Estimating probabilities from experimental frequencies

Inés Samengo*
Centro Atómico Bariloche and Instituto Balseiro, 8400 San Carlos de Bariloche, Río Negro, Argentina
(Received 9 November 2001; published 4 April 2002)

Estimating the probability distribution **q** governing the behavior of a certain variable by sampling its value a finite number of times most typically involves an error. Successive measurements allow the construction of a histogram, or frequency count **f**, of each of the possible outcomes. In this work, the probability that the true distribution be **q**, given that the frequency count **f** was sampled, is studied. Such a probability may be written as a Gibbs distribution. A thermodynamic potential, which allows an easy evaluation of the mean Kullback-Leibler divergence between the true and measured distribution, is defined. For a large number of samples, the expectation value of any function of **q** is expanded in powers of the inverse number of samples. As an example, the moments, the entropy, and the mutual information are analyzed.

DOI: 10.1103/PhysRevE.65.046124 PACS number(s): 02.50.Tt

I. ESTIMATING PROBABILITIES FROM EXPERIMENTAL FREQUENCIES

The estimation of probability distributions from a limited number of samples typically involves an error. Consider, for example, a random variable that can be either 0 or 1, both values with probability 1/2. An experimenter measures the variable, say, four times. If n_0 (or n_1) is the number of trials the result was 0 (or 1), the possible outcomes are $n_0=j, n_1=4-j$, where j may vary between 0 and 4. Each of those possibilities has a probability $3/2j!(4-j)!$ of occurring. If the experimenter estimates the underlying probability from the frequencies, his or her claim will be that the probability of getting a zero is $n_0/4$. However, in view that n_0 depends on the particular outcome of the four trials, only a fraction 3/16 of the times will this procedure give the correct result, that is $f_0=q_0=1/2$.

In the above example, there are three probability distributions involved. First, there is the *true* underlying probability **q**, actually governing the outcome of the experiment. In vector notation $\mathbf{q}=(q_0,q_1)$, and in the particular instance above $\mathbf{q}=(1/2,1/2)$. Then, there is the frequency count $\mathbf{f}=(f_0,f_1)$, where f_i is obtained by dividing n_i by the total number of measurements N (four in the example). And finally, there is the probability that $\mathbf{f}=\mathbf{q}$. To define this last probability, one has to consider all possible samples of N trials and evaluate how often the condition $\mathbf{f}=\mathbf{q}$ is fulfilled.

More generally, one can define the probability of measuring a particular **f**, while the underlying **q** remains fixed. This means to consider a probability distribution of all the possible frequency counts. The independent variable is the vector **f**, which varies in a discrete set, and the dependent variable is $p(\mathbf{f}|\mathbf{q})$.

The frequency count **f** is an estimation of the underlying **q**. In many applications, however, one is interested not quite in **q**, but rather in some function of **q**. Treves and Panzeri [1], for example, have quantified the mean error that an experimenter makes when evaluating the mutual information in the frequency count **f**, as an approximation to that in the true (and unknown) **q**. Their analysis was made in the same spirit as above, that is, they have considered **q** fixed, while the value of **f** depended on the particular outcome of N measurements. They have obtained a clean analytical result under an independence approximation. Their approach may be naturally generalized to situations where **q** is a probability density, that is, varies in a continuous set [2].

However, what the experimenter knows is not the true **q**, but one particular **f**, obtained after N observations. His or her aim is to estimate the most probable value of **q** (or of some function of **q**) from the knowledge of **f**. More generally, the experimenter may be interested in the whole distribution $P(\mathbf{q}|\mathbf{f})$, that is, the probability that the true distribution be **q**, given that he or she has measured **f**. This means to settle the problem the other way round as was studied by Treves and Panzeri and in the example above. It actually corresponds to Wolpert and Wolf's approach [3] in the estimation of entropies.

In the following section, the properties of the distribution $P(\mathbf{q}|\mathbf{f})$ are studied. In Sec. III, $P(\mathbf{q}|\mathbf{f})$ is written as a Gibbs' distribution, where the inverse number of samples plays the role of an effective temperature, and the Kullback-Leibler divergence between **f** and **q** is the equivalent of the energy of state **q**. As a consequence, a thermodynamic potential is defined, thus allowing the calculation of the mean Kullback-Leibler divergence between **f** and **q** by simple derivation. This inspires the expansion made in Sec. IV, where the expectation value of an arbitrary function of **q** can be written as a power series in the inverse number of samples. The case of the entropy, the mutual information, or any moment of the distribution **q** is shown in the examples of Sec. V. Next, in Sec. VI the analytical results are confronted with numerical simulations. Finally, in Sec. VII, the main results are summarized and discussed.

II. THE PROBABILITY DISTRIBUTION FOR THE TRUE PROBABILITY DISTRIBUTION

Consider the random variable X taking values from the set $\mathbf{x}=(x_1,\ldots,x_S)$ with probabilities $\mathbf{q}=(q_1,\ldots,q_S)$. In principle, there is no need that $x_1,\ldots,x_S$ be numerical values, it

*Electronic address: samengo@cab.cnea.gov.ar

suffices them to be any exclusive and exhaustive set of categories.

An experimenter makes N observations of the value of X and builds a histogram $\mathbf{n}=(n_1,\ldots,n_S)$, where n_i is the number of times the outcome was x_i. The experimenter considers the frequencies $\mathbf{f}=(f_1,\ldots,f_S)=(n_1/N,\ldots,n_S/N)$ as an estimation of the true underlying probability distribution $\mathbf{q}$. If the measurements are taken independently, the probability of measuring $\mathbf{f}$ given that the data are sorted according to $\mathbf{q}$ is equal to the probability of observing each x_i a number n_i of times, that is,

$$p(\mathbf{f}|\mathbf{q})=N!\Pi_i\frac{q_i^{n_i}}{n_i!}=\frac{N!}{\Pi_i(Nf_i)!}\exp\left(N\sum_i f_i\ln q_i\right). \quad (1)$$

However, the knowledge the experimenter has at hand is $\mathbf{f}$, not $\mathbf{q}$. He or she may, therefore, wonder what is the probability that the true distribution be $\mathbf{q}$, given that the outcome of the experiment was $\mathbf{f}$. This means to evaluate a probability density $P(\mathbf{q}|\mathbf{f})$, whose independent variable $\mathbf{q}$ runs over all the possible distributions of the data. That is, all vectors in $\Re^S$ such that

$$\sum_i q_i=1,$$
$$0\leqslant q_i\leqslant 1\,\forall i. \quad (2)$$

The set of all $\mathbf{q}$ obeying Eqs. (2) constitutes the domain $\mathcal{D}$ where $P(\mathbf{q}|\mathbf{f})$ is defined. It is a finite portion of an $(S-1)$-dimensional plane embedded in $\Re^S$, and is normal to the vector $(1,1,\ldots,1)$.

Notice that since each f_i is the ratio of two natural numbers, the set of possible frequencies $\mathbf{f}$ is discrete. The domain $\mathcal{D}$, on the contrary, contains a continuum of distributions $\mathbf{q}$. Consequently, $p(\mathbf{f}|\mathbf{q})$ is a probability, whereas $P(\mathbf{q}|\mathbf{f})$ is a density.

Bayes's rule states that

$$P(\mathbf{q}|\mathbf{f})=\frac{p(\mathbf{f}|\mathbf{q})P(\mathbf{q})}{p(\mathbf{f})}, \quad (3)$$

where $P(\mathbf{q})$ is the prior probability distribution for $\mathbf{q}$, and

$$p(\mathbf{f})=\int_{\mathcal{D}}P(\mathbf{f}|\mathbf{q})P(\mathbf{q})dS_{\mathbf{q}}. \quad (4)$$

Here, $dS_{\mathbf{q}}$ is a volume element in $\mathcal{D}$.

The prior $P(\mathbf{q})$ contains all additional pieces of knowledge about $\mathbf{q}$, apart from the experimental data. Here, the assumption is made that there is no *a priori* knowledge. However, it turns out to be crucial to specify *what* is it that is not known [5]. A prior that is uniform over $\mathcal{D}$, as was used by Wolpert and Wolf [3], is certainly not uniform over any nonlinear function of $\mathbf{q}$; for example, the log likelihood. Thus, not knowing anything about $\mathbf{q}$ implies knowing something about $\ln q$, which in turn may result in awkward scaling properties. In this work, the power prior

$$P_\beta(\mathbf{q})=\frac{\Pi_{i=1}^S q_i^{\beta-1}}{\mathcal{Z}_\beta} \quad (5)$$

is repeatedly used, with $\mathcal{Z}_\beta=\sqrt{S}[\Gamma(\beta)]^S/\Gamma(S\beta)$ (notice that when $\beta\to 0,\mathcal{Z}_\beta\to\sqrt{S}$). However, as was shown in Ref. [5] choosing any of these priors results in a surprisingly peaked *a priori* distribution of the possible entropies. Hence, the choice of the prior is a delicate issue and, in any particular application, it should be done carefully. Here, no attempt will be made to instruct on the way such a choice should be made, but since the results that follow are strongly grounded on the Bayesian inference, their validity is, at most, as good as the prior [3].

Replacing Eqs. (1) and (4) in Eq. (3),

$$P(\mathbf{q}|\mathbf{f})=\frac{\exp[-ND(\mathbf{f},\mathbf{q})]P(\mathbf{q})}{\mathcal{Z}}, \quad (6)$$

where D is the Kullback-Leibler divergence between $\mathbf{f}$ and $\mathbf{q}$,

$$D(\mathbf{f},\mathbf{q})=\sum_i f_i\ln\left(\frac{f_i}{q_i}\right), \quad (7)$$

and it quantifies the mean information for discriminating in favor of $\mathbf{f}$ against $\mathbf{q}$, given the data [4]. The function $\mathcal{Z}$ reads

$$\mathcal{Z}=\int_{\mathcal{D}}dS_{\mathbf{q}}P(\mathbf{q})\exp[-ND(\mathbf{f},\mathbf{q})]. \quad (8)$$

In the remainder of the section, the properties of $P(\mathbf{q}|\mathbf{f})$ are studied for the particular $P_\beta(\mathbf{q})$ defined in Eq. (5). In doing so, the integral

$$\int_{\mathcal{D}}\Pi_{i=1}^S q_i^{m_i}dS_{\mathbf{q}}=\sqrt{S}\,\frac{\Pi_i\Gamma(m_i+1)}{\Gamma\left(S+\sum_i m_i\right)} \quad (9)$$

is frequently encountered. Equation (9) was first derived in [3], and an alternative proof may be found in the Appendix.

For the priors in Eq. (5), the function $\mathcal{Z}$, Eq. (8), may be calculated analytically and it reads

$$\mathcal{Z}=\exp[N\mathcal{H}(\mathbf{f})]\sqrt{S}\,\frac{\Pi_{j=1}^S\Gamma(Nf_k+\beta)}{\Gamma(N+S\beta)}, \quad (10)$$

where $\mathcal{H}$ is the entropy of a distribution

$$\mathcal{H}(\mathbf{f})=-\sum_{i=1}^S f_i\ln f_i. \quad (11)$$

Thus, replacing Eq. (10) in Eq. (6)

$$P(\mathbf{q}|\mathbf{f})=\frac{\Gamma(N+S\beta)}{\sqrt{S}}\Pi_i\frac{q_i^{Nf_i+\beta-1}}{\Gamma(Nf_i+\beta)}. \quad (12)$$

The most probable $\mathbf{q}^M=(q_1^M,\ldots,q_S^M)$ is obtained by maximizing Eq. (12) under the normalization constraint. The result is

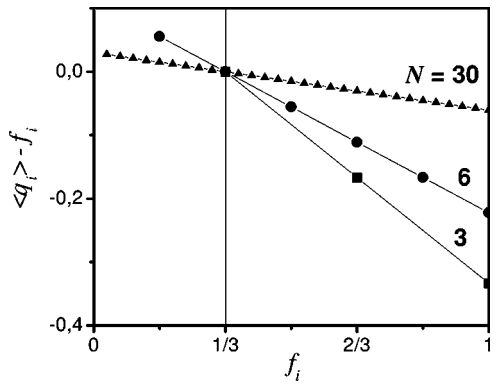

FIG. 1. Difference between $\langle q_i \rangle$ and f_i as a function of f_i. The value of β has been set to 1. The three lines correspond to $N=3$, 6, and 30. Here, X may take three values ($S=3$). When $f_i<1/3$, the expectation value of q_i is larger than the measured frequency f_i. As N increases, the effect becomes less important.

$$q_i^M = \frac{Nf_i+\beta-1}{N+S(\beta-1)}. \tag{13}$$

Thus, if $P(\mathbf{q})$ is uniform in $\mathcal{D}(\beta=1)$, then the most probable $\mathbf{q}$ is $\mathbf{f}$. With the maximum likelihood prior ($\beta\to 0$), the most probable $\mathbf{q}$ is shifted from $\mathbf{f}$ towards lower counts. The Krichevsky-Trofimov estimator [8] ($\beta=1/2$) and the Shurmann-Grassber [9] $\beta=1/S$ lie in between.

Using Eq. (9) the expectation value of each component q_i may be calculated,

$$\langle q_i \rangle = \frac{Nf_i+\beta}{N+S\beta}. \tag{14}$$

For the uniform prior $\beta=1$, this equation reduces to Laplace's estimator of probabilities. In Fig. 1 the difference between $\langle q_i \rangle$ and the frequency count f_i is shown for $\beta=1$. It is seen that when f_i is smaller than $1/S$, $\langle q_i \rangle$ is larger than f_i. On the other hand, if $f_i>1/S$ then $\langle q_i \rangle < f_i$. That is, the mean value of q_i is displaced from the frequency count so as to approach the flat distribution $1/S$. Of course, the larger the number of samples N, the smaller the effect. Changing the value of β is equivalent to rescaling the vertical axis of Fig. 1.

Typically, one wants to make a guess about the true $\mathbf{q}$. Here, two possible estimators have been calculated: the maximum $\mathbf{q}^M$ and the mean $\langle \mathbf{q} \rangle$. By using the maximum, one is choosing the value that is most probably correct. But of course, eventually one will also make an error. If one measures the error as $(\mathbf{q}^M-\mathbf{q})^2$, and averages it with $P(\mathbf{q}|\mathbf{f})$, its mean turns out to be larger than if one had chosen $\langle \mathbf{q} \rangle$ [3]. Hence, although $\mathbf{q}^M$ is the estimator that gives the correct answer most frequently, if one cares for the typical size of the errors, $\langle \mathbf{q} \rangle$ is a better choice.

When using $\langle \mathbf{q} \rangle$ as an estimator, the covariance matrix Σ_{ij} may be of interest. By means of Eq. (9) it is easy to show that for $i \neq j$,

$$\Sigma_{ij} = \langle (q_i - \langle q_i \rangle)(q_j - \langle q_j \rangle) \rangle = -\frac{(Nf_i+\beta)(Nf_j+\beta)}{(N+S\beta)^2(N+S\beta+1)} \to -\frac{f_i f_j}{N} \quad \text{when } N \gg S, \tag{15}$$

whereas for $i=j$,

$$\Sigma_{ii} = \langle (q_i - \langle q_i \rangle)^2 \rangle = \frac{(Nf_i+\beta)[N(1-f_i)+\beta(S-1)]}{(N+S\beta)^2(N+S\beta+1)} \to \frac{f_i(1-f_i)}{N} \quad \text{when } N \gg S. \tag{16}$$

The negative sign in Eq. (15) derives from the normalization condition: since the sum of all q_i is fixed to unity, if one of them surpasses its mean, it is to be expected that some other component will be below. In contrast, Eq. (16) shows that Σ_{ii} is always positive.

The expectation value of $\mathbf{q}$, Eq. (14), together with the covariance matrix equations (15) and (16) are useful for the Gaussian approximation to $P(\mathbf{q}|\mathbf{f})$, centered in its mean:

$$P(\mathbf{q}|\mathbf{f}) = K \exp\left[-\frac{1}{2} (\mathbf{q}-\langle \mathbf{q} \rangle)^t \tilde{\Sigma}^{-1} (\mathbf{q}-\langle \mathbf{q} \rangle) \right], \tag{17}$$

where the superscript t means transposed and K is a normalization constant. Equation (17) is only defined in the plane containing $\mathcal{D}$, normal to the vector $(1,1,\ldots,1)$. Actually, Σ does not have an inverse in the entire space $\Re^S$, since the direction $(1,1,\ldots,1)$ is one of its eigenvectors with eigenvalue equal to zero. However, Σ being a symmetric matrix, it can be diagonalized by an orthogonal basis. Hence, the $S-1$ remaining eigenvectors lie in the plane containing $\mathcal{D}$. The restriction of Σ into that subspace is $\tilde{\Sigma}$, and its inverse is the matrix appearing in the exponent of Eq. (17).

In order to normalize the approximation (17) an integral of a Gaussian function in $\mathcal{D}$ is needed. This is certainly not an easy task. If, however, one can assume that the distribution is sufficiently peaked so that $P(\mathbf{q}|\mathbf{f}) \approx 0$ for $\mathbf{q}$ in the border of $\mathcal{D}$, then the domain $\mathcal{D}$ can be extended to the whole plane normal to $(1,1,\ldots,1)$. In that case, $K^{-1} = \sqrt{2\pi \Pi_j \lambda_j}$, where λ_j are the $S-1$ eigenvalues of $\tilde{\Sigma}$. While the calculation of all the λ_j is a difficult problem, it is quite straightforward to show that when $N \gg S$, all the λ_j are proportional to $1/N$. Therefore, the square root of each eigenvalue is a useful measure of the width of $P(\mathbf{q}|\mathbf{f})$ in the direction of its eigenvector.

However, the Gaussian approximation (17) is not useful for other purposes, as for instance, calculating mean values, since it lacks from analytical expressions such as Eq. (9). As a consequence, in what follows, the full Eq. (12) is used.

Equation (9) allows the evaluation of all moments of $P(q_i|\mathbf{f})$,

$$\langle q_i^k \rangle = \frac{\Gamma(Nf_i+k+\beta)\Gamma(N+S\beta)}{\Gamma(Nf_i+\beta)\Gamma(N+S\beta+k)}. \tag{18}$$

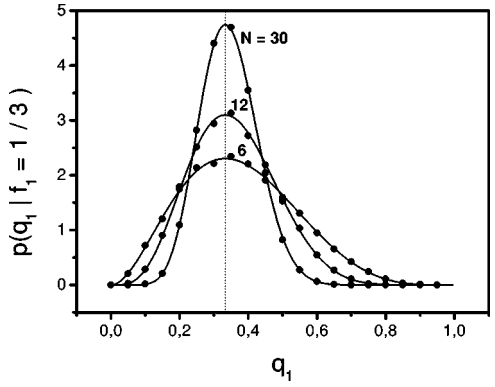

FIG. 2. Probability distribution $P(q_1|f_1)$ for the case $f_1=1/3$, $\beta=1$, and $S=2$. Different curves correspond to several values of the number of samples N. The full line depicts the analytical result Eq. (19), while the dots are the numerical simulations (see Sec. VI).

Since the moments are the coefficients of the Taylor expansion of the Fourier transform of a distribution, the single-component distribution reads

$$P(q_i|\mathbf{f})=P(q_i|f_i)=\frac{q^{Nf_i+\beta-1}(1-q)^{N(1-f_i)+\beta(S-1)-1}}{B[Nf_i+\beta,N(1-f_i)+\beta(S-1)]}, \tag{19}$$

where $B(x,y)=\Gamma(x)\Gamma(y)/\Gamma(x+y)$. Figure 2 displays the distribution $P(q_i|f_i)$ for three different values of N and $\beta=1$. In all cases, when N is large, the distribution is symmetrical, and reaches its maximum value of $q_i=f_i=1/3$. In fact, it may be shown analytically that when $N\gg 1$,

$$\lim_{N\gg 1} P(q_i|f_i)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp[-(q_i-f_i)^2/2\sigma^2], \tag{20}$$

where $\sigma=[f_i(1-f_i)/N]^{1/2}$. That is, the distribution tends to a Gaussian function centered at the experimental frequency, and with a mean dispersion that diminishes with the square root of the number of samples. Notice that in this limit, $P(\mathbf{q}|\mathbf{f})$ does not depend on β.

It may be seen in Fig. 2 that for smaller values of N, the distribution is no longer symmetrical. In fact, since $S=2$ and $f_1=1/3<1/S$, the tail in $P(q_1|f_1)$ extends to the right, resulting in a positive $\langle q_i\rangle-f_i$, as predicted by Eq. (18).

III. THE INVERSE NUMBER OF SAMPLES AS AN EFFECTIVE TEMPERATURE

Equation (6) states that $P(\mathbf{q}|\mathbf{f})$ is completely analogous to a Gibbs distribution, where the number of samples N plays the role of the inverse of the temperature, $D(\mathbf{f},\mathbf{q})$ is the equivalent to the energy of the state $\mathbf{q}$, and $P(\mathbf{q})$ is the density of states. This analogy was first pointed out in the context of machine learning [6], and since then, several times in learning theory (see, for example, [7]). In these cases, when fluctuations were neglected, the probability distribution under study had the form of Eq. (6). In the present context, no approximations are needed to write down Eq. (6).

The exponential factor in Eq. (6) depends on $\mathbf{q}$ and $\mathbf{f}$ only in the combination $D(\mathbf{f},\mathbf{q})$, diminishing exponentially as the divergence between the two distributions grows. Its maximum is attained when $D=0$. It can be shown [4] that for any $\mathbf{f}$ and $\mathbf{q}$, $D(\mathbf{f},\mathbf{q})\geqslant 0$, and the equality holds only when $\mathbf{f}=\mathbf{q}$.

Defining the thermodynamic potential

$$F=-\ln\mathcal{Z}, \tag{21}$$

it follows that

$$\langle D\rangle=\frac{\partial F}{\partial N}, \tag{22}$$

$$\sigma_D^2=\langle D^2-\langle D\rangle^2\rangle=-\frac{\partial^2 F}{\partial N^2}, \tag{23}$$

where the mean values $\langle(\cdot)\rangle$ are defined by $\int_{\mathcal{D}}(\cdot)P(\mathbf{q}|\mathbf{f})dS_{\mathbf{q}}$.

For example, when the prior is given by Eq. (5),

$$\langle D\rangle=\mathcal{H}(\mathbf{f})-\Psi(N+S\beta)+\sum_i f_i\Psi(Nf_i+\beta), \tag{24}$$

where $\Psi(x)=d\ln\Gamma(x)/dx$ is the digamma function [10]. It is easy to show that

$$\lim_{N\gg S}\langle D\rangle=\frac{S-1}{2N}+\mathcal{O}(1/N^2). \tag{25}$$

Here, both N and Nf_i have been supposed large for all i. Since f_i is of the order of $1/S$, the above limit holds when $N\gg S$. Equation (25) states that for a large number of samples, the expected value of the divergence between the experimental frequencies and the true distribution does not depend on the measured $\mathbf{f}$. It grows linearly with the number of items and decreases as $1/N$.

Accordingly,

$$\sigma_D^2=-\Psi^1(N+S\beta)+\sum_{i=1}^{S} f_i^2\Psi^1(Nf_i+\beta), \tag{26}$$

where $\Psi^1(x)=d\Psi(x)/dx$ is the first polygamma function [10]. Taking the limit of a large number of samples,

$$\lim_{N\gg S}\sigma_D^2=\frac{S-1}{2N^2}+\mathcal{O}(1/N^3). \tag{27}$$

In the limit $N\gg S$, the mean quadratic dispersion does not depend on the measured f_i.

IV. ESTIMATION OF FUNCTIONALS OF q, FOR A LARGE NUMBER OF SAMPLES.

Many times, one is interested in the value of some function $W(\mathbf{q})$. For instance, if X takes numerical values, W may be the mean $\overline{X}=\Sigma_i x_i q_i$. Or, in some other application, W may be the entropy of the distribution $\mathbf{q}$ [see Eq. (11)]. If the

set X is the Cartesian product of two other sets $X=Z^1\times Z^2$ such that $\forall x_i\in X: x_i=(z_a^1,z_b^2)$, where $z_a^1\in Z^1$ and $z_b^2\in Z^2$, then W may be the mutual information I between Z^1 and Z^2,

$$I=\sum_{ab} q_{ab}\ln\left[\frac{q_{ab}}{q_{a.}q_{.b}}\right], \tag{28}$$

where

$$q_{a.}=\sum_b q_{ab},$$

$$q_{.b}=\sum_a q_{ab}. \tag{29}$$

Since $\mathbf{q}$ is unknown, an interesting guess for $W(\mathbf{q})$ is its Bayesian estimation

$$\langle W\rangle=\int_{\mathcal{D}} W(\mathbf{q})P(\mathbf{q}|\mathbf{f}), \tag{30}$$

which has the appealing property of minimizing the mean square error [3]. The zero-order guess for $\langle W\rangle$ is $W(\mathbf{f})$. In what follows, a systematic method to improve this value is derived.

In the preceding section the expectation value of the divergence between the true and the measured distribution was calculated, as well as the size of the fluctuations, for the priors in Eq. (5). As the number of samples increases, both the expected divergence and the fluctuations diminish as $1/N$. Since a small divergence means that the two distributions are necessarily very similar, only the $\mathbf{q}$ that are very near $\mathbf{f}$ have a nonvanishing probability—for D sufficiently small, this argument holds for any definition of similarity.

As a consequence, it is reasonable to expand $W(\mathbf{q})$ in its Taylor series in the neighborhood of $\mathbf{f}$. Hence, Eq. (30) reads

$$\langle W\rangle=\left\langle \sum_{k=0}^{\infty}\frac{1}{k!}\left(\sum_{i=1}^{S}(q_i-f_i)\frac{\partial}{\partial q_i}\right)^k W\Big|_{\mathbf{f}}\right\rangle. \tag{31}$$

Since $P(\mathbf{q}|\mathbf{f})$ decreases dramatically as $\mathbf{q}$ departs from $\mathbf{f}$, the higher-order terms (large k) in Eq. (31) should become negligible, at least, for large N.

In the first place, the mean values of Eq. (31) are evaluated for the special case of the power law priors. This involves, basically, the computation of integrals in $\mathcal{D}$ of $\Pi_{i=1}^{S}(q_i-f_i)^{k_i}$ for a set of non-negative indices $(k_i,k_2,\ldots k_S)$ that sum up to K. This can be done using Eq. (9). Of course, the term $k=0$—that is, the raw guess—does not depend on N. It may be shown that only $k=1$ and $k=2$ are proportional to $1/N$. Specifically,

$$\langle q_i-f_i\rangle=\frac{\beta(1-Sf_i)}{N+S\beta}\to\frac{\beta(1-Sf_i)}{N}\quad\text{when}\quad N\gg S. \tag{32}$$

In the same way if $i\neq j$,

$$\langle (q_i-f_i)(q_j-f_j)\rangle = -\frac{Nf_if_j-\beta[\beta+(1+S\beta)(Sf_if_j-f_i-f_j)]}{(N+S\beta)(N+S\beta+1)} \to -\frac{f_if_j}{N}\quad\text{when}\quad N\gg S, \tag{33}$$

whereas when $i=j$,

$$\langle (q_i-f_i)^2\rangle=\frac{Nf_i(1-f_i)+\beta[1+\beta+f_i(1+S\beta)(Sf_i-2)]}{(N+S\beta)(N+S\beta+1)} \to\frac{f_i(1-f_i)}{N}\quad\text{when}\quad N\gg S. \tag{34}$$

Summarizing, to first order in $1/N$,

$$\langle W\rangle\approx W(\mathbf{f})+\sum_{i=1}^{S}\frac{\partial W}{\partial q_i}\Big|_{\mathbf{f}}\frac{\beta(1-Sf_i)}{N}+\frac{1}{2}\sum_{i=1}^{S}\frac{\partial^2 W}{\partial q_i^2}\Big|_{\mathbf{f}}\frac{f_i(1-f_i)}{N} -\sum_{i=1}^{S}\sum_{j<i}\frac{\partial^2 W}{\partial q_i\partial q_j}\Big|_{\mathbf{f}}\frac{f_if_j}{N}. \tag{35}$$

This general formula allows the calculation of the first correction of the expectation value of an arbitrary function $W(\mathbf{q})$, whenever the prior is given by Eq. (5).

Now, consider the more general case of an arbitrary prior. If $P(\mathbf{q})$ is not given by Eq. (5), then one can still proceed as above, but by replacing $W(\mathbf{q})$ by the product $W(\mathbf{q})P(\mathbf{q})$, and setting $\beta=1$.

V. EXAMPLES

Here, the expansion (35) is applied to a few particular cases. Wolpert and Wolf [3] have already calculated the first two examples exactly (Secs. V A and V B) in the particular case of $\beta=1$. Their results, once expanded up to first order in $1/N$ are now compared to Eq. (35) for verification. The advantage of Eq. (35) is that, in contrast to Wolpert and Wolf's approach, it applies to any function W. The counterpart, of course, is that it gives no more than the first correction to $\langle W\rangle$. Section V C deals with the calculation of moments.

A. The mean value of the entropy

First, the function $W(\mathbf{q})$ is taken to be the entropy $\mathcal{H}$ of the distribution $\mathbf{q}$, defined in Eq. (11), for $\mathbf{q}=\mathbf{f}$. It is easy to verify that $\partial\mathcal{H}/\partial q_i=-(1+\ln q_i)$, whereas $\partial^2\mathcal{H}/\partial q_i\partial q_j=-\delta_{ij}/q_i$, where δ_{ij} is Kroeneker delta function: $\delta_{ij}=1$ if $i=j$ and $\delta_{ij}=0$ if $i\neq j$. Replacing in Eq. (35) and keeping only up to the first order in $1/N$ one arrives at

$$\langle\mathcal{H}\rangle=\left(1-\frac{\beta S}{N}\right)\mathcal{H}(\mathbf{f})+\frac{\beta}{N}\sum_{i=1}^{S}\ln\left(\frac{1}{f_i}\right)-\frac{S-1}{2N}+\mathcal{O}(1/N^2). \tag{36}$$

For the case of $\beta=1$, this same expression is obtained by expanding the exact result obtained in Ref. [3],

$$\langle \mathcal{H} \rangle_{[3]} = -\sum_{i=1}^{S} \frac{Nf_i+1}{N+S}[\Phi^{(1)}(Nf_i+2)-\Phi^{(1)}(N+S+1)], \tag{37}$$

where $\Phi^{(1)}(x)=d\ln\Gamma(x)/dx$ is the digamma function [10].

B. The mean value of the mutual information

Now W is taken to be the mutual information between two sets, as defined by Eq. (28). Replacing in Eq. (35),

$$\langle I \rangle = I(\mathbf{f})\left(1-\beta\frac{S_1S_2}{N}\right)+\frac{S_1S_2+1-S_1-S_2}{2N} + \frac{\beta}{N}\sum_{ab}\ln\left(\frac{f_{ab}}{f_{a.}f_{.b}}\right), \tag{38}$$

where S_1 and S_2 are the number of elements in the sets Z^1 and Z^2. When $\beta=1$, Eq. (38) coincides with the expansion up to first order in $1/N$ of the exact result derived in [3],

$$\langle I \rangle_{[3]} = \sum_{ab}\frac{Nf_{ab}+1}{N+S_1S_2}[\Phi^{(1)}(Nf_{ab}+2)-\Phi^{(1)}(N+S_1S_2+1)] - \sum_{a}\frac{Nf_{a.}+S_2}{N+S_1S_2}[\Phi^{(1)}(Nf_{a.}+S_2+1)-\Phi^{(1)}(N + S_1S_2+1)] - \sum_{b}\frac{Nf_{.b}+S_1}{N+S_1S_2}[\Phi^{(1)}(Nf_{.b}+S_1+1) - \Phi^{(1)}(N+S_1S_2+1)]. \tag{39}$$

The quantities $f_{a.}$ and $f_{.b}$ in Eqs. (38) and (39) are defined as in Eq. (29).

In contrast to the result obtained in [1], the first-order correction to the mutual information does bear a dependence on the values of the individual probabilities f_{ab}. There is no conflict, however, between the two results, since the mean value in Eq. (38) involves the distribution $P(\mathbf{q}|\mathbf{f})$. The approach in [1], instead, uses $p(\mathbf{f}|\mathbf{q})$, while the true $\mathbf{q}$ is fixed. In the present approach, the mean value $\langle I \rangle$ can be either higher or lower than $I(\mathbf{f})$.

C. The mean value of functions of X

Consider a function $g:\{x_1,\ldots,x_S\}\to\Re$ that maps the possible values of X into real numbers. For example, if X takes numerical values, then g_k can be such that $g_k(x_i)=x_i^k$. For each such g, another function $G:\mathcal{D}\to\Re$ is defined, namely, $G(\mathbf{q})=\Sigma_i g(x_i)q_i$. In the example above, G_k is the k moment of the distribution $\mathbf{q}$. The expectation value $\langle G \rangle$ is easily calculated using Eq. (35) and reads

$$\langle G \rangle = G(\mathbf{f})\left(1-\frac{\beta S}{N}\right)+\frac{\beta}{N}\sum_{i=1}^{S} g(x_i). \tag{40}$$

In particular, for the g_k considered above, this is the first-order correction to all moments of $\mathbf{q}$.

VI. NUMERICAL SIMULATIONS

In this section, Eq. (35) is confronted with the result of numerical simulations. Once again, and just to follow previous studies, $W(\mathbf{q})$ is set equal to the mutual information. However, in contrast to what was done up to now [1–3], the simulations are performed strictly within the present framework. That is, the measured frequency $\mathbf{f}$ is kept fixed, and the probability for the true $\mathbf{q}$ is evaluated.

The procedure to measure numerically $P(\mathbf{q}|\mathbf{f})$ is now explained. As before, X takes values in a set of S elements. Hence, $\mathbf{f}$ and $\mathbf{q}$ are S-dimensional vectors. The value of $\mathbf{f}$ is fixed. The domain $\mathcal{D}$ is discretized into a number J of cells. Each cell corresponds to a vector $\mathbf{q}$ that will be visited by the program. The larger the number of cells J, the better the sampling of the domain $\mathcal{D}$. For each one of these cells, the value of X is measured N times. The outcomes are sorted with the distribution $\mathbf{q}$ of the actual cell. If the frequency count thus obtained equals $\mathbf{f}$, the counter of the selected cell is increased (there is a counter for each cell in $\mathcal{D}$). The comparison between the frequency count and the (fixed) $\mathbf{f}$ is done with precision ϵ. The procedure is repeated M times (M large) in order to have enough counts. This algorithm allows to construct a histogram for the probability that a given $\mathbf{q}\in\mathcal{D}$ generates the selected $\mathbf{f}$.

For simplicity, in the results below, the number of trials M is the same for all cells. This is equivalent to using a uniform prior in $\mathcal{D}(\beta=1)$. A simulation with a nonuniform prior can be carried out by choosing a different M for each cell.

The two parameters that determine the precision of the simulations are J and ϵ. If D_J is the Kullback-Leibler divergence between two neighboring $\mathbf{q}$ cells, whenever $1/N \ll D_J$ then the only vector $\mathbf{q}$ that produces frequency counts equal to $\mathbf{f}$ is $\mathbf{q}=\mathbf{f}$. That is, for N sufficiently large, the discretized system behaves as if $N=\infty$. Notice that for large J, two neighboring cells correspond to $\mathbf{q}$ and $\mathbf{q}+\delta\mathbf{q}$, with each $\delta q_i \propto J^{S-1}$. Thus, the Kullback-Leibzig distance between the two is $\approx S/J^{S-1}$. This means that when N reaches J^{S-1}/S, the simulation starts to behave as if N were actually infinite.

On the other hand, if ϵ is not small enough, one mistakenly counts coincidences with $\mathbf{f}$, just because the criterion used in the comparison is too brute. In other words, a large ϵ allows cells $\mathbf{q}$ too far away from $\mathbf{f}$ to give rise to frequency counts equal to $\mathbf{f}$. That is, the system behaves as if N where smaller than its actual value.

The dots in Fig. 2 show the result of the above procedure for a single component q_1. As observed, there is very good agreement with the full line, showing the analytical result, Eq. (12).

To evaluate the expectation value of a certain function, one simply needs to calculate the sum

$$\langle W \rangle|_{\text{numerical}} = \sum_{\text{cells in } \mathcal{D}} W(\mathbf{q})P(\mathbf{q}|\mathbf{f}), \tag{41}$$

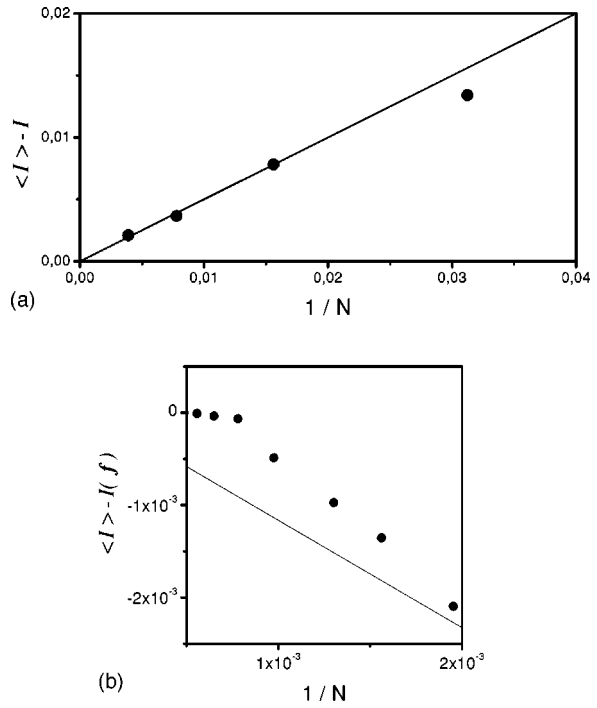

FIG. 3. Difference between the expectation value of the mutual information $\langle I\rangle$ and the measured $I(\mathbf{f})$ as a function of the inverse number of samples $1/N$. The $\beta=1$ prior was considered. The full line represents the analytical result, Eq. (38), and the dots the simulations. In (a), $f_{11}=f_{12}=f_{21}=f_{22}=1/4$ and $I(\mathbf{f})=0$. For each cell in $\mathcal{D}$, 30 000 sets of N samples have been sorted. In (b), $f_{11}=0.4$, $f_{12}=0.1$, $f_{21}=0.1$, and $f_{22}=0.4$, so $I(\mathbf{f})=0.192\,745$. For each cell in $\mathcal{D}$, 10 000 sets of N samples have been sorted. In both cases, each axis in $\mathbf{q}$ space has been divided into 20 intervals, in order to discretize $\mathcal{D}$, while the parameter ϵ was set to 0.0125.

using the $P(\mathbf{q}|\mathbf{f})$ obtained with the algorithm explained above. Figure 3 depicts the result for the mutual information with $\beta=1$. The dots represent the simulations, Eq. (41), whereas the full line shows the analytical result (38). The computational time required to evaluate $P(\mathbf{q}|\mathbf{f})$ increases exponentially with the number of dimensions S. Hence, in the present comparison it is desirable to keep S as small as possible. However, in order to define a mutual information, two sets Z^1 and Z^2 are needed with S_1 and S_2 elements each. In Fig. 3, $S_1=2$ and $S_2=2$, thus making a three-dimensional domain $\mathcal{D}$.

In (a) the selected $\mathbf{f}$ had no mutual information: $I(\mathbf{f})=0$. The graph shows that the expectation value of I is positive. With the chosen parameters (see the caption of the figure), the analytical result (38) coincides exactly with that derived by Treves and Panzeri [1], that is, $\langle I\rangle=(S_1-1)(S_2-1)/2N$. Since for $I(\mathbf{f})=0$, Eq. (38) reduces to $\langle I\rangle=S_1S_2+1-S_1-S_2/2N$ for some particular choices of S_I and S_J, the two expressions may coincide. It should be kept in mind, however, that this is just a coincidence, and the two mean values have different meanings.

In contrast, in case (b) the value of $I(\mathbf{f})$ is large (see the caption for details). In this case, the simulations confirm the phenomenon that was pointed out in the preceding section, namely, that the expectation value $\langle I\rangle$ may be lower than the measured $I(\mathbf{f})$.

It may be seen that for large N, all the dots concentrate in $\langle I\rangle=I(\mathbf{f})$. This is, as pointed out before, due to the discretization of $\mathcal{D}$. If the number of cells J is increased, one needs to go to a larger N to find such a saturation. On the contrary, for smaller N, the simulated $\langle I\rangle$ lies below its theoretical value. This is a manifestation of the finite nature of ϵ, and the phenomenon becomes less evident as ϵ is lowered.

VII. DISCUSSION

In this work, the probability density $P(\mathbf{q}|\mathbf{f})$ for the true distribution $\mathbf{q}$ given the experimental frequencies $\mathbf{f}$ is analyzed. Such a density, it is shown, may be written as a Gibbs distribution, where the inverse number of samples plays the role of an effective temperature, and the Kullback-Leibzig divergence between $\mathbf{f}$ and $\mathbf{q}$ is the equivalent of the energy of state $\mathbf{q}$. Its study is not only for academic purposes, but eventually also practical. In the ideal situation, it would be valuable to calculate $P(\mathbf{q}|\mathbf{f})$ while an experiment is being carried out, in order to know when the number of samples is already enough. The experimenter may thus decide to give an end to the sampling process when the width of $P(\mathbf{q}|\mathbf{f})$ reaches some acceptable value. For example, someone interested in measuring the public opinion prior to an election may wonder how many subjects need to be polled in order to have a reliable estimation of the forthcoming result. Many times, however, experiments comes to an end because of other factors (a deadline or a floor in the the amount of money, patience, or students). An estimation of the width of $P(\mathbf{q}|\mathbf{f})$ is valuable even in these cases, just to provide error bars.

One possibility is to write down the full $P(\mathbf{q}|\mathbf{f})$. However, being a function of many variables, this may not be very practical. A convenient parameter measuring the width of $P(\mathbf{q}|\mathbf{f})$ in several directions is the square root of the corresponding eigenvalues of $\tilde{\Sigma}$. These have been shown to diminish asymptotically as $1/N$. From the information-theoretical point of view, a more appealing parameter is the mean divergence D and its mean quadratic fluctuations. As is shown in Eq. (24), for small N, such a width depends on the value of $\mathbf{f}$. If $N\gg S$, however, both $\langle D\rangle$ and σ_D become independent of $\mathbf{f}$ and decrease as $1/N$ [Eq. (25)]. Yet another route is to work with the function $W(\mathbf{q})$ one is interested in. By means of Eq. (35), it is possible to decide whether the term proportional to $1/N$ is only a small correction to $W(\mathbf{f})$ or, on the contrary, the two terms are comparable. In the latter case, more measurements should be carried out.

Although some of the expressions presented here are valid for an arbitrary prior, much of the work deals with the particular case of Eq. (5). The use of a prior that is essentially a linear combination of functions of the form (5) has been proposed [5], specifically, to be used in the inference of entropies. For this case, the partition function should be constructed by applying the same linear superposition to Eq. (10), and the same holds for Eqs. (13)–(19). The calculation of $\langle D\rangle$ and σ_D as derivatives of F is still valid, whereas Eq. (12) should also be averaged.

The analysis of $P(\mathbf{q}|\mathbf{f})$ carried out in Sec. II, and the statistical mechanical description of Sec. III are valid even for small N. The fact that $\langle D\rangle\to 1/N$ for large N inspires the

expansion of $\langle W \rangle$ of Sec. IV. It should be clear, nevertheless, that such an expansion is *only* convergent when $N \gg S$. Actually, Eq. (12) is the first-order term in powers of S/N, and there is no reason to think that the higher-order terms will be negligible, if such a condition does not hold. Moreover, it is necessary to have $Nf_i \gg 1$ for all i. When N is large enough, one can always define the number of categories S as to have them all well populated. But for $N \approx S$ this may well not be the case. The consequences may, in fact, be quite dramatic. For instance, in the example of the entropy (Sec. V A) one can explicitly see that f_i appears in the denominator of Eq. (36). In other words, the result is meaningless if there are empty categories.

However, when the condition $N \gg S$ does hold, Eq. (12) may serve to draw nontrivial conclusions. For instance, it is usually supposed that limited sampling, on average, flaws the data introducing false correlations. This work shows that this is not necessarily the case: limited sampling may sometimes, on average, lower the correlations. This is clear in the simulations of Sec. VI, where finite sampling results, on average, in a downward bias of the mutual information.

ACKNOWLEDGMENTS

I thank Ilya Nemenman for his very useful comments and suggestions. I also thank Alessandro Treves and Stefano Panzeri for a critical reading of the manuscript. This work has been carried out with a grant from Fundación Antorchas.

APPENDIX: INTEGRATING A POWER DISTRIBUTION IN $\mathcal{D}$

Here Eq. (9) is derived. An alternative and more general line of reasoning may be found in [3].

The aim is to calculate

$$I^S_{\mathbf{m}} = \int_{\mathcal{D}} \Pi_{i=1}^S dq_i q_i^{m_i} = \int_0^1 dq_1 q_1^{m_1} \int_0^1 dq_2 q_2^{m_2} \ldots dq_S q_S^{m_S} \delta\left[\lambda_S \left(1 - \sum_{j=1}^S q_j\right)\right], \tag{A1}$$

where λ_S is a constant ensuring that when all m_i vanish, $I^S_{\mathbf{0}}$ is the volume of $\mathcal{D}$. The superscript in $I^S_{\mathbf{m}}$ indicates the dimension of the vectors $\mathbf{m}$ and $\mathbf{q}$.

If X can only take two values, then $S=2$. In this case [11],

$$I^2_{\mathbf{m}} = \int_0^1 dq_1 q_1^{m_1} \int_0^1 dq_2 q_2^{m_2} \delta[\lambda_2(1-q_1-q_2)], = \frac{1}{\lambda_2} \int_0^1 dq_1 q_1^{m_1} (1-q_1)^{m_2} = \frac{1}{\lambda_2} \frac{m_1! m_2!}{(m_1+m_2+1)!}. \tag{A2}$$

Now, the hypothesis is made for arbitrary S,

$$I^S_{\mathbf{m}} = \frac{1}{\lambda_S} \frac{\Pi_{i=1}^S m_i!}{\left(S-1+\sum_{j=1}^S m_j\right)!}. \tag{A3}$$

To prove it, one proceeds by complete induction. Equation (A3) is assumed true for a given $\mathbf{m}=(m_1, \ldots m_S)$ and the aim is to prove it for $(m_i, \ldots, m_{S+1})$. Hence

$$I^{S+1}_{(m_1,\ldots,m_{S+1})} = \int_{\mathcal{D}} (\Pi_{i=1}^{S+1} dq_i q_i^{m_i}) = \frac{\lambda_S}{\lambda_{S+1}} I^{S-1}_{(m_1,m_2,\ldots,m_{S-1})} \int_0^{1-\Sigma_{i=1}^S} dq_S q_S^{m_S} \times \left(1-\sum_{j=1}^S\right)^{m_{S+1}} \Theta\left(1-\sum_{j=1}^S\right) \tag{A4}$$

$$= \frac{\lambda_S}{\lambda_{S+1}} I^S_{(m_1,m_2,\ldots,m_{S-1}),m_S+m_{S+1}+1} \frac{m_S! m_{S+1}!}{(m_S+m_{S+1}+1)!} \tag{A5}$$

$$= \frac{1}{\lambda_{S+1}} \frac{\Pi_{i=1}^{S+1} m_i!}{\left[(S+1)-1+\sum_{j=1}^{S+1} m_j\right]!}, \tag{A6}$$

where $\Theta(x)$ is Heaviside step function: $\Theta(x)=1$ if $x \geqslant 1$ and $\Theta(x)=0$ if $x<0$. When passing from Eq. (A4) to Eq. (A5), use was made of the result (A2). Accordingly, Eq. (A6) derives from the inductive hypothesis (A3). Since Eq. (A6) coincides with Eq. (A3) when S is replaced by $S+1$, the hypothesis (A3) is proved true.

Finally, to determine λ_S one evaluates

$$I^S_{\mathbf{0}} = \frac{1}{\lambda_S (S-1)!}. \tag{A7}$$

The volume of $\mathcal{D}$ is $\sqrt{S}/(S-1)!$, as can be verified, once again, by complete induction. Then $\lambda_S = 1/\sqrt{S}$.

[1] Alessandro Treves and Stefano Panzeri, Neural Comput. **7**, 399 (1995).
[2] Stefano Panzeri and Alessandro Treves, Network **7**, 87 (1996).
[3] David H. Wolpert and David R. Wolf, Phys. Rev. E **52**, 6841 (1995).
[4] S. Kullback, *Information Theory and Statistics* (Dover, New York, 1968).
[5] Ilya Nemenman, Fariel Shafee, and William Bialek, e-print physics/0108025.
[6] H.S. Seung, H. Sompolinsky, and N. Tishby, Phys. Rev. A **45**, 6056 (1992).
[7] William Bialek, Ilya Nemenman, and Naftali Tishby, Neural Comput. **13**, 2409 (2001).
[8] F. Willems, Y. Shtarkov, and T. Tjalkens, IEEE Trans. Inf. Theory **41**, 653 (1995).
[9] T. Schurmann and P. Grassberg, Chaos **6**, 414 (1996).
[10] *Handbook of Mathematical Functions*, edited by M. Abramowitz and I. A. Stegun (Dover, New York, 1972).
[11] I. S. Gradshteyn and I. M. Ryzhik, *Tables of Integrals, Series and Products* (Academic Press, San Diego, 1994).