Bayesian Entropy Estimation for Countable Discrete Distributions

Evan Archer* earcher@utexas.edu
Institute for Computational Engineering Sciences,
The University of Texas at Austin, Austin, TX 78712, USA

Il Memming Park* memming@austin.utexas.edu
Center for Perceptual Systems
The University of Texas at Austin, Austin, TX 78712, USA

Jonathan W. Pillow pillow@utexas.edu
Department of Psychology, Section of Neurobiology,
Division of Statistics and Scientific Computation, and Center for Perceptual Systems
The University of Texas at Austin, Austin, TX 78712, USA

Editor:

Abstract

We consider the problem of estimating Shannon's entropy H from discrete data, in cases where the number of possible symbols is unknown or even countably infinite. The Pitman-Yor process, a generalization of Dirichlet process, provides a tractable prior distribution over the space of countably infinite discrete distributions, and has found major applications in Bayesian non-parametric statistics and machine learning. Here we show that it also provides a natural family of priors for Bayesian entropy estimation, due to the fact that moments of the induced posterior distribution over H can be computed analytically. We derive formulas for the posterior mean (Bayes' least squares estimate) and variance under Dirichlet and Pitman-Yor process priors. Moreover, we show that a fixed Dirichlet or Pitman-Yor process prior implies a narrow prior distribution over H, meaning the prior strongly determines the entropy estimate in the under-sampled regime. We derive a family of continuous mixing measures such that the resulting mixture of Pitman-Yor processes produces an approximately flat prior over H. We show that the resulting Pitman-Yor Mixture (PYM) entropy estimator is consistent for a large class of distributions. We explore the theoretical properties of the resulting estimator, and show that it performs well both in simulation and in application to real data.

Keywords: entropy, information theory, Bayesian estimation, Bayesian nonparametrics, Dirichlet process, Pitman–Yor process, neural coding

1. Introduction

Shannon's discrete entropy appears as a basic statistic in many fields, from probability theory to engineering and even ecology. While entropy may best be known as a theoretical quantity, its accurate estimation from data is an important step in disparate applications. Entropy is employed in the study of information processing in neuroscience (Barbieri et al., 2004; Rolls et al., 1999; Shlens et al., 2007; Strong et al., 1998). It is also used in statistics and machine learning for estimating dependency structure and inferring causal relations (Chow and Liu, 1968; Schindler et al., 2007), for example in molecular biology (Hausser and Strimmer, 2009); as a tool in the study of complexity and dynamics in physics (Letellier, 2006); and as a measure of diversity in ecology (Chao and Shen, 2003) and genetics (Farach et al., 1995). Each of these studies, confronted with data arising from an unknown

∗. EA and IP contributed equally.

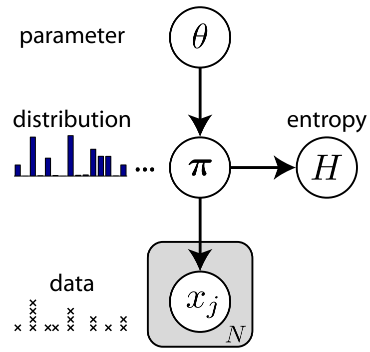

Figure 1: Graphical model illustrating the ingredients for Bayesian entropy estimation. Arrows indicate conditional dependencies between variables, and the gray "plate" denotes multiple copies of a random variable (with the number of copies N indicated at bottom). For entropy estimation, the joint probability 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 $p(H|\boldsymbol{\pi}) = \delta(H - \sum_i \pi_i \log \pi_i)$. The Bayes least squares estimator corresponds to the posterior mean: $\mathbb{E}[H|\mathbf{x}] = \iint p(H|\boldsymbol{\pi})p(\boldsymbol{\pi}, \theta|\mathbf{x})d\boldsymbol{\pi}\, d\theta$.

discrete distribution, seeks to estimate the entropy rather than the distribution itself. Estimating the entropy is much easier than estimating the full distribution. In fact, in many cases, entropy can be accurately estimated with fewer samples than the number of distinct symbols. However, entropy estimation remains a difficult problem: there is no unbiased estimator for entropy, and the maximum likelihood estimator is severely biased for small datasets. Many previous studies have focused upon methods for computing and reducing this bias (Grassberger, 2008; Miller, 1955; Paninski, 2003; Panzeri and Treves, 1996; Strong et al., 1998). In this paper we instead take a Bayesian approach, building upon the work of Nemenman et al. (2002). Our basic strategy is to place a prior over the space of discrete probability distributions, and then perform inference using the induced posterior distribution over entropy. (See Fig. 1).

We focus here on the under-sampled regime, where the number of unique symbols observed in the data is small in comparison with the unknown (perhaps countably infinite) number of possible symbols. The Pitman-Yor process (PYP), a two-parameter generalization of the Dirichlet process (DP) (Goldwater et al., 2006; Ishwaran and James, 2003; Pitman and Yor, 1997), provides an attractive family of priors in this setting, 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 (Dudok de Wit, 1999; Newman, 2005; Zipf, 1949). We show that a PYP prior with fixed hyperparameters imposes a narrow prior distribution over entropy, leading to severe bias and overly narrow posterior credible intervals given a small dataset. Our approach, inspired by Nemenman et al. (2002), is to introduce a family of mixing measures over Pitman-Yor processes 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 compare various estimators with 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 $\mathcal{X}$, with cardinality $\mathcal{A}$, that is, $p(x_j = i) = \pi_i$. We wish to estimate the entropy of $\boldsymbol{\pi}$,

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

We are interested in the under-sampled regime, $N \ll \mathcal{A}$, where many of the symbols remain unobserved. We will see that a naive approach to entropy estimation in this regime results in seriously biased estimators, and briefly review approaches for correcting this bias. We then consider Bayesian techniques for entropy estimation in general before introducing the NSB method upon which the remainder of the article will build.

2.1 Plugin estimator and bias-correction methods

Perhaps the most straightforward entropy estimation technique is to estimate the distribution $\boldsymbol{\pi}$ and then use the plugin formula (1) to evaluate its entropy. The empirical distribution $\hat{\boldsymbol{\pi}} = (\hat{\pi}_1, \ldots, \hat{\pi}_{\mathcal{A}})$ is computed by normalizing the observed counts $\mathbf{n} := (n_1, \ldots, n_{\mathcal{A}})$ of each symbol,

$$\hat{\pi}_k = n_k/N, \qquad n_k = \sum_{i=1}^{N} \mathbf{1}_{\{x_i = k\}}, \tag{2}$$

for each $k \in \mathcal{X}$. Plugging this estimate for $\boldsymbol{\pi}$ into (1), we obtain the so-called "plugin" estimator:

$$\hat{H}_{\text{plugin}} = -\sum \hat{\pi}_i \log \hat{\pi}_i, \tag{3}$$

which is also the maximum-likelihood estimator under discrete (or multinomial) likelihood. Despite its simplicity and desirable asymptotic properties, $\hat{H}_{\text{plugin}}$ exhibits substantial negative bias in the undersampled regime. There exists a large literature on methods for removing this bias, much of which considers the setting in which $\mathcal{A}$ is known and finite. One popular and well-studied method involves taking a series expansion of the bias (Grassberger, 2008; Miller, 1955; Panzeri and Treves, 1996; Treves and Panzeri, 1995) and then subtracting it from the plugin estimate. Other recent proposals include minimizing an upper bound over a class of linear estimators (Paninski, 2003), and a James-Stein estimator (Hausser and Strimmer, 2009). Recent work has also considered countably infinite alphabets. The coverage-adjusted estimator (CAE) (Chao and Shen, 2003; Vu et al., 2007) 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 $\mathbf{x}$. In a similar spirit, Zhang (2012) proposed an estimator based on the Good-Turing estimate of population size.

2.2 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:

$$\hat{H}(\mathbf{x}) = \mathbb{E}[H|\mathbf{x}] = \int H(\boldsymbol{\pi}) p(H|\boldsymbol{\pi}) p(\boldsymbol{\pi}|\mathbf{x})\, \mathrm{d}\boldsymbol{\pi}, \tag{4}$$

where $p(\boldsymbol{\pi}|\mathbf{x})$ is the posterior over $\boldsymbol{\pi}$ under some prior $p(\boldsymbol{\pi})$ and discrete likelihood $p(\mathbf{x}|\boldsymbol{\pi})$, and

$$p(H|\boldsymbol{\pi}) = \delta(H + \sum_i \pi_i \log \pi_i), \tag{5}$$

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 with the categorical distribution. It takes the form

$$p_{Dir}(\boldsymbol{\pi}) \propto \prod_{i=1}^{\mathcal{A}} \pi_i^{\alpha-1}, \tag{6}$$

for $\boldsymbol{\pi}$ on the $\mathcal{A}$-dimensional simplex ($\pi_i \geq 1$, $\sum \pi_i = 1$), where $\alpha > 0$ is a "concentration" parameter (Hutter, 2002). Many previously proposed estimators can be viewed as Bayesian under a Dirichlet prior with particular fixed choice of α. See Hausser and Strimmer (2009) for a historical overview of entropy estimators arising from specific choices of α.

2.3 Nemenman-Shafee-Bialek (NSB) estimator

In a seminal paper, Nemenman et al. (2002) showed that for finite distributions with known $\mathcal{A}$, Dirichlet priors with fixed α impose a narrow prior distribution over entropy. In the undersampled regime, Bayesian estimators based on such priors are severely biased. Moreover, they have undesirably narrow posterior credible intervals, reflecting narrow prior uncertainty rather than strong evidence from the data. (These estimators generally give incorrect answers with high confidence!). To address this problem, Nemenman et al. (2002) suggested a mixture-of-Dirichlets prior:

$$p(\boldsymbol{\pi}) = \int p_{\mathrm{Dir}}(\boldsymbol{\pi}|\alpha)p(\alpha)d\alpha, \tag{7}$$

where $p_{\mathrm{Dir}}(\boldsymbol{\pi}|\alpha)$ denotes a $\mathrm{Dir}(\alpha)$ prior on $\boldsymbol{\pi}$, and $p(\alpha)$ denotes a set of mixing weights, given by

$$p(\alpha) \propto \frac{d}{d\alpha}\mathbb{E}[H|\alpha] = \mathcal{A}\psi_1(\mathcal{A}\alpha + 1) - \psi_1(\alpha + 1), \tag{8}$$

where $\mathbb{E}[H|\alpha]$ denotes the expected value of H under a $\mathrm{Dir}(\alpha)$ prior, and $\psi_1(\cdot)$ denotes the tri-gamma function. To the extent that $p(H|\alpha)$ resembles a delta function, (7) and (8) imply a uniform prior for H on $[0, \log \mathcal{A}]$. The BLS estimator under the NSB prior can be written:

$$\begin{aligned}\hat{H}_{nsb} = \mathbb{E}[H|\mathbf{x}] &= \iint H(\boldsymbol{\pi})p(\boldsymbol{\pi}|\mathbf{x},\alpha)\, p(\alpha|\mathbf{x})\, \mathrm{d}\boldsymbol{\pi}\, \mathrm{d}\alpha \\ &= \int \mathbb{E}[H|\mathbf{x},\alpha]\frac{p(\mathbf{x}|\alpha)p(\alpha)}{p(\mathbf{x})}\, d\alpha,\end{aligned} \tag{9}$$

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 Pólya distribution (Minka, 2003):

$$\begin{aligned}p(\mathbf{x}|\alpha) &= \int p(\mathbf{x}|\boldsymbol{\pi})p(\boldsymbol{\pi}|\alpha)d\boldsymbol{\pi} \\ &= \frac{(N!)\Gamma(\mathcal{A}\alpha)}{\Gamma(\alpha)^{\mathcal{A}}\Gamma(N+\mathcal{A}\alpha)}\prod_{i=1}^{\mathcal{A}}\frac{\Gamma(n_i+\alpha)}{n_i!}.\end{aligned} \tag{10}$$

The NSB estimate $\hat{H}_{nsb}$ and its posterior variance are fast to compute via 1D numerical integration in α using closed-form expressions for the first two moments of the posterior distribution of H given α. The forms for these moments are discussed in Nemenman et al. (2002); Wolpert and Wolf (1995), but the full formulae are not explicitly shown. Here we state the results:

$$\mathbb{E}[H|\mathbf{x},\alpha] = \psi_0(\tilde{N}+1) - \sum_i \frac{\tilde{n}_i}{\tilde{N}} \psi_0(\tilde{n}_i + 1) \tag{11}$$

$$\mathbb{E}[H^2|\mathbf{x},\alpha] = \sum_{i \neq k} \frac{\tilde{n}_i \tilde{n}_k}{(\tilde{N}+1)\tilde{N}} I_{i,k} + \sum_i \frac{(\tilde{n}_i+1)\tilde{n}_i}{(\tilde{N}+1)\tilde{N}} J_i \tag{12}$$

$$I_{i,k} = \Big(\psi_0(\tilde{n}_k+1) - \psi_0(\tilde{N}+2)\Big)\Big(\psi_0(\tilde{n}_i+1) - \psi_0(\tilde{N}+2)\Big) - \psi_1(\tilde{N}+2)$$

$$J_i = (\psi_0(\tilde{n}_i+2) - \psi_0(\tilde{N}+2))^2 + \psi_1(\tilde{n}_i+2) - \psi_1(\tilde{N}+2),$$

where $\tilde{n}_i = n_i + \alpha$ are counts plus prior "pseudocount" α, $\tilde{N} = \sum \tilde{n}_i$ is the total of counts plus pseudocounts, and ψ_n is the polygamma of n-th order (i.e., ψ_0 is the digamma function). Finally, $\mathrm{var}[H|\mathbf{n},\alpha] = \mathbb{E}[H^2|\mathbf{n},\alpha] - \mathbb{E}[H|\mathbf{n},\alpha]^2$. We derive these formulae in the Appendix, and in addition provide an alternative derivation using a size-biased sampling formulae discussed in Section 3.

2.4 Asymptotic NSB estimator

Nemenman *et al.* have proposed an extension of the NSB estimator to countably infinite distributions (or distributions with unknown cardinality), using a zeroth order approximation to $\hat{H}_{nsb}$ in the limit $\mathcal{A} \to \infty$ which we refer to as $\hat{H}_{nsb_\infty}$ (Nemenman, 2011; Nemenman et al., 2004),

$$\hat{H}_{nsb_\infty} = 2\log(N) + \psi_0(N-K) - \psi_0(1) - \log(2), \tag{13}$$

where K is the number of distinct symbols in the sample. Unfortunately, $\hat{H}_{nsb_\infty}$ increases unboundedly with N (as noted by Vu et al. (2007)), and performs poorly for the examples we consider.

3. Dirichlet and Pitman-Yor Process Priors

To construct a prior over unknown or countably infinite discrete distributions, we borrow tools from nonparametric Bayesian statistics. The Dirichlet Process (DP) and Pitman-Yor process (PYP) define stochastic processes whose samples are countably infinite discrete distributions (Kingman, 1975; Pitman and Yor, 1997). 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] We use DP and PYP to define a prior distribution on the infinite-dimensional simplex. The prior distribution over $\boldsymbol{\pi}$ under the DP or PYP is technically called the GEM distribution or the two-parameter Poisson-Dirichlet distribution, but we will abuse terminology by referring to both the process and its associated weight distribution by the same symbol, DP or PY (Ishwaran and Zarepour, 2002). The DP distribution over $\boldsymbol{\pi}$ results from a limit of the (finite) Dirichlet distribution where alphabet size grows and concentration parameter shrinks: $\mathcal{A} \to \infty$ and $\alpha \to 0$ s.t. $\alpha\mathcal{A} \to \alpha'$. The PYP distribution over $\boldsymbol{\pi}$ generalizes the DP to allow power-law tails (and includes DP as a special case) (Kingman, 1975; Pitman and Yor, 1997). For PY(d,α) with $d \neq 0$, the tails approximately follow a power-law: $\pi_i \propto (i)^{-\frac{1}{d}}$ (pp. 867, (Pitman and Yor, 1997)) [2] Many natural phenomena such as city size, language, spike responses, etc., also exhibit power-law tails (Newman, 2005; Zipf, 1949). Fig. 2 shows two such examples, along with a sample drawn from the best-fitting DP and PYP distributions. Let PY(d,α) denote the PYP 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 Dirichlet process, DP(α). We can draw samples

1. Here, we will assume the base measure is non-atomic, so that the atoms ϕ_i's are distinct with probability one. 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 (Goldwater et al., 2006; Teh, 2006).

$\boldsymbol{\pi} \sim \mathrm{PY}(d, \alpha)$ from an infinite sequence of independent Beta random variables in a process known as "stick-breaking" (Ishwaran and James, 2001):

$$\beta_i \sim \mathrm{Beta}(1-d, \alpha + id), \qquad \tilde{\pi}_i = \prod_{j=1}^{i-1}(1-\beta_j)\beta_i, \tag{14}$$

where $\tilde{\pi}_i$ is known as the i'th *size-biased permutation* from $\boldsymbol{\pi}$ (Pitman, 1996). 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 (Pitman and Yor, 1997).

3.1 Expectations over DP and PYP priors

A key virtue of PYP priors for our purposes is a mathematical property called *invariance under size-biased sampling*, which allows us to convert expectations over $\boldsymbol{\pi}$ on the infinite-dimensional simplex (which are required for computing the mean and variance of H given data) into one- or two-dimensional integrals with respect to the distribution of the first two size-biased samples (Perman et al., 1992; Pitman, 1996).

Proposition 1 (Expectations with first two size-biased samples) *For $\boldsymbol{\pi} \sim PY(d, \alpha)$,*

$$\mathbb{E}_{(\boldsymbol{\pi}|d,\alpha)}\left[\sum_{i=1}^{\infty} f(\pi_i)\right] = \mathbb{E}_{(\tilde{\pi}_1|d,\alpha)}\left[\frac{f(\tilde{\pi}_1)}{\tilde{\pi}_1}\right], \tag{15}$$

$$\mathbb{E}_{(\boldsymbol{\pi}|d,\alpha)}\left[\sum_{i,j\neq i} g(\pi_i, \pi_j)\right] = \mathbb{E}_{(\tilde{\pi}_1,\tilde{\pi}_2|d,\alpha)}\left[\frac{g(\tilde{\pi}_1, \tilde{\pi}_2)}{\tilde{\pi}_1\tilde{\pi}_2}(1-\tilde{\pi}_1)\middle|\boldsymbol{\pi}\right], \tag{16}$$

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

The first result (15) appears in (Pitman and Yor, 1997), and an analogous proof can be constructed for (16) (see Appendix). The direct consequence of this lemma is that the first two moments of $H(\boldsymbol{\pi})$ under the PYP and DP priors have closed forms,

$$\mathbb{E}[H|d,\alpha] = \psi_0(\alpha+1) - \psi_0(1-d), \tag{17}$$

$$\mathrm{var}[H|d,\alpha] = \frac{\alpha+d}{(\alpha+1)^2(1-d)} + \frac{1-d}{\alpha+1}\psi_1(2-d) - \psi_1(2+\alpha). \tag{18}$$

The derivation can be found in the appendix.

3.2 Expectations over DP and PYP posteriors

A useful property of PYP priors (for multinomial observations) is that the posterior $p(\boldsymbol{\pi}|\mathbf{x}, d, \alpha)$ takes the form of a mixture of a Dirichlet distribution (over the observed symbols) and a Pitman-Yor process (over the unobserved symbols) (Ishwaran and James, 2003). This makes the integrals over the infinite-dimensional simplex tractable, and as a result we obtain closed form solutions for the posterior mean and variance of H. Let K be the number of unique symbols observed in N samples, i.e., $K = \sum_{i=1}^{\mathcal{A}} \mathbf{1}_{\{n_i>0\}}$. Further, let $\alpha_i = n_i - d$, $N = \sum n_i$, and $A = \sum \alpha_i = \sum_i n_i - Kd = N - Kd$. Now, following (Ishwaran and Zarepour, 2002) we write the posterior as an infinite random vector $\boldsymbol{\pi}_{\mathrm{post}} = (p_1, p_2, p_3, \ldots, p_K, p_*\boldsymbol{\pi})$, where

$$\begin{aligned}(p_1, p_2, \ldots, p_K, p_*) &\sim \mathrm{Dir}(n_1 - d, \ldots, n_K - d, \alpha + Kd) \\ \boldsymbol{\pi} := (\pi_1, \pi_2, \pi_3, \ldots) &\sim \mathrm{PY}(d, \alpha + Kd).\end{aligned} \tag{19}$$

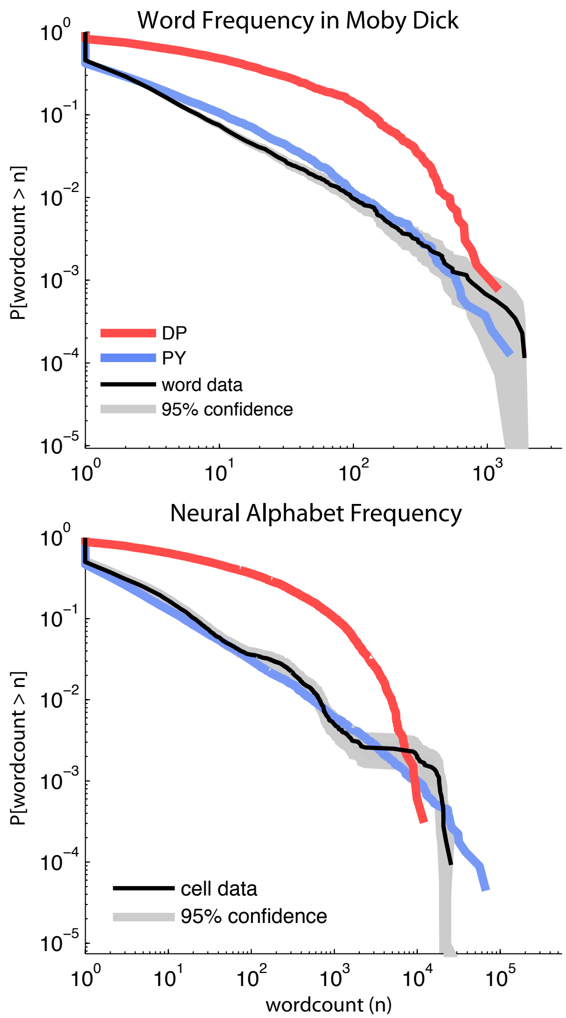

Figure 2: Frequency distributions of words in natural language (above) and neural spike patterns (below). We compare samples from the DP (red) and PYP (blue) priors for two datasets with heavy tails (black). In both cases, we compare the empirical CDF estimated from data to distributions drawn from DP and PYP using the ML values of α and (d, α), respectively. For both datasets, PYP better captures the heavy-tailed behavior of the data. **Top:** Frequency of $N = 217826$ words in the novel Moby Dick by Herman Melville. **Bottom:** Frequencies among $N = 1.2 \times 10^6$ neural spike words from 27 simultaneously-recorded retinal ganglion cells, binarized and binned at 10ms.

The posterior mean $E[H|\mathbf{x}, d, \alpha]$ is given by,

$$\mathbb{E}[H|\alpha, d, \mathbf{x}] = \psi_0(\alpha + N + 1) - \frac{\alpha + Kd}{\alpha + N}\psi_0(1 - d) - \frac{1}{\alpha + N}\left[\sum_{i=1}^{K}(n_i - d)\psi_0(n_i - d + 1)\right]. \quad (20)$$

The variance, $\mathrm{var}[H|\mathbf{x}, d, \alpha]$, also has an analytic closed form which is fast to compute. As we discuss in detail in Appendix A.4, $\mathrm{var}[H|\mathbf{x}, d, \alpha]$ may be expressed in terms of the first two moments of p_*, $\boldsymbol{\pi}$, and $\mathbf{p} = (p_1, \ldots, p_K)$ appearing in the posterior (19). Applying the law of total variance and using the independence properties of the posterior, we find:

$$\begin{aligned}\operatorname{var}[H|d,\alpha] = \mathbb{E}_{p_*}[(1-p_*)^2] \operatorname*{var}_{\mathbf{p}}[H(\mathbf{p})] + \mathbb{E}_{p_*}[p_*^2] \operatorname*{var}_{\boldsymbol{\pi}}[H(\boldsymbol{\pi})] \\ + \mathbb{E}_{p_*}[\Omega^2(p_*)] - \mathbb{E}_{p_*}[\Omega(p_*)]^2,\end{aligned} \tag{21}$$

where $\Omega(p_*) := (1-p_*)\mathbb{E}_{\mathbf{p}}\left[H(\mathbf{p})\right] + p_*\mathbb{E}_{\boldsymbol{\pi}}\left[H(\boldsymbol{\pi})\right] + h(p_*)$. To specify $\Omega(p_*)$, we let $\mathbf{A} := \mathbb{E}_{\mathbf{p}}\left[H(\mathbf{p})\right]$, $\mathbf{B} := \mathbb{E}_{\boldsymbol{\pi}}\left[H(\boldsymbol{\pi})\right]$ so that,

$$\begin{aligned}\mathbb{E}[\Omega] &:= \mathbb{E}_{p_*}[1-p_*]\mathbb{E}_{\mathbf{p}}\left[H(\mathbf{p})\right] + \mathbb{E}_{p_*}[p_*]\mathbb{E}_{\boldsymbol{\pi}}\left[H(\boldsymbol{\pi})\right] + h(p_*), \\ \mathbb{E}[\Omega^2] &:= 2\mathbb{E}_{p_*}[p_* h(p_*)][\mathbf{B}-\mathbf{A}] + 2\mathbf{A}\mathbb{E}_{p_*}[h(p_*)] + \mathbb{E}_{p_*}[h^2(p_*)] \\ &\quad + \mathbb{E}_{p_*}[p_*^2]\left[\mathbf{B}^2 - 2\mathbf{AB}\right] + 2\mathbb{E}_{p_*}[p_*]\mathbf{AB} + \mathbb{E}_{p_*}[(1-p_*)^2]\mathbf{A}^2.\end{aligned}$$

4. Entropy inference under DP and PYP priors

The posterior expectations computed in Section 3.2 provide a class of entropy estimators for distributions with countably-infinite support. For each choice of (d,α), $\mathbb{E}[H|\alpha,d,\mathbf{x}]$ is the posterior mean under a $PY(d,\alpha)$ prior, analygous to the fixed-α Dirichlet priors discussed in Section 2.2. Unfortunately, fixed $PY(d,\alpha)$ priors also carry the same difficulties as fixed Dirichlet priors. A fixed-parameter $\mathrm{PY}(d,\alpha)$ prior on $\boldsymbol{\pi}$ results in a highly concentrated prior distribution on entropy (Fig. 3). We address this problem by introducing a mixture prior $p(d,\alpha)$ on $\mathrm{PY}(d,\alpha)$ under which the implied prior on entropy is flat [3] We then define the BLS entropy estimator under this mixture prior, the Pitman-Yor Mixture (PYM) estimator, and discuss some of its theoretical properties. Finally, we turn to the computation of PYM, discussing methods for sampling, and numerical quadrature integration.

4.1 Pitman-Yor process mixture (PYM) prior

One way of constructing a flat mixture prior is to follow the approach of Nemenman et al. (2002) by setting $p(d,\alpha)$ proportional to the derivative of the expected entropy (17). Unlike NSB, we have two parameters through which to control the prior expected entropy. For instance, large prior (expected) entropies can arise either from large values of α (as in the DP) or from values of d near 1 (see Fig. 3A). We can explicitly control this trade-off by reparametrizing PYP as follows,

$$h = \psi_0(\alpha+1) - \psi_0(1-d), \quad \gamma = \frac{\psi_0(1) - \psi_0(1-d)}{\psi_0(\alpha+1) - \psi_0(1-d)}, \tag{22}$$

where $h > 0$ is equal to the expected prior entropy (17), and $\gamma \in [0,\infty)$ captures prior beliefs about tail behavior (Fig. 4A). For $\gamma = 0$, we have the DP (i.e., $d = 0$, giving $\boldsymbol{\pi}$ with exponential tails), while for $\gamma = 1$ we have a $\mathrm{PY}(d,0)$ process (i.e., $\alpha = 0$, yielding $\boldsymbol{\pi}$ with power-law tails). Where required, the inverse transformation to standard 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 distribution over H on $[0,\infty]$ by setting $p(h,\gamma) = q(\gamma)$ for all h, where q is any density on $[0,\infty)$. We call this the Pitman-Yor process mixture (PYM) prior. The induced prior on entropy is thus:

$$p(H) = \iint p(H|\boldsymbol{\pi})p(\boldsymbol{\pi}|\gamma,h)p(\gamma,h)\,\mathrm{d}\gamma\,\mathrm{d}h, \tag{23}$$

3. Notice, however, that by constructing a flat prior on entropy, we do not obtain an objective prior. Here, we are not interested in estimating the underlying high-dimensional probabilities $\{\pi_i\}$, but rather in estimating a single statistic. An objective prior on the model parameters is not necessarily optimal for estimating entropy: entropy is not a parameter in our model. In fact, Jefferys' prior for multinomial observations is exactly a Dirichlet distribution with a fixed $\alpha = 1/2$. As mentioned in the text, such Bayesian priors heavily bias entropy estimates.

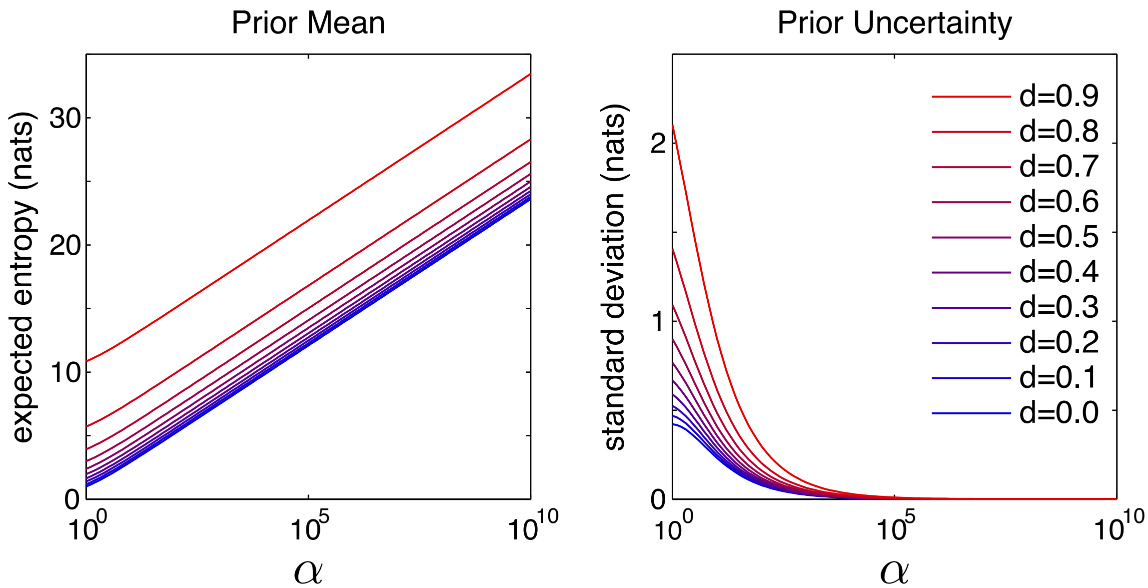

Figure 3: Prior entropy mean and variance (17) and 18 as a function of α and d. Note that entropy is approximiately linear in $\log \alpha$. For large values of α, $p(H(\boldsymbol{\pi})|d, \alpha)$ is highly concentrated around the mean.

where $p(\boldsymbol{\pi}|\gamma, h)$ denotes a PYP on $\boldsymbol{\pi}$ with parameters γ, h. The ability to adapt $q(\gamma)$ to a given problem greatly enhances PYM's flexibility. The PYM mixture priors resulting from two different choices of $q(\gamma)$ are both approximately flat on H, but each favors distributions with different tail behavior. Fig. 4B shows samples from this prior under three different choices of $q(\gamma)$, for h uniform on $[0, 3]$. For the experiments, we use $q(\gamma) = \exp(-\frac{10}{1-\gamma})\mathbf{1}_{\{\gamma<1\}}$ which yields good results by weighting less on extremely heavy-tailed distributions. Combined with the likelihood, the posterior $p(d, \alpha|\mathbf{x}) \propto p(\mathbf{x}|d, \alpha)p(d, \alpha)$ quickly concentrates as more data are given, as demonstrated in Fig. 5.

4.2 The Pitman-Yor Mixture Entropy Estimator

Now that we have determined a prior on the infinite simplex, we turn to the problem of inference given observations $\mathbf{x}$. The Bayes least squares entropy estimator under the mixture prior $p(d, \alpha)$, the Pitman-Yor Mixture (PYM) estimator, takes the form

$$\hat{H}_{\mathrm{PYM}} = \mathbb{E}[H|\mathbf{x}] = \int \mathbb{E}[H|\mathbf{x}, d, \alpha] \frac{p(\mathbf{x}|d, \alpha)p(d, \alpha)}{p(\mathbf{x})} \,\mathrm{d}(d, \alpha), \tag{24}$$

where $\mathbb{E}[H|\mathbf{x}, d, \alpha]$ is the expected posterior entropy for a fixed (d, α) (see Section 3.2). The quantity $p(\mathbf{x}|d, \alpha)$ is the evidence, given by

$$p(\mathbf{x}|d, \alpha) = \frac{\left(\prod_{l=1}^{K-1}(\alpha + ld)\right)\left(\prod_{i=1}^{K}\Gamma(n_i - d)\right)\Gamma(1 + \alpha)}{\Gamma(1 - d)^K \Gamma(\alpha + N)}. \tag{25}$$

We can obtain posterior credible intervals regions for $\hat{H}_{\mathrm{PYM}}$ by estimating the posterior variance $\mathbb{E}[(H - \hat{H}_{\mathrm{PYM}})^2|\mathbf{x}]$. The estimate takes the same form as (24), except that we replace $\mathbb{E}[H|\mathbf{x}, d, \alpha]$ with $\mathrm{var}[H|\mathbf{x}, d, \alpha]$ in the integrand.

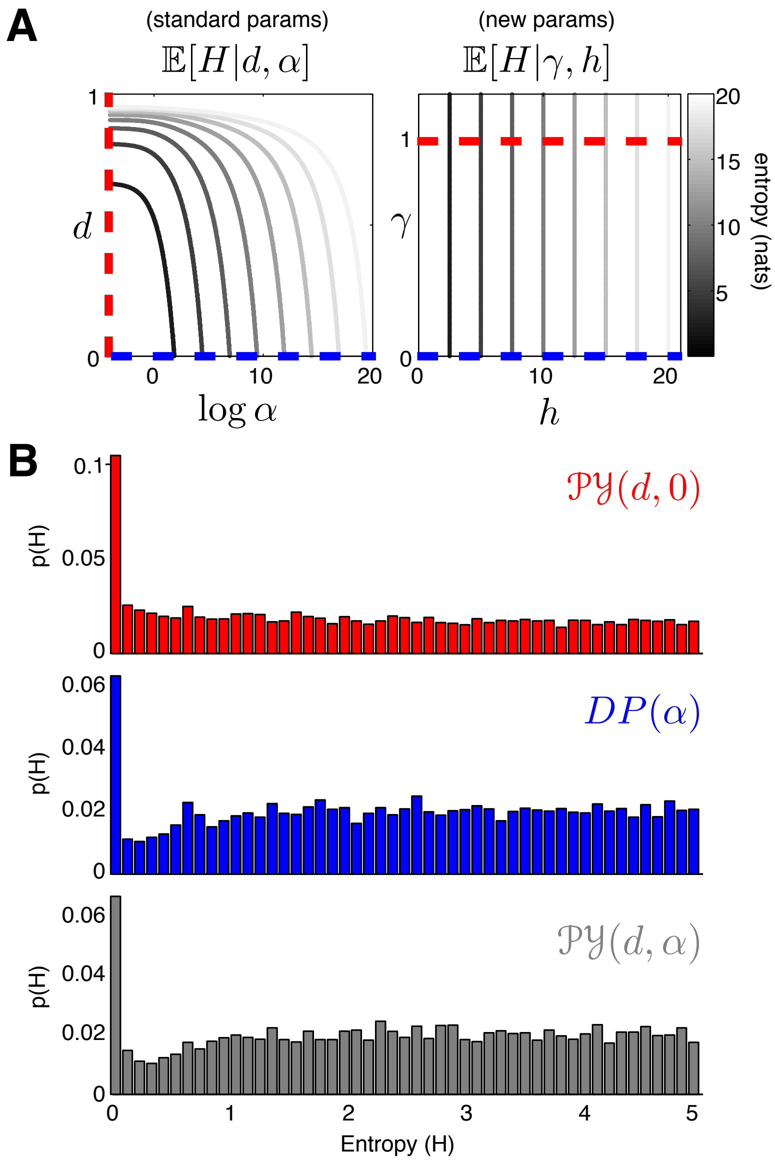

Figure 4: Prior over expected entropy under Pitman-Yor process prior. **(A)** Left: expected entropy as a function of the natural parameters (d, α). Right: expected entropy as a function of transformed parameters (h, γ). **(B)** Sampled prior distributions ($N = 5e3$) over entropy implied by three different PYP priors: $\mathrm{PY}(d, 0)$ (red), $\mathrm{PY}(d, \alpha)$ (grey), and $\mathrm{PY}(0, \alpha) = \mathrm{DP}(\alpha)$ (blue). We show the distributions only on the range from 0 to 5 nats; sampling from PY becomes prohibitively expensive with increasing expected entropy, especially as $d \to 1$.

4.3 Computation

Due to the improperness of the prior $p(d, \alpha)$ and the requirement of integrating over all $\alpha > 0$ (eq. (24)), it is not obvious that the PYM estimate $\hat{H}_{\mathrm{PYM}}$ is computationally tractable. In this

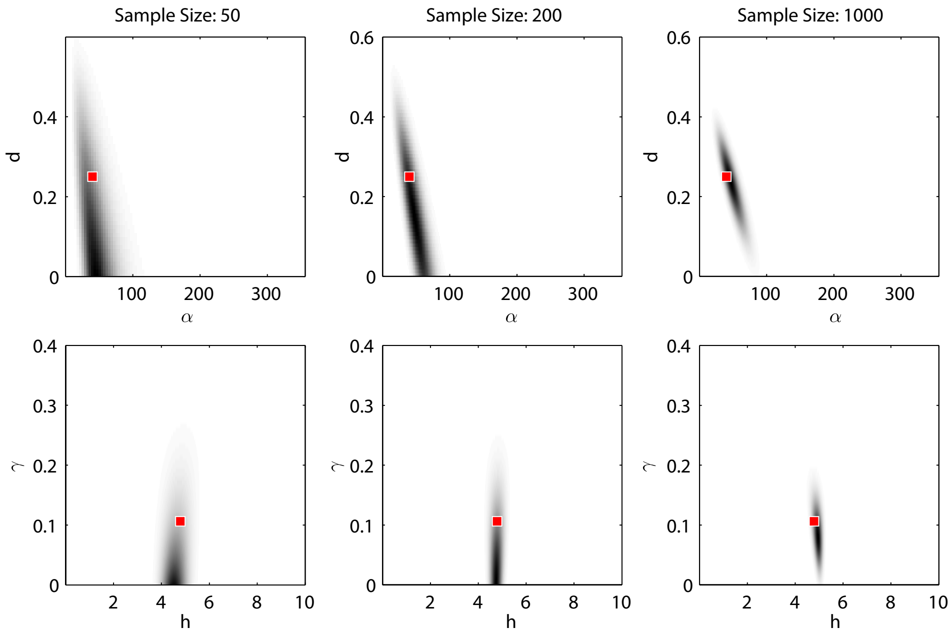

Figure 5: Convergence of $p(d, \alpha|\mathbf{x})$ for increasing sample size. Both parameterization (d, α) and (γ, h) are shown. Data are simulated from a $\mathrm{PY}(0.25, 40)$ whose parameters are indicated by the red dot.

section we discuss techniques for efficient and accurate computation of $\hat{H}_{\mathrm{PYM}}$. First, we outline a compressed data representation we call the "multiplicities" representation, which substantially reduces computational cost. Then, we outline a fast method for performing the numerical integration over a suitable range of α and d.

4.3.1 Multiplicities

Computation of the expected entropy $\mathbb{E}[H|\mathbf{x}, d, \alpha]$ can be carried out more efficiently using a representation in terms of *multiplicities* (also known as the *empirical histogram distribution function* (Paninski, 2003)), 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 M is the largest number of samples for any symbol. Note that the inner product $[0, 1, \ldots, M]^\top \mathbf{m} = N$, is 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 M), rather than the total number of symbols K. In practice we compute all expressions not explicitly involving $\boldsymbol{\pi}$ using the multiplcities representation. For instance, in terms of the multiplicities the evidence takes the compressed form,

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

4.3.2 Heuristic for Integral Computation

In principle the PYM integral over α is supported on the range $[0, \infty)$. In practice, however, the posterior is concentrated on a relatively small region of parameter speace. It is generally unncessary to consider the full integral over a semi-infinite domain. Instead, we select a subregion of $[0, 1] \times [0, \infty)$ which supports the posterior up to ϵ probability mass. We illustrate the concentration of the posterior visually in figure 5. We compute the hessian at the MAP parameter value, $(d_{\mathrm{MAP}}, \alpha_{\mathrm{MAP}})$. Using the inverse hessian as the covariance of a Gaussian approximation to the posterior, we select the grid which spans ± 6 std. We use numerial integration (Gauss-Legendre quadrature) on this region to compute the integral. When the hessian is rank-deficient (which may occur, for instance, when the $\alpha_{\mathrm{MAP}} = 0$ or $d_{\mathrm{MAP}} = 0$), we use Gauss-Legendre quadature perform the integral in d over $[0, 1)$, but employ a Fourier-Chebyshev numerical quadrature routine to integrate α over $[0, \infty)$ (Boyd, 1987).

4.4 Sampling the full posterior over H

The closed-form expressions for the conditional moments derived in the previous section allow us to compute PYM and its variance by 2-dimensional numerical integraton. PYM's posterior mean and variance provide essentially a Gaussian approximation to the posterior, and corresponding credible regions. However, in some situations (see Fig. 6), variance-based credible intervals are a poor approximation to the true posterior credible intervals. In such situations we may wish to examine the full posterior distribution over H. In what follows we describe methods for exactly sampling the posterior and argue that the posterior variance provides a good approximation to the true credible interval in many situations. Stick-breaking, as described by (14), provides a straightforward

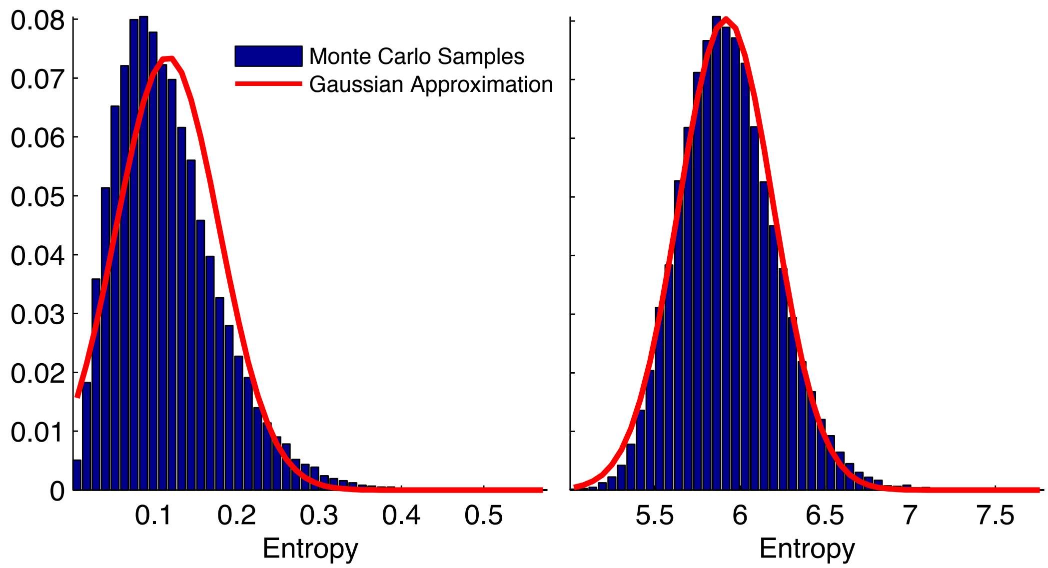

Figure 6: The posterior distributions of entropy for two datasets of 100 samples drawn from distinct distributions and the Gaussian approximation to each distribution based upon the posterior mean and variance. **(left):** Simulated example with low entropy. Notice that the true posterior is highly asymmetric, and that the Gaussian approximation does not respect the positivity of H. **(right)** Simulated example with higher entropy. The Gaussian approximation is a much closer approximation to the true distribution.

algorithm for sampling distributions $\boldsymbol{\pi} \sim \mathrm{PY}(d, \alpha)$. With large enough N_s, stick-breaking samples

$\{\tilde{\pi}_i\}_{i=1}^{N_s}$ approximate π to arbitrary accuracy[4]. Even so, in practice sampling $\boldsymbol{\pi}$ is difficult when it is heavy-tailed. When sampling $\pi \sim \mathrm{PY}(d, \alpha)$ for d near 1, where $\boldsymbol{\pi}$ is likely to be heavy-tailed, N_s may need to be intractably large to assure $\sum_{i>N_s} \tilde{\pi}_i < \epsilon$. In such sitations, truncation may result in serverely biased samples of entropy. We address this problem by directly estimating the entropy of the tail, $\mathrm{PY}(d, \alpha + N_s d)$, using (17). As shown in Fig. 3, the prior variance of PY becomes arbitrarily small as for large α. For sampling, N_s need only be large enough to make the variance of the tail entropy small. The resulting sample is the entropy of the (finite) samples plus the expected entropy of the tail, $H(\boldsymbol{\pi}^*) + \mathbb{E}[H|d, \alpha + Kd]$.[5] Sampling entropy is most useful for very small amounts of data drawn from distributions with low expected entropy. In Fig. 5 we illustrate the posterior distributions of entropy in two simulated experiments. In general, as the expected entropy and sample size increase, the posterior becomes more approximately Gaussian.

5. Theoretical properties of PYM

Having defined PYM and discussed its practical computation, we now consider first establish conditions under which (24) is defined (that is, finite), and also prove some basic facts about its asymptotic properties. While $\hat{H}_{\mathrm{PYM}}$ is a Bayesian estimator, we wish to build connection to the literature by showing frequentist properties. Note that the prior expectation $\mathbb{E}[H]$ does not exist for the improper prior defined above, since $p(h = \mathbb{E}[H]) \propto 1$ on $[0, \infty]$. It is therefore reasonable to ask what conditions on the data are sufficient to obtain finite posterior expectation $\hat{H}_{\mathrm{PYM}} = E[H|\mathbf{x}]$. We give an answer to this question in the following short proposition (proofs of all statements may be found in the appendix),

Theorem 2 *Given a fixed dataset* $\mathbf{x}$ *of* N *samples,* $\hat{H}_{PYM} < \infty$ *for any prior distribution* $p(d, \alpha)$ *if* $N - K \geq 2$.

In other words, we require 2 coincidences in the data for $\hat{H}_{\mathrm{PYM}}$ to be finite. When no coincidences have occurred in $\mathbf{x}$, we have no evidence regarding the support of the $\boldsymbol{\pi}$, and our resulting entropy estimate is unbounded. In fact, in the absence of coincidences, no entropy estimator can give a reasonable estimate without prior knowledge or assumptions about $\mathcal{A}$. Concerns about inadequate numbers of coincidences are peculiar to the undersampled regime; as we collect more data, we will almost surely observe each letter infinitely often. We now turn to asymptotic considerations, establishing consistency of $\hat{H}_{\mathrm{PYM}}$ in the limit of large N for a broad class of distributions. It is known that the plugin is consistent for any distribution (finite or countably infinite), although the rate of convergence can be arbitrarily slow (Antos and Kontoyiannis, 2001). Therefore, we establish consistency by showing asymptotic convergence to the plugin estimator. For clarity, we explicitly denote a quantity's dependence upon sample size N by a introducing a subscript. Thus, $\mathbf{x}$ and K become $\mathbf{x}_N$ and K_N, respectively. As a first step, we show that $\mathbb{E}[H|\mathbf{x}_N, d, \alpha]$ converges to the plugin estimator.

Theorem 3 *Assuming* $\mathbf{x}_N$ *drawn from a fixed, finite or countably infinite discrete distribution* $\boldsymbol{\pi}$ *such that* $\frac{K_N}{N} \xrightarrow{P} 0$,

$$|\mathbb{E}[H|\mathbf{x}_N, d, \alpha] - \mathbb{E}[H_{\mathrm{plugin}}|\mathbf{x}_N]| \xrightarrow{P} 0$$

The assumption $K_N/N \to 0$ is more general than it may seem. For any discrete distribution it holds that $K_N \to \mathbb{E}[K_N]$ a.s., and $\mathbb{E}[K_N]/N \to 0$ a.s. (Gnedin et al., 2007), and so $K_N/N \to 0$ in probability for an arbitrary distribution. As a result, (20) shares its asymptotic behavior with $\hat{H}_{\mathrm{plugin}}$, in particular consistency. As (20) is consistent for each value of α and d, it is intuitively

4. Bounds on N_s, the number of samples (i.e., sticks), necessary to reach ϵ on average are provided in (Ishwaran and James, 2001).
5. Due to the generality of the expecation formula (15), this method may be applied to sample the distributions of other additive functionals of PY.

plausible that PYM, as a mixture of such values, should be consistent as well. However, while (20) alone is well-behaved, it is not clear that PYM should be. Since $\mathbb{E}[H|\mathbf{x}, d, \alpha] \to \infty$ as $\alpha \to \infty$, care must be taken when integrating over $p(d, \alpha|\mathbf{x})$. Our main consistency result is,

Theorem 4 *For any proper prior or bounded improper prior $p(d, \alpha)$, if data $\mathbf{x}_N$ are drawn from a fixed, countably infinite discrete distribution $\boldsymbol{\pi}$ such that for some constant $C > 0$, $K_N = o(N^{1-1/C})$ in probability, then*

$$|\mathbb{E}[H|\mathbf{x}_N] - \mathbb{E}[H_{\text{plugin}}|\mathbf{x}_N]| \xrightarrow{P} 0$$

Intuitively, the asymptotic behavior of K_N/N is tightly related to the tail behavior of the distribution (Gnedin et al., 2007). In particular, $K_N \sim cN^a$ with $0 < a < 1$ if and only if $\pi_i \sim c' i^{\frac{1}{a}}$ where c and c' are constants (Gnedin et al., 2007). The class of distributions such that $K_N = o(N^{1-1/C})$ a.s. includes the class of power-law or thinner tailed distributions, i.e., $\pi_i = O(i^a)$ for some $a > 1$. We conclude this section with some remarks on the role of the prior in Theorem 4 as well as the significance of asymptotic results in general. While consistency is an important property for any estimator, we emphasize that PYM is designed to address the undersampled regime. Indeed, since $\hat{H}_{\text{plugin}}$ is consistent and has an optimal rate of convergence for a large class of distributions (Antos and Kontoyiannis, 2001; Vu et al., 2007; Zhang, 2012), asymptotic properties provide little reason to use $\hat{H}_{\text{PYM}}$. Nevertheless, notice that Theorem 4 makes very weak assumptions about $p(d, \alpha)$. In particular, the result is not dependant upon the form of the PYM prior introduced in the previous section: it holds for any probability distribution $p(d, \alpha)$, or even a bounded improper prior. Thus, we can view Theorem 4 as a statement about a class of PYM estimators. Almost any prior we choose on (d, α) results in a consistent estimator of entropy.

6. Results

We compare $\hat{H}_{\text{PYM}}$ to other proposed entropy estimators using several example datasets. Each plot in Figs 7, 8, 9, and 10 shows convergence as well as small sample performance. We compare our estimators, DPM ($d = 0$ only) and PYM ($\hat{H}_{\text{PYM}}$), with other enumerable-support estimators: coverage-adjusted estimator (CAE) (Chao and Shen, 2003; Vu et al., 2007), asymptotic NSB (ANSB, section 2.4) (Nemenman, 2011), James-Stein (JS) (Hausser and Strimmer, 2009), Grassberger's asymptotic bias correction (GR08) (Grassberger, 2008), and Good-Turing estimator (Zhang, 2012). Note that like ANSB, DPM is an asymptotic (Poisson-Dirichlet) limit of NSB, and hence behaves close to NSB assuming a large number of symbols. We also compare with plugin (3) and a standard bias correction methods assuming finite support: Miller-Maddow bias correction (MiMa) (Miller, 1955). To make comparisons more straightforward, we do not apply jackknife-based bias correction to any of the estimators. PYM performs well as expected when the data are truly generated by a Pitman-Yor process (Fig. 7). Credible intervals for DPM tend to be smaller than PYM, although both shrink quickly (indicating high confidence). When the tail of the distribution is exponentially decaying, ($d = 0$ case; Fig. 7 top), DPM shows slightly improved performance. When the tail has a strong power-law decay, (Fig. 7 bottom), PYM performs better than DPM. Most of the other estimators are consistently biased down, with the exception of JS and ANSB. Although Pitman-Yor process PY(d, α) has a power-law tail controlled by d, the high probability portion is modulated by α, and does not strictly folllow a power-law distribution as a whole. In Fig. 8, we evaluate the performance for $p_i \propto i^{-2}$ and $p_i \propto i^{-1.5}$. PYM and DPM has slight negative bias, but the credible interval covers the true entropy for all sample sizes. For small sample sizes, most estimators are negatively biased, again except for JS and ANSB (which does not show up in the plot since it is severely biased upwards). Notably CAE performs very well in moderate sample sizes. In Fig. 9, we compute the entropy per word of in the novel *Moby Dick* by Herman Melville, and entropy per time bin of a population of retinal ganglion cells from monkey retina (Pillow et al., 2005). These real-world

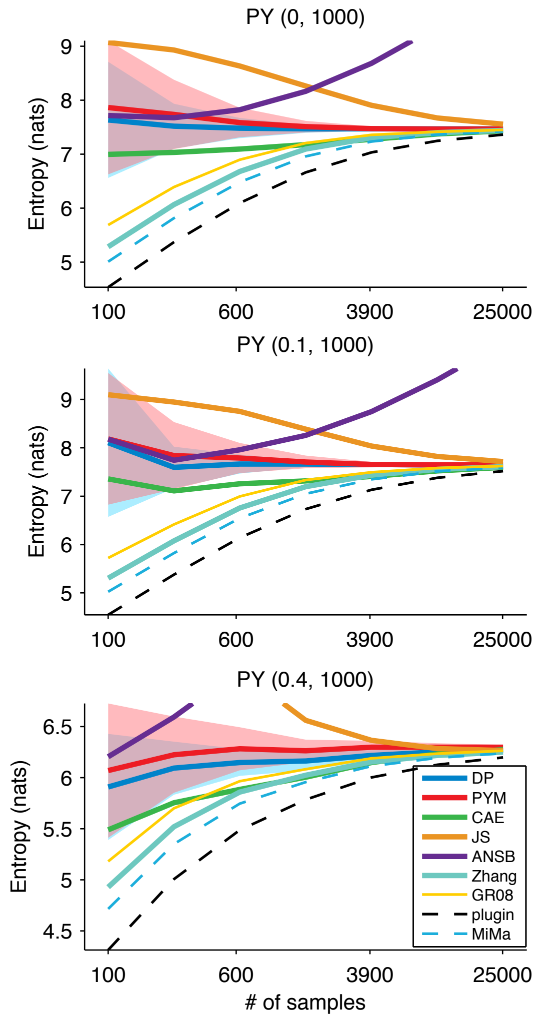

Figure 7: Comparison of estimators on stick-breaking distributions. Poisson-Dirichlet distribution with $(d = 0, \alpha = 1000)$ (top), $(d = 0.1, \alpha = 1000)$ (middle), $(d = 0.4, \alpha = 100)$ (bottom). We compare our estimators (DP, PYM) with other enumerable support estimators (CAE, ANSB, JS, Zhang, GR08), and finite support estimators (plugin, MiMa). Solid lines are averaged over 10 realizations. Shaded area represent two standard deviations credible intervals averaged over 10 realizations.

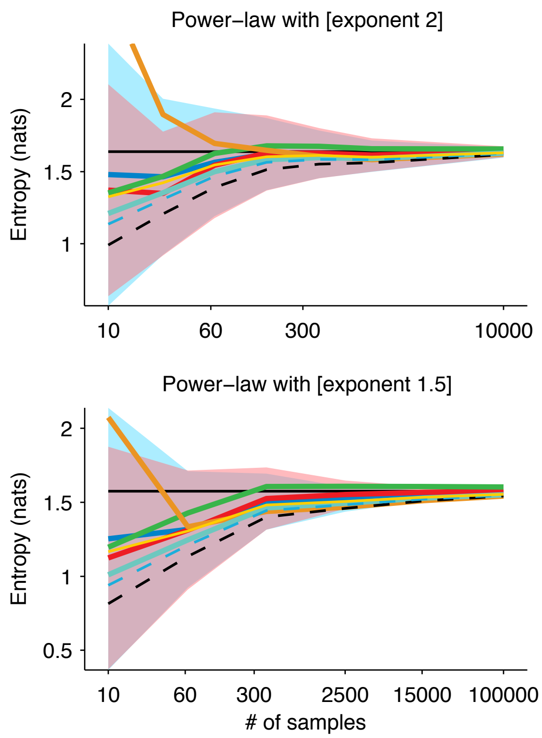

Figure 8: Comparison of estimators on power-law distributions.

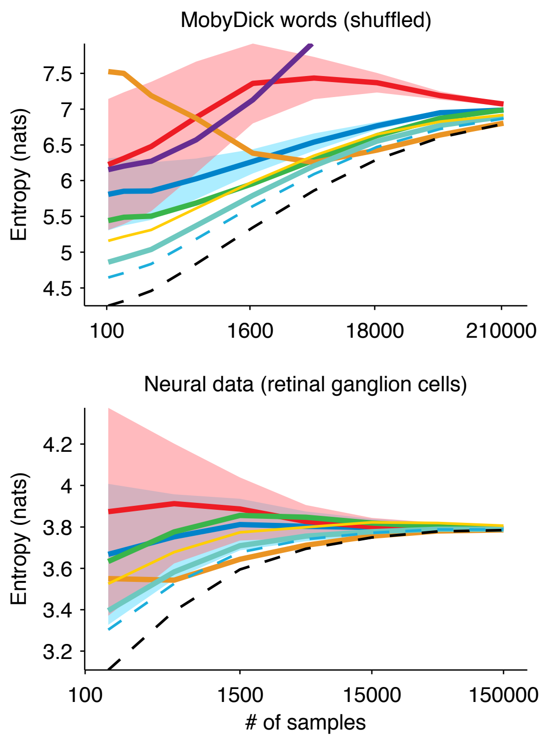

Figure 9: Comparison of estimators on real data sets.

datasets have heavy, approximately power-law tails [6] as pointed out earlier in Fig. 2. For Moby Dick, PYM slightly overestimates, while DPM slightly underestimates, yet either method is closer to the entropy estimated by the full data available than other estimators. DPM is overly confident (its credible interval is too narrow), while PYM becomes overly confident with more data. The neural data were preprocessed to be a binarized response (10 ms time bins) of 8 simultaneously recorded off-response retinal ganglion cells. PYM, DPM, and CAE all perform well on this dataset, with both PYM and DPM bracketing the asymptotic value with their credible intervals. Finally, we applied the denumerable support estimators to finite support distributions (Fig. 10). The power-law $p_n \propto n^{-1}$ has the heaviest tail among the simulations we consider, but notice that it does not define a proper distribution (the probability mass does not integrate), and so we use a truncated $1/n$ distribution with the first 1000 symbols (Fig. 10 top). Initially PYM shows the least bias, but DPM provides a better estimate for increasing sample size. Notice, however, that for both estimates the credible intervals consistently cover the true entropy. Interestingly, the finite support estimators perform poorly compared to DPM, CAE and PYM. For the uniform distribution over 1000 symbols, both DPM and PYM have slight upward bias, while CAE shows almost perfect performance (Fig. 10 middle). For Poisson distribution, a theoretically enumerable support distribution on the natural number, the tail decays so quickly that the effective support (due to machine precision) is very small (26 in this case). All the estimators, with the exception of JS and ANSB, work quite well. Note that JS performs poorly for both uniform and Poisson distribution (it shows severe upward biased). The novel Moby Dick provides the most challenging data: no estimator seems to have converged, even with the full data. Surprisingly, the Good-Turing estimator (Zhang, 2012) tends to perform similarly to the Grassberger and Miller-Maddow bias-correction methods. Among such the bias-correction methods, Grassberger's method tended to show the best performance, outperforming Zhang's method.

7. Conclusion

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 DP and PY prior for fixed parameters. Inspired by the work of (Nemenman et al., 2002), we defined a novel 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 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. A MATLAB implementation of the PYM estimator is available at `https://github.com/pillowlab/PYMentropy`.

Appendix A. Derivations of Dirichlet and PY moments

In this Appendix we present as propositions a number of technical moment derivations used in the text.

6. We emphasize that we use the term "power-law" in a heuristic, descriptive sense only. We did not fit explicit power-law models to the datasets in questions, and neither do we rely upon the properties of power-law distributions in our analyses.

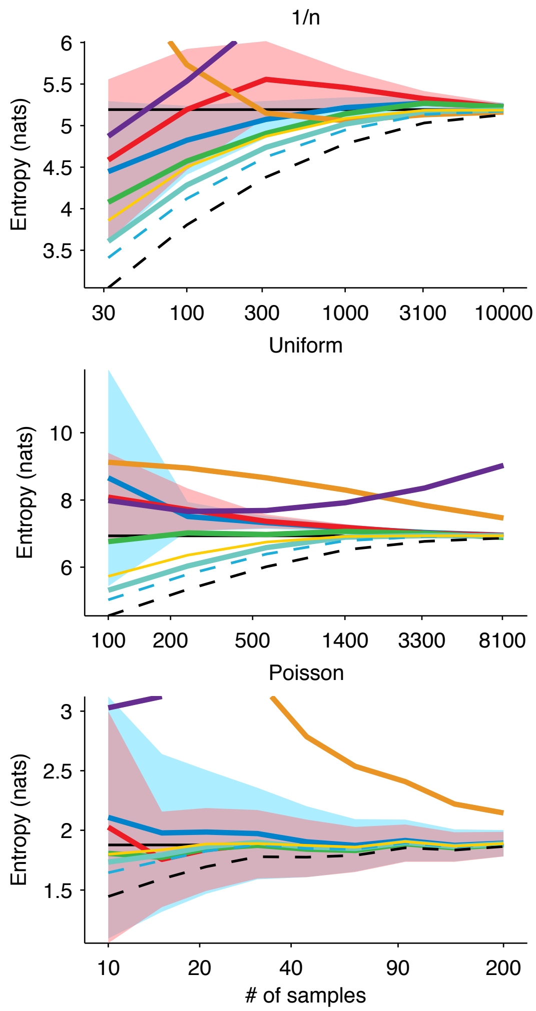

Figure 10: Comparison of estimators on finite support distributions. Black solid line indicates the true entropy. Poisson distribution ($\lambda = e$) has a countably infinite tail, but a very thin one—all probability mass was concentrated in 26 symbols within machine precision.

A.1 Mean entropy of finite Dirichlet

Proposition 5 (Replica trick for entropy (Wolpert and Wolf, 1995)) *For $\boldsymbol{\pi} \sim \mathrm{Dir}(\alpha_1, \alpha_2, \ldots, \alpha_{\mathcal{A}})$, such that $\sum_{i=1}^{\mathcal{A}} \alpha_i = A$, and letting $\vec{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_{\mathcal{A}})$, we have*

$$rCl\mathbb{E}[H(\boldsymbol{\pi})|\vec{\alpha}] \quad = \quad \psi_0(A+1) - \sum_{i=1}^{\mathcal{A}} \frac{\alpha_i}{A}\psi_0(\alpha_i+1) \tag{26}$$

Proof First, let c be the normalizer of Dirichlet, $c = \frac{\prod_j \Gamma(\alpha_j)}{\Gamma(A)}$ and let $\mathcal{L}$ denote the Laplace transform (on π to s). Now,

$$\begin{aligned}
c\mathbb{E}[H|\vec{\alpha}] &= \int \left(-\sum_i \pi_i \log_2 \pi_i\right) \delta(\textstyle\sum_i \pi_i - 1) \displaystyle\prod_j \pi_j^{\alpha_j - 1} d\boldsymbol{\pi} \\
&= -\sum_i \int \left(\pi_i^{\alpha_i} \log_2 \pi_i\right) \delta(\textstyle\sum_i \pi_i - 1) \displaystyle\prod_{j\neq i} \pi_j^{\alpha_j - 1} d\boldsymbol{\pi} \\
&= -\sum_i \int \left(\frac{d}{d(\alpha_i)} \pi_i^{\alpha_i}\right) \delta(\textstyle\sum_i \pi_i - 1) \displaystyle\prod_{j\neq i} \pi_j^{\alpha_j - 1} d\boldsymbol{\pi} \\
&= -\sum_i \frac{d}{d(\alpha_i)} \int \pi_i^{\alpha_i} \delta(\textstyle\sum_i \pi_i - 1) \displaystyle\prod_{j\neq i} \pi_j^{\alpha_j - 1} d\boldsymbol{\pi} \\
&= -\sum_i \frac{d}{d(\alpha_i)} \mathcal{L}^{-1}\left[\mathcal{L}(\pi_i^{\alpha_i}) \prod_{j\neq i} \mathcal{L}(\pi_j^{\alpha_j - 1})\right](1) \\
&= -\sum_i \frac{d}{d(\alpha_i)} \mathcal{L}^{-1}\left[\frac{\Gamma(\alpha_i+1)\prod_{j\neq i}\Gamma(\alpha_j)}{s^{\sum_k(\alpha_k)+1}}\right](1) \\
&= -\sum_i \frac{d}{d(\alpha_i)} \left[\frac{\Gamma(\alpha_i+1)}{\Gamma(\sum_k(\alpha_k)+1)}\right] \prod_{j\neq i}\Gamma(\alpha_j) \\
&= -\sum_i \frac{\Gamma(\alpha_i+1)}{\Gamma(\sum_k \alpha_k+1)} \left[\psi_0(\alpha_i+1) - \psi_0(A+1)\right] \prod_{j\neq i}\Gamma(\alpha_j) \\
&= \left[\psi_0(A+1) - \sum_{i=1}^{\mathcal{A}} \frac{\alpha_i}{A}\psi_0(\alpha_i+1)\right] \frac{\prod_j \Gamma(\alpha_j)}{\Gamma(A)}.
\end{aligned}$$

■

A.2 Variance entropy of finite Dirichlet

We derive $\mathbb{E}[H^2(\boldsymbol{\pi})|\vec{\alpha}]$. In practice we compute $\mathrm{var}[H(\boldsymbol{\pi})|\vec{\alpha}] = \mathbb{E}[H^2(\boldsymbol{\pi})|\vec{\alpha}] - \mathbb{E}[H(\boldsymbol{\pi})|\vec{\alpha}]^2$.

Proposition 6 *For $\boldsymbol{\pi} \sim \mathrm{Dir}(\alpha_1, \alpha_2, \ldots, \alpha_{\mathcal{A}})$, such that $\sum_{i=1}^{\mathcal{A}} \alpha_i = A$, and letting $\vec{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_{\mathcal{A}})$, we have*

$$\begin{aligned}
\mathbb{E}[H^2(\boldsymbol{\pi})|\vec{\alpha}] &= \sum_{i\neq k} \frac{\alpha_i \alpha_k}{(A+1)(A)} I_{ik} + \sum_i \frac{\alpha_i(\alpha_i+1)}{(A+1)(A)} J_i \qquad (27) \\
I_{ik} &= \left(\psi_0(\alpha_k+1) - \psi_0(A+2)\right)\left(\psi_0(\alpha_i+1)\right. \\
&\quad \left.-\psi_0(A+2)\right) - \psi_1(A+2)
\end{aligned}$$

$$J_i = (\psi_0(\alpha_i + 2) - \psi_0(A + 2))^2 + \psi_1(\alpha_i + 2) \\ - \psi_1(A + 2)$$

Proof We can evaluate the second moment in a manner similar to the mean entropy above. First, we split the second moment into square and cross terms. To evaluate the integral over the cross terms, we apply the "replica trick" twice. Letting c be the normalizer of Dirichlet, $c = \frac{\prod_j \Gamma(\alpha_j)}{\Gamma(A)}$ we have,

$$\begin{aligned}
c\mathbb{E}[H^2|\vec{\alpha}] &= \int \left(-\sum_i \pi_i \log_2 \pi_i \right)^2 \delta(\textstyle\sum_i \pi_i - 1) \displaystyle\prod_j \pi_j^{\alpha_j - 1} d\boldsymbol{\pi} \\
&= \sum_i \int \left(\pi_i^2 \log_2{}^2 \pi_i \right) \delta(\textstyle\sum_i \pi_i - 1) \displaystyle\prod_j \pi_j^{\alpha_j - 1} d\boldsymbol{\pi} \\
&+ \sum_{i \neq k} \int \left(\pi_i \log_2 \pi_i \right) \left(\pi_k \log_2 \pi_k \right) \delta(\textstyle\sum_i \pi_i - 1) \displaystyle\prod_j \pi_j^{\alpha_j - 1} d\boldsymbol{\pi} \\
&= \sum_i \int \pi_i^{\alpha_i + 1} \log_2{}^2 \pi_i \delta(\textstyle\sum_i \pi_i - 1) \displaystyle\prod_{j \neq i} \pi_j^{\alpha_j - 1} d\boldsymbol{\pi} \\
&+ \sum_{i \neq k} \int \left(\pi_i^{\alpha_i} \log_2 \pi_i \right) \left(\pi_k^{\alpha_k} \log_2 \pi_k \right) \delta(\textstyle\sum_i \pi_i - 1) \displaystyle\prod_{j \notin \{i,k\}} \pi_j^{\alpha_j - 1} d\boldsymbol{\pi} \\
&= \sum_i \frac{d^2}{d(\alpha_i + 1)^2} \int \pi_i^{\alpha_i + 1} \delta(\textstyle\sum_i \pi_i - 1) \displaystyle\prod_{j \neq i} \pi_j^{\alpha_j - 1} d\boldsymbol{\pi} \\
&+ \sum_{i \neq k} \frac{d}{d\alpha_i} \frac{d}{d\alpha_k} \int \left(\pi_i^{\alpha_i} \right) \left(\pi_k^{\alpha_k} \right) \delta(\textstyle\sum_i \pi_i - 1) \displaystyle\prod_{j \notin \{i,k\}} \pi_j^{\alpha_j - 1} d\boldsymbol{\pi}
\end{aligned}$$

Assuming $i \neq k$, these will be the cross terms.

$$\begin{aligned}
&\int (\pi_i \log_2 \pi_i)(\pi_k \log_2 \pi_k) \delta(\textstyle\sum_i \pi_i - 1) \displaystyle\prod_j \pi_j^{\alpha_j - 1} d\boldsymbol{\pi} \\
&= \frac{d}{d\alpha_i} \frac{d}{d\alpha_k} \int (\pi_i^{\alpha_i})(\pi_k^{\alpha_k}) \delta(\textstyle\sum_i \pi_i - 1) \displaystyle\prod_{j \notin \{i,k\}} \pi_j^{\alpha_j - 1} d\boldsymbol{\pi} \\
&= \frac{d}{d\alpha_i} \frac{d}{d\alpha_k} \left[\frac{\Gamma(\alpha_i + 1)\Gamma(\alpha_k + 1)}{\Gamma(A + 2)} \right] \prod_{j \notin \{i,k\}} \Gamma(\alpha_j) \\
&= \frac{d}{d\alpha_k} \left[\frac{\alpha_i \Gamma(\alpha_k + 1)}{\Gamma(A + 2)} \psi_0(\alpha_i + 1) \right. \\
&\qquad \left. - \frac{\alpha_i \Gamma(\alpha_k + 1)}{\Gamma(A + 2)} \psi_0(A + 2) \right] \prod_{j \neq k} \Gamma(\alpha_j) \\
&= \frac{d}{d\alpha_k} \left[\frac{\alpha_i \psi_0(\alpha_k + 1)}{\Gamma(A + 2)} \psi_0(\alpha_i + 1) \right. \\
&\qquad \left. - \frac{\alpha_i \Gamma(\alpha_k + 1)}{\Gamma(A + 2)} \psi_0(A + 2) \right] \prod_{j \neq k} \Gamma(\alpha_j) \\
&= \frac{\alpha_i \alpha_k}{\Gamma(A + 2)} \left[(\psi_0(\alpha_k + 1) - \psi_0(A + 2)) \right. \\
&\qquad \left. (\psi_0(\alpha_i + 1) - \psi_0(A + 2)) - \psi_1(A + 2) \right] \prod_j \Gamma(\alpha_j)
\end{aligned}$$

$$= \frac{\alpha_i \alpha_k}{(A+1)(A)} \left[(\psi_0(\alpha_k + 1) - \psi_0(A+2)) \right.$$
$$\left. (\psi_0(\alpha_i + 1) - \psi_0(A+2)) - \psi_1(A+2) \right] \frac{\prod_j \Gamma(\alpha_j)}{\Gamma(A)}$$

$$\frac{d^2}{d(\alpha_i+1)^2} \int \pi_i^{\alpha_i+1} \delta(\textstyle\sum_i \pi_i - 1) \displaystyle\prod_{j \neq i} \pi_j^{\alpha_j - 1} d\boldsymbol{\pi}$$
$$= \frac{d^2}{d(\alpha_i+1)^2} \left[\frac{\Gamma(\alpha_i+2)}{\Gamma(A+2)} \right] \prod_{j \neq i} \Gamma(\alpha_j)$$
$$= \frac{d^2}{dz^2} \left[\frac{\Gamma(z+1)}{\Gamma(z+c)} \right] \prod_{j \neq i} \Gamma(\alpha_j), \quad \{c = A + 2 - (\alpha_i + 1)\}$$
$$= \frac{\Gamma(1+z)}{\Gamma(c+z)} \left[(\psi_0(1+z) - \right.$$
$$\left. \psi_0(c+z))^2 + \psi_1(1+z) - \psi_1(c+z) \right] \prod_{j \neq i} \Gamma(\alpha_j)$$
$$= \frac{z(z-1)}{(c+z-1)(c+z-2)} \left[(\psi_0(1+z) - \psi_0(c+z))^2 \right.$$
$$\left. + \psi_1(1+z) - \psi_1(c+z) \right] \frac{\prod_j \Gamma(\alpha_j)}{\Gamma(A)}$$
$$= \frac{(\alpha_i+1)(\alpha_i)}{(A+1)(A)} \left[(\psi_0(\alpha_i+2) - \psi_0(A+2))^2 + \psi_1(\alpha_i+2) \right.$$
$$\left. - \psi_1(A+2) \right] \frac{\prod_j \Gamma(\alpha_j)}{\Gamma(A)}$$

Summing over all terms and adding the cross and square terms, we recover the desired expression for $\mathbb{E}[H^2(\boldsymbol{\pi})|\vec{\alpha}]$. ■

A.3 Prior entropy mean and variance under PY

We derive the prior entropy mean and variance of a PY distribution with fixed parameters α and d, $\mathbb{E}_{\boldsymbol{\pi}}[H(\boldsymbol{\pi})|d,\alpha]$ and var $\boldsymbol{\pi}[H(\boldsymbol{\pi})|d,\alpha]$. We first prove our Proposition 1 (mentioned in (Pitman and Yor, 1997)). This proposition establishes the identity $\mathbb{E}\left[\sum_{i=1}^{\infty} f(\pi_i) \middle| \alpha\right] = \int_0^1 \frac{f(\tilde{\pi}_1)}{\tilde{\pi}_1} p(\tilde{\pi}_1|\alpha) d\tilde{\pi}_1$ which will allow us to compute expectations over PY using only the distribution of the first size biased sample, $\tilde{\pi}_1$.

Proof [Proof of Proposition 1] First we validate (15). Writing out the general form of the size-biased sample,

$$p(\tilde{\pi}_1 = x|\boldsymbol{\pi}) = \sum_{i=1}^{\infty} \pi_i \delta(x - \pi_i),$$

we see that

$$\mathbb{E}_{\tilde{\pi}_1}\left[\frac{f(\tilde{\pi}_1)}{\tilde{\pi}_1}\right] = \int_0^1 \frac{f(x)}{x} p(\tilde{\pi}_1 = x) dx$$
$$= \int_0^1 \mathbb{E}_{\boldsymbol{\pi}}\left[\frac{f(x)}{x} p(\tilde{\pi}_1 = x|\boldsymbol{\pi})\right] dx$$

$$
\begin{aligned}
&= \int_0^1 \mathbb{E}_{\boldsymbol{\pi}} \left[\sum_{i=1}^{\infty} \frac{f(x)}{x} \pi_i \delta(x - \pi_i) \right] dx \\
&= \mathbb{E}_{\boldsymbol{\pi}} \left[\int_0^1 \sum_{i=1}^{\infty} \frac{f(x)}{x} \pi_i \delta(x - \pi_i) dx \right] \\
&= \mathbb{E}_{\boldsymbol{\pi}} \left[\sum_{i=1}^{\infty} \int_0^1 \frac{f(x)}{x} \pi_i \delta(x - \pi_i) dx \right] \\
&= \mathbb{E}_{\boldsymbol{\pi}} \left[\sum_{i=1}^{\infty} f(\pi_i) \right],
\end{aligned}
$$

where the interchange of sums and integrals is justified by Fubini's theorem. A similar method validates (16). We will need the second size-biased sample in addition to the first. We begin with the sum inside the expectation on the left hand side of (16),

$$
\begin{aligned}
&\sum_i \sum_{j \neq i} g(\pi_i, \pi_j) && (28) \\
&= \frac{\sum_i \sum_{j \neq i} g(\pi_i, \pi_j)}{p(\tilde{\pi}_1 = \pi_i, \tilde{\pi}_2 = \pi_j)} p(\tilde{\pi}_1 = \pi_i, \tilde{\pi}_2 = \pi_j) && (29) \\
&= \sum_i \sum_{j \neq i} \frac{g(\pi_i, \pi_j)}{\pi_i \pi_j} (1 - \pi_i) p(\tilde{\pi}_1 = \pi_i, \tilde{\pi}_2 = \pi_j) && (30) \\
&= \mathbb{E}_{\tilde{\pi}_1, \tilde{\pi}_2} \left[\frac{g(\tilde{\pi}_1, \tilde{\pi}_2)}{\tilde{\pi}_1 \tilde{\pi}_2} (1 - \tilde{\pi}_1) \middle| \boldsymbol{\pi} \right] && (31)
\end{aligned}
$$

where the joint distribution of size biased samples is given by,

$$
\begin{aligned}
p(\tilde{\pi}_1 = \pi_i, \tilde{\pi}_2 = \pi_j) &= p(\tilde{\pi}_1 = \pi_i) p(\tilde{\pi}_2 = \pi_j | \tilde{\pi}_1 = \pi_i) \\
&= \pi_i \cdot \frac{\pi_j}{1 - \pi_i}.
\end{aligned}
$$

■

As this identity is defined for any additive functional f of $\boldsymbol{\pi}$; we can employ it to compute the first two moments of entropy. For PYP (and DP when $d = 0$), the first size-biased sample is distributed according to:

$$\tilde{\pi}_1 \sim \text{Beta}(1 - d, \alpha + d) \tag{32}$$

Proposition 1 gives the mean entropy directly. Taking $f(x) = -x \log(x)$ we have,

$$\mathbb{E}[H(\boldsymbol{\pi}) | d, \alpha] = -\mathbb{E}_{\alpha}[\log(\pi_1)] = \psi_0(\alpha + 1) - \psi_0(1 - d),$$

The same method may be used to obtain the prior variance, although the computation is more involved. For the variance, we will need the second size-biased sample in addition to the first. The second size-biased sample is given by,

$$\tilde{\pi}_2 = (1 - \tilde{\pi}_1) v_2, \quad v_2 \sim \text{Beta}(1 - d, \alpha + 2d) \tag{33}$$

We will compute the second moment explicitly, splitting $H(\boldsymbol{\pi})^2$ into square and cross terms,

$$\mathbb{E}[(H(\boldsymbol{\pi}))^2 \,|d,\alpha] = \mathbb{E}\left[\left(-\sum_i \pi_i \log(\pi_i)\right)^2 \middle| d,\alpha\right]$$

$$= \mathbb{E}\left[\sum_i (\pi_i \log(\pi_i))^2 \middle| d,\alpha\right] \tag{34}$$

$$+ \mathbb{E}\left[\sum_i \sum_{j\neq i} \pi_i \pi_j \log(\pi_i)\log(\pi_j) \middle| d,\alpha\right] \tag{35}$$

The first term follows directly from (15),

$$\mathbb{E}\left[\sum_i (\pi_i \log(\pi_i))^2 \middle| d,\alpha\right] = \int_0^1 x(-\log(x))^2 p(x|d,\alpha)\,\mathrm{d}x$$

$$= B^{-1}(1-d,\alpha+d)\int_0^1 x\log^2(x) x^{1-d}(1-x)^{\alpha+d-1}\,\mathrm{d}x$$

$$= \frac{1-d}{\alpha+1}\left[(\psi_0(2-d)-\psi_0(2+\alpha))^2 + \psi_1(2-d) - \psi_1(2+\alpha)\right] \tag{36}$$

The second term of (35), requires the first two size biased samples, and follows from (16) with $g(x,y) = \log(x)\log(y)$. For the PYP prior, it is easier to integrate on V_1 and V_2, rather than the size biased samples. The second term is then (note that we let $\gamma = B^{-1}(1-d,\alpha+2d)$ and $\zeta = B^{-1}(1-d,\alpha+d)$),

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}\left[\log(\tilde{\pi}_1)\log(\tilde{\pi}_2)(1-\pi_1)|\boldsymbol{\pi}\right]|\alpha\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\log(V_1)\log((1-V_1)V_2)(1-V_1)|\boldsymbol{\pi}\right]|\alpha\right] \\
&= \zeta\gamma \int_0^1\int_0^1 \log(v_1)\log((1-v_1)v_2)(1-v_1)v_1^{1-d}(1-v_1)^{\alpha+d-1} \\
&\qquad\qquad \times v_2^{1-d}(1-v_2)^{\alpha+2d-1}\,\mathrm{d}v_1\,\mathrm{d}v_2 \\
&= \zeta\left[\int_0^1 \log(v_1)\log(1-v_1)(1-v_1)v_1^{1-d}(1-v_1)^{\alpha+d-1}\,\mathrm{d}v_1\right. \\
&\quad + \gamma\int_0^1 \log(v_1)(1-v_1)v_1^{1-d}(1-v_1)^{\alpha+d-1} \\
&\qquad\qquad \left.\times \int_0^1 \log(v_2)v_2^{1-d}(1-v_2)^{\alpha+2d-1}\,\mathrm{d}v_1\,\mathrm{d}v_2\right] \\
&= \frac{\alpha+d}{\alpha+1}\left[(\psi_0(1-d)-\psi_0(2+\alpha))^2 - \psi_1(2+\alpha)\right]
\end{aligned}$$

Finally combining the terms, the variance of the entropy under PYP prior is

$$\begin{aligned}
&\mathrm{var}[H(\boldsymbol{\pi})|d,\alpha] = \qquad (37)\\
&\frac{1-d}{\alpha+1}\left[(\psi_0(2-d)-\psi_0(2+\alpha))^2 + \psi_1(2-d)-\psi_1(2+\alpha)\right] \\
&\qquad + \frac{\alpha+d}{\alpha+1}\left[(\psi_0(1-d)-\psi_0(2+\alpha))^2 - \psi_1(2+\alpha)\right] \\
&\qquad - (\psi_0(1+\alpha)-\psi_0(1-d))^2 \\
&= \frac{\alpha+d}{(\alpha+1)^2(1-d)} + \frac{1-d}{\alpha+1}\psi_1(2-d) - \psi_1(2+\alpha) \qquad (38)
\end{aligned}$$

We note that the expectations over the finite Dirichlet may also be derived using this formula by letting the $\tilde{\boldsymbol{\pi}}$ be the first size-biased sample of a finite Dirichlet on $\Delta_{\mathcal{A}}$.

A.4 Posterior Moments of PYP

First, we discuss the form of the PYP posterior, and introduce independence properties that will be important in our derivation of the mean. We recall that the PYP posterior, $\boldsymbol{\pi}_{\text{post}}$, of (19) has three stochastically independent components: Bernoulli p_*, PY $\boldsymbol{\pi}$, and Dirichlet $\mathbf{p}$.

Component expectations: From the above derivations for expectations under the PYP and Dirichlet distributions as well as the Beta integral identities (see e.g., (Archer et al., 2012)), we find expressions for $\mathbb{E}_{\mathbf{p}}\left[H(\mathbf{p})|d,\alpha\right]$, $E_{\boldsymbol{\pi}}\left[H(\boldsymbol{\pi})|d,\alpha\right]$, and $\mathbb{E}_{p_*}\left[H(p_*)\right]$.

$$\begin{aligned}
\mathbb{E}[H(\boldsymbol{\pi})|d,\alpha] &= \psi_0(\alpha+1) - \psi_0(1-d) \\
\mathbb{E}_{p_*}[H(p_*)] &= \psi_0(\alpha+N+1) - \frac{\alpha+Kd}{\alpha+N}\psi_0(\alpha+Kd+1) \\
&\quad - \frac{N-Kd}{\alpha+N}\psi_0(N-Kd+1) \\
\mathbb{E}_{\mathbf{p}}[H(\mathbf{p})|d,\alpha] &= \psi_0(N-Kd+1) - \sum_{i=1}^{K}\frac{n_i-d}{N-Kd}\psi_0(n_i-d+1)
\end{aligned}$$

where by a slight abuse of notation we define the entropy of p_* as $H(p_*) = -(1-p_*)\log(1-p_*) - p_*\log p_*$. We use these expectations below in our computation of the final posterior integral.
Derivation of posterior mean: We now derive the analytic form of the posterior mean, (20).

$$\begin{aligned}
\mathbb{E}[H(\pi_{\text{post}})|d,\alpha] &= \mathbb{E}\left[-\sum_{i=1}^{K} p_i\log p_i - p_*\sum_{i=1}^{\infty}\pi_i\log p_*\pi_i \middle| d,\alpha\right] \\
&= \mathbb{E}\left[-(1-p_*)\sum_{i=1}^{K}\frac{p_i}{1-p_*}\log\left(\frac{p_i}{1-p_*}\right)\right. \\
&\qquad \left. -(1-p_*)\log(1-p_*) - p_*\sum_{i=1}^{\infty}\pi_i\log\pi_i - p_*\log p_* \middle| d,\alpha\right] \\
&= \mathbb{E}\left[-(1-p_*)\sum_{i=1}^{K}\frac{p_i}{1-p_*}\log\left(\frac{p_i}{1-p_*}\right)\right. \\
&\qquad \left. -p_*\sum_{i=1}^{\infty}\pi_i\log\pi_i + H(p_*) \middle| d,\alpha\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[-(1-p_*)\sum_{i=1}^{K}\frac{p_i}{1-p_*}\log\left(\frac{p_i}{1-p_*}\right)\right.\right. \\
&\qquad \left.\left. -p_*\sum_{i=1}^{\infty}\pi_i\log\pi_i + H(p_*) \;\middle|\; p_*\right] \middle| d,\alpha\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[(1-p_*)H(\mathbf{p}) + p_*H(\boldsymbol{\pi}) + H(p_*) \;\middle|\; p_*\right]\middle| d,\alpha\right] \\
&= \mathbb{E}_{p_*}\left[(1-p_*)\mathbb{E}_{\mathbf{p}}\left[H(\mathbf{p})|d,\alpha\right] + p_*\mathbb{E}_{\boldsymbol{\pi}}\left[H(\boldsymbol{\pi})|d,\alpha\right] + H(p_*)\right]
\end{aligned}$$

using the formulae for $\mathbb{E}_{\mathbf{p}}\left[H(\mathbf{p})|d,\alpha\right]$, $\mathbb{E}_{\boldsymbol{\pi}}\left[H(\boldsymbol{\pi})|d,\alpha\right]$, and $\mathbb{E}_{p_*}[H(p_*)]$ and rearranging terms, we obtain (20),

$$\mathbb{E}[H(\pi_{\text{post}})|d,\alpha] = \frac{A}{\alpha+N}\mathbb{E}_{\mathbf{p}}[H(\mathbf{p})]$$

$$
\begin{aligned}
&+ \frac{\alpha + Kd}{\alpha + N}\mathbb{E}_{\boldsymbol{\pi}}[H(\boldsymbol{\pi})] + \mathbb{E}_{p_*}[H(p_*)] \\
= &\frac{A}{\alpha + N}\left[\psi_0(A+1) - \sum_{i=1}^{K}\frac{\alpha_i}{A}\psi_0(\alpha_i + 1)\right] \\
&+ \frac{\alpha + Kd}{\alpha + N}\left[\psi_0(\alpha + Kd + 1) - \psi_0(1-d)\right] + \\
&\psi_0(\alpha + N + 1) - \frac{\alpha + Kd}{\alpha + N}\psi_0(\alpha + Kd + 1) - \frac{A}{\alpha + N}\psi_0(A+1) \\
= &\,\psi_0(\alpha + N + 1) - \frac{\alpha + Kd}{\alpha + N}\psi_0(1-d) - \\
&\qquad \frac{A}{\alpha + N}\left[\sum_{i=1}^{K}\frac{\alpha_i}{A}\psi_0(\alpha_i + 1)\right] \\
= &\,\psi_0(\alpha + N + 1) - \frac{\alpha + Kd}{\alpha + N}\psi_0(1-d) - \\
&\qquad \frac{1}{\alpha + N}\left[\sum_{i=1}^{K}(n_i - d)\psi_0(n_i - d + 1)\right]
\end{aligned}
$$

Derivation of posterior variance: We continue the notation from the subsection above. In order to exploit the independence properties of π_{post} we first apply the law of total variance to obtain (39),

$$
\begin{aligned}
\operatorname{var}[H(\pi_{\text{post}})|d,\alpha] = &\operatorname*{var}_{p_*}\left[\mathbb{E}_{\boldsymbol{\pi},\mathbf{p}}[H(\pi_{\text{post}})]\Big| d,\alpha\right] \\
&+ \mathbb{E}_{p_*}\left[\operatorname*{var}_{\boldsymbol{\pi},\mathbf{p}}[H(\pi_{\text{post}})]\Big| d,\alpha\right]
\end{aligned} \tag{39}
$$

We now seek expressions for each term in (39) in terms of the expectations already derived. *Step 1:* For the right-hand term of (39), we use the independence properties of π_{post} to express the variance in terms of PYP, Dirichlet, and Beta variances,

$$
\mathbb{E}_{p_*}\left[\operatorname*{var}_{\boldsymbol{\pi},\mathbf{p}}[H(\pi_{\text{post}})|p_*]\Big| d,\alpha\right] \tag{40}
$$

$$
\begin{aligned}
&= \mathbb{E}_{p_*}\left[(1-p_*)^2\operatorname*{var}_{\mathbf{p}}[H(\mathbf{p})] + p_*^2\operatorname*{var}_{\boldsymbol{\pi}}[H(\boldsymbol{\pi})]\Big| d,\alpha\right] \\
&= \frac{(N-Kd)(N-Kd+1)}{(\alpha+N)(\alpha+N+1)}\operatorname*{var}_{\mathbf{p}}[H(\mathbf{p})] \\
&\qquad + \frac{(\alpha+Kd)(\alpha+Kd+1)}{(\alpha+N)(\alpha+N+1)}\operatorname*{var}_{\boldsymbol{\pi}}[H(\boldsymbol{\pi})]
\end{aligned} \tag{41}
$$

Step 2: In the left-hand term of (39) the variance is with respect to the Beta distribution, while the inner expecation is precisely the posterior mean we derived above. Expanding, we obtain,

$$
\begin{aligned}
&\operatorname*{var}_{p_*}\left[\mathbb{E}_{\boldsymbol{\pi},\mathbf{p}}[H(\pi_{\text{post}})|p_*]\Big| d,\alpha\right] \\
&= \operatorname*{var}_{p_*}\left[(1-p_*)\mathbb{E}_{\mathbf{p}}\left[H(\mathbf{p})\right] + p_*\mathbb{E}_{\boldsymbol{\pi}}\left[H(\boldsymbol{\pi})|p_*\right] + h(p_*)\Big| d,\alpha\right]
\end{aligned} \tag{42}
$$

To evaluate this integral, we introduce some new notation,

$$\begin{aligned}\mathbf{A} &:= \mathbb{E}_{\mathbf{p}}\left[H(\mathbf{p})\right] \\ \mathbf{B} &:= \mathbb{E}_{\boldsymbol{\pi}}\left[H(\boldsymbol{\pi})\right] \\ \Omega(p_*) &:= (1-p_*)\mathbb{E}_{\mathbf{p}}\left[H(\mathbf{p})\right] + p_*\mathbb{E}_{\boldsymbol{\pi}}\left[H(\boldsymbol{\pi})\right] + h(p_*) \\ &= (1-p_*)\mathbf{A} + p_*\mathbf{B} + h(p_*)\end{aligned}$$

so that

$$\begin{aligned}\Omega^2(p_*) &= 2p_*h(p_*)[\mathbf{B}-\mathbf{A}] + 2\mathbf{A}h(p_*) + h^2(p_*) \\ &\quad + p_*^2[\mathbf{B}^2 - 2\mathbf{AB}] + 2p_*\mathbf{AB} + (1-p_*)^2\mathbf{A}^2 \qquad (43)\end{aligned}$$

and we note that

$$\operatorname*{var}_{p_*}\left[\mathbb{E}_{\boldsymbol{\pi},\mathbf{p}}[H(\pi_{\text{post}})|p_*]\middle| d,\alpha\right] = \mathbb{E}_{p_*}[\Omega^2(p_*)] - \mathbb{E}_{p_*}[\Omega(p_*)]^2 \qquad (44)$$

The components composing $\mathbb{E}_{p_*}[\Omega(p_*)]$, as well as each term of (43) can be found in (Archer et al., 2012). Although less elegant than the posterior mean, the expressions derived above permit us to compute (39) numerically from its component expecatations, without sampling.

Appendix B. Proof of Proposition 2

In this Appendix we give a proof for Proposition 2.
Proof PYM is given by

$$\hat{H}_{PYM} = \frac{1}{p(\mathbf{x})}\int_0^\infty \int_0^1 H_{(d,\alpha)} p(\mathbf{x}|d,\alpha)p(d,\alpha)\,\mathrm{d}\alpha\,\mathrm{d}d$$

where we have written $H_{(d,\alpha)} := \mathbb{E}[H|d,\alpha,\mathbf{x}]$. Note that $p(\mathbf{x}|d,\alpha)$ is the evidence, given by (25). We will assume $p(d,\alpha) = 1$ for all α and d to show conditions under which $H_{(d,\alpha)}$ is integrable for any prior. Using the identity $\frac{\Gamma(x+n)}{\Gamma(x)} = \prod_{i=1}^n (x+i-1)$ and the log convexity of the Gamma function we have,

$$\begin{aligned}p(\mathbf{x}|d,\alpha) &\leq \prod_{i=1}^K \frac{\Gamma(n_i - d)}{\Gamma(1-d)} \frac{\Gamma(\alpha+K)}{\Gamma(\alpha+N)} \\ &\leq \frac{\Gamma(n_i-d)}{\Gamma(1-d)}\frac{1}{\alpha^{N-K}}\end{aligned}$$

Since $d \in [0,1)$, we have from the properties of the digamma function,

$$\psi_0(1-d) = \psi_0(2-d) - \frac{1}{1-d} \geq \psi_0(1) - \frac{1}{1-d} = -\gamma - \frac{1}{1-d},$$

and thus the upper bound,

$$\begin{aligned}H_{(d,\alpha)} &\leq \psi_0(\alpha+N+1) + \frac{\alpha+Kd}{\alpha+N}\left(\gamma + \frac{1}{1-d}\right) \qquad (45) \\ &\quad - \frac{1}{\alpha+N}\left[\sum_{i=1}^K (n_i-d)\psi_0(n_i-d+1)\right]. \qquad (46)\end{aligned}$$

Although second term is unbounded in d notice that $\frac{\Gamma(n_i-d)}{\Gamma(1-d)} = \prod_{i=1}^{n_i}(i-d)$; thus, so long as $N-K \geq 1$, $H_{(\alpha,d)}p(\mathbf{x}|d,\alpha)$ is integrable in d. For the integral over alpha, it suffices to choose $\alpha_0 \gg N$ and consider the tail, $\int_{\alpha_0}^{\infty} H_{(d,\alpha)}p(\mathbf{x}|d,\alpha)p(d,\alpha)\,\mathrm{d}\alpha$. From (45) and the asymptotic expansion $\psi(x) = \log(x) - \frac{1}{2x} - \frac{1}{12x^2} + O(\frac{1}{x^4})$ as $x \to \infty$ we see that in the limit of $\alpha \gg N$,

$$H_{(d,\alpha)} \leq \log(\alpha + N + 2) + \frac{c}{\alpha},$$

where c is a constant depending on K, N, and d. Thus, we have

$$\begin{aligned}
&\int_{\alpha_0}^{\infty} H_{(d,\alpha)}p(\mathbf{x}|d,\alpha)p(d,\alpha)\,\mathrm{d}\alpha \\
&\qquad \leq \frac{\prod_{i=1}^{K}\Gamma(n_i-d)}{\Gamma(1-d)} \int_{\alpha_0}^{\infty} \left(\log(\alpha+N+2) + \frac{c}{\alpha}\right) \frac{1}{\alpha^{N-K}}\,\mathrm{d}\alpha
\end{aligned}$$

and so $H_{(d,\alpha)}$ is integrable in α so long as $N - K \geq 2$. ∎

Appendix C. Proofs of Consistency Results

Proof [proof of Theorem 3] We have,

$$\begin{aligned}
&\lim_{N\to\infty} \mathbb{E}[H|\alpha,d,\mathbf{x}_N] \\
&\quad = \lim_{N\to\infty} \left[\psi_0(\alpha+N+1) - \frac{\alpha+K_N d}{\alpha+N}\psi_0(1-d) - \right. \\
&\qquad\qquad \left. \frac{1}{\alpha+N}\left[\sum_{i=1}^{K_N}(n_i-d)\psi_0(n_i-d+1)\right]\right] \\
&\quad = \lim_{N\to\infty}\left[\psi_0(\alpha+N+1) - \sum_{i=1}^{K_N}\frac{n_i}{N}\psi_0(n_i-d+1)\right] \\
&\quad = -\lim_{N\to\infty}\sum_{i=1}^{K_N}\frac{n_i}{N}\left[\psi_0(n_i-d+1) - \psi_0(\alpha+N+1)\right]
\end{aligned}$$

although we have made no assumptions about the tail behavior of $\boldsymbol{\pi}$, so long as $\pi_k > 0$, $\mathbb{E}[n_k] = \mathbb{E}[\sum_{i=1}^{\infty}\mathbf{1}_{\{x_i=k\}}] = \sum_{i=1}^{\infty}P\{x_i = k\} = \lim_{N\to\infty} N\pi_k \to \infty$, and we may apply the asymptotic expansion $\psi(x) = \log(x) - \frac{1}{2x} - \frac{1}{12x^2} + O(\frac{1}{x^4})$ as $x \to \infty$ to find,

$$\lim_{N\to\infty}\mathbb{E}[H|\alpha,d,\mathbf{x}_N] = H_{\text{plugin}}$$

∎

We now turn to the proof of consistency for PYM. Although consistency is an intuitively plausible property for PYM, due to the form of the estimator our proof involves a rather detailed technical argument. Because of this, we break the proof of Theorem 4 into two parts. First, we prove a supporting Lemma.

Lemma 7 *If the data $\mathbf{x}_N$ have at least two coincidences, and are sampled from a distribution such that, for some constant $C > 0$, $K_N = o(N^{1-1/C})$ in probability, the following sequence of integrals converge.*

$$\int_0^{K_N+c}\int_0^1 \mathbb{E}[H|\alpha,d,\mathbf{x}_N]\frac{p(\mathbf{x}_N|\alpha,d)p(\alpha,d)}{p(\mathbf{x}_N)}\mathrm{d}\alpha\mathrm{d}d \xrightarrow{P} \mathbb{E}[\hat{H}_{plugin}|\mathbf{x}_N]$$

where $c > 0$ is an arbitrary constant.

Proof
Notice first that $E[H|\alpha, d, \mathbf{x}_N]$ is monotonically increasing in α, and so

$$\begin{aligned}\int_{\alpha=0}^{K_N+c}\int_{d=0}^{1} & \mathbb{E}[H|\alpha, d, \mathbf{x}_N]\frac{p(\mathbf{x}_N|\alpha,d)}{p(\mathbf{x}_N)}\,\mathrm{d}\alpha\,\mathrm{d}d \\ &\leq \int_{\alpha=0}^{K_N+c}\int_{d=0}^{1}\mathbb{E}[H|K_N+c, d, \mathbf{x}_N]\frac{p(\mathbf{x}_N|\alpha,d)}{p(\mathbf{x}_N)}\,\mathrm{d}\alpha\,\mathrm{d}d\end{aligned}$$

We have that,

$$\begin{aligned}\mathbb{E}[H|K_N+c, d, \mathbf{x}_N] = & \ \psi_0(K_N+c+N+1) \\ & - \frac{(1+d)K_N+c}{K_N+N+c}\psi_0(1-d) \\ & - \frac{1}{K_N+c+N}\left(\sum_{i=1}^{K_N}(n_i-d)\psi_0(n_i-d+1)\right)\end{aligned} \tag{47}$$

As a consequence of Proposition 2, $\int_{d=0}^{1}(1+d)\psi(1-d)\frac{p(\mathbf{x}|\alpha,d)}{p(\mathbf{x}_N)}\,\mathrm{d}d < \infty$, and so the second term is bounded and controlled by K_N/N. We let

$$A(d,N) := -\frac{(1+d)K_N+c}{K_N+N+c}\psi_0(1-d)$$

and, since $\lim_{N\to\infty}\int_{d=0}^{1}A(d,N)\frac{p(\mathbf{x}|\alpha,d)}{p(\mathbf{x}_N)}\,\mathrm{d}d = 0$, we focus on the remaining terms of (47). We also let $B\mathbf{n}) := \sum_{i=1}^{K_N}\left(\frac{n_i-1}{N}\log\left(\frac{n_i}{N}\right)\right)$, and note that $\lim_{N\to\infty}B = \hat{H}_{\text{plugin}}$. We find that,

$$\begin{aligned}&\mathbb{E}[H|K_N+c, d, \mathbf{x}_N] \\ &\quad \leq \log(N+K_N+c+1) + A(d,N) \\ &\quad -\sum_{i=1}^{K_N}\left(\frac{n_i-1}{K_N+N+c}\log(n_i)\right) \\ &\quad = \log(N+K_N+c+1) + A(d,N) - \\ &\quad \frac{N}{K_N+N+c}\left[\sum_{i=1}^{K_N}\left(\frac{n_i-1}{N}\log\left(\frac{n_i}{N}\right)\right) + \frac{N-K_N}{N}\log(N)\right] \\ &\quad = \log\left(1+\frac{K_N+c+1}{N}\right) + A(d,N) \\ &\qquad + \log(N)\left[\frac{2K_N+c}{N+K_N+c}\right] + \frac{N}{K_N+N+c}B \\ &\quad = \log\left(1+\frac{K_N+c+1}{N}\right) + A(d,N) \\ &\qquad + \frac{1}{1+(K_N+c)/N}\frac{2K_N+c}{N^{1-1/C}}\frac{\log(N)}{N^{1/C}} + \frac{N}{K_N+N+c}B \\ &\quad \to \hat{H}_{\text{plugin}} + o(1)\end{aligned}$$

As a result,

$$\int_{\alpha=0}^{K_N+c}\int_{d=0}^{1}\mathbb{E}[H|\alpha,d,\mathbf{x}_N]\frac{p(\mathbf{x}_N|\alpha,d)}{p(\mathbf{x}_N)}\,\mathrm{d}d\,\mathrm{d}\alpha$$
$$\leq\left[\hat{H}_{\text{plugin}}\int_{\alpha=0}^{K_N+c}\int_{d=0}^{1}\frac{p(\mathbf{x}_N|\alpha,d)}{p(\mathbf{x}_N)}\,\mathrm{d}d\,\mathrm{d}\alpha+o(1)\right]$$
$$\rightarrow\hat{H}_{\text{plugin}}$$

For the lower bound, we let $H_{(\alpha,d,N)} := \mathbb{E}[H|\alpha,d,\mathbf{x}_N]\mathbf{1}_{[0,K_N+c]}(\alpha)$. Notice that $\exp(-H_{(\alpha,d,N)})\leq 1$, so by dominated convergence $\lim_{N\to\infty}\mathbb{E}[\exp(-H_{(\alpha,d,N)})]=\exp(-\hat{H}_{\text{plugin}})$ by Proposition 2. And so by Jensen's inequality,

$$\exp(-\lim_{N\to\infty}\mathbb{E}[H_{(\alpha,d,N)}])\leq\lim_{N\to\infty}\mathbb{E}[\exp(-H_{(\alpha,d,N)})]=\exp(-\hat{H}_{\text{plugin}})$$
$$\implies\lim_{N\to\infty}\mathbb{E}[H_{(\alpha,d,N)}]\geq\hat{H}_{\text{plugin}},$$

and the lemma follows. ■

We now turn to the proof of our primary consistency result.

Proof [proof of Theorem 4]

$$\iint\mathbb{E}[H|\alpha,d,\mathbf{x}_N]\frac{p(\mathbf{x}_N|\alpha,d)p(\alpha,d)}{p(\mathbf{x}_N)}\mathrm{d}\alpha\mathrm{d}d$$
$$=\int_0^{\alpha_0}\int_0^1\mathbb{E}[H|\alpha,d,\mathbf{x}_N]\frac{p(\mathbf{x}_N|\alpha,d)p(\alpha,d)}{p(\mathbf{x}_N)}\mathrm{d}\alpha\mathrm{d}d$$
$$+\int_{\alpha_0}^{\infty}\int_0^1\mathbb{E}[H|\alpha,d,\mathbf{x}_N]\frac{p(\mathbf{x}_N|\alpha,d)p(\alpha,d)}{p(\mathbf{x}_N)}\mathrm{d}\alpha\mathrm{d}d$$

If we let $\alpha_0 = K_N+1$, by Lemma 7,

$$\int_0^{\alpha_0}\int_0^1\mathbb{E}[H|\alpha,d,\mathbf{x}_N]\frac{p(\mathbf{x}_N|\alpha,d)p(\alpha,d)}{p(\mathbf{x}_N)}\mathrm{d}\alpha\mathrm{d}d\rightarrow\mathbb{E}[H_{\text{plugin}}|\mathbf{x}_N].$$

Therefore, it remains to show that

$$\int_{\alpha_0}^{\infty}\int_0^1\mathbb{E}[H|\alpha,d,\mathbf{x}_N]\frac{p(\mathbf{x}_N|\alpha,d)p(\alpha,d)}{p(\mathbf{x}_N)}\mathrm{d}\alpha\mathrm{d}d\rightarrow 0$$

For finite support distributions where $K_N\to K<\infty$, this is trivial. Hence, we only consider infinite support distributions where $K_N\to\infty$. In this case, there exists N_0 such that for all $N\geq N_0$, $p([0,K_N+1],[0,1))\neq 0$. Since $p(\alpha,d)$ has a decaying tail as $\alpha\to\infty$, $\exists N_0\forall N\geq N_0$, $p(K_N+1,d)\leq 1$, thus, it is sufficient demonstrate convergence under an improper prior $p(\alpha,d)=1$. Using,

$$\mathbb{E}[H|\alpha,d,\mathbf{x}_N]\leq\psi_0(N+\alpha+1)\leq N+\alpha$$

we bound

$$\int_{\alpha_0}^{\infty}\int_0^1\mathbb{E}[H|\alpha,d,\mathbf{x}_N]\frac{p(\mathbf{x}_N|\alpha,d)}{p(\mathbf{x}_N)}\mathrm{d}\alpha\mathrm{d}d$$
$$\leq\frac{\int_{\alpha_0}^{\infty}\int_0^1(N+\alpha-1)p(\mathbf{x}_N|\alpha,d)\mathrm{d}\alpha\mathrm{d}d}{p(\mathbf{x}_N)}$$
$$+\frac{\int_{\alpha_0}^{\infty}\int_0^1 p(\mathbf{x}_N|\alpha,d)\mathrm{d}\alpha\mathrm{d}d}{p(\mathbf{x}_N)}$$

We focus upon the first term on the RHS since its boundedness implies that of the smaller second term. Recall, that $p(\mathbf{x}) = \int_{\alpha=0}^{\infty}\int_{d=0}^{1} p(\mathbf{x}|\alpha,d)\,\mathrm{d}d\,\mathrm{d}\alpha$. We seek an upper bound for the numerator and a lower bound for $p(\mathbf{x}_N)$. *Upper Bound:* First we integrate over d to find the upper bound of the numerator. (For the following display only we let $\gamma(d) = \left(\prod_{i=1}^{K_N}\Gamma(n_i - d)\right)$.

$$
\begin{aligned}
&\int_{\alpha_0}^{\infty}\int_0^1 (N+\alpha-1)p(\mathbf{x}_N|\alpha,d)\mathrm{d}\alpha\mathrm{d}d \\
&= \int_{\alpha_0}^{\infty}\int_{d=0}^{1} \frac{\left(\prod_{l=1}^{K_N-1}(\alpha+ld)\right)\gamma(d)\Gamma(1+\alpha)(N+\alpha-1)}{\Gamma(1-d)^{K_N}\Gamma(\alpha+N)}\,\mathrm{d}d\,\mathrm{d}\alpha \\
&\leq \int_{d=0}^{1}\frac{\gamma(d)}{\Gamma(1-d)^{K_N}}\,\mathrm{d}d\int_{\alpha_0}^{\infty}\frac{\Gamma(\alpha+K_N)(N+\alpha-1)}{\Gamma(\alpha+N)}\,\mathrm{d}\alpha
\end{aligned}
$$

Fortunately, the first integral on d will cancel with a term from the lower bound of $p(\mathbf{x}_N)$. Using[7], $\frac{(N+\alpha-1)\Gamma(\alpha+K_N)}{\Gamma(\alpha+N)} = \frac{\mathrm{Beta}(\alpha+K_N,N-K-1)}{\Gamma(N-K-1)}$,

$$
\begin{aligned}
&\int_{\alpha_0}^{\infty}\frac{(N+\alpha-1)\Gamma(\alpha+K)}{\Gamma(\alpha+N)}\,\mathrm{d}\alpha \\
&= \frac{1}{\Gamma(N-K-1)}\int_{\alpha_0}^{\infty}\mathrm{Beta}(\alpha+K,N-K-1)\,\mathrm{d}\alpha \\
&= \frac{1}{\Gamma(N-K-1)}\int_{\alpha_0}^{\infty}\int_0^1 t^{\alpha+K-1}(1-t)^{N-K-2}\,\mathrm{d}t\,\mathrm{d}\alpha \\
&= \frac{1}{\Gamma(N-K-1)}\int_{t=0}^{1}\frac{t^{\alpha_0+K-1}}{\log(\frac{1}{t})}(1-t)^{N-K-2}\,\mathrm{d}t \\
&\leq \frac{1}{\Gamma(N-K-1)}\int_{t=0}^{1}\frac{t^{\alpha_0+K-1}}{(1-t)}(1-t)^{N-K-2}\,\mathrm{d}t \\
&= \frac{1}{\Gamma(N-K-1)}\mathrm{Beta}(\alpha_0+K,N-K-2) \\
&= \frac{1}{\Gamma(N-K-1)}\frac{\Gamma(\alpha_0+K)\Gamma(N-K-2)}{\Gamma(N+\alpha_0-2)} \\
&= \frac{\Gamma(\alpha_0+K)}{\Gamma(N+\alpha_0-2)(N-K-2)}
\end{aligned}
$$

Lower Bound: Again, we first integrate d,

$$
\begin{aligned}
&\int_{\alpha=0}^{\infty}\int_{d=0}^{1} p(\mathbf{x}|\alpha,d)\,\mathrm{d}d\,\mathrm{d}\alpha \\
&= \int_{\alpha=0}^{\infty}\int_{d=0}^{1}\frac{\left(\prod_{l=1}^{K-1}(\alpha+ld)\right)\left(\prod_{i=1}^{K}\Gamma(n_i-d)\right)\Gamma(1+\alpha)}{\Gamma(1-d)^K\Gamma(\alpha+N)}\,\mathrm{d}d\,\mathrm{d}\alpha \\
&= \int_{d=0}^{1}\frac{\left(\prod_{i=1}^{K}\Gamma(n_i-d)\right)}{\Gamma(1-d)^K}\,\mathrm{d}d\int_{\alpha=0}^{\infty}\frac{\alpha^{K-1}\Gamma(1+\alpha)}{\Gamma(\alpha+N)}\,\mathrm{d}\alpha
\end{aligned}
$$

7. Note that in the argument for the inequalities we use K rather than K_N for concision.

So, since $\frac{\Gamma(1+\alpha)}{\Gamma(\alpha+N)} = \frac{\mathrm{Beta}(1+\alpha,N-1)}{\Gamma(N-1)}$, then

$$\begin{aligned}
\Gamma(N-1)\int_{\alpha=0}^{\infty} \frac{\alpha^{K-1}\Gamma(1+\alpha)}{\Gamma(\alpha+N)}\,\mathrm{d}\alpha \\
&\geq \int_{\alpha=0}^{\infty} \alpha^{K-1}\mathrm{Beta}(1+\alpha, N-1)\,\mathrm{d}\alpha \\
&= \int_{\alpha=0}^{\infty} \alpha^{K-1}\int_{t=0}^{1} t^{\alpha}(1-t)^{N-2}\,\mathrm{d}t\,\mathrm{d}\alpha \\
&= \int_{t=0}^{1}(1-t)^{N-2}\int_{\alpha=0}^{\infty} \alpha^{K-1}t^{\alpha}\,\mathrm{d}\alpha\,\mathrm{d}t \\
&= \Gamma(K)\int_{t=0}^{1}(1-t)^{N-2}\log\left(\frac{1}{t}\right)^{-K}\mathrm{d}t \\
&\geq \Gamma(K)\int_{t=0}^{1}(1-t)^{N-K-2}t^{K}\,\mathrm{d}t \\
&= \Gamma(K)\mathrm{Beta}(N-K-1, K+1)
\end{aligned}$$

where we've used the fact that $\log(\frac{1}{t})^{-1} \geq \frac{t}{1-t}$. Finally, we obtian the bound,

$$\int_{\alpha=0}^{\infty} \frac{\alpha^{K_N-1}\Gamma(1+\alpha)}{\Gamma(\alpha+N)}\,\mathrm{d}\alpha \geq \frac{\Gamma(K)\Gamma(N-K-1)\Gamma(K+1)}{\Gamma(N-1)\Gamma(N)}.$$

Now, we apply the upper and lower bounds to bound PYM. We have,

$$\begin{aligned}
&\frac{\int_{\alpha_0}^{\infty}\int_0^1 (N+\alpha-1)p(\mathbf{x}_N|\alpha,d)\mathrm{d}\alpha\mathrm{d}d}{p(\mathbf{x}_N)} \\
&\leq \frac{\Gamma(\alpha_0+K_N)}{(N-K_N-2)\Gamma(N+\alpha_0-2)}\frac{\Gamma(N-1)\Gamma(N)}{\Gamma(K_N)\Gamma(N-K_N-1)\Gamma(K_N+1)} \\
&= \frac{1}{(N-K_N-2)}\frac{\Gamma(\alpha_0+K_N)}{\Gamma(K_N)}\frac{\Gamma(N-1)}{\Gamma(N+\alpha_0-2)} \\
&\qquad\qquad \times \frac{\Gamma(N)}{\Gamma(N-K_N-1)\Gamma(K_N+1)} \\
&\to \frac{N}{(N-K_N-2)}\left(\frac{K_N}{N}\right)^{\alpha_0}\frac{N^{N-1/2}}{(N-K_N-1)^{N-K_N-3/2}(K_N+1)^{K_N+1/2}} \\
&= \frac{N^2}{(K_N+1)^{1/2}(N-K_N-2)}\left(\frac{K_N}{N}\right)^{\alpha_0}\left(\frac{N}{N-K_N-1}\right)^{N-3/2} \\
&\qquad\qquad \times \left(\frac{N-K_N-1}{K_N+1}\right)^{K_N} \\
&\to \frac{N}{(K_N+1)^{1/2}}\left(\frac{K_N}{N}\right)^{\alpha_0}\left(\frac{N}{K_N}\right)^{K_N}
\end{aligned}$$

Where we have applied the asymptotic expansion for the Beta function,

$$\mathrm{Beta}(x,y) \sim \sqrt{2\pi}\frac{x^{x-\frac{1}{2}}y^{y-\frac{1}{2}}}{(x+y)^{x+y-\frac{1}{2}}}$$

a consequence of Stirling's formula. Finally, we take $\alpha_0 := K_N + (C+1)/2$ so that the limit becomes,

$$\begin{aligned} &\to \frac{N}{K_N^{1/2}} \left(\frac{K_N}{N}\right)^{(C+1)/2} \\ &= \frac{K_N^{C/2}}{N^{C/2-1/2}} \end{aligned}$$

which tends to 0 with increasing N since, by assumption, $K_N = o(N^{1-1/C})$. ■

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). Parts of this manuscript were presented at the Advances in Neural Information Processing Systems (NIPS) 2012 conference.

References

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

E. Archer, I. M. Park, and J. Pillow. Bayesian estimation of discrete entropy with mixtures of stick-breaking priors. In P. Bartlett, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2024–2032. NIPS, 2012.

R. Barbieri, L. Frank, D. Nguyen, M. Quirk, V. Solo, M. Wilson, and E. Brown. Dynamic analyses of information encoding in neural ensembles. *Neural Computation*, 16:277–307, 2004.

J. Boyd. Exponentially convergent fourier-chebshev quadrature schemes on bounded and infinite intervals. *Journal of scientific computing*, 2(2):99–109, 1987.

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.

C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *Information Theory, IEEE Transactions on*, 14(3):462–467, 1968.

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.

M. Farach, M. Noordewier, S. Savari, L. Shepp, A. Wyner, and J. Ziv. On the entropy of DNA: Algorithms and measurements based on memory and rapid convergence. In *Proceedings of the sixth annual ACM-SIAM symposium on Discrete algorithms*, pages 48–57. Society for Industrial and Applied Mathematics, 1995.

A. Gnedin, B. Hansen, and J. Pitman. Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws. *Probability Surveys*, 4:146–171, 2007.

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.

P. Grassberger. Entropy estimates from insufficient samplings. *arXiv preprint*, January 2008.

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.

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

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

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

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.

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

C. Letellier. Estimating the shannon entropy: recurrence plots versus symbolic dynamics. *Physical review letters*, 96(25):254102, 2006.

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

T. Minka. Estimating a Dirichlet distribution. Technical report, MIT, 2003.

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

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

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.

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

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

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

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.

J. W. Pillow, L. Paninski, V. J. Uzzell, E. P. Simoncelli, and E. J. Chichilnisky. Prediction and decoding of retinal ganglion cell responses with a probabilistic spiking model. *The Journal of Neuroscience*, 25:11003–11013, 2005.

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

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.

E. T. Rolls, M. J. Tovée, and S. Panzeri. The neurophysiology of backward visual masking: Information analysis. *Journal of Cognitive Neuroscience*, 11(3):300–311, May 1999.

K. H. Schindler, M. Palus, M. Vejmelka, and J. Bhattacharya. Causality detection based on information-theoretic approaches in time series analysis. *Physics Reports*, 441:1–46, 2007.

J. Shlens, M. B. Kennel, H. D. I. Abarbanel, and E. J. Chichilnisky. Estimating information rates with confidence intervals in neural spike trains. *Neural Comput*, 19(7):1683–1719, Jul 2007.

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.

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.

A. Treves and S. Panzeri. The upward bias in measures of information derived from limited data samples. *Neural Computation*, 7:399–407, 1995.

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

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

Z. Zhang. Entropy estimation in Turing's perspective. *Neural Computation*, pages 1–22, 2012.

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