Hidden network-size assumptions in the maximum-entropy method

V. Rostami
Forschungszentrum Jülich INM-6
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 binary networks at one specific instant, or short window, in time. It is possible to generalize it to multi-state neuron models and network 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 reasons; sometimes not completely clear ones. For our purposes we can simply 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 not specifically chosen, 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.

V1 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.
V2 In this note we would like to show that there is a clash between the "representative random sample" spirit and the "maximally noncommital" spirit of maximum-entropy models applied to such a sample of neurons.

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 "random" and "representative" of some population, then we must choose which of them has a "maximally noncommittal" distribution: the sample or the population. We can't choose both. Which choice is most meaningful? It goes without saying that if either is "maximally noncommittal", the other must be somehow "committal".

***

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 network. It can informally be put this way:

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

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

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 subnetwork, the maximum-entropy method assumes that the uniform over the larger network is uniform, and therefore factorizable. The new distribution of the subnetwork will not be uniform, but that of the full network will still be factorizable into the one for the subnetwork 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 network. 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 network and the subnetwork. 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 network. 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 subnetwork *cannot give us any insights about the network 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 network with N units, 2^N possible states
2. expect averages of $\widehat{\boldsymbol{x}}$ active neurons and $\widehat{\boldsymbol{x}\boldsymbol{x}}$ 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 network with M units, 2^M possible states
2. expect that any subpopulation of N units has $\widehat{\boldsymbol{x}}$ active neurons and $\widehat{\boldsymbol{x}\boldsymbol{x}}$ 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.

1.1 ***

1.1.1 ***

**

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