
Hidden assumptions on population size in the maximum-entropy method

V. Rostami
Forschungszentrum Jülich INM-6
Hitlerland
v.rostami@fz-juelich.de


P.G.L. Porta Mana
Mussoliniland

E. Torre
Moneylaunderingland

Abstract

1 Implicit assumptions in the maximum-entropy method

The maximum-entropy method is used in neuroscience blahblah***

 **Version 1** The present note has three nested purposes.

The first and more general is to remind our readers that the maximum-entropy method rests on subtle assumptions that may be contradictory with other assumptions we want to make on the data under study.


As a concrete example we show – and this is the second purpose – that a sample of a neuronal population cannot be assigned a “maximally noncommittal” maximum-entropy distribution and at the same time be considered a “random representative” of the population.

The proof of the example above rests on some probability relations from classical random sampling. These relations are somewhat hard to find explicitly stated in the neuroscientific literature; the third, minor purpose of this note is to state them explicitly.

Let’s clarify at once that our discussion pertains to a binary population at one specific instant or short window in time. It is possible to similarly discuss multi-state neuron models and population dynamics; but for simplicity neurons are here assumed to be in a fixed, active “1” or inactive “0” state; and evolution, change, time correlations, and similar concepts do not concern us.

Maximum-entropy models are used for a variety of (sometimes not completely clear) reasons. For our purposes let’s say that the maximum-entropy method is assumed to generate a “maximally noncommittal” [1] probability distribution. We intend the quoted expression only as an umbrella term, also because it means very little without further clarifications.

The neurons recorded in these kinds of experiments are usually picked out according to an unknown process, and from their observation we expect to learn something about all other neurons in the same brain area. That is, they are assumed to be a “representative random sample” of the neurons in that area. Our discussion hinges on this often just tacitly understood assumption.

 **Version 2** In this note we would like to show that there is a contradiction in applying maximum-entropy to a “representative random sample” of neurons to generate a “maximally noncommittal” distribution for that sample.

The contradiction is this. Assume that our sample of neurons is representative of some population. If we assign a maximum-entropy distribution to the sample, then we cannot assign a maximum-entropy distribution to the full population. Vice versa, If we assign a maximum-entropy distribution to the full population, then we cannot assign a maximum-entropy distribution to the sample. This impossibility

holds at least for maximum-entropy distribution with moment constraints, and even if the orders of the moments constrained in the sample and in the full population are different.

In other words, if our sample is a “random representative” of some population, then we must choose which has a “maximally noncommittal” distribution: the sample or the population. We can’t choose both. It goes without saying that if one is “maximally noncommittal”, the other must be somehow “committal”. Which choice is most meaningful?

In the rest of the paper we will mathematically show the contradiction above and generate a maximum-entropy distribution for the full population rather than the sample. We will also present some probability relations relevant to random sampling. These relations 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.

Our notation follows ISO and ANSI standards [2–4] but for the use of the comma “,” to denote logical conjunction. Probability notation follows Jaynes [5].

2 Setup

We have a population of N binary neurons. Let’s assume that the identities of the neurons can be distinguished, by their spike shapes for example; but other details, like their locations, are unknown. The neurons have a joint state $(X_1, \dots, X_N) =: \mathbf{X}$ having fixed but unknown binary values $(R_1, \dots, R_N) =: \mathbf{R} \in \{0, 1\}^N$. A particular sample of n neurons from this population has joint state $(x_1, \dots, x_n) =: \mathbf{x}$ having fixed binary values $(r_1, \dots, r_n) =: \mathbf{r}$. We will consider various averages for the population and the sample. For this purpose we introduce a general averaging operator $\overline{}$ defined by


$$\overline{X} := \frac{1}{N}(X_1 + X_2 + \dots + X_N), \quad \overline{x} := \frac{1}{n}(x_1 + x_2 + \dots + x_n), \quad (1)$$

$$\overline{XX} := \binom{N}{2}^{-1}(X_1X_2 + X_1X_3 + \dots + X_{N-1}X_N), \quad \overline{xx} := \binom{n}{2}^{-1}(x_1x_2 + x_1x_3 + \dots + x_{n-1}x_n), \quad (2)$$

$$\overline{XXX} := \binom{N}{3}^{-1}(X_1X_2X_3 + \dots + X_{N-2}X_{N-1}X_N), \quad \overline{xxx} := \binom{n}{3}^{-1}(x_1x_2x_3 + \dots + x_{n-2}x_{n-1}x_n), \quad (3)$$

and so on; analogously for the quantities \mathbf{R} and \mathbf{r} . These formulae say that \overline{X} is the fraction of active neurons, \overline{XX} the fraction of simultaneously active pairs out of all $\binom{N}{2}$ pairs, \overline{XXX} the fraction of simultaneously active triplets, and so on.

 Add formula $\underbrace{\overline{X \cdots X}}_{m \text{ factors}} = \binom{N\overline{X}}{m} / \binom{N}{m}$.

 Remove this A sample can be “random” in two different ways. Consider the population of neurons named $\{1, 2, 3\}$ for example. The sample $\{1, 3\}$ is random because neurons 1 and 3 were chosen according to an unknown process. The sample $\{X_1, X_2\}$ is random because we don’t know the identity of neurons X_1 and X_2 . Given the probability for the population, the probabilities for the states of these two samples are different: the former is obtained by marginalization, the latter by symmetrization and marginalization.

Our uncertainty about the state of the population is completely expressed by the probability distribution

$$P(X_1 = R_1, X_2 = R_2, \dots, X_N = R_N | I) \quad \text{or} \quad P(\mathbf{X} = \mathbf{R} | I), \quad \mathbf{R} \in \{0, 1\}^N, \quad (4)$$

where I denotes our initial state of knowledge (the evidence and assumptions backing our probability assignments). Our uncertainty about the state of the sample is likewise expressed by

$$P(x_1 = r_1, x_2 = r_2, \dots, x_n = r_n | I) \quad \text{or} \quad P(\mathbf{x} = \mathbf{r} | I), \quad \mathbf{r} \in \{0, 1\}^n. \quad (5)$$

3 Initial assumptions: the probability of random representative samples

Before any experimental observations are made we need to make an initial probability assignment, no matter what kinds of predictions we are interested in. This initial assignment will be modified by the experimental observations.

We yet know very little about the physical details of the individual neurons, like their locations. Our state of knowledge is therefore symmetric, or “exchangeable”, under their permutations. This symmetry must be reflected in our initial probability, and the *representation theorem for finite exchangeability* states that it must obey

$$P(\mathbf{X} = \mathbf{R} | I) = \left(\frac{N}{N\bar{\mathbf{R}}} \right)^{-1} P(\bar{\mathbf{X}} = \bar{\mathbf{R}} | I), \quad (6)$$

the latter being the probability for the population average \mathbf{X} . Proof and generalizations to non-binary and continuum cases are given by de Finetti [6], Ericson [7], Diaconis [8], Heath & Sudderth [9]. This theorem is intuitive: owing to symmetry, all states with $N\bar{\mathbf{R}}$ active neurons must have equal probabilities.

By marginalization we obtain the probability for the state of the sample:

$$P(\mathbf{x} = \mathbf{r} | I) = \left(\frac{n}{n\bar{\mathbf{r}}} \right)^{-1} P(\bar{\mathbf{x}} = \bar{\mathbf{r}} | I), \quad (7)$$

with

$$P(\bar{\mathbf{x}} = \bar{\mathbf{r}} | I) = \sum_{N\bar{\mathbf{R}}=0}^N P(\bar{\mathbf{x}} = \bar{\mathbf{r}} | \bar{\mathbf{X}} = \bar{\mathbf{R}}, I) P(\bar{\mathbf{X}} = \bar{\mathbf{R}} | I), \quad (8)$$

$$P(\bar{\mathbf{x}} = \bar{\mathbf{r}} | \bar{\mathbf{X}} = \bar{\mathbf{R}}, I) = \left(\frac{n}{n\bar{\mathbf{r}}} \right) \left(\frac{N-n}{N\bar{\mathbf{R}}-n\bar{\mathbf{r}}} \right) \left(\frac{N}{N\bar{\mathbf{R}}} \right)^{-1}. \quad (9)$$

The latter conditional probability is a hypergeometric distribution, typical of “drawing without replacement” problems. The combinatorial proof of the formulae above is in fact the same as for this class of problems [5 ch. 3; 10 § 4.8.3; 11 § II.6].

The symmetry present in the probabilities above is not a physical property of the neuronal population. It only expresses the symmetry of our initial ignorance about the population, and does not imply any sort of physical similarity between the neurons. Subsequent observations may in fact break this symmetry. The initial symmetry should intuitively also apply to the sample. This is indeed the case: the probability for the state of the sample (25) automatically satisfies the representation theorem (22) as well.

Our initial knowledge (or ignorance) I has two very important properties derivable from the formulae above. First, *our initial expectations for all averages of the sample and the full population are the same*:

$$E(\underbrace{\bar{\mathbf{x}} \cdots \bar{\mathbf{x}}}_{m \text{ factors}} | I) = E(\underbrace{\bar{\mathbf{X}} \cdots \bar{\mathbf{X}}}_{m \text{ factors}} | I) = \left(\frac{n}{m} \right)^{-1} \sum_{n\bar{\mathbf{r}}=0}^n \left(\frac{n\bar{\mathbf{r}}}{m} \right) P(\bar{\mathbf{x}} = \bar{\mathbf{r}} | I) \equiv \left(\frac{N}{m} \right)^{-1} \sum_{N\bar{\mathbf{R}}=0}^N \left(\frac{N\bar{\mathbf{R}}}{m} \right) P(\bar{\mathbf{X}} = \bar{\mathbf{R}} | I). \quad (10)$$

Second, upon observation of a sample average, say $\bar{\mathbf{x}}\bar{\mathbf{x}} = a$, the updated expectations for such average in the population *and in any new sample* will usually be shifted towards the observed value, as follows from Bayes’s theorem and the formulae above.

 [Add paragraph on moments](#)

These two properties express the fact that the sample is *representative* of the population. Our initial knowledge I thus assumes that we can make predictions about the population by observing a sample. Contrast this with an assumption that assigns an independent probability to each neuron: in this case, the observation of one sample does not change our initial probability about another; we can’t learn from observation.

4 Enter maximum-entropy: quandary

Formulae (22)–(27) are constraints on our initial probability assignment, but do not determine it numerically yet. The probability $P(\bar{\mathbf{X}} = \bar{\mathbf{R}} | I)$ for the population average needs to be numerically specified, and by (25) it will determine that of the sample average, $P(\bar{\mathbf{x}} = \bar{\mathbf{r}} | I)$. If we numerically

specify the latter, the former is not completely specified, because eq. (25) linearly constrains $N + 1$ unknowns by only $n + 1$ equations in this case.

We may want to specify the probability by enforcing several of its sample expectations to have specific values, for example $E(\bar{x}) = a$, $E(\bar{x}\bar{x}) = b$. But this is still an underdetermined linear problem because several distributions can have the same desired expectations, as clear from eqs (27).

It is at this point that the maximum-entropy method enters the scene. It selects one distribution, purported to be “maximally noncommittal”, among those that have the desired expectations. But we are in a quandary: formulae (27) allow us to apply the method to find the probability of the population $P(\bar{X} = \bar{R} | I)$, or of the sample $P(\bar{x} = \bar{r} | I)$; *the two applications, however, are inequivalent*. They lead to numerically different $P(\bar{x} = \bar{r} | I)$. Here is an example.

Suppose our desired constraints are $E(\bar{x}) = c_1$ and $E(\bar{x}\bar{x}) = c_2$, the expectations being given by the fourth term of eq. (27). Applying maximum-entropy to the distribution of the average of the full population we find

$$P(\bar{X} = \bar{R} | I') = \Lambda \left(\frac{N}{N\bar{R}} \right) \exp \left[\Lambda_2 \frac{\bar{R}(N\bar{R} - 1)}{N - 1} + \Lambda_1 \bar{R} \right] \quad (11)$$

with such parameters Λ , Λ_1 , Λ_2 as to satisfy normalization and constraints. We call this state of knowledge I' . Marginalization by eq. (23) then gives

$$P(\bar{x} = \bar{r} | I') = \Lambda \sum_{N\bar{R}=0}^N \binom{n}{n\bar{r}} \binom{N-n}{N\bar{R}-n\bar{r}} \exp \left[\Lambda_2 \frac{\bar{R}(N\bar{R} - 1)}{N - 1} + \Lambda_1 \bar{R} \right]. \quad (12)$$

Applying maximum-entropy at the sample level, using the third term of eq. (27), we find

$$P(\bar{x} = \bar{r} | I'') = \lambda \binom{n}{n\bar{r}} \exp \left[\lambda_2 \frac{\bar{r}(n\bar{r} - 1)}{n - 1} + \lambda_1 \bar{r} \right] \quad (13)$$

with such parameters λ , λ_1 , λ_2 as to satisfy the constraints. We call this state of knowledge I'' . For example, with $N = 5000$, $n = 100$, $c_1 = 0.45$, $c_2 = 0.35$, we find $\Lambda_1 = -13\,321.9$, $\Lambda_2 = 13\,321.5$, $\lambda_1 = -269.3$, $\lambda_2 = 268.9$, and appropriate normalization constants. The resulting distributions $P(\bar{x} = \bar{r} | I')$, $P(\bar{x} = \bar{r} | I'')$ are shown in fig. 1; they are clearly different.

The inequivalence of the two maximum-entropy applications is generally valid for small number of constraints, even if the two states of knowledge I' , I'' use different numbers m' , m'' and values of constraints. The equality of the third and fourth term of eq. (27) in this case implies $n + 1$ equations in $m' + m'' + 1$ unknowns (one normalization is taken care of), and in general can only be satisfied if $n \leq m' + m''$.

Which application of the maximum-entropy method should we choose, then?

5 Discussion

We prefer not to answer the question that closed the preceding section, because the optimal answer can only be given case by case, depending on the inferences we are trying to make, on the assumptions needed and negligible, and on our computational power. We would just like to stress that the answer is not unique, and some answers may imply assumptions we were not aware of.

The probability distribution for the sample average obtained by applying maximum-entropy at the population level has by construction a lower entropy than the one obtained by application at the sample level. It says that some sample states x should be considered less likely than in the other distribution. The lower entropy signals the presence of an additional physical assumption behind the former distribution. Such assumption, intuitively, is that the sample is part of a larger population, of which it is representative. And this assumption may affect our predictions.

The quandary we found in the present note has at least two bright sides. The first is that the two different maximum-entropy applications may give numerically similar results in some cases. One could thus be used as an approximation of the other, depending on the circumstances – e.g. if probability tails are unimportant. See fig. 2 for an example. The second is that the possibility of

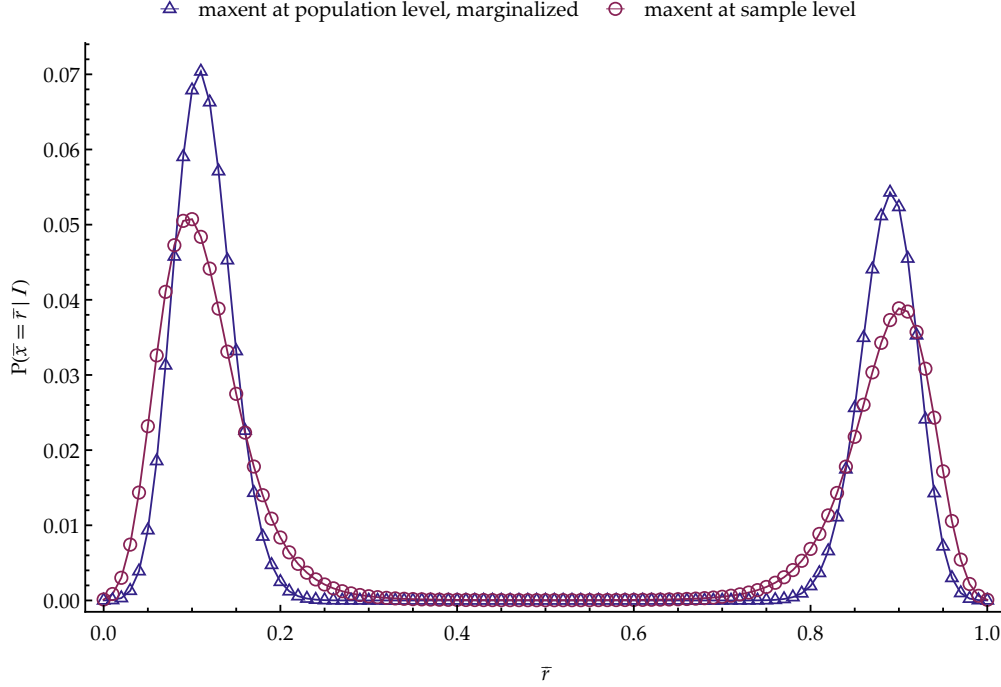


Figure 1: Example of discrepancy in $P(\bar{x} = \bar{r} | I)$ constructed by maximum-entropy at the population level followed by sample marginalization (blue triangles), and at the sample level (red circles). The constraints are the same in either case: $E(\bar{x} | I) = 0.45$, $E(\bar{x}\bar{x} | I) = 0.35$, and $N = 5\,000$, $n = 100$.

using two different distributions is not a physical contradiction. Similar situations arise in statistical mechanics. It is known that if a system is described by a maximum-entropy Gibbs state, its subsystems need not be [12]. A quandary somehow similar to ours also appears in the statistical description of an equilibrium state at the end of a non-equilibrium process: we can describe our knowledge about it either by a Gibbs distribution, or by the Liouville-evolved Gibbs distribution associated with equilibrium state at the beginning of the process. The two descriptions differ – even though the final physical state is exactly the same [13 § 4]. The appearance of this difference is natural: in one case we can make sharper predictions about the state thanks to our knowledge of its preceding dynamics.

✚ but in this case both distributions are immensely sharp, and this doesn't affect predictions

✚ Importance of the formulae relating sample and population.

Acknowledgments

PGLPM thanks Mari & Miri for continuous encouragement and affection, Buster Keaton for filling life with awe and inspiration, and the developers and maintainers of \LaTeX , Emacs, AUCTeX , MiKTeX , arXiv, biorXiv, PhilSci, Hal archives, Python, Inkscape, Sci-Hub for making a free and unfiltered scientific exchange possible.

References

- [1] E. T. Jaynes: *Information theory and statistical mechanics*. In: Ford [14] (1963), 181–218. Repr. in ref. [15 ch. 4, pp. 39–76]; <http://bayes.wustl.edu/etj/node1.html>.
- [2] *Quantities and units*, 3rd ed. International Organization for Standardization. Geneva (1993).
- [3] *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. New York (1993).
- [4] *Guide for the Use of the International System of Units (SI): NIST special publication 811, 1995 edition*. National Institute of Standards and Technology. Washington, D.C. (1995). <http://physics.nist.gov/cuu/Uncertainty/bibliography.html>.

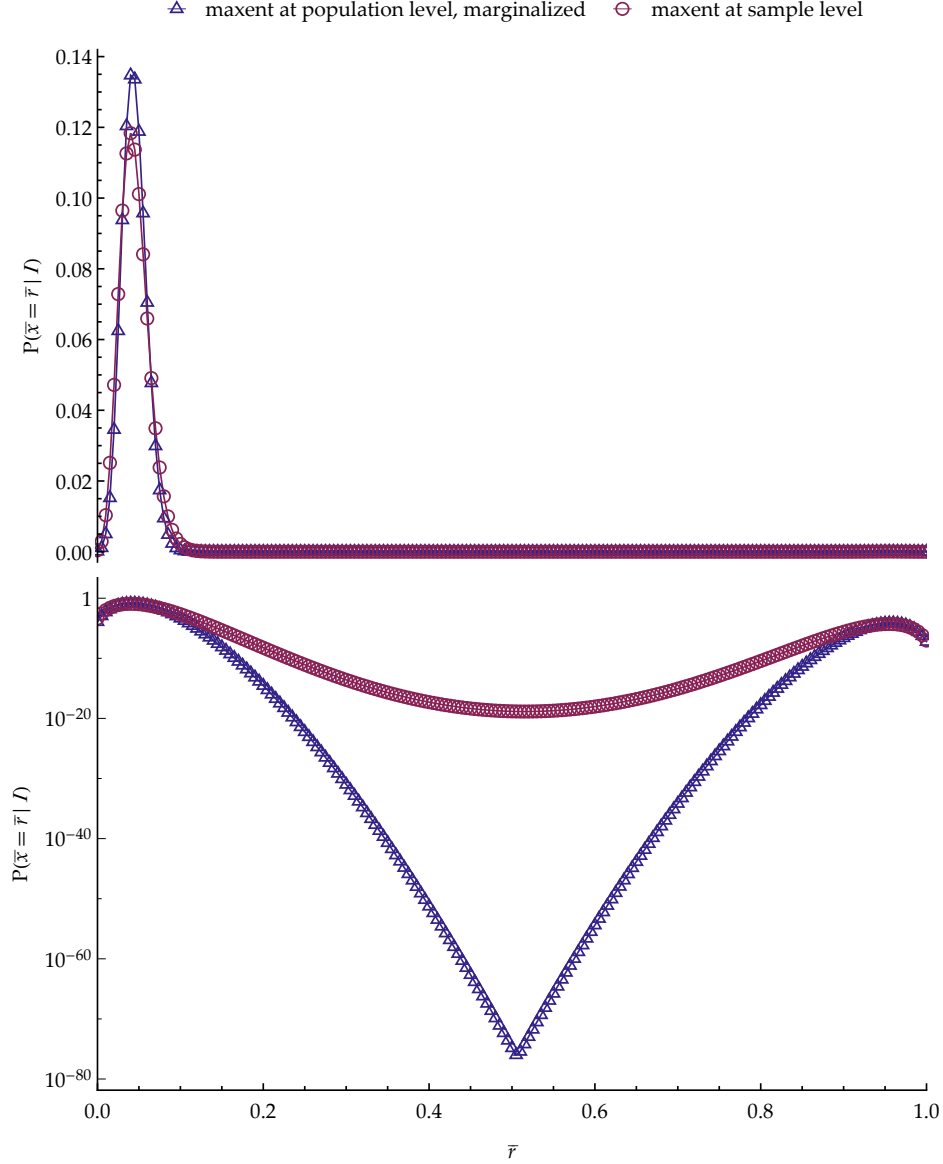


Figure 2: Example of discrepancy in $P(\bar{x} = \bar{r} | I)$ constructed by maximum-entropy at the population level followed by sample marginalization (blue triangles), and at the sample level (red circles). The constraints are the same in either case: $E(\bar{x} | I) = 0.045$, $E(\bar{x}\bar{x} | I) = 0.0025$, and $N = 5\,000$, $n = 200$.

- [5] E. T. Jaynes: *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge (2003). Ed. by G. Larry Bretthorst; <http://omega.albany.edu:8008/JaynesBook.html>, <http://omega.albany.edu:8008/JaynesBookPdf.html>, <http://www-biba.inrialpes.fr/Jaynes/prob.html>; first publ. 1994.
- [6] B. de Finetti: *La probabilità e la statistica nei rapporti con l'induzione, secondo i diversi punti di vista*. In: de Finetti [16] (1959), 1–115. Transl. as ref. [17].
- [7] W. A. Ericson: *A Bayesian approach to two-stage sampling*. Tech. rep. University of Michigan, Ann Arbor, USA (1976). <http://hdl.handle.net/2027.42/4819>.
- [8] P. Diaconis: *Finite forms of de Finetti's theorem on exchangeability*. *Synthese* **36**² (1977), 271–281. <http://statweb.stanford.edu/~cgates/PERSI/year.html>.
- [9] D. Heath, W. Sudderth: *De Finetti's theorem on exchangeable variables*. *American Statistician* **30**⁴ (1976), 188–189.

- [10] S. Ross: *A First Course in Probability*, 8th ed. Pearson, Upper Saddle River, USA (2010). First publ. 1976.
- [11] W. Feller: *An Introduction to Probability Theory and Its Applications. Vol. I*, 3rd ed. Wiley, New York (1968). First publ. 1950.
- [12] C. Maes, F. Redig, A. Van Moffaert: *The restriction of the Ising model to a layer*. J. Stat. Phys. **96**¹ (1999), 69–107. [arXiv:math/9810094](https://arxiv.org/abs/math/9810094).
- [13] E. T. Jaynes: *Inferential scattering*. (1993). <http://bayes.wustl.edu/etj/node1.html>; extensively rewritten version of a paper first publ. 1985 in ref. [18], pp. 377–398.
- [14] K. W. Ford, ed.: *Statistical Physics*. Benjamin, New York (1963).
- [15] E. T. Jaynes: *E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics*, reprint. Kluwer, Dordrecht (1989). Ed. by R. D. Rosenkrantz. First publ. 1983.
- [16] B. de Finetti, ed.: *Induzione e statistica*, reprint. Springer, Berlin (2011). First publ. 1959.
- [17] B. de Finetti: *Probability, statistics and induction: their relationship according to the various points of view*. In: de Finetti [19] (1972), p. 9, 147–227. Transl. of ref. [6].
- [18] C. R. Smith, W. T. Grandy Jr., eds.: *Maximum-Entropy and Bayesian Methods in Inverse Problems*. D. Reidel, Dordrecht (1985).
- [19] B. de Finetti: *Probability, Induction and Statistics: The art of guessing*. Wiley, London (1972).

arXiv eprints available at <http://arxiv.org/>.

 *** From here down old text to be discarded ***

6 ***

What do we mean by saying that we can learn something about the population by observing the sample? At the very least we mean that our uncertainty about the rest of the population can change upon observing the sample. That is, the probability for the state \mathbf{y} of the rest of the population is conditionally dependent on the sample's values:

$$P(\mathbf{y} = \mathbf{s} | \mathbf{x} = \mathbf{r}, I) \neq P(\mathbf{y} = \mathbf{s} | I) \quad \text{for some } \mathbf{s}, \mathbf{r}, \quad (14)$$

This also means that the probability for the population cannot be factorized into that of the sample times that of the complement.

What do we mean when we say that the sample is representative of the population? It means that we expect some collective properties of the sample, like the fraction of active neurons, or the fraction of simultaneously active pairs, to be roughly equal to those of the full population.

For example, \mathbf{X} can represent the state of a population at a particular time. We call the neurons “units” to lend some generality to our discussion. We shall make statements about the whole population of N units and about a subpopulation of n units; the word “population” will always refer to the *whole* population. The subpopulation states and their values are denoted by lowercase letters: $(x_1, \dots, x_n) \equiv \mathbf{x}$ and $(r_1, \dots, r_n) \equiv \mathbf{r}$; but note that $x_i \equiv X_{j_i}$ and $r_i \equiv R_{j_i}$ for some distinct j_1, \dots, j_n . We shall also make statements about the population-averaged state, or *population average*:

$$\bar{X} := (X_1 + \dots + X_N)/N, \quad (15)$$

and the subpopulation-averaged state, or *subpopulation average*:

$$\bar{\mathbf{x}} := (x_1 + \dots + x_n)/n. \quad (16)$$

The quantities $N\bar{X}$ and $n\bar{\mathbf{x}}$ represent the total number of active units in the population and the subpopulation. Quantities like $\bar{\mathbf{R}}$ and $\bar{\mathbf{r}}$ are defined analogously. The averaging operators $\bar{\cdot}$ and $\bar{\cdot}$ are also extended to averages of $\binom{n}{m}$ or $\binom{N}{m}$ products of m states; e.g.,

$$\overline{X X} := \binom{N}{2}^{-1} (X_1 X_2 + X_1 X_3 + \dots + X_{N-1} X_N), \quad (17)$$

$$\overline{\mathbf{x} \mathbf{x} \mathbf{x}} := \binom{n}{3}^{-1} (x_1 x_2 x_3 + x_1 x_2 x_4 + \dots + x_{n-2} x_{n-1} x_n), \quad (18)$$

and so on.

6.1 Assumptions

Our uncertainty about the population state is represented by the joint probability distribution of the individual states, from which we can derive all other probabilities of interest. We denote it by

$$P(X_1 = R_1, X_2 = R_2, \dots, X_N = R_N | I) \quad \text{or} \quad P(\mathbf{X} = \mathbf{R} | I). \quad (19)$$

Such probability is conditional on our state of knowledge, i.e. the evidence and assumptions backing our probability assignments, denoted by the proposition I .

In the present discussion, I is a state of knowledge that leads to two specific properties in our probability assignments:

1. *Permutation symmetry*, expressed as the invariance of the joint distribution (19) under arbitrary permutations of the units's labels:

$$P(X_1 = R_1, X_2 = R_2, \dots, X_N = R_N | I) = P(X_1 = R_{\pi(1)}, X_2 = R_{\pi(2)}, \dots, X_N = R_{\pi(N)} | I) \quad \text{for any permutation } \pi. \quad (20)$$

This property can reflect two very different states of knowledge: physical homogeneity of the population, or symmetry in our ignorance about the population. This property is called *finite exchangeability* in the Bayesian literature and its basis, consequences, and alternatives to it are discussed in § ??.

2. The population average \bar{X} has a particular distribution Q :

$$P(\bar{X} = A | I) = Q(A), \quad A \in \{0, \frac{1}{N}, \frac{2}{N}, \dots, 1\}. \quad (21)$$

For the moment we are not concerned about the specific form of Q and about how it was assigned: it could, e.g., arise from maximum-entropy arguments [e.g.: 1–10] used with data on the population.

6.2 Formulae

The state of knowledge I has the following six (not independent) main consequences for our probability assignments:

1. Probability for the population state:

$$P(\mathbf{X} = \mathbf{R} | I) = \left(\frac{N}{N\bar{\mathbf{R}}} \right)^{-1} Q(\bar{\mathbf{R}}). \quad (22)$$

2. Probability for the state \mathbf{x} of any subpopulation of n units:

$$P(\mathbf{x} = \mathbf{r} | I) = \sum_{NA=0}^N \binom{N-n}{NA-n\bar{\mathbf{r}}} \binom{N}{NA}^{-1} Q(A). \quad (23)$$

Note that the only summands contributing to this sum are those for which $n\bar{\mathbf{r}} \leq NA \leq N$; the others are zero because by definition $\binom{M}{y} = 0$ if $y < 0$. This remark applies to all the sums of this kind in the rest of this Note.

3. Probability for the subpopulation state conditional on a population state:

$$P(\mathbf{x} = \mathbf{r} | \mathbf{X} = \mathbf{R}, I) = \binom{N-n}{N\bar{\mathbf{R}}-n\bar{\mathbf{r}}}. \quad (24)$$

4. Probability for the subpopulation average \bar{x} :

$$P(\bar{x} = a | I) = \binom{n}{na} \sum_{NA=0}^N \binom{N-n}{NA-na} \binom{N}{NA}^{-1} Q(A),$$

$$a \in \{0, \frac{1}{n}, \frac{2}{n}, \dots, 1\}. \quad (25)$$

5. Probability for the subpopulation average conditional on the population average:

$$P(\bar{x} = a | \bar{X} = A, I) = \binom{n}{na} \binom{N-n}{NA-na} \binom{N}{NA}^{-1}. \quad (26)$$

6. The product of the states of any m distinct units from a given subpopulation,

$$x_{i_1} x_{i_2} \cdots x_{i_m}, \quad 1 \leq i_1 < i_2 < \cdots < i_m \leq n$$

has an expectation equal to that of the subpopulation average of such products, is independent of the subpopulation size n :

$$E(x_{i_1} \cdots x_{i_m} | I) = E(\underbrace{\bar{x} \cdots \bar{x}}_{m \text{ factors}} | I) = E(\underbrace{\bar{X} \cdots \bar{X}}_{m \text{ factors}} | I), \quad (27a)$$

and has an explicit expression in terms of Q :

$$E(x_{i_1} \cdots x_{i_m} | I) = \binom{N}{m}^{-1} \sum_{NA=0}^N \binom{NA}{m} Q(A)$$

$$\equiv \sum_{NA=0}^N \binom{N-m}{NA-m} \binom{N}{NA}^{-1} Q(A). \quad (27b)$$

A useful relation connects the expectation of a product (27) and the m th factorial moment [11] of the probability distributions for the averages. The m th factorial moment of the subpopulation average \bar{x} is defined by

$$E[\underbrace{n\bar{x}(n\bar{x}-1)\cdots(n\bar{x}-(m-1))}_{m \text{ factors}} | I] \equiv E\left[\frac{(n\bar{x})!}{(n\bar{x}-m)!} | I\right], \quad (28)$$

an analogous definition holding for \bar{X} . We have that

$$E(x_{i_1} \cdots x_{i_m} | I) = \frac{(n-m)!}{n!} E\left[\frac{(n\bar{x})!}{(n\bar{x}-m)!} | I\right] = \frac{(N-m)!}{N!} E\left[\frac{(N\bar{X})!}{(N\bar{X}-m)!} | I\right]. \quad (29)$$

As a consequence of the above relation, the first three moments of the probability distributions $P(\bar{x} = a | I)$ and $P(\bar{X} = A | I)$, are related by

$$E(\bar{x} | I) = E(\bar{X} | I), \quad (30a)$$

$$E(\bar{x}^2 | I) = E(\bar{X} | I) \frac{N-n}{(N-1)n} +$$

$$E(\bar{X}^2 | I) \frac{N(n-1)}{(N-1)n}, \quad (30b)$$

$$E(\bar{x}^3 | I) = E(\bar{X} | I) \frac{(N-n)(N-2n)}{(N-1)(N-2)n^2} +$$

$$E(\bar{X}^2 | I) \frac{3N(N-n)(n-1)}{(N-1)(N-2)n^2} +$$

$$E(\bar{X}^3 | I) \frac{N^2(n-1)(n-2)}{(N-1)(N-2)n^2}. \quad (30c)$$

Relations for higher moments can be obtained recursively from eq. (29). In general, this means that the two sets of first m moments are related by a homogeneous linear transformation,

$$\mathbb{E}(\bar{x}^m | I) = \sum_{l=1}^m M_{ml}(n, N) \mathbb{E}(\bar{X}^l | I), \quad (31)$$

with a universal, lower-triangular transformation matrix $M_{ml}(n, N)$ that depends only on n, N , and the condition of symmetry (20).

As intuition suggests, we have

$$\mathbb{E}(\bar{x}^m | I) \xrightarrow{n \rightarrow N} \mathbb{E}(\bar{X}^m | I), \quad \mathbb{E}(\bar{x}^m | I) \xrightarrow{n \rightarrow 1} \mathbb{E}(\bar{X} | I), \quad (32)$$

the latter because $x_i^m = x_i$, since states are $\{0, 1\}$ -valued.

The core of the six mathematical relations above are eqs (23) and (25). The latter expresses the probability for the subpopulation average as a mixture of hypergeometric distributions [12 ch. 3; 13 § 4.8.3; 14 § II.6], with parameters $N, N\bar{X}, n$, weighted by the probabilities $P(\bar{X} = A | I)$ [cf. 15 § 4, esp. eq. (22)]. The connection between this mixture representation and the condition of symmetry (20) is well-known in the Bayesian literature [15–20].

7 Examples of inferential use of the formulae

7.1 From network to subnetwork

Let us illustrate with an example how the probability distribution for the subnetwork average \bar{x} , determined by eq. (25), changes with the subnetwork size n . Choose a network-average distribution $P(\bar{X} = A | I)$ belonging to the exponential family [21 § 4.5.3; see also 22]:

$$P(\bar{X} = A | I) = Q(A) \propto \binom{N}{NA} \exp[\lambda_2 NA (NA - 1)/2 + \lambda_1 NA]. \quad (33)$$

This is the form obtained from the principle of maximum relative entropy [e.g.: 1–10] with first and second moments as constraints and the reference distribution Q_0 defined by $Q_0(A) = 2^{-N} \binom{N}{NA}$, corresponding to a uniform probability distribution for the network state X .

The probability distribution of eq. (33) is plotted in fig. 3, together with the resulting subnetwork-average distributions $P(\bar{x} = a | I)$, for the case in which $N = 1000$ units, $\lambda_1 = -2.55$, $\lambda_2 = 0.005$, and $n = 10, 50, 100, 250$. The distributions become broader as n decreases, and the minimum of the original distribution disappears; at the same time the finite-difference

$$\frac{P(\bar{x} = a + 1/n | I) - P(\bar{x} = a | I)}{1/n}$$

presents a sharp jump at this minimum when $n \approx 100$.

To the eye familiar with maximum-entropy distributions, the subnetwork-average distributions of fig. 3 do not look like maximum-entropy ones with second-moment constraints. In fact, they are not and *cannot* be:

$$P(\bar{x} = a | I) \neq \kappa \binom{n}{na} \exp[\kappa_2 na (na - 1)/2 + \kappa_1 na] \quad (34)$$

for any $\kappa, \kappa_1, \kappa_2$, unless $n = 2$. This impossibility holds more generally for any number of constraints m and subnetwork size n such that $m < n$. The reason is simple: suppose we have assigned a maximum-entropy distribution with m moment constraints as the distribution for the network average. If we want the same kind of distribution for a subnetwork of size n , we are free to play with $m + 1$ parameters (normalization included), but we must also satisfy the $n + 1$ equations corresponding to the marginalization (25). This is generally impossible unless $m \geq n$. (Impossibilities of a similar kind appear in statistical mechanics, see e.g. ref. [23].)

This fact can be significant for recent works [e.g., 24–32] in which a maximum-entropy probability distribution with second- or third-moment constraints is assigned to relatively small subnetworks

($n < 200$) of neurons. If we assume that such subnetwork is part of a larger network, and assume the condition of symmetry (20), then the larger network *cannot* be assigned a maximum-entropy distribution with the same number of constraints. Vice versa, if we assign such a maximum-entropy

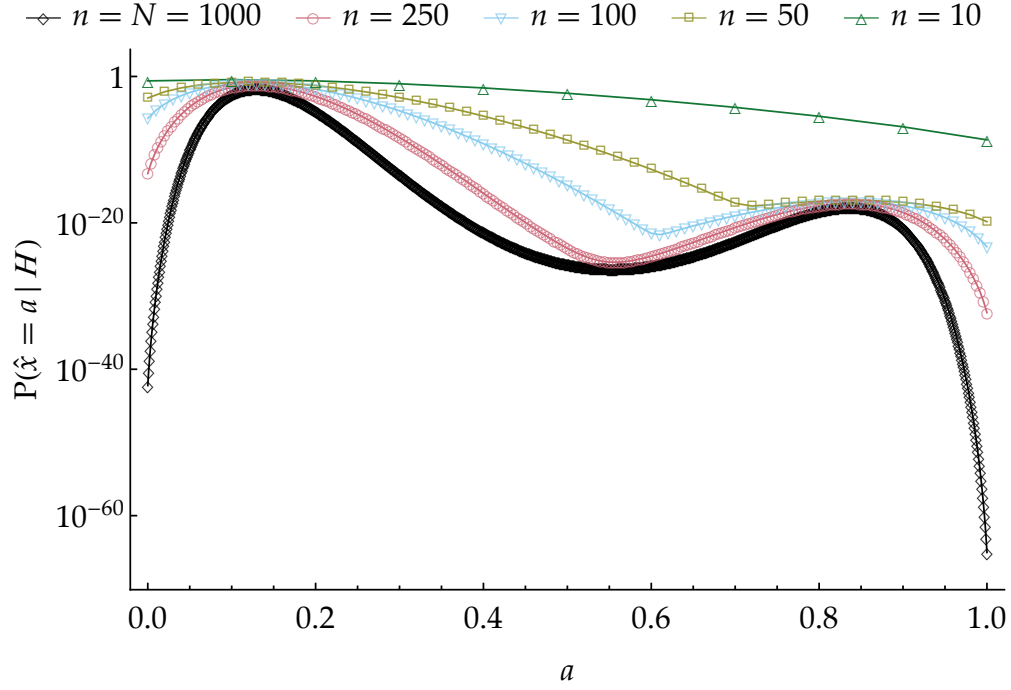


Figure 3: Probability distributions $P(\bar{x} = a | I)$ for different subnetwork sizes n , obtained from a network probability distribution $P(\bar{X} = A | I)$ having the maximum-entropy form (33).

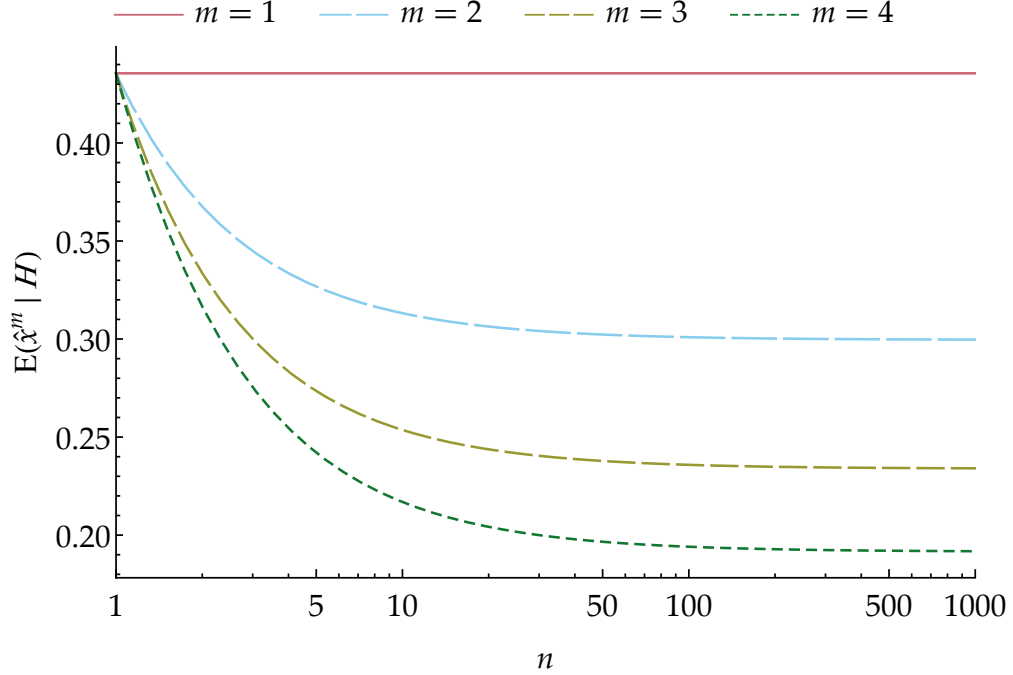


Figure 4: Moments of the probability distributions $P(\bar{x} = a | I)$ as functions of the subnetwork size n .

distribution to the larger network, then none of its subnetwork of enough large size n can be assigned a similar maximum-entropy distribution. See ref. [33] for a broader discussion of this fact and of its consequences.

The dependence of the first four moments $E(\bar{x}^m | I)$ as a function of size n is shown in fig. 4. The moments become practically constant when $n \approx 100$ or larger. The expectations of m -tuple products of states $E(x_{i_1} \cdots x_{i_m} | I)$, proportional to the factorial moments, are not shown as they do not depend on n .

7.2 From subnetwork to network

We have seen that, given the condition of symmetry (20), the probability $P(\bar{X} = A | I)$ for the network average determines that of each subnetwork average, $P(\bar{x} | I)$, by the marginalization eq. (25). The reverse is trivially not true, since eq. (25), as a linear mapping from \mathbf{R}^{N+1} to \mathbf{R}^{n+1} , with N larger than n , is onto but not into. Assigning a probability distribution $P(\bar{x} = a | I)$ to a subnetwork average \bar{x} does not determine a network distribution $P(\bar{X} = A | I)$: it only restricts the set of possible ones; this set can in principle be determined via linear-programming methods [34–38].

Analogous situations appear in the truth-valued logical calculus: if the composite proposition $A \Rightarrow B$ is assigned the truth-value “true”, then assigning A the value “true” also determines the value of B , whereas assigning B the value “true” leaves the value of A undetermined.

The same linear-programming methods show that any inference from subnetwork properties to network ones must necessarily start from some assumptions I that assign a probability distribution $P(X = \mathbf{R} | I)$ for the network states. The approaches to this task and reformulations of it have become uncountable: they include exchangeable models, parametric and non-parametric models, hierarchical models, general linear models, models via sufficiency, maximum-entropy models, and whatnot [e.g.: 39; 40; 12; 21; 41–48]. We now show two examples, based on a maximum-entropy approach, that to our knowledge have not yet been explored in the neuroscientific literature. For a concrete application see [49].

First example: moment constraints for the network. Consider a state of knowledge I' leading to the following properties:

1. the expectations of the single and pair averages \bar{x} and $\bar{x}\bar{x}$ of a particular subnetwork have given values
$$E(\bar{x} | I') = c_1, \quad E(\bar{x}_i \bar{x}_j | I') = c_2; \quad (35)$$
2. the network probability distribution $P(X = \mathbf{R} | I')$ has maximum relative entropy with respect to the uniform one, given the constraints above.

Then the probability distribution for the network conditional on I' is completely determined: it satisfies the symmetry property (20) and is defined by

$$p(X = \mathbf{R} | I') = K \exp[\Lambda_2 N \bar{\mathbf{R}} (N \bar{\mathbf{R}} - 1)/2 + \Lambda_1 N \bar{\mathbf{R}}]$$

with K, Λ_m , such that the distribution is normalized and

$$K \sum_{NA=0}^N \binom{N-m}{NA-m} \exp[\Lambda_2 NA (NA-1)/2 + \Lambda_1 NA] = c_m, \quad m = 1, 2. \quad (36)$$

We omit the full proof of this statement: it is a standard application of the maximum-entropy procedure [e.g.: 1–4; 6–10], combined with the equality (27) of subnetwork and network expectations, e.g.

$$c_2 = E(\bar{x}\bar{x} | I') = \binom{N}{2}^{-1} \sum_{NA=0}^N \binom{NA}{2} P(\bar{X} = A | I'), \quad (37)$$

and with relations (22), (24). This example is easily generalized to any number m of constraints such that $m \leq n$.

Note again that, as remarked in § 7.1, the subnetwork from which the averages in the expectations (35) are calculated has a probability distribution $P(\bar{x} = a | I)$ determined by the marginalization (25) and does *not* have a maximum-entropy form with the same number of constraints.

Second example: subnetwork-distribution constraint. Consider another state of knowledge I'' leading to the following properties:

1. the average \bar{x} of a particular subnetwork has a probability distribution q :

$$P(\bar{x} = a | I'') = q(a); \quad (38)$$

2. the probability distribution for the network, $P(X = \mathbf{R} | I'')$, has maximum relative entropy with respect to the uniform one, given the constraint above.

Then the probability distribution for the network given I'' is completely determined and satisfies the symmetry property (20):

$$P(X = \mathbf{R} | I'') = \exp \left[\sum_{na=0}^n \Lambda_a \binom{n}{na} \binom{N-n}{N\bar{\mathbf{R}}-na} \right]$$

with Λ_a such that


$$\sum_{NA=0}^N \binom{n}{na} \binom{N-n}{NA-na} \exp \left[\sum_{na=0}^n \Lambda_a \binom{n}{na} \binom{N-n}{NA-na} \right] = q(a) \quad (39)$$

(the normalization constraint being unnecessary since q is normalized). This result is just another application of the maximum-entropy procedure with $n + 1$ (linear) constraints given by eq. (23), where the left-hand side is now given and equal to $q(a)$.

This example is equivalent to the generalization of the previous one with n moment constraints, since knowledge of $P(\bar{x} = a | I'')$ is equivalent to knowledge its first n moments.

In this note we would like to analyse and warn about a subtle assumption behind the maximum-entropy method when it is applied to a population. It can informally be put this way:


the maximum-entropy method assumes that the population it is applied to is completely isolated from any larger population.

 **Version 1** The maximum-entropy method does not construct a probability distribution out of nothing, but starting from a uniform distribution. A uniform distribution is an innocuous assumption for a set of non-composite events, like the outcomes of a die roll, and also for some sets of composite events, like the outcomes of the roll of two dice. In the latter case multiplicities appear.

When applied to a subpopulation, the maximum-entropy method assumes that the uniform over the larger population is uniform, and therefore factorizable. The new distribution of the subpopulation will not be uniform, but that of the full population will still be factorizable into the one for the subpopulation and the rest.

A uniform distribution, however, is not the right one when we suppose that learning about an event may tell us something about a related event. For example, consider 1 000 tosses of a particular coin and assume a uniform distribution over the possible 2^{1000} outcomes. If we learn that the first 999 tosses yielded all “heads”, the probability calculus tells us that the probability for the 1 000th toss is still 50%/50%. It is a consequence of our choice of a uniform distribution: we have implicitly declared all tosses to be completely independent, completely *irrelevant* to one another. This fact is well-known in sampling theory. A more telling example in fact is that of a presidential election with two candidates: each citizen will vote for one or the other. We do survey sampling on a large number of citizens to guess the election’s outcome. If we assumed a uniform distribution over the possible combinations of choices of all citizens, our sampling would be completely irrelevant for the choices of the rest of the population.

The latter example has many similarities with that of a neuronal binary population. When we record the neuronal activity of a sample of neurons from a brain area, we assume that our measurements can tell us something – no matter how vague or imprecise – about the whole brain area. This means that we are not assuming a uniform distribution over all possible states of the area.

 **Version 2** This may come as a surprise. The method simply requires a number of exhaustive and mutually exclusive events, and if these are composite events the final distribution may have a

multiplicity factor. When we consider the 2^N states of N units we are not excluding that these might be marginals of 2^M states of M units. Each one has the same multiplicity 2^{M-N} , but this constant multiplicity factor disappears by normalization. So the method applies just as in the case of N units only, right?

Right, but

Right, and that is where the problem lies. This way of counting of multiplicities assumes an underlying

Wrong. In our reasoning we have made subtle assumptions of independence between the full population and the subpopulation. The problem is that the counting of multiplicities is not based on simple enumeration, but already involves probability considerations. Consider three cases with a full population of two units, $M = 2$, of which we consider one unit, $N = 1$.

- First case: all four states are *possible*. The two states of the first unit have multiplicity 2 each. The usual maximum-entropy distribution obtains.
- Second case: only the states with at most one active unit are possible. The state $x_1 = 1$ of the first unit has multiplicity 2, and $x_1 = 0$ has multiplicity 1. The maximum-entropy distribution has multiplicity factors.
- Third case: states with at most one active unit are, say, 10^9 times more probable than the state with no active units. But all four states are *possible*. By enumeration this case is like the first: multiplicities (1, 1). But by common sense it is more similar to the second: multiplicities (2, 1) for most practical purposes.

This simple example shows that the multiplicity inspection that must precede a maximum-entropy application already involves probability considerations at the level of the full population. The usual reasoning by enumeration implicitly assumes a uniform distribution or at least a *factorizable* distribution.

***If the distribution is factorizable, however, it means that examination of the subpopulation *cannot give us any insights about the population it is a part of*. This is obviously contrary to the reason why we made neuronal observations.

Consider the following ways of proceeding. We:

1. have a population with N units, 2^N possible states
2. expect averages of \bar{x} active neurons and $\overline{x\bar{x}}$ active pairs
3. use maximum-entropy to choose a probability distribution for the states of the N units conforming to our expectations.

We:

1. have a population with M units, 2^M possible states
2. expect that any subpopulation of N units has \bar{x} active neurons and $\overline{x\bar{x}}$ active pairs
3. use maximum-entropy to choose a probability distribution for the states of the M units conforming to our expectations
4. marginalize to find the probability distribution for the states of N units.

8 Sketched proofs

Variants of the following derivations and combinatorial considerations can be found e.g. in [50 chs I–IV; 14 ch. II; 12 ch. 3]; see also [51].

To derive the joint probability distribution (22) from that for the network average (21), consider that if the network total is $N\bar{X}$, then $N\bar{X}$ out of N units are active, and there are $\binom{N}{N\bar{X}}$ possible states for which this can be true; therefore

$$P(X = R | H) = \left(\frac{N}{N\bar{R}} \right)^{-1} Q(\bar{R}). \quad (22)_r$$

An analogous reasoning for n and \bar{x} leads to an analogous equality,

$$P(\mathbf{x} = \mathbf{r} | H) = \binom{n}{n\bar{r}}^{-1} P(\bar{\mathbf{x}} = \bar{\mathbf{r}} | I), \quad (40)$$

for the subnetwork.

Let us next consider the probability $P(\bar{\mathbf{x}} = a | \bar{X} = A, I)$ for the subnetwork average $\bar{\mathbf{x}}$ conditional on the network average \bar{X} . There are $\binom{N}{N\bar{X}}$ possible network states if the network average is \bar{X} , i.e. if $N\bar{X}$ units are in state 1; the conditional probability of each is therefore $1/\binom{N}{N\bar{X}}$, owing to the symmetry assumption (20). Now consider the subnetwork of the first n units. The conditional probability of having $n\bar{x}$ specific ones in state 1 is the sum of the probabilities of all states for which $N\bar{X} - n\bar{x}$ of the remaining $N - n$ units are in state 1; there are $\binom{N-n}{N\bar{X}-n\bar{x}}$ such states, all equally probable. Finally, there are $\binom{n}{n\bar{x}}$ possible ways, all equally probable, in which $n\bar{x}$ of the first n units can be in state 1. In formulae,

$$\begin{aligned} P(\bar{\mathbf{x}} = a | \bar{X} = A, I) &= \sum_{\mathbf{r}} \sum_{\mathbf{R}=\bar{X}} P(\mathbf{x} = \mathbf{r} | \mathbf{X} = \mathbf{R}, I) P(\mathbf{X} = \mathbf{R} | \bar{X} = A, I), \\ &= \sum_{\mathbf{r}} \sum_{\mathbf{R}=\bar{X}} P(\mathbf{x} = \mathbf{r} | \mathbf{X} = \mathbf{R}, I) \binom{N}{NA}^{-1}, \\ &= \sum_{\mathbf{r}} \binom{N-n}{NA-na} \binom{N}{NA}^{-1}, \\ &= \binom{n}{na} \binom{N-n}{NA-na} \binom{N}{NA}^{-1}, \end{aligned} \quad (41)$$

which is the conditional probability (26). Note that this is just a derivation of the hypergeometric distribution [12 ch. 3; 13 § 4.8.3; 14 § II.6], which describes the probability of, say, drawing a proportion of \bar{x} blue balls in n drawings without replacement from an urn with N balls, a fraction \bar{X} of which are blue.

The probability of a subnetwork average $\bar{\mathbf{x}}$ is then, by marginalization,

$$\begin{aligned} P(\bar{\mathbf{x}} = a | I) &= \sum_{NA=0}^N P(\bar{\mathbf{x}} = a | \bar{X} = A, I) P(\bar{X} = A | I), \\ &= \sum_{NA=0}^N \binom{n}{na} \binom{N-n}{NA-na} \binom{N}{NA}^{-1} Q(A). \end{aligned} \quad (42)$$

which proves the subnetwork-average formula (25). This formula, combined with eqs (40) and (22), leads to the conditional probability (24).

The independence of the expectation of products of states from the subnetwork size is trivial by marginalization:

$$\begin{aligned} \sum_{\mathbf{x}} x_1 \cdots x_m P(\mathbf{x} = \mathbf{r} | I) &= \sum_{\mathbf{r}, \mathbf{R}} r_1 \cdots r_m P(\mathbf{x} = \mathbf{r} | \mathbf{X} = \mathbf{R}, I) P(\mathbf{X} = \mathbf{R} | I), \\ &= \sum_{\mathbf{r}, \mathbf{R}} r_1 \cdots r_m \delta(R_1 - r_1) \cdots \delta(R_m - r_m) P(\mathbf{X} = \mathbf{R} | I), \\ &= \sum_{\mathbf{R}} R_1 \cdots R_m P(\mathbf{X} = \mathbf{R} | I). \end{aligned} \quad (43)$$

All such m -fold products have the same expectation by symmetry, therefore their subnetwork average will do, too, being an average of equal terms.

Now consider the sum of all distinct products of states of two units in the subnetwork:

$$x_1 x_2 + x_1 x_3 + \cdots + x_{n-1} x_n.$$

The terms in this sum are either 0 or 1. The non-vanishing ones are those with index pairs chosen from the $n\bar{x}$ units of the subnetwork which are in state 1, and there are $\binom{n\bar{x}}{2}$ such choices, so the sum above is equal to $\binom{n\bar{x}}{2}$. The sum has $\binom{n}{2}$ terms, so their average is $\binom{n\bar{x}}{2} / \binom{n}{2}$. Generalizing the argument to products of m units, we have that

$$\overline{x_{i_1} \cdots x_{i_m}} = \binom{n\bar{x}}{m} \binom{n}{m}^{-1}. \quad (44)$$

Then, using eq. (25),

$$\begin{aligned} E(\overline{x_{i_1} \cdots x_{i_m}} | I) &= \frac{(n-m)!}{n!} E\left(m! \binom{n\bar{x}}{m} | I\right) \\ &\equiv \frac{(n-m)!}{n!} \sum_{na=0}^n m! \binom{na}{m} P(\bar{x} = a | I), \\ &= \frac{(n-m)!}{n!} \sum_{NA=0}^N Q(A) \left[\sum_{na=0}^n m! \binom{na}{m} \binom{n}{na} \binom{N-n}{NA-na} \binom{N}{NA}^{-1} \right]. \end{aligned} \quad (45)$$

The expression in brackets is the m th factorial moment of the hypergeometric function, and is given by [11]

$$\sum_{na=0}^n m! \binom{na}{m} \binom{n}{na} \binom{N-n}{N\bar{X}-na} \binom{N}{N\bar{X}}^{-1} = m! \binom{n}{m} \binom{na}{m} \binom{N}{m}, \quad (46)$$

which combined with the previous equation yields the second line of eq. (27); its last equality comes from the identity

$$\binom{N}{M} \binom{M}{m} = \binom{N}{m} \binom{N-m}{M-m}, \quad (47)$$

easily derived by writing the binomial coefficients in terms of factorials. Finally, eqs (30), relating the moments of the distributions for subnetwork and network averages, is obtained from the definition of moments,

$$E(\bar{x}^m | I) := \sum_{na=0}^n a^m P(\bar{x} = a | H), \quad E(\bar{X}^m | I) := \sum_{NA=0}^N A^m P(\bar{X} = A | H), \quad (48)$$

replaced in the equalities for the factorial moments (29), by recursively solving in terms of the moments of the network distribution.