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
who knows

E. Torre
Chocolateland

Abstract

***

1 Implicit assumptions in the maximum-entropy method

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

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 noncommital" [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 in a way beyond our control, 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.

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.

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

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 Probability of random representative samples

2.1 Setup

Consider a population of N binary neurons with states $(X_1, \dots, X_N)$ having fixed but unknown binary values $(R_1, \dots, R_N)$, with R_i in $\{0, 1\}$, vectorially written $\boldsymbol{X} = \boldsymbol{R}$. 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*:

$$\overline{\boldsymbol{X}} \coloneqq (X_1 + \dots + X_N)/N, \tag{1}$$

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

$$\widehat{\boldsymbol{x}} \coloneqq (x_1 + \dots + x_n)/n. \tag{2}$$

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

$$\overline{\boldsymbol{X}\boldsymbol{X}} \coloneqq \tbinom{N}{2}^{-1}(X_1X_2 + X_1X_3 + \dots + X_{N-1}X_N), \tag{3}$$

$$\widehat{\boldsymbol{x}\boldsymbol{x}\boldsymbol{x}} \coloneqq \tbinom{n}{3}^{-1}(x_1x_2x_3 + x_1x_2x_4 + \dots + x_{n-2}x_{n-1}x_n), \tag{4}$$

and so on.

2.2 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

$$\mathrm{P}(X_1 = R_1, X_2 = R_2, \dots, X_N = R_N \mid H) \quad \text{or} \quad \mathrm{P}(\boldsymbol{X} = \boldsymbol{R} \mid H). \tag{5}$$

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

In the present discussion, H 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 (5) under arbitrary permutations of the units's labels:

   $$\mathrm{P}(X_1 = R_1, X_2 = R_2, \dots, X_N = R_N \mid H) = \mathrm{P}(X_1 = R_{\pi(1)}, X_2 = R_{\pi(2)}, \dots, X_N = R_{\pi(N)} \mid H) \quad \text{for any permutation } \pi. \tag{6}$$

   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:

   $$\mathrm{P}(\overline{\boldsymbol{X}} = A \mid H) = Q(A), \qquad A \in \left\{0, \tfrac{1}{N}, \tfrac{2}{N}, \dots, 1\right\}. \tag{7}$$

   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.: 6; 1; 7–14] used with data on the population.

2.3 Formulae

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

1. Probability for the population state:

$$\mathrm{P}(\boldsymbol{X} = \boldsymbol{R} \mid H) = \binom{N}{N\overline{\boldsymbol{R}}}^{-1} Q(\overline{\boldsymbol{R}}). \tag{8}$$

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

$$\mathrm{P}(\boldsymbol{x} = \boldsymbol{r} \mid H) = \sum_{NA=0}^{N} \binom{N-n}{NA-n\widehat{\boldsymbol{r}}} \binom{N}{NA}^{-1} Q(A). \tag{9}$$

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

$$\mathrm{P}(\boldsymbol{x} = \boldsymbol{r} \mid \boldsymbol{X} = \boldsymbol{R}, H) = \binom{N-n}{N\overline{\boldsymbol{R}} - n\widehat{\boldsymbol{r}}}. \tag{10}$$

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

$$\mathrm{P}(\widehat{\boldsymbol{x}} = a \mid H) = \binom{n}{na} \sum_{NA=0}^{N} \binom{N-n}{NA-na} \binom{N}{NA}^{-1} Q(A), \qquad a \in \left\{0, \tfrac{1}{n}, \tfrac{2}{n}, \dots, 1\right\}. \tag{11}$$

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

$$\mathrm{P}(\widehat{\boldsymbol{x}} = a \mid \overline{\boldsymbol{X}} = A, H) = \binom{n}{na} \binom{N-n}{NA-na} \binom{N}{NA}^{-1}. \tag{12}$$

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:

$$\mathrm{E}(x_{i_1} \cdots x_{i_m} \mid H) = \mathrm{E}(\underbrace{\widehat{\boldsymbol{x} \cdots \boldsymbol{x}}}_{m \text{ factors}} \mid H) = \mathrm{E}(\underbrace{\overline{\boldsymbol{X} \cdots \boldsymbol{X}}}_{m \text{ factors}} \mid H), \tag{13a}$$

   and has an explicit expression in terms of Q:

$$\begin{aligned} \mathrm{E}(x_{i_1} \cdots x_{i_m} \mid H) &= \binom{N}{m}^{-1} \sum_{NA=0}^{N} \binom{NA}{m} Q(A) \\ &\equiv \sum_{NA=0}^{N} \binom{N-m}{NA-m} \binom{N}{NA}^{-1} Q(A). \end{aligned} \tag{13b}$$

A useful relation connects the expectation of a product (13) and the mth *factorial moment* [15] of the probability distributions for the averages. The mth factorial moment of the subpopulation average $\widehat{\boldsymbol{x}}$ is defined by

$$\mathrm{E}[\underbrace{n\widehat{\boldsymbol{x}}\,(n\widehat{\boldsymbol{x}} - 1) \cdots \big(n\widehat{\boldsymbol{x}} - (m-1)\big)}_{m \text{ factors}} \mid H] \equiv \mathrm{E}\left[\frac{(n\widehat{\boldsymbol{x}})!}{(n\widehat{\boldsymbol{x}} - m)!} \mid H\right], \tag{14}$$

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

$$\mathrm{E}(x_{i_1}\cdots x_{i_m}|\,H) = \tfrac{(n-m)!}{n!}\mathrm{E}\left[\frac{(n\widehat{\boldsymbol{x}})!}{(n\widehat{\boldsymbol{x}}-m)!}|\,H\right] = \tfrac{(N-m)!}{N!}\mathrm{E}\left[\frac{(N\overline{\boldsymbol{X}})!}{(N\overline{\boldsymbol{X}}-m)!}|\,H\right]. \tag{15}$$

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

$$\mathrm{E}(\widehat{\boldsymbol{x}}|\,H) = \mathrm{E}(\overline{\boldsymbol{X}}|\,H), \tag{16a}$$

$$\begin{aligned}\mathrm{E}(\widehat{\boldsymbol{x}}^2|\,H) = \mathrm{E}(\overline{\boldsymbol{X}}|\,H)\frac{N-n}{(N-1)n} + \\ \mathrm{E}(\overline{\boldsymbol{X}}^2|\,H)\frac{N\,(n-1)}{(N-1)n},\end{aligned} \tag{16b}$$

$$\begin{aligned}\mathrm{E}(\widehat{\boldsymbol{x}}^3|\,H) = \mathrm{E}(\overline{\boldsymbol{X}}|\,H)\frac{(N-n)(N-2n)}{(N-1)(N-2)n^2} + \\ \mathrm{E}(\overline{\boldsymbol{X}}^2|\,H)\frac{3N\,(N-n)(n-1)}{(N-1)(N-2)n^2} + \\ \mathrm{E}(\overline{\boldsymbol{X}}^3|\,H)\frac{N^2\,(n-1)(n-2)}{(N-1)(N-2)n^2}.\end{aligned} \tag{16c}$$

Relations for higher moments can be obtained recursively from eq. (15). In general, this means that the two sets of first m moments are related by a homogeneous linear transformation,

$$\mathrm{E}(\widehat{\boldsymbol{x}}^m|\,H) = \sum_{l=1}^{m} M_{ml}(n,N)\mathrm{E}(\overline{\boldsymbol{X}}^l|\,H), \tag{17}$$

with a universal, lower-triangular transformation matrix $M_{ml}(n,N)$ that depends only on n, N, and the condition of symmetry (6).

As intuition suggests, we have

$$\mathrm{E}(\widehat{\boldsymbol{x}}^m|\,H) \xrightarrow{n\to N} \mathrm{E}(\overline{\boldsymbol{X}}^m|\,H), \qquad \mathrm{E}(\widehat{\boldsymbol{x}}^m|\,H) \xrightarrow{n\to 1} \mathrm{E}(\overline{\boldsymbol{X}}|\,H), \tag{18}$$

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

The core of the six mathematical relations above are eqs (9) and (11). The latter expresses the probability for the subpopulation average as a mixture of hypergeometric distributions [5 ch. 3; 16 § 4.8.3; 17 § II.6], with parameters $N, N\overline{\boldsymbol{X}}, n$, weighted by the probabilities $\mathrm{P}(\overline{\boldsymbol{X}} = A|\,H)$ [cf. 18 § 4, esp. eq. (22)]. The connection between this mixture representation and the condition of symmetry (6) is well-known in the Bayesian literature [18–23].

3 Examples of inferential use of the formulae

3.1 From network to subnetwork

Let us illustrate with an example how the probability distribution for the subnetwork average $\widehat{\boldsymbol{x}}$, determined by eq. (11), changes with the subnetwork size n. Choose a network-average distribution $\mathrm{P}(\overline{\boldsymbol{X}} = A|\,H)$ belonging to the exponential family [24 § 4.5.3; see also 25]:

$$\mathrm{P}(\overline{\boldsymbol{X}} = A|\,H) = Q(A) \propto \binom{N}{NA}\exp[\lambda_2\,NA\,(NA-1)/2 + \lambda_1\,NA]. \tag{19}$$

This is the form obtained from the principle of maximum relative entropy [e.g.: 6; 1; 7–14] 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. (19) is plotted in fig. 1, together with the resulting subnetwork-average distributions $\mathrm{P}(\widehat{\boldsymbol{x}} = a|\, H)$, 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

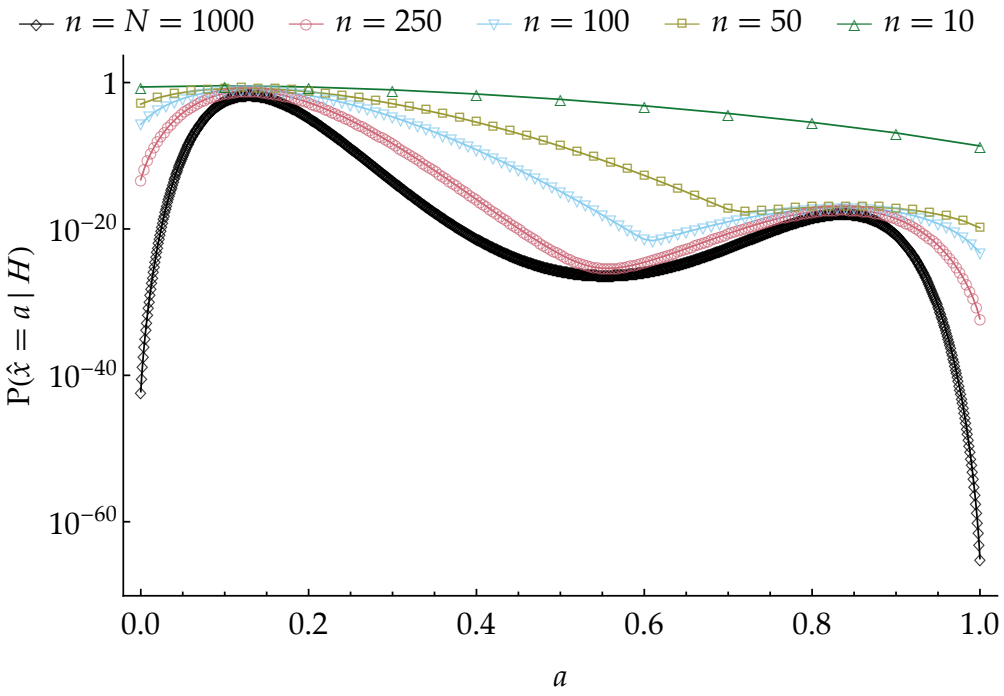

Figure 1: Probability distributions $\mathrm{P}(\widehat{\boldsymbol{x}} = a|\, H)$ for different subnetwork sizes n, obtained from a network probability distribution $\mathrm{P}(\overline{\boldsymbol{X}} = A|\, H)$ having the maximum-etropy form (19).

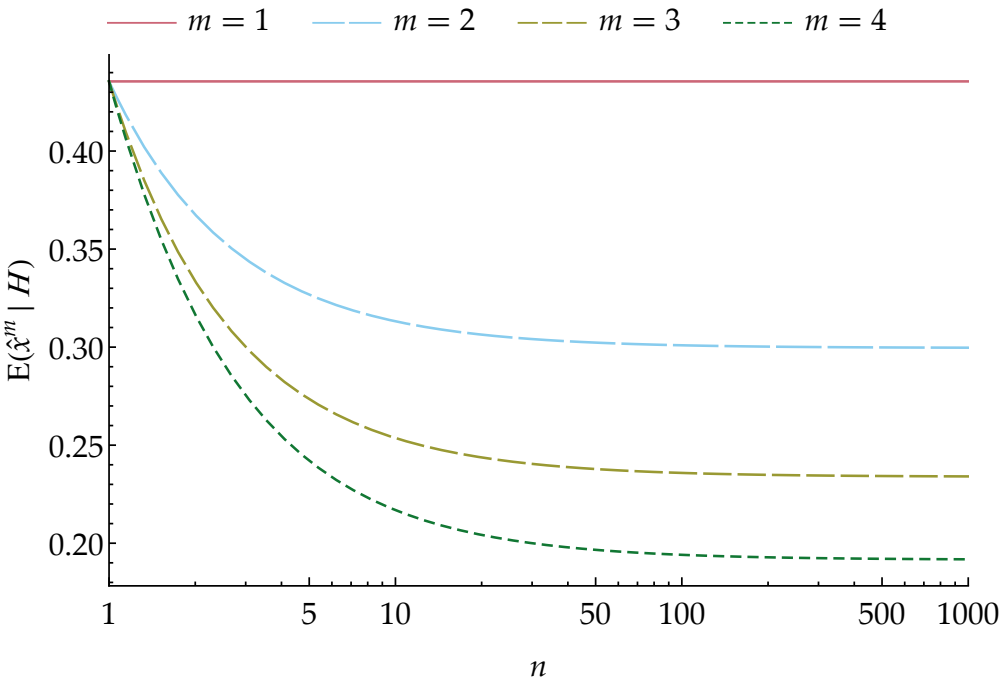

Figure 2: Moments of the probability distributions $\mathrm{P}(\widehat{\boldsymbol{x}} = a|\, H)$ as functions of the subnetwork size n.

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

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

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

$$\mathrm{P}(\widehat{\boldsymbol{x}} = a \mid H) \neq \kappa \binom{n}{na} \exp[\kappa_2\, na\,(na-1)/2 + \kappa_1\, na] \tag{20}$$

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 (11). This is generally impossible unless $m \geqslant n$. (Impossibilities of a similar kind appear in statistical mechanics, see e.g. ref. [26].)

This fact can be significant for recent works [e.g., 27–35] 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 (6), 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. [36] for a broader discussion of this fact and of its consequences.

The dependence of the first four moments $\mathrm{E}(\widehat{\boldsymbol{x}}^m \mid H)$ as a function of size n is shown in fig. 2. The moments become practically constant when $n \approx 100$ or larger. The expectations of m-tuple products of states $\mathrm{E}(x_{i_1} \cdots x_{i_m} \mid H)$, proportional to the factorial moments, are not shown as they do not depend on n.

3.2 From subnetwork to network

We have seen that, given the condition of symmetry (6), the probability $\mathrm{P}(\overline{\boldsymbol{X}} = A \mid H)$ for the network average determines that of each subnetwork average, $\mathrm{P}(\widehat{\boldsymbol{x}} \mid H)$, by the marginalization eq. (11). The reverse is trivially not true, since eq. (11), 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 $\mathrm{P}(\widehat{\boldsymbol{x}} = a \mid H)$ to a subnetwork average $\widehat{\boldsymbol{x}}$ does not determine a network distribution $\mathrm{P}(\overline{\boldsymbol{X}} = A \mid H)$: it only restricts the set of possible ones; this set can in principle be determined via linear-programming methods [37–41].

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.: 42; 43; 5; 24; 44–51]. 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 [52].

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

1. the expectations of the single and pair averages $\widehat{\boldsymbol{x}}$ and $\widehat{\boldsymbol{x}\boldsymbol{x}}$ of a particular subnetwork have given values

$$\mathrm{E}(\widehat{\boldsymbol{x}} \mid H') = c_1, \qquad \mathrm{E}(\widehat{x_i x_j} \mid H') = c_2; \tag{21}$$

2. the network probability distribution $\mathrm{P}(\boldsymbol{X} = \boldsymbol{R}|\, H')$ has maximum relative entropy with respect to the uniform one, given the constraints above.

Then the probability distribution for the network conditional on H' is completely determined: it satisfies the symmetry property (6) and is defined by

$$\mathrm{p}(\boldsymbol{X} = \boldsymbol{R}|\, H') = K \exp[\varLambda_2 N\overline{\boldsymbol{R}}\,(N\overline{\boldsymbol{R}} - 1)/2 + \varLambda_1 N\overline{\boldsymbol{R}}]$$

with $K, \varLambda_m$, such that the distribution is normalized and

$$K \sum_{NA=0}^{N} \binom{N-m}{NA-m} \exp[\varLambda_2 NA\,(NA-1)/2 + \varLambda_1 NA] = c_m, \quad m = 1, 2. \tag{22}$$

We omit the full proof of this statement: it is a standard application of the maximum-entropy procedure [e.g.: 6; 1; 7; 8; 10–14], combined with the equality (13) of subnetwork and network expectations, e.g.

$$c_2 = \mathrm{E}(\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}}|\, H') = \binom{N}{2}^{-1} \sum_{NA=0}^{N} \binom{NA}{2} \mathrm{P}(\overline{\boldsymbol{X}} = A|\, H'), \tag{23}$$

and with relations (8), (10). This example is easily generalized to any number m of constraints such that $m \leqslant n$.

Note again that, as remarked in § 3.1, the subnetwork from which the averages in the expectations (21) are calculated has a probability distribution $\mathrm{P}(\widehat{\boldsymbol{x}} = a|\, H)$ determined by the marginalization (11) and does *not* have a maximum-entropy form with the same number of constraints.

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

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

$$\mathrm{P}(\widehat{\boldsymbol{x}} = a|\, H'') = q(a); \tag{24}$$

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

Then the probability distribution for the network given H'' is completely determined and satisfies the symmetry property (6):

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

with $\varLambda_a$ such that

$$\sum_{NA=0}^{N} \binom{n}{na}\binom{N-n}{NA-na} \exp\left[\sum_{na=0}^{n} \varLambda_a \binom{n}{na}\binom{N-n}{NA-na}\right] = q(a) \tag{25}$$

(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. (9), 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 $\mathrm{P}(\widehat{\boldsymbol{x}} = a|\, H'')$ 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 $2^{1\,000}$ 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 $x_1 = 1$ 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 $\widehat{\boldsymbol{xx}}$ 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 $\widehat{\boldsymbol{x}}$ active neurons and $\widehat{\boldsymbol{xx}}$ 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.

3.3 ***

3.3.1 ***

***

Acknowledgments

Use unnumbered third level headings for the acknowledgments. All acknowledgments go at the end of the paper. Do not include acknowledgments in the anonymized submission, only in the final paper.

References