occur in each of several time bins. This is essentially the “direct” method of Strong *et al.* [12]. Precision is better than for the binned estimates derived from embedding (top three rows, second through fifth columns), but accuracy is dramatically worse, especially for time bins of  $0.25$  s or less.

The final row in Fig. 6 shows information estimates calculated via another unbinned approach, the metric-space method of Victor and Purpura [17]. In this approach, spike trains are embedded in a metric space via a range of different candidate “spike time metrics,” parametrized by a quantity  $q$  that indicates the relative importance of spike timing and spike count. Clustering (parametrized by an averaging expo-

nent  $z$ ) of responses is performed directly in the metric space. The information estimate is based on how faithfully the response clusters reflect the original stimuli, at the optimal value of  $q$ . As shown here, the estimators have a precision that is comparable to those derived from binning, but, as previously noted [17], there is a modest downward bias, even for large sample sizes.

The simulation of Figs. 5 and 6 shows that binned entropy estimates, even when debiased, depend strongly on maximal embedding dimension  $D$  and bin width. Since these simulations are based on Poisson trains (for which there is no contribution of spike timing *per se*), an accurate estimate of information can be obtained by choosing a large bin width. For non-Poisson spike trains, the contribution of spike timing will only be evident when the bins are fine enough to capture the informative temporal structure of the spike train. As a consequence, it may be difficult to choose a bin small enough to capture the temporal structure, and large enough to eliminate bias due to limited data size. In this regime, the binless estimators are expected to have a considerable advantage. This is illustrated in Fig. 7.

Figure 7 shows the analysis of simulated spike trains generated by a gamma process whose coefficient of variation is

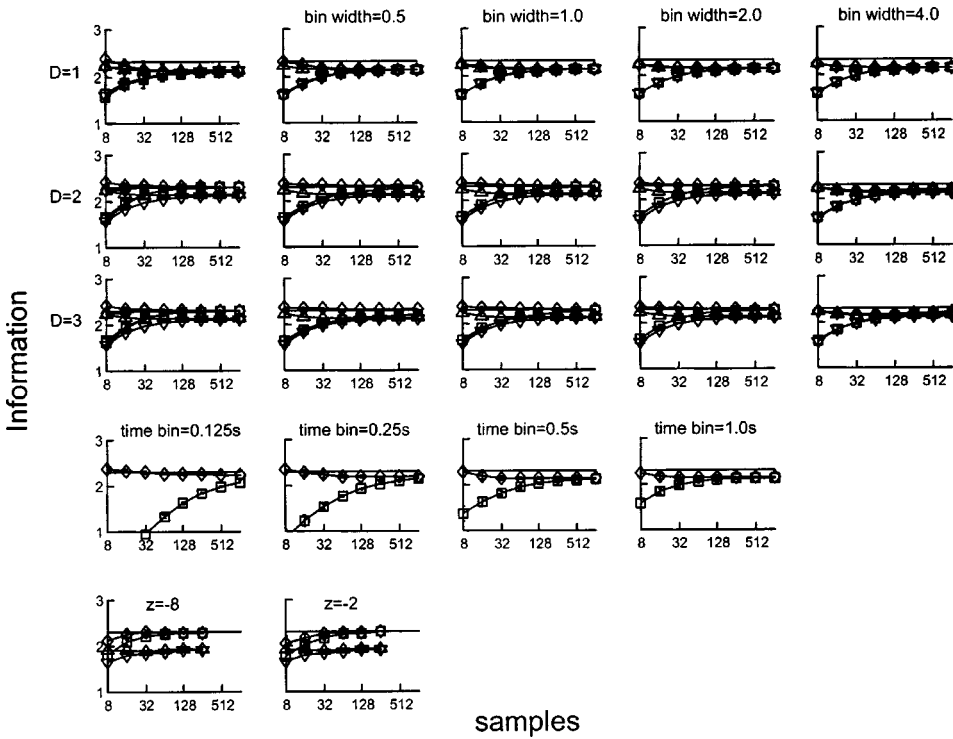


FIG. 7. Information estimates for simulated data consisting of highly regular spike trains. Spike trains are generated by a gamma process of order 64 (and thus the interspike intervals have a coefficient of variation of 0.125). The mean firing rates and other details are otherwise identical to the simulation of Figs. 5 and 6. (This is identical to the simulation of Figs. 4E–H of Victor and Purpura [17].) The solid line (no symbols) indicates the information estimated by the binless approach with a sample size of 4096 and  $D=4$  (the value of  $D$  in  $\{1, \dots, 6\}$  that provided the maximum information). Display conventions otherwise as in Fig. 6.

0.125. Because firing rates are more regular, the contribution of spike counts to information is greater than in Fig. 5. Moreover, the regularity of the spike trains *per se* should provide an additional but modest contribution to the information. This is because (given the extreme regularity of the spike trains) even a single short interval is unlikely at the lower firing rates. This increment,  $\approx 0.2$  bits, is evident for the estimates based on  $D=2$  or 3, both for the binned and unbinned estimates. (Failure to identify this incremental information for  $D=1$  is consistent with the fact that the timing contribution reflects information carried by pairs of spikes.) Note that for the binned estimates based on embedding, this incremental information can only be seen for a bin width less than or equal to 2. That is, there is a very narrow range of bins (ca. 2) that is both large enough to avoid a large bias for Poisson data (Fig. 6), and small enough to capture the additional information in the regularity of the spike trains (Fig. 7). For information calculated by direct binning, the situation is worse (fourth row of Fig. 7). Only a time bin width less than or equal to 0.25 s begins to capture the information in the regularity of the spike trains, but these time bins lead to unacceptable bias for Poisson data (Fig. 6). The spike metric estimators (last row in Fig. 7) provide an acceptable estimate, but they are negatively biased for the Poisson trains (Fig. 6).

Figure 8 considers responses that are inhomogeneous Poisson processes with identical mean firing rates. Consequently, information is carried only by spike timing. For  $D=1$ , neither the binless estimator nor the binned estimator (based on embedded spike trains) captures the full information. This is consistent with the fact that the response geometry is that of a circle (of phases), and any one-dimensional projection entails ambiguity. With  $D \geq 2$ , the binless estimators closely approximate their asymptotic value, even for a

sample size of less than or equal to 64. The binned estimators for  $D \geq 2$  straddle this asymptotic value widely, but do not approach it very closely, even for 1024 samples of each response type. Direct time binning (next to last row) results in estimates that are either downwardly biased because the bin size is too large (bin width greater than or equal to 0.5 s), or far from the asymptotic value because of limited data (bin width less than or equal to 0.25 s). Metric-space estimates (last row) converge more rapidly but have some downward bias.

Figure 9 considers inhomogeneous Poisson spike trains that differ in mean rate and in their transient firing envelopes. Thus, information is carried both in the spike counts and in spike timing, and is not uniformly distributed in time. As is typical of cortical responses [4], the systematic difference between the times of the onsets of the transients is comparable to the typical interspike interval. This represents a particularly severe challenge for binned approaches. Other than a nonzero contribution of  $I_{\text{count}}$ , the behavior of the estimators is generally similar to that of Fig. 8. Binless estimates appear to achieve a maximum with  $D=3$  or 4 and asymptote with a sample size of less than or equal to 64, while the binned estimators are highly sensitive to the choice of bin size, and have not reached asymptotic behavior even with a sample size of 1024.

## DISCUSSION

### The hybrid topology of spike trains

Entropy and information are usually defined and estimated for probability distributions on a discrete set or on a Euclidean space. The space of spike trains has a distinct hybrid topology [18], and this has implications for the estimation of information. Spike trains have a discrete character,

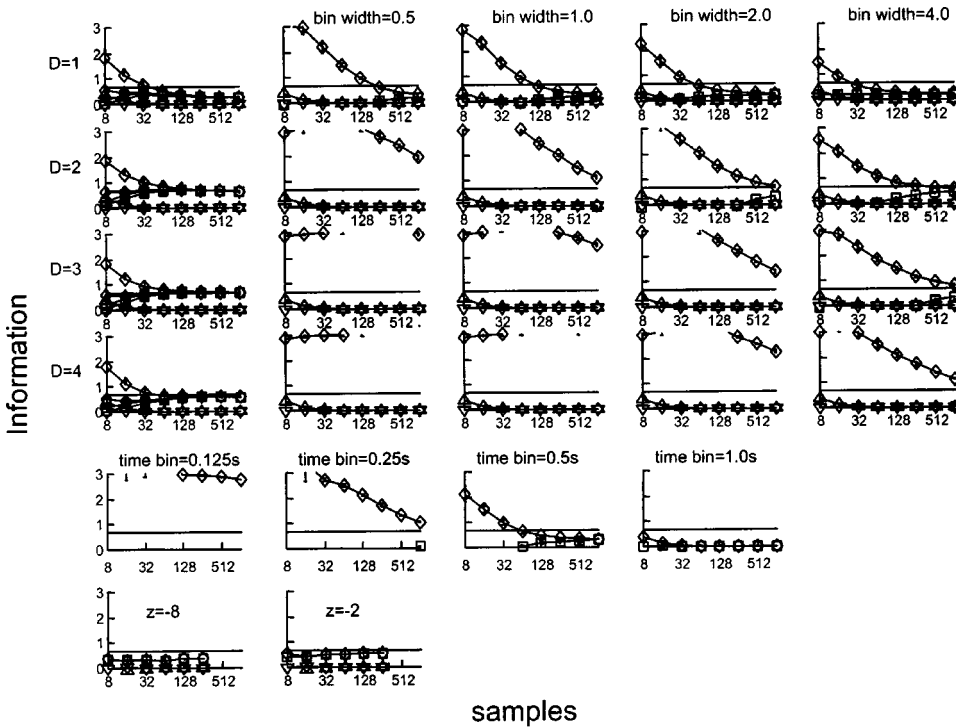


FIG. 8. Information estimates for simulated data consisting of sinusoidally modulated inhomogeneous Poisson trains (eight equally spaced phases, mean rate 10 impulses/s, modulation depth 5 impulses/s, duration 1 s). The solid line (no symbols) indicates the information estimated by the binless approach with a sample size of 4096 and  $D=3$  (the value of  $D$  in  $\{1, \dots, 6\}$  that provided the maximum information). First four rows: estimates based on embedding in dimensions 1, 2, 3, and 4, followed by the binless estimator or binned estimators of a range of bin widths. Fifth row: estimates based on direct binning. Sixth row: estimates based on spike metrics. Display conventions otherwise as in Fig. 6.

because the number of spikes in any spike train must be an integer. Spike trains also have a continuous character, owing to the continuous nature of time: two spike trains may be considered to be “close” to each other if they have the same number of spikes, and the corresponding spikes occur at nearly identical times. Reducing a spike train to a discrete series of integers (via binning) destroys this topology, in that small shifts in the time of a spike (that cause a spike to cross a bin boundary) result in as much of a change as moving a spike to an arbitrarily distant bin. One of the appeals of information measures is that they are independent of smooth, invertible, transformations of the underlying space. However, they are *not* independent of transformations that destroy the topology. Thus, since formal information is only preserved when the topology of the response space is preserved, approaches that ignore the continuous aspects of the topology might not even converge to the correct answer. The present approach both respects and exploits this natural hybrid topology of spike trains, and is thus more likely to be robust and efficient than procedures that ignore it.

The numerical examples presented above indicate that these theoretical considerations are highly relevant for the typical size of experimental datasets (Figs. 5–9). Of the estimators considered, the binless approach provides the most robust and rapidly converging estimators for information in spike trains. Direct binning of the spike trains provides the least useful estimators. Estimators that use an embedding that reflects the hybrid topology but then calculate information by binning the embedded data have an intermediate level of performance. The benefit of exploiting the underlying topology of a distribution applies not only to information calculations, but also to simple estimates of entropy (Figs. 1–4).

### Variations

Within the framework of binless estimators applied to spike trains, there are a number of reasonable variations on the particular implementation proposed here. The time warping transformation [Eq. (17)] is not an essential step: it improves convergence but entails a modest penalty in computation time and the scatter of the information estimates.

The choice of embedding functions (here, the Legendre polynomials) is a generic one, not necessarily the most appropriate for all situations. Fourier coefficients might be particularly appropriate for periodic stimuli and Laguerre polynomials might be particularly appropriate for responses that have initial transients. Principal components are another reasonable choice. While stratification by spike count is critical in deriving a rigorously valid estimator, this step is not a prerequisite for estimators that have practical utility. For example, one could lump together all spike trains regardless of count, and simply add an embedding coordinate equal to the spike count, ignoring the fact that the distribution is discrete. A strategy of this sort expresses the notion that spike trains that differ in just one spike should be considered “similar,” although each spike is a discrete event. If all spike trains contain sufficient spikes, one anticipates only small biases due to this discretization, and a gain in the precision of the estimator since all spike trains are pooled. Numerical studies suggest that this regime is reached once there are three or four spikes in each train. Hybrid schemes based on lumping together spike trains once the spike count exceeds some criterion may also fill a practical niche.

The basic binless information estimate [Eq. (12)] and the entropy estimate [Eq. (10)] that underlies it are based on the Euclidean distance to the nearest neighbor. Analogous estimators can be constructed based on  $k$ th-nearest-neighbor dis-

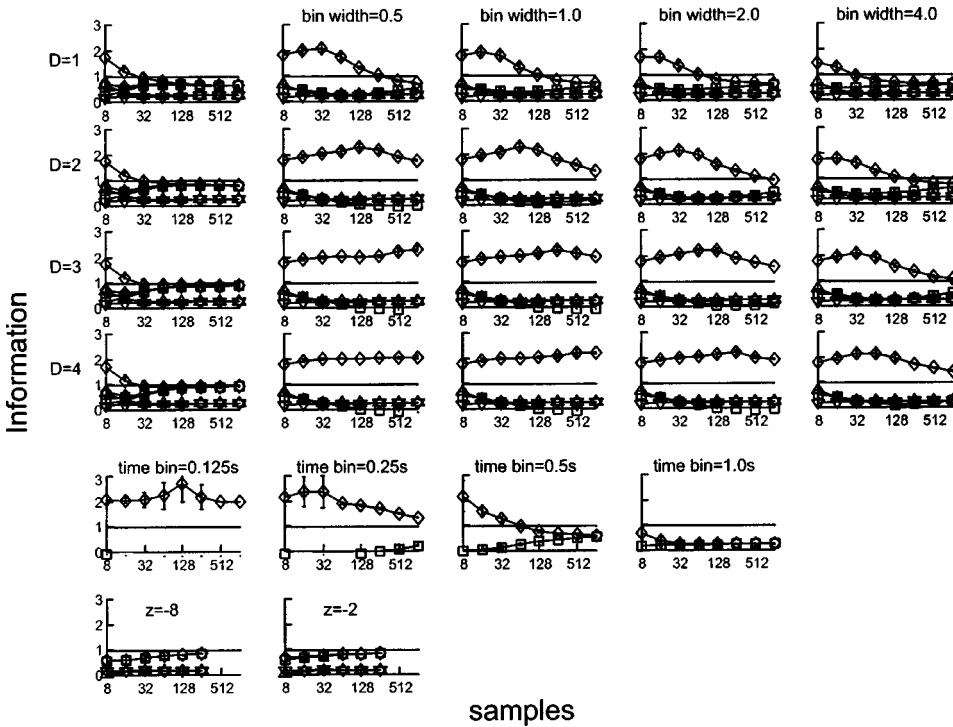


FIG. 9. Information estimates for simulated data consisting of transiently modulated Poisson trains. Each of the four stimuli elicit responses with a basal firing rate of 10 impulses/s and an exponential transient. The onsets of the transients are at 0.10, 0.15, 0.20, and 0.25 s; the heights of the transients are 128, 48, 16, and 4 impulses/s, and their time constants are  $\frac{1}{16}$ ,  $\frac{1}{8}$ ,  $\frac{1}{4}$ , and  $\frac{1}{2}$  s (The larger responses occur earlier and decay more rapidly.) The solid line (no symbols) indicates the information estimated by the binless approach with a sample size of 4096 and  $D=4$  (the value of  $D$  in  $\{1, \dots, 6\}$  that provided the maximum information). Display conventions otherwise as in Fig. 8.

tances [25]. These have been studied in detail, but the exact result of Kozachenko and Leonenko [19] was not demonstrated beyond dimension 1 [30]. Numerical studies (not shown) suggest that this modification generally leads to estimates that converge less rapidly as a function of the number of samples, but have a somewhat lower variance. Since this strategy would also need to handle many more special cases than just the “singleton” situation (see the Appendix), it is unlikely that it would provide a significant practical advantage.

### Comparison to other approaches

The proposed approach applies to neural responses of limited duration, elicited by single stimuli. Another typical experimental situation is that of prolonged responses elicited by rapid presentation of multiple stimuli (typically in a pseudorandom sequence). In those situations, existing methods (the reconstruction method of Bialek *et al.* [21], and the direct method of Strong *et al.* [12] are clearly appropriate. The rapid sequential presentation of stimuli in the latter approach acts to destroy whatever temporal structure might be generated by the neural response to a single transiently presented stimulus. Thus, it makes sense that bin-based methods work well, and the topology of the response space is less crucial.

The goals of these two kinds of experiments are different. In the case considered here, the intent is to determine how faithfully a neuron can transmit information about a particular stimulus set considered relevant to the neuron’s function (such as a set of gratings of various contrasts). In the pseudorandom sequence approach, the intent is to determine the maximal amount of information that the neuron can transmit.

We previously introduced [17] another binless approach based on a metric-space embedding. Similar to the present approach, the metric-space approach explicitly considers the

topology of spike trains. Also, it is designed for the limited-duration situation and not to the pseudorandom sequence situation. As shown here (Figs. 6–9), the spike metric method converges at least as rapidly as the present method, but is somewhat downwardly biased. These approaches have somewhat different goals. By virtue of the work of Kozachenko and Leonenko [19], the present method is demonstrably unbiased, and, with sufficient data, will converge to the amount of information present. However, it provides little insight into *how* this information is carried. In contrast, the metric-space method is based on comparing various families of biologically motivated (but stereotyped) metrics. Thus, it is capable of determining which aspects of a spike train are informative, but there is no guarantee that it will extract the maximal amount of information that is present.

### Extensions

Although we have focused on the estimation of information carried by a single neuron’s spike trains elicited by a discrete set of stimuli, the proposed approach is not limited to this setting. For example, if the stimuli  $a$  are drawn from a continuous distribution, Eq. (11) can be replaced with

$$I = H_{\text{diff}}(a) + H_{\text{diff}}(x) - H_{\text{diff}}(a, x), \quad (20)$$

where  $H_{\text{diff}}(a, x)$  is the differential entropy of the joint distribution of  $a$  and  $x$ , and  $H_{\text{diff}}(a)$  and  $H_{\text{diff}}(x)$  are the differential entropies of the marginal distributions of  $a$  and  $x$ , respectively. Binless estimators for  $H_{\text{diff}}(a, x)$  along the lines of Eqs. (12) and (14) can be constructed in the product space of the domain of  $a$  and the embedding of  $n$ -element spike trains, and binless estimators for the marginal entropies  $H_{\text{diff}}(a)$  and  $H_{\text{diff}}(x)$  follow from the projection of this product space onto the domain of  $a$  and the spike train embed-

ding. The key quantities in this estimator are the ratio of the Euclidean nearest-neighbor distances in the product space to the Euclidean nearest-neighbor distances in each of the two projections.

It is also straightforward to extend this approach to multineuronal responses, at least in principle. As in the single-neuron case, there is a contribution  $I_{\text{count}}$  due to spike counts alone, but this stratification would need to be performed independently for the spike count associated with each of the  $m$  neurons. Each neuron's spike trains are then independently embedded. Within each such response subset, the estimate (14) can then be used to determine the additional contribution of  $I_{\text{timing}}$ . Since there would be many response subsets (one for each  $m$ -tuple), it might be useful (though not rigorously justifiable) to lump together subsets with similar  $m$ -tuples.

Finally, it is straightforward to apply this approach to responses that are continuous functions of time. For such data sets, the stratification stage is superfluous, and the estimate (12) can be used directly, once an embedding of the data into a low-dimensional space is accomplished. Such an embedding could be accomplished in several ways, including orthogonal functions or principal components.

## APPENDIX

### Two implementation details

In a practical implementation, undefined terms may arise in Eq. (14), the estimate of  $I_{\text{timing}}(n)$ , via two routes: “zero distances” and “singletons.” A Euclidean distance of zero between two spike trains can arise either because of finite measurement accuracy of the spike times, or because the embedding procedure happens to map distinct spike trains to the same point. Another problem is that some stimuli may elicit only one spike train containing  $n$  spikes. These “singleton” spike trains will have no nearest neighbors within their stimulus class from which to calculate  $\lambda^*$  in Eq. (14). We now describe how we deal with these eventualities.

To deal with “zero distances,” the spike trains containing exactly  $n$  spikes that are unique (i.e., at a nonzero Euclidean distance from all other spike trains) are placed into a set  $C_n$ . The remaining spike trains, each of which is at a distance of zero from at least one spike train, are grouped into maximal disjoint subsets  $Z_{n,1}, Z_{n,2}, \dots, Z_{n,b(n)}$  of spike trains, with each of these subsets containing spike trains at a Euclidean distance of zero from each other. That is,  $C_n$  contains the spike trains that are all distinct while each  $Z_{n,m}$  is a set of spike trains that appear to be identical. Information  $I_{\text{timing}}(n)$  related to the timing of spikes within the  $n$ -spike trains can now be subdivided into a discrete component, corresponding to the partition of spike trains into disjoint sets  $C_n, Z_{n,1}, Z_{n,2}, \dots, Z_{n,b(n)}$ , and a continuous component, corresponding to distinctions within each of these subsets. However, since spikes within the subsets  $Z_{n,1}, Z_{n,2}, \dots, Z_{n,b(n)}$  all appear identical, the only continuous contribution to information comes from  $C_n$ . In sum,

$$I_{\text{timing}}(n) = I_{\text{partition}}(n) + \frac{N(x \in C_n)}{N(n)} I_{\text{continuous}}(n), \quad (21)$$

where  $I_{\text{partition}}$  is the information associated with the discrete partitioning of  $n$ -spike trains into the disjoint subsets  $C_n, Z_{n,1}, Z_{n,2}, \dots, Z_{n,b(n)}$ ,  $I_{\text{continuous}}(n)$  is the binless estimate [Eq. (14)] of transmitted information restricted to  $n$ -spike trains within the set  $C_n$ , and  $N(x \in C_n)$  is the number of spike trains in  $C_n$ . [If no spike trains are at a distance of zero from each other, then only  $C_n$  is nonempty, and  $I_{\text{timing}}(n) = I_{\text{continuous}}(n)$  and  $I_{\text{partition}}(n) = 0$ .]

With  $C_n$  written as  $Z_{n,0}$  for notational convenience,  $I_{\text{partition}}(n)$  can be estimated by a plug-in estimate

$$I_{\text{partition}} \approx - \sum_{m=0}^{b(n)} \sum_{k=1}^S \frac{N(x \in Z_{n,m}, a_k)}{N(n)} \log_2 N(x \in Z_{n,m}, a_k) + \sum_{m=0}^{b(n)} \frac{N(x \in Z_{n,m})}{N(n)} \log_2 N(x \in Z_{n,m}) + \sum_{k=1}^S \frac{N(n, a_k)}{N(n)} \log_2 \frac{N(n, a_k)}{N(n)} + I_{\text{partition, basis}}, \quad (22)$$

where  $N(x \in Z_{n,m}, a_k)$  is the number of observations of a spike train in  $Z_{n,m}$  elicited by a stimulus  $a_k$ .  $I_{\text{partition, bias}}$  is the bias estimate for this discrete partitioning. Since there are  $b(n) + 1$  response categories  $C_n, Z_{n,1}, Z_{n,2}, \dots, Z_{n,b(n)}$ , the classical estimate [analogous to Eq. (16)] for this bias is

$$I_{\text{partition, bias}} \approx - \frac{[s(n) - 1]b(n)}{2N(n) \ln 2}, \quad (23)$$

where  $s(n)$  is the number of stimuli that elicit spike trains with  $n$  spikes.

The strategy for dealing with singletons builds on the above idea. Singletons arise when one or more spike trains, say  $x_{j_1}, \dots, x_{j_u}$ , are the sole examples in  $C_n$  of the responses to their respective stimuli  $a_{k(j_1)}, \dots, a_{k(j_u)}$ . In this case, the nearest-neighbor distance  $\lambda_{j_i}^*$  from each  $x_{j_i}$  to another spike train elicited by the same stimulus is undefined. This eventuality is a direct consequence of having a limited amount of data, so our strategy is based on considering two ways of extrapolating to what the dataset might plausibly consist of, had additional data been available. One extreme is that additional observations of responses to each stimulus  $a_{k(j_i)}$  would only yield responses that coincide with the observed singleton  $x_{j_i}$ . In this case, each  $x_{j_i}$  should be considered to constitute a singleton set along with the above  $Z_{n,m}$ .  $I_{\text{continuous}}$  is then recomputed according to Eq. (14) with the  $u$  singletons removed from  $C_n$ , and  $I_{\text{timing}}(n)$  is then recomputed according to Eqs. (21)–(23), with the list of zero-distance sets  $Z$  used in the calculation of  $I_{\text{partition}}$  augmented by the  $u$  singletons  $\{x_{j_i}\}$ . The other extreme is that additional observations would indicate that each of the  $u$  singleton responses,  $x_{j_i}$  is completely uninformative. In this case,  $I_{\text{timing}}$  is then recomputed with the singletons removed from  $C_n$ , and  $N(x \in C_n)$  is reduced by  $u$  in Eqs. (21) and (22), but the list of zero-distance sets is left unchanged. With reasonably large datasets, the two extremes yield very similar values for  $I_{\text{timing}}(n)$ , as the numerical examples have shown.

### A chain rule for bias estimates

To deal with the eventualities of “zero spikes” and “singletons,” the partitioning of spike trains according to spike count  $n$  is followed by a second partitioning into the subsets  $Z_{n,m}$ . Both stages of partitioning contribute a discrete component to the overall information. The contribution of the second stage contains a separate component,  $I_{\text{partition}}(n)$ , corresponding to each number of spikes  $n$ . The chain rule for information [24] implies that it is equivalent to add the information associated with each of these two stages or to consider the entire partitioning as a single step. A simple counting argument shows that the chain rule also extends to bias estimates, either via the classical bias correction

or via the jackknife. That is, the bias estimate does not depend on whether the bias is estimated at each of the two stages and then added (as we have done here), or by considering the two stages of partitioning as a single step.

### ACKNOWLEDGMENTS

This work was presented in part at the 2001 meeting of the Society for Neuroscience, and was supported by NIH NEI EY9314. The author thanks Bruce Knight, Peter Latham, Partha Mitra, Rodrigo Quian Quiroga, Peter Grassberger, Satish Iyengar, and especially Liam Paninski for helpful discussions.

- 
- [1] F. Rieke, D. Warland, R. R. de Ruyter van Steveninck, and W. Bialek, *Spikes: Exploring the Neural Code* (MIT, Cambridge, MA, 1997).
- [2] F. Théunissen and J. P. Miller, *J. Comput. Neurosci.* **2**, 149 (1995).
- [3] M. N. Shadlen and W. T. Newsome, *J. Neurosci.* **18**, 3870 (1998).
- [4] T. J. Gawne, *Exp. Brain Res.* **133**, 293 (2002).
- [5] W. Softky, *Neuroscience* **58**, 13 (1994).
- [6] K. Sen, J. C. Jorge-Rivera, E. Marder, and L. F. Abbott, *J. Neurosci.* **16**, 6307 (1996).
- [7] C. M. Gray, P. Konig, A. K. Engel, and W. Singer, *Nature (London)* **338**, 334 (1989).
- [8] M. Meister, L. Lagnado, and D. A. Baylor, *Science* **270**, 1207 (1995).
- [9] C. Shannon, *Bell Syst. Tech. J.* **27**, 379 (1948).
- [10] M. W. Oram, M. C. Wiener, R. Lestienne, and B. J. Richmond, *J. Neurophysiol.* **81**, 3021 (1999).
- [11] S. B. Laughlin, R. R. de Ruyter van Steveninck, and J. C. Anderson, *Nat. Neurosci.* **1**, 36 (1998).
- [12] S. P. Strong, R. Koberle, R. R. de Ruyter van Steveninck, and W. Bialek, *Phys. Rev. Lett.* **80**, 197 (1998).
- [13] A. G. Carlton, *Psychol. Bull.* **71**, 108 (1969).
- [14] G. A. Miller, *Information Theory in Psychology; Problems and Methods II-B* (Free Press, Glencoe, IL, 1955), p. 95.
- [15] A. Treves and S. Panzeri, *Neural Comput.* **7**, 399 (1995).
- [16] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap* (Chapman and Hall, New York, 1993).
- [17] J. D. Victor and K. P. Purpura, *Network* **8**, 127 (1997).
- [18] J. A. McFadden, *J. Soc. Ind. Appl. Math.* **13**, 988 (1995).
- [19] L. F. Kozachenko and N. N. Leonenko, *Probl. Peredachi Inf.* **23**, 9 (1987) [*Probl. Inf. Transm.* **23**, 95 (1987)].
- [20] G. Werner and V. B. Mountcastle, *J. Neurophysiol.* **28**, 359 (1965).
- [21] W. Bialek, F. Rieke, R. R. de Ruyter van Steveninck, and D. Warland, *Science* **252**, 1854 (1991).
- [22] D. S. Reich, J. D. Victor, B. W. Knight, T. Ozaki, and E. Kaplan, *J. Neurophysiol.* **77**, 2836 (1997).
- [23] Z. F. Mainen and T. J. Sejnowski, *Science* **268**, 1503 (1995).
- [24] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).
- [25] J. Beirlant and E. J. Dudewicz, *Int. J. Math. Stat. Sci.* **6**, 17 (1997).
- [26] P. Grassberger, *Phys. Lett. A* **128**, 369 (1988).
- [27] J. D. Victor, *Neural Comput.* **12**, 2797 (2000).
- [28] R. FitzHugh, *J. Gen. Physiol.* **40**, 925 (1957).
- [29] G. Gestri, *Biol. Cybern.* **31**, 97 (1978).
- [30] A. B. Tsybakov and E. C. van der Meulen, *Scand. J. Stat.* **23**, 75 (1966).