# A Derivation of the probability distribution and of its approximations

## A.1 Notation

$N = 1000$: size of the full network.

$n = 65$: size of the sample.

$T = 417\,641$: number of bins in the recording.

Bins are indexed by $t \in \{1, \dots, T\}$.

$A \in \{0, 1, \dots, N\}$: levels of total activity of the full network.

$A_t \in \{0, 1, \dots, N\}$: total activity of the full network at bin $t$.

$\bar{A} \coloneqq (A_1, A_2, \dots, A_T)$: sequence of activities of full network during the recording.

$F_A \in \{0, 1/T, \dots, 1\}$: relative frequency of activity level $A$ of full network during the recording. That is, activity level $A$ appears in $TF_A$ out of $T$ bins: $TF_A = \sum_t \delta(A_t = A)$.

$\boldsymbol{F} \coloneqq (F_0, F_1, \dots, F_N)$: frequency distribution.

Note that $\bar{A}$ completely determines $\boldsymbol{F}$. We write the $\boldsymbol{F}$ determined by $\bar{A}$ as $\boldsymbol{F}(\bar{A})$.

$\nu_A \in [0, 1]$: long-run relative frequency of activity level $A$ of full network during all hypothetical repetitions of the same recording in the same conditions.

$a \in \{0, 1, \dots, n\}$: levels of total activity of the sample.

$a_t \in \{0, 1, \dots, n\}$: total activity of the sample at bin $t$.

$\bar{a} \coloneqq (a_1, a_2, \dots, a_T)$: sequence of activities of sample during the recording.

$f_a \in \{0, 1/T, \dots, 1\}$: relative frequency of activity level $a$ of sample during the recording.

$\boldsymbol{f} \coloneqq (f_0, f_1, \dots, f_n)$: frequency distribution.

Note that $\bar{a}$ completely determines $\boldsymbol{f}$. We write the $\boldsymbol{f}$ determined by $\bar{a}$ as $\boldsymbol{f}(\bar{a})$.

$J_{a\,A} \in \{0, 1/T, \dots, 1\}$: joint relative frequency of activity levels $a$ and $A$ for sample and full network during the recording. That is, the pair $(a, A)$ appears in $TJ_{a\,A}$ out of $T$ bins: $TJ_{a\,A} = \sum_t \delta(a_t = a)\,\delta(A_t = A)$.

Note that the full network can't have fewer active neurons or fewer silent neurons than its sample (for example, if $A_t = 0$ then $a_t = 0$, and if $A_t = N$ then $a_t = n$), so $J_{a\,A} \equiv 0$ for $a > A$ or $n - a > N - A$.

Obviously $F_A \equiv \sum_a J_{a\,A}$ and $f_a \equiv \sum_A J_{a\,A}$.

$\boldsymbol{J} \coloneqq (J_{0\,0}, J_{0\,1}, \dots, J_{n\,N})$: joint frequency distribution.

Note that $(\bar{a}, \overline{A})$ together completely determine $\boldsymbol{J}$. We write the $\boldsymbol{J}$ determined by $(\bar{a}, \overline{A})$ as $\boldsymbol{J}(\bar{a}, \overline{A})$.

$G_{a\,A} \coloneqq \binom{n}{a}\binom{N-n}{A-a}\binom{N}{A}^{-1} \equiv \binom{A}{a}\binom{N-A}{n-a}\binom{N}{n}^{-1}$: hypergeometric distribution.

$I$: background information and assumptions; includes knowledge of $N$ and $n$.

## A.2 Derivation

We want $\mathrm{p}(\boldsymbol{F} \mid \bar{a}, I)$. This can be obtained as the marginal of $\mathrm{p}(\boldsymbol{J} \mid \bar{a}, I)$. If this probability distribution is represented by a set of Monte Carlo samples $\{\boldsymbol{J}^{(i)}\}$, then $\{\boldsymbol{F}^{(i)} \equiv \sum_a J^{(i)}_{a\,A}\}$ are automatically samples of $\mathrm{p}(\boldsymbol{F} \mid \bar{a}, I)$.

By the theorem of total probability,

$$\mathrm{p}(\boldsymbol{J} \mid \bar{a}, I) = \sum_{\overline{A}} \mathrm{p}(\boldsymbol{J} \mid \bar{a}, \overline{A}, I)\, \mathrm{p}(\overline{A} \mid \bar{a}, I), \tag{35}$$

the sum being over all possible sequences $\{\overline{A}\}$ of total activities.

The first factor is a singular distribution:

$$\begin{aligned}\mathrm{p}(\boldsymbol{J} \mid \bar{a}, \overline{A}, I) &= \delta\big[\boldsymbol{J} = \boldsymbol{J}(\bar{a}, \overline{A})\big] \\ &\equiv \prod_{a\,A} \delta\big[T J_{a\,A} = \textstyle\sum_t \delta(a_t = a)\, \delta(A_t = A)\big].\end{aligned} \tag{36}$$

The second factor is derived from Bayes's theorem:

$$\mathrm{p}(\overline{A} \mid \bar{a}, I) = \frac{\mathrm{p}(\bar{a} \mid \overline{A}, I)\, \mathrm{p}(\overline{A} \mid I)}{\sum_{\overline{A}} \mathrm{p}(\bar{a} \mid \overline{A}, I)\, \mathrm{p}(\overline{A} \mid I)}. \tag{37}$$

In the last formula, let's first derive $\mathrm{p}(\bar{a} \mid \overline{A}, I)$. We make this

**assumption**: if $A_t$ is known, then knowledge of $a_{t'}$ or $A_{t'}$ with $t' \neq t$ is irrelevant for our inferences about $a_t$. (38)

It isn't a realistic assumption, but it's the one behind the maximum-entropy method in this particular application. From this assumption we have, using the product rule,

$$\mathrm{p}(\bar{a} \mid \bar{A}, I) = \prod_t \mathrm{p}(a_t \mid A_t, I), \tag{39}$$

and from simple sampling theory $\mathrm{p}(a_t \mid A_t, I) = G_{a_t\,A_t}$ for each $t$, so that

$$\mathrm{p}(\bar{a} \mid \bar{A}, I) = \prod_t G_{a_t\,A_t}. \tag{40}$$

The latter product can be rewritten by grouping together all $t$ in which the same $(a, A)$ pair appears: there are $T J_{a\,A}(\bar{a}, \bar{A})$ such $t$. Then we consider all such pairs:

$$\mathrm{p}(\bar{a} \mid \bar{A}, I) = \prod_{a,A} {G_{a\,A}}^{T J_{a\,A}(\bar{a},\bar{A})} \tag{41}$$

where $\prod_{a,A} \coloneqq \prod_a \prod_A$.

Now let's derive $\mathrm{p}(\bar{A} \mid I)$ in formula (37). We assume that the value of this probability has the same value for all sequences $\bar{A}$ having the same frequency distribution $\boldsymbol{F}(\bar{A})$; that is, it functionally depends on $\bar{A}$ only through $\boldsymbol{F}(\bar{A})$. More specifically we assume it can be written hierarchically as

$$\mathrm{p}(\bar{A} \mid I) = \pi(\boldsymbol{F}) = \int \mathrm{d}\boldsymbol{\nu}\; q(\boldsymbol{\nu}) \prod_A {\nu_A}^{T F_A} \tag{42}$$

for some density function $q(\boldsymbol{\nu})\,\mathrm{d}\boldsymbol{\nu}$. The integral expression is simply a mathematical way to write $\pi(\boldsymbol{F})$, and doesn't need to be further interpreted (if we can solve the integral analytically, it just disappears). But it's also the formula for infinite exchangeability, and from this point of view $\boldsymbol{\nu}$ can be interpreted as the long-run frequency distribution of the activities in all experiments performed in the same conditions (imagining to joint together their recording times). This is the point of view underlying the maximum-entropy method.

Replacing (41) and (42) into (37) and rearranging we find

$$\begin{aligned} \mathrm{p}(\bar{A} \mid \bar{a}, I) &= \frac{1}{Z} \int \mathrm{d}\boldsymbol{\nu}\; q(\boldsymbol{\nu}) \prod_{a,A} \Bigl[{G_{a\,A}}^{T J_{a\,A}(\bar{a},\bar{A})}\Bigr] \prod_A {\nu_A}^{T F_A} \\ &\equiv \frac{1}{Z} \int \mathrm{d}\boldsymbol{\nu}\; q(\boldsymbol{\nu}) \prod_{a,A} \bigl(G_{a\,A}\, \nu_A\bigr)^{T J_{a\,A}(\bar{a},\bar{A})}, \end{aligned} \tag{43}$$

where $Z \coloneqq \sum_{\bar{A}} \mathrm{p}(\bar{a} \mid \bar{A}, I)\, \mathrm{p}(\bar{A} \mid I)$ is the normalization factor and the second equality comes from rewriting

$$\nu_A{}^{T F_A} \equiv \nu_A{}^{T \sum_a J_{a\,A}} \equiv \prod_a \nu_A{}^{T J_{a\,A}}. \tag{44}$$

Note that the normalization factor is independent of $\bar{A}$ and only depends on $\bar{a}$, which is given and fixed in our inference.

We can now return to our initial probability distribution (35). Replacing (36) and (43) into it and rearranging we have

$$\mathrm{p}(\boldsymbol{J} \mid \bar{a}, I) = \frac{1}{Z} \int \mathrm{d}\boldsymbol{\nu}\; q(\boldsymbol{\nu}) \sum_{\bar{A}} \delta[\boldsymbol{J} = \boldsymbol{J}(\bar{a}, \bar{A})] \prod_{a,A} \bigl(G_{a\,A}\, \nu_A\bigr)^{T J_{a\,A}(\bar{a},\bar{A})}. \tag{45}$$

This expression can be considered as the marginal distribution of the joint density function

$$\mathrm{p}(\boldsymbol{J}, \boldsymbol{\nu} \mid \bar{a}, I) = \frac{1}{Z}\, q(\boldsymbol{\nu}) \sum_{\bar{A}} \delta[\boldsymbol{J} = \boldsymbol{J}(\bar{a}, \bar{A})] \prod_{a,A} \bigl(G_{a\,A}\, \nu_A\bigr)^{T J_{a\,A}(\bar{a},\bar{A})}. \tag{46}$$

If we represent this density by a set of Monte Carlo samples for $(\boldsymbol{J}, \boldsymbol{\nu})$, we automatically also have samples for the distribution for $\boldsymbol{J}$, that for $\boldsymbol{F}$, and that for $\boldsymbol{\nu}$.

Let's consider the sum over all possible sequences, $\sum_{\bar{A}}$, in the joint density (46). Owing to the delta in the summand, the only $\bar{A}$s that contribute to the sum are those for which $\boldsymbol{J}(\bar{a}, \bar{A}) = \boldsymbol{J}$. This also means that all terms in the sum have the same numerical value, because their values depend on $\bar{A}$ only through $\boldsymbol{J}(\bar{a}, \bar{A})$, which is fixed. We must therefore only find how many terms there are in the sum, and multiply the value of one term for the number of terms.

With $\bar{a}$ and $\boldsymbol{J}$ fixed, we are asking how many $\bar{A}$s satisfy $\boldsymbol{J}(\bar{a}, \bar{A}) = \boldsymbol{J}$. Consider the problem from the following point of view. We have a grid of boxes with $(n+1)$ rows and $(N+1)$ columns, indexed by $(a, A)$. A sequence of $T$ balls go into the boxes: if the $t$th ball goes into the $(a, A)$ box, it means that at the $t$th time bin the activities of sample and full network were $a_t = a$ and $A_t = A$. In the typical combinatorial problem we would ask in how many ways we can fill the boxes, with $T J_{a\,A}$ balls in the box $(a, A)$, by throwing the $T$ balls. The number would be given by all $T!$ possible permutations of the balls, but considering permutations of two or more balls within the same box as equivalent, thus finding the multinomial coefficient

$$\binom{T}{T\boldsymbol{J}} \equiv \binom{T}{TJ_{0\,0},\dots,TJ_{n\,N}} := \frac{T!}{\prod_{a,A}(TJ_{a\,A})!}. \tag{47}$$

In our case, however, we have one constraint: $\bar{a}$ is fixed, which means that the *row* in which each ball must fall is determined and fixed. The $t$th ball must perforce fall into row $a_t$. In considering all possible permutations we must therefore exclude those that change the rows of the balls. This means that only within-row permutations are allowed. Within each row the counting proceeds as usual, so that for row $a$, which has a total of $T\sum_A J_{a\,A} \equiv T f_a$ balls, we have

$$\binom{Tf_a}{TJ_{a\,0},\dots,TJ_{a\,N}} = \frac{(Tf_a)!}{\prod_A (TJ_{a\,A})!}. \tag{48}$$

We therefore find that the number of terms in the sum of (46) is

$$\prod_a \binom{Tf_a}{TJ_{a\,0},\dots,TJ_{a\,N}}. \tag{49}$$

If $T$ is large the logarithm of the multinomial coefficient can be approximated using the Shannon entropy $H$, which for a generic distribution $(x_i)$ satisfies the bounds[44]

$$TH(x_0,\dots,x_N) - \ln\binom{T+N}{T} \leqslant \ln\binom{T}{Tx_0,\dots,Tx_N} \leqslant TH(x_0,\dots,x_N). \tag{50}$$

We therefore approximate the multiplicity (49) as

$$\exp\Bigl[T\sum_a f_a\, H\Bigl(\frac{J_{a\,0}}{f_a},\dots,\frac{J_{a\,N}}{f_a}\Bigr)\Bigr] \equiv \exp\Bigl[-T\sum_a f_a \sum_A \frac{J_{a\,A}}{f_a}\ln\frac{J_{a\,A}}{f_a}\Bigr] \equiv$$

$$\exp\Bigl[-T\sum_{a,A} J_{a\,A}\ln J_{a\,A} + T\sum_a f_a \ln f_a\Bigr] \equiv \exp[T\,H(\boldsymbol{J}) - T\,H(\boldsymbol{f})], \tag{51}$$

where the second line is obtained by combining the sums, simplifying, and using the definition of Shannon entropy. 🧩 Is it possible that this approximation affects the result? Check this!

[44] Lemma 2.2 pp. 429–430 in I. Csiszár et al. (2004): *Information theory and statistics: a tutorial*. Foundations and Trends in Communications and Information Theory **1**$^4$, 417–528. http://www.renyi.hu/~csiszar/.

The sum over $\bar{A}$ in formula (46) can therefore be replaced by any single summand multiplied by the multiplicity (51) above. The marginal frequency distribution $\boldsymbol{f}$ is still fixed, determined by $\bar{a}$, so we still have the constraint $\sum_A J_{a\,A} = f_a$ for each $a$, which we compactly write $\delta(\sum \boldsymbol{J} = \boldsymbol{f})$. We find

$$\begin{aligned}
&\mathrm{p}(\boldsymbol{J}, \boldsymbol{\nu} \mid \bar{a}, I) \\
&\qquad = \frac{1}{Z}\, \delta(\textstyle\sum \boldsymbol{J} = \boldsymbol{f})\; q(\boldsymbol{\nu})\; \exp[T\, H(\boldsymbol{J}) - T\, H(\boldsymbol{f})]\; \displaystyle\prod_{a,A} \bigl(G_{a\,A}\, \nu_A\bigr)^{T J_{a\,A}(\bar{a},\bar{A})} \\
&\qquad = \frac{1}{Z}\, \delta(\textstyle\sum \boldsymbol{J} = \boldsymbol{f})\; q(\boldsymbol{\nu})\; \exp\bigl\{T\bigl[H(\boldsymbol{J}) - H(\boldsymbol{f}) + \sum_{a,A} J_{a\,A} \ln(G_{a\,A}\, \nu_A)\bigr]\bigr\},
\end{aligned} \tag{52}$$

where the second equality comes from re-expressing the product over $(a, A)$ in exponential-logarithm form.

Now we note that $G_{a\,A}\, \nu_A$ is a normalized distribution in $(a, A)$ owing to the properties of the hypergeometric distribution and of $\boldsymbol{\nu}$. We denote it $\boldsymbol{G} \cdot \boldsymbol{\nu}$. We can also use the definition of relative entropy

$$H[(x_i);(y_i)] \coloneqq \textstyle\sum_i x_i \ln \dfrac{x_i}{y_i} \equiv -H[(x_i)] - \textstyle\sum_i x_i \ln y_i, \tag{53}$$

to combine the first and last terms within the exponential. The logarithm of our joint density (46) can then finally be written as

$$\boxed{\begin{aligned}\ln \mathrm{p}(\boldsymbol{J}, \boldsymbol{\nu} \mid \bar{a}, I) = \ln \delta(\textstyle\sum \boldsymbol{J} = \boldsymbol{f}) + \ln q(\boldsymbol{\nu}) - T\, H(\boldsymbol{J}; \boldsymbol{G} \cdot \boldsymbol{\nu}) \\ - T\, H(\boldsymbol{f}) - \ln Z(\bar{a})\end{aligned}} \tag{54}$$

The last two terms are constants: they only depend on $\bar{a}$ and $\boldsymbol{f}(\bar{a})$, which are given and fixed in our problem. They can therefore be discarded in Monte Carlo sampling. In sampling, the delta term is taken care of by restricting the sampling to the set of allowed $\boldsymbol{J}$.

The density $q(\boldsymbol{\nu})$ remains to be specified. We use an entropic density

$$q(\boldsymbol{\nu}) \propto \exp[L\, H(\boldsymbol{\nu}; \boldsymbol{r})] \tag{55}$$

where $\boldsymbol{r}$ is a reference distribution – we take the uniform one for simplicity – and $L$ is of order unity, say less than 10. This density is motivated my the consideration of multiplicities: a frequency distribution, like $\boldsymbol{\nu}$, can be realized in a number of ways given by a multinomial coefficient. This coefficient, by formula (50), approximately equal to the exponential of the Shannon entropy of the distribution. So the density (55) keeps track of this multiplicity, but regulates its importance via the parameter $L$. We can also choose $L = 0$, which leads to $q(\boldsymbol{\nu})$ being uniform in $\boldsymbol{\nu}$.

### A.3 Approximations

We consider two successive approximations of the log-density (54), based on the fact that $T$ is large.

The relative entropy $H(\boldsymbol{J}; \boldsymbol{G} \cdot \boldsymbol{\nu})$ is always positive, and vanishes if only if $J = \boldsymbol{G} \cdot \boldsymbol{\nu}$. This term is multiplied by $-T$ and appears in an exponential. We therefore approximate it with a delta.

This delta restricts $\boldsymbol{J}$ to the form $\boldsymbol{G} \cdot \boldsymbol{\nu}$, making it completely determined by $\boldsymbol{\nu}$. The density (54) thus collapses onto the $\boldsymbol{\nu}$-space. More precisely a subspace, since $\boldsymbol{J}$ must also satisfy the constraint $\sum_a \boldsymbol{J} = \boldsymbol{f}$, which becomes $\sum_a \boldsymbol{G} \cdot \boldsymbol{\nu} = \boldsymbol{f}$, constraining $\boldsymbol{\nu}$.

Such a constraint is actually too strong, in the sense that no distribution $\boldsymbol{\nu}$ can satisfy it – it leads to negative values $nu_a$ for some $a$s. This problem is mitigated by considering a softer constraint between $\boldsymbol{J}$ and $\boldsymbol{f}$, for example taking into account errors in spike sorting or just considering the first $m$ moments of $\boldsymbol{f}$. We express the softer constraint by replacing the delta with a normal N whose covariance matrix $\boldsymbol{\sigma}$ is determined from the spike-sorting error. With this approximation we reduce our problem to the density

$$\boxed{\ln \mathrm{p}(\boldsymbol{\nu} \mid \bar{a}, I) \propto \ln \mathrm{N}(\textstyle\sum_a \boldsymbol{G} \cdot \boldsymbol{\nu} \mid \boldsymbol{f}, \boldsymbol{\sigma}) + \ln q(\boldsymbol{\nu}),} \tag{56}$$

with $q(\boldsymbol{\nu})$ given by the entropic density (55). We can call this the 'maximum-entropy approximation with Bayesian correction'. It can also be represented by a set of Monte Carlo samples $\{\boldsymbol{\nu}^{(i)}\}$.

As a further approximation we can simply consider the mode of (56). This is found by maximizing the relative entropy (55) under the constraints about $\boldsymbol{f}$. The solution is the one given by the maximum-entropy method applied to the full network with subnetwork constraints.

Figure A shows comparison of the expected values (means of the Monte Carlo samples) for the density (54) (for $\boldsymbol{F}$), the density (56) (for $\boldsymbol{\nu}$

and using the first six moments of $f$ as constraints), and the maximum-entropy solution (six moments as constraints).

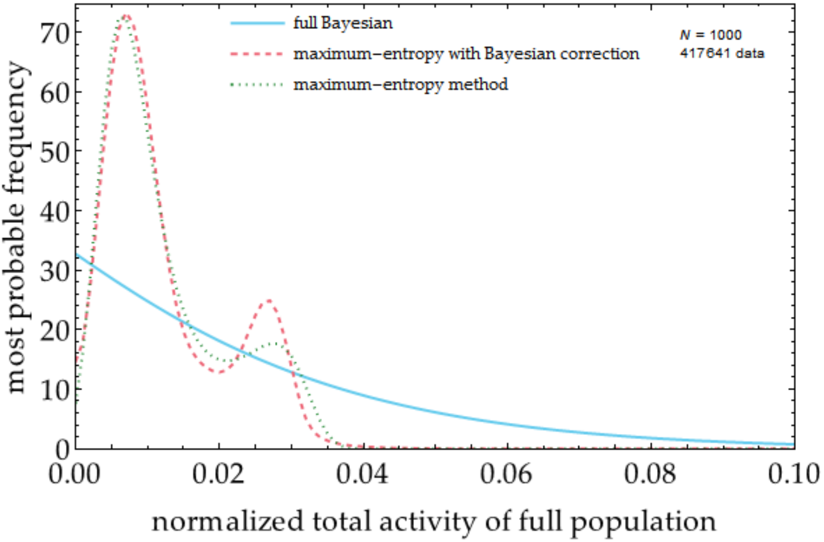

Figure 8 Comparison of (54), (56), and the maximum-entropy approximation (the vertical-axis label is wrong).