Maximum-entropy distributions and representative random samples: a dilemma

V. Rostami
Hitlerland
v.rostami@fz-juelich.de

P.G.L. Porta Mana
Mussoliniland
pgl@portamana.org

E. Torre
Moneylaunderingland

Abstract

Better title? ***

1 Introduction: maximum-entropy and sampling in neuroscience

This note is mainly addressed to neuroscientists interested in maximum-entropy methods, but its discussion of "random sampling" formulae may be useful to all neuroscientists involved in understanding the physical and dynamical characteristics of brain areas via neuronal recording.

Recent electrophysiological techniques [***] in fact allow experimenters to record samples of even a couple hundreds neurons from specific brain areas. These samples are usually picked out according to an unknown process, but from their observation we expect to learn something about all other neurons in the same brain area. That is, we assumed that they are a "representative random sample".

We don't elaborate on the various important purposes of such recordings here [**?*****], but stress that these sample sizes can be considered "large", because their statistical analysis requires considerable computational power.

These computational costs are one, probably not the earliest, of the many reasons why maximum-entropy methods have been introduced in neuroscience. It would be useful to somehow compress the statistical wealth of large neuronal recordings into few quantities, like sample moments for example. This compression would also entail interesting physico-biological properties of neuronal activity. The standard maximum-entropy method [1–3] accomplishes this kind of compression: it associates a unique probability distribution with few experimental quantities [***]. But this is only one of its uses. It is also used for various information-theoretic purposes or to generate reference probability distributions [***]. In all these uses the maximum-entropy distribution is chosen as the "maximally noncommittal" one [4]. This adjective means little without further technical characterizations. Different works give different characterizations, but in this paper we will use the quoted expression as an umbrella term for all of them. Our results will not depend on the specific characterization of "noncommittal".

In this note we want to show that we have to face a dilemma if we want to apply the maximum-entropy method to a representative random sample to find a maximally noncommittal distribution.

The dilemma is this. We can apply the maximum-entropy method to the sample, using a specified set of experimental constraints, and generate a probability distribution for its state. But our sample is representative of a larger population. We can apply the maximum-entropy method to the larger population, using the same constraints, and generate a probability distribution for its larger state, and then find the distribution for the sample by marginalization. Either application seems to have some commendable features. However, *the distributions obtained by these two applications differ*. It goes without saying that if one is "maximally noncommittal", the other must be somehow "committal". Which choice is most meaningful?

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

In the rest of the paper we mathematically formulate this dilemma. To this purpose we 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, so they may be of interest on their own.

Our discussion pertains to a neurons modelled as binary units 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.

Our notation follows ISO and ANSI standards [5–7] but for the use of the comma "," to denote logical conjunction. Probability notation follows Jaynes [8]. We intend "probability" as the degree of belief that "would be agreed by all rational men if there were any rational men" [9].

2 Setup

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

$$\overline{\boldsymbol{X}} \coloneqq \tfrac{1}{N}(X_1 + X_2 + \cdots + X_N), \tag{1}$$

$$\overline{\boldsymbol{X}\boldsymbol{X}} \coloneqq \tbinom{N}{2}^{-1}(X_1X_2 + X_1X_3 + \cdots + X_{N-1}X_N), \tag{2}$$

$$\overline{\boldsymbol{X}\boldsymbol{X}\boldsymbol{X}} \coloneqq \tbinom{N}{3}^{-1}(X_1X_2X_3 + \cdots + X_{N-2}X_{N-1}X_N), \tag{3}$$

and so on These formulae say that $\overline{\boldsymbol{X}}$ is the fraction of active neurons, $\overline{\boldsymbol{X}\boldsymbol{X}}$ the fraction of simultaneously active pairs out of all $\binom{N}{2}$ pairs, $\overline{\boldsymbol{X}\boldsymbol{X}\boldsymbol{X}}$ the fraction of simultaneously active triplets, and so on. Products of states like $X_i \cdots X_j$ also have values in $\{0,1\}$; from this it can combinatorially be proved that

$$\underbrace{\overline{\boldsymbol{X}\cdots\boldsymbol{X}}}_{m \text{ factors}} = \binom{N}{m}^{-1}\binom{N\overline{\boldsymbol{X}}}{m}. \tag{4}$$

Analogous formulae hold for quantities like $\boldsymbol{x}$, $\overline{\boldsymbol{R}}$, $\bar{\boldsymbol{r}}$.

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

$$\mathrm{P}(X_1 = R_1, X_2 = R_2, \dots, X_N = R_N \mid K) \quad \text{or} \quad \mathrm{P}(\boldsymbol{X} = \boldsymbol{R} \mid K), \quad \boldsymbol{R} \in \{0,1\}^N, \tag{5}$$

where K denotes our state of knowledge, i.e. the evidence and assumptions backing this particular probability assignment. Our uncertainty about the state of the sample is likewise expressed by

$$\mathrm{P}(x_1 = r_1, x_2 = r_2, \dots, x_n = r_n \mid K) \quad \text{or} \quad \mathrm{P}(\boldsymbol{x} = \boldsymbol{r} \mid K), \quad \boldsymbol{r} \in \{0,1\}^n. \tag{6}$$

3 Initial assumptions: the probability of random representative samples

We need to make an initial probability assignment before any experimental observations are made, 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, their locations for example. Our initial state of knowledge I is therefore symmetric, or "exchangeable", under their permutations. This symmetry must be reflected in our initial probability: the *representation theorem for finite exchangeability* states that it must obey

$$\mathrm{P}(\boldsymbol{X} = \boldsymbol{R} \mid I) = \binom{N}{N\overline{\boldsymbol{R}}}^{-1} \mathrm{P}(\overline{\boldsymbol{X}} = \overline{\boldsymbol{R}} \mid I), \tag{7}$$

the latter being the probability for the population average $\boldsymbol{X}$. Proof and generalizations to non-binary and continuum cases are given by de Finetti [10], Ericson [11], Diaconis [12], Heath & Sudderth [13]. This theorem is intuitive: owing to symmetry, all states with $N\overline{\boldsymbol{R}}$ active neurons must have equal probabilities.

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

$$\mathrm{P}(\boldsymbol{x} = \boldsymbol{r} \mid I) = \binom{n}{n\bar{\boldsymbol{r}}}^{-1} \mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}} \mid I), \tag{8}$$

with

$$\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}} \mid I) = \sum_{N\overline{\boldsymbol{R}}=0}^{N} \mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}} \mid \overline{\boldsymbol{X}} = \overline{\boldsymbol{R}}, I)\, \mathrm{P}(\overline{\boldsymbol{X}} = \overline{\boldsymbol{R}} \mid I), \tag{9}$$

$$\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}} \mid \overline{\boldsymbol{X}} = \overline{\boldsymbol{R}}, I) = \binom{n}{n\bar{\boldsymbol{r}}}\binom{N-n}{N\overline{\boldsymbol{R}} - n\bar{\boldsymbol{r}}}\binom{N}{N\overline{\boldsymbol{R}}}^{-1} =: \mathit{\Pi}(\bar{\boldsymbol{r}} \mid \overline{\boldsymbol{R}}). \tag{10}$$

The latter conditional probability is a hypergeometric distribution $\mathit{\Pi}(\bar{\boldsymbol{r}} \mid \overline{\boldsymbol{R}})$, typical of "drawing without replacement" problems. The combinatorial proof of the formulae above is in fact the same as for this class of problems [8 ch. 3; 14 § 4.8.3; 15 § II.6].

The conditional probability $\mathit{\Pi}(\bar{\boldsymbol{r}}, \overline{\boldsymbol{R}})$ connects the spaces of the sample average $\overline{\boldsymbol{X}} \in \{0, \dots, N\}$ and of the population average $\bar{\boldsymbol{x}} \in \{0, \dots, n\}$. It is a coarsening projector of any probability p for $\overline{\boldsymbol{X}}$ onto a marginal probability p_* for $\bar{\boldsymbol{x}}$:

$$p_*(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}}) = \sum_{N\overline{\boldsymbol{R}}=0}^{N} \mathit{\Pi}(\bar{\boldsymbol{r}} \mid \overline{\boldsymbol{R}}) p(\overline{\boldsymbol{X}} = \overline{\boldsymbol{R}}). \tag{11}$$

Therefore it also a pulls back expectations of functions f of the sample average $\bar{\boldsymbol{x}}$ to expectations of functions f^* the population average $\overline{\boldsymbol{X}}$:

$$\begin{gathered} f^*(\overline{\boldsymbol{X}}) \coloneqq \sum_{n\bar{\boldsymbol{r}}=0}^{n} f(\bar{\boldsymbol{r}}) \mathit{\Pi}(\bar{\boldsymbol{r}} \mid \overline{\boldsymbol{X}}), \\ \mathrm{E}[f(\bar{\boldsymbol{x}})] = \mathrm{E}[f^*(\overline{\boldsymbol{X}})] = \sum_{n\bar{\boldsymbol{r}}=0}^{n} f(\bar{\boldsymbol{r}})\, \mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}} \mid I) = \sum_{N\overline{\boldsymbol{R}}=0}^{N} f^*(\overline{\boldsymbol{R}})\, \mathrm{P}(\overline{\boldsymbol{X}} = \overline{\boldsymbol{R}} \mid I). \end{gathered} \tag{12}$$

A look at a plot of the hypergeometric distribution $\mathit{\Pi}(\bar{\boldsymbol{r}} \mid \overline{\boldsymbol{R}})$, for example fig. 1, reveals that it is a sort of "fuzzy identity matrix" between the $\overline{\boldsymbol{X}}$-space $\{0, \dots, N\}$ and $\bar{\boldsymbol{x}}$-space $\{0, \dots, n\}$. When $n = N$ it is the identity matrix. We thus have that

$$\mathrm{P}(\bar{\boldsymbol{x}} = a) \approx \mathrm{P}(\overline{\boldsymbol{X}} = a), \qquad \mathrm{E}[f(\bar{\boldsymbol{x}})] \approx \mathrm{E}[f(\overline{\boldsymbol{X}})]. \tag{13}$$

These are only very approximate equalities; they may miss important features of the two probability distributions. If the distribution for the population average $\overline{\boldsymbol{X}}$ is bimodal, for example, the bimodality can be lost in the distribution for the sample average $\bar{\boldsymbol{x}}$, owing to the coarsening effect of $\mathit{\Pi}(\bar{\boldsymbol{r}} \mid \overline{\boldsymbol{R}})$.

The approximate equalities above express the fact that *our uncertainty about the sample is representative of our uncertainty about the population and about other samples*, and vice versa. Upon observation of a sample average, say $\bar{\boldsymbol{x}}\bar{\boldsymbol{x}} = a$, 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.

Note that formulae (13) say more than the limits $\mathrm{P}(\bar{\boldsymbol{x}} = a) \to \mathrm{P}(\overline{\boldsymbol{X}} = a)$, $\mathrm{E}[f(\bar{\boldsymbol{x}})] \to \mathrm{E}[f(\overline{\boldsymbol{X}})]$ as $n \to N$. These limits are trivially valid because the sample becomes the full population as $n \to N$. In particular, these limits hold even in cases where the conditional probability $\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}} \mid \overline{\boldsymbol{X}} = \overline{\boldsymbol{R}})$ is not a fuzzy identity and our uncertainties about sample and about population can differ wildly.

The symmetry present in the probabilities above is not a physical property of the neuronal population. It only expresses the symmetry of our initial uncertainty about the population, and does not imply

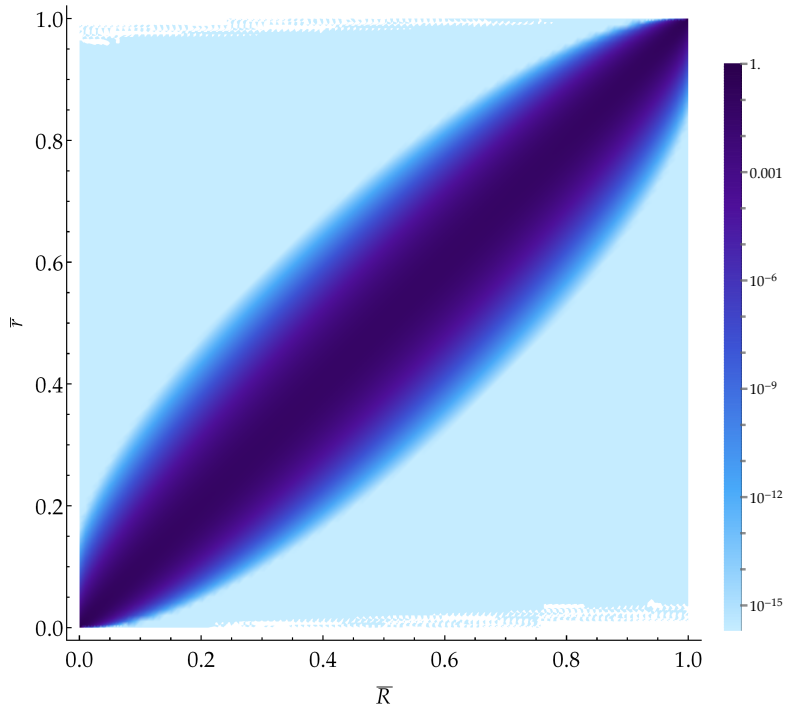

Figure 1: Log-plot of the hypergeometric distribution $\binom{n}{n\bar{\boldsymbol{r}}}\binom{N-n}{N\bar{\boldsymbol{R}}-n\bar{\boldsymbol{r}}}\binom{N}{N\bar{\boldsymbol{R}}}^{-1} =: \Pi(\bar{\boldsymbol{r}}|\,\bar{\boldsymbol{R}})$ for $N = 5000$, $n = 200$. (Band artifacts may appear in the colourbar depending on your PDF viewer.)

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, and this is indeed the case: the probability for the state of the sample (9) automatically satisfies the representation theorem (7) as well.

For functions representing averaged products, $f(\bar{\boldsymbol{x}}) := \overline{\boldsymbol{x}\dots\boldsymbol{x}} = \binom{\bar{\boldsymbol{x}}}{m}/\binom{n}{m}$, formulae (12) have the useful form

$$\Bigl(\underbrace{\overline{\boldsymbol{x}\cdots\boldsymbol{x}}}_{m\text{ factors}}\Bigr)^* = \underbrace{\overline{\boldsymbol{X}\cdots\boldsymbol{X}}}_{m\text{ factors}},$$

$$\mathrm{E}(\overline{\boldsymbol{x}\cdots\boldsymbol{x}}|\,I) = \mathrm{E}(\overline{\boldsymbol{X}\cdots\boldsymbol{X}}|\,I) = \binom{n}{m}^{-1}\sum_{n\bar{\boldsymbol{r}}=0}^{n}\binom{n\bar{\boldsymbol{r}}}{m}\mathrm{P}(\bar{\boldsymbol{x}}=\bar{\boldsymbol{r}}|\,I) = \binom{N}{m}^{-1}\sum_{N\bar{\boldsymbol{R}}=0}^{N}\binom{N\bar{\boldsymbol{R}}}{m}\mathrm{P}(\bar{\boldsymbol{X}}=\bar{\boldsymbol{R}}|\,I). \tag{14}$$

The proof uses the expression for the mth factorial moment of the hypergeometric distribution [16]. Thus, the averages of activity products *are the same for the sample and for the full population*. Similar relations can be found for the raw moments $\mathrm{E}(\bar{\boldsymbol{x}}^m)$ and $\mathrm{E}(\bar{\boldsymbol{X}}^m)$, which can be written in terms of the product expectations via eq. (4).

4 Enter maximum-entropy: quandary

Formulae (7)–(10) are constraints on our initial probability assignment, but do not determine it numerically. The probability $\mathrm{P}(\bar{\boldsymbol{X}} = \bar{\boldsymbol{R}}|\,I)$ for the population average needs to be numerically specified, and by (9) it will determine that of the sample average, $\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}}|\,I)$. If we numerically specify the latter, the former is not completely specified, because eq. (9) linearly constrains $N + 1$ unknowns by only $n + 1$ equations in this case.

We may want to specify the probability by enforcing the sample expectations of several functions to have specific values, for example $\mathrm{E}(\bar{\boldsymbol{x}}) = c_1$, $\mathrm{E}(\overline{\boldsymbol{x}\boldsymbol{x}}) = c_2$. This is still an underdetermined linear problem: several distributions can have the same desired expectations, as clear from eqs (14).

The maximum-entropy method is brought into play to solve this indeterminacy. It selects one distribution, purported to be "maximally noncommittal", among those that have the desired expectations. But we are in a quandary: formulae (12) allow us to apply the method to find the probability of the population $\mathrm{P}(\overline{\boldsymbol{X}} = \overline{\boldsymbol{R}} \mid I)$, or of the sample $\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}} \mid I)$. *The two applications, however, are inequivalent*. They lead to numerically different $\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}} \mid I)$.

Suppose we want to constrain the sample expectations of a vector function $\boldsymbol{f} = (f_1, \dots, f_m)$ to the vector values $\boldsymbol{c} = (c_1, \dots, c_m)$, that is, $\mathrm{E}[\boldsymbol{f}(\bar{\boldsymbol{x}})] = \boldsymbol{c}$. Application of maximum-entropy at the population level, denoted by I', gives

$$\mathrm{P}(\overline{\boldsymbol{X}} = \overline{\boldsymbol{R}} \mid I') = K \binom{N}{N\overline{\boldsymbol{R}}} \exp\Big[\boldsymbol{\Lambda}^\intercal \sum_{n\bar{\boldsymbol{r}}=0}^{n} \boldsymbol{f}(\bar{\boldsymbol{r}}) \Pi(\bar{\boldsymbol{r}} \mid \overline{\boldsymbol{R}})\Big], \tag{15}$$

and then by marginalization with eq. (8)

$$\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}} \mid I') = K \sum_{N\overline{\boldsymbol{R}}=0}^{N} \Pi(\bar{\boldsymbol{r}} \mid \overline{\boldsymbol{R}}) \binom{N}{N\overline{\boldsymbol{R}}} \exp\Big[\boldsymbol{\Lambda}^\intercal \sum_{n\bar{\boldsymbol{r}}=0}^{n} \boldsymbol{f}(\bar{\boldsymbol{r}}) \Pi(\bar{\boldsymbol{r}} \mid \overline{\boldsymbol{R}})\Big], \tag{16}$$

where K is a normalization constant and $\boldsymbol{\Lambda}^\intercal = (\Lambda_1, \dots, \Lambda_m)^\intercal$ are Lagrange multipliers such that

$$\boldsymbol{c} = K \sum_{n\bar{\boldsymbol{r}}=0}^{n} \sum_{N\overline{\boldsymbol{R}}=0}^{N} \boldsymbol{f}(\bar{\boldsymbol{r}}) \Pi(\bar{\boldsymbol{r}} \mid \overline{\boldsymbol{R}}) \binom{N}{N\overline{\boldsymbol{R}}} \exp\Big[\boldsymbol{\Lambda}^\intercal \sum_{n\bar{\boldsymbol{r}}=0}^{n} \boldsymbol{f}(\bar{\boldsymbol{r}}) \Pi(\bar{\boldsymbol{r}} \mid \overline{\boldsymbol{R}})\Big]. \tag{17}$$

Application of maximum-entropy at the sample level, denoted by I'', gives

$$\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}} \mid I'') = \kappa \binom{n}{n\bar{\boldsymbol{r}}} \exp[\boldsymbol{\lambda}^\intercal \boldsymbol{f}(\bar{\boldsymbol{r}})] \tag{18}$$

where κ is a normalization constant and $\boldsymbol{\lambda}^\intercal$ are Lagrange multipliers such that

$$\boldsymbol{c} = \kappa \sum_{n\bar{\boldsymbol{r}}=0}^{n} \boldsymbol{f}(\bar{\boldsymbol{r}}) \binom{n}{n\bar{\boldsymbol{r}}} \exp[\boldsymbol{\lambda}^\intercal \boldsymbol{f}(\bar{\boldsymbol{r}})]. \tag{19}$$

The probabilities for the sample average obtained from application at the population level (16) and at the sample level (18) will be approximately equal, by our previous observation (13) and also by the fact that they must satisfy the same expectations for $\boldsymbol{f}$. But the latter may miss important features present in the former, like additional modes or particular tail behaviour.

We show two examples of this discrepancy in figs 2 and 3, for $N = 5000$, $n = 200$. In the first example the constraints are $\mathrm{E}(\bar{\boldsymbol{x}}) = c_1$ and $\mathrm{E}(\overline{\boldsymbol{x}\boldsymbol{x}}) = c_2$, with $c_1 = 0.0478$ and $c_2 = 0.00257$. The distribution obtained at the sample level is broader than the one obtained at the population level; the tails of the two distributions are very different. The second example includes two additional constraints $\mathrm{E}(\overline{\boldsymbol{x}\boldsymbol{x}\boldsymbol{x}}) = c_3$, $\mathrm{E}(\overline{\boldsymbol{x}\boldsymbol{x}\boldsymbol{x}\boldsymbol{x}}) = c_4$ with $c_3 = 0.000148$, $c_4 = 8.81 \times 10^{-6}$. The distribution obtained at the population level has two modes, replaced by only one in the distribution obtained at the sample level; the tails are very different also in this case. The constraints used in these examples have neurobiologically realistic values [17].

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 $\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}} \mid I') = \mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}} \mid I'')$ in this case requires solving n equations in $m' + m''$ unknowns (normalization is taken care of), so in general it can only be satisfied if $n \leqslant m' + m''$.

Where should we apply the maximum-entropy method then, on the sample or on the population?

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.

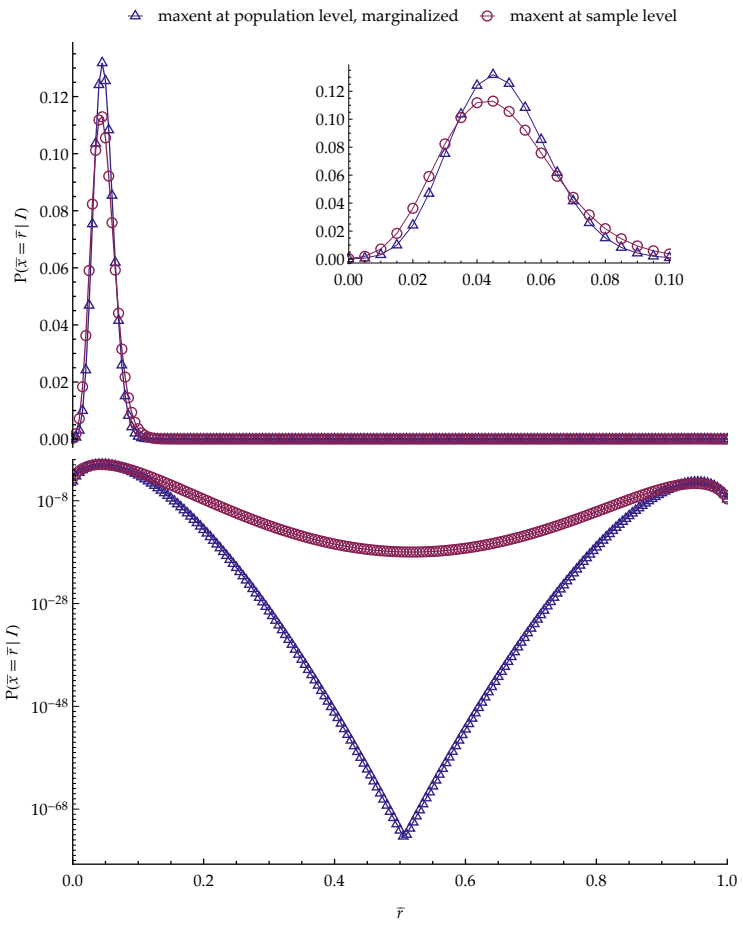

Figure 2: Plots of $\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}})$ constructed by maximum-entropy at the population level followed by sample marginalization (blue triangles), eq. (56), and at the sample level (red circles), eq. (57). In this case $N = 5000$, $n = 200$, $\mathrm{E}(\bar{\boldsymbol{x}}) = 0.0478$, $\mathrm{E}(\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}) = 0.00257$.

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 $\boldsymbol{x}$ 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. It is a very natural assumption in the neuroscientific context and therefore we find it preferable. In fact, also physical neuronal network model add an external input to the neurons, mimicking their embedding in a larger network

The possibility of 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 [18]. 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 [19 § 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.

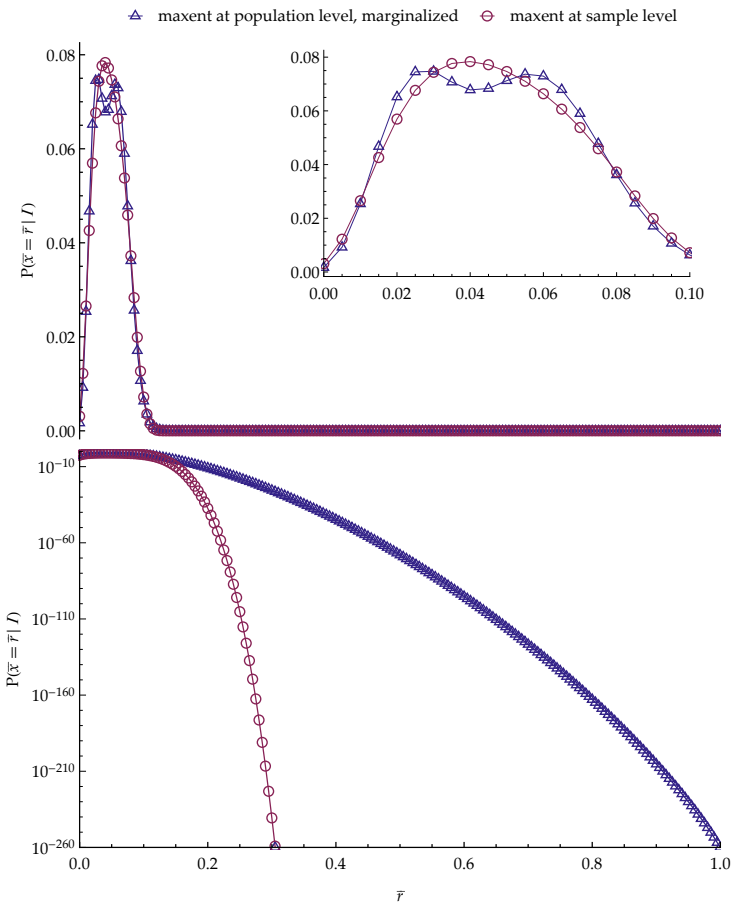

Figure 3: Linear and log-plots of $P(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}})$ constructed by maximum-entropy at the population level followed by sample marginalization (blue triangles), eq. (56), and at the sample level (red circles), eq. (57). In this case $N = 5000$, $n = 200$, $E(\bar{\boldsymbol{x}}) = 0.0478$, $E(\overline{\boldsymbol{xx}}) = 0.00257$, $E(\overline{\boldsymbol{xxx}}) = 1.48 \times 10^{-4}$, $E(\overline{\boldsymbol{xxxx}}) = 8.81 \times 10^{-6}$.

Discussion of case when individual $X_i \cdots X_j$ are constrained. Likely breakdown of maximum-entropy in this case [20]. Hints at full Bayesian treatment of which maximum-entropy is a limit? [21 ***]

Acknowledgments

To be added after review.

References

[1] E. T. Jaynes: *Probability Theory: With Applications in Science and Engineering: A Series of Informal Lectures*. (1954–1974). http://bayes.wustl.edu/etj/science.pdf.html; lecture notes written 1954–1974; earlier version of ref. [**jaynes1994_r1996**].

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

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

[4] E. T. Jaynes: *Information theory and statistical mechanics*. In: Ford [22] (1963), 181–218. Repr. in ref. [23 ch. 4, pp. 39–76]; http://bayes.wustl.edu/etj/node1.html.

[5] *Quantities and units*, 3rd ed. International Organization for Standardization. Geneva (1993).

[6] *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).

[7] *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.

[8] 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.

[9] I. J. Good: *How to estimate probabilities*. J. Inst. Maths. Applics **2**[4] (1966), 364–383.

[10] B. de Finetti: *La probabilitá e la statistica nei rapporti con l'induzione, secondo i diversi punti di vista*. In: de Finetti [24] (1959), 1–115. Transl. as ref. [25].

[11] 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.

[12] 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.

[13] D. Heath, W. Sudderth: *De Finetti's theorem on exchangeable variables*. American Statistician **30**[4] (1976), 188–189.

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

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

[16] R. B. Potts: *Note on the factorial moments of standard distributions*. Aust. J. Phys. **6**[4] (1953), 498–499.

[17] V. Rostami, P. G. L. Porta Mana, M. Helias: *Pairwise maximum-entropy models and their Glauber dynamics: bimodality, bistability, non-ergodicity problems, and their elimination via inhibition*. (2016). arXiv:1605.04740. Accepted for publ. in PloS CB.

[18] 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.

[19] E. T. Jaynes: *Inferential scattering*. (1993). http://bayes.wustl.edu/etj/node1.html; extensively rewritten version of a paper first publ. 1985 in ref. [26], pp. 377–398.

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

[21] D. J. C. MacKay: *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge (2003). http://www.inference.phy.cam.ac.uk/mackay/itila/; first publ. 1995.

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

[23] 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.

[24] B. de Finetti, ed.: *Induzione e statistica*, reprint. Springer, Berlin (2011). First publ. 1959.

[25] B. de Finetti: *Probability, statistics and induction: their relationship according to the various points of view*. In: de Finetti [27] (1972), p. 9, 147–227. Transl. of ref. [10].

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

[27] B. de Finetti: *Probability, Induction and Statistics: The art of guessing*. Wiley, London (1972).

*** 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 $\boldsymbol{y}$ of the rest of the population is conditionally dependent on the sample's values:

$$\mathrm{P}(\boldsymbol{y} = \boldsymbol{s} \mid \boldsymbol{x} = \boldsymbol{r}, I) \neq \mathrm{P}(\boldsymbol{y} = \boldsymbol{s} \mid I) \qquad \text{for some } \boldsymbol{s}, \boldsymbol{r}, \tag{20}$$

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, $\boldsymbol{X}$ 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: $(x_1, \dots, x_n) \equiv \boldsymbol{x}$ and $(r_1, \dots, r_n) \equiv \boldsymbol{r}$; 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*:

$$\overline{\boldsymbol{X}} \coloneqq (X_1 + \cdots + X_N)/N, \tag{21}$$

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

$$\bar{\boldsymbol{x}} \coloneqq (x_1 + \ldots + x_n)/n. \tag{22}$$

The quantities $N\overline{\boldsymbol{X}}$ and $n\bar{\boldsymbol{x}}$ represent the total number of active units in the population and the subpopulation. Quantities like $\overline{\boldsymbol{R}}$ and $\bar{\boldsymbol{r}}$ 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.,

$$\overline{\boldsymbol{X}\boldsymbol{X}} \coloneqq \tbinom{N}{2}^{-1}(X_1X_2 + X_1X_3 + \cdots + X_{N-1}X_N), \tag{23}$$

$$\overline{\boldsymbol{x}\boldsymbol{x}\boldsymbol{x}} \coloneqq \tbinom{n}{3}^{-1}(x_1x_2x_3 + x_1x_2x_4 + \cdots + x_{n-2}x_{n-1}x_n), \tag{24}$$

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

$$\mathrm{P}(X_1 = R_1, X_2 = R_2, \dots, X_N = R_N \mid I) \quad \text{or} \quad \mathrm{P}(\boldsymbol{X} = \boldsymbol{R} \mid I). \tag{25}$$

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 (25) under arbitrary permutations of the units's labels:

   $$\mathrm{P}(X_1 = R_1, X_2 = R_2, \dots, X_N = R_N \mid I) = \mathrm{P}(X_1 = R_{\pi(1)}, X_2 = R_{\pi(2)}, \dots, X_N = R_{\pi(N)} \mid I) \quad \text{for any permutation } \pi. \tag{26}$$

   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 $\overline{\boldsymbol{X}}$ has a particular distribution Q:

   $$\mathrm{P}(\overline{\boldsymbol{X}} = A \mid I) = Q(A), \qquad A \in \left\{0, \tfrac{1}{N}, \tfrac{2}{N}, \dots, 1\right\}. \tag{27}$$

   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:

$$\mathrm{P}(\boldsymbol{X} = \boldsymbol{R}|\, H) = \binom{N}{N\overline{\boldsymbol{R}}}^{-1} Q(\overline{\boldsymbol{R}}). \tag{28}$$

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

$$\mathrm{P}(\boldsymbol{x} = \boldsymbol{r}|\, I) = \sum_{NA=0}^{N} \binom{N-n}{NA-n\bar{\boldsymbol{r}}} \binom{N}{NA}^{-1} Q(A). \tag{29}$$

   Note that the only summands contributing to this sum are those for which $n\bar{\boldsymbol{r}} \leqslant NA \leqslant N$; 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:

$$\mathrm{P}(\boldsymbol{x} = \boldsymbol{r}|\, \boldsymbol{X} = \boldsymbol{R}, I) = \binom{N-n}{N\overline{\boldsymbol{R}} - n\bar{\boldsymbol{r}}}. \tag{30}$$

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

$$\mathrm{P}(\bar{\boldsymbol{x}} = a|\, I) = \binom{n}{na} \sum_{NA=0}^{N} \binom{N-n}{NA-na} \binom{N}{NA}^{-1} Q(A),$$
$$a \in \left\{0, \tfrac{1}{n}, \tfrac{2}{n}, \ldots, 1\right\}. \tag{31}$$

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

$$\mathrm{P}(\bar{\boldsymbol{x}} = a|\, \overline{\boldsymbol{X}} = A, I) = \binom{n}{na} \binom{N-n}{NA-na} \binom{N}{NA}^{-1}. \tag{32}$$

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

$$x_{i_1} x_{i_2} \cdots x_{i_m}, \qquad 1 \leqslant i_1 < i_2 < \cdots < i_m \leqslant n$$

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

$$\mathrm{E}(x_{i_1} \cdots x_{i_m}|\, I) = \mathrm{E}(\underbrace{\overline{\boldsymbol{x} \cdots \boldsymbol{x}}}_{m \text{ factors}}|\, I) = \mathrm{E}(\underbrace{\overline{\boldsymbol{X} \cdots \boldsymbol{X}}}_{m \text{ factors}}|\, I), \tag{33a}$$

   and has an explicit expression in terms of Q:

$$\begin{aligned} \mathrm{E}(x_{i_1} \cdots x_{i_m}|\, I) &= \binom{N}{m}^{-1} \sum_{NA=0}^{N} \binom{NA}{m} Q(A) \\ &\equiv \sum_{NA=0}^{N} \binom{N-m}{NA-m} \binom{N}{NA}^{-1} Q(A). \end{aligned} \tag{33b}$$

A useful relation connects the expectation of a product (33) and the mth *factorial moment* [11] of the probability distributions for the averages. The mth factorial moment of the subpopulation average $\bar{\boldsymbol{x}}$ is defined by

$$\mathrm{E}[\underbrace{n\bar{\boldsymbol{x}}\,(n\bar{\boldsymbol{x}} - 1) \cdots \big(n\bar{\boldsymbol{x}} - (m-1)\big)}_{m \text{ factors}}|\, I] \equiv \mathrm{E}\left[\frac{(n\bar{\boldsymbol{x}})!}{(n\bar{\boldsymbol{x}} - m)!}|\, I\right], \tag{34}$$

an analogous definition holding for $\overline{\boldsymbol{X}}$. We have that

$$\mathrm{E}(x_{i_1}\cdots x_{i_m}|\, I) = \tfrac{(n-m)!}{n!}\mathrm{E}\left[\frac{(n\bar{\boldsymbol{x}})!}{(n\bar{\boldsymbol{x}}-m)!}|\, I\right] = \tfrac{(N-m)!}{N!}\mathrm{E}\left[\frac{(N\overline{\boldsymbol{X}})!}{(N\overline{\boldsymbol{X}}-m)!}|\, I\right]. \tag{35}$$

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

$$\mathrm{E}(\bar{\boldsymbol{x}}|\, I) = \mathrm{E}(\overline{\boldsymbol{X}}|\, I), \tag{36a}$$

$$\begin{aligned}\mathrm{E}(\bar{\boldsymbol{x}}^2|\, I) = \mathrm{E}(\overline{\boldsymbol{X}}|\, I)\frac{N-n}{(N-1)n} + \\ \mathrm{E}(\overline{\boldsymbol{X}}^2|\, I)\frac{N\,(n-1)}{(N-1)n},\end{aligned} \tag{36b}$$

$$\begin{aligned}\mathrm{E}(\bar{\boldsymbol{x}}^3|\, I) = \mathrm{E}(\overline{\boldsymbol{X}}|\, I)\frac{(N-n)(N-2n)}{(N-1)(N-2)n^2} + \\ \mathrm{E}(\overline{\boldsymbol{X}}^2|\, I)\frac{3N\,(N-n)(n-1)}{(N-1)(N-2)n^2} + \\ \mathrm{E}(\overline{\boldsymbol{X}}^3|\, I)\frac{N^2\,(n-1)(n-2)}{(N-1)(N-2)n^2}.\end{aligned} \tag{36c}$$

Relations for higher moments can be obtained recursively from eq. (35). In general, this means that the two sets of first m moments are related by a homogeneous linear transformation,

$$\mathrm{E}(\bar{\boldsymbol{x}}^m|\, I) = \sum_{l=1}^{m} M_{ml}(n,N)\mathrm{E}(\overline{\boldsymbol{X}}^l|\, I), \tag{37}$$

with a universal, lower-triangular transformation matrix $M_{ml}(n,N)$ that depends only on n, N, and the condition of symmetry (26).

As intuition suggests, we have

$$\mathrm{E}(\bar{\boldsymbol{x}}^m|\, I) \xrightarrow{n\to N} \mathrm{E}(\overline{\boldsymbol{X}}^m|\, I), \qquad \mathrm{E}(\bar{\boldsymbol{x}}^m|\, I) \xrightarrow{n\to 1} \mathrm{E}(\overline{\boldsymbol{X}}|\, I), \tag{38}$$

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

The core of the six mathematical relations above are eqs (29) and (31). 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 $N, N\overline{\boldsymbol{X}}, n$, weighted by the probabilities $\mathrm{P}(\overline{\boldsymbol{X}} = A|\, I)$ [cf. 15 § 4, esp. eq. (22)]. The connection between this mixture representation and the condition of symmetry (26) 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 $\bar{\boldsymbol{x}}$, determined by eq. (31), changes with the subnetwork size n. Choose a network-average distribution $\mathrm{P}(\overline{\boldsymbol{X}} = A|\, I)$ belonging to the exponential family [21 § 4.5.3; see also 22]:

$$\mathrm{P}(\overline{\boldsymbol{X}} = A|\, I) = Q(A) \propto \binom{N}{NA}\exp[\lambda_2\, NA\,(NA-1)/2 + \lambda_1\, NA]. \tag{39}$$

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 $\boldsymbol{X}$.

The probability distribution of eq. (39) is plotted in fig. 4, together with the resulting subnetwork-average distributions $\mathrm{P}(\bar{\boldsymbol{x}} = a|\, I)$, 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

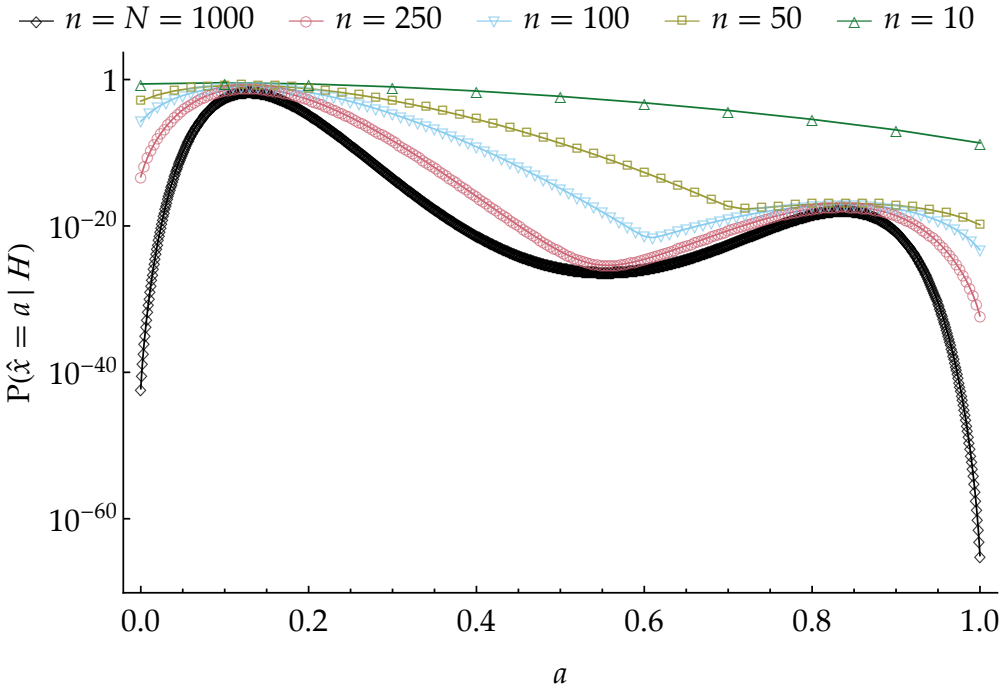

Figure 4: Probability distributions $\mathrm{P}(\bar{\boldsymbol{x}} = a|\, I)$ for different subnetwork sizes n, obtained from a network probability distribution $\mathrm{P}(\bar{\boldsymbol{X}} = A|\, I)$ having the maximum-etropy form (39).

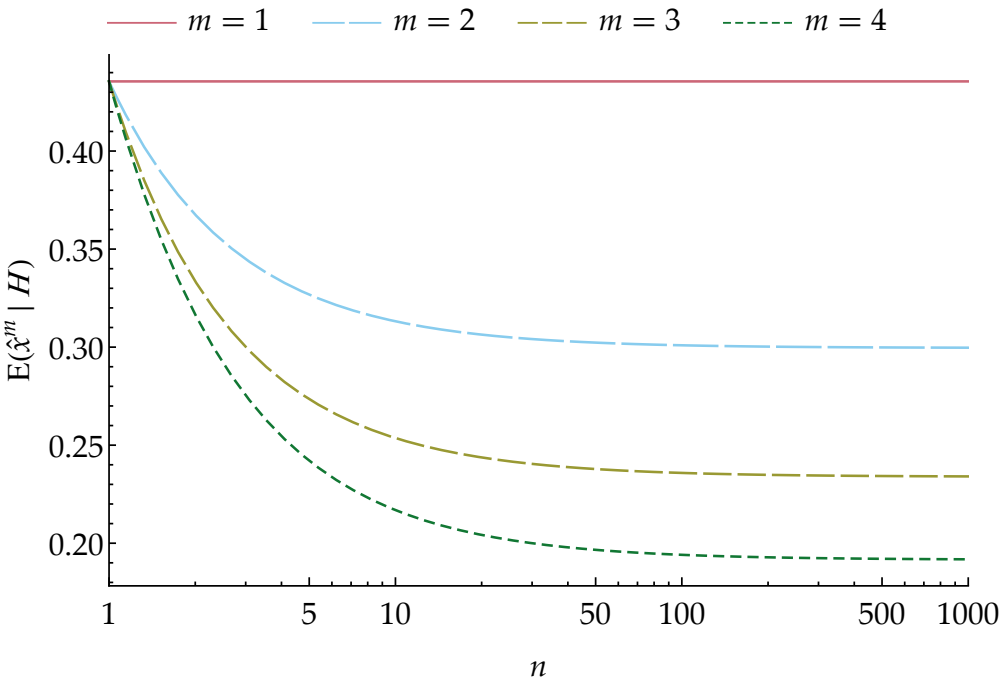

Figure 5: Moments of the probability distributions $\mathrm{P}(\bar{\boldsymbol{x}} = a|\, I)$ as functions of the subnetwork size n.

original distribution disappears; at the same time the finite-difference

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

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. 4 do not look like maximum-entropy ones with second-moment constraints. In fact, they are not and *cannot* be:

$$\mathrm{P}(\bar{\boldsymbol{x}} = a \mid I) \neq \kappa \binom{n}{na} \exp[\kappa_2\, na\,(na-1)/2 + \kappa_1\, na] \tag{40}$$

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 (31). This is generally impossible unless $m \geqslant n$. (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 (26), 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 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 $\mathrm{E}(\bar{\boldsymbol{x}}^m \mid I)$ as a function of size n is shown in fig. 5. The moments become practically constant when $n \approx 100$ or larger. The expectations of m-tuple products of states $\mathrm{E}(x_{i_1} \cdots x_{i_m} \mid I)$, 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 (26), the probability $\mathrm{P}(\overline{\boldsymbol{X}} = A \mid I)$ for the network average determines that of each subnetwork average, $\mathrm{P}(\bar{\boldsymbol{x}} \mid I)$, by the marginalization eq. (31). The reverse is trivially not true, since eq. (31), 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 $\mathrm{P}(\bar{\boldsymbol{x}} = a \mid I)$ to a subnetwork average $\bar{\boldsymbol{x}}$ does not determine a network distribution $\mathrm{P}(\overline{\boldsymbol{X}} = A \mid I)$: 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 $\mathrm{P}(\boldsymbol{X} = \boldsymbol{R} \mid I)$ 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 $\bar{\boldsymbol{x}}$ and $\overline{\boldsymbol{x}\boldsymbol{x}}$ of a particular subnetwork have given values

$$\mathrm{E}(\bar{\boldsymbol{x}} \mid I') = c_1, \qquad \mathrm{E}(\overline{x_i x_j} \mid I') = c_2; \tag{41}$$

2. the network probability distribution $\mathrm{P}(\boldsymbol{X} = \boldsymbol{R}|\, I')$ 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 (26) and is defined by

$$\mathrm{p}(\boldsymbol{X} = \boldsymbol{R}|\, I') = K \exp[\boldsymbol{\Lambda}_2 N\overline{\boldsymbol{R}}\,(N\overline{\boldsymbol{R}} - 1)/2 + \boldsymbol{\Lambda}_1 N\overline{\boldsymbol{R}}]$$

with $K, \boldsymbol{\Lambda}_m$, such that the distribution is normalized and

$$K \sum_{NA=0}^{N} \binom{N-m}{NA-m} \exp[\boldsymbol{\Lambda}_2 NA\,(NA-1)/2 + \boldsymbol{\Lambda}_1 NA] = c_m, \quad m = 1, 2. \tag{42}$$

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 (33) of subnetwork and network expectations, e.g.

$$c_2 = \mathrm{E}(\overline{\boldsymbol{x}\boldsymbol{x}}|\, I') = \binom{N}{2}^{-1} \sum_{NA=0}^{N} \binom{NA}{2} \mathrm{P}(\overline{\boldsymbol{X}} = A|\, I'), \tag{43}$$

and with relations (28), (30). This example is easily generalized to any number m of constraints such that $m \leqslant n$.

Note again that, as remarked in § 7.1, the subnetwork from which the averages in the expectations (41) are calculated has a probability distribution $\mathrm{P}(\bar{\boldsymbol{x}} = a|\, I)$ determined by the marginalization (31) 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 $\bar{\boldsymbol{x}}$ of a particular subnetwork has a probability distribution q:

$$\mathrm{P}(\bar{\boldsymbol{x}} = a|\, I'') = q(a); \tag{44}$$

2. the probability distribution for the network, $\mathrm{P}(\boldsymbol{X} = \boldsymbol{R}|\, I'')$, 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 (26):

$$\mathrm{P}(\boldsymbol{X} = \boldsymbol{R}|\, I'') = \exp\left[\sum_{na=0}^{n} \boldsymbol{\Lambda}_a \binom{n}{na}\binom{N-n}{N\overline{\boldsymbol{R}} - na}\right]$$

with $\boldsymbol{\Lambda}_a$ such that

$$\sum_{NA=0}^{N} \binom{n}{na}\binom{N-n}{NA-na} \exp\left[\sum_{na=0}^{n} \boldsymbol{\Lambda}_a \binom{n}{na}\binom{N-n}{NA-na}\right] = q(a) \tag{45}$$

(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. (29), 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 $\mathrm{P}(\bar{\boldsymbol{x}} = a|\, I'')$ 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 1000 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 1000th 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 $x_1 = 1$ 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 $\boldsymbol{x}$ active neurons and $\overline{\boldsymbol{x}\boldsymbol{x}}$ 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 $\bar{\boldsymbol{x}}$ active neurons and $\overline{\boldsymbol{x}\boldsymbol{x}}$ 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 (28) from that for the network average (27), consider that if the network total is $N\overline{\boldsymbol{X}}$, then $N\overline{\boldsymbol{X}}$ out of N units are active, and there are $\binom{N}{N\overline{\boldsymbol{X}}}$ possible states for which this can be true; therefore

$$\mathrm{P}(\boldsymbol{X} = \boldsymbol{R} \mid H) = \binom{N}{N\overline{\boldsymbol{R}}}^{-1} Q(\overline{\boldsymbol{R}}). \tag{7$_\mathrm{r}$}$$

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

$$\mathrm{P}(\boldsymbol{x} = \boldsymbol{r} \mid H) = \binom{n}{n\bar{\boldsymbol{r}}}^{-1} \mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}} \mid I), \tag{46}$$

for the subnetwork.

Let us next consider the probability $\mathrm{P}(\bar{\boldsymbol{x}} = a \mid \overline{\boldsymbol{X}} = A, I)$ for the subnetwork average $\bar{\boldsymbol{x}}$ conditional on the network average $\overline{\boldsymbol{X}}$. There are $\binom{N}{N\overline{\boldsymbol{X}}}$ possible network states if the network average is $\overline{\boldsymbol{X}}$, i.e. if $N\overline{\boldsymbol{X}}$ units are in state 1; the conditional probability of each is therefore $1/\binom{N}{N\overline{\boldsymbol{X}}}$, owing to the symmetry assumption (26). Now consider the subnetwork of the first n units. The conditional probability of having $n\bar{\boldsymbol{x}}$ specific ones in state 1 is the sum of the probabilities of all states for which $N\overline{\boldsymbol{X}} - n\bar{\boldsymbol{x}}$ of the remaining $N - n$ units are in state 1; there are $\binom{N-n}{N\overline{\boldsymbol{X}}-n\bar{\boldsymbol{x}}}$ such states, all equally probable. Finally, there are $\binom{n}{n\bar{\boldsymbol{x}}}$ possible ways, all equally probable, in which $n\bar{\boldsymbol{x}}$ of the first n units can be in state 1. In formulae,

$$\begin{aligned}
\mathrm{P}(\bar{\boldsymbol{x}} = a \mid \overline{\boldsymbol{X}} = A, I) &= \sum_{\boldsymbol{r}}^{\bar{\boldsymbol{r}}=a} \sum_{\boldsymbol{R}}^{\overline{\boldsymbol{R}}=A} \mathrm{P}(\boldsymbol{x} = \boldsymbol{r} \mid \boldsymbol{X} = \boldsymbol{R}, I)\, \mathrm{P}(\boldsymbol{X} = \boldsymbol{R} \mid \overline{\boldsymbol{X}} = A, I), \\
&= \sum_{\boldsymbol{r}}^{\bar{\boldsymbol{r}}=a} \sum_{\boldsymbol{R}}^{\overline{\boldsymbol{R}}=A} \mathrm{P}(\boldsymbol{x} = \boldsymbol{r} \mid \boldsymbol{X} = \boldsymbol{R}, I) \binom{N}{NA}^{-1}, \\
&= \sum_{\boldsymbol{r}}^{\bar{\boldsymbol{r}}=a} \binom{N-n}{NA-na} \binom{N}{NA}^{-1}, \\
&= \binom{n}{na} \binom{N-n}{NA-na} \binom{N}{NA}^{-1},
\end{aligned} \tag{47}$$

which is the conditional probability (32). 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 $\bar{\boldsymbol{x}}$ blue balls in n drawings without replacement from an urn with N balls, a fraction $\overline{\boldsymbol{X}}$ of which are blue.

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

$$\begin{aligned}\mathrm{P}(\bar{\boldsymbol{x}} = a \mid I) &= \sum_{NA=0}^{N} \mathrm{P}(\bar{\boldsymbol{x}} = a \mid \overline{\boldsymbol{X}} = A, I)\, \mathrm{P}(\overline{\boldsymbol{X}} = A \mid I), \\ &= \sum_{NA=0}^{N} \binom{n}{na}\binom{N-n}{NA-na}\binom{N}{NA}^{-1} Q(A).\end{aligned} \tag{48}$$

which proves the subnetwork-average formula (31). This formula, combined with eqs (46) and (28), leads to the conditional probability (30).

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

$$\begin{aligned}\sum_{\boldsymbol{x}} x_1 \cdots x_m \mathrm{P}(\boldsymbol{x} = \boldsymbol{r} \mid I) &= \sum_{\boldsymbol{r},\boldsymbol{R}} r_1 \cdots r_m \mathrm{P}(\boldsymbol{x} = \boldsymbol{r} \mid \boldsymbol{X} = \boldsymbol{R}, I)\, \mathrm{P}(\boldsymbol{X} = \boldsymbol{R} \mid I), \\ &= \sum_{\boldsymbol{r},\boldsymbol{R}} r_1 \cdots r_m\, \delta(R_1 - r_1) \cdots \delta(R_m - R_m) \mathrm{P}(\boldsymbol{X} = \boldsymbol{R} \mid I), \\ &= \sum_{\boldsymbol{R}} R_1 \cdots R_m \mathrm{P}(\boldsymbol{X} = \boldsymbol{R} \mid I).\end{aligned} \tag{49}$$

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 $n\bar{\boldsymbol{x}}$ units of the subnetwork which are in state 1, and there are $\binom{n\bar{\boldsymbol{x}}}{2}$ such choices, so the sum above is equal to $\binom{n\bar{\boldsymbol{x}}}{2}$. The sum has $\binom{n}{2}$ terms, so their average is $\binom{n\bar{\boldsymbol{x}}}{2}/\binom{n}{2}$. Generalizing the argument to products of m units, we have that

$$\overline{x_{i_1} \cdots x_{i_m}} = \binom{n\bar{\boldsymbol{x}}}{m}\binom{n}{m}^{-1}. \tag{50}$$

Then, using eq. (31),

$$\begin{aligned}&\mathrm{E}\big(\overline{x_{i_1} \cdots x_{i_m}} \mid I\big) \\ &\quad= \tfrac{(n-m)!}{n!}\mathrm{E}\Big(m!\,\tbinom{n\bar{\boldsymbol{x}}}{m} \mid I\Big) \\ &\quad\equiv \frac{(n-m)!}{n!} \sum_{na=0}^{n} m! \binom{na}{m} \mathrm{P}(\bar{\boldsymbol{x}} = a \mid I), \\ &\quad= \frac{(n-m)!}{n!} \sum_{NA=0}^{N} Q(A) \left[\sum_{na=0}^{n} m! \binom{na}{m}\binom{n}{na}\binom{N-n}{NA-na}\binom{N}{NA}^{-1}\right].\end{aligned} \tag{51}$$

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

$$\sum_{na=0}^{n} m! \binom{na}{m}\binom{n}{na}\binom{N-n}{N\overline{\boldsymbol{X}}-na}\binom{N}{N\overline{\boldsymbol{X}}}^{-1} = m! \binom{n}{m}\binom{na}{m}\binom{N}{m}, \tag{52}$$

which combined with the previous equation yields the second line of eq. (33); its last equality comes from the identity

$$\binom{N}{M}\binom{M}{m} = \binom{N}{m}\binom{N-m}{M-m}, \tag{53}$$

easily derived by writing the binomial coefficients in terms of factorials. Finally, eqs (36), relating the moments of the distributions for subnetwork and network averages, is obtained from the definition of moments,

$$\mathrm{E}\big(\bar{\boldsymbol{x}}^m \mid I\big) \coloneqq \sum_{na=0}^{n} a^m \mathrm{P}(\bar{\boldsymbol{x}} = a \mid H), \quad \mathrm{E}\big(\overline{\boldsymbol{X}}^m \mid I\big) \coloneqq \sum_{NA=0}^{N} A^m \mathrm{P}(\overline{\boldsymbol{X}} = A \mid H), \tag{54}$$

replaced in the equalities for the factorial moments (35), by recursively solving in terms of the moments of the network distribution.

****

Here is an example.

Suppose our desired constraints are $\mathrm{E}(\bar{\boldsymbol{x}}) = c_1$ and $\mathrm{E}(\overline{\boldsymbol{x}\boldsymbol{x}}) = c_2$, the expectations being given by the fourth term of eq. (14). Applying maximum-entropy to the distribution of the average of the full population we find

$$\mathrm{P}(\bar{\boldsymbol{X}} = \bar{\boldsymbol{R}} \mid I') = \boldsymbol{\Lambda} \binom{N}{N\bar{\boldsymbol{R}}} \exp\left[\boldsymbol{\Lambda}_2 \bar{\boldsymbol{R}} \, \frac{N\bar{\boldsymbol{R}} - 1}{N-1} + \boldsymbol{\Lambda}_1 \, \bar{\boldsymbol{R}}\right] \tag{55}$$

with such parameters $\boldsymbol{\Lambda}$, $\boldsymbol{\Lambda}_1$, $\boldsymbol{\Lambda}_2$ as to satisfy normalization and constraints. We call this state of knowledge I'. Marginalization by eq. (8) then gives

$$\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}} \mid I') = \boldsymbol{\Lambda} \sum_{N\bar{\boldsymbol{R}}=0}^{N} \binom{n}{n\bar{\boldsymbol{r}}} \binom{N-n}{N\bar{\boldsymbol{R}} - n\bar{\boldsymbol{r}}} \exp\left[\boldsymbol{\Lambda}_2 \bar{\boldsymbol{R}} \, \frac{N\bar{\boldsymbol{R}} - 1}{N-1} + \boldsymbol{\Lambda}_1 \, \bar{\boldsymbol{R}}\right]. \tag{56}$$

Applying maximum-entropy to the distribution of the average of the sample, using the third term of eq. (14), we find

$$\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}} \mid I'') = \boldsymbol{\lambda} \binom{n}{n\bar{\boldsymbol{r}}} \exp\left[\boldsymbol{\lambda}_2 \bar{\boldsymbol{r}} \, \frac{n\bar{\boldsymbol{r}} - 1}{n-1} + \boldsymbol{\lambda}_1 \, \bar{\boldsymbol{r}}\right] \tag{57}$$

with such parameters $\boldsymbol{\lambda}$, $\boldsymbol{\lambda}_1$, $\boldsymbol{\lambda}_2$ as to satisfy the constraints. We call this state of knowledge I''.

The resulting distributions $\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}} \mid I')$, $\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}} \mid I'')$ are different. Figure 2 shows one example, for $N = 5000$, $n = 100$, $c_1 = 0.45$, $c_2 = 0.35$. The two distributions have slightly different modes and the one obtained from maximum-entropy at the population level is more peaked. Another example is shown in fig. 3, for $N = 5000$, $n = 200$, and neurobiologically more realistic constraints $c_1 = 0.045$, $c_2 = 0.0025$ [33]. The discrepancy is maybe not as large as in the previous example, but the tails of the distributions are still very different.

****

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.

One of the uses of maximum-entropy methods

number. It is large from large population of neurons [[]]. The recorded neurons are usually picked out from a specific brain area according to an unknown process, and from their observation in relation to stimuli or behavioral task, we strive for understanding the mechanism behind neuronal coding LM still don't know what people mean by this, can we find an unambiguous term or rephrase?. That is, they are assumed to be a "representative random sample" of the neurons in that area. This assumption says that from the sample we can learn about the population. The relation of population activity with stimuli or behaviour is a separate problem, so "this is" doesn't really fit...

Huge amounts of parallel recordings from hundreds of neurons, available today, are accompanied by the statistical challenges to be employed for shedding light on the mechanism behind neuronal coding. The principle of maximum-entropy was proposed as a way to overcome these challenges by constructing a full description of the neuronal firing patterns. The maximum-entropy method employs only first and second order statistics of the representative random sample to construct a "maximally noncommittal" [2] probability distribution of the firing patterns [[]]. LM: "noncommittal" doesn't mean anything to me to be honest, even if Jaynes used that term. I would like to openly say this, backed up by references, as done some paragraphs below.
It seems they have used maximum-entropy for other purposes too, sometimes unclear ones. And with all kinds of constraints. So we shouldn't be too specific here in relation to purpose and constraints. I'd rather give references and the readers can form their own ideas about its use.

Using the pairwise maximum-entropy models many studies investigated the questions such as, which level of interaction ambiguous word (pairwise or higher-order) between neurons is suffiecient to

explain the observed experimental data [[]]? Or how collective behavior of neurons can be related to stimuli or behavioral tasks [[]]? add more **

Here we show that there is a quandary in applying maximum-entropy to a representative random sample of neurons to generate a maximally noncommital distribution for that sample.

The quandary 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.

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" [2] 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 noncommital" 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.