Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

A simple method for estimating the entropy of neural activity

Michael J Berry II[1], Gašper Tkačik[2], Julien Dubuis[3], Olivier Marre[4] and Rava Azeredo da Silveira[5,6]

[1] Department of Molecular Biology and Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544, USA
[2] Institute of Science and Technology Austria, Am Campus 1, A-3400 Klosterneuburg, Austria
[3] Department of Physics, Princeton University, Princeton, NJ 08544, USA
[4] Department of Molecular Biology, Princeton University, Princeton, NJ 08544, USA
[5] Laboratoire de Physique Statistique, Ecole Normale Supérieure, F-75005 Paris, France
[6] Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544, USA
E-mail: berry@princeton.edu, gasper.tkacik@ist.ac.at, julien.dubuis@gmail.com, olivier.marre@gmail.com and rava@ens.fr

Received 10 October 2012
Accepted 6 February 2013
Published 12 March 2013

Online at stacks.iop.org/JSTAT/2013/P03015
doi:10.1088/1742-5468/2013/03/P03015

Abstract. The number of possible activity patterns in a population of neurons grows exponentially with the size of the population. Typical experiments explore only a tiny fraction of the large space of possible activity patterns in the case of populations with more than 10 or 20 neurons. It is thus impossible, in this undersampled regime, to estimate the probabilities with which most of the activity patterns occur. As a result, the corresponding entropy—which is a measure of the computational power of the neural population—cannot be estimated directly. We propose a simple scheme for estimating the entropy in the undersampled regime, which bounds its value from both below and above. The lower bound is the usual 'naive' entropy of the experimental frequencies. The upper bound results from a hybrid approximation of the entropy which makes use of the naive estimate, a maximum entropy fit, and a coverage adjustment. We apply our simple scheme to artificial data, in order to check their accuracy; we also compare its performance to those of several previously defined entropy estimators. We then apply it to actual measurements of neural activity in populations with

1742-5468/13/P03015+20$33.00

J. Stat. Mech. (2013) P03015

up to 100 cells. Finally, we discuss the similarities and differences between the proposed simple estimation scheme and various earlier methods.

Keywords: neuronal networks (theory), computational biology, neural code

Contents

1. Introduction

While simpler organisms appear to estimate probability densities rather effortlessly, statisticians have a notoriously difficult time doing so. Since Shannon's fundamental work [15], they have developed a number of techniques for estimating a reduced quantity, the entropy, which depends upon the entire probability distribution. Concretely, the central difficulty is one of *undersampling*: if the number of possible configurations (or states) is large, typical experiments do not explore these sufficiently thoroughly to estimate the probability with which each occurs. In particular, the probability of the unobserved configurations remains undetermined by the experiment and, if these constitute an appreciable fraction of all possible configurations, then they may contribute to the entropy significantly. The various estimation techniques developed so far propose different ways to address this difficulty.

Shannon's entropy [15] is defined as

$$H = -\sum_{\mu} p_{\mu} \log_2 (p_{\mu}) \,, \tag{1}$$

where the sum runs over all possible configurations and p_μ denotes the (true) probability with which the configuration labeled by μ occurs. A naive estimator, also referred to as the *maximum likelihood estimator* because it makes use of the maximum likelihood estimates of the probabilities, is given by

$$\hat{H}_{\mathrm{MLE}} = -\sum_{\mu} \frac{m_\mu}{M} \log_2 \left(\frac{m_\mu}{M} \right), \tag{2}$$

where m_μ is the number of times configuration μ is seen in the experiment and M is the total number of configurations seen in the experiment; i.e., m_μ/M is the frequency of occurrence of configuration μ. (Throughout, we designate estimates by a circumflex accent.) This estimator assigns a vanishing probability of unobserved configurations and, as a result, is negatively biased. Miller calculated the bias for the case in which the number of data points, M, is larger that the number of configurations that occur with non-vanishing probability, Ω [9]; the resulting, so-called, Miller–Madow correction reads

$$\hat{H}_{\mathrm{MM}} = \hat{H}_{\mathrm{MLE}} + \frac{\Omega - 1}{2M}. \tag{3}$$

Clearly, this correction is useful only if the data are voluminous enough. In the context of neural recordings, this is rarely the case, and several more effective estimators have been devised.

The method of jackknifing was introduced in [13] and [18], and it was later applied to Shannon's entropy in [22]. It amounts to evaluating naive estimations of the entropy for different subdivisions of the data, and then to extrapolating the trend from small to large subdivisions in such a way as to obtain an estimate of the entropy in the limit of a large data set, in which the naive estimate and actual value match. A very similar method was applied in the context of neural recordings more recently [16]. Subsequently, more refined methods were introduced, also based upon estimating probabilities. In his beautiful piece of work [11], Paninski lays out a number of general results and puts forth the so-called *best upper bound* estimator. Yet a different kind of estimator is the *coverage adjusted estimator* [3, 20], which addresses specifically the difficulty associated with the fact that, in the undersampled regime, a number of configurations are unobserved in a typical experiment. As these estimators share some commonalities with the estimator we propose below, we relegate a more thorough discussion to the final section of the paper. Other methods of estimation do not rely upon explicitly approximating the probabilities, but rather follow a Bayesian approach, in which the key question is that of the choice of priors [21, 10, 1]. Finally, some rigorous approaches [2, 19] offer estimators and bounds, but these may fail in the highly undersampled regime most often encountered in the context of neural data.

Estimating the entropy of the activity in a population of neurons is one way of characterizing the 'coding power' of the population: the larger the entropy, the more 'configurations' can be represented by the population. However, undersampling is particularly severe in the context of neural recordings. If we restrict ourselves to a single time bin and assume that each of N neurons can emit up to k spikes, then the population can choose between $(k+1)^N$ configurations (and this must be further raised to the power of T if we allow T time bins). As a result, we would need at least as many as $(k+1)^N$ data points (and, in fact, many more) in order to have some chance of observing improbable configurations. For a cortical microcolumn with as few as 10^4 neurons, and assuming a

J. Stat. Mech. (2013) P03015

short time bin that allows no more than a single spike, we would have to measure twice as many as 10^{3010} data points. This number vastly exceeds the number of particles in the Universe. Thus, whenever we estimate the entropy of neural recordings, we generally have to do so in an extremely undersampled regime, in which a number of methods developed previously fail as they assume an asymptotic behavior.

Here, we put forth a new method for estimating the entropy in neural data. Our prescription bears some similarity with coverage adjusted methods [3, 20], but differs from these in ways guided by our intuition about neural population activity. We propose this scheme as a simple, and potentially useful, estimation device to the practitioner. Indeed, we see as the main merits of this new method its conceptual and calculational simplicity as well as the fact that it can be easily modified or refined in a number of ways. The reminder of the paper is laid out as follows. In section 2, we set up the problem and describe our estimation method; we then apply this method, first, to artificial data, for which one knows the true entropy, as a control, and, second, to real neural data. In section 3, we relate the method to earlier work and we describe briefly how it can be refined.

2. Results

2.1. Setting up the problem

We consider the activity of a population of N neurons in a time bin so short that each neuron fires at most one spike. Thus, the activity of the population can be represented by an N-dimensional vector (or pattern) with binary entries; there are 2^N possible vectors. We label the neurons by the index $i = 1, \ldots, N$ and the activity vectors by the index $\mu = 1, \ldots, 2^N$. We further assume that, in a given experiment, we record a number, M, of activity patterns. Among these M patterns, there are m_μ patterns μ; that is, the activity pattern labeled by μ is seen m_μ times in the experiment. By definition, $\sum_{\mu=1}^{2^N} m_\mu = M$. We also define a number, M_1, as the total number of singletons observed in the experiment,

$$M_1 = \sum_{\mu=1}^{2^N} m_\mu \delta_{m_\mu,1}, \tag{4}$$

where $\delta_{m,n}$ is the Kronecker function, which is non-vanishing and equal to one only if $m = n$. Finally, we assume a strongly undersampled regime, with $M \ll 2^N$. In that regime, a large fraction of the pattern counts, m_μ, are vanishing, i.e., most of the possible activity patterns are unobserved in the experiment.

Our aim is to estimate the entropy of the population activity. If the true probability with which each pattern occurs, p_μ, were known, then we would obtain the entropy as

$$H = -\sum_{\mu=1}^{2^N} p_\mu \log_2 (p_\mu) . \tag{5}$$

However, as we have just mentioned, most patterns are unobserved, so that the probabilities are unknown. We note that if the unknown probabilities are of the order of 2^{-N}, and if the number of such activity patterns is of the order of 2^{N-K} with $K \sim \mathcal{O}(1)$, then their contribution to the entropy is appreciable, of the order of $N/2^K$. If, for the

J. Stat. Mech. (2013) P03015

moment, we ignore this problem and assign to each pattern a naive 'counts probability', defined as

$$\tilde{p}_\mu \equiv \frac{m_\mu}{M}, \tag{6}$$

we obtain a naive estimate of the entropy—in fact, the entropy of the experimental counts—as

$$\hat{H}_< \equiv -\sum_{\mu=1}^{2^N} \tilde{p}_\mu \log_2\left(\tilde{p}_\mu\right) = -\sum_{\mu=1}^{2^N} \frac{m_\mu}{M} \log_2\left(\frac{m_\mu}{M}\right). \tag{7}$$

We denote this quantity by $H_<$ because it represents a lower bound to the actual entropy. The reason is that it neglects all the unobserved patterns, which, as we have noted above, can contribute significantly to the total entropy; in essence, this naive estimate assumes a much *narrower* probability distribution than it likely is. Many of the estimation schemes proposed hitherto advance ways to 'adjust the *coverage*' of the probability distribution in a way to compensate for its artifactual 'narrowing' and, thus, correct for the *bias* in the estimation. Hereafter, we propose another simple procedure.

For the sake of precision, we mention that $\hat{H}_<$ is, in fact, an *approximate* lower bound, as the count probabilities, m_μ/M, are only approximations of the true probabilities, p_μ; the former are expected to match the latter up to a standard deviation of order $\sqrt{m_\mu}/M$. We do not dwell on this distinction hereafter, because it is well known (see, e.g., [11]) that the resulting uncertainty in the naive estimate of the entropy is largely overwhelmed by the estimation bias, described above.

2.2. Singleton entropy estimator

The entropy is a sum of terms $-p_\mu \log_2\left(p_\mu\right)$ over all possible patterns $\mu = \{1, \ldots, 2^N\}$. While it is, in principle, not required to estimate the full probability distribution in order to estimate such a reduced quantity, many effective estimators do, and so shall we. The central idea, in our estimation scheme, is to treat those patterns that are observed reliably (in an experiment) and those that are not on an *un*equal footing. Thus, we divide the set of possible patterns into two groups; the first group is made up of patterns that are observed at least twice (in an experiment),

$$\text{group A} \equiv \{\mu \text{ such that } m_\mu \geq 2\}, \tag{8}$$

and the second group is that of the unlikely patterns with at most one count,

$$\text{group B} \equiv \{\mu \text{ such that } m_\mu = 0 \text{ or } 1\}. \tag{9}$$

Part of our rationale for thus dividing the possible patterns is that, in an undersampled regime, seeing a pattern twice is already a guarantee that it occurs with relatively high probability, whereas a single count does not carry the same meaning. Indeed, all we can conclude from a singleton is that its true probability is at most of the order of $1/M$. In particular, two different patterns observed a single time each may occur with vastly different true probabilities, while a singleton pattern and an unobserved pattern may occur with very similar true probabilities.

J. Stat. Mech. (2013) P03015

We then estimate the entropy of group A and that of group B, and express the full entropy estimate as the sum of the two terms,

$$\hat{H}_{>} = \hat{H}_{\mathrm{A}} + \hat{H}_{\mathrm{B}}. \tag{10}$$

We label the entropy estimate by the subscript '>' because, as we shall see, it is an approximate upper bound to the value of the actual entropy. Since the probability distribution over group A can be estimated faithfully, we calculate its entropy as the 'naive entropy' (the entropy of the experimental counts)

$$\hat{H}_{\mathrm{A}} = -\sum_{\mu\in \text{ group A}} \frac{m_\mu}{M} \log_2\left(\frac{m_\mu}{M}\right). \tag{11}$$

We now turn to group B: as we cannot estimate the probability distribution over group B, we instead bound the corresponding entropy from above. A coarse treatment would assume a uniform distribution over group B, but we can do better by constraining the entropy by moments of the probability distribution—even in a severely undersampled regime, moments can be estimated, while individual probabilities cannot. To do so, we calculate the firing rate of each cell, as

$$r_i \equiv \frac{1}{M_1} \sum_{\mu\in \text{ group B}} \Delta_i(\mu), \tag{12}$$

where the function $\Delta_i(\mu)$ is equal to 1 if cell i is active in pattern μ and equal to 0 otherwise. We can then use the *independent probability distribution*,

$$p_\mu^{\mathrm{indep}} \equiv \frac{1}{Z} \prod_{i=1}^{N} \left\{ r_i \Delta_i(\mu) + (1-r_i)\left[1-\Delta_i(\mu)\right] \right\}, \tag{13}$$

where Z is a normalizing factor, adjusted so that

$$\sum_{\mu\in \text{ group B}} p_\mu^{\mathrm{indep}} = \frac{M_1}{M}, \tag{14}$$

i.e., the probabilities sum to the fractional weight of group B. An upper bound to the entropy over group B is then obtained as

$$\hat{H}_{\mathrm{B}} \equiv -\sum_{\mu\in \text{ group B}} p_\mu^{\mathrm{indep}} \log_2\left(p_\mu^{\mathrm{indep}}\right). \tag{15}$$

In practice, in an undersampled regime one expects that group B represents a significant fraction of the 2^N possible patterns and, hence, it is difficult to compute the normalization factor, Z, directly by summing over all patterns in group B. However, the independent distribution provides a great simplification, as one can sum it easily over *all* patterns,

$$\sum_{\mu=1}^{2^N} p_\mu^{\mathrm{indep}} = \frac{1}{Z}, \tag{16}$$

$$-\sum_{\mu=1}^{2^N} p_\mu^{\mathrm{indep}} \log_2\left(p_\mu^{\mathrm{indep}}\right) = -\frac{1}{Z}\sum_{i=1}^{N}\left[r_i \log_2(r_i) + (1-r_i)\log_2(1-r_i)\right] + \frac{\log_2(Z)}{Z}. \tag{17}$$

J. Stat. Mech. (2013) P03015

Since summing over group A is straightforward (as it is expected to be made up of comparatively few patterns), one can use these identities to calculate Z and $\hat{H}_{\rm B}$. By writing the sum over the patterns in group B, in equation (15), as a difference between a total sum and a sum over group A only,

$$\hat{H}_{\rm B} \equiv -\sum_{\text{all } \mu} p_\mu^{\rm indep} \log_2\left(p_\mu^{\rm indep}\right) + \sum_{\mu \in \text{ group B}} p_\mu^{\rm indep} \log_2\left(p_\mu^{\rm indep}\right), \tag{18}$$

we can calculate the partition function and the entropy estimate as

$$Z = \left(\sum_{\mu=1}^{2^N} p_\mu^{\rm indep}\right)^{-1} = \left(\sum_{\mu \in \text{ group A}} p_\mu^{\rm indep} + \frac{M_1}{M}\right)^{-1}, \tag{19}$$

$$\begin{aligned}\hat{H}_{\rm B} = &-\frac{1}{Z}\sum_{i=1}^{N}\left[r_i \log_2(r_i) + (1-r_i)\log_2(1-r_i)\right] \\ &+ \frac{\log_2(Z)}{Z} + \sum_{\mu \in \text{ group A}} p_\mu^{\rm indep} \log_2\left(p_\mu^{\rm indep}\right).\end{aligned} \tag{20}$$

Finally, we arrive at an approximate upper bound to the total entropy,

$$\begin{aligned}\hat{H}_{>} = &-\sum_{\mu \in \text{ group A}} \frac{m_\mu}{M} \log_2\left(\frac{m_\mu}{M}\right) \\ &- \frac{1}{Z}\sum_{i=1}^{N}\left[r_i \log_2(r_i) + (1-r_i)\log_2(1-r_i)\right] \\ &+ \frac{\log_2(Z)}{Z} + \sum_{\mu \in \text{ group A}} p_\mu^{\rm indep} \log_2\left(p_\mu^{\rm indep}\right).\end{aligned} \tag{21}$$

In sum, we bound the entropy below by $\hat{H}_{<}$, the naive entropy defined in equation (7), and above by $\hat{H}_{>}$, given in equation (21). As sampling becomes better and, as a result, the fractional weight of group B, M_1/M, decreases, $\hat{H}_{<}$ and $\hat{H}_{>}$ move toward each other. As illustrated in the following sections, we extrapolate the bounds to the point $M_1/M = 0$, at which we obtain a tight interval which likely contains the actual entropy.

2.3. Illustration of the singleton method with artificial data

As a way to test our simple estimation scheme, we applied it to artificial data for which we had access to accurate estimations of the entropy obtained by direct methods (see below). We refer to the data we examined as 'artificial data' because population activity patterns were sampled from a given probability distribution rather than being measured experimentally. Still, the probability distribution was itself inferred from neural data; it was not an arbitrary distribution.

Specifically, we started with simultaneous recordings of spike trains from populations of ganglion cells in response to a natural movie [8]. Next, we fitted a maximum entropy distribution constrained by the firing rates and pairwise correlations of all neurons, so

J. Stat. Mech. (2013) P03015

that it took the form

$$p_{\mu,N} = \frac{1}{Z_N} \exp\left(\sum_{i=1}^{N} h_{i,N} s_i^\mu + \sum_{i,j=1}^{N} J_{ij,N} s_i^\mu s_j^\mu\right), \tag{22}$$

where s_i^μ denotes the activity of cell i in pattern μ and $h_{i,N}$ and $J_{ij,N}$ are fitting parameters. The parameters of the maximum entropy distribution were obtained by a gradient descent algorithm, which made use of Monte Carlo sampling for the estimation of its moments [17]. We applied this step of the procedure successively for different values of the total number of ganglion cells under scrutiny, N, so as to illustrate our estimation scheme for different population sizes; the subscript N acts as a reminder that the numerical values of the parameters $h_{i,N}$, $J_{ij,N}$ and Z_N depend upon the population size. For each choice of N, we then drew $M = 11\,270\,000$ activity patterns from the probability distribution $p_{\mu,N}$. These constituted our 'artificial data sets'.

We then calculated the estimated lower bound to the entropy, $\hat{H}_<$, from equation (7) and the estimated upper bound to the entropy, $\hat{H}_>$, from equation (21). For each choice of N, we obtained a final estimated range of the entropy by extrapolating the values of $\hat{H}_<$ and $\hat{H}_>$ to a high sampling limit. To achieve this, we divided the set of M activity patterns (for a given N) into K subsets, with $K = 2$, 3, 4 or 5, by assigning each of the M patterns randomly to one of the K subsets. Each choice of K yielded a different value of the fraction of singletons, M_1/M, which decreased monotonically as K increased. (This amounts to saying that, for larger data sets, unobserved and rare patterns represented a smaller fraction of the probability weight.) The estimated lower and upper bounds to the entropy, $\hat{H}_<$ and $\hat{H}_>$, also varied with K, and we plotted these as a function of M_1/M. For a given value of K, all of these quantities varied among the data subsets; we used their average values. Finally, we extrapolated the values of $\hat{H}_<$ and $\hat{H}_>$ to the limit of perfect sampling, $M_1/M = 0$, with a quadratic polynomial. We expect that the extrapolated values of $\hat{H}_<$ and $\hat{H}_>$ will converge at $M_1/M = 0$, and choose their average to be our best estimate of the entropy.

Figure 1 illustrates the outcome of the procedure just outlined in the cases of $N = 20$, 40, 60, 80 and 100 cells. Notice that the scales of the abscissa and the ordinate vary substantially from panel to panel; in particular, M_1/M takes a much larger range of values when sampling responses from larger populations. In all cases, the extrapolated values of $\hat{H}_<$ and $\hat{H}_>$ at $M_1/M = 0$ match or differ very little, by about 0.1%. Furthermore, our simple estimation scheme yields entropy values very close to those obtained by a numerical integration of the heat capacity (see [17] for details on the latter). When the population is sufficiently small, we can integrate over all the possible activity patterns numerically and, as a result, calculate the true entropy directly. We did so for the case $N = 20$ (figure 1(A)) and found that the estimated entropy overestimated the true entropy by 0.03%, while the heat capacity integration overestimated the true entropy by 0.1%. Interestingly, the lower and upper bounds lie more or less symmetrically below and above the estimated entropy (with the upper bound slightly closer to the estimated entropy than the lower bound). Thus, a coarse estimation of the entropy as the average between the lower and upper bounds, without extrapolation, would yield a much more reliable value than the naive entropy estimate alone.

J. Stat. Mech. (2013) P03015

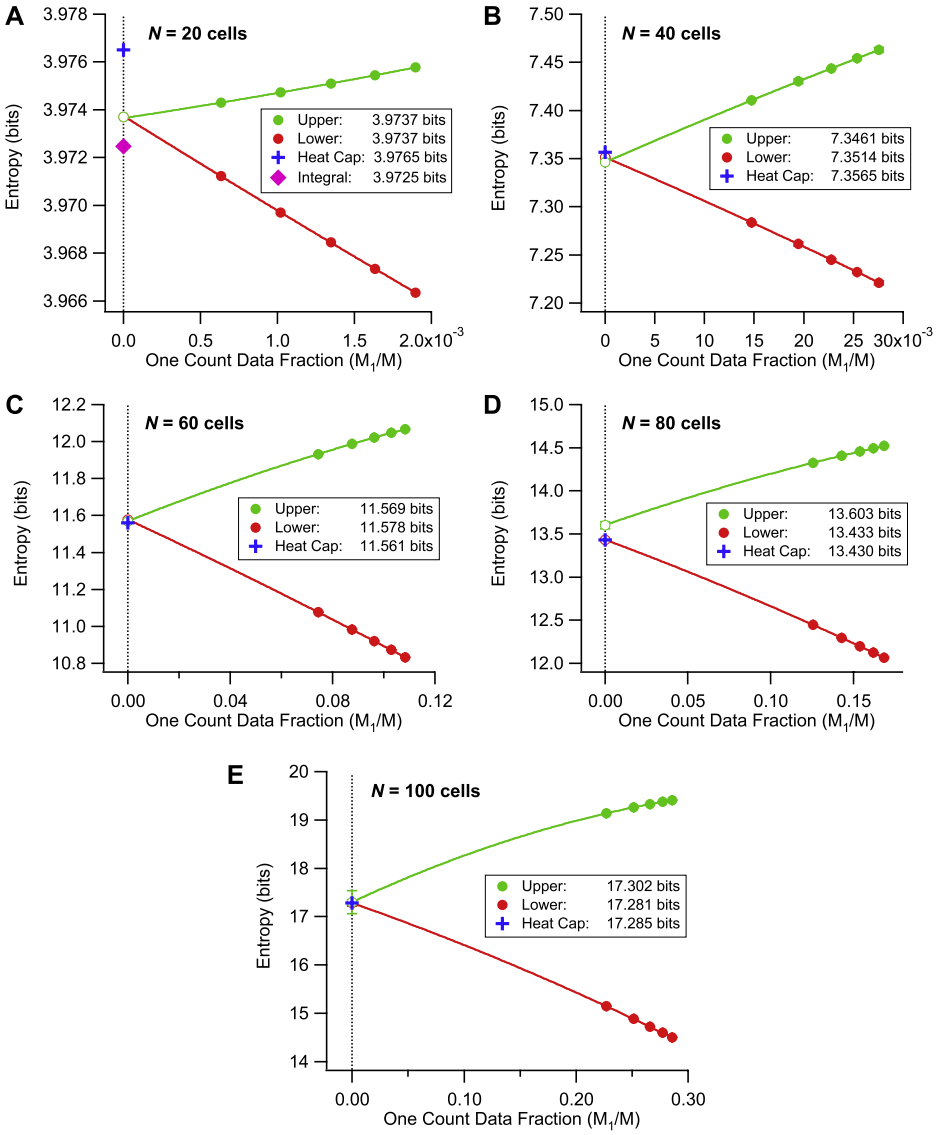

Figure 1. Singleton entropy estimator applied to artificial data. (A) Upper bound ($\hat{H}_>$, green) and lower bound ($\hat{H}_<$, red) entropies plotted against the fraction of singletons (M_1/M), for $N = 20$ cells, from a pairwise maximum entropy model (solid circles). Quadratic extrapolations (lines) converge on estimates at $M_1/M = 0$ (open circles). Entropies estimated by integration of the heat capacity (blue cross) and calculated by numerical integration over the probability distribution (pink diamond) are similar. (B)–(E) Upper bound ($\hat{H}_>$, green) and lower bound ($\hat{H}_<$, red) entropies plotted against the fraction of singletons (M_1/M) for larger populations with $N = 40$, 60, 80 and 100 cells, respectively. The error bars (often not visible) are the standard deviation of entropy values across all K data subsets. The quadratic extrapolations (lines), extrapolated entropies (open circles) and entropy estimated by heat capacity integration (blue cross) are as in panel (A). The error bars on the extrapolated values derive from uncertainties in the parameters of the quadratic fits, according to the CurveFit command in IgorPro 6.04.

J. Stat. Mech. (2013) P03015

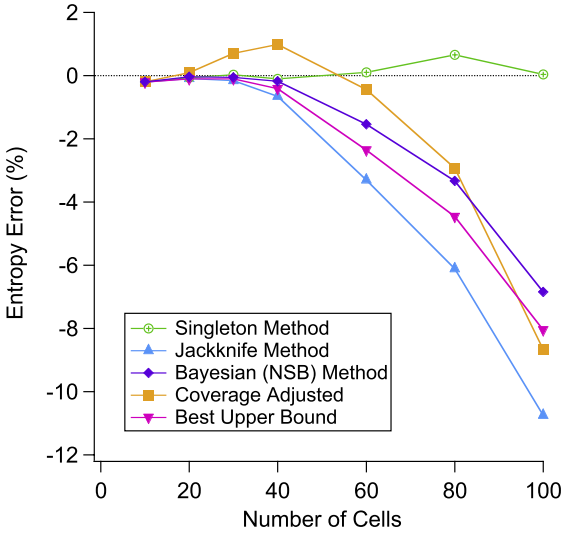

Figure 2. Comparison of the singleton entropy estimator to other entropy estimators. Errors of entropy estimates from the singleton method (green circles), jackknife method (blue triangles), Bayesian (NSB) method (purple diamonds), coverage adjusted method (orange squares) and best upper bound (inverted pink triangles) are plotted versus the number of cells, N. The error is defined as the difference between the estimated entropy and the value derived from heat capacity integration.

2.4. Comparison of the singleton method with other entropy estimators

In the case of artificial data, discussed in the above section, we had access to a reliable estimate of the entropy from numerical integration of the heat capacity corresponding to the maximum entropy probability distribution [17]. We used it as a benchmark: we evaluated the performance of various entropy estimators by referencing their outcomes with respect to the estimate of the entropy from heat capacity integration, for different choices of population size, N (figure 2). In this way, we compared the outcomes of the singleton method to the outcomes of four previously proposed entropy estimators, namely, the classic jackknife estimator [13, 18, 22, 16], a Bayesian (NSB) estimator [10], and the more recent best upper bound estimator [11] and coverage adjusted estimator [3, 20]. (We discuss the structure of the latter two and their similarities with the singleton method in section 3. We mention here that, following [20], we did not use the actual total number of possible patterns in the best upper bound estimator; rather, we used a naive estimate of this quantity—the total number of distinct observed patterns. In [20], this version of the estimator is referred to as 'BUB−'. The best upper bound method was devised originally to provide a rigorous bound to the entropy. In the undersampled cases of interest here, this bound becomes loose. The naive choice in the BUB− method addresses this issue, but because of it the method should be interpreted as yielding an approximation to the entropy rather than a bound.)

For each method, we calculated the estimate of the entropy in the cases of $N = 20$, 40, 60, 80 and 100 cells. While the singleton method yielded values which were consistently close to the estimate from heat capacity integration, the performance of the other

J. Stat. Mech. (2013) P03015

estimators—jackknife, Bayesian, best upper bound and coverage adjusted—degraded progressively for larger population sizes.

We note that the difference between the outcome of the singleton method and that of the heat capacity integration was of the order of one part in a thousand for several values of N and was below one part in a hundred for all conditions tested. This discrepancy is comparable to the relative difference between the true value of the entropy and the estimate from heat capacity, when we have access to the former in the case of $N = 20$ (figure 1(A)). We note also that our extrapolation of the lower bound to the entropy is reminiscent of the jackknife method. The crucial difference, however, is that we use M_1/M as the extrapolation variable instead of the inverse data size, K.

2.5. Illustration of the singleton method with neural data

Following a similar procedure to the one we applied to artificial data, we submitted real neural data to the singleton method. For this, we used recordings of the activity of retinal ganglion cells in response to a set of five different natural movies [12]; we divided the spike raster of the population into $M = 426\,000$ time bins of 20 ms each. We calculated the lower and upper bounds to the entropy (equations (7) and (21), respectively) for populations of $N = 20$, 40, 60, 80 and 100 cells. As in the case of artificial data, for each choice of N we subdivided the data into K subsets, with $K = 2$–5, and, using the successive subdivisions, we extrapolated the values of the lower and upper bounds to the limit of perfect sampling. Our results are illustrated in figure 3; notice again the different scales of the abscissae and ordinates for different numbers of neurons.

As a consistency check of the singleton method, the extrapolated values of the bounds converge to neighboring values. As we take the average between the two values to be our best estimate of the entropy, the difference between the two values provides us with an estimate of the reliability of the method. Here, we find that this difference is less than one part in a hundred.

3. Discussion

3.1. Summary and variations

In this study, we introduced a simple entropy estimator, which we called the singleton estimator, useful in the undersampled regime commonly encountered in the context of neural recordings. The singleton method finds an estimate of the entropy by extrapolating lower and upper bounds to the entropy from the undersampled regime to the fully sampled limit. In all the cases we examined, we checked the consistency of the method, finding that the extrapolated lower and upper bounds very nearly matched. In the case of artificial data, in which we had access to either the true entropy (from direct integration of the probability distribution) or a reliable estimate of the entropy (from integration of the heat capacity), we found that the singleton method yielded accurate estimates of the entropy—with accuracy often of the order of one part in a thousand. Furthermore, a comparison with other recently proposed entropy estimators suggested that the singleton method was the most reliable: all the other methods underestimated the entropy by larger amounts, ranging up to as much as 7–10% for $N = 100$ cells. However, our preliminary

J. Stat. Mech. (2013) P03015

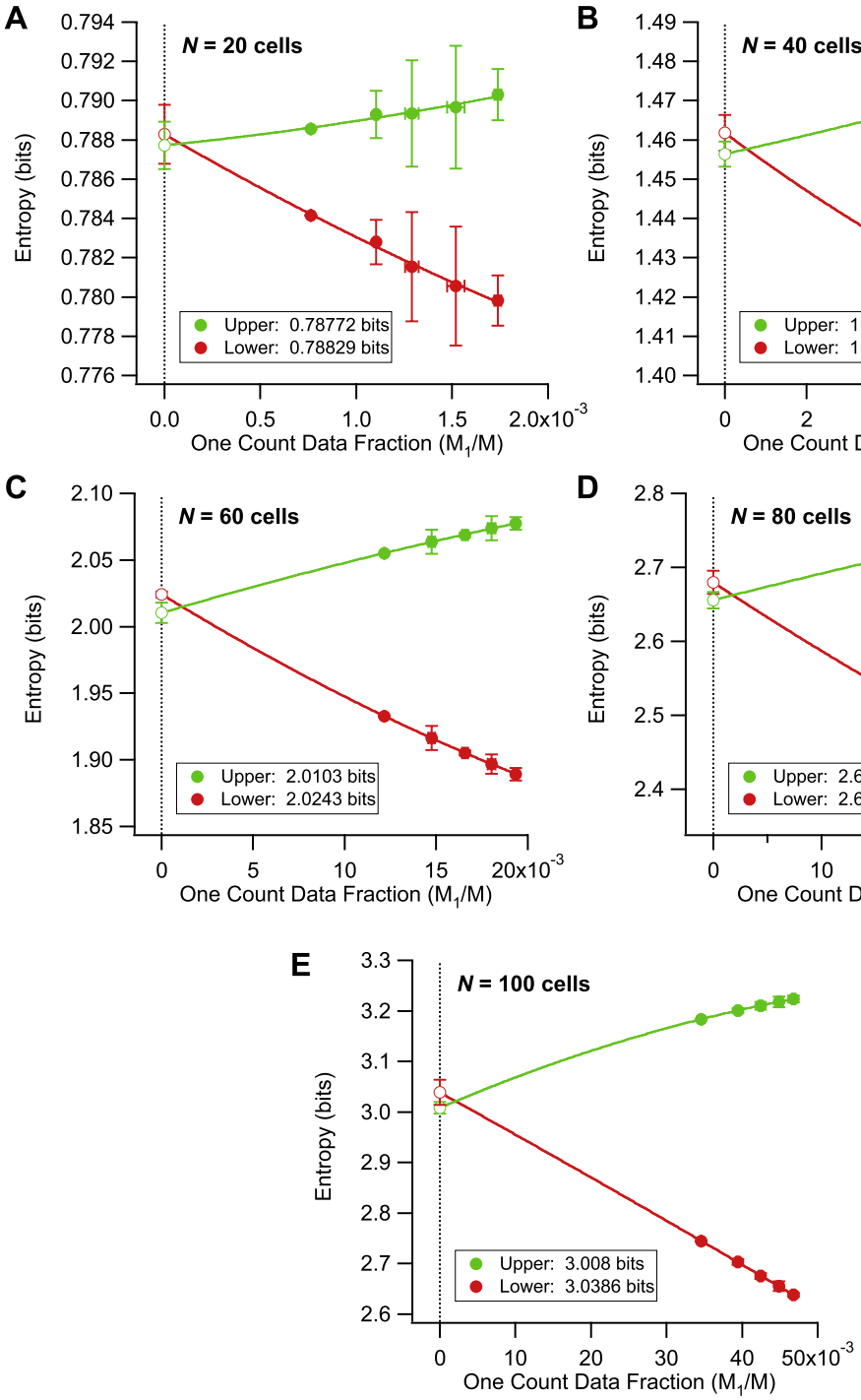

Figure 3. Singleton entropy estimator applied to real data. Upper bound ($\hat{H}_>$, green) and lower bound ($\hat{H}_<$, red) entropies plotted against the fraction of singletons (M_1/M), for $N = 20$ (A), 40 (B), 60 (C), 80 (D) and 100 (E) real neurons (solid circles). The error bars are the standard deviation of entropy values across all K data subsets. The quadratic extrapolations (lines) converge on the estimates at $M_1/M = 0$ (open circles). The error bars on the extrapolated values derive from uncertainties in the parameters of the quadratic fits, according to the CurveFit command in IgorPro 6.04.

J. Stat. Mech. (2013) P03015

evaluations of an improved Bayesian estimator newly proposed by Archer, Memming Park and Pillow [1], using the same data set, indicate a significantly improved performance.

The key idea in our simple estimation method is to divide the data into two groups. The first group (group A) is made up of patterns with reliably estimated probabilities; its entropy can thus be estimated in a naive fashion. The second group (group B) contains the rare patterns not sampled by the experiment and is the cause of the 'unknown bias difficulty'. Rather than estimating its entropy, we bound it from above in a way that is constrained by the firing rates of the neurons. Using this 'independent neurons prescription' yields a very simple form of a probability distribution which can then be integrated to obtain the entropy bound, even if the distribution is so undersampled that most of the patterns belong to group B. Extrapolating the resulting upper bound, together with the lower bound from a naive (maximum likelihood) estimation of the entropy, yields a tight bound around the estimated entropy. We emphasize that our aim, here, is to put forth simple heuristics to estimate a numerical value for the entropy of neural activity; we are not concerned with rates of convergence as a function of data volume, nor with asymptotics otherwise. Furthermore, we have disregarded the question of the variance of estimations (but see below), because the bias is known to overwhelm the variance. However, we mention that the naive entropy estimate over group A of course varies from experiment to experiment; as a result, our upper bound is an 'approximate upper bound'. The uncertainty in the estimate of the entropy of group A can be evaluated via the variability of the counts, m_μ.

Our simple estimation scheme may be modified or refined in a number of ways. For example, one may wonder whether it would make more sense to define group B as made up of all the patterns represented not at most a single time in the data, but at most q times in the data. Doing so would make the estimate of the entropy of group A more reliable, because counts are more faithfully representative of probabilities in the case of large counts. But it would also shift some of the probability weight from group A to group B, which causes the bias and on the entropy of which our handle is a relatively rough upper bound. We expect that, in most cases, it would be unfavorable to thus trade bias against a reduced variance—but it is a quantitative question, to be studied in specific instances of data sets. Still, conceptually, one can say that one-counts are 'morally different' from q-counts: if M is sufficiently large, seeing a given pattern at least twice in an experiment indicates that this pattern is likely; but if it is seen only a single time, there is no way to argue that it is likelier than an unobserved pattern. We note that a tacit assumption is that patterns are drawn independently; thus, we assume that spike trains are binned over a time scale that largely ensures independence among bins.

3.2. Relation with previous work

3.2.1. Paninski's best upper bound estimator. Paninski's *best upper bound* estimator [11] is obtained by replacing a sum over all possible patterns by a sum over the data set,

$$\hat{H}_{\mathrm{BUB}} = \sum_{j=1}^{M} a_j h_j, \tag{23}$$

where h_j is equal to the number of patterns that occur exactly j times in the data set,

$$h_j = \sum_{\mu=1}^{\Omega} \delta_{m_\mu, j}. \tag{24}$$

For the particular choice of coefficients $a_j = -(j/M)\log_2(j/M)$, we retrieve the maximum likelihood estimator, $\hat{H}_{\mathrm{MLE}}$. The average of $\hat{H}_{\mathrm{BUB}}$ defined in equation (23) is then obtained as

$$\begin{aligned}\left\langle \hat{H}_{\mathrm{BUB}} \right\rangle &= \sum_{j=1}^{M} a_j \langle h_j \rangle \\ &= \sum_{j=1}^{M} a_j \sum_{\mu=1}^{\Omega} \langle \delta_{m_\mu, j} \rangle \\ &= \sum_{\mu=1}^{\Omega} \sum_{j=1}^{M} a_j \binom{M}{j} p_\mu^j (1-p_\mu)^{M-j}. \end{aligned} \tag{25}$$

For a good estimator, one would like this polynomial form to approximate well the form of the true entropy,

$$H = -\sum_{\mu=1}^{\Omega} p_\mu \log_2(p_\mu). \tag{26}$$

In essence, Paninski derives coefficients, a_j, which minimize an error functional. Thus, his approach is very different from ours; however, it is similar in that he also treats the rare patterns differently from the likely patterns. Specifically, in his error functional, he weighs the 'local bias',

$$-x\log_2(x) - \sum_{j=1}^{M} a_j \binom{M}{j} x^j (1-x)^{M-j}, \tag{27}$$

where $x \in [0,1]$, with a prefactor

$$f(x) = \begin{cases} \Omega & \text{if } x < 1/\Omega, \\ 1/x & \text{if } x \geq 1/\Omega. \end{cases} \tag{28}$$

In other words, the rare patterns are penalized in a more stringent way than the likely patterns. In our simple scheme, the entropy bound is tightened by submitting the estimation of the probabilities of the rare patterns to the constraints of the measured firing rates of the neurons.

3.2.2. Coverage adjusted estimators. Entropy estimators that are conceptually closer to our proposed scheme are the so-called *coverage adjusted estimators* [3, 20]. In these, the set of possible patterns in also divided into two groups but, unlike our prescription, here the first group (call it group A′) is made up of all the observed patterns, while the second group (call it group B′) contains all the unobserved patterns. The entropy is then estimated by making use of two tricks. We note that we cannot take as the estimated values of the

J. Stat. Mech. (2013) P03015

probabilities, $\hat{p}_\mu$, of patterns in group A′ their maximum likelihood values,

$$\hat{p}_\mu = \frac{m_\mu}{M}, \tag{29}$$

because this choice would assign a vanishing probability weight to group B′. We must suppress these values by the 'coverage' of group A′, i.e., by the estimated probability weight of group A′ . A simple scheme, discussed in the section 3.2.3, indicates that the frequency of singletons, M_1/M, provides a good estimate of the total probability weight of the *unobserved* patterns. This observation in turn yields a simple prescription for suppressing the probabilities of *observed* patterns by an adequate amount, as

$$\hat{p}_\mu = \left(1 - \frac{M_1}{M}\right) \frac{m_\mu}{M}. \tag{30}$$

However, this prescription is not sufficient to yield an estimate of the entropy, because it does not carry any information on the contribution of group B′ to the entropy; indeed; since group B′ is made up of *unobserved* patterns, we would be hard-pressed to make any precise statement about these. However, we can still make an adjustment proposed by Horvitz and Thomson [6], consistent with the existence of group B′, as follows. They considered a problem in which one would like to estimate the total sum of Ω numbers x_μ,

$$S = \sum_{\mu=1}^{\Omega} x_\mu. \tag{31}$$

The twist, here, is not only that the samples, $\sigma \equiv (x_{\mu_1}, x_{\mu_2}, \ldots, x_{\mu_M})$, are under-representative, with $M < \Omega$, but also that the labels, $\mu_1, \ldots, \mu_M$, are drawn with (in general, non-uniform) probability, p_μ. The naive estimator

$$\hat{S} = \sum_{\substack{\text{labels } \mu \text{ observed} \\ \text{in the sample } \sigma}} x_\mu \tag{32}$$

is biased because a given value x_μ will be under-represented or over-represented depending upon the probability with which it is present in the sample. Horvitz and Thomson proposed instead the estimator given by

$$\hat{S} = \sum_{\substack{\text{labels } \mu \text{ observed} \\ \text{in the sample } \sigma}} \frac{x_\mu}{\pi_\mu}, \tag{33}$$

where π_μ is the probability with which label μ is observed in the sample. This estimator can be trivially rewritten as

$$\hat{S} = \sum_{i=1}^{M} \beta_i x_{\mu_i}, \tag{34}$$

with $\beta_i = (m_{\mu_i} \pi_{\mu_i})^{-1}$; as before, m_μ denotes the number of times the label μ appears in the sample (i.e., the count). Using this form and that of the probability of a given (ordered) sample,

$$p(\sigma) = p_{\mu_1} \cdot p_{\mu_2} \cdot \cdots \cdot p_{\mu_M}, \tag{35}$$

we can show readily that this estimator is unbiased,

$$\begin{aligned}\left\langle \hat{S} \right\rangle &= \sum_{\text{all } \sigma} p\left(\sigma\right) \sum_{i=1}^{M} \beta_i x_{\mu_i} \\ &= M \sum_{\text{all } \sigma} p\left(\sigma\right) \frac{x_\mu}{m_\mu \pi_\mu} \\ &= M \sum_{\text{all } \sigma} \sum_{m=1}^{M} \binom{M-1}{m-1} p_\mu^m \left(1-p_\mu\right)^{M-m} \frac{x_\mu}{m\pi_\mu}. \end{aligned} \tag{36}$$

However, this expression is none other than the full sum in equation (31), since

$$\sum_{m=1}^{M} \binom{M-1}{m-1} p_\mu^m \left(1-p_\mu\right)^{M-m} \frac{M}{m} = 1-\left(1-p_\mu\right)^M, \tag{37}$$

which amounts, precisely, to the probability that label μ appears in the sample, π_μ.

We can translate these two prescriptions—coverage adjustment and Horvitz–Thomson adjustment—in the case of entropy estimation, by setting $x_\mu = -p_\mu \log_2\left(p_\mu\right)$, and we obtain the *coverage adjusted estimator* of entropy, as

$$\hat{H}_{\text{CAE}} = -\sum_{\mu \text{ in group A}'} \frac{\hat{p}_\mu \log_2\left(\hat{p}_\mu\right)}{1-\left(1-\hat{p}_\mu\right)^M}, \tag{38}$$

where $\hat{p}_\mu$ is given in equation (30). It is worth noting the similarities and differences between this estimator and the one we propose in this paper. As in our case, the coverage adjusted estimator divides the data into two groups—'known' and 'unknown'—and it adjusts the values of naive probability estimates as a function of the weight of singletons (one-counts in the data). In contrast to our case, it includes the singletons in the 'known' group and does not rely upon any estimate of probabilities within the 'unknown' group. In our scheme, the 'unknown' group is, in fact, partially known: from singletons, we derive enough knowledge about the firing properties of neurons to be able to approximate the probabilities in the (largely) 'unknown' group and, as a result, bound its entropy.

3.2.3. The Good–Turing Bayesian scheme. In naive estimates of the entropy, the normalized count (i.e., the frequency), m_μ/M, is equated to the probability of a given pattern. Good points out that, in a Bayesian framework, one can make a finer estimate; he reports the method in [5], where he also mentions that it was suggested to him by Turing. In essence, the method is rather simple and goes as follows. Consider a pattern that has appeared m times in the data. If the set of true probabilities is known to us, we can calculate the probability that this pattern is the pattern labeled by μ, as

$$\rho_\mu \equiv \frac{\binom{M}{m} p_\mu^m \left(1-p_\mu\right)^{M-m}}{\sum_{\nu=1}^{\Omega} \binom{M}{m} p_\nu^m \left(1-p_\nu\right)^{M-m}}. \tag{39}$$

J. Stat. Mech. (2013) P03015

As a result, the 'average estimated probability' associated with this pattern, which we denote by $\langle \hat{p}\,(m)\rangle$, is obtained as

$$\begin{aligned}\langle \hat{p}\,(m)\rangle &= \sum_{\mu=1}^{\Omega} p_\mu \rho_\mu \\ &= \frac{\sum_{\mu=1}^{\Omega} \binom{M}{m} p_\mu^{m+1} (1-p_\mu)^{M-m}}{\sum_{\nu=1}^{\Omega} \binom{M}{m} p_\nu^{m} (1-p_\nu)^{M-m}} \\ &= \frac{m+1}{M+1} \frac{\sum_{\mu=1}^{\Omega} \binom{M+1}{m+1} p_\mu^{m+1} (1-p_\mu)^{(M+1)-(m+1)}}{\sum_{\nu=1}^{\Omega} \binom{M}{m} p_\nu^{m} (1-p_\nu)^{M-m}}. \end{aligned} \tag{40}$$

Now, the denominator is simply the average number of patterns that appear m times in an M-point data set and, similarly, the numerator is the average number of patterns that appear $m+1$ times in an $(M+1)$-point data set. If we estimate these from the data—estimates which are more robust than estimates of the probabilities themselves—we obtain the estimate of the probability of pattern μ, if it has appeared m times in the data, as

$$\hat{p}_\mu\,(m) = \frac{m+1}{M+1} \frac{n_{M+1}\,(m+1)}{n_M\,(m)} \approx \frac{m+1}{M} \frac{n_M\,(m+1)}{n_M\,(m)}, \tag{41}$$

where $n_M\,(m)$ is the number of different patterns that appear m times each in the data.

This estimate ought to be contrasted with the naive estimate,

$$\hat{p}_\mu = \frac{m}{M}. \tag{42}$$

The above arguments suggest that we would be better off using the Good–Turing form, rather than the naive one, in calculating our bounds. Another point of interest also emerges from this framework: the Good–Turing form implies that the total probability weight of all the patterns that appear m times each in the data is estimated at $n_M\,(m+1)\,/M$. In particular, the total weight of the unobserved patterns is estimated to be the frequency of singletons, M_1/M. This conclusion provides a justification for the coverage adjustment discussed in section 3.2.2, but it differs from our simple procedure in which we took the frequency of singletons, M_1/M, to represent the *combined* probability weights of the singletons *and* the unobserved patterns. We attempted to refine our estimation procedure by using the Good–Turing estimates of activity pattern probabilities, instead of their naive (count) estimates, to obtain a tighter upper bound, $\hat{H}_>$. However, there was considerable sampling variability in the values of $n_M(m)$ for intermediate values of m, yielding a prohibitive variability in $\hat{H}_>$ (data not shown).

3.2.4. Maximum entropy models and the reliable interaction model. In past years, maximum entropy models [7] have emerged as promising formalisms for summarizing the statistical properties of large, correlated neural populations; more specifically, for predicting the probabilities with which population activity patterns occur and for estimating the corresponding entropy. In these models, some activity-dependent quantities, $f_a\,(\mu)$, are averaged over the data, and this average is used to constrain

J. Stat. Mech. (2013) P03015

the entropy, which is maximized otherwise. This procedure yields an estimate of the probability, as

$$\hat{p}_{\{f_a\},\mu} = \exp\left(-\sum_a \lambda_a f_a(\mu)\right), \tag{43}$$

where the values of the prefactors, λ_a, are chosen so that the average of $f_a(\mu)$ over the distribution $\hat{p}_{\{f_a\},\mu}$ matches its average over the data. The resulting probability distribution is the one that maximizes entropy while fixing the average of the functions $f_a(\mu)$. Such a procedure thus provides an upper bound to the entropy of the output of a neural population.

If the population is large or if the data are not voluminous enough, the derivation of a maximum entropy model from neural data can become computationally very difficult. First, the functions $f_a(\mu)$ have to be chosen to be 'simple enough' so that the data offer a sufficiently efficient sampling to calculate their average reliably. In many cases, $f_a(\mu)$ have been chosen as low-order products of the single-cell activities,

$$f_0(\mu) = 1, \tag{44}$$

$$f_{1,i}(\mu) = s_i(\mu), \tag{45}$$

$$f_{2,ij}(\mu) = s_i(\mu)\, s_j(\mu), \tag{46}$$

where $s_i(\mu)$ is the activity of cell i in pattern μ. Note that the above choice for $f_0(\mu)$ will yield a probability distribution normalized to unity. In most studies, n-point higher-order terms do not appear because the data are not voluminous enough for a reliable estimate of the corresponding moments for all choices of subsets of n neurons. Second, the numerical optimization of the parameters in the model requires a thorough sampling of the estimated probability distribution, which becomes computationally prohibitive for large enough neural populations.

While pairwise maximum entropy models have been applied to populations of more than one hundred neurons [17, 14], careful inspection reveals that higher-order interactions may become statistically significant for populations with more than 40 neurons. In order to address this point, an alternative approach has been introduced recently, under the name of the *reliable interaction model* [4]. Thus, the reliable interaction model takes well-sampled probabilities as its constraints and estimates the probabilities of rare patterns by extrapolation, assuming a form

$$\hat{p}_{\{f_a\},\mu} = \exp\left(\alpha_0 + \sum_{i=1}^{N} \alpha_i s_i(\mu) + \sum_{i,j=1}^{N} \alpha_{ij} s_i(\mu)\, s_j(\mu) + \sum_{i,j,k=1}^{N} \alpha_{ijk} s_i(\mu)\, s_j(\mu)\, s_i(\mu) + \cdots\right). \tag{47}$$

In general, such an approach would fail. The reason why it works well for neural data is that, if the binning in time is sufficiently fine, the activity patterns are very sparse: only a small minority of patterns, each with very few spikes across the population, occur reliably. As a result, only a small subset of higher-order coefficients ('interactions') are found to be non-vanishing and the rest can be neglected. Presumably, also, the structure of correlation in neural systems is such that only a moderate number of higher-order interaction terms

J. Stat. Mech. (2013) P03015

are non-vanishing, but these are highly relevant in shaping the distribution of activity. In practice, non-vanishing interactions appear up to fourth order and these are so sparse that the total number of non-vanishing interactions sums to less than the corresponding number of interactions in the pairwise maximum entropy model [4].

We note that, in fact, the reliable interaction model can be viewed as resulting from an approximate maximum entropy procedure, with the particular choice of functions

$$f_0(\mu) = \prod_i \delta_{i_\mu,0}, \tag{48}$$

$$f_{1,i}(\mu) = \delta_{i_\mu,1} \prod_{j\neq i} \delta_{j_\mu,0}, \tag{49}$$

$$f_{2,ij}(\mu) = \delta_{i_\mu,1}\delta_{j_\mu,1} \prod_{k\neq i,j} \delta_{k_\mu,0}, \tag{50}$$

$$f_{3,ijk}(\mu) = \delta_{i_\mu,1}\delta_{j_\mu,1}\delta_{k_\mu,1} \prod_{l\neq i,j,k} \delta_{l_\mu,0}, \tag{51}$$

$$\ldots, \tag{52}$$

where $\delta_{i_\mu,0(1)}$ is a Kronecker delta equal to 1 if the ith neuron in pattern μ is silent (active) and equal to 0 otherwise. If we set, by convention, that $s_i(\mu) = 1$ if the ith neuron in pattern μ is active and $s_i(\mu) = 0$ if it is silent, a maximum entropy procedure yields the form of equation (47) up to higher-order corrections (which come with the particular structure that derives from expanding the products of terms $\delta_{i_\mu,0} = 1 - s_i(\mu)$). An important point to note, here, is that the constants in the distribution of equation (47) are chosen so as to match the pattern probabilities that can be estimated reliably directly from the data, but they do not ensure the normalization of the distribution. While the reliable interaction model captures well some aspects of the neural activity statistics [4], it does not represent a well-defined probabilistic model. A direct implementation of the normalization constraint suffers from the same sampling difficulty as plagues the usual maximum entropy approaches. Perhaps, this issue can be addressed by one of the adjustment methods described above.

In sum, maximum entropy models and the reliable interaction model have complementary strengths: the former are normalized and, hence, yield an estimate of the entropy, but are intractable for large neural populations; the latter is tractable even in the context of large populations, provided that the activity is sparse, but is non-normalized and, hence, cannot be used to estimate entropy. Our simple scheme for estimating entropy may be viewed as a hybrid of these two approaches. We rely upon the frequently occurring patterns to make direct estimates of probabilities and we adjust for the rare patterns with an independent maximum entropy model.

References

[1] Archer E, Memming Park I and Pillow J W, *Bayesian estimation of discrete entropy with mixtures of stick-breaking priors*, 2012 *Adv. Neural Inform. Proc. Syst.* **25** 2024–32

[2] Batu T, Dasgupta S, Kumar R and Rubinfeld R, *The complexity of approximating the entropy*, 2005 *SIAM J. Comput.* **35** 132–50

[3] Chao A and Shen T J, *Nonparametric estimation of shannon's index of diversity when there are unseen species in sample*, 2003 *Environ. Ecol. Stat.* **10** 429–43

J. Stat. Mech. (2013) P03015

[4] Ganmor E, Segev R and Schneidman E, *Sparse low-order interaction network underlies a highly correlated and learnable neural population code*, 2011 *Prc. Natl Acad. Sci.* **108** 9679–84
[5] Good I J, *The population frequencies of species and the estimation of population parameters*, 1953 *Biometrika* **40** 237–64
[6] Horvitz D G and Thompson D J, *A generalization of sampling without replacement from a finite universe*, 1952 *J. Am. Stat.* **47** 663–85
[7] Jaynes E T, *Information theory and statistical mechanics*, 1957 *Phys. Rev.* **106** 62–79
[8] Marre O, Amodei D, Deshmukh N, Sadeghi K, Soo F, Holy T E and Berry M J II, *Mapping a complete neural population in the retina*, 2012 *J. Neurosci.* **32** 14859–73
[9] Miller G, *Note on the bias of information estimates*, Information Theory in Psychology II-B ed H Quastler (Glencoe, IL: Free Press)
[10] Nemenman I, Bialek W and de Ruyter van Steveninck R, *Entropy and information in neural spike trains: progress on the sampling problem*, 2004 *Phys. Rev.* E **69** 056111
[11] Paninski L, *Estimation of entropy and mutual information*, 2003 *Neural Comput.* **15** 1191–253
[12] Quah C, *Noise correlations and population coding in salamander retinal ganglion cells*, 2012 *B.S. Thesis*
[13] Quenouille M, *Notes on bias in estimation*, 1956 *Biometrika* **43** 353–60
[14] Schwartz G, Macke Y, Amodei D, Tang H and Berry M J II, *Low-error discrimination using a correlated population code*, 2012 *J. Neurophys.* **108** 1069–88
[15] Shannon C E, *A mathematical theory of communication*, 1948 *Bell System Tech. J.* **27** 379–423
[16] Strong S P, Koberle R and Steveninck R R, *Entropy and information in neural spike trains*, 1998 *Phys. Rev. Lett.* **80** 197–200
[17] Tkacik G, Schneidman E, Berry M J II and Bialek W, *Spin glass models for a network of real neurons*, 2009 arXiv:0912.5409
[18] Tukey J, *Bias and confidence in not quite large samples*, 1958 *Ann. Math. Stat.* **29** 614
[19] Valiant G and Valiant P, *Estimating the unseen: an n/log(n)-sample estimator for entropy and support size, shown optimal via new clts*, 2011 *STOC: 43rd ACM Symp. on Theory of Computing*
[20] Vu V Q, Yu B and Kass R E, *Coverage-adjusted entropy estimation*, 2007 *Stat. Med.* **26** 4039–60
[21] Wolpert D H and Wolf D R, *Estimating functions of probability distributions from a finite set of samples*, 2011 *Phys. Rev.* E **52** 6841–54
[22] Zahl S, *Jackknifing an index of diversity*, 1977 *Ecology* **58** 907–13

J. Stat. Mech. (2013) P03015