Entropy and Inference, Revisited

Ilya Nemenman,[1,2] Fariel Shafee,[3] and William Bialek[1,3]
[1]NEC Research Institute, 4 Independence Way, Princeton, New Jersey 08540
[2]Institute for Theoretical Physics, University of California, Santa Barbara, CA 93106
[3]Department of Physics, Princeton University, Princeton, New Jersey 08544
nemenman@itp.ucsb.edu, {fshafee/wbialek}@princeton.edu

Abstract

We study properties of popular near–uniform (Dirichlet) priors for learning undersampled probability distributions on discrete nonmetric spaces and show that they lead to disastrous results. However, an Occam–style phase space argument expands the priors into their infinite mixture and resolves most of the observed problems. This leads to a surprisingly good estimator of entropies of discrete distributions.

Learning a probability distribution from examples is one of the basic problems in data analysis. Common practical approaches introduce a family of parametric models, leading to questions about model selection. In Bayesian inference, computing the total probability of the data arising from a model involves an integration over parameter space, and the resulting "phase space volume" automatically discriminates against models with larger numbers of parameters—hence the description of these volume terms as Occam factors [1, 2]. As we move from finite parameterizations to models that are described by smooth functions, the integrals over parameter space become functional integrals and methods from quantum field theory allow us to do these integrals asymptotically; again the volume in model space consistent with the data is larger for models that are smoother and hence less complex [3]. Further, at least under some conditions the relevant degree of smoothness can be determined self–consistently from the data, so that we approach something like a model independent method for learning a distribution [4].

The results emphasizing the importance of phase space factors in learning prompt us to look back at a seemingly much simpler problem, namely learning a distribution on a discrete, nonmetric space. Here the probability distribution is just a list of numbers $\{q_i\}$, $i = 1, 2, \cdots, K$, where K is the number of bins or possibilities. We do not assume any metric on the space, so that a priori there is no reason to believe that any q_i and q_j should be similar. The task is to learn this distribution from a set of examples, which we can describe as the number of times n_i each possibility is observed in a set of $N = \sum_{i=1}^{K} n_i$ samples. This problem arises in the context of language, where the index i might label words or phrases, so that there is no natural way to place a metric on the space, nor is it even clear that our intuitions about similarity are consistent with the constraints of a metric space. Similarly, in bioinformatics the index i might label n–mers of the the DNA or amino acid sequence, and although most work in the field is based on metrics for sequence comparison one might like an alternative approach that does not rest on such assumptions. In the analysis of neural responses, once we fix our time resolution the response becomes a set of discrete "words," and estimates of the information content in the response are determined by the probability distribution on this discrete space. What all of these examples have in common is that we often need to draw some conclusions with data sets that are *not* in the asymptotic limit $N \gg K$. Thus, while we might use a large corpus to sample the distribution of words in English by brute force (reaching $N \gg K$ with K the size of the vocabulary), we can hardly do the same for three or four word phrases.

In models described by continuous functions, the infinite number of "possibilities" can never be overwhelmed by examples; one is saved by the notion of smoothness. Is there some nonmetric analog of this notion that we can apply in the discrete case? Our intuition is that information theoretic quantities may play this role. If we have a joint distribution of two variables, the analog of a smooth distribution would be one which does not have too much mutual information between these variables. Even more simply, we might say that smooth distributions have large entropy. While the idea of "maximum entropy inference" is common [5], the interplay between constraints on the entropy and the volume in the space of models seems not to have been considered. As we shall explain, phase space factors alone imply that seemingly sensible, more or less uniform priors on the space of discrete probability distributions correspond to disastrously singular prior hypotheses about the entropy of the underlying distribution. We argue that reliable inference outside the asymptotic regime $N \gg K$ requires a more uniform prior on the entropy, and we offer one way of doing this. While many distributions are consistent with the data when $N \leq K$, we provide empirical evidence that this flattening of the entropic prior allows us to make surprisingly reliable statements about the entropy itself in this regime.

At the risk of being pedantic, we state very explicitly what we mean by uniform or nearly uniform priors on the space of distributions. The natural "uniform" prior is given by

$$\mathcal{P}_{\mathrm{u}}(\{q_i\}) = \frac{1}{Z_{\mathrm{u}}}\,\delta\left(1 - \sum_{i=1}^{K} q_i\right), \quad Z_{\mathrm{u}} = \int_{\mathcal{A}} dq_1 dq_2 \cdots dq_K\,\delta\left(1 - \sum_{i=1}^{K} q_i\right) \qquad (1)$$

where the delta function imposes the normalization, Z_{u} is the total volume in the space of models, and the integration domain $\mathcal{A}$ is such that each q_i varies in the range $[0, 1]$. Note that, because of the normalization constraint, an *individual* q_i chosen from this distribution in fact is not uniformly distributed—this is also an example of phase space effects, since in choosing one q_i we constrain all the other $\{q_{j\neq i}\}$. What we mean by uniformity is that all distributions that obey the normalization constraint are equally likely a priori.

Inference with this uniform prior is straightforward. If our examples come independently from $\{q_i\}$, then we calculate the probability of the model $\{q_i\}$ with the usual Bayes rule: [1]

$$P(\{q_i\}|\{n_i\}) = \frac{P(\{n_i\}|\{q_i\})\mathcal{P}_{\mathrm{u}}(\{q_i\})}{P_{\mathrm{u}}(\{n_i\})}, \quad P(\{n_i\}|\{q_i\}) = \prod_{i=1}^{K} (q_i)^{n_i}. \qquad (2)$$

If we want the best estimate of the probability q_i in the least squares sense, then we should compute the conditional mean, and this can be done exactly, so that [6, 7]

$$\langle q_i \rangle = \frac{n_i + 1}{N + K}. \qquad (3)$$

Thus we can think of inference with this uniform prior as setting probabilities equal to the observed frequencies, but with an "extra count" in every bin. This sensible procedure was first introduced by Laplace [8]. It has the desirable property that events which have not been observed are not automatically assigned probability zero.

[1] If the data are unordered, extra combinatorial factors have to be included in $P(\{n_i\}|\{q_i\})$. However, these cancel immediately in later expressions.

A natural generalization of these ideas is to consider priors that have a power–law dependence on the probabilities, the so called Dirichlet family of priors:

$$\mathcal{P}_\beta(\{q_i\}) = \frac{1}{Z(\beta)} \delta\left(1 - \sum_{i=1}^{K} q_i\right) \prod_{i=1}^{K} q_i^{\beta-1}\,, \tag{4}$$

It is interesting to see what typical distributions from these priors look like. Even though different q_i's are not independent random variables due to the normalizing δ–function, generation of random distributions is still easy: one can show that if q_i's are generated successively (starting from $i = 1$ and proceeding up to $i = K$) from the Beta–distribution

$$P(q_i) = B\left(\frac{q_i}{1-\sum_{j<i} q_j}; \beta, (K-i)\beta\right), \quad B\,(x; a, b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}\,, \tag{5}$$

then the probability of the whole sequence $\{q_i\}$ is $\mathcal{P}_\beta(\{q_i\})$. Fig. 1 shows some typical distributions generated this way. They represent different regions of the range of possible entropies: low entropy ($\sim$ 1 bit, where only a few bins have observable probabilities), entropy in the middle of the possible range, and entropy in the vicinity of the maximum, $\log_2 K$. When learning an unknown distribution, we usually have no a priori reason to expect it to look like only one of these possibilities, but choosing β pretty much fixes allowed "shapes." This will be a focal point of our discussion.

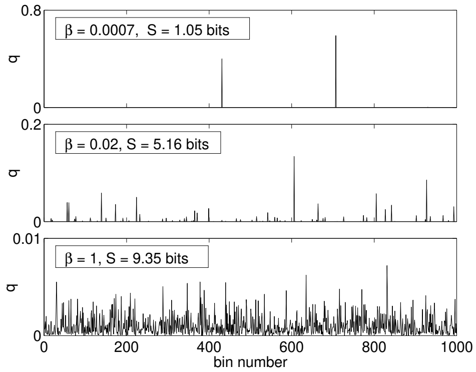

Figure 1: Typical distributions, $K = 1000$.

Even though distributions look different, inference with all priors Eq. (4) is similar [6, 7]:

$$\langle q_i \rangle_\beta = \frac{n_i + \beta}{N + \kappa}\,, \quad \kappa = K\beta. \tag{6}$$

This simple modification of the Laplace's rule, Eq. (3), which allows us to vary probability assigned to the outcomes not yet seen, was first examined by Hardy and Lidstone [9, 10]. Together with the Laplace's formula, $\beta = 1$, this family includes the usual maximum likelihood estimator (MLE), $\beta \to 0$, that identifies probabilities with frequencies, as well as the Jeffreys' or Krichevsky–Trofimov (KT) estimator, $\beta = 1/2$ [11, 12, 13], the Schurmann–Grassberger (SG) estimator, $\beta = 1/K$ [14], and other popular choices.

To understand why inference in the family of priors defined by Eq. (4) is unreliable, consider the entropy of a distribution drawn at random from this ensemble. Ideally we would like to compute this whole a priori distribution of entropies,

$$\mathcal{P}_\beta(S) = \int dq_1 dq_2 \cdots dq_K\, P_\beta(\{q_i\})\, \delta\left[S + \sum_{i=1}^{K} q_i \log_2 q_i\right], \tag{7}$$

but this is quite difficult. However, as noted by Wolpert and Wolf [6], one can compute the moments of $\mathcal{P}_\beta(S)$ rather easily. Transcribing their results to the present notation (and correcting some small errors), we find:

$$\xi(\beta) \equiv \langle\, S[n_i = 0]\,\rangle_\beta \;=\; \psi_0(\kappa+1) - \psi_0(\beta+1)\,, \tag{8}$$

$$\sigma^2(\beta) \equiv \langle\,(\delta S)^2[n_i = 0]\rangle_\beta \;=\; \frac{\beta+1}{\kappa+1}\,\psi_1(\beta+1) - \psi_1(\kappa+1)\,, \tag{9}$$

where $\psi_m(x) = (d/dx)^{m+1} \log_2 \Gamma(x)$ are the polygamma functions.

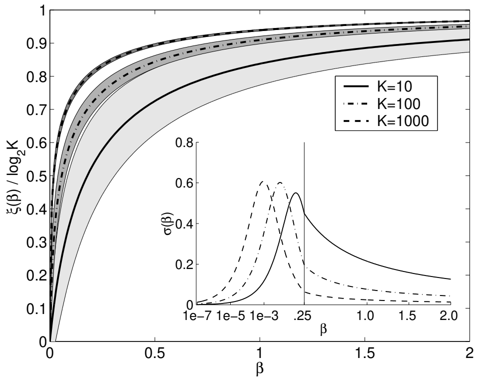

Figure 2: $\xi(\beta)/\log_2 K$ and $\sigma(\beta)$ as functions of β and K; gray bands are the region of $\pm\sigma(\beta)$ around the mean. Note the transition from the logarithmic to the linear scale at $\beta = 0.25$ in the insert.

This behavior of the moments is shown on Fig. 2. We are faced with a striking observation: a priori distributions of entropies in the power–law priors are extremely peaked for even moderately large K. Indeed, as a simple analysis shows, their maximum standard deviation of approximately 0.61 bits is attained at $\beta \approx 1/K$, where $\xi(\beta) \approx 1/\ln 2$ bits. This has to be compared with the possible range of entropies, $[0, \log_2 K]$, which is asymptotically large with K. Even worse, for any fixed β and sufficiently large K, $\xi(\beta) = \log_2 K - O(K^0)$, and $\sigma(\beta) \propto 1/\sqrt{\kappa}$. Similarly, if K is large, but κ is small, then $\xi(\beta) \propto \kappa$, and $\sigma(\beta) \propto \sqrt{\kappa}$. This paints a lively picture: varying β between 0 and ∞ results in a smooth variation of ξ, the a priori expectation of the entropy, from 0 to $S_{\max} = \log_2 K$. Moreover, for large K, the standard deviation of $\mathcal{P}_\beta(S)$ is always negligible relative to the possible range of entropies, and it is negligible even absolutely for $\xi \gg 1$ ($\beta \gg 1/K$). Thus a seemingly innocent choice of the prior, Eq. (4), leads to a disaster: *fixing β specifies the entropy almost uniquely*. Furthermore, the situation persists even after we observe some data: *until the distribution is well sampled, our estimate of the entropy is dominated by the prior!*

Thus it is clear that all commonly used estimators mentioned above have a problem. While they may or may not provide a reliable estimate of the distribution $\{q_i\}$[2], they are definitely a poor tool to learn entropies. Unfortunately, often we are interested precisely in these entropies or similar information–theoretic quantities, as in the examples (neural code, language, and bioinformatics) we briefly mentioned earlier.

Are the usual estimators really this bad? Consider this: for the MLE ($\beta = 0$), Eqs. (8, 9) are formally wrong since it is impossible to normalize $\mathcal{P}_0(\{q_i\})$. However, the prediction that $\mathcal{P}_0(S) = \delta(S)$ still holds. Indeed, $S_{\rm ML}$, the entropy of the ML distribution, is zero even for $N = 1$, let alone for $N = 0$. In general, it is well known that $S_{\rm ML}$ always underestimates the actual value of the entropy, and the correction

$$S = S_{\rm ML} + \frac{K^*}{2N} + O\left(\frac{1}{N^2}\right) \tag{10}$$

is usually used (cf. [14]). Here we must set $K^* = K - 1$ to have an asymptotically correct result. Unfortunately in an undersampled regime, $N \ll K$, this is a disaster. To alleviate the problem, different authors suggested to determine the dependence $K^* = K^*(K)$ by various (rather ad hoc) empirical [15] or pseudo–Bayesian techniques [16]. However, then there is no principled way to estimate both the residual bias and the error of the estimator.

The situation is even worse for the Laplace's rule, $\beta = 1$. We were unable to find any results in the literature that would show a clear understanding of the effects of the prior on the entropy estimate, $S_{\rm L}$. And these effects are enormous: the a priori distribution of the entropy has $\sigma(1) \sim 1/\sqrt{K}$ and is almost δ-like. This translates into a very certain, but nonetheless possibly wrong, estimate of the entropy. We believe that this type of error

[2] In any case, the answer to this question depends mostly on the "metric" chosen to measure reliability. Minimization of bias, variance, or information cost (Kullback–Leibler divergence between the target distribution and the estimate) leads to very different "best" estimators.

(cf. Fig. 3) has been overlooked in some previous literature.

The Schurmann–Grassberger estimator, $\beta = 1/K$, deserves a special attention. The variance of $\mathcal{P}_\beta(S)$ is maximized near this value of β (cf. Fig. 2). Thus the SG estimator results in the most uniform a priori expectation of S possible for the power–law priors, and consequently in the least bias. We suspect that this feature is responsible for a remark in Ref. [14] that this β was empirically the best for studying printed texts. But even the SG estimator is flawed: it is biased towards (roughly) $1/\ln 2$, and it is still a priori rather narrow.

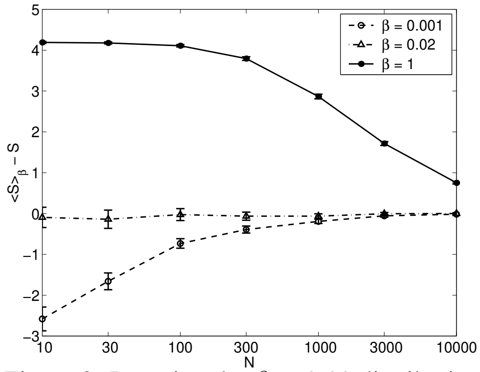

Figure 3: Learning the $\beta = 0.02$ distribution from Fig. 1 with $\beta = 0.001, 0.02, 1$. The actual error of the estimators is plotted; the error bars are the standard deviations of the posteriors. The "wrong" estimators are very certain but nonetheless incorrect.

Summarizing, we conclude that simple power–law priors, Eq. (4), must not be used to learn entropies when there is no strong a priori knowledge to back them up. On the other hand, they are the only priors we know of that allow to calculate $\langle q_i \rangle$, $\langle S \rangle$, $\langle \chi^2 \rangle$, …exactly [6]. Is there a way to resolve the problem of peakedness of $\mathcal{P}_\beta(S)$ without throwing away their analytical ease? One approach would be to use $\mathcal{P}^{\rm flat}_\beta(\{q_i\}) = \frac{\mathcal{P}_\beta(\{q_i\})}{\mathcal{P}_\beta(S[q_i])}\, \mathcal{P}^{\rm actual}(S[q_i])$ as a prior on $\{q_i\}$. This has a feature that the a priori distribution of S deviates from uniformity only due to our actual knowledge $\mathcal{P}^{\rm actual}(S[q_i])$, but not in the way $\mathcal{P}_\beta(S)$ does. However, as we already mentioned, $\mathcal{P}_\beta(S[q_i])$ is yet to be calculated.

Another way to a flat prior is to write $\mathcal{P}(S) = 1 = \int \delta(S - \xi) d\xi$. If we find a family of priors $\mathcal{P}(\{q_i\}, \mathrm{parameters})$ that result in a δ-function over S, and if changing the parameters moves the peak across the whole range of entropies uniformly, we may be able to use this. Luckily, $\mathcal{P}_\beta(S)$ is almost a δ-function! [3] In addition, changing β results in changing $\xi(\beta) = \langle\, S[n_i = 0]\,\rangle_\beta$ across the whole range $[0, \log_2 K]$. So we may hope that the prior [4]

$$\mathcal{P}(\{q_i\}; \beta) = \frac{1}{Z}\, \delta\left(1 - \sum_{i=1}^{K} q_i\right) \prod_{i=1}^{K} q_i^{\beta-1} \frac{d\xi(\beta)}{d\beta}\, \mathcal{P}(\beta) \tag{11}$$

may do the trick and estimate entropy reliably even for small N, and even for distributions that are atypical for any one β. We have less reason, however, to expect that this will give an equally reliable estimator of the atypical distributions themselves.[2] Note the term $d\xi/d\beta$ in Eq. (11). It is there because ξ, not β, measures the position of the entropy density peak.

Inference with the prior, Eq. (11), involves additional averaging over β (or, equivalently,

---

[3] The approximation becomes not so good as $\beta \to 0$ since $\sigma(\beta)$ becomes $O(1)$ before dropping to zero. Even worse, $\mathcal{P}_\beta(S)$ is skewed at small β. This accumulates an extra weight at $S = 0$. Our approach to dealing with these problems is to ignore them while the posterior integrals are dominated by β's that are far away from zero. This was always the case in our simulations, but is an open question for the analysis of real data.

[4] Priors that are formed as weighted sums of the different members of the Dirichlet family are usually called *Dirichlet mixture priors*. They have been used to estimate probability distributions of, for example, protein sequences [17]. Equation (11), an *infinite* mixture, is a further generalization, and, to our knowledge, it has not been studied before.

ξ), but is nevertheless straightforward. The a posteriori moments of the entropy are

$$\widehat{S^m} = \frac{\int d\xi\, \rho(\xi, \{n_i\}) \langle\, S^m[n_i]\,\rangle_{\beta(\xi)}}{\int d\xi\, \rho(\xi, [n_i])}\,, \quad \text{where} \tag{12}$$

$$\rho(\xi, [n_i]) = \mathcal{P}\left(\beta\left(\xi\right)\right) \frac{\Gamma(\kappa(\xi))}{\Gamma(N + \kappa(\xi))} \prod_{i=1}^{K} \frac{\Gamma(n_i + \beta(\xi))}{\Gamma(\beta(\xi))}\,. \tag{13}$$

Here the moments $\langle\, S^m[n_i]\,\rangle_{\beta(\xi)}$ are calculated at fixed β according to the (corrected) formulas of Wolpert and Wolf [6]. We can view this inference scheme as follows: first, one sets the value of β and calculates the expectation value (or other moments) of the entropy at this β. For small N, the expectations will be very close to their a priori values due to the peakedness of $\mathcal{P}_\beta(S)$. Afterwards, one integrates over $\beta(\xi)$ with the density $\rho(\xi)$, which includes our a priori expectations about the entropy of the distribution we are studying [$\mathcal{P}\left(\beta\left(\xi\right)\right)$], as well as the evidence for a particular value of β [Γ-terms in Eq. (13)].

The crucial point is the behavior of the evidence. If it has a pronounced peak at some $\beta_{\rm cl}$, then the integrals over β are dominated by the vicinity of the peak, $\widehat{S}$ is close to $\xi(\beta_{\rm cl})$, and the variance of the estimator is small. In other words, data "selects" some value of β, much in the spirit of Refs. [1] – [4]. However, this scenario may fail in two ways. First, there may be no peak in the evidence; this will result in a very wide posterior and poor inference. Second, the posterior density may be dominated by $\beta \to 0$, which corresponds to MLE, the best possible fit to the data, and is a discrete analog of overfitting. While all these situations are possible, we claim that generically the evidence is well–behaved. Indeed, while small β increases the fit to the data, it also increases the phase space volume of all allowed distributions and thus decreases probability of each particular one [remember that $\langle q_i \rangle_\beta$ has an extra β counts in each bin, thus distributions with $q_i < \beta/(N+\kappa)$ are strongly suppressed]. The fight between the "goodness of fit" and the phase space volume should then result in some non–trivial β_{cl}, set by factors $\propto N$ in the exponent of the integrand.

Figure 4 shows how the prior, Eq. (11), performs on some of the many distributions we tested. The left panel describes learning of distributions that are typical in the prior $\mathcal{P}_\beta(\{q_i\})$ and, therefore, are also likely in $\mathcal{P}(\{q_i\};\beta)$. Thus we may expect a reasonable performance, but the real results exceed all expectations: for all three cases, the actual relative error drops to the 10% level at N as low as 30 (recall that $K = 1000$, so we only have ~ 0.03 data points per bin on average)! To put this in perspective, simple estimates like fixed β ones, MLE, and MLE corrected as in Eq. (10) with K^* equal to the number of nonzero n_i's produce an error so big that it puts them off the axes until $N > 100$. [5] Our results have two more nice features: the estimator seems to know its error pretty well, and it is almost completely unbiased.

One might be puzzled at how it is possible to estimate anything in a 1000–bin distribution with just a few samples: the distribution is completely unspecified for low N! The point is that we are not trying to learn the distribution — in the absence of additional prior information this would, indeed, take $N \gg K$ — but to estimate just one of its characteristics. It is less surprising that one number can be learned well with only a handful of measurements. In practice the algorithm builds its estimate based on the number of coinciding samples (multiple coincidences are likely only for small β), as in the Ma's approach to entropy estimation from simulations of physical systems [18].

What will happen if the algorithm is fed with data from a distribution $\{\tilde{q}_i\}$ that is strongly atypical in $\mathcal{P}(\{q_i\};\beta)$? Since there is no $\{\tilde{q}_i\}$ in our prior, its estimate may suffer. Nonetheless, for any $\{\tilde{q}_i\}$, there is some β which produces distributions with the same mean entropy as $S[\tilde{q}_i]$. Such β should be determined in the usual fight between the "goodness of fit" and

[5] More work is needed to compare our estimator to more complex techniques, like in Ref. [15, 16].

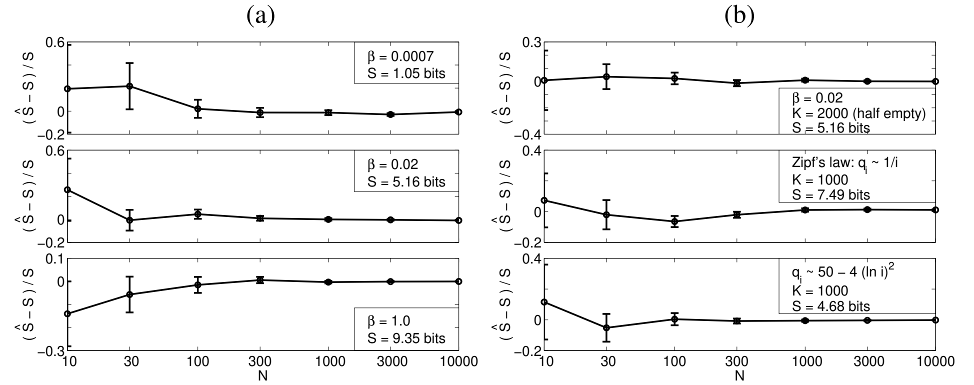

Figure 4: Learning entropies with the prior Eq. (11) and $\mathcal{P}(\beta) = 1$. The actual relative errors of the estimator are plotted; the error bars are the relative widths of the posteriors. (a) Distributions from Fig. 1. (b) Distributions atypical in the prior. Note that while $\widehat{S}$ may be safely calculated as just $\langle S \rangle_{\beta_{\rm cl}}$, one has to do an honest integration over β to get $\widehat{S^2}$ and the error bars. Indeed, since $\mathcal{P}_\beta(S)$ is almost a δ-function, the uncertainty at any fixed β is very small (see Fig. 3).

the Occam factors, and the correct value of entropy will follow. However, there will be an important distinction from the "correct prior" cases. The value of β indexes available phase space volumes, and thus the smoothness (complexity) of the model class [19]. In the case of discrete distributions, smoothness is the absence of high peaks. Thus data with faster decaying Zipf plots (plots of bins' occupancy vs. occupancy rank i) are rougher. The priors $\mathcal{P}_\beta(\{q_i\})$ cannot account for all possible roughnesses. Indeed, they only generate distributions for which the expected number of bins ν with the probability mass less than some q is given by $\nu(q) = KB(q, \beta, \kappa - \beta)$, where B is the familiar incomplete Beta function, as in Eq. (5). This means that the expected rank ordering for small and large ranks is

$$q_i \approx 1 - \left[\frac{\beta B(\beta, \kappa - \beta)(K-1)\, i}{K}\right]^{1/(\kappa - \beta)}, \quad i \ll K, \tag{14}$$

$$q_i \approx \left[\frac{\beta B(\beta, \kappa - \beta)(K - i + 1)}{K}\right]^{1/\beta}, \quad K - i + 1 \ll K. \tag{15}$$

In an undersampled regime we can observe only the first of the behaviors. Therefore, any distribution with q_i decaying faster (rougher) or slower (smoother) than Eq. (14) for some β cannot be explained well with fixed $\beta_{\rm cl}$ for different N. So, unlike in the cases of learning data that are typical in $\mathcal{P}_\beta(\{q_i\})$, we should expect to see $\beta_{\rm cl}$ growing (falling) for qualitatively smoother (rougher) cases as N grows.

Figure 4(b) and Tbl. 1 illustrate these points. First, we study the $\beta = 0.02$ distribution from Fig. 1. However, we added a 1000 extra bins, each with $q_i = 0$. Our estimator performs remarkably well, and $\beta_{\rm cl}$ does not drift because the ranking law remains the same. Then we turn to the famous Zipf's distribution, so common in Nature. It has $n_i \propto 1/i$, which is qualitatively smoother than our prior allows. Correspondingly, we get an upwards drift in $\beta_{\rm cl}$. Finally, we analyze a "rough" distribution, which has $q_i \propto 50 - 4(\ln i)^2$, and $\beta_{\rm cl}$ drifts downwards. Clearly, one would want to predict the dependence $\beta_{\rm cl}(N)$ analytically, but this requires calculation of the predictive information (complexity) for the

N	1/2 full	Zipf	rough
units	$\cdot 10^{-2}$	$\cdot 10^{-1}$	$\cdot 10^{-3}$
10	1.7	1907	16.8
30	2.2	0.99	11.5
100	2.4	0.86	12.9
300	2.2	1.36	8.3
1000	2.1	2.24	6.4
3000	1.9	3.36	5.4
10000	2.0	4.89	4.5

Table 1: $\beta_{\rm cl}$ for solutions shown on Fig. 4(b).

involved distributions [19] and is a work for the future. Notice that, the entropy estimator for atypical cases is almost as good as for typical ones. A possible exception is the 100–1000 points for the Zipf distribution—they are about two standard deviations off. We saw similar effects in some other "smooth" cases also. This may be another manifestation of an observation made in Ref. [4]: smooth priors can easily adapt to rough distribution, but there is a limit to the smoothness beyond which rough priors become inaccurate.

To summarize, an analysis of a priori entropy statistics in common power–law Bayesian estimators revealed some very undesirable features. We are fortunate, however, that these minuses can be easily turned into pluses, and the resulting estimator of entropy is precise, knows its own error, and gives amazing results for a very large class of distributions.

Acknowledgements

We thank Vijay Balasubramanian, Curtis Callan, Adrienne Fairhall, Tim Holy, Jonathan Miller, Vipul Periwal, Steve Strong, and Naftali Tishby for useful discussions. I. N. was supported in part by NSF Grant No. PHY99-07949 to the Institute for Theoretical Physics.

References

[1] D. MacKay, *Neural Comp.* **4**, 415–448 (1992).

[2] V. Balasubramanian, *Neural Comp.* **9**, 349–368 (1997).

[3] W. Bialek, C. Callan, and S. Strong, *Phys. Rev. Lett.* **77**, 4693–4697 (1996).

[4] I. Nemenman and W. Bialek, *Advances in Neural Inf. Processing Systems* **13**, 287–293 (2001).

[5] J. Skilling, in *Maximum entropy and Bayesian methods,* J. Skilling ed. (Kluwer Academic Publ., Amsterdam, 1989), pp. 45–52.

[6] D. Wolpert and D. Wolf, *Phys. Rev. E* **52**, 6841–6854 (1995).

[7] I. Nemenman, Ph.D. Thesis, Princeton, (2000), ch. 3, http://arXiv.org/abs/physics/0009032.

[8] P. de Laplace, marquis de, *Essai philosophique sur les probabilités* (Courcier, Paris, 1814), trans. by F. Truscott and F. Emory, *A philosophical essay on probabilities* (Dover, New York, 1951).

[9] G. Hardy, *Insurance Record* (1889), reprinted in *Trans. Fac. Actuaries* **8** (1920).

[10] G. Lidstone, *Trans. Fac. Actuaries* **8**, 182–192 (1920).

[11] H. Jeffreys, *Proc. Roy. Soc. (London) A* **186**, 453–461 (1946).

[12] R. Krichevskii and V. Trofimov, *IEEE Trans. Inf. Thy.* **27**, 199–207 (1981).

[13] F. Willems, Y. Shtarkov, and T. Tjalkens, *IEEE Trans. Inf. Thy.* **41**, 653–664 (1995).

[14] T. Schurmann and P. Grassberger, *Chaos* **6**, 414–427 (1996).

[15] S. Strong, R. Koberle, R. de Ruyter van Steveninck, and W. Bialek, *Phys. Rev. Lett.* **80**, 197–200 (1998).

[16] S. Panzeri and A. Treves, *Network: Comput. in Neural Syst.* **7**, 87–107 (1996).

[17] K. Sjlander, K. Karplus, M. Brown, R. Hughey, A. Krogh, I. S. Mian, and D. Haussler, *Computer Applications in the Biosciences (CABIOS)* **12**, 327–345 (1996).

[18] S. Ma, *J. Stat. Phys.* **26**, 221 (1981).

[19] W. Bialek, I. Nemenman, N. Tishby, *Neural Comp.* **13**, 2409-2463 (2001).