Bayesian entropy estimation for infinite neural alphabets

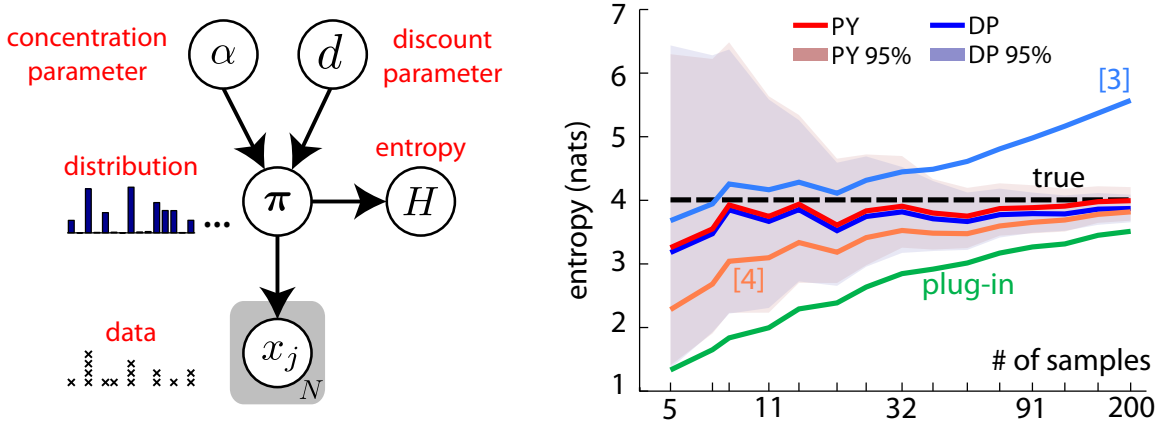
Evan Archer, Il Memming Park, & Jonathan Pillow

Abstract:
Shannon entropy quantifies the information that may be conveyed by a vector of neural responses, and has featured prominently in the analysis of neural codes. However, entropy is notoriously difficult to estimate from data, particularly in the "undersampled" regime, where the number of possible response patterns (or "words") K is larger than the number of observed responses ("samples") N. Here we describe a Bayesian method for estimating entropy from datasets where the number of possible words (i.e., the size of the neural alphabet) is arbitrarily large or unknown. Our approach follows that of Nemenman *et al* [1], who formulated a Bayesian entropy estimator using a "mixture-of-Dirichlets" prior over the space of discrete distributions on K bins. Here we extend this approach in several directions. First, we formulate two priors over discrete, countably infinite distributions using mixtures of Dirichlet processes (DP) or Pitman-Yor (PY) processes [2]. (These processes play a central role in nonparametric Bayesian statistics, and are useful when the number of parameters or "bins" is not known *a priori*.) We analytically derive a set of mixing weights over these processes so that the resulting improper prior over entropy is approximately flat across a semi-infinite range. Secondly, we consider the posterior over entropy given a dataset (which contains some observed number of words but an unknown number of unobserved words), and show that the posterior mean can be efficiently computed via a simple 1D or 2D numerical integral. Remarkably, for most datasets the expected entropy given data is finite, even though the distributions have positive probability on infinitely many bins and the prior is improper. We compare our approach to previous methods, including an approximate Bayesian entropy estimator [3] and a frequentist "coverage-adjusted" estimator designed for unknown or infinite K [4].

Additional Details: The PY process specifies a prior distribution over discrete probability distributions $\boldsymbol{\pi}$ (where $\sum_{i=1}^{\infty} \boldsymbol{\pi}_i = 1$, and $\boldsymbol{\pi}_i \geq 0$) with parameters α and d which specify the concentration and tail behavior of likely distributions, respectively. The DP is a special case where $d = 0$. Following [1], we specify a prior over PY parameters, $p(\alpha, d)$, such that the resulting mixture prior over distributions, $p(\boldsymbol{\pi}) = \iint p(\boldsymbol{\pi}|\alpha, d)p(\alpha, d)d\alpha\, dd$, is approximately flat over entropy, $H(\boldsymbol{\pi}) = -\sum_{i=1}^{\infty} \boldsymbol{\pi}_i \log \boldsymbol{\pi}_i$. Given this prior and data $\mathbf{x}$ drawn from an unknown discrete distribution, we compute the posterior mean (the Bayes' least squares entropy estimate) through numerical evaluation of the integral:

$$\hat{H} = \mathbb{E}[H|\mathbf{x}] = \iint \mathbb{E}\left[H|\alpha, d\right] p(\alpha, d|\mathbf{x})p(\alpha, d)d\alpha\, dd.$$

This relies on a closed-form analytic expression for $E[H|\alpha, d] = \int H(\boldsymbol{\pi})p(\boldsymbol{\pi}|\alpha, d)d\boldsymbol{\pi}$, which we derive. A graphical model and comparison of estimators using samples from a $\mathcal{PY}(0.5, 5)$ is shown below.



References: [1] Nemenman, Shafee & Bialek 2001. [2] Pitman & Yor 1997. [3] Nemenman, Bialek & de Ruyter van Steveninck, R. 2004. [4] Vu, Yu, & Kass 2007.