

Bayesian inference for a neuronal network from subnetwork data


P.G.L. Porta Mana

Kavli Institute, Trondheim, Norway <piero.mana@ntnu.no>

Draft of 20 July 2019 (first drafted 15 May 2019)

This work shows how to build a maximum-entropy probabilistic model for the total activity of a network of neurons, given only some activity data or statistics – for example, empirical moments – of a *subnetwork* thereof. This kind of model is useful because neuronal recordings are always limited to a very small sample of a network of neurons. The model is applied to two sets of neuronal data available in the literature. In some cases it makes interesting forecasts about the larger network – for example, two low-regime modes in the frequency distribution for the total activity – that are not visible in the sample data or in maximum-entropy models applied only to the sample. For the two datasets, the maximum-entropy probability model applied only to the subnetwork is compared with the marginal probability distribution obtained from the maximum-entropy model applied to the full network. On a linear probability scale no large differences are visible, but on a logarithmic scale the two distributions show very different behaviours, especially in the tails.

Note: Dear Reader & Peer, this manuscript is being peer-reviewed by you. Thank you.

 **comment about the possibility of drawing conclusions about a brain area using different sets of neurons (eg because of recording across many sessions)**

1 Introduction

What correlations are important for the description of the multi-neuronal activity in a specific brain area? How does such activity change when external stimuli or experimental conditions change? Does such activity range over all its mathematically possible values, or only over a subset thereof?

Answering this kind of questions always engages an element of uncertainty. Our answers therefore involve experimental data, such as neuronal recordings from specific brain areas, and probabilities or degrees of belief, based on prior knowledge, about biological conditions and mechanisms that cannot be experimentally ascertained. Such degrees of belief are often formulated as simplified ‘models’ to be mathematically more tractable.

Despite remarkable advances in recording technologies, the best experimental measurements of neuronal activity can still only record a very small sample of neurons compared to the numbers that constitute a functionally distinguished brain region. Many probabilistic models focus on such samples only, somehow neglecting, in their assumptions, that the recorded neurons are a sample from a larger network. This kind of isolation assumptions sometimes escape attention, being subtly hidden in the mathematics. Some probabilistic models try to take also unrecorded neurons into account, but become very complex in doing so.

In the present work we give an answer to this question: How much can the total activity of a large neuronal network be, if we have observed the activity of a very small sample thereof? We'll quantify our degrees of belief about the possible answers by combining, in a straightforward way, the maximum-entropy method and basic sampling relations of the probability calculus. Later on we'll show that our degrees of belief can be quantified by exclusively using the probability calculus. This derivation will provide a more accurate quantification, revealing that the maximum-entropy answer is only a first approximation.

We apply our approach to two concrete data sets: the activity of 65 neurons recorded from a rat's Medial Entorhinal Cortex [✚ ref](#), and the activity of 159 neurons recorded from a macaque's Motor Cortex [✚ ref](#) [✚ add recording length](#). In the first data set, the most interesting finding is that the most probable frequency distribution for the total activity of the full network has two very distinct modes, both at low activities, see fig. 1. The analogous frequency distribution for the second data set doesn't have two modes but still presents one prominent shoulder in its one low-activity mode. Note that these guessed features of the full network aren't observed in the sample, nor can they be inferred by the application of maximum-entropy *to the sample alone*. [✚ Monte Carlo sampling: say something about the deviation from the most probable frequency distribution](#)

[✚ shall we give a two-sentence summary of the main idea here?](#)

The maximum-entropy or minimum-relative-entropy method (jaynes1957jaynes1963sivia1996_r2006; hobsonetal1973; jaynes1985b; grandy1980) has been used for different kinds of estimations of the neuronal activity of various brain areas and about other phenomena of importance to the neurosciences, for example gene and protein interaction (martignonet1995; bohteetal2000; shlensetal2006;

schneidmanetal2006; tkaciketal2006; mackeetal2009b; tkaciketal2009; roudietal2009c; barreiroetal2010; gerwinnetal2010; mackeetal2011b; ganmoreetal2011; cohenetal2011; granotatedgiatal2013; mackeetal2013; tkaciketal2014b; shimazakietal2015; moraetal2015; lezonetal2006; weigtetal2009). This method is often used to test whether some statistics of the data, for example second-order time correlations, is sufficient for quantifying our degree of belief about some quantities of the system. It can be considered as an approximation of probabilistic models based on various assumptions of inferential sufficiency (jaynes1986d_r1996; portamana2017).

✚ Say something more about advantages of such question/answer: for example we can make statements about total activity of brain area even across recordings when sampled neurons aren't the same.

✚ Add this?: from sampling theory we know that important features of the full network may not be visible in a sample because smoothed out. But using sampling theory in the inverse direction we can infer such full-network features from the sample.

✚ To be continued after structure of the rest of the article is clear. Orig intro is on p. 40

Our notation and terminology follow iso standards (iso1993; iso2006; iso2006b) and Jaynes (jaynes1994_r2003) for degrees of belief. We often simply say 'belief' for 'degree of belief'.

2 Model: maximum-entropy and sampling

Let's introduce some context and notation for our problem.

The context we consider is as follows. During an experimental session we have recorded the spiking activities of n neurons for a certain amount of time. These neurons are our 'sample' or 'subnetwork'. Their spikes are binned into T time bins and binarized to $\{0, 1\}$ values in each bin. Call a_t the number of neurons that fire during time bin t , divided by the total number of neurons n ; this is the *normalized total activity* of the sample, or just 'activity' for short. It is also the network-averaged activity of the neurons. Obviously $a_t \in \{0, 1/n, \dots, (n-1)/n, 1\}$; if $a_t = 0$, no neuron spikes during bin t ; if $a_t = 1$, all spike at some point during bin t . For brevity, let's say 'at t ' for 'during time bin t '. From the activities $\{a_t\}$ we can count how often the activity levels $a = 0$, $a = 1/n$, and so on appeared during the recording, obtaining the distribution of measured

relative frequencies (f_a) =: f . We can also consider the sample activity at time bins *outside* of the recorded range. Such activity is unknown to us, of course.

✚ maybe move this paragraph to the intro? Present-day technologies enable the recording of neuronal activity from small brain areas ✚ be more specific. For many animal species, the neurons that are recorded within the area are not – and at present cannot be – specifically chosen from among the rest, owing to several limiting factors; for example, limitations in how precisely electrodes are inserted or neurons are targeted by viruses. In fact, the set of recorded neurons may even change during very long recordings or across experimental sessions. We assume that there’s an area, comprising a network of N neurons, from which other sets of neurons could have been or will be recorded***have the same probability of being recorded, even in other experimental sessions, as the set of n neurons that was actually recorded in this session. This is our ‘full network’. ✚ how about calling it ‘the pool’? The normalized total activity of these N neurons at t is A_t . The relative frequencies of the various activity levels during the recording were (F_A) =: F . We don’t know the values A_t at each t , or the frequency distribution F . We only know for certain that $A_t \in \{0, 1/N, \dots, (N-1)/N, 1\}$ and that $NA_t \geq na_t$ for obvious reasons. For the time being we assume that we know N ; in §*** we discuss the consequences of our lack of precise knowledge about this number.

Our questions concern general features of the total activity A of the full network during and after the recording, and across sessions under the same study conditions. For example: what was its frequency distribution during the recording? How much does this frequency distribution change across sessions? How much total activity should we expect at any time bin during a recording? We cannot answer these questions with certainty; we can only give distributions of probability or degrees of belief over their possible answers. The approach presented here gives such probability distributions.

✚ Move this to intro? We want to stress the usefulness of making quantified guesses about the full-network activity. First of all, this seems to be the primary idea behind recording a sample. Second, it allows us to make comparisons across experimental sessions; such comparisons would be difficult or meaningless if made with the recorded samples, which generally comprise non-overlapping sets of neurons and differ in

size.

The idea behind our approach is easily summarized:

- (a) we build a distribution $p_{\text{me}}(A)$ for the total activity of the full network using the maximum-entropy method;
- (b) the constrained averages used in the maximum-entropy method for the full network are, in turn, determined via sampling theory from the constrained averages for the sample.

Let's discuss these points in detail.

Regarding (a), we assume that you're familiar with the maximum-entropy method. We actually use the *minimum-relative-entropy* method (**hobsonetal1973**), but call it 'maximum-entropy' for brevity. It amounts to a pair of prescriptions: choose the distribution, among those satisfying specific convex constraints, such as fixed expectations, that has minimum relative entropy with respect to a reference distribution, often taken to be the uniform one; and judge those expectations to be equal to measured averages. We add two remarks about this method that are seldom made in the literature. First, the distribution $p_{\text{me}}(A)$ given by this method is the zeroth-order approximation (**debruijn1958_r1961tierneyetal1986; strawderman2000**) of four different distributions for the full network:

- the most probable *frequency* distribution for the total activity across the *recorded* bins,
- the *belief* distribution for the value of the total activity at any time bin among those *recorded*,
- the most probable *frequency* distribution for the total activity in a very long run of *new* time bins,
- the *belief* distribution for the value of the total activity at a *new* time bin.

The maximum-entropy distribution is thus an approximation of our belief distribution about four completely different quantities. Note that the four distributions above numerically differ in higher-order approximations. We discuss their differences in §***. Second, the maximum-entropy method based on the Shannon entropy implicitly makes some assumptions about the probabilities for the long-run frequency distributions (**jaynes1986d_r1996; portamana2009; portamana2017**). We discuss these assumptions in §***.

In our case, to apply the method for the total activity of the full network we need to fix some of the latter's averages, for example its moments. But we don't have any measured moments for the full network to equate the distribution moments to. Here enters point (b): the probability calculus gives an exact, linear relation between the first m moments for the full network and the first m for the sample (**portamanaetal2015**); the ones determine the others and vice versa at every time bin. This relation is a classical result of sampling theory (**whitworth1867_r1965feller1950_r1968jaynes1994_r2003whitworth1897**).

Combining this result with the maximum-entropy prescription 'moments = measured moments' for the sample, we have that the measured moments for the sample determine the moments for the full network:

$$\overbrace{\text{measured moments} \rightarrow \text{sample moments} \rightarrow \text{full-network moments}}^{\text{maximum-entropy prescription}} \quad \underbrace{\hspace{10em}}_{\text{sampling theory}}$$

These two steps are more straightforward if instead of power moments we use *normalized factorial moments* (**potts1953**). The m th normalized factorial moment of a distribution $p(a)$ for the activity of the sample neurons is defined as the average

$$\sum_a \binom{na}{m} \binom{n}{m}^{-1} p(a), \quad 1 \leq m \leq n \quad (1)$$

This moment can be interpreted as the expectation of the number of distinct m -tuples of simultaneously spiking neurons (within a bin's time width), normalized by the number of distinct m -tuples. For example, with $m = 2$, if $na = 4$ neurons spike in a network of $n = 5$, we have $\binom{4}{2} = 6$ distinct pairs of simultaneously spiking neurons, and the total number of distinct pairs is $\binom{5}{2} = 10$. The normalized number of spiking pairs is therefore $6/10$. Note that the first m factorial moments linearly determine the first m power moments and vice versa, because $\binom{na}{m}$ is a polynomial in a of degree m ; so fixing the ones is equivalent to fixing the others. Now we have this extremely convenient property: *the first n normalized factorial moments for a sample and for the full network are numerically identical*:

$$\sum_a \binom{na}{m} \binom{n}{m}^{-1} p(a) = \sum_A \binom{NA}{m} \binom{N}{m}^{-1} p(A), \quad 1 \leq m \leq n, \quad (2)$$

where $p(A)$ is the distribution for the full-network activity.

We can therefore apply the maximum-entropy method to obtain a distribution for the full network, by constraining its m th factorial moment to be equal to the sample's recorded average of $\binom{na}{m} \binom{n}{m}^{-1}$, for as many m as we please with $1 \leq m \leq n$. The constraints on the maximum-entropy distribution $p_{\text{me}}(A)$ for the full network are

$$\frac{1}{T} \sum_t \binom{na_t}{m} \binom{n}{m}^{-1} \equiv \sum_a \binom{na}{m} \binom{n}{m}^{-1} f_a = \sum_A \binom{NA}{m} \binom{N}{m}^{-1} p_{\text{me}}(A) \quad (3)$$

for all m we wish.

The calculation amounts to a convex optimization (**meadetal1984pressetal1988_r2007fangetal1997; boydetal2004_r2009; portamana2017b**) and for the numbers N, n, m considered in the next section it can be done on a modern computer without approximations of normalization constants or potential functions.

The number of moments used with this method depends on the questions and hypotheses that a researcher is exploring; for example, hypotheses of sufficient statistics, such as the sufficiency of pairwise correlations to quantify our degree of belief about network activity. In the present work we only want to introduce the general method without entering into biological questions of this kind.

The particular case in which n moments are constrained is especially important: it corresponds to fully constraining the marginal frequency distribution for the activity of the sample neurons. In this case, our belief about the full-network activity is based on all available measured frequency data. Note that application of the maximum-entropy method *at the sample level* is trivial and meaningless in this case – it just gives back the measured frequency distribution. But application of the method *at the level of the full network* is not trivial.

Using the full frequency distribution of the sample may be a bad idea, however, because the maximum-entropy distribution may become a bad approximation of some of the four distributions described above. It is preferable to use a moderately high number of moments smaller than n . We explain this point in §***.


In the next section we apply the method just described to the data sets from two actual recordings, using *** moments, and discuss the properties of the resulting distributions.

3 Application: two data sets

We apply the approach just described to two data sets publicly available in the literature:

- The first, from **stensolaetal2012**, consists of $n = 65$ neurons (27 of which classified as grid cells) from rat Medial Entorhinal Cortex, recorded for about 20 minutes. Their spikes are binned into $T = 417\,641$ bins of 3 ms width.
- The second, from **rostamietal2016_r2017**, consists of $n = 159$ neurons from macaque Motor Cortex, recorded for about 15 minutes. Their spikes are binned into $T = 300\,394$ bins of 3 ms width.

For concreteness's sake we'll consider the maximum-entropy distribution $p_{\text{me}}(A)$ as *the most probable frequency distribution* for the full-network activity during the recording; but remember that it is also the approximation of three other distributions, as discussed in the previous section.

We first calculate the distribution by using six moments. This number already provides almost as much information as the full frequency distribution of the sample, and at the same time illustrates the use of the approach in questions of statistic sufficiency (typically limited to two or three moments). Figure 1 shows the resulting densities (that is, distribution $\times N$) for full-network sizes $N = n$, $N = 1\,000$, $N = 5\,000$, $N = 10\,000$  *motivate?*. The case $N = n$ corresponds to applying the maximum-entropy method at the sample level; it can be observed that with six moments it reproduces almost exactly the measured frequency distribution.

The distribution for the full-network is sharper than the measured frequency distribution for the sample; the sharper the larger N is. Most remarkably, it has two distinct low-activity modes for the first data set. But also the second data set presents, upon closer inspection, a small shoulder on the right of the mode, suggestive of two activity regimes. We frame no hypotheses about the biological cause of these two modes (they could stem from the presence of different kinds of cells or modules). These features are clearly not present in the sample or in the maximum-entropy distribution at the sample level. The application of the probability calculus thus reveals interesting possible features of the full network.

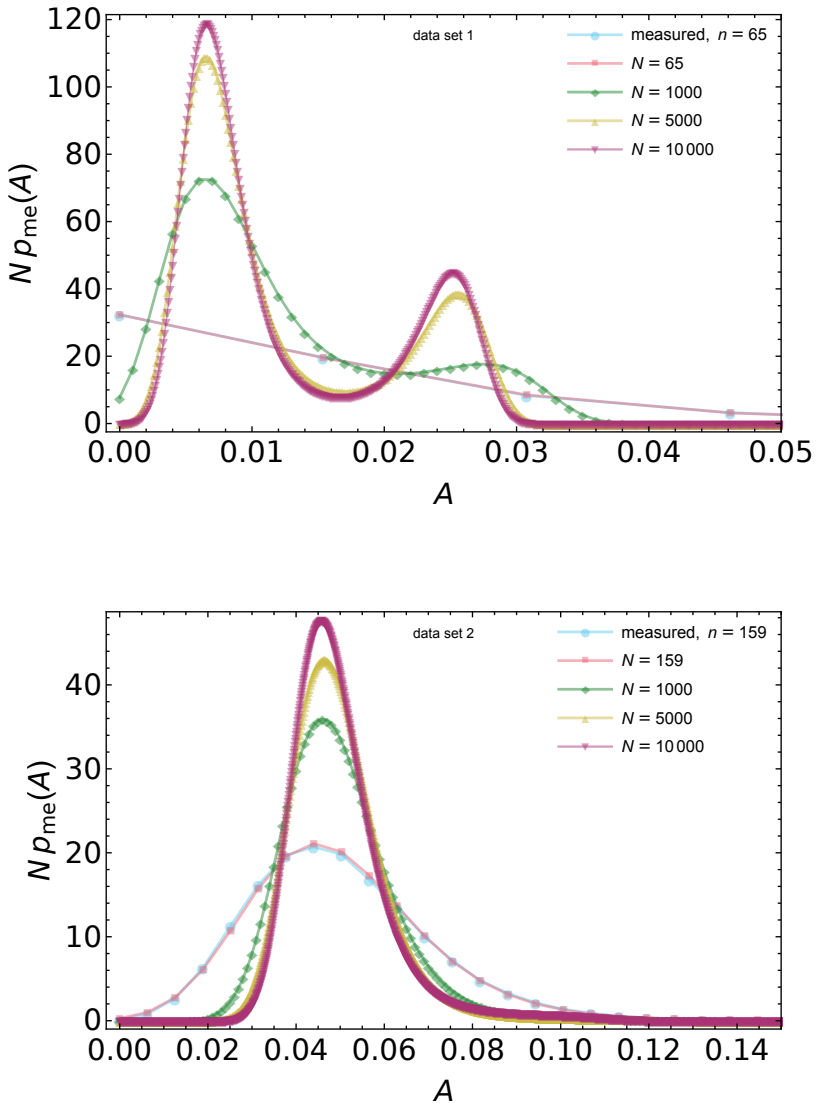


Figure 1 ***

To illustrate how our approach can be applied to studies of sufficient statistics, fig. 2 shows the results for two constrained moments (equivalent to constraining means and correlations only) and for four constrained moments, with $N = 10\,000$. In either data set we obtain two very distinct distributions. For the first data set in particular, four moments lead to a bimodal distribution, whereas two to a unimodal one, showing that two moments would not be sufficient statistics. For comparison, the two distributions obtained applying maximum-entropy *at the sample level* are shown as an inset in the plot for the first data set: their differences aren't so glaring as in the full-network application.

4 Marginalization to the sample level: the issue with pairwise correlations and other statistics

In the neurosciences the maximum-entropy method is often applied to questions of sufficient statistics. In this section we discuss how our proposed application bears on this kind of questions. For concreteness we use an example with specific statistics: first and second moments (equivalent to mean and correlations), versus first-to-fourth moments. But our discussion holds for any sets of statistics.

The question is as follows. We want to assess our degree of belief about the frequencies of the activities of the sample, or about the sample activity in a new time bin. For this assessment we should use all measured data we have. But it sometimes happens that dropping part of the data – for example, the measured third- and higher-order moments – leaves our assessment almost unchanged: the dropped data are *informationally irrelevant*, or almost so. The remaining data – for example, first and second moments – are *informationally sufficient*. This informational quasi-sufficiency is interesting because it may hint at peculiar biological or physical mechanisms or dependencies.

We can approximately check how much some data are irrelevant by simply dropping them from the conditional of our degree of belief, and see if the latter changes appreciably. When our degree of belief is built with the maximum-entropy method, we simply drop some of the data used as constraints and see how the resulting distribution changes. For example, we may only retain the first two moments (means and correlations) or the first four moments. In the neurosciences this check

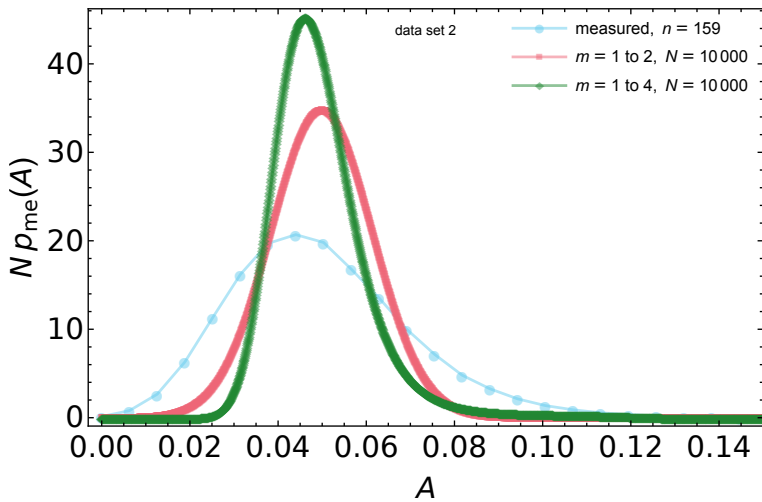


Figure 2 ***

is often done with maximum-entropy distributions constructed *for the sample*, ignoring the full network it's sampled from.

The method introduced in the present work reveals an issue with this application at the sample level, however. The issue doesn't come from the method itself but from a general fact of probability theory. If a probability distribution has some sufficient statistics (**bernardoetal1994fortinietal2000**), then its marginals *cannot* have the same sufficient statistics, and vice versa, except for trivial cases such as uniform distributions. This generally also holds among different marginals: if a marginal has some sufficient statistics, another won't have it. This impossibility is known in statistical mechanics: marginals of Gibbs states aren't Gibbs states (**maesetal1999**). Mathematically it corresponds to the general impossibility of solving a system of n equations in m unknowns when $n > m$ (**portamanaetal2015portamanaetal2018b**).

In our example this means that if means and pairwise correlations or the first four moments are informationally sufficient for a particular sample from a brain area, then they *cannot* be sufficient for the full network of neurons in that area, or for a different sample. Vice versa, if we assume that they are informationally sufficient for the full network, then we must expect them *not* to be sufficient for any recorded sample thereof. Now, questions about sufficient statistics are meant to be addressed to a whole brain area, not just to some specific, casually recorded sample. Thus, paradoxically, maximum-entropy methods applied at the sample level give an answer opposite to the one we might think they're giving.

Questions about informational sufficiency are therefore correctly addressed by applying the maximum-entropy method at the full-network level. If a comparison with the measured frequencies of the sample activities is desired, then the marginal distribution to the sample size should be used.

But does the application at the full-network level lead to appreciably different results from that at the sample level? after all we're interested in an approximate informational sufficiency, not in an analytically exact one. A correct answer can only be given case by case; fig. 3 gives a graphical answer to this question in the case of our first data set. The upper plot shows the maximum-entropy distributions for a full network of 10 000 neurons, constructed from two moments (---) and from four moments (—). The lower plot shows the maximum-entropy distributions for the sample, from two moments (▼) and from four moments (□), and the

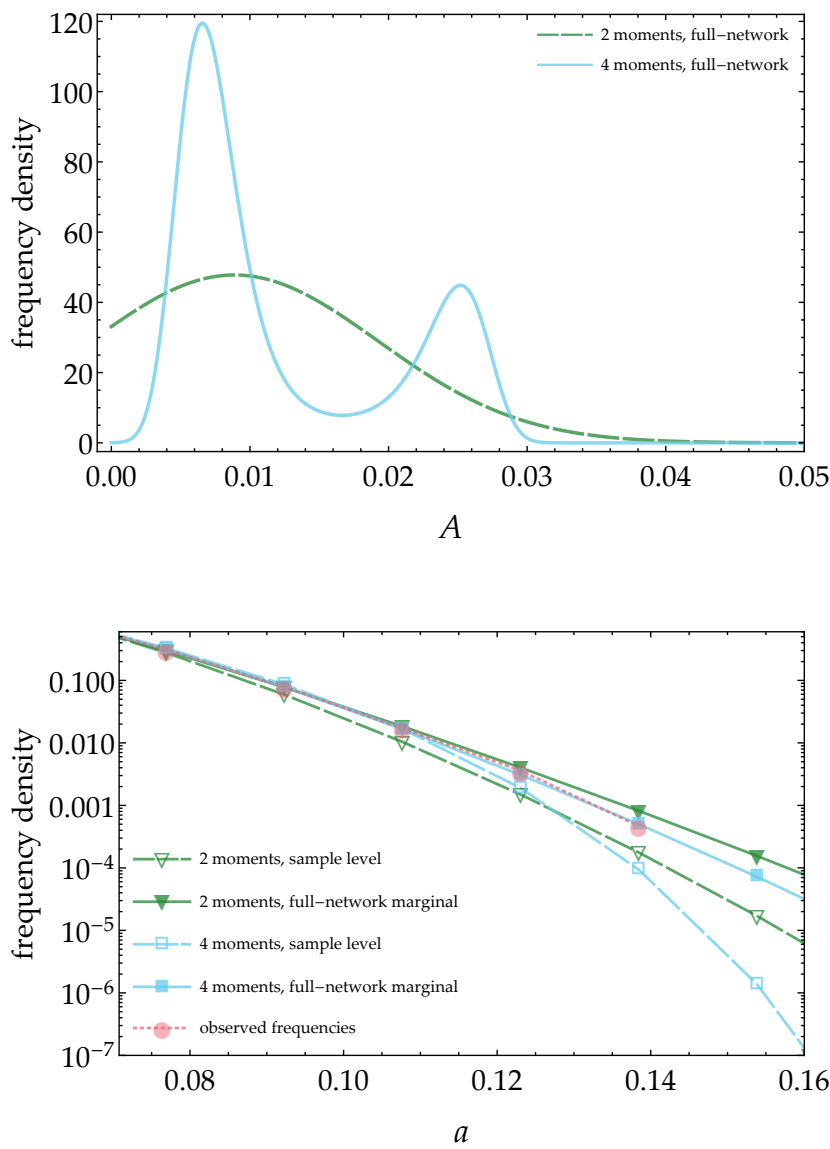


Figure 3 ***

distributions obtained by marginalizing the full-network distribution to the sample size, from two moments (▼) and from four moments (■). The measured frequency distribution (●) is also shown. We note the following:

- The application to the full-network (upper plot) leads to two completely different distributions (even different number of modes); clearly the two sets of statistics are not even approximately equivalent for inferential purposes.
- The application at the sample level leads to deceptively similar distributions (▼ and ■), which can only be distinguished on a logarithmic scale. This could lead to the erroneous conclusion that the two statistics are approximately equivalent.
- The difference between marginals from the full-network distribution and the distributions obtained at the sample level (filled vs empty markers) is larger than the difference between different statistics (triangles vs squares).
- The marginals of the application to the full network are closer to the measured frequencies than the distributions obtained at the sample level (although this closeness is not a valid criterion for their goodness).

Therefore, the maximum-entropy application at the sample level and at the full-network level lead to different results; the latter is not only more meaningful but also superior because it shows clearly the different informational sufficiency of the two sets of statistics.

5 Assumptions and corrections: beyond the maximum-entropy approximation

In § 2 we mentioned that maximum-entropy distributions come from a zeroth-order Laplace approximation of the densities obtained from the probability calculus. Zeroth-order means that we are simply considering the mode of such densities. Let's be more concrete.

Consider the question: What is the relative-frequency distribution for the different activity levels for the full-network, across the recorded bins? Call this distribution $(F_A) =: F$: activity level A appeared in $T \times F(A)$ out of T time bins. We don't know F , and our beliefs about its possible

values have a distribution $p(F | DI)$ conditional on data D and prior knowledge I , which we approximate with a continuous distribution

$$p(F | DI) dF. \quad (4)$$

In appendix*** we give a summary of how this distribution is derived. Note that this is a probability distribution over frequency distributions, so we must be careful with terminology to avoid confusion.

Now consider the related question: What is the activity A_t of the full network at some specific time bin t ? If we don't know the exact sequence of activities during the recording then we don't know A_t ; if we know the frequency distribution F of the activity levels, then our degree of belief about A_t is F_{A_t} , out of combinatorial reasons. But if we are also uncertain about F , with belief distribution (4), then our degree of belief about A_t by marginalization is

$$p(A_t | DI) = \int F_{A_t} p(F | DI) dF; \quad (5)$$

our belief distribution for A_t is thus simply the mean $\int F p(F | DI) dF$.

Formulae (4) and (5) show that the most probable frequency distribution F (if it's unique) and the probability distribution for A_t are generally different: the former is the mode of $p(F | DI)$, the latter its mean.

If the probability density $p(F | DI)$ is extremely peaked at some frequency distribution F^* , it can be approximately treated as a delta density,

$$p(F | DI) dF \approx \delta(F - F^*) dF, \quad (6)$$

and consequently the probability for the activity at bin t , by eq. (5), is simply the mode F^* itself:

$$p(A_t | DI) \approx F_{A_t}^*. \quad (7)$$

The maximum-entropy distribution is exactly this mode F^* , for *particular* kinds of probability densities $p(F | DI)$ which we'll discuss shortly. Formulae (6) and (7) show why the maximum-entropy distribution is an approximate answer to two different questions.

If the probability density $p(F | DI)$ is not sufficiently peaked, however, its mode and mean can be quite different. The maximum-entropy approximation can then be inadequate for two reasons. First, despite being the mode it can be a poor representative of all the possible frequency

distributions around the mode. Second, it can be a bad approximation of the probability distribution (5).

Let's therefore examine the full density $p(F | DI)$ given by the probability calculus. Its derivation is given in appendix***; here we discuss its mathematical features and examine the answers it provides. Its expression,

$$p(F | DI) \propto \exp \left[-\frac{|\mathbf{CF} - \hat{\mathbf{c}}|^2}{2\sigma^2} - \frac{|\Delta \mathbf{F}|^2}{2\delta^2} - L H(\mathbf{F}; \mathbf{R}) \right], \quad (8)$$

consists of three terms:

- The density

$$\exp[-L H(\mathbf{F}; \mathbf{R})] \quad (9)$$

expresses our pre-data beliefs about the possible frequency distributions. Here $H(\mathbf{F}; \mathbf{R}) := \sum_A F_A \ln(F_A/R_A)$ is the relative entropy or discrimination information (**kullback1987; jaynes1963; hobson1969; hobsonetal1973**), \mathbf{R} is a reference distribution, and L a positive parameter. This density expresses a belief proportional to the number of ways in which the frequency distribution \mathbf{F} can be realized if we assume that the activities appear with frequencies \mathbf{R} in the long run, as can be seen using Stirling's approximation. For \mathbf{R} we choose the uniform distribution, but any biologically sensible distribution, such as one decreasing exponentially with the activity, leads to the same final results. The parameter L roughly represents the number of data points necessary to change our initial belief, and we set it equal to 50. This density is similar to an 'entropic prior' (**rodriguez1991; skilling1998; catichaetal2004; portamana2017neumann2007**).

- The previous density is corrected by the normal density $\exp\left(-\frac{|\Delta \mathbf{F}|^2}{2\delta^2}\right)$, where the matrix Δ is the discrete equivalent of a fourth-order derivative operator. This density expresses our belief, hinging on biological arguments, that the frequency distribution should be somewhat smooth, without huge discontinuities between one activity level and the next. The fourth derivative is chosen because it still allows for the presence of modulated maxima and minima more than derivatives of lower order do. The standard deviation δ is such that the smoothness of the maximum-entropy distribution, taken as a touchstone, is within two standard deviations.

• The normal density $\exp\left(-\frac{|\mathbf{CF}-\hat{\mathbf{c}}|^2}{2\sigma^2}\right)$ is the likelihood of the frequency F in view of our data. The matrix \mathbf{C} expresses the linear constraints on F , and the tuple $\hat{\mathbf{c}}$ the measured values of such constraints. For example, if we consider the full frequency distribution f of the sample, then \mathbf{C} effects the marginalization to a sample of size n through a hypergeometric distribution (sampling without replacement):

$$C_{a,A} := \binom{n}{na} \binom{N-n}{NA-na} \binom{N}{NA}^{-1} \quad (10)$$

and $\hat{\mathbf{c}} := f$ are the measured frequencies. If we only consider some factorial moments, instead, then

$$C_{m,A} := \binom{NA}{m} \binom{N}{m}^{-1} \quad (11)$$

as for eq. (2), and $\hat{\mathbf{c}}$ is the tuple of measured factorial moments. The standard deviations σ reflect our beliefs about the measurement results, which can come for example from the spike-sorting procedure. Their values are discussed in appendix***.

The post-data belief density (8) has a maximum-entropy distribution as a mode approximation when σ is small, L is large, and σL is small: the mode will be the constrained F that minimizes the relative entropy $H(F; \mathbf{R})$. It should be noted that a pre-data belief different from the entropic one (9) will lead to a different mode approximation. For example, a Dirichlet density (which is equivalent to (9) with F and \mathbf{R} exchanged) would lead to a maximum-entropy distribution with *Burg's* entropy (burg1975) instead of Shannon's (jaynes1986d_r1996; portamana2009).

The properties of the belief density (8) in the case of our first data set are shown in fig. 6 for $N = 1\,000$ and in fig. 7 for $N = 5\,000$. Only the first six factorial moments were considered as our data, but it turns out that they are informationally sufficient, leading to practically the same result as specifying the full frequency distribution for the sample.

The top plots in the figures show the mean $\int F p(F | DI) dF$ (—), which is also our belief about the activity at any single time bin, eq. (5), is clearly different from the maximum-entropy mode. The latter is therefore not a good approximation for the full density (8), which is quite broad, as can be seen from the ten samples (—) drawn from it. This broadness

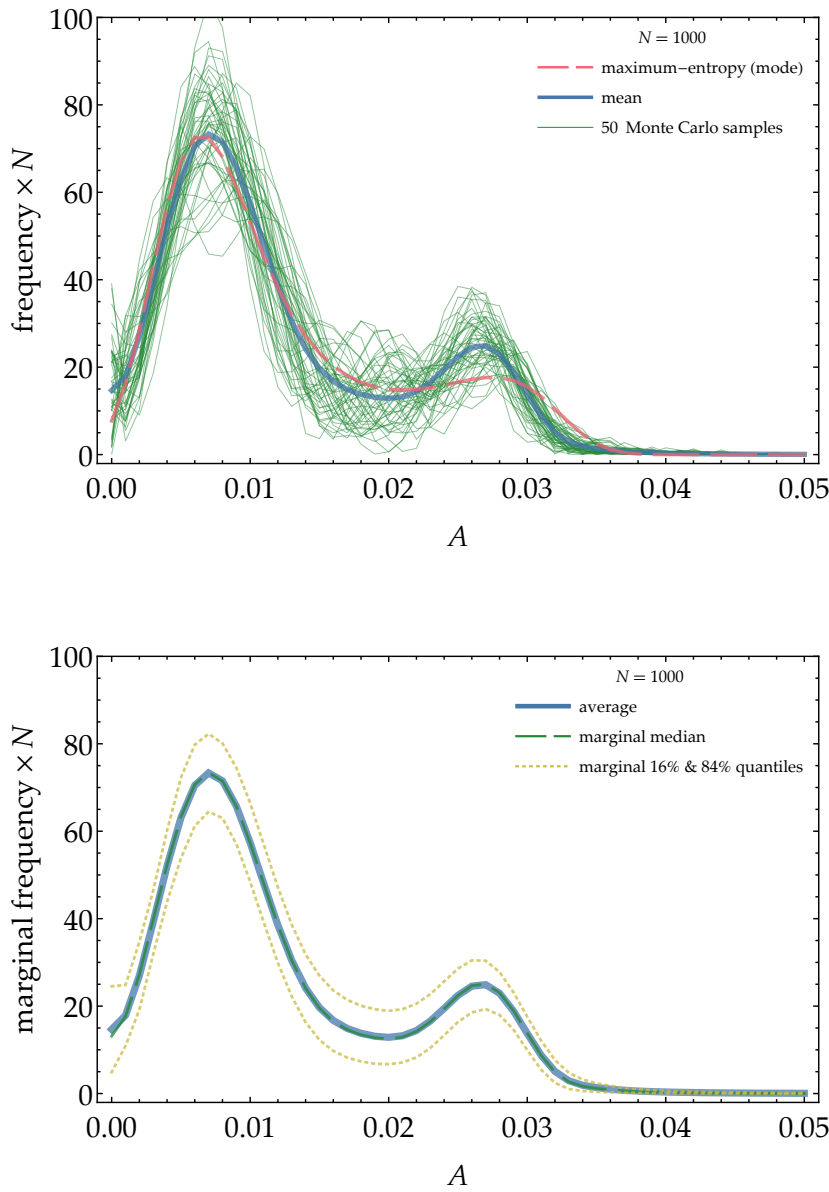


Figure 4 Entropic prior

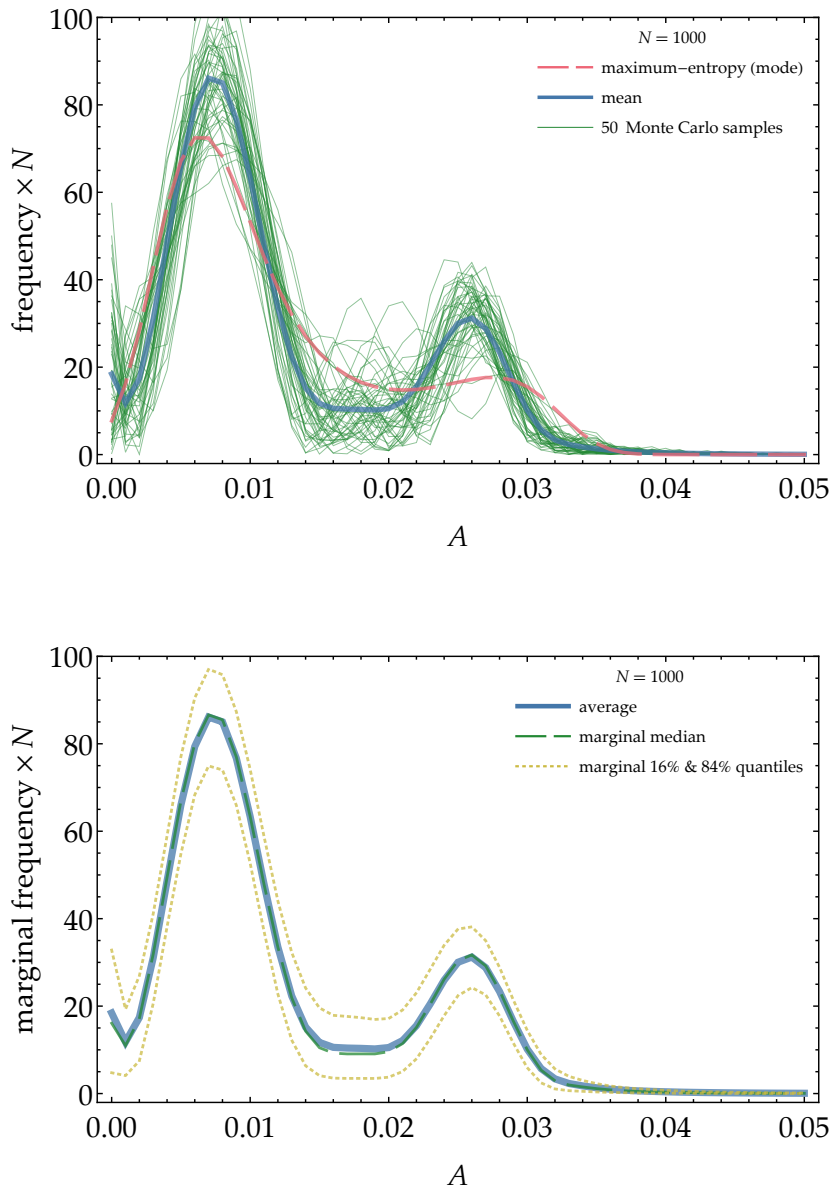


Figure 5 Prior uniform in dF

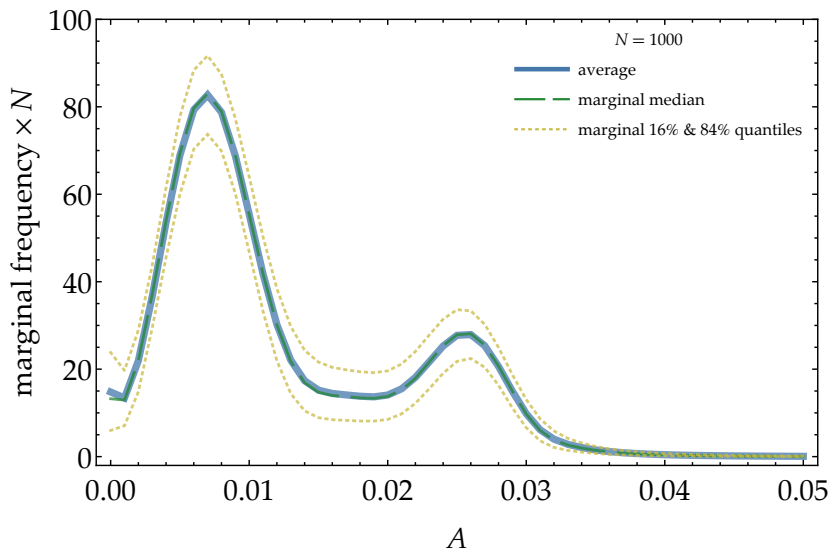
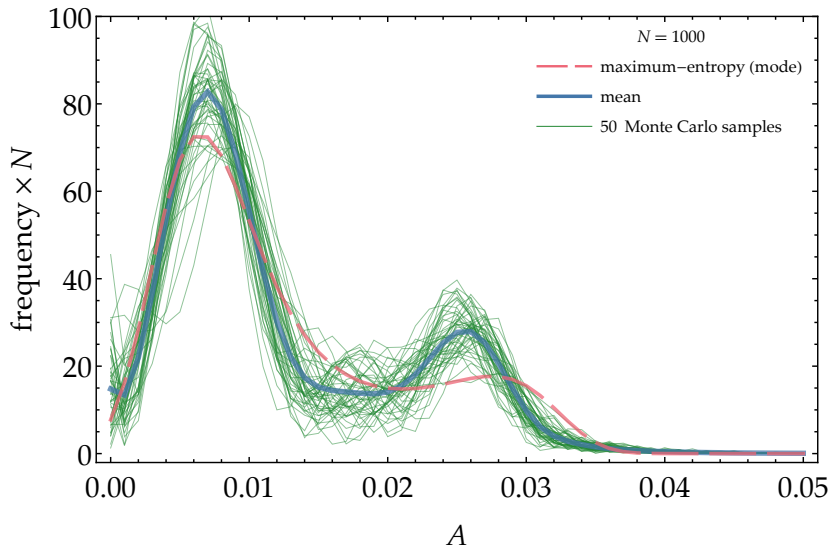


Figure 6 Dirichlet prior (entropic with reversed relative entropy)

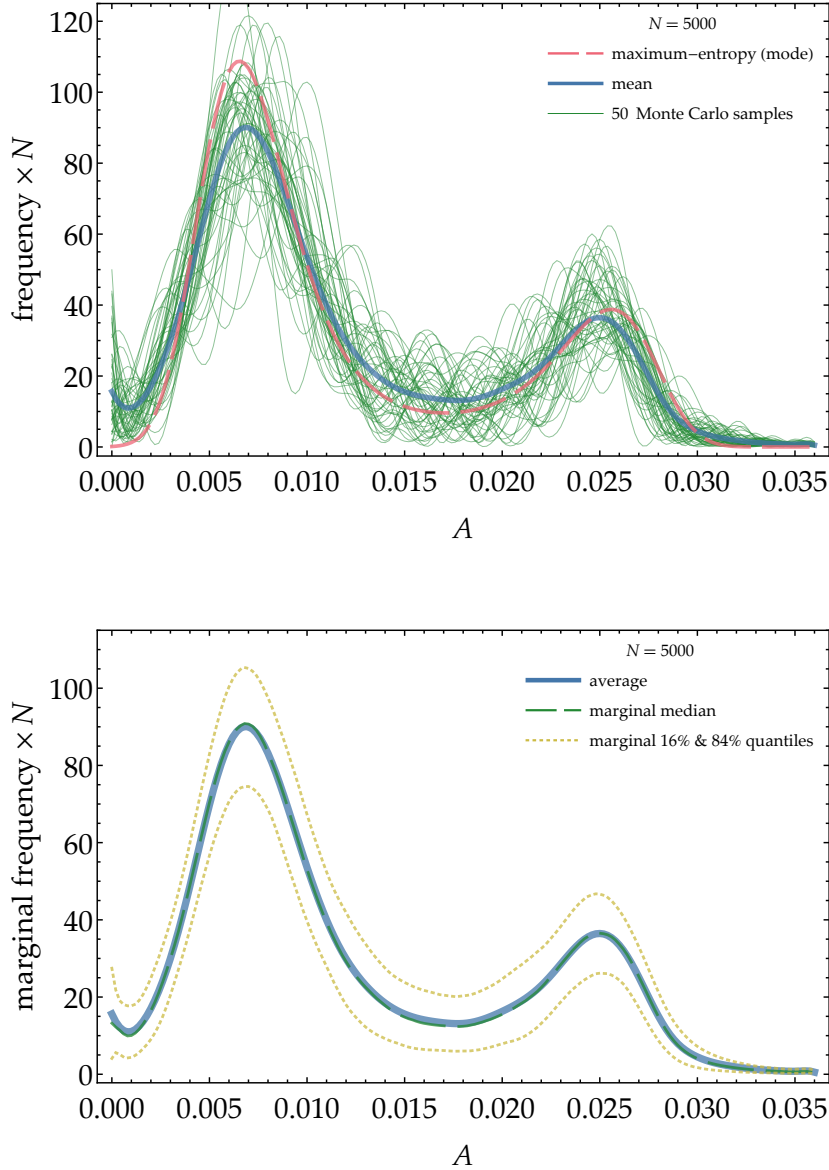


Figure 7 Entropic prior

is not surprising: we're making guesses about thousands of neurons from the observation of fewer than 100. Yet, the mean and the maximum-entropy mode are qualitatively similar. In fact, for $N = 1\,000$ the second peak of the mean is even higher than the corresponding maximum-entropy one. The samples also show that generally two high-frequency regions of activity are expected, especially for $N = 5\,000$.

The bottom plots in the figures show our belief distributions for the marginal frequency of each activity level, $p(F_A \mid DI)$, with the means (—), medians (---), and 16% and 84% quantiles (···· corresponding to one standard deviation for a normal distribution). The coincidence of means and medians indicate that these marginals are very close to normal distributions. This happens thanks to the marginalization, which integrates out $N - 2$ quantities. Note, however, that the full distribution for F is far from normal, as indicated by the pronounced differences between mode and mean.

These results caution us against taking maximum-entropy solutions too literally. The full probabilistic analysis allows us to see more clearly the extent of our uncertainty and to make better-informed guesses.

6 Summary and discussion

 OLD TEXT

7 Mathematical development

$$\begin{aligned} \mathbb{E}\left[\frac{s(s-1)}{n(n-1)} \mid I\right] &= \mathbb{E}\left[\frac{S(S-1)}{N(N-1)} \mid I\right], \\ \mathbb{E}\left[\frac{s(s-1)(s-2)}{n(n-1)(n-2)} \mid I\right] &= \mathbb{E}\left[\frac{S(S-1)(S-2)}{N(N-1)(N-2)} \mid I\right], \end{aligned} \quad (12a)$$

$$\text{and in general } \mathbb{E}\left[\binom{s}{r} \binom{n}{r}^{-1} \mid I\right] = \mathbb{E}\left[\binom{S}{r} \binom{N}{r}^{-1} \mid I\right], \quad r < N.$$

Note that from the first r factorial moments we can calculate the first r power moments and vice versa; but the simple equalities above don't hold for the power moments (**portamanaetal2015**).

$$p(s \mid I) = \sum_s \binom{n}{s} \binom{N-n}{S-s} \binom{N}{S}^{-1} p(S \mid I). \quad (12b)$$

Suppose we have recorded the firing activity of a hundred neurons, sampled from a particular brain area. What are we to do with such data? Gerstein et al. (**gersteinetal1985**) posed this question very tersely (our emphasis):

The principal conceptual problems are (1) *defining cooperativity or functional grouping* among neurons and (2) *formulating quantitative criteria* for recognizing and characterizing such cooperativity.

These questions have a long history, of course; see for instance the 1966 review by Moore et al. (**mooreetal1966**). The neuroscientific literature has offered several mathematical definitions of 'cooperativity' or 'functional grouping' and criteria to quantify it.

One such quantitative criterion relies on the maximum-entropy or relative-maximum-entropy method (**jaynes1957**; **jaynes1963**; **hobsonetal1973**; **sivia1996_r2006**; **meadeetal1984**). This criterion has been used in neuroscience at least since the 1990s, applied to data recorded from brain areas as diverse as retina and motor cortex (**mackay1991**; **martignonetal1995**; **bohteetal2000**; **amarietal2003**; **schneidmanetal2006**; **shlensetal2006**; **mackeetal2009b**; **roudietal2009c**; **tkaciketal2009**; **gerwinnetal2009**; **mackeetal2011**; **mackeetal2011b**; **ganmoreetal2011**; **granotatedgietal2013**; **tkaciketal2014b**; **moraetal2015**; **shimazakietal2015**), and it has been subjected to mathematical and conceptual scrutiny (**tkaciketal2006**; **roudietal2009**;

roudietal2009b; barreiroetal2010; barreiroetal2010b; mackeetal2013; rostamietal2016_r2017).

‘Cooperativity’ can be quantified and characterized with maximum-entropy methods in several ways. The simplest way roughly proceeds along the following steps. Consider the recorded activity of a sample of n neurons.

1. The activity of each neuron, a continuous signal, is divided into T time bins and binarized in intensity, and thus transformed into a sequence of digits ‘0’s (inactive) and ‘1’s (active) (**caianiello1961; caianiello1986**).

Let the variable $a_i(t) \in \{0, 1\}$ denote the activity of the i th sampled neuron at time bin t . Collectively denote the n activities with $\mathbf{a}(t) := (a_1(t), \dots, a_n(t))$. The network-averaged activity at that bin is $a(t) := \sum_i a_i(t)/n$. If we count the number of distinct pairs of active neurons at that bin we combinatorially find $\binom{na(t)}{2} \equiv na(t)[na(t) - 1]/2$. There can be at most $\binom{n}{2}$ simultaneously active pairs, so the network-averaged pair activity is $\overline{aa}(t) := \binom{n}{2}^{-1} \binom{na(t)}{2}$. With some combinatorics we see that the network-averaged activity of m -tuples of neurons is

$$\underbrace{\overline{a \cdots a}}_{m \text{ terms}}(t) = \binom{n}{m}^{-1} \binom{na(t)}{m}. \quad (13)$$

For brevity let us agree to simply call ‘activity’ the average a , ‘pair-activity’ the average \overline{aa} , and so on.

2. Construct a sequence of relative-maximum-entropy distributions for the activity a , using this sequence of constraints:
 - the time average of the activity: $\widehat{a} := \sum_t a(t)/T$;
 - the time averages of the activity and of the pair-activity $\widehat{\overline{aa}} := \sum_t \overline{aa}(t)/T$;
 - ...
 - the time averages of the activity, of the pair-activity, and so on, up to the k -activity.

Call the resulting distributions $p_1(a), p_2(a), \dots, p_k(a)$. The time-bin dependence is now absent because these distributions can be interpreted as referring to any one of the time bins t , or to a new time bin (in the future or in the past) containing new data.

We also have the empirical frequency distribution of the total activity, $f(a)$, counted from the time bins.

3. Now compare the distributions above with one another and with the frequency distribution, using some probability-space distance like the relative entropy or discrimination information (**kullback1987; jaynes1963; hobson1969; hobsonetal1973**). If we find, say, that such distance is very high between p_1 and f , very low between p_2 and f , and is more or less the same between all p_m and f for $m \geq 2$, then we can say that there is a ‘pairwise cooperativity’, and that any higher-order cooperativity is just a reflection or consequence of the pairwise one. The reason is that the information from higher-order simultaneous activities did not lead to appreciable changes in the distribution obtained from pair activities.

The protocol above needs to be made precise by specifying various parameters, such as the width of the time bins or the probability distance used.

We hurry to say that the description just given is just *one* way to quantify and characterize cooperativity and functional grouping, not *the only* way. It can surely be criticized from many points of view. Yet, it is quantitative and bears a more precise meaning than an undefined, vague notion of ‘cooperativity’. Two persons who apply this procedure to the same data will obtain the same numbers. Different protocols can be based on the maximum-entropy method, for instance protocols that take into account the activities or pair activities of specific neurons rather than network averages, or even protocols that take into account time dependence.

The purpose of the present work is not to assess the merits of maximum-entropy methods with respect to other methods. Its main purpose is to show that there is a problem in the way the maximum-entropy method itself, as sketched above, is applied to the activity of the recorded neurons. We believe that this problem is at the root of some quirks about this method that were pointed out in the literature (**roudi2009b**). This problem extends also to more complex versions of the method, possibly except versions that use ‘hidden’ neurons (**smolensky1986; kulkarnietal2007; huang2015; dunnetal2017**). The problem is that the recorded neurons are a *sample* from a larger, unrecorded network, but the maximum-entropy method as applied above is treating them as isolated from the rest of the brain. Hence, the results it provides cannot

be rightfully extrapolated. We will give a mathematical proof of this. Let us first analyse this issue in more detail.

Suppose that the neurons were recorded with electrodes covering an area of some square millimetres (**berenyi et al 2014**). This recording is a sample of the activity of the neuronal network under the recording device, which can amount to tens of thousands of neurons (**abeles 1991**). We could even consider the recorded neurons as a sample of a brain area more extended than the recording device.

The characterization of the cooperativity of the recorded sample would have little meaning if we did not expect its results to generalize to a larger, unrecorded network – at the very least the network under the recording device. In other words, we expect that the conclusions drawn with the maximum-entropy methods about the sampled neurons should somehow extrapolate to unrecorded neurons in some larger area, from which the recorded neurons were sampled. In statistical terms we are assuming that the recorded neurons are a *representative sample* of some larger neuronal network. Probability theory tells us how to make inferences from a sample to the larger network from which it is sampled (see references below).

We can apply the maximum-entropy method to the sample, as described in the above protocol, to generate probability distributions for the activity of the sample. But, given that our sample is representative of a larger network, we can also apply the maximum-entropy method to the larger (unrecorded) network. The constraints are the same: the time averages of the sampled data, since they constitute representative data about the larger network as well. The method thus yields a probability distribution for the larger network, and the distribution for the sample is obtained by marginalization. The problem is that *the distributions obtained from these two applications differ*. Which choice is most meaningful?

In this work we develop the second way of applying the maximum-entropy method, at the level of the larger network, and show that its results differ from the application at the sample level. We also consider the case where the size of the larger network is unknown.

To apply the maximum-entropy method to the larger, unsampled network, it is necessary to use probability relations relevant to sampling (**ghosh et al 1997 freedman et al 1978_r2007 gelman et al 1995_r2014 jaynes 1994_r2003**). The relations we present are well-known in survey sampling and in the pedagogic problem of drawing from an urn without replacement,

yet they are somewhat hard to find explicitly written in the neuroscientific literature. We present and discuss them in the next section. A minor purpose of this paper is to make these relations more widely known, because they can be useful independently of maximum-entropy methods.

The notation and terminology in the present work follow ISO and ANSI standards (**iso1993**; **ieee1993**; **nist1995**; **iso2006**; **iso2006b**) but for the use of the comma ‘,’ to denote logical conjunction. Probability notation follows Jaynes (**jaynes1994_r2003**). By ‘probability’ we mean a degree of belief which ‘would be agreed by all rational men if there were any rational men’ (**good1966**).

8 Probability relations between network and sample

We have already introduced the notation for the sample neurons. We introduce an analogous notation for the N neurons constituting the larger network, but using the corresponding capital letters: $A_i(t)$ is the activity of the i th neuron at time bin t , $A(t) := \sum_i A_i(t)/N$ is the activity at that bin averaged over the larger network, and so on.

The probability relations between sample and larger network are valid at every time bin. As we mentioned above, the maximum-entropy distribution refers to any time bin or to a new bin. For these reasons we will now omit the time-bin argument ‘(t)’ from our expressions.

If K denotes our state of knowledge – the evidence and assumptions backing our probability assignments – our uncertainty about the full activity of the larger network is expressed by the joint probability distribution

$$p(A_1, A_2, \dots, A_N | K) \quad \text{or} \quad p(A | K), \quad A \in \{0, 1\}^N. \quad (14)$$

Our uncertainty about the state of the sample is likewise expressed by

$$p(a_1, a_2, \dots, a_n | K) \quad \text{or} \quad p(a | K), \quad a \in \{0, 1\}^n. \quad (15)$$

The theory of statistical sampling is covered in many excellent texts, for example Ghosh & Meeden (**ghoshetal1997**) or Freedman, Pisani, & Purves (**freedmanetal1978_r2007**); summaries can be found in Gelman et al. (**gelmanetal1995_r2014**) and Jaynes (**jaynes1994_r2003**).

We need to make an initial probability assignment for the state of the full network before any experimental observations are made. This initial assignment will be modified by our experimental observations, and these can involve just a sample of the network. Our state of knowledge and initial probability assignment should reflect that samples are somehow representative of the whole network.

In this state of knowledge, denoted I , we know that the neurons in the network are biologically or functionally similar, for example in morphology or the kind of input or output they receive or give. But we are completely ignorant about the physical details of the individual neurons. Our ignorance is therefore symmetric under permutations of neuron identities. This ignorance is represented by a probability distribution that is symmetric under permutations of neuron identities; such a distribution is usually called *finitely exchangeable* (ericson1969ghoshetal1997). We stress that this probability assignment is just an expression of the symmetry of our *ignorance* about the state of the network, not an expression of some biologic or physical symmetry or identity of the neurons.

The *representation theorem for finite exchangeability* states that, in the state of knowledge I , the symmetric distribution for the full activity is completely determined by the distribution for its network-average:

$$p(A | I) \equiv \sum_A p(A | A, I) p(A | I) = \binom{N}{NA}^{-1} p(A | I). \quad (16)$$

The equivalence on the left is just an application of the law of total probability; the equality on the right is the statement of the theorem. This result is intuitive: owing to symmetry, we must assign equal probabilities to all $\binom{N}{NA}$ activity vectors with NA active neurons; the probability of each activity vector is therefore given by that of the average activity divided by the number of possible vector values. Proof of this theorem and generalizations to non-binary and continuum cases are given by de Finetti (definetti1959b), Kendall (kendall1967), Ericson (ericson1976), Diaconis & Freedman (diaconis1977; diaconisetal1980), Heath & Sudderth (heathetal1976).

Our uncertainties about the full network and the sample are connected via the conditional probability

$$p(a | A, I) = \binom{n}{na} \binom{N-n}{NA-na} \binom{N}{NA}^{-1} =: G_{aA}, \quad (17)$$

which is a hypergeometric distribution, typical of ‘drawing without replacement’ problems. The combinatorial proof of this expression is in fact the same as for this class of problems (jaynes1994_r2003ross1976_r2010feller1950_r1968).

Using the conditional probability above we obtain the probability for the activity of the sample:

$$p(a | I) = \sum_A p(a | A, I) p(A | I) = \sum_A G_{aA} p(A | I). \quad (18)$$

It should be proved that the probability distribution for the full activity of the sample is also symmetric and completely determined by the distribution of its network-averaged activity:

$$p(a | I) = \binom{n}{na}^{-1} p(a | I). \quad (19)$$

This is intuitively clear: our initial symmetric ignorance should also apply to the sample. The distribution for the sample (18) indeed satisfies the same representation theorem (16) as the distribution for the full network.

The conditional probability $p(a | A, I) \equiv G_{aA}$, besides relating the distributions for the network and sample activities via marginalization, also allows us to express the expectation value of any function of the sample activity, c_a , in terms of the distribution for the full network, as follows:

$$E(c | I) \equiv \sum_a c_a p(a | I) = \sum_a c_a \sum_A G_{aA} p(A | I) = \sum_A (\sum_a c_a G_{aA}) p(A | I), \quad (20)$$

where the second step uses eq. (18). The last expression shows that the expectation of the function c_a is equal to the expectation of the function $c^*(A) := \sum_a c_a G_{aA}$.

The final expression in eq. (20) is important for our maximum-entropy application: the requirement that the function c , defined for the sample, have a value \hat{c} obtained from observed data, *translates into a linear constraint for the distribution of the full network*:

$$\hat{c} = E(c | I) \equiv \sum_A (\sum_a c_a G_{aA}) p(A | I). \quad (21)$$

In particular, when the function c is the m -activity of the sample, $c_a = \overline{a \cdots a} \equiv \binom{na}{m} / \binom{n}{m}$, we find

$$\begin{aligned} E(\underbrace{\overline{a \cdots a}}_{m \text{ factors}} | I) &\equiv \sum_a \binom{n}{m}^{-1} \binom{na}{m} p(a | I) = \\ &\quad \binom{N}{m}^{-1} \sum_A \binom{NA}{m} p(A | I) \equiv E(\underbrace{\overline{A \cdots A}}_{m \text{ factors}} | I), \quad (22) \end{aligned}$$

that is, *the expected values of the m -activities of the sample and of the full network are equal*. The proof of the middle equality uses the expression for the m th factorial moment of the hypergeometric distribution and can be found in **potts1953**. Similar relations can be found for the raw moments $E(a^m)$ and $E(A^m)$, which can be written in terms of the product expectations using eq. (13).

Thus, in a maximum-entropy application, when we require the expectation of the m -activity of a sample to have a particular value, we are also requiring the expectation of the m -activity of the full network to have the same value.

These expectation equalities between sample and full network should not be surprising: we intuitively *expect* that the proportion of coloured balls sampled from an urn should be roughly equal to the proportion of coloured ball contained in the urn. The formulae in the present section formalize and mathematically express our intuition. The hypergeometric distribution G_{aA} plays an important role in this formalization. A look at its plot, fig. 8, reveals that it is a sort of ‘fuzzy identity transformation’, or fuzzy Kronecker delta, between the A -space $\{0, \dots, N\}$ and a -space $\{0, \dots, n\}$. From eq. (19) we thus have that

$$p(a = a | I) \approx p(A = a | I), \quad E[c_a | I] \approx E[c_A | I], \quad (23)$$

where c is any smooth function defined on $[0, 1]$. These approximate equalities express the intuitive fact that *our uncertainty about the sample is representative of our uncertainty about the network and about other samples*, and vice versa. When $n = N$, G_{aA} becomes the identity matrix and the approximate equalities above become exact – of course, since we have sampled the full network.

But the approximate equalities above may miss important features of the two probability distributions. In the next section we will in fact

emphasize their differences. If the distribution for the network average A is bimodal, for example, the bimodality can be lost in the distribution for the sample average a , owing to the coarsening effect of G_{aA} .

9 Maximum-entropy: sample level vs full-network level

In the previous section we have seen that observations about a sample can be used as constraints on the distribution for the activity of the full network. Let us use such constraints with the maximum-entropy method. Suppose that we want to constrain m functions of the sample activity, vectorially written $c := (c_1, \dots, c_m)$, to m values $\hat{c} := (\hat{c}_1, \dots, \hat{c}_m)$. These functions are typically k -activities $\overline{a \dots a}$, and the values are typically the time averages of the observed sample, as discussed in § 1: $\hat{c} = \sum_t c[a(t)]/T$.

Let us apply the relative-maximum-entropy method (sivia1996_r2006; meadetal1984) directly to sampled neurons; denote this approach by I_s . Then we apply the method to the full network of neurons, most of which are unsampled; denote this approach by I_p .

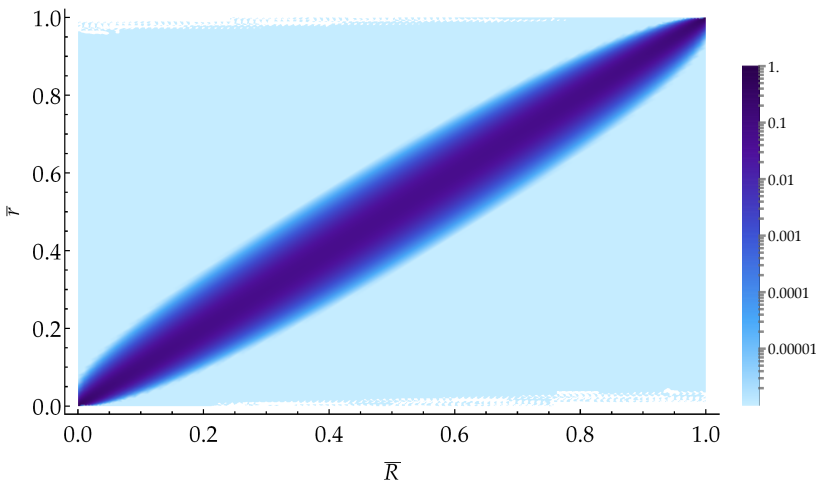


Figure 8 Log-density plot of the hypergeometric distribution $G_{aA} := \binom{n}{na} \binom{N-n}{NA-na} \binom{N}{NA}^{-1}$ for $N = 5000$, $n = 200$. (Band artifacts may appear in the colourbar depending on your PDF viewer.)

Applied directly to the sampled neurons, the method yields the distribution

$$p(a \mid I_s) = \frac{1}{z(I)} \binom{n}{na} \exp[I^\top c_a] \quad (24)$$

where $z(I)$ is a normalization constant. The binomial in front of the exponential appears because we must account for the multiplicity by which the network-average activity a can be realized: $a = 0$ can be realized in only one way (all neurons inactive), $a = 1/n$ can be realized in n ways (one active neuron out of n), and so on. This term is analogous to the ‘density of states’ in front of the Boltzmann factor in statistical mechanics (callen1960_r1985). The m Lagrange multipliers $I := (I_1, \dots, I_m)$ must satisfy the m constraint equations

$$\hat{c} = E(c \mid I_s) \equiv \frac{1}{z(I)} \sum_a c_a \binom{n}{na} \exp[I^\top c_a]. \quad (25)$$

Applied to the full network, using the constraint expression (21) derived in the previous section, the method yields the distribution for the full-network activity

$$p(A \mid I_p) = \frac{1}{\zeta(\lambda)} \binom{N}{NA} \exp(\lambda^\top \sum_a c_a G_{aA}). \quad (26)$$

The m Lagrange multipliers $\lambda := (\lambda_1, \dots, \lambda_m)$ must satisfy the m constraint equations

$$\hat{c} = E(c \mid I_p) \equiv \frac{1}{\zeta(\lambda)} \sum_a \sum_A c_a G_{aA} \binom{N}{NA} \exp(\lambda^\top \sum_a c_a G_{aA}). \quad (27)$$

We obtain the distribution for the sample activity by marginalization, using eq. (19):

$$p(a \mid I_p) = \frac{1}{\zeta(\lambda)} \sum_A G_{aA} \binom{N}{NA} \exp(\lambda^\top \sum_a c_a G_{aA}). \quad (28)$$

The distributions for the sample activity, eqs (28) and (24), obtained with the two approaches I_s and I_p , are different. From the discussion in the previous section we expect them to be vaguely similar; yet they cannot be exactly equal, because their equality would require the $2m$ quantities λ and I to satisfy the constraint equations (27) and (25), and in addition also the n equations $p(a \mid I_p) = p(a \mid I_s)$, $a = 1/n, \dots, 1$ (one

equation is taken care of by the normalization of the distributions). We would have a set of $2m + n$ equations in $2m$ unknowns.

Hence, *the applications of maximum-entropy at the sample level and at the full-network level are inequivalent*. They lead to numerically different distributions for the sample activity a .

The distribution obtained at the sample level will show different features from the one obtained at the network level, like displaced or additional modes or particular tail behaviour. We show an example of this discrepancy in fig. 9, for $N = 10\,000$, $n = 200$, and the two constraints

$$E(a) = 0.0478, \quad E(\overline{a\overline{a}}) = 0.00257, \quad (29)$$

which come from the actual recording of circa 200 neurons from macaque motor cortex ([rostamietal2016_r2017](#)). The distribution obtained at the network level (blue triangles) has a higher and displaced mode and a quite different behaviour for activities around 0.5 than the distribution obtained at the sample level (red squares).

In our discussion we have so far assumed the size N of the larger network to be known. This is rarely the case, however. We usually are uncertain about N and can only guess its order of magnitude. In such a state of knowledge I_u our ignorance about the possible value of N is expressed by a probability distribution $p(N = N \mid I_u) = h(N)$, and the marginal distribution for the sample activity (28) is modified, by the law of total probability, to

$$p(a \mid I_u) = \sum_N p(a \mid N, I_u) p(N \mid I_u) = \sum_N \left\{ \frac{1}{\zeta(\lambda_N)} \sum_A G_{aA}^{(N)} \binom{N}{NA} \exp[\lambda_N^\top \sum_a c_a G_{aA}^{(N)}] \right\} h(N), \quad (30)$$

where the Lagrange multipliers λ_N and the summation range for A depend on N .


As a proof of concept, fig. 9 also shows such a distribution (yellow circles) for the same constraints as above, and a probability distribution for N inspired by Jeffreys ([jeffreys1939_r1983](#)):


$$h(N) \propto 1/N, \quad N \in \{1\,000, 2\,000, \dots, 10\,000\}. \quad (31)$$


10 Derivation from the probability calculus

There are three inequivalent main routes that lead to a probability distribution of the maximum-entropy form (24) or (26). The distribution carries a different interpretation under each route (jaynes1979bjaynes1982jaynes1982bjaynes1986d_r1996jaynes1994_r2003).

(a) One route is the choice of the distribution having the highest Shannon entropy, given only a quantitative assessment some of its properties, such as expectations. The numerical choice of the value of such properties is a (subjective) assumption.

In the two other routes the maximum-entropy distribution is obtained as an *approximation* of a distribution obtained via the probability calculus, using data coming from a set of T measurements – as in our present case.  [refs here](#) Also in this case some (subjective) assumptions are necessary: they concern our beliefs about the long-run relative frequencies of the measurement outcomes:¹

(b) In one case we consider all possible sets of measurement outcomes to be *roughly* equally likely; this leads to a probability for the frequencies ν proportional to a multinomial coefficient $\binom{L}{\nu}$, with L large but smaller than T . (We cannot assume the sets of measurement outcomes to be exactly equally likely, because this is equivalent to their independence: cf. jaynes1994_r2003portamana2009portamana2017.) The exact expression is  [equation here](#)

(c) In the other case we assume that the measurements have a *sufficient statistics*: the same as appears in the exponential of the maximum-entropy distribution. The exact expression is  [equation here](#)

It's important to keep in mind that the approximate equivalence of these three routes only holds under very specific assumptions – which *have physical and biological meanings and consequences*. In particular,

¹Such assumptions are always necessary at the beginning of an inference: 'Now the axioms of probability enable us to infer any probability-conclusion *only* from probability-premisses. In other words, the calculus of probability does not enable us to infer any probability-value unless we have some probabilities or probability relations *given*. Such data cannot be supplied by the mathematician. E.g. the rules of arithmetic and the axioms of the probability-calculus are utterly impotent to determine, on the supposed knowledge that the throw of a coin must yield either head or tail and cannot yield both, the probability that it will yield head or that it will yield tail. We must assume that the two co-exclusive and co-exhaustive possibilities are *equally probable*, before we can estimate the probability of either as being a half of certitude' (johnson1924).

route (c) implies that we can discard other empirical statistics of the data, if they are known; whereas route (b) requires us to specify all known empirical statistics, because using only a subset of them may lead to different results. Route (a) is also supposed to be used with all known data. Moreover, the approximate equivalence of route (a) with routes (b) and (c) *only holds if T is much larger than the possible values of the activity A* . Finally, we also obtain very different expressions depending on whether we're asking about *the activity in one of the recorded time bins* or about *the activity in a new time bin*. Works that use maximum-entropy distributions are often very vague about the latter point.

Here are the distributions for our degrees of belief about three different quantities, assuming an entropic pre-data distribution for the long-run frequencies of the full network:

The frequency distribution F of the full-network activity during the recording

$$\begin{aligned}
 P(F | f, I) &\propto \int d\mathbf{v} \sum_{\boldsymbol{\phi}} \delta(\sum_A \phi_{aA} = f_a) \delta(\sum_a \phi_{aA} = F_A) \times \\
 &\quad \binom{T}{T\boldsymbol{\phi}} \left[\prod_{aA} (G_{aA} v_A)^{T\phi_{aA}} \right] \exp[-L H(\mathbf{v}; \mathbf{R})] \\
 &\propto \delta(\sum_A G_{aA} F_A = f_a) \exp[-L H(F; \mathbf{R})]
 \end{aligned} \tag{32}$$

The long-run frequency distribution \mathbf{v} of the full-network activity

$$\begin{aligned}
 P(\mathbf{v} | f, I) &\propto \binom{T}{Tf} \left[\prod_a (\sum_A G_{aA} v_A)^{Tf_a} \right] \exp[-L H(\mathbf{v}; \mathbf{R})] \\
 &\propto \exp[-T H(f; \mathbf{G}\mathbf{v}) - L H(\mathbf{v}; \mathbf{R})]
 \end{aligned} \tag{33}$$

Note that if some f_a are zero, then the first exponential may be badly approximated by a delta, because the constraints lie on a facet of the simplex of frequencies $\{\mathbf{v}\}$.

The activity A' of the full-network in a new time bin

$$\begin{aligned}
P(A' | f, I) &\propto \int d\mathbf{v} \, v_{A'} \binom{T}{Tf} \left[\prod_a (\sum_A G_{aA} v_A)^{Tf_a} \right] \exp[-L H(\mathbf{v}; \mathbf{R})] \\
&\approx \int d\mathbf{v} \, v_{A'} \exp[-T H(f; \mathbf{G}\mathbf{v}) - L H(\mathbf{v}; \mathbf{R})]
\end{aligned} \tag{34}$$

Note that if some f_a are zero, then the first exponential may be badly approximated by a delta, because the constraints lie on a facet of the simplex of frequencies $\{\mathbf{v}\}$.

The maximum-relative-entropy distribution is, in the first two cases, an approximation of the most probable frequency distribution F or \mathbf{v} ; in the third case, an approximation of the probability distribution for A' .

11 Assumptions and limitations

Main assumptions behind this belief distribution:

We are approximating our state of knowledge with a finitely exchangeable one. In turn, this is numerically approximated by an infinitely exchangeable one for T large. But we don't really have an exchangeable belief: our degree of belief that the activity at the next time bin will differ from the activity at the present one is roughly proportional to the difference in the two subsequent activities.

The formulae say that we have equal beliefs about the underlying network states having the same activity. This isn't really our belief, for we believe there are interactions between the neurons and subnetworks thereof.

12 Discussion

The purpose of the present work was to point out and show, in a simple set-up, that the maximum-entropy method can be applied to recorded neuronal data in a way that accounts for the larger network from which the data are sampled, eqs (26)–(28). This application leads to results that differ from the standard application which only considers the sample in isolation, eqs (24)–(25). We gave a numerical example of this difference.

We have also shown how to extend the new application when the size of the larger network is unknown, eq. (30).

The latter formula, in particular, shows that the standard way of applying maximum-entropy implicitly assumes that *no* larger network exists beyond the recorded sample of neurons. One could in fact object to the application at the network level, and say that the traditional way of applying maximum-entropy, eq. (24), yields different results because it does not make assumptions about the size N of a possibly existing larger network. Such a state of uncertainty, however, is correctly formalized according to the laws of probability by introducing a probability distribution for N , and is expressed by eq. (30). This expression cannot generally be equal to (24) unless the distribution for N gives unit probability to $N = n$; that is, unless the sample *is* the full network, and no larger network exists.

The standard maximum-entropy approach therefore assumes that the recorded neurons constitute a special subnetwork, isolated from the larger network of neurons in which it is embedded, and which was also present under the recording device. This assumption is unrealistic. The maximum-entropy approach at the network level does not make such assumption and is therefore preferable. It may reveal features in a data set that were unnoticed by the standard maximum-entropy approach.

The difference in the resulting distributions between the applications at the sample and at the network levels appears in the use of Boltzmann machines with hidden units (**lerouxetal2008**), although by a different conceptual route. It also appears in statistical mechanics: if a system is statistically described by a maximum-entropy Gibbs state, its subsystems cannot be described by a Gibbs state (**maesetal1999**). A somewhat similar situation also appears in the statistical description of the final state of a non-equilibrium process starting and ending in two equilibrium states: we can describe our knowledge about the final state either by (1) a Gibbs distribution, calculated from the final equilibrium macrovariables, or (2) by the distribution obtained from the Liouville evolution of the Gibbs distribution assigned to the initial state. The two distributions differ (even though the final *physical* state is obviously exactly the same (**jaynes1985d_r1993**)), and the second allows us to make sharper predictions about the final physical state thanks to our knowledge of its preceding dynamics. In this example, though, both distributions are usually extremely sharp and practically lead to the same predictions. In


neuroscientific applications, the difference in predictions of the sample vs full-network applications can instead be very relevant.

The idea of the new application leads in fact to more questions. For instance:

- Do the standard and new applications lead to different or contrasting conclusions about ‘cooperativity’, when applied to real data sets?
- How to extend the new application to the ‘inhomogeneous’ case ([schneidmanetal2006](#); [shlensetal2006](#); [roudietal2009b](#)), in which expectations for individual neurons or groups of neurons are constrained?
- What is the mathematical relation between the new application and maximum-entropy models with hidden neurons ([smolensky1986](#); [kulkarnietal2007](#); [huang2015](#); [dunnetal2017](#))?

Owing to space limitations we must leave a thorough investigation of these questions to future work.

Finally, we would like to point out the usefulness and importance of the probability formulae that relate our states of knowledge about a network and its samples, presented in § 8. This kind of formulae is essential in neuroscience, where we try to understand properties of extended brain areas from partial observations. The formulae presented here reflect a simple, symmetric state of ignorance. More work is needed ([levinaetal2017](#)) to extend these formulae to account for finer knowledge of the cerebral cortex and its network properties.

 **orig intro** Experimental technologies to record the activity of many neurons at the same time in different species and brain areas are rapidly advancing. These experimental advancements are paralleled by advances in theoretical and computational methods for analyzing the data accumulated using the recording technologies. Such theoretical methods usually take the form of probabilistic models that try to describe the multi-neuronal activity of the recorded neurons. With such probabilistic models one aims to address numerous issues: What correlations are important in describing the multi-neuronal pattern? How does the pattern of activity covary with external stimuli or experimental conditions? What dimensionality does the neural data live in and how is this related to the underlying network interactions? The probabilistic models can also be used to make predictions about the structure of the neural code, by studying the properties of the fitted model, or by generating synthetic data from it.

In general, despite the rapid advances in recording technology, the best experimental measurements of neuronal activity still only provide data from a small subset of neurons that comprise a neuronal network. A wealth of studies on building probabilistic models of neural data focuses on describing such subsets, ignoring the fact that the observed neurons is a small group in a much bigger set of hidden neurons. Some other studies do include hidden variables which, amongst other things, aim to model the global features of the unrecorded neurons, but we still lack a through understanding of how to include the role of hidden neurons in probabilistic models, how our inferences about the recorded activity would be affected by them, and what we can we say about the rest of the network by studying the heavily subsampled recordings. In this paper, we aim to address these questions in the case of a simple maximum entropy model, namely the homogeneous maximum entropy model.

The maximum entropy approach has been used in a variety of setting for building statistical models of complex systems and datasets, ranging from neuronal activity in the retina, in the cortex, protein sequences, gene regulatory networks and natural images. The general idea is, for a dataset, to write down the distribution that maximizes the entropy of state variables, given some low order statistics. Now given the fact that the recorded neurons are a fraction of the neurons in the network, several quantitative questions arise that we will address in this paper: given the data from the sampled neurons, can we build a maximum

entropy model over the whole network? Once we build such a network level maximum entropy model, can we see features in the neural activity which cannot be directly seen from a model build from the sampled neurons? Since we can marginalize down the network level maximum entropy model to the sampled network, how does this marginalized maximum entropy model match the sample level model?

All these questions can be answered in the case of the homogeneous maximum entropy model. First we show how to go from the sample level maximum entropy level to the network level, by assuming different sizes of the network and also by assuming an uninformative prior over the size of the network. This is done by inferring the statistics of correlation functions at the network level from those of the sample level by using simple counting arguments. We then find that, when applied to experimental recordings from the Medial Entorhinal Cortex of rats and the monkey visual cortex, this network level maximum entropy model may exhibit features that the sample level model does not predict. Specifically, we observed modes in the distribution of the activity in the network level model that do not show up in the sample level. We study how the assumed size of the full network affects the appearance of these modes and find that there is a minimum size of the full network for which such modes can be observed. We then compared the distribution found by marginalizing the full network maximum entropy model down to the sample level, and the distribution fit directly to the sample level. For the two datasets that we tested, we found that the two distributions match each other to a large degree but that there are also differences between them. We quantify how these differences also depend on the assumed size of the network and find that . . . (WE SHOULD TEST SHI) This predicts that for a large enough network (DO WE PREDICT THAT IF THE FULL NETWORK GETS BIGGER THE DIFFERENCE ALSO GET BIGGER)?. . .

The rest of the paper is organized as follows. We first describe how to go from the sample level maximum entropy model to the full network maximum entropy model. In section 2, we apply this to the two experimental datasets and study the effect of the assumed size of the network as well as the moments that we use for building the maximum entropy model. In section 3 we compare the distributions found from marginalizing the maximum entropy model down to the sample level and the original sample level model.