Entropy and Inference, Revisited

Ilya Nemenman,1,2 Fariel Shafee,3 and William Bialek1,3
1NEC Research Institute, 4 Independence Way, Princeton, New Jersey 08540
2Institute for Theoretical Physics, University of California, Santa Barbara, CA 93106
3Department 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 \( \{q_i\} \), \( i = 1, 2, \ldots, K \), 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 de-
terminated 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

$$ P_u(\{q_i\}) = \frac{1}{Z_u} \delta \left(1 - \sum_{i=1}^{K} q_i\right), \quad Z_u = \int_{\mathcal{A}} dq_1 dq_2 \cdots dq_K \delta \left(1 - \sum_{i=1}^{K} q_i\right) \tag{1} $$

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:

$$ P(\{q_i\}|\{n_i\}) = \frac{P(\{n_i\}|\{q_i\}) P_u(\{q_i\})}{P_u(\{n_i\})}, \quad P(\{n_i\}|\{q_i\}) = \prod_{i=1}^{K} (q_i)^{n_i}. \tag{2} $$

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]

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

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.