# Maximum-entropy distributions for a neuronal population from subpopulation constraints

P.G.L. Porta Mana
<piero.mana@ntnu.no>

Y. Roudi
<yasser.roudi@ntnu.no>

V. Rostami
<vrostami@uni-koeln.de>

E. Torre
<torre@ibk.baug.ethz.ch>

Draft of 20 April 2019 (first drafted 4 November 2015)

This work shows how to build a maximum-entropy probabilistic model for the total activity of a population of neurons, given only some activity statistics – for example, empirical moments – of a *subpopulation* thereof. This kind of model is useful because neuronal recordings are always limited to a very small sample of a population of neurons. The model is applied to two sets of neuronal data available in the literature. In some cases it predicts the larger population to have interesting features – for example, two modes in the probability for the total activity at low regimes – that are not visible in the sample or in a maximum-entropy model built for the sample alone. For the two datasets, the maximum-entropy probability model applied only to the subpopulation is compared with the marginal probability distribution obtained from the maximum-entropy model applied to the full population. 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.*

This document is designed for screen reading and two-up printing on A4 or Letter paper

## 1 Why polls?

When political elections are approaching, mass media are constantly bubbling with results and discussions of pre-election polls. Why do we conduct pre-election polls and other kinds of survey sampling? The simplest, intuitive answer is 'because they give us an idea of what the final election outcomes will be'. Can this answer be made more quantitative?

Yes. In the simplest cases, the probability calculus give us simple and precise quantitative relations between our expectations from a population – of voters, for instance – and from a sample thereof. Suppose there are $N$ voters, $S$ of which prefer one political candidate and $N - S$ of which prefer another. We don't know $S$ or $N - S$. Now take a sample of $n$ voters, assuming that every set of $n$ voters is equally likely to be sampled. In this sample, $s$ voters prefer the first candidate, and $n - s$ the second. The probability calculus tells us that our expectations of $s/n$ and $S/N$ are equal:

$$\mathrm{E}(s/n \mid I) = \mathrm{E}(S/N \mid I), \tag{1}$$

where $I$ represents our state of knowledge in this example.

A more precise answer to our original question is then: 'because we *expect* to see the same proportion of political preferences in a sample as in the full population'.

The equality above extends to other quantities beyond the means $S/N$ and $s/n$: it holds for all normalized factorial moments (Potts 1953; Broca 2005):

$$\begin{multline}\mathrm{E}\Bigl[\tfrac{s\,(s-1)}{n\,(n-1)} \Bigm| I\Bigr] = \mathrm{E}\Bigl[\tfrac{S\,(S-1)}{N\,(N-1)} \Bigm| I\Bigr], \\ \mathrm{E}\Bigl[\tfrac{s\,(s-1)\,(s-2)}{n\,(n-1)\,(n-2)} \Bigm| I\Bigr] = \mathrm{E}\Bigl[\tfrac{S\,(S-1)\,(S-2)}{N\,(N-1)\,(N-2)} \Bigm| I\Bigr], \quad \dots \\ \mathrm{E}\Bigl[\tbinom{s}{r}\tbinom{n}{r}^{-1} \Bigm| I\Bigr] = \mathrm{E}\Bigl[\tbinom{S}{r}\tbinom{N}{r}^{-1} \Bigm| I\Bigr], \tag{2}\end{multline}$$

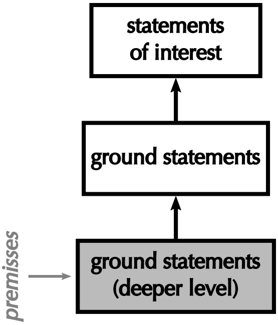
with $r < N$.

## 2 Maximum-entropy method

A necessary step of every plausible inference is the assignment of some probabilistic premisses; as many as necessary to arrive at our desired probabilistic conclusions given the evidence. The necessity of this step was tersely stated by W. E. Johnson a century ago:

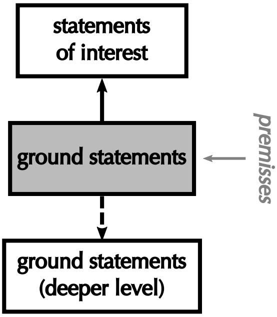
> Now the axioms of probability enable us to infer any probability-conclusion *only* from probability-premisses. In other words, the calculus of probability does not enable us to infer any probability-value unless we have some probabilities or probability relations *given*. Such data cannot be supplied 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. (Johnson 1924 Appendix on eduction, § 5, p. 182)

These premisses may represent either actual knowledge or just



some hypotheses whose consequences we want to explore. They can be assigned at different depths of the logical analysis of our inference problem.

We can for example assign probabilistic premisses at a very deep analytical level, as schematically illustrated on the left. They determine the probabilities of statements at more superficial levels, including the statements which we're ultimately interested in or which we can directly observe.



Or we can assign probabilistic premisses at a level of superficial or intermediate depth, as schematically illustrated on the right. In this case they determine the probabilities of statements at more superficial levels, and partly constrain the probabilities of statements at deeper levels.

The choice of the analytical depth at which to assign the premisses depends on many factors: on the difficulty of assessing our state of uncertainty at that depth, on the cost of the ensuing calculations, and especially on the general purpose of our inference. Assignments of probabilistic premisses at different levels can still turn out to be completely equivalent for the probabilities of the statements of interest.

Here's an example. Suppose that $n$ neurons have been recorded in a sequence of time bins. We're given the time averages of the numbers of active neurons and of neuron pairs in the sequence. We don't know the exact sequence of activities or the number $T$ of time bins, although we know that the latter is large. We want to forecast how many of the $n$ neurons will be active in a *new* time bin.

In a deep approach to this inference problem we assign the probability for the possible value of $T$, and the probabilities for several hypotheses about the sequence of activities in $T+1$ or more time bins. One hypothesis could be, for instance, that the $n$ neurons are a sample of a larger population of $N$ neurons, and a second hypothesis could be that for this population there are some sufficient statistics, out of biological reasons. We could of course deepen this approach further, with no end, down to assigning the probabilities for the constitution and motion of the molecules of the neurons and their environment, the probabilities for the position and response of the recording electrodes, and so on. From all these probabilities and the data we could finally calculate the probability for the activity in the new time bin.

At the other end, the shallowest approach to this inference problem is to directly assign the probabilities for the number of active neurons in the new time bin, somehow using the data we were given. (It must be remarked that such assignment partly constrains, in a backward fashion, all assignments we could do at a deeper level.)

The maximum-entropy method was proposed by Jaynes (1957a; 1963; 2003; Sivia 2006; Hobson et al. 1973) as a way of assigning probability premisses at a more direct level, as in the approach just described. This point is made especially clear in Jaynes (2003), § 11.1:

> If we knew the numbers $N$, $L$, $W$, [some quantities relevant to our inference] then this could be solved by direct application of Bayes' theorem; without that information, we could still introduce the unknowns $N$, $L$, $W$ as nuisance parameters and use Bayes' theorem, eliminating them at the end. [. . .] However, the Bayesian solution would not really address our problem; it only transfers it to the problem of assigning priors to $N$, $L$, $W$, leaving us back in essentially the same situation; how do we assign informative probabilities?

The idea behind this method is to translate *all* the data we're given into constraints – such as expectations – for our probability assignment, resolving the leftover freedom by choosing the constrained probability distribution having highest Shannon entropy, as a measure of its uncertainty or broadness. It can be shown that the use of this method is approximately equivalent, under specific conditions, to a probability assignment at a deeper level based on an assumption of exchangeability together with a multinomial probability distribution for long-run frequencies, or with an assumption about sufficient statistics (Jaynes 1996; Rodríguez 1989; 2002; Skilling 1989a; 1990; Bernardo et al. 2000 § 4.5; Diaconis et al. 1981; Porta Mana 2017; 2009).

Sometimes this method is applied with only *part* of the available data used as constraint; such an altered approach can be a heuristic way of testing for sufficient statistics.

Owing to its directness, the maximum-entropy method is straightforward to use – and to misuse. In our previous example, for instance, it would give silly results if the number of time bins were not much larger than the possible activity values (it has thus been variously misused in the neuroscientific literature 🧩 refs, especially to Tkacik).

In the present work we show how to use the maximum-entropy method to assign a probability distribution for the total activity at a new time bin of a population of neurons, *most of which can be unrecorded*, given specific statistics about the activity of a recorded *sample* of the full population. This kind of probability assignment is extremely important in neuroscience: we try to understand the workings of brain areas that may include tens of thousands of neurons, but we can record only few tens or hundreds of them.

Our approach also corrects a slight misuse of the maximum-entropy method as sometimes appears in the literature. When this method is applied to a set of neurons, it implicitly assumes that the set is completely isolated from other neurons. We'll reveal this implicit assumption in several ways. The probability distribution thus constructed is therefore affected by this hidden assumption. We'll study in which circumstances the application of the method at the sample level and at the full-population level give similar or different probabilities.

The maximum-entropy method can be used this way thanks to a very simple ***

## Yasser's 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 margianzlied 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 population (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.

Luca: 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.

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. (1985) 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. (1966). 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 (Jaynes 1957a; 1963; Hobson et al. 1973; Sivia 2006; Mead et al. 1984). 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 (MacKay 1991; Martignon et al. 1995; Bohte et al. 2000; Amari et al. 2003; Schneidman et al. 2006; Shlens et al. 2006; Macke et al. 2009; Roudi et al. 2009c; Tkačik et al. 2009; Gerwinn et al. 2009; Macke et al. 2011b,a; Ganmor et al. 2011; Granot-Atedgi et al. 2013; Tkačik et al. 2014; Mora et al. 2015; Shimazaki et al. 2015), and it has been subjected to mathematical and conceptual scrutiny (Tkačik et al. 2006; Roudi et al. 2009a,b; Barreiro et al. 2010a,b; Macke et al. 2013; Rostami et al. 2017).

'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) (cf. Caianiello 1961; 1986).

   Let the variable $s_i(t) \in \{0, 1\}$ denote the activity of the $i$th sampled neuron at time bin $t$. Collectively denote the $n$ activities with $\boldsymbol{s}(t) \coloneqq \bigl(s_1(t), \dots, s_n(t)\bigr)$. The population-averaged activity at that bin is $s(t) \coloneqq \sum_i s_i(t)/n$. If we count the number of distinct pairs of active neurons at that bin we combinatorially find $\binom{ns(t)}{2} \equiv ns(t)\,[ns(t) - 1]/2$. There can be at most $\binom{n}{2}$ simultaneously active pairs, so the population-averaged pair activity is $\overline{ss}(t) \coloneqq \binom{n}{2}^{-1}\binom{ns(t)}{2}$. With some combinatorics we see that the population-averaged activity of $m$-tuples of neurons is

   $$\underbrace{\overline{s \cdots s}}_{m\text{ terms}}(t) = \binom{n}{m}^{-1}\binom{ns(t)}{m}. \tag{3}$$

   For brevity let us agree to simply call 'activity' the average $s$, 'pair-activity' the average $\overline{ss}$, and so on.

2. Construct a sequence of relative-maximum-entropy distributions for the activity $s$, using this sequence of constraints:

   - the time average of the activity: $\widehat{s} \coloneqq \sum_t s(t)/T$;
   - the time averages of the activity and of the pair-activity $\widehat{\overline{ss}} \coloneqq \sum_t \overline{ss}(t)/T$;
   - ...
   - the time averages of the activity, of the pair-activity, and so on, up to the $k$-activity.

   Call the resulting distributions $p_1(s), p_2(s), \dots, p_k(s)$. 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, $f(s)$, 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 (Kullback 1987; Jaynes 1963; Hobson 1969; Hobson et al. 1973). 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 $m \geqslant 2$, 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 population 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 (Roudi et al. 2009b). This problems extends also to more complex versions of the method, possibly except versions that use 'hidden' neurons (Smolensky 1986; Kulkarni et al. 2007; Huang 2015; Dunn et al. 2017). The problem is that the recorded neurons are a *sample* from a larger, unrecorded population, 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 (cf. Berényi et al. 2014). This recording is a sample of the activity of the neuronal population 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 population – at the very least the population 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 population. Probability theory tells us how to make inferences from a sample to the larger population 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 population, we can also apply the maximum-entropy method to the larger (unrecorded) population. The constraints are the same: the time averages of the sampled data, since they constitute representative data about the larger population as well. The method thus yields a probability distribution for the larger population, 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 population, and show that its results differ from the application at the sample level. We also consider the case where the size of the larger population is unknown.

To apply the maximum-entropy method to the larger, unsampled population, it is necessary to use probability relations relevant to sampling (Ghosh et al. 1997; Freedman et al. 2007 parts I, VI; Gelman et al. 2014 ch. 8; Jaynes 2003 ch. 3). 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 (iso 1993; ieee 1993; nist 1995; iso 2006a,b) but for the use of the comma ',' to denote logical conjunction. Probability notation follows Jaynes (Jaynes 2003). By 'probability' we mean a degree of belief which 'would be agreed by all rational men if there were any rational men' (Good 1966).

## 3 Probability relations between population and sample

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

The probability relations between sample and larger population 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 population is expressed by the joint probability distribution

$$\mathrm{p}(S_1, S_2, \dots, S_N \mid K) \quad \text{or} \quad \mathrm{p}(\boldsymbol{S} \mid K), \quad \boldsymbol{S} \in \{0,1\}^N. \tag{4}$$

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

$$\mathrm{p}(s_1, s_2, \dots, s_n \mid K) \quad \text{or} \quad \mathrm{p}(\boldsymbol{s} \mid K), \quad \boldsymbol{s} \in \{0,1\}^n. \tag{5}$$

The theory of statistical sampling is covered in many excellent texts, for example Ghosh & Meeden (1997) or Freedman, Pisani, & Purves (2007 parts I, VI); summaries can be found in Gelman et al. (2014 ch. 8) and Jaynes (2003 ch. 3).

We need to make an initial probability assignment for the state of the full population 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 population. Our state of knowledge and initial probability assignment should reflect that samples are somehow representative of the whole population.

In this state of knowledge, denoted $I$, we know that the neurons in the population 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* (Ericson 1969a; Ghosh et al. 1997 ch. 1). We stress that this probability assignment is just an expression of the symmetry of our *ignorance* about the state of the population, 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 population-average:

$$\mathrm{p}(\boldsymbol{S} \mid I) \equiv \sum_{S} \mathrm{p}(\boldsymbol{S} \mid S, I)\, \mathrm{p}(S \mid I) = \binom{N}{NS}^{-1} \mathrm{p}(S \mid I). \tag{6}$$

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 $\binom{N}{NS}$ activity vectors with $NS$ 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 (1959), Kendall (1967), Ericson (1976), Diaconis & Freedman (1977; 1980), Heath & Sudderth (1976).

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

$$\mathrm{p}(s \mid S, I) = \binom{n}{ns}\binom{N-n}{NS-ns}\binom{N}{NS}^{-1} =: G_{sS}, \tag{7}$$

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 (Jaynes 2003 ch. 3; Ross 2010 § 4.8.3; Feller 1968 § II.6).

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

$$\mathrm{p}(s \mid I) = \sum_{S} \mathrm{p}(s \mid S, I)\, \mathrm{p}(S \mid I) = \sum_{S} G_{sS}\, \mathrm{p}(S \mid I). \tag{8}$$

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 population-averaged activity:

$$\mathrm{p}(\boldsymbol{s} \mid I) = \binom{n}{ns}^{-1} \mathrm{p}(s \mid I). \tag{9}$$

This is intuitively clear: our initial symmetric ignorance should also apply to the sample. The distribution for the sample (8) indeed satisfies the same representation theorem (6) as the distribution for the full population.

The conditional probability $\mathrm{p}(s \mid S, I) \equiv G_{sS}$, besides relating the distributions for the population and sample activities via marginalization, also allows us to express the expectation value of any function of the sample activity, $\boldsymbol{c}_s$, in terms of the distribution for the full population, as follows:

$$\mathrm{E}(\boldsymbol{c} \mid I) \equiv \sum_{s} \boldsymbol{c}_s\, \mathrm{p}(s \mid I) = \sum_{s} \boldsymbol{c}_s \sum_{S} G_{sS}\, \mathrm{p}(S \mid I) = \sum_{S} \bigl(\textstyle\sum_s \boldsymbol{c}_s G_{sS}\bigr)\, \mathrm{p}(S \mid I), \tag{10}$$

where the second step uses eq. (8). The last expression shows that the expectation of the function $\boldsymbol{c}_s$ is equal to the expectation of the function $\boldsymbol{c}^*(S) := \sum_s \boldsymbol{c}_s\, G_{sS}$.

The final expression in eq. (10) is important for our maximum-entropy application: the requirement that the function $\boldsymbol{c}$, defined for the sample, have a value $\widehat{\boldsymbol{c}}$ obtained from observed data, *translates into a linear constraint for the distribution of the full population*:

$$\widehat{\boldsymbol{c}} = \mathrm{E}(\boldsymbol{c} \mid I) \equiv \sum_{S} \bigl(\textstyle\sum_s \boldsymbol{c}_s G_{sS}\bigr)\, \mathrm{p}(S \mid I). \tag{11}$$

In particular, when the function $\boldsymbol{c}$ is the $m$-activity of the sample, $\boldsymbol{c}_s = \overline{\boldsymbol{s} \dots \boldsymbol{s}} \equiv \binom{ns}{m}/\binom{n}{m}$, we find

$$\mathrm{E}(\underbrace{\overline{\boldsymbol{s} \cdots \boldsymbol{s}}}_{m \text{ factors}} \mid I) \equiv \sum_s \binom{n}{m}^{-1} \binom{ns}{m} \mathrm{p}(s \mid I) = \binom{N}{m}^{-1} \sum_S \binom{NS}{m} \mathrm{p}(S \mid I) \equiv \mathrm{E}(\underbrace{\overline{\boldsymbol{S} \cdots \boldsymbol{S}}}_{m \text{ factors}} \mid I), \qquad (12)$$

that is, *the expected values of the m-activities of the sample and of the full population 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 Potts (1953). Similar relations can be found for the raw moments $\mathrm{E}(s^m)$ and $\mathrm{E}(S^m)$, which can be written in terms of the product expectations using eq. (3).

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 population to have the same value.

These expectation equalities between sample and full population 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 $G_{sS}$ plays an important role in this formalization. A look at its plot, fig. 1, reveals that it is a sort of 'fuzzy identity transformation', or fuzzy Kronecker delta, between the $S$-space $\{0, \dots, N\}$ and $s$-space $\{0, \dots, n\}$. From eq. (9) we thus have that

$$\mathrm{p}(s = a \mid I) \approx \mathrm{p}(S = a \mid I), \qquad \mathrm{E}[\boldsymbol{c}_s \mid I] \approx \mathrm{E}[\boldsymbol{c}_S \mid I], \qquad (13)$$

where $\boldsymbol{c}$ 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 population and about other samples*, and vice versa. When $n = N$, $G_{sS}$ becomes the identity matrix and the approximate equalities above become exact – of course, since we have sampled the full population.

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 population average $S$ is bimodal, for example, the bimodality can be lost in the distribution for the sample average $s$, owing to the coarsening effect of $G_{sS}$.

## 4 Maximum-entropy: sample level vs full-population 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 population. Let us use such constraints with the maximum-entropy method. Suppose that we want to constrain $m$ functions of the sample activity, vectorially written $\boldsymbol{c} \coloneqq (c_1, \dots, c_m)$, to $m$ values $\widehat{\boldsymbol{c}} \coloneqq (\widehat{c}_1, \dots, \widehat{c}_m)$. These functions are typically $k$-activities $\overline{\boldsymbol{s} \dots \boldsymbol{s}}$, and the values are typically the time averages of the observed sample, as discussed in § **??**: $\widehat{\boldsymbol{c}} = \sum_t \boldsymbol{c}[s(t)]/T$.

Let us apply the relative-maximum-entropy method (Sivia 2006; Mead et al. 1984) directly to sampled neurons; denote this approach by $I_{\mathrm{s}}$. Then we apply the method to the full population of neurons, most of which are unsampled; denote this approach by $I_{\mathrm{p}}$.

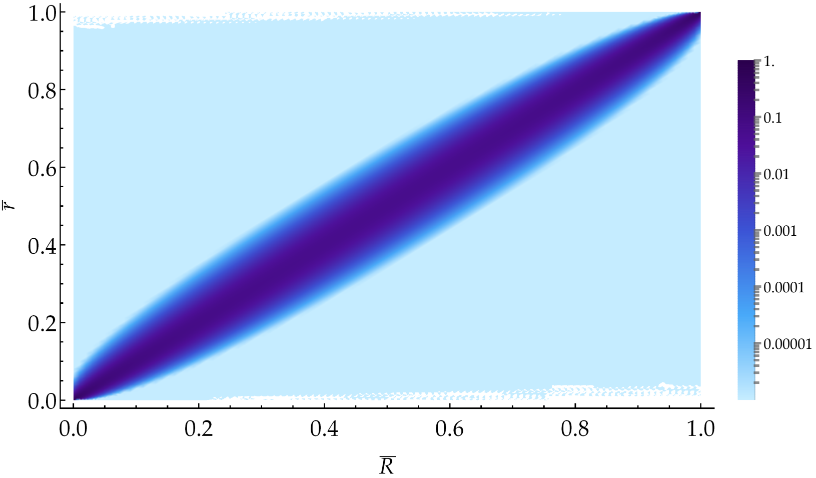


Figure 1 Log-density plot of the hypergeometric distribution $G_{sS} \coloneqq \binom{n}{ns}\binom{N-n}{NS-ns}\binom{N}{NS}^{-1}$ for $N = 5000$, $n = 200$. (Band artifacts may appear in the colourbar depending on your PDF viewer.)

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

$$\mathrm{p}(s \mid I_\mathrm{s}) = \frac{1}{z(\boldsymbol{l})} \binom{n}{ns} \exp[\boldsymbol{l}^\intercal \boldsymbol{c}_s] \tag{14}$$

where $z(\boldsymbol{l})$ is a normalization constant. The binomial in front of the exponential appears because we must account for the multiplicity by which the population-average activity $s$ can be realized: $s = 0$ can be realized in only one way (all neurons inactive), $s = 1/n$ 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 (Callen 1985 ch. 16). The $m$ Lagrange multipliers $\boldsymbol{l} \coloneqq (l_1, \dots, l_m)$ must satisfy the $m$ constraint equations

$$\widehat{\boldsymbol{c}} = \mathrm{E}(\boldsymbol{c} \mid I_\mathrm{s}) \equiv \frac{1}{z(\boldsymbol{l})} \sum_s \boldsymbol{c}_s \binom{n}{ns} \exp[\boldsymbol{l}^\intercal \boldsymbol{c}_s]. \tag{15}$$

Applied to the full population, using the constraint expression (11) derived in the previous section, the method yields the distribution for the full-population activity

$$\mathrm{p}(S \mid I_\mathrm{p}) = \frac{1}{\zeta(\boldsymbol{\lambda})} \binom{N}{NS} \exp\bigl(\boldsymbol{\lambda}^\intercal \textstyle\sum_s \boldsymbol{c}_s G_{sS}\bigr). \tag{16}$$

The $m$ Lagrange multipliers $\boldsymbol{\lambda} \coloneqq (\lambda_1, \dots, \lambda_m)$ must satisfy the $m$ constraint equations

$$\widehat{\boldsymbol{c}} = \mathrm{E}(\boldsymbol{c} \mid I_\mathrm{p}) \equiv \frac{1}{\zeta(\boldsymbol{\lambda})} \sum_s \sum_S \boldsymbol{c}_s G_{sS} \binom{N}{NS} \exp\bigl(\boldsymbol{\lambda}^\intercal \textstyle\sum_s \boldsymbol{c}_s G_{sS}\bigr). \tag{17}$$

We obtain the distribution for the sample activity by marginalization, using eq. (9):

$$\mathrm{p}(s \mid I_\mathrm{p}) = \frac{1}{\zeta(\boldsymbol{\lambda})} \sum_S G_{sS} \binom{N}{NS} \exp\bigl(\boldsymbol{\lambda}^\intercal \textstyle\sum_s \boldsymbol{c}_s G_{sS}\bigr). \tag{18}$$

The distributions for the sample activity, eqs (18) and (14), obtained with the two approaches $I_\mathrm{s}$ and $I_\mathrm{p}$, 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 $\boldsymbol{\lambda}$ and $\boldsymbol{l}$ to satisfy the constraint equations (17) and (15), and in addition also the $n$ equations $\mathrm{p}(s \mid I_\mathrm{p}) = \mathrm{p}(s \mid I_\mathrm{s})$, $s = 1/n, \dots, 1$ (one

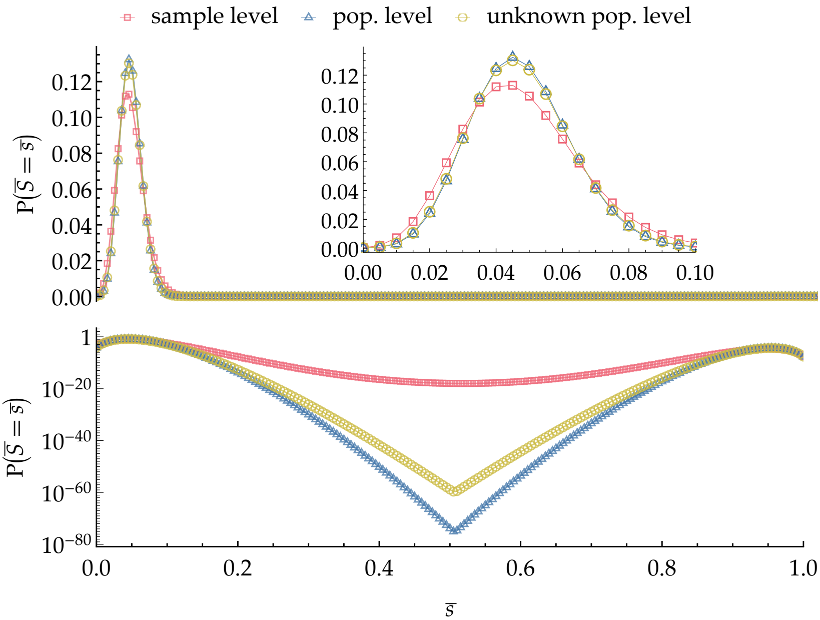


Figure 2 Linear and log-plots of $\mathrm{p}(s)$ for a sample of $n = 200$ and constraints as in eq. (19), constructed by: **red squares:** maximum-entropy at the sample level, eq. (14); **blue triangles:** maximum-entropy at the population level, eq. (18) with $N = 10\,000$, followed by sample marginalization; **yellow circles:** maximum-entropy at the population level with unknown population size, eq. (20), according to the distribution (21) for the population.

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-population level are inequivalent*. They lead to numerically different distributions for the sample activity $\boldsymbol{s}$.

The distribution obtained at the sample level will show different features from the one obtained at the population level, like displaced or additional modes or particular tail behaviour. We show an example of this discrepancy in fig. 2, for $N = 10\,000$, $n = 200$, and the two constraints

$$\mathrm{E}(s) = 0.0478, \qquad \mathrm{E}(\overline{ss}) = 0.00257, \tag{19}$$

which come from the actual recording of circa 200 neurons from macaque motor cortex (Rostami et al. 2017). The distribution obtained at the population 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 population 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 $I_\mathrm{u}$ our ignorance about the possible value of $N$ is expressed by a probability distribution $\mathrm{p}(N = N \mid I_\mathrm{u}) = h(N)$, and the marginal distribution for the sample activity (18) is modified, by the law of total probability, to

$$\begin{aligned}\mathrm{p}(s \mid I_\mathrm{u}) = \sum_N \mathrm{p}(s \mid N, I_\mathrm{u})\, \mathrm{p}(N \mid I_\mathrm{u}) = \\ \sum_N \biggl\{ \frac{1}{\zeta(\boldsymbol{\lambda}_N)} \sum_S G^{(N)}_{sS} \tbinom{N}{NS} \exp\Bigl[\boldsymbol{\lambda}_N{}^{\intercal} \textstyle\sum_s \boldsymbol{c}_s G^{(N)}_{sS}\Bigr] \biggr\}\, h(N),\end{aligned} \qquad (20)$$

where the Lagrange multipliers $\boldsymbol{\lambda}_N$ and the summation range for $S$ depend on $N$.

As a proof of concept, fig. 2 also shows such a distribution (yellow circles) for the same constraints as above, and a probability distribution for $N$ inspired by Jeffreys (Jeffreys 1983 § 4.8):

$$h(N) \propto 1/N, \qquad N \in \{1\,000,\ 2\,000,\ \dots,\ 10\,000\}. \qquad (21)$$

## 5 Derivation from the probability calculus

There are three inequivalent main routes that lead to a probability distribution of the maximum-entropy form (14) or (16). The distribution carries a different interpretation under each route (Jaynes 1979 pp. 52–55, 72–77; 1982a pp. 25–28; 1982b § I; 1996; 2003 § 11.1).

(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 $\boldsymbol{\nu}$ proportional to a multinomial coefficient $\binom{L}{L\boldsymbol{\nu}}$, 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. Jaynes 2003 § 6.7; Porta Mana 2009 § B; 2017 § 2.) 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, 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 $S$*. 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.

## 6 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

[1] Such 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' (Johnson 1924 *Appendix on eduction*, § 5, p. 182).

neuronal data in a way that accounts for the larger population from which the data are sampled, eqs (16)–(18). This application leads to results that differ from the standard application which only considers the sample in isolation, eqs (14)–(15). We gave a numerical example of this difference. We have also shown how to extend the new application when the size of the larger population is unknown, eq. (20).

The latter formula, in particular, shows that the standard way of applying maximum-entropy implicitly assumes that *no* larger population exists beyond the recorded sample of neurons. One could in fact object to the application at the population level, and say that the traditional way of applying maximum-entropy, eq. (14), yields different results because it does not make assumptions about the size $N$ of a possibly existing larger population. 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. (20). This expression cannot generally be equal to (14) unless the distribution for $N$ gives unit probability to $N = n$; that is, unless the sample *is* the full population, and no larger population 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 population 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 population levels appears in the use of Boltzmann machines with hidden units (Le Roux et al. 2008), 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 (Maes et al. 1999). 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 (Jaynes

1993 § 4)), 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-population 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 (Schneidman et al. 2006; Shlens et al. 2006; Roudi et al. 2009b), 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 (Smolensky 1986; Kulkarni et al. 2007; Huang 2015; Dunn et al. 2017)?

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 population and its samples, presented in § 3. This kind of formulae is essential in neuroscience, where we try to understand properties of extended brain regions from partial observations. The formulae presented here reflect a simple, symmetric state of ignorance. More work is needed (cf. Levina et al. 2017) to extend these formulae to account for finer knowledge of the cerebral cortex and its network properties.

# Bibliography

('de *X*' is listed under D, 'van *X*' under V, and so on, regardless of national conventions.)

Abeles, M. (1991): *Corticonics: Neural circuits of the cerebral cortex*. (Cambridge University Press, Cambridge).

Amari, S.-i., Nakahara, H., Wu, S., Sakai, Y. (2003): *Synchronous firing and higher-order interactions in neuron pool*. Neural Comp. **15**$^{1}$, 127–142.

Barreiro, A. K., Gjorgjieva, J., Rieke, F. M., Shea-Brown, E. T. (2010a): *When are microcircuits well-modeled by maximum entropy methods?* arXiv:1011.2797.

Barreiro, A. K., Shea-Brown, E. T., Rieke, F. M., Gjorgjieva, J. (2010b): *When are microcircuits well-modeled by maximum entropy methods?* BMC Neurosci. **11**$^{\text{Suppl. 1}}$, P65.

Berényi, A., Somogyvári, Z., Nagy, A. J., Roux, L., Long, J. D., Fujisawa, S., Stark, E., Leonardo, A., et al. (2014): *Large-scale, high-density (up to 512 channels) recording of local circuits in behaving animals*. J. Neurophysiol. **111**$^{5}$, 1132–1149. http://www.buzsakilab.com/content/PDFs/Berenyi2013.pdf.

Bernardo, J.-M., Smith, A. F. (2000): *Bayesian Theory*, reprint. (Wiley, New York). First publ. 1994.

Bohte, S. M., Spekreijse, H., Roelfsema, P. R. (2000): *The effects of pair-wise and higher-order correlations on the firing rate of a postsynaptic neuron*. Neural Comp. **12**$^{1}$, 153–179.

Broca, D. S. (2005): *Mean and variance through factorial moments*. Teach. Stat. **27**$^{2}$, 55–57.

Caianiello, E. R. (1961): *Outline of a theory of thought-processes and thinking machines*. J. Theor. Biol. **1**$^{2}$, 204–235.

— (1986): *Neuronic equations revisited and completely solved*. In: Palm, Aertsen (1986), 147–160.

Callen, H. B. (1985): *Thermodynamics and an Introduction to Thermostatistics*, 2nd ed. (Wiley, New York). First publ. 1960.

de Finetti, B. (1959): *La probabilità e la statistica nei rapporti con l'induzione, secondo i diversi punti di vista*. In: de Finetti (2011), 1–115. Transl. in de Finetti (1972), ch. 9, pp. 147–227.

— (1972): *Probability, Induction and Statistics: The art of guessing*. (Wiley, London).

— ed. (2011): *Induzione e statistica*, reprint. (Springer, Berlin). First publ. 1959.

Diaconis, P. (1977): *Finite forms of de Finetti's theorem on exchangeability*. Synthese **36**$^{2}$, 271–281. http://statweb.stanford.edu/~cgates/PERSI/year.html.

Diaconis, P., Freedman, D. (1980): *Finite exchangeable sequences*. Ann. Prob. **8**$^{4}$, 745–764.

— (1981): *Partial exchangeability and sufficiency*. In: Ghosh, Roy (1981), 205–236. http://statweb.stanford.edu/~cgates/PERSI/year.html. Also publ. 1982 as technical report https://www.stat.berkeley.edu/~aldous/206-Exch/Papers/diaconis_freedman_PES.pdf.

Dunn, B., Battistin, C. (2017): *The appropriateness of ignorance in the inverse kinetic Ising model*. J. Phys. A **50**$^{12}$, 124002.

Ericson, W. A. (1969a): *Subjective Bayesian models in sampling finite populations*. J. Roy. Stat. Soc. B **31**$^{2}$, 195–224. http://www.stat.cmu.edu/~brian/905-2008/papers/Ericson-JRSSB-1969.pdf. See also discussion in Sampford, Scott, Stone, Lindley, Smith, Kerridge, Godambe, Kish, et al. (1969).

— (1969b): *A note on the posterior mean of a population mean*. J. Roy. Stat. Soc. B **31**$^{2}$, 332–334.

Ericson, W. A. (1976): *A Bayesian approach to two-stage sampling*. Tech. rep. AFFDL-TR-75-145. (University of Michigan, Ann Arbor, USA). http://hdl.handle.net/2027.42/4819.

Feller, W. (1968): *An Introduction to Probability Theory and Its Applications. Vol. I*, 3rd ed. (Wiley, New York). First publ. 1950.

Ford, K. W., ed. (1963): *Statistical Physics*. (Benjamin, New York).

Fougère, P. F., ed. (1990): *Maximum Entropy and Bayesian Methods: Dartmouth, U.S.A., 1989*. (Kluwer, Dordrecht).

Freedman, D. A., Pisani, R., Purves, R. (2007): *Statistics*, 4th ed. (Norton, London). First publ. 1978.

Ganmor, E., Segev, R., Schneidman, E. (2011): *Sparse low-order interaction network underlies a highly correlated and learnable neural population code*. Proc. Natl. Acad. Sci. (USA) **108**$^{23}$, 9679–9684. http://www.weizmann.ac.il/neurobiology/labs/schneidman/The_Schneidman_Lab/Publications.html.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., Rubin, D. B. (2014): *Bayesian Data Analysis*, 3rd ed. (Chapman & Hall/CRC, Boca Raton, USA). First publ. 1995.

Gerstein, G. L., Perkel, D. H., Dayhoff, J. E. (1985): *Cooperative firing activity in simultaneously recorded populations of neurons: detection and measurement*. J. Neurosci. **5**$^{4}$, 881–889.

Gerwinn, S., Berens, P., Bethge, M. (2009): *A joint maximum-entropy model for binary neural population patterns and continuous signals*. Adv. Neural Information Processing Systems (NIPS) **22**, 620–628.

Ghosh, J. K., Roy, J., eds. (1981): *Statistics: Applications and New Directions*. (Indian Statistical Institute, Calcutta).

Ghosh, M., Meeden, G. (1997): *Bayesian Methods for Finite Population Sampling*, reprint. (Springer, Dordrecht).

Good, I. J. (1966): *How to estimate probabilities*. J. Inst. Maths. Applics **2**$^{4}$, 364–383.

Grandy Jr., W. T., Schick, L. H., eds. (1991): *Maximum Entropy and Bayesian Methods: Laramie, Wyoming, 1990*. (Kluwer, Dordrecht).

Granot-Atedgi, E., Tkačik, G., Segev, R., Schneidman, E. (2013): *Stimulus-dependent maximum entropy models of neural population codes*. PLoS Comput. Biol. **9**$^{3}$, e1002922.

Heath, D., Sudderth, W. (1976): *De Finetti's theorem on exchangeable variables*. American Statistician **30**$^{4}$, 188–189.

Hobson, A. (1969): *A new theorem of information theory*. J. Stat. Phys. **1**$^{3}$, 383–391.

Hobson, A., Cheng, B.-K. (1973): *A comparison of the Shannon and Kullback information measures*. J. Stat. Phys. **7**$^{4}$, 301–310.

Huang, H. (2015): *Effects of hidden nodes on network structure inference*. J. Phys. A **48**$^{35}$, 355002.

ieee (1993): *ANSI/IEEE Std 260.3-1993: American National Standard: Mathematical signs and symbols for use in physical sciences and technology*. Institute of Electrical and Electronics Engineers.

iso (1993): *Quantities and units*, 3rd ed. International Organization for Standardization.

— (2006a): *ISO 3534-1:2006: Statistics – Vocabulary and symbols – Part 1: General statistical terms and terms used in probability*. International Organization for Standardization.

— (2006b): *ISO 3534-2:2006: Statistics – Vocabulary and symbols – Part 2: Applied statistics*. International Organization for Standardization.

Jaynes, E. T. (1957a): *Information theory and statistical mechanics*. Phys. Rev. **106**$^{4}$, 620–630. http://bayes.wustl.edu/etj/node1.html, see also Jaynes (1957b).

— (1957b): *Information theory and statistical mechanics. II*. Phys. Rev. **108**$^{2}$, 171–190. http://bayes.wustl.edu/etj/node1.html, see also Jaynes (1957a).

— (1963): *Information theory and statistical mechanics*. In: Ford (1963), 181–218. Repr. in Jaynes (1989), ch. 4, 39–76. http://bayes.wustl.edu/etj/node1.html.

— (1979): *Where do we stand on maximum entropy?* In: Levine, Tribus (1979), 15–118. http://bayes.wustl.edu/etj/node1.html; repr. with an introduction in Jaynes (1989), pp. 210–314.

— (1982a): *Prior information in inference*. (). http://bayes.wustl.edu/etj/node2.html.

— (1982b): *On the rationale of maximum-entropy methods*. Proc. IEEE **70**$^{9}$, 939. http://bayes.wustl.edu/etj/node1.html.

— (1989): *E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics*, reprint. (Kluwer, Dordrecht). Edited by R. D. Rosenkrantz. First publ. 1983.

— (1993): *Inferential scattering*. http://bayes.wustl.edu/etj/node1.html. Extensively rewritten version of a paper first publ. 1985 in Smith, Grandy (1985), pp. 377–398.

— (1996): *Monkeys, kangaroos, and $N$*. http://bayes.wustl.edu/etj/node1.html. First publ. 1986. (Errata: in equations (29)–(31), (33), (40), (44), (49) the commas should be replaced by gamma functions, and on p. 19 the value 0.915 should be replaced by 0.0915).

— (2003): *Probability Theory: The Logic of Science*. (Cambridge University Press, Cambridge). Ed. by G. Larry Bretthorst. First publ. 1994. https://archive.org/details/XQUHIUXHIQUHIQXUIHX2, http://www-biba.inrialpes.fr/Jaynes/prob.html.

Jeffreys, H. (1983): *Theory of Probability*, third ed. with corrections. (Oxford University Press, London). First publ. 1939.

Johnson, W. E. (1924): *Logic. Part III: The Logical Foundations of Science*. (Cambridge University Press, Cambridge). https://archive.org/details/logic03john.

Kendall, D. G. (1967): *On finite and infinite sequences of exchangeable events*. Studia Sci. Math. Hung. **2**, 319–327.

Kulkarni, J. E., Paninski, L. (2007): *Common-input models for multiple neural spike-train data*. Netw. **18**$^{4}$, 375–407.

Kullback, S. (1987): *The Kullback-Leibler distance*. American Statistician **41**$^{4}$, 340–341.

Le Roux, N., Bengio, Y. (2008): *Representational power of restricted Boltzmann machines and deep belief networks*. Neural Comp. **20**$^{6}$, 1631–1649.

Levina, A., Priesemann, V. (2017): *Subsampling scaling*. Nat. Comm. **8**, 15140.

Levine, R. D., Tribus, M., eds. (1979): *The Maximum Entropy Formalism: A Conference Held at the Massachusetts Institute of Technology on May 2–4, 1978*. (MIT Press, Cambridge, USA).

MacKay, D. J. C. (1991): *Maximum entropy connections: neural networks*. In: Grandy, Schick (1991), 237–244.

Macke, J. H., Buesing, L., Cunningham, J. P., Yu, B. M., Shenoy, K. V., Sahani, M. (2011a): *Empirical models of spiking in neural populations*. Adv. Neural Information Processing Systems (NIPS) **24**, 1350–1358.

Macke, J. H., Murray, I., Latham, P. E. (2013): *Estimation bias in maximum entropy models*. Entropy **15**$^{8}$, 3109–3129.

Macke, J. H., Opper, M., Bethge, M. (2009): *The effect of pairwise neural correlations on global population statistics*. Tech. rep. 183. (Max-Planck-Institut für biologische Kybernetik, Tübingen). http://www.kyb.tuebingen.mpg.de/publications/attachments/MPIK-TR-183_%5B0%5D.pdf.

Macke, J. H., Opper, M., Bethge, M. (2011b): *Common input explains higher-order correlations and entropy in a simple model of neural population activity*. Phys. Rev. Lett. **106**$^{20}$, 208102.

Maes, C., Redig, F., Van Moffaert, A. (1999): *The restriction of the Ising model to a layer*. J. Stat. Phys. **96**$^{1}$, 69–107.

Martignon, L., Von Hassein, H., Grün, S., Aertsen, A., Palm, G. (1995): *Detecting higher-order interactions among the spiking events in a group of neurons*. Biol. Cybern. **73**$^{1}$, 69–81.

Mead, L. R., Papanicolaou, N. (1984): *Maximum entropy in the problem of moments*. J. Math. Phys. **25**$^{8}$, 2404–2417. http://bayes.wustl.edu/Manual/MeadPapanicolaou.pdf.

Moore, G. P., Perkel, D. H., Segundo, J. P. (1966): *Statistical analysis and functional interpretation of neuronal spike data*. Annu. Rev. Physiol. **28**, 493–522.

Mora, T., Deny, S., Marre, O. (2015): *Dynamical criticality in the collective activity of a population of retinal neurons*. Phys. Rev. Lett. **114**$^{7}$, 078105.

NIST (1995): *Guide for the Use of the International System of Units (SI): NIST special publication 811, 1995 edition*. National Institute of Standards and Technology. http://physics.nist.gov/cuu/Uncertainty/bibliography.html.

Palm, G., Aertsen, A., eds. (1986): *Brain Theory*. (Springer, Berlin).

Porta Mana, P. G. L. (2009): *On the relation between plausibility logic and the maximum-entropy principle: a numerical study*. arXiv:0911.2197. Presented as invited talk at the 31st International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering 'MaxEnt 2011', Waterloo, Canada.

— (2017): *Maximum-entropy from the probability calculus: exchangeability, sufficiency*. Open Science Framework doi:10.17605/osf.io/xdy72, arXiv:1706.02561.

Potts, R. B. (1953): *Note on the factorial moments of standard distributions*. Aust. J. Phys. **6**$^{4}$, 498–499.

Rodríguez, C. C. (1989): *The metrics induced by the Kullback number*. In: Skilling (1989b), 415–422.

— (2002): *Entropic priors for discrete probabilistic networks and for mixtures of Gaussians models*. Am. Inst. Phys. Conf. Proc. **617**, 410–432.

Ross, S. (2010): *A First Course in Probability*, 8th ed. (Pearson, Upper Saddle River, USA). First publ. 1976.

Rostami, V., Porta Mana, P. G. L., Grün, S., Helias, M. (2017): *Bistability, non-ergodicity, and inhibition in pairwise maximum-entropy models*. PLoS Comput. Biol. **13**$^{10}$, e1005762. See also the slightly different version arXiv:1605.04740.

Roudi, Y., Aurell, E., Hertz, J. A. (2009a): *Statistical physics of pairwise probability models*. Front. Comput. Neurosci. **3**, 22.

Roudi, Y., Nirenberg, S., Latham, P. E. (2009b): *Pairwise maximum entropy models for studying large biological systems: when they can work and when they can't*. PLoS Comput. Biol. **5**$^{5}$, e1000380.

Roudi, Y., Tyrcha, J., Hertz, J. (2009c): *Ising model for neural data: model quality and approximate methods for extracting functional connectivity*. Phys. Rev. E **79**$^{5}$, 051915.

Rumelhart, D. E., McClelland, J. L., PDP Research Group, eds. (1999): *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations*, 12th printing. (MIT Press, Cambridge, USA).

Sampford, M. R., Scott, A., Stone, M., Lindley, D. V., Smith, T. M. F., Kerridge, D. F., Godambe, V. P., Kish, L., et al. (1969): *Discussion on professor Ericson's paper*. J. Roy. Stat. Soc. B **31**$^{2}$, 224–233. http://www.stat.cmu.edu/~brian/905-2008/papers/Ericson-JRSSB-1969.pdf. See Ericson (1969b).

Schneidman, E., Berry II, M. J., Segev, R., Bialek, W. (2006): *Weak pairwise correlations imply strongly correlated network states in a neural population*. Nature **440**$^{7087}$, 1007–1012. http://www.weizmann.ac.il/neurobiology/labs/schneidman/The_Schneidman_Lab/Publications.html.

Shimazaki, H., Sadeghi, K., Ishikawa, T., Ikegaya, Y., Toyoizumi, T. (2015): *Simultaneous silence organizes structured higher-order interactions in neural populations*. Sci. Rep. **5**, 9821.

Shlens, J., Field, G. D., Gauthier, J. L., Grivich, M. I., Petrusca, D., Sher, A., Litke, A. M., Chichilnisky, E. J. (2006): *The structure of multi-neuron firing patterns in primate retina*. J. Neurosci. **26**$^{32}$, 8254–8266. See also correction in Shlens, Field, Gauthier, Grivich, Petrusca, Sher, Litke, Chichilnisky (2008).

— (2008): *Correction, the structure of multi-neuron firing patterns in primate retina*. J. Neurosci. **28**$^{5}$, 1246. See Shlens, Field, Gauthier, Grivich, Petrusca, Sher, Litke, Chichilnisky (2006).

Sivia, D. S. (2006): *Data Analysis: A Bayesian Tutorial*, 2nd ed. (Oxford University Press, Oxford). Written with J. Skilling. First publ. 1996.

Skilling, J. (1989a): *Classic maximum entropy*. In: Skilling (1989b), 45–52.

— ed. (1989b): *Maximum Entropy and Bayesian Methods: Cambridge, England, 1988*. (Kluwer, Dordrecht).

— (1990): *Quantified maximum entropy*. In: Fougère (1990), 341–350.

Smith, C. R., Grandy Jr., W. T., eds. (1985): *Maximum-Entropy and Bayesian Methods in Inverse Problems*. (Reidel, Dordrecht).

Smolensky, P. (1986): *Information processing in dynamical systems: foundations of harmony theory*. In: Rumelhart, McClelland, PDP Research Group (1999), ch. 6, 194–281.

Tkačik, G., Mora, T., Marre, O., Amodei, D., Palmer, S. E., Berry II, M. J., Bialek, W. (2014): *Thermodynamics and signatures of criticality in a network of neurons*. Proc. Natl. Acad. Sci. (USA) **112**$^{37}$, 11508–11513.

Tkačik, G., Schneidman, E., Berry II, M. J., Bialek, W. (2006): *Ising models for networks of real neurons*. arXiv:q-bio/0611072.

— (2009): *Spin glass models for a network of real neurons*. arXiv:0912.5409.