Bayesian estimation of discrete entropy with mixtures of stick-breaking priors

Evan Archer[*124], **Il Memming Park**[*234], **& Jonathan W. Pillow**[234]
1. Institute for Computational and Engineering Sciences
2. Center for Perceptual Systems, 3. Dept. of Psychology,
4. Division of Statistics & Scientific Computation
The University of Texas at Austin

Abstract

We consider the problem of estimating Shannon's entropy H in the under-sampled regime, where the number of possible symbols may be unknown or countably infinite. Dirichlet and Pitman-Yor processes provide tractable prior distributions over the space of countably infinite discrete distributions, and have found major applications in Bayesian non-parametric statistics and machine learning. Here we show that they provide natural priors for Bayesian entropy estimation, due to the analytic tractability of the moments of the induced posterior distribution over entropy H. We derive formulas for the posterior mean and variance of H given data. However, we show that a fixed Dirichlet or Pitman-Yor process prior implies a narrow prior on H, meaning the prior strongly determines the estimate in the under-sampled regime. We therefore define a family of continuous mixing measures such that the resulting mixture of Dirichlet or Pitman-Yor processes produces an approximately flat prior over H. We explore the theoretical properties of the resulting estimators and show that they perform well on data sampled from both exponential and power-law tailed distributions.

1 Introduction

An important statistical problem in the study of natural systems is to estimate the entropy of an unknown discrete distribution on the basis of an observed sample. This is often much easier than the problem of estimating the distribution itself; in many cases, entropy can be accurately estimated with fewer samples than the number of distinct symbols. Entropy estimation remains a difficult problem, however, as there is no unbiased estimator for entropy, and the maximum likelihood estimator exhibits severe bias for small datasets. Previous work has tended to focus on methods for computing and reducing this bias [1–5]. Here, we instead take a Bayesian approach, building on a framework introduced by Nemenman *et al* [6]. The basic idea is to place a prior over the space of probability distributions that might have generated the data, and then perform inference using the induced posterior distribution over entropy. (See Fig. 1).

We focus on the setting where our data are a finite sample from an unknown, or possibly even countably infinite, number of symbols. A Bayesian approach requires us to consider distributions over the infinite-dimensional simplex, Δ_∞. To do so, we employ the Pitman-Yor (PYP) and Dirichlet (DP) processes [7–9]. These processes provide an attractive family of priors for this problem, since: (1) the posterior distribution over entropy has analytically tractable moments; and (2) distributions drawn from a PYP can exhibit power-law tails, a feature commonly observed in data from social, biological, and physical systems [10–12]. However, we show that a fixed PYP prior imposes a narrow

[*] These authors contributed equally.

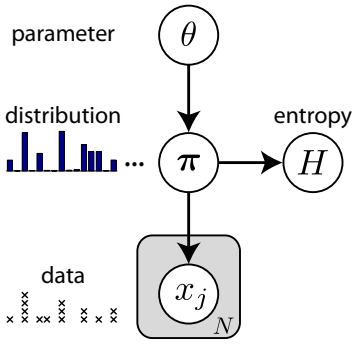

Figure 1: Graphical model illustrating the ingredients for Bayesian entropy estimation. Arrows indicate conditional dependencies between variables, and the gray "plate" denotes multiple copies of a random variable (with the number of copies N indicated at bottom). For entropy estimation, the joint probability distribution over entropy H, data $\mathbf{x} = \{x_j\}$, discrete distribution $\boldsymbol{\pi} = \{\pi_i\}$, and parameter θ factorizes as: $p(H, \mathbf{x}, \boldsymbol{\pi}, \theta) = p(H|\boldsymbol{\pi})p(\mathbf{x}|\boldsymbol{\pi})p(\boldsymbol{\pi}|\theta)p(\theta)$. Entropy is a deterministic function of $\boldsymbol{\pi}$, so $p(H|\boldsymbol{\pi}) = \delta(H - \sum_i \pi_i \log \pi_i)$.

prior over entropy, leading to severe bias and overly narrow credible intervals for small datasets. We address this shortcoming by introducing a set of mixing measures such that the resulting Pitman-Yor Mixture (PYM) prior provides an approximately non-informative (i.e., flat) prior over entropy.

The remainder of the paper is organized as follows. In Section 2, we introduce the entropy estimation problem and review prior work. In Section 3, we introduce the Dirichlet and Pitman-Yor processes and discuss key mathematical properties relating to entropy. In Section 4, we introduce a novel entropy estimator based on PYM priors and derive several of its theoretical properties. In Section 5, we show applications to data.

2 Entropy Estimation

Consider samples $\mathbf{x} := \{x_j\}_{j=1}^N$ drawn *iid* from an unknown discrete distribution $\boldsymbol{\pi} := \{\pi_i\}_{i=1}^{\mathcal{A}}$ on a finite or (countably) infinite alphabet $\mathcal{X}$. We wish to estimate the entropy of $\boldsymbol{\pi}$,

$$H(\boldsymbol{\pi}) = -\sum_{i=1}^{\mathcal{A}} \pi_i \log \pi_i, \tag{1}$$

where we identify $\mathcal{X} = \{1, 2, \ldots, \mathcal{A}\}$ as the set of alphabets without loss of generality (where the alphabet size $\mathcal{A}$ may be infinite), and $\pi_i > 0$ denotes the probability of observing symbol i. We focus on the setting where $N \ll \mathcal{A}$.

A reasonable first step toward estimating H is to estimate the distribution $\boldsymbol{\pi}$. The sum of observed counts $n_k = \sum_{i=1}^N \mathbf{1}_{\{x_i=k\}}$ for each $k \in \mathcal{X}$ yields the empirical distribution $\hat{\boldsymbol{\pi}}$, where $\hat{\pi}_k = n_k/N$. Plugging this estimate for $\boldsymbol{\pi}$ into eq. 1, we obtain the so-called "plugin" estimator: $\hat{H}_{\text{plugin}} = -\sum \hat{\pi}_i \log \hat{\pi}_i$, which is also the maximum-likelihood estimator. It exhibits substantial negative bias in the undersampled regime.

2.1 Bayesian entropy estimation

The Bayesian approach to entropy estimation involves formulating a prior over distributions $\boldsymbol{\pi}$, and then turning the crank of Bayesian inference to infer H using the posterior distribution. Bayes' least squares (BLS) estimators take the form:

$$\hat{H}(\mathbf{x}) = \mathbb{E}[H|\mathbf{x}] = \int H(\boldsymbol{\pi}) p(\boldsymbol{\pi}|\mathbf{x}) \, \mathrm{d}\boldsymbol{\pi} \tag{2}$$

where $p(\boldsymbol{\pi}|\mathbf{x})$ is the posterior over $\boldsymbol{\pi}$ under some prior $p(\boldsymbol{\pi})$ and *categorical* likelihood $p(\mathbf{x}|\boldsymbol{\pi}) = \prod_j p(x_j|\boldsymbol{\pi})$, where $p(x_j = i) = \pi_i$. The conditional $p(H|\boldsymbol{\pi}) = \delta(H - \sum_i \pi_i \log \pi_i)$, since H is deterministically related to $\boldsymbol{\pi}$. To the extent that $p(\boldsymbol{\pi})$ expresses our true prior uncertainty over the unknown distribution that generated the data, this estimate is optimal in a least-squares sense, and the corresponding credible intervals capture our uncertainty about H given the data.

For distributions with known finite alphabet size $\mathcal{A}$, the Dirichlet distribution provides an obvious choice of prior due to its conjugacy to the discrete (or multinomial) likelihood. It takes the form $p(\boldsymbol{\pi}) \propto \prod_{i=1}^{\mathcal{A}} \pi_i^{\alpha-1}$, for $\boldsymbol{\pi}$ on the $\mathcal{A}$-dimensional simplex ($\pi_i \geq 1$, $\sum \pi_i = 1$), with concentration

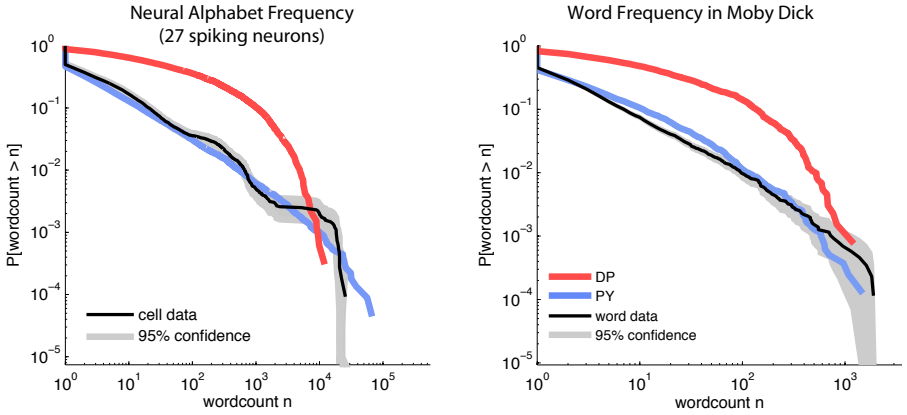

Figure 2: Power-law frequency distributions from neural signals and natural language. We compare samples from the DP (red) and PYP (blue) priors for two datasets with heavy tails (black). In both cases, we compare the empirical CDF with distributions sampled given d and α fixed to their ML estimates. For both datasets, the PYP better captures the heavy-tailed behavior of the data. **Left:** Frequencies among $N = 1.2e6$ neural spike words from 27 simultaneously-recorded retinal ganglion cells, binarized and binned at 10 ms [18]. **Right:** Frequency of $N = 217826$ words in the novel Moby Dick by Herman Melville.

parameter α [13]. Many previously proposed estimators can be viewed as Bayesian estimators with a particular fixed choice of α. (See [14] for an overview).

2.2 Nemenman-Shafee-Bialek (NSB) estimator

In a seminal paper, Nemenman *et al* [6] showed that Dirichlet priors impose a narrow prior over entropy. In the under-sampled regime, Bayesian estimates using a fixed Dirichlet prior are severely biased, and have small credible intervals (i.e., they give highly confident wrong answers!). To address this problem, [6] suggested a mixture-of-Dirichlets prior:

$$p(\boldsymbol{\pi}) = \int p_{\mathrm{Dir}}(\boldsymbol{\pi}|\alpha)p(\alpha)d\alpha, \tag{3}$$

where $p_{\mathrm{Dir}}(\boldsymbol{\pi}|\alpha)$ denotes a $\mathrm{Dir}(\alpha)$ prior on $\boldsymbol{\pi}$. To construct an approximately flat prior on entropy, [6] proposed the mixing weights on α given by,

$$p(\alpha) \propto \frac{d}{d\alpha}\mathbb{E}[H|\alpha] = \mathcal{A}\psi_1(\mathcal{A}\alpha + 1) - \psi_1(\alpha + 1), \tag{4}$$

where $\mathbb{E}[H|\alpha]$ denotes the expected value of H under a $\mathrm{Dir}(\alpha)$ prior, and $\psi_1(\cdot)$ denotes the trigamma function. To the extent that $p(H|\alpha)$ resembles a delta function, eq. 3 implies a uniform prior for H on $[0, \log \mathcal{A}]$.The BLS estimator under the NSB prior can then be written as,

$$\hat{H}_{nsb} = \mathbb{E}[H|\mathbf{x}] = \iint H(\boldsymbol{\pi})p(\boldsymbol{\pi}|\mathbf{x},\alpha)\, p(\alpha|\mathbf{x})\, \mathrm{d}\boldsymbol{\pi}\, \mathrm{d}\alpha = \int \mathbb{E}[H|\mathbf{x},\alpha]\frac{p(\mathbf{x}|\alpha)p(\alpha)}{p(\mathbf{x})}\, d\alpha, \tag{5}$$

where $E[H|\mathbf{x},\alpha]$ is the posterior mean under a $\mathrm{Dir}(\alpha)$ prior, and $p(\mathbf{x}|\alpha)$ denotes the evidence, which has a Polya distribution. Given analytic expressions for $E[H|\mathbf{x},\alpha]$ and $p(\mathbf{x}|\alpha)$, this estimate is extremely fast to compute via 1D numerical integration in α. (See Appendix for details).

Next, we shall consider the problem of extending this approach to infinite-dimensional discrete distributions. Nemenman *et al* proposed one such extension using an approximation to $\hat{H}_{nsb}$ in the limit $\mathcal{A} \to \infty$,which we refer to as $\hat{H}_{nsb_\infty}$ [15, 16]. Unfortunately, $\hat{H}_{nsb_\infty}$ increases unboundedly with N (as noted by [17]), and it performs poorly for the examples we consider.

3 Stick-Breaking Priors

To construct a prior over countably infinite discrete distributions we employ a class of distributions from nonparametric Bayesian statistics known as *stick-breaking* processes [19]. In particular, we

focus on two well-known subclasses of stick-breaking processes: the Dirichlet Process (DP) and Pitman-Yor process (PYP). Both are stochastic processes whose samples are discrete probability distributions [7, 20]. A sample from a DP or PYP may be written as $\sum_{i=1}^{\infty} \pi_i \delta_{\phi_i}$, where $\boldsymbol{\pi} = \{\pi_i\}$ denotes a countably infinite set of 'weights' on a set of atoms $\{\phi_i\}$ drawn from some base probability measure, where δ_{ϕ_i} denotes a delta function on the atom ϕ_i.[1] The prior distribution over $\boldsymbol{\pi}$ under the DP and PYP is technically called the GEM distribution or the two-parameter Poisson-Dirichlet distribution, but we will abuse terminology and refer to it more simply as script notation $\mathcal{DP}$ or $\mathcal{PY}$. The DP weight distribution $\mathcal{DP}(\alpha)$ may be described as a limit of the finite Dirichlet distributions where the alphabet size grows and concentration parameter shrinks, $\mathcal{A} \to \infty$ and $\alpha' \to 0$, such that $\frac{\alpha'}{\mathcal{A}} \to \alpha$ [20]. The PYP generalizes the DP to allow power-law tails, and includes DP as a special case [7].

Let $\mathcal{PY}(d, \alpha)$ denote the PYP weight distribution with *discount* parameter d and *concentration* parameter α (also called the "Dirichlet parameter"), for $d \in [0, 1), \alpha > -d$. When $d = 0$, this reduces to the DP weight distribution, denoted $\mathcal{DP}(\alpha)$. The name "stick-breaking" refers to the fact that the weights of the DP and PYP can be sampled by transforming an infinite sequence of independent Beta random variables in a procedure known as "stick-breaking" [21]. Stick-breaking provides samples $\boldsymbol{\pi} \sim \mathcal{PY}(d, \alpha)$ according to:

$$\beta_i \sim \text{Beta}(1 - d, \alpha + id) \qquad \tilde{\pi}_i = \prod_{k=1}^{i-1} (1 - \beta_k)\beta_i, \tag{6}$$

where $\tilde{\pi}_i$ is known as the i'th *size-biased sample* from $\boldsymbol{\pi}$. (The $\tilde{\pi}_i$ sampled in this manner are not strictly decreasing, but decrease on average such that $\sum_{i=1}^{\infty} \tilde{\pi}_i = 1$ with probability 1). Asymptotically, the tails of a (sorted) sample from $\mathcal{DP}(\alpha)$ decay exponentially, while for $\mathcal{PY}(d, \alpha)$ with $d \neq 0$, the tails approximately follow a power-law: $\pi_i \propto (i)^{-\frac{1}{d}}$ ([7], pp. 867)[2]. Many natural phenomena such as city size, language, spike responses, etc., also exhibit power-law tails [10, 12]. (See Fig. 2).

3.1 Expectations over DP and PY weight distributions

A key virtue of PYP priors is a mathematical property called *invariance under size-biased sampling*, which allows us to convert expectations over $\boldsymbol{\pi}$ on the infinite-dimensional simplex to one or two-dimensional integrals with respect to the distribution of the first two size-biased samples [23, 24]. These expectations are required for computing the mean and variance of H under the prior (or posterior) over $\boldsymbol{\pi}$.

Proposition 1 (Expectations with first two size-biased samples)**.** *For $\boldsymbol{\pi} \sim \mathcal{PY}(d, \alpha)$ and arbitrary integrable functionals f and g of $\boldsymbol{\pi}$,*

$$\mathbb{E}_{(\boldsymbol{\pi}|d,\alpha)}\left[\sum_{i=1}^{\infty} f(\pi_i)\right] = \mathbb{E}_{(\tilde{\pi}_1|d,\alpha)}\left[\frac{f(\tilde{\pi}_1)}{\tilde{\pi}_1}\right], \tag{7}$$

$$\mathbb{E}_{(\boldsymbol{\pi}|d,\alpha)}\left[\sum_{i,j\neq i} g(\pi_i, \pi_j)\right] = \mathbb{E}_{(\tilde{\pi}_1,\tilde{\pi}_2|d,\alpha)}\left[g(\tilde{\pi}_1, \tilde{\pi}_2)(1 - \tilde{\pi}_1)\right], \tag{8}$$

where $\tilde{\pi}_1$ and $\tilde{\pi}_2$ are the first two size-biased samples from $\boldsymbol{\pi}$.

The first result (eq. 7) appears in [7], and we construct an analogous proof for eq. 8 (see Appendix). The direct consequence of this lemma is that first two moments of $H(\boldsymbol{\pi})$ under the $\mathcal{DP}$ and $\mathcal{PY}$ priors have closed forms , which can be obtained using (from eq. 6): $\tilde{\pi}_1 \sim \text{Beta}(1 - d, \alpha + d)$, and $\tilde{\pi}_2/(1-\tilde{\pi}_1)|\tilde{\pi}_1 \sim \text{Beta}(1-d, \alpha+2d)$, with $f(\pi_i) = -\pi_i \log(\pi_i)$ for $\mathbb{E}[H]$, and $f(\pi_i) = \pi_i^2(\log \pi_i)^2$ and $g(\pi_i, \pi_j) = \pi_i\pi_j(\log \pi_i)(\log \pi_j)$ for $\mathbb{E}[H^2]$.

[1] Here, we will assume the base measure is non-atomic, so that the atoms ϕ_i are distinct with probability 1. This allows us to ignore the base measure, making entropy of the distribution equal to the entropy of the weights $\boldsymbol{\pi}$.

[2] Note that the power-law exponent is given incorrectly in [9, 22].

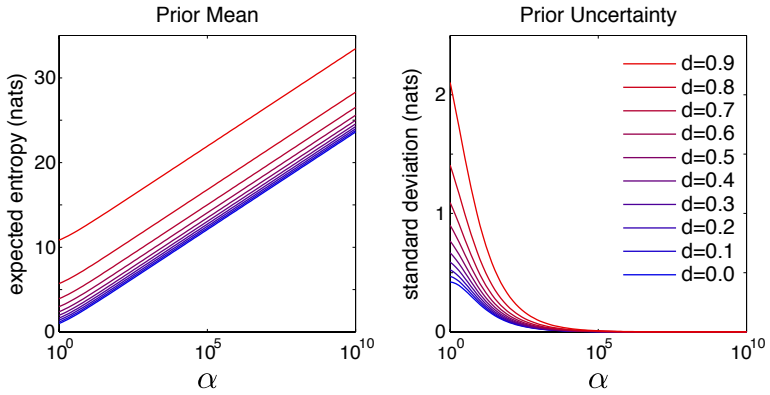

Figure 3: Prior mean and standard deviation over entropy H under a fixed PY prior, as a function of α and d. Note that expected entropy is approximately linear in $\log \alpha$. Small prior standard deviations (right) indicate that $p(H(\boldsymbol{\pi})|d, \alpha)$ is highly concentrated around the prior mean (left).

3.2 Posterior distribution over weights

A second desirable property of the $\mathcal{PY}$ distribution is that the posterior $p(\boldsymbol{\pi}_{\text{post}}|\mathbf{x}, d, \alpha)$ takes the form of a (finite) Dirichlet mixture of point masses and a $\mathcal{PY}$ distribution [8]. This makes it possible to apply the above results to the posterior mean and variance of H.

Let n_i denote the count of symbol i in an observed dataset. Then let $\alpha_i = n_i - d$, $N = \sum n_i$, and $A = \sum \alpha_i = \sum_i n_i - Kd = N - Kd$, where $K = \sum_{i=1}^{\mathcal{A}} \mathbf{1}_{\{n_i>0\}}$ is the number of unique symbols observed. Given data, the posterior over (countably infinite) discrete distributions, written as $\boldsymbol{\pi}_{\text{post}} = (p_1, p_2, p_3, \dots, p_K, p_*\boldsymbol{\pi})$, has the distribution (given in [19]):

$$\begin{aligned}(p_1, p_2, p_3, \dots, p_K, p_*) &\sim \text{Dir}(n_1 - d, n_2 - d, \dots, n_K - d, \alpha + Kd) \\ \boldsymbol{\pi} := (\pi_1, \pi_2, \pi_3, \dots) &\sim \mathcal{PY}(d, \alpha + Kd).\end{aligned} \tag{9}$$

4 Bayesian entropy inference with PY priors

4.1 Fixed PY priors

Using the results of the previous section (eqs. 7 and 8), we can derive the prior mean and variance of H under a $\mathcal{PY}(d, \alpha)$ prior on $\boldsymbol{\pi}$:

$$\mathbb{E}[H(\boldsymbol{\pi})|d, \alpha] = \psi_0(1 + \alpha) - \psi_0(1 - d), \tag{10}$$

$$\text{var}[H(\boldsymbol{\pi})|d, \alpha] = \frac{\alpha + d}{(1 + \alpha)^2(1 - d)} + \frac{1 - d}{1 + \alpha}\psi_1(2 - d) - \psi_1(2 + \alpha), \tag{11}$$

where ψ_n is the polygamma of n-th order (i.e., ψ_0 is the digamma function). Fig. 3 shows these functions for a range of d and α values. These reveal the same phenomenon that [6] observed for finite Dirichlet distributions: a $\mathcal{PY}$ prior with fixed (d, α) induces a narrow prior over H. In the undersampled regime, Bayesian estimates under $\mathcal{PY}$ priors will therefore be strongly determined by the choice of (d, α), and posterior credible intervals will be unrealistically narrow.[3]

4.2 Pitman-Yor process mixture (PYM) prior

The narrow prior on H induced by fixed $\mathcal{PY}$ priors suggests a strategy for constructing a non-informative prior: mix together a family of $\mathcal{PY}$ distributions with some hyper-prior $p(d, \alpha)$ selected to yield an approximately flat prior on H. Following the approach of [6], we setting $p(d, \alpha)$ proportional to the derivative of the expected entropy. This leaves one extra degree of freedom, since large

[3]The only exception is near the corner $d \to 1$ and $\alpha \to -d$. There, one can obtain arbitrarily large prior variance over H for given mean. However, these such priors have very heavy tails and seem poorly suited to data with finite or exponential tails; we do not explore them further here.

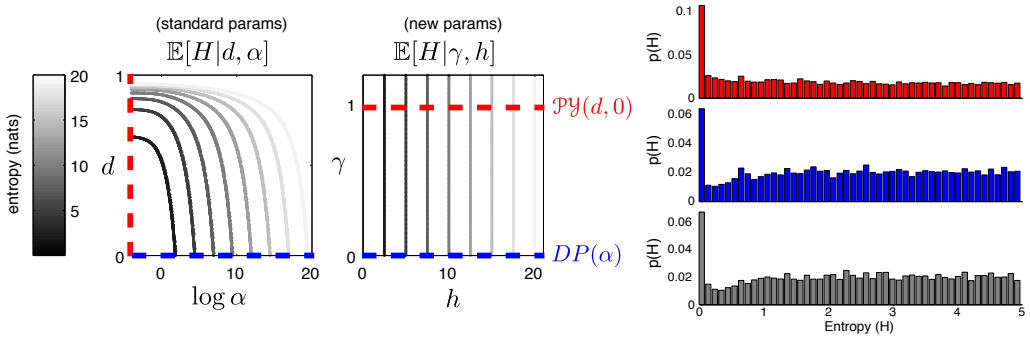

Figure 4: Expected entropy under Pitman-Yor and Pitman-Yor Mixture priors. **(A)** Left: expected entropy as a function of the natural parameters (d, α). Right: expected entropy as a function of transformed parameters (h, γ). **(B)** Sampled prior distributions ($N = 5e3$) over entropy implied by three different $\mathcal{PY}$ mixtures: (1) $p(\gamma, h) \propto \delta(\gamma - 1)$ (red), a mixture of $\mathcal{PY}(d, 0)$ distributions; (2) $p(\gamma, h) \propto \delta(\gamma)$ (blue), a mixture of $\mathcal{DP}(\alpha)$ distributions; and (3) $p(\gamma, h) \propto \exp(-\frac{10}{1-\gamma})$ (grey), which provides a tradeoff between (1) & (2). Note that the implied prior over H is approximately flat.

prior entropies can arise either from large values of α (as in the $\mathcal{DP}$) or from values of d near 1. (See Fig. 4A). We can explicitly control this trade-off by reparametrizing the $\mathcal{PY}$ distribution, letting

$$h = \psi_0(1+\alpha) - \psi_0(1-d), \qquad \gamma = \frac{\psi_0(1) - \psi_0(1-d)}{\psi_0(1+\alpha) - \psi_0(1-d)}, \tag{12}$$

where $h > 0$ is equal to the expected entropy of the prior (eq. 10) and $\gamma > 0$ captures prior beliefs about tail behavior of $\boldsymbol{\pi}$. For $\gamma = 0$, we have the $\mathcal{DP}$ ($d = 0$); for $\gamma = 1$ we have a $\mathcal{PY}(d, 0)$ process (i.e., $\alpha = 0$). Where required, the inverse transformation to standard $\mathcal{PY}$ parameters is given by: $\alpha = {\psi_0}^{-1}\left(h(1-\gamma) + \psi_0(1)\right) - 1$, $d = 1 - {\psi_0}^{-1}\left(\psi_0(1) - h\gamma\right)$, where ${\psi_0}^{-1}(\cdot)$ denotes the inverse digamma function.

We can construct an (approximately) flat improper prior over H on $[0, \infty]$ by setting $p(h, \gamma) = q(\gamma)$, where q is any density on $[0, \infty]$. The induced prior on entropy is thus:

$$p(H) = \iint p(H|\boldsymbol{\pi}) p_{\mathcal{PY}}(\boldsymbol{\pi}|\gamma, h) p(\gamma, h) d\gamma \, dh, \tag{13}$$

where $p_{\mathcal{PY}}(\boldsymbol{\pi}|\gamma, h)$ denotes a $\mathcal{PY}$ distribution on $\boldsymbol{\pi}$ with parameters γ, h. Fig. 4B shows samples from this prior under three different choices of $q(\gamma)$, for h uniform on $[0, 3]$. We refer to the resulting prior distribution over $\boldsymbol{\pi}$ as the Pitman-Yor mixture (PYM) prior. All results in the figures are generated using the prior $q(\gamma) \propto \max(1 - \gamma, 0)$.

4.3 Posterior inference

Posterior inference under the PYM prior amounts to computing the two-dimensional integral over the hyperparameters (d, α),

$$\hat{H}_{\text{PYM}} = \mathbb{E}[H|\mathbf{x}] = \int \mathbb{E}[H|\mathbf{x}, d, \alpha] \frac{p(\mathbf{x}|d, \alpha) p(\alpha, d)}{p(\mathbf{x})} \, \mathrm{d}(d, \alpha) \tag{14}$$

Although in practice we parametrize our prior using the variables γ and h, for clarity and consistency with other literature we present results in terms of d and α. Just as the case with the prior mean, the posterior mean $E[H|\mathbf{x}, d, \alpha]$ is given by a convenient analytic form (derived in the Appendix),

$$\mathbb{E}[H|\alpha, d, \mathbf{x}] = \psi_0(\alpha + N + 1) - \frac{\alpha + Kd}{\alpha + N}\psi_0(1-d) - \frac{1}{\alpha + N}\left[\sum_{i=1}^{K}(n_i - d)\psi_0(n_i - d + 1)\right]. \tag{15}$$

The evidence, $p(\mathbf{x}|d, \alpha)$, is given by

$$p(\mathbf{x}|d, \alpha) = \frac{\left(\prod_{l=1}^{K-1}(\alpha + ld)\right)\left(\prod_{i=1}^{K}\Gamma(n_i - d)\right)\Gamma(1+\alpha)}{\Gamma(1-d)^K \Gamma(\alpha + N)}. \tag{16}$$

We can obtain confidence regions for $\hat{H}_{\text{PYM}}$ by computing the posterior variance $\mathbb{E}[(H - \hat{H}_{\text{PYM}})^2|\mathbf{x}]$. The estimate takes the same form as eq. 14, except that we substitute $\text{var}[H|\mathbf{x}, d, \alpha]$ for $\mathbb{E}[H|\mathbf{x}, d, \alpha]$. Although $\text{var}[H|\mathbf{x}, d, \alpha]$ has an analytic closed form that is fast to compute, it is a lengthy expression that we do not have space to reproduce here; we provide it in the Appendix.

4.4 Computation

In practice, the two-dimensional integral over α and d is fast to compute numerically. Computation of the integrand can be carried out more efficiently using a representation in terms of *multiplicities* (also known as the *empirical histogram distribution function* [4]), the number of symbols that have occurred with a given frequency in the sample. Letting $m_k = |\{i : n_i = k\}|$ denote the total number of symbols with exactly k observations in the sample gives the compressed statistic $\mathbf{m} = [m_0, m_1, \ldots, m_M]^\top$, where n_{max} is the largest number of samples for any symbol. Note that the inner product $[0, 1, \ldots, n_{\text{max}}] \cdot \mathbf{m} = N$, the total number of samples.

The multiplicities representation significantly reduces the time and space complexity of our computations for most datasets, as we need only compute sums and products involving the number symbols with distinct frequencies (at most n_{max}), rather than the total number of symbols K. In practice, we compute all expressions not explicitly involving $\boldsymbol{\pi}$ using the multiplicities representation. For instance, in terms of the multiplicities, the evidence takes the compressed form

$$p(\mathbf{x}|d, \alpha) = p(m_1, \ldots, m_M|d, \alpha) = \frac{\Gamma(1+\alpha)\prod_{l=1}^{K-1}(\alpha + ld)}{\Gamma(\alpha + n)} \prod_{i=1}^{M} \left(\frac{\Gamma(i-d)}{i!\Gamma(1-d)} \right)^{m_i} \frac{M!}{m_i!}. \quad (17)$$

4.5 Existence of posterior mean

Given that the PYM prior with $p(h) \propto 1$ on $[0, \infty]$ is improper, the prior expectation $E[H]$ does not exist. It is therefore reasonable to ask what conditions on the data are sufficient to obtain finite posterior expectation $E[H|\mathbf{x}]$. We give an answer to this question in the following short proposition, the proof of which we provide in Appendix B.

Theorem 1. *Given a fixed dataset* $\mathbf{x}$ *of* N *samples and any bounded (potentially improper) prior* $p(\gamma, h)$, $\hat{H}_{PYM} < \infty$ *when* $N - K \geq 2$.

This result says that the BLS entropy estimate is finite whenever there are at least two "coincidences", i.e., two fewer unique symbols than samples, even though the prior expectation is infinite.

5 Results

We compare PYM to other proposed entropy estimators using four example datasets in Fig. 5. The Miller-Maddow estimator is a well-known method for bias correction based on a first-order Taylor expansion of the entropy functional. The CAE ("Coverage Adjusted Estimator") addresses bias by combining the Horvitz-Thompson estimator with a nonparametric estimate of the proportion of total probability mass (the "coverage") accounted for by the observed data $\mathbf{x}$ [17, 25]. When $d = 0$, PYM becomes a DP mixture (DPM). It may also be thought of as NSB with a very large $\mathcal{A}$, and indeed the empirical performance of NSB with large $\mathcal{A}$ is nearly identical to that of DPM. All estimators appear to converge except $\hat{H}_{nsb_\infty}$, the asymptotic extension of NSB discussed in Section 2.2, which increases unboundedly with data size. In addition PYM performs competitively with other estimators. Note that unlike frequentist estimators, PYM error bars in Fig. 5 arise from direct compution of the posterior variance of the entropy.

6 Discussion

In this paper we introduced PYM, a novel entropy estimator for distributions with unknown support. We derived analytic forms for the conditional mean and variance of entropy under a $\mathcal{DP}$ and $\mathcal{PY}$ prior for fixed parameters. Inspired by the work of [6], we defined a novel $\mathcal{PY}$ mixture prior, PYM, which implies an approximately flat prior on entropy. PYM addresses two major issues with NSB: its dependence on knowledge of $\mathcal{A}$ and its inability (inherited from the Dirichlet distribution) to

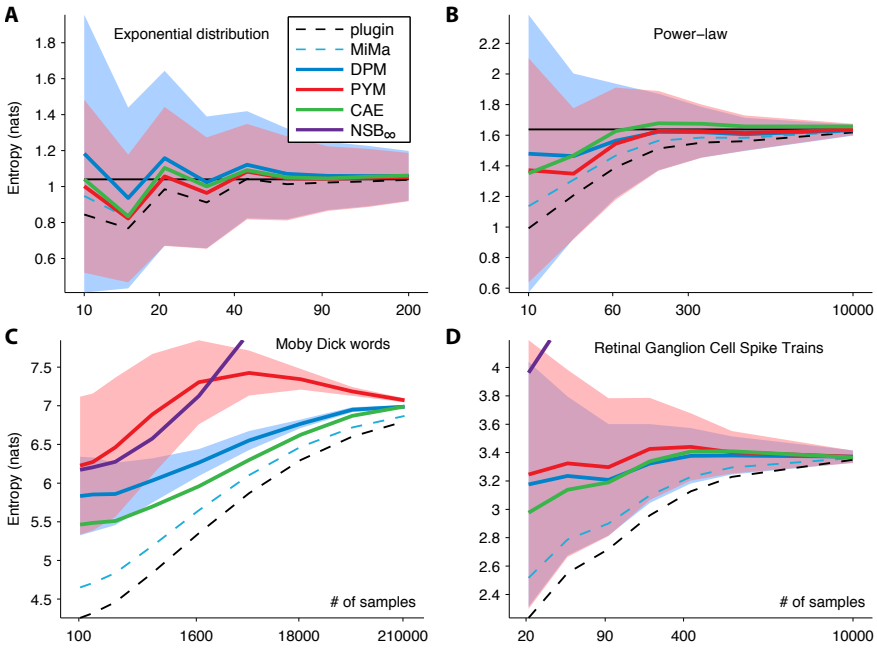

Figure 5: Convergence of entropy estimators with sample size, on two simulated and two real datasets. We write "MiMa" for "Miller-Maddow" and "NSB$_\infty$" for $\hat{H}_{nsb_\infty}$. Note that DPM ("DP mixture") is simply a PYM with $\gamma = 0$. Credible intervals are indicated by two standard deviation of the posterior for DPM and PYM. **(A)** Exponential distribution $\pi_i \propto e^{-i}$. **(B)** Power law distribution with exponent 2 ($\pi_i \propto i^{-2}$). **(C)** Word frequency from the novel Moby Dick. **(D)** Neural words from 8 simultaneously-recorded retinal ganglion cells. Note that for clarity $\hat{H}_{nsb_\infty}$ has been cropped from B and D. All plots are average of 16 Monte Carlo runs.

account for the heavy-tailed distributions which abound in biological and other natural data. We have shown that PYM performs well in comparison to other entropy estimators, and indicated its practicality in example applications to data.

We note, however, that despite its strong performance in simulation and in many practical examples, we cannot assure that PYM will always be well-behaved. There may be specific distributions for which the PYM estimate is so heavily biased that the credible intervals fail to bracket the true entropy. This reflects a general state of affairs for entropy estimation on countable distributions: any convergence rate result must depend on restricting to a subclass of distributions [26]. Rather than working within some analytically-defined subclass of discrete distributions (such as, for instance, those with finite "entropy variance" [17]), we work within the space of distributions parametrized by $\mathcal{PY}$ which spans both the exponential and power-law tail distributions. Although $\mathcal{PY}$ parameterizes a large class of distributions, its structure allows us to use the $\mathcal{PY}$ parameters to understand the qualitative features of the distributions made likely under a choice of prior. We feel this is a key feature for small-sample inference, where the choice of prior is most relevant. Moreover, in a forthcoming paper, we demonstrate the consistency of PYM, and show that its small-sample flexibility does not sacrifice desirable asymptotic properties.

In conclusion, we have defined the PYM prior through a reparametrization that assures an approximately flat prior on entropy. Moreover, although parametrized over the space of countably-infinite discrete distributions, the computation of PYM depends primarily on the first two conditional moments of entropy under $\mathcal{PY}$. We derive closed-form expressions for these moments that are fast to compute, and allow the efficient computation of both the PYM estimate and its posterior credible interval. As we demonstrate in application to data, PYM is competitive with previously proposed estimators, and is especially well-suited to neural applications, where heavy-tailed distributions are commonplace.

Acknowledgments

We thank E. J. Chichilnisky, A. M. Litke, A. Sher and J. Shlens for retinal data, and Y. .W. Teh for helpful comments on the manuscript. This work was supported by a Sloan Research Fellowship, McKnight Scholar's Award, and NSF CAREER Award IIS-1150186 (JP).

References

[1] G. Miller. Note on the bias of information estimates. *Information theory in psychology: Problems and methods*, 2:95–100, 1955.

[2] S. Panzeri and A. Treves. Analytical estimates of limited sampling biases in different information measures. *Network: Computation in Neural Systems*, 7:87–107, 1996.

[3] R. Strong, S. Koberle, de Ruyter van Steveninck R., and W. Bialek. Entropy and information in neural spike trains. *Physical Review Letters*, 80:197–202, 1998.

[4] L. Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15:1191–1253, 2003.

[5] P. Grassberger. Entropy estimates from insufficient samplings. *arXiv preprint*, January 2008, arXiv:0307138 [physics].

[6] I. Nemenman, F. Shafee, and W. Bialek. Entropy and inference, revisited. *Adv. Neur. Inf. Proc. Sys.*, 14, 2002.

[7] J. Pitman and M. Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855–900, 1997.

[8] H. Ishwaran and L. James. Generalized weighted chinese restaurant processes for species sampling mixture models. *Statistica Sinica*, 13(4):1211–1236, 2003.

[9] S. Goldwater, T. Griffiths, and M. Johnson. Interpolating between types and tokens by estimating power-law generators. *Adv. Neur. Inf. Proc. Sys.*, 18:459, 2006.

[10] G. Zipf. *Human behavior and the principle of least effort.* Addison-Wesley Press, 1949.

[11] T. Dudok de Wit. When do finite sample effects significantly affect entropy estimates? *Eur. Phys. J. B - Cond. Matter and Complex Sys.*, 11(3):513–516, October 1999.

[12] M. Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary physics*, 46(5):323–351, 2005.

[13] M. Hutter. Distribution of mutual information. *Adv. Neur. Inf. Proc. Sys.*, 14:399, 2002.

[14] J. Hausser and K. Strimmer. Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *The Journal of Machine Learning Research*, 10:1469–1484, 2009.

[15] I. Nemenman, W. Bialek, and R. Van Steveninck. Entropy and information in neural spike trains: Progress on the sampling problem. *Physical Review E*, 69(5):056111, 2004.

[16] I. Nemenman. Coincidences and estimation of entropies of random variables with large cardinalities. *Entropy*, 13(12):2013–2023, 2011.

[17] V. Q. Vu, B. Yu, and R. E. Kass. Coverage-adjusted entropy estimation. *Statistics in medicine*, 26(21):4039–4060, 2007.

[18] J. W. Pillow, J. Shlens, L. Paninski, A. Sher, A. M. Litke, and E. P. Chichilnisky, E. J. Simoncelli. *Nature*, 454:995–999, 2008.

[19] H. Ishwaran and M. Zarepour. Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics*, 30(2):269–283, 2002.

[20] J. Kingman. Random discrete distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 37(1):1–22, 1975.

[21] H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, March 2001.

[22] Y. Teh. A hierarchical Bayesian language model based on Pitman-Yor processes. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 985–992, 2006.

[23] M. Perman, J. Pitman, and M. Yor. Size-biased sampling of poisson point processes and excursions. *Probability Theory and Related Fields*, 92(1):21–39, March 1992.

[24] J. Pitman. Random discrete distributions invariant under size-biased permutation. *Advances in Applied Probability*, pages 525–539, 1996.

[25] A. Chao and T. Shen. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environmental and Ecological Statistics*, 10(4):429–443, 2003.

[26] A. Antos and I. Kontoyiannis. Convergence properties of functional estimates for discrete distributions. *Random Structures & Algorithms*, 19(3-4):163–193, 2001.

[27] D. Wolpert and D. Wolf. Estimating functions of probability distributions from a finite set of samples. *Physical Review E*, 52(6):6841–6854, 1995.