Journal of Machine Learning Research 15 (2014) 2833-2868 Submitted 4/13; Revised 4/14; Published 10/14

Bayesian Entropy Estimation for Countable Discrete Distributions

Evan Archer* EARCHER@UTEXAS.EDU
Center for Perceptual Systems
The University of Texas at Austin, Austin, TX 78712, USA

Max Planck Institute for Biological Cybernetics
Spemannstrasse 41
72076 Tübingen, Germany

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: Lawrence Carin

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 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 measures for mixing Pitman-Yor processes to produce 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. Finally, 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 an important quantity in many fields, from probability theory to engineering, ecology, and neuroscience. While entropy may be best known for its role in information theory, the practical problem of estimating entropy from samples arises in many applied settings. For example, entropy provides an important tool for quantifying the information carried by neural signals, and there is an extensive literature in neuroscience devoted to estimating the entropy of neural spike trains (Strong et al., 1998; Barbieri et al., 2004; Shlens et al., 2007; Rolls et al., 1999;

∗. EA and IP contributed equally.

Knudson and Pillow, 2013). Entropy is also used for estimating dependency structure and inferring causal relations in statistics and machine learning (Chow and Liu, 1968; Schindler et al., 2007), as well as in molecular biology (Hausser and Strimmer, 2009). Entropy also arises 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).

In these settings, researchers are confronted with data arising from an unknown discrete distribution, and seek to estimate its entropy. One reason for estimating the entropy, as opposed to estimating the full distribution, is that it may be infeasible to collect enough data to estimate the full distribution reliably. The problem is not just that we may not have enough data to estimate the probability of an event accurately. In the so-called "undersampled regime" we may not even observe all events that have non-zero probability. In general, estimating a distribution in this setting is a hopeless endeavor. Estimating the entropy, by contrast, is much easier. In fact, in many cases, entropy can be accurately estimated with fewer samples than the number of distinct

Nonetheless, entropy estimation remains a difficult problem. There is no unbiased estimator for entropy, and the maximum likelihood estimator is severely biased for small data sets (Paninski, 2003). Many previous studies have taken a frequentist approach and focused on methods for computing and reducing this bias (Miller, 1955; Panzeri and Treves, 1996; Strong et al., 1998; Paninski, 2003; Grassberger, 2008). Here, we instead take a Bayesian approach to entropy estimation, building upon an approach introduced by Nemenman and colleagues (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. Figure 1 shows a graphical model illustrating the dependencies between the basic quantities of interest.

When there are few samples relative to the total number of symbols, entropy estimation is especially difficult. We refer to this informally as the "under-sampled" regime. In this regime, it is common for many symbols with non-zero probability to remain unobserved, and often we can only bound or estimate the *support* of the distribution (i.e., the number of symbols with non-zero probability). Previous Bayesian approaches to entropy estimation (Nemenman et al., 2002) required *a priori* knowledge of the support. Here we overcome this limitation by formulating a prior over the space of countably-infinite discrete distributions. As we will show, the resulting estimator is consistent even when the support of the true distribution is finite.

Our approach relies on Pitman-Yor process (PYP), a two-parameter generalization of the Dirichlet process (DP) (Pitman and Yor, 1997; Ishwaran and James, 2003; Goldwater et al., 2006), which provides a prior distribution over the space of countably infinite discrete distributions. The PYP provides an attractive family of priors in this setting because: (1) the induced posterior distribution over entropy given data has analytically tractable moments; and (2) distributions sampled from a PYP can exhibit power-law tails, a feature commonly observed in data from social, biological and physical systems (Zipf, 1949; Dudok de Wit, 1999; Newman, 2005).

However, 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 data set. Our approach, inspired by Nemenman and colleagues (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 compare various estimators with applications to data.

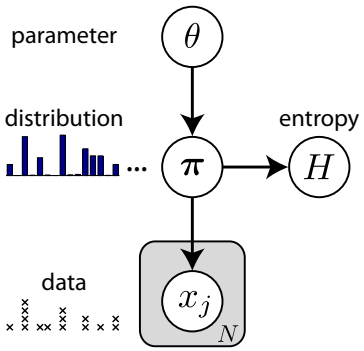

Figure 1: Graphical model illustrating the ingredients for Bayesian entropy estimation. Arrows indicate conditional dependencies between variables, and the gray "plate" denotes multiple copies of a random variable (with the number of copies N indicated at bottom). For entropy estimation, the joint probability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

2. Entropy Estimation

Consider samples $\mathbf{x} := \{x_j\}_{j=1}^N$ drawn *iid* from an unknown discrete distribution $\boldsymbol{\pi} := \{\pi_i\}_{i=1}^{\mathcal{A}}$, $p(x_j = i) = \pi_i$, on a finite or (countably) infinite alphabet $\mathcal{X}$ with cardinality $\mathcal{A}$. 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 so-called "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 severely biased estimators and briefly review approaches for correcting this bias. We then consider Bayesian techniques for entropy estimation in general before introducing the Nemenman–Shafee–Bialek (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 categorical (or multinomial) likelihood.

Despite its simplicity and desirable asymptotic properties, $\hat{H}_{\text{plugin}}$ exhibits substantial negative bias in the under-sampled 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 (Miller, 1955; Treves and Panzeri, 1995; Panzeri and Treves, 1996; Grassberger, 2008) 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). Recently, Wolpert and colleagues have considered entropy estimation in the case of unknown alphabet size (Wolpert and DeDeo, 2013). In that paper, the authors infer entropy under a (finite) Dirichlet prior, but treat the alphabet size itself as a random variable that can be either inferred from the data or integrated out.

Other recent work considers 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 proposed an estimator based on the Good-Turing estimate of population size (Zhang, 2012).

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

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

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

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

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

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

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

2.3 Nemenman-Shafee-Bialek (NSB) Estimator

In a seminal paper, Nemenman and colleagues showed that for finite distributions with known $\mathcal{A}$, Dirichlet priors with fixed a impose a narrow prior distribution over entropy (Nemenman et al., 2002). In the under-sampled regime, Bayesian estimates based on such highly informative priors are essentially determined by the value of a. 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 and colleagues suggested a mixture-of-Dirichlets prior

$$p(\boldsymbol{\pi}) = \int p_{\mathrm{Dir}}(\boldsymbol{\pi}|a) p(a) \, \mathrm{d}a, \tag{4}$$

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

$$p(a) \propto \frac{d}{da}\mathbb{E}[H|a] = \mathcal{A}\psi_1(\mathcal{A}a+1) - \psi_1(a+1), \tag{5}$$

where $\mathbb{E}[H|a]$ denotes the expected value of H under a $\mathrm{Dir}(a)$ prior, and $\psi_1(\cdot)$ denotes the tri-gamma function. To the extent that $p(H|a)$ resembles a delta function, (4) and (5) imply a uniform prior for H on $[0, \log \mathcal{A}]$. The BLS estimator under the NSB prior can be written

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

where $E[H|\mathbf{x},a]$ is the posterior mean under a $\mathrm{Dir}(a)$ prior, and $p(\mathbf{x}|a)$ denotes the evidence, which has a Pólya distribution (Minka, 2003)

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

The NSB estimate $\hat{H}_{nsb}$ and its posterior variance are easily computable via 1D numerical integration in a using closed-form expressions for the first two moments of the posterior distribution of H given a. The forms for these moments are discussed in previous work (Wolpert and Wolf, 1995; Nemenman et al., 2002), but the full formulae have to our knowledge never been explicitly shown. Here we state the results,

$$\mathbb{E}[H|\mathbf{x},a] = \psi_0(\tilde{N}+1) - \sum_i \frac{\tilde{n}_i}{\tilde{N}}\psi_0(\tilde{n}_i+1) \tag{6}$$

$$\mathbb{E}[H^2|\mathbf{x},a] = \sum_{i\neq k}\frac{\tilde{n}_i\tilde{n}_k}{(\tilde{N}+1)\tilde{N}}I_{i,k} + \sum_i \frac{(\tilde{n}_i+1)\tilde{n}_i}{(\tilde{N}+1)\tilde{N}}J_i \tag{7}$$

$$\begin{aligned}I_{i,k} &= \Big(\psi_0(\tilde{n}_k+1) - \psi_0(\tilde{N}+2)\Big)\Big(\psi_0(\tilde{n}_i+1) - \psi_0(\tilde{N}+2)\Big) - \psi_1(\tilde{N}+2) \\ J_i &= (\psi_0(\tilde{n}_i+2) - \psi_0(\tilde{N}+2))^2 + \psi_1(\tilde{n}_i+2) - \psi_1(\tilde{N}+2),\end{aligned}$$

where $\tilde{n}_i = n_i + a$ are counts plus prior "pseudocount" a, $\tilde{N} = \sum \tilde{n}_i$ 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, $\mathrm{var}[H|\mathbf{n},a] = \mathbb{E}[H^2|\mathbf{n},a] - \mathbb{E}[H|\mathbf{n},a]^2$. 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 and colleagues 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 asymptotic-NSB (ANSB) (Nemenman et al., 2004; Nemenman, 2011),

$$\hat{H}_{ansb} = 2\log(N) + \psi_0(N-K) - \psi_0(1) - \log(2), \tag{8}$$

where K is the number of distinct symbols in the sample. Note that the ANSB estimator is designed specifically for an extremely under-sampled regime ($K \sim N$), which we refer to as the "ANSB approximation regime". The fact that ANSB diverges with N in the well-sampled regime (Vu et al.,

2007) is therefore consistent with its design. In our experiments with ANSB in subsequent sections, we follow the work of Nemenman (2011) to define the ANSB approximation regime to be that region such that $E[K_N]/N > 0.9$, where K_N is the number of unique symbols appearing in a sample of size N.

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 (Ferguson, 1973; 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 now $\boldsymbol{\pi} = \{\pi_i\}$ denotes a countably infinite set of 'weights' on a set of atoms $\{\phi_i\}$ drawn from some base probability measure, where δ_{ϕ_i} is 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[2] 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 $a \to 0$ s.t. $a\mathcal{A} \to \alpha$. 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 $\mathrm{PY}(d, \alpha)$ with $d \neq 0$, the tails approximately follow a power-law: $\pi_i \propto (i)^{-\frac{1}{d}}$ (pp. 867, Pitman and Yor (1997)).[3] Many natural phenomena such as city size, language, spike responses, etc., also exhibit power-law tails (Zipf, 1949; Newman, 2005). Figure 2 shows two such examples, along with a sample drawn from the best-fitting DP and PYP distributions.

Let $\mathrm{PY}(d, \alpha)$ 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, $\mathrm{DP}(\alpha)$. To gain intuition for the DP and PYP, it is useful to consider typical samples $\boldsymbol{\pi}$ with weights $\{\pi_i\}$ sorted in decreasing order of probability, so that $\pi_{(1)} > \pi_{(2)} > \cdots$. The concentration parameter α controls how much of the probability mass is concentrated in the first few samples, that is, in the head instead of the tail of the sorted distribution. For small α the first few weights carry most of the probability mass whereas, for large α, the probability mass is more spread out so that $\boldsymbol{\pi}$ is more uniform. As noted above the discount parameter d controls the shape of the tail. Larger d gives heavier power-law tails, while $d = 0$ yields exponential tails.

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

$$\beta_i \sim \mathrm{Beta}(1 - d, \alpha + id), \qquad \tilde{\pi}_i = \prod_{j=1}^{i-1} (1 - \beta_j)\beta_i, \tag{9}$$

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).

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. GEM stands for "Griffiths, Engen and McCloskey", after three researchers who considered these ideas early on (Ewens, 1990).
3. The power-law exponent has been given incorrectly in previous work (Goldwater et al., 2006; Teh, 2006).

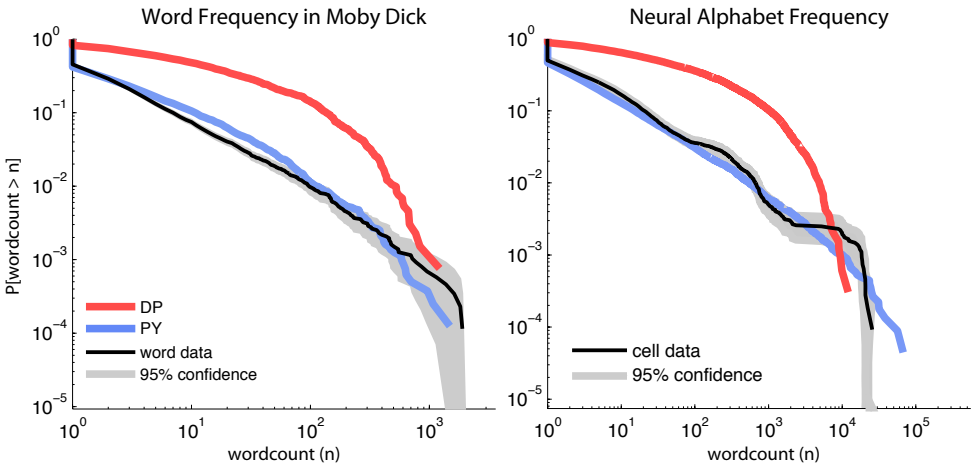

Figure 2: Empirical cumulative distribution functions of words in natural language (**left**) and neural spike patterns (**right**). We compare samples from the DP (red) and PYP (blue) priors for two data sets with heavy tails (black). In both cases, we compare the empirical CDF estimated from data to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

3.1 Expectations over DP and PYP Priors

For our purposes, a key virtue of PYP priors is a mathematical property called *invariance under size-biased sampling*. This property 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 $\boldsymbol{\pi} \sim PY(d, \alpha)$,*

$$\mathbb{E}_{(\boldsymbol{\pi}|d,\alpha)}\left[\sum_{i=1}^{\infty} f(\pi_i)\right] = \mathbb{E}_{(\tilde{\pi}_1|d,\alpha)}\left[\frac{f(\tilde{\pi}_1)}{\tilde{\pi}_1}\right], \tag{10}$$

$$\mathbb{E}_{(\boldsymbol{\pi}|d,\alpha)}\left[\sum_{i,j\neq i} g(\pi_i, \pi_j)\right] = \mathbb{E}_{(\tilde{\pi}_1,\tilde{\pi}_2|d,\alpha)}\left[\frac{g(\tilde{\pi}_1, \tilde{\pi}_2)}{\tilde{\pi}_1\tilde{\pi}_2}(1 - \tilde{\pi}_1)\right], \tag{11}$$

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

The first result (10) appears in (Pitman and Yor, 1997), and an analogous proof can be constructed for (11) (see Appendix).

The direct consequence of this proposition is that the first two moments of $H(\boldsymbol{\pi})$ under the PYP and DP priors have closed forms[4]

$$\mathbb{E}[H|d, \alpha] = \psi_0(\alpha + 1) - \psi_0(1 - d), \tag{12}$$

$$\mathrm{var}[H|d, \alpha] = \frac{\alpha + d}{(\alpha + 1)^2(1 - d)} + \frac{1 - d}{\alpha + 1}\psi_1(2 - d) - \psi_1(2 + \alpha). \tag{13}$$

4. Note that (12) and (13) follow from (6) and (7), respectively, under the PY limit.

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., $K = \sum_{i=1}^{\mathcal{A}} \mathbf{1}_{\{n_i>0\}}$.[5] 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 colleagues (Ishwaran and Zarepour, 2002), we write the posterior as an infinite random vector $\boldsymbol{\pi}|\mathbf{x}, d, \alpha = (p_1, p_2, p_3, \ldots, p_K, p_*\boldsymbol{\pi}')$, where

$$\begin{aligned}(p_1, p_2, \ldots, p_K, p_*) &\sim \mathrm{Dir}(n_1 - d, \ldots, n_K - d, \alpha + Kd) \\ \boldsymbol{\pi}' := (\pi_1, \pi_2, \pi_3, \ldots) &\sim \mathrm{PY}(d, \alpha + Kd).\end{aligned} \tag{14}$$

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

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

The variance, $\mathrm{var}[H|\mathbf{x}, d, \alpha]$, may also be expressed in an easily-computable closed-form. 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 (14). Applying the law of total variance and using the independence properties of the posterior, we find

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

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

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

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, analogous to the fixed-α Dirichlet priors discussed in Section 2.2. Unfortunately, fixed $PY(d, \alpha)$ priors 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 (Figure 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.[6] We then define the BLS entropy estimator under this mixture prior,

5. We note that the quantity K has been studied in Bayesian nonparametrics in its own right, for instance to study species diversity in ecological applications (Favaro et al., 2009).

6. 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, Jeffreys' prior for multinomial observations is exactly a Dirichlet distribution with a fixed $\alpha = 1/2$. As mentioned in the text, such Bayesian priors are highly informative about the entropy.

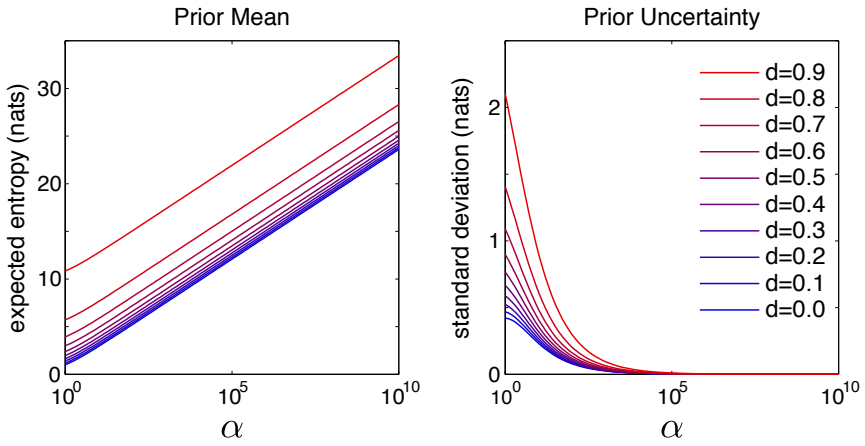

Figure 3: Prior entropy mean and variance (12) and 13 as a function of α and d. Note that entropy is approximately linear in $\log \alpha$. For large values of α, $p(H(\boldsymbol{\pi})|d, \alpha)$ is highly concentrated around the mean.

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 and colleagues (Nemenman et al., 2002), setting $p(d, \alpha)$ proportional to the derivative of the expected entropy (12). 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 Figure 3A). We can explicitly control this trade-off by reparameterizing PYP as follows

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

where $h > 0$ is equal to the expected prior entropy (12), and $\gamma \in [0, \infty)$ captures prior beliefs about tail behavior (Figure 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 PY$(d, 0)$ process (i.e., $\alpha = 0$, yielding $\boldsymbol{\pi}$ with power-law tails). In the limit where $\alpha \to -1$ and $d \to 1$, $\gamma \to \infty$. 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

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

where $p(\boldsymbol{\pi}|\gamma, h)$ denotes a PYP on $\boldsymbol{\pi}$ with parameters γ, h. We compare only three choices of $q(\gamma)$ here. However, the prior $q(\gamma)$ is not fixed but may be adapted to reflect prior beliefs about the data set at hand. A $q(\gamma)$ that places probability mass on larger γ (near 1) results in a prior that prefers heavy-tailed behavior and high entropy, whereas weight on small γ prefers exponential-tailed distributions. As a result, priors with more mass on large γ will also tend to yield wider credible

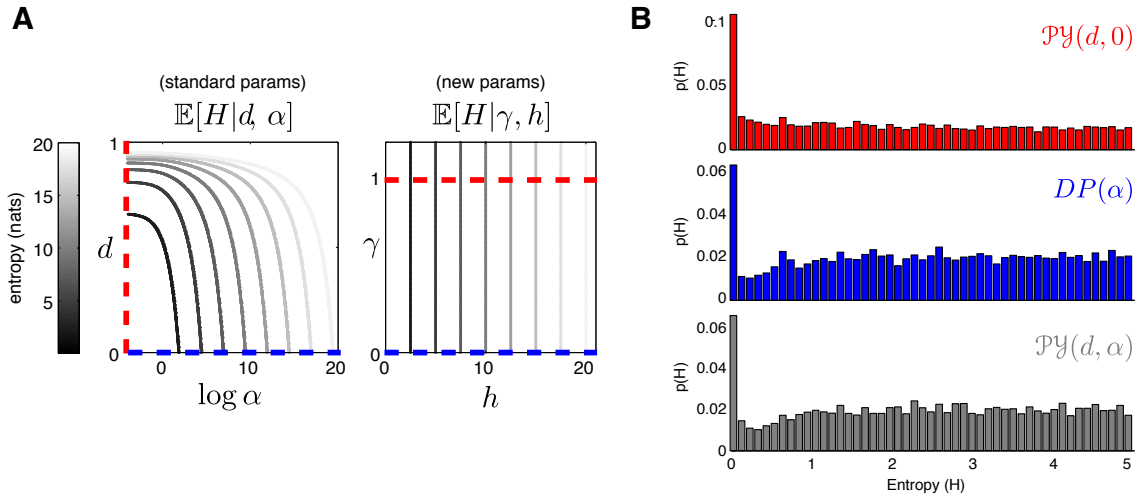

Figure 4: Prior over expected entropy under Pitman-Yor process prior. **(A)** Left: expected entropy as a function of the natural parameters (d, α). Right: expected entropy as a function of transformed parameters (h, γ). The dotted red and blue lines indicate the contours on which the PY$(d, 0)$ and DP(α) priors are defined, respectively. **(B)** Sampled prior distributions $(N = 5e3)$ over entropy implied by three different PYM priors, each with a different mixing density over α and d. We formulate each prior in the transformed parameters γ and h. We place a uniform prior on h and show three different choices of prior $q(\gamma)$. Each resulting PYM prior is a mixture of Pitman-Yor processes: PY$(d, 0)$ (red) uses a mixing density over d: $q(\gamma) = \delta_{\gamma-1}$; PY$(0, \alpha) =$ DP(α) (blue) uses a mixing density over α: $q(\gamma) = \delta_{\gamma}$; and PY$(d, \alpha)$ (grey) uses a mixture over both hyperparameters: $q(\gamma) = \exp(-\frac{10}{1-\gamma})\mathbf{1}_{\{\gamma<1\}}$. Note that for all of these examples, the "true" $p(H)$ is an improper prior supported on $[0, \infty)$. We visualize the sampled distributions only on the range from 0 to 5 nats, since sampling from PY becomes prohibitively expensive with increasing expected entropy (especially as $d \to 1$).

intervals and higher estimates of entropy. PYM mixture priors resulting from different choices of $q(\gamma)$ are all approximately flat on H, but each favors distributions with different tail behavior; the ability to select $q(\gamma)$ greatly enhances the flexibility of PYM, allowing the practitioner to adapt it to her own data.

Figure 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.[7] 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 (see Figure 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

7. In particular, the restriction $\gamma < 1$ omits the corner $d \to 1$ and $\alpha \to -d$. In this region, one can obtain arbitrarily large prior variance over H for a given mean. However, such priors have very heavy tails and seem poorly-suited to data with finite or exponential tails, and we therefore do not explore them further here.

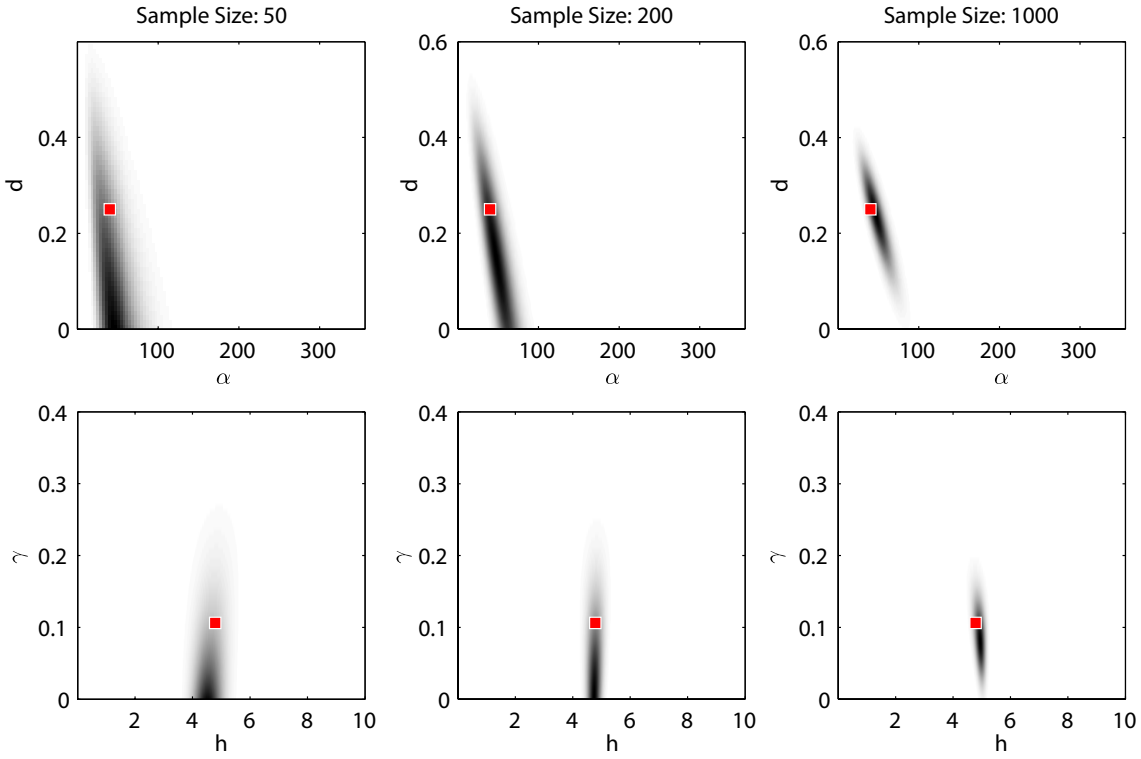

Figure 5: Convergence of $p(d, \alpha|\mathbf{x})$ for increasing sample size. Both parameterization (d, α) and (γ, h) are shown. Data are simulated from a PY$(0.25, 40)$ whose parameters are indicated by the red dot.

Pitman-Yor Mixture (PYM) estimator, takes the form

$$\hat{H}_{\mathrm{PYM}} = \mathbb{E}[H|\mathbf{x}] = \int \mathbb{E}[H|\mathbf{x}, d, \alpha] \frac{p(\mathbf{x}|d, \alpha)p(d, \alpha)}{p(\mathbf{x})} \, \mathrm{d}(d, \alpha), \tag{16}$$

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

$$p(\mathbf{x}|d, \alpha) = \frac{\left(\prod_{l=1}^{K-1}(\alpha + ld)\right)\left(\prod_{i=1}^{K}\Gamma(n_i - d)\right)\Gamma(1 + \alpha)}{\Gamma(1 - d)^K \Gamma(\alpha + N)}. \tag{17}$$

We can obtain posterior credible intervals 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 (16), except that we replace $\mathbb{E}[H|\mathbf{x}, d, \alpha]$ with $\mathrm{var}[H|\mathbf{x}, d, \alpha]$ in the integrand.

4.3 Computation

Because of the improper prior $p(d, \alpha)$, and because by (16) it must be integrated over all $\alpha > 0$, it is not obvious that the PYM estimate $\hat{H}_{\mathrm{PYM}}$ is computationally tractable. In this 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

We can compute the expected entropy $\mathbb{E}[H|\mathbf{x}, d, \alpha]$ more efficiently by using a representation in terms of *multiplicities*, a compressed statistic often used under other names (e.g., the *empirical histogram distribution function* as discussed by Paninski 2003). Multiplicities are 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 dot 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 data sets as we need only compute sums and products involving the number of 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 multiplicities representation. For instance, when expressed 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 concentrates on a relatively small region of parameter space. It is generally unnecessary 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. The posterior is unimodal in each variable α and d separately (see Appendix D); however, we do not have a proof for the unimodality of the evidence. Nevertheless, if there are multiple modes, they must lie on a strictly decreasing line of d as a function of α and, in practice, we find the posterior to be unimodal. We illustrate the concentration of the evidence visually in Figure 5.

We compute the hessian at the MAP parameter value, $(d_{\text{MAP}}, \alpha_{\text{MAP}})$. Using the inverse hessian as the covariance of a Gaussian approximation to the posterior, we select a grid spanning ± 6 std. We use numerical integration (Gauss-Legendre quadrature) on this region to compute the integral. When the hessian is rank-deficient (which may occur, for instance, when the $\alpha_{\text{MAP}} = 0$ or $d_{\text{MAP}} = 0$), we use Gauss-Legendre quadrature to 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 integration. PYM's posterior mean and variance provide, essentially, a Gaussian approximation to the posterior, and corresponding credible regions. However, in some situations (see Figure 6), variance-based credible intervals are a poor approximation to the true posterior credible intervals. In these cases we may wish to examine the full posterior distribution over H.

Stick-breaking, as described by (9), provides a straightforward algorithm for sampling distributions $\boldsymbol{\pi} \sim \text{PY}(d, \alpha)$. With enough stick-breaking samples, it is always possible to approximate π to arbitrary

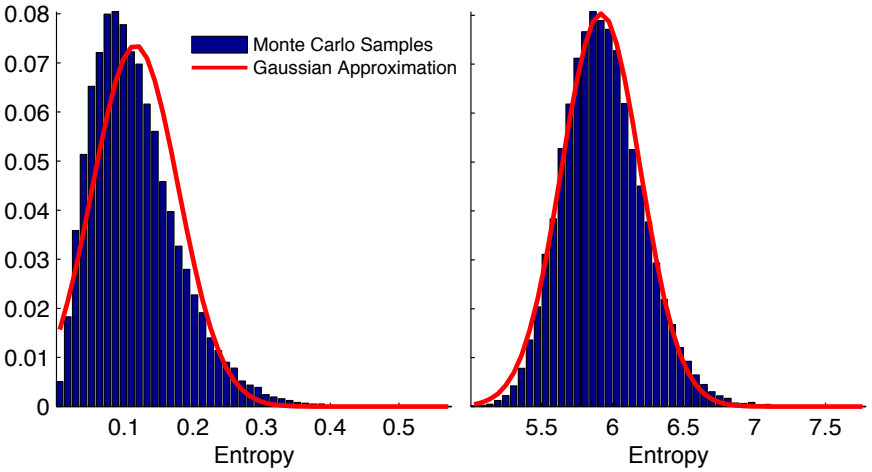

Figure 6: The posterior distributions of entropy for two data sets of 100 samples drawn from distinct distributions and the Gaussian approximation to each distribution based upon the posterior mean and variance. **(left)** Simulated example with low entropy. Notice that the true posterior is highly asymmetric, 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.

accuracy.[8] Even so, sampling $\pi \sim \mathrm{PY}(d, \alpha)$ for d near 1, where $\boldsymbol{\pi}$ is likely to be heavy-tailed, may require intractably large number of samples to obtain a good approximation.

We address this problem by directly estimating the entropy of the tail, $\mathrm{PY}(d, \alpha + N_s d)$, using (12). As shown in Figure 3, the prior variance of PY becomes arbitrarily small as for large α. We need only enough samples to assure that the variance of the tail entropy is small. The resulting final sample is the entropy of the (finite) samples plus the expected entropy of the tail, $H(\boldsymbol{\pi}^*) + \mathbb{E}[H|d, \alpha + Kd]$.[9]

Sampling entropy is most useful for very small amounts of data drawn from distributions with low expected entropy. In Figure 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 establish conditions under which (16) is defined (i.e., the right–hand of the equation 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 connections 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),

8. Bounds on the number of samples necessary to reach ϵ on average have been considered by Ishwaran and James (2001).
9. Due to the generality of the expectation formula (10), this method may be applied to sample the distributions of other additive functionals of PY.

Theorem 2 *Given a fixed data set* $\mathbf{x}$ *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 under-sampled regime; as $N \to \infty$, 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 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 infinite 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). Hence, $K_N/N \to 0$ in probability for an arbitrary distribution. As a result, the right–hand–side of (15) shares its asymptotic behavior, in particular consistency, with $\hat{H}_{\mathrm{plugin}}$. As (15) is consistent for each value of α and d, it is intuitively plausible that $\hat{H}_{\mathrm{PYM}}$, as a mixture of such values, should be consistent as well. However, while (15) alone is well-behaved, it is not clear that $\hat{H}_{\mathrm{PYM}}$ 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*

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

Intuitively, the asymptotic behavior of K_N/N is tightly related to the tail behavior of the distribution (Gnedin et al., 2007). In particular, $K_N \sim cN^b$ with $0 < b < 1$ if and only if $\pi_i \sim c' i^{\frac{1}{b}}$ where c and c' are constants, and we assume π_i is non-increasing (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., $\pi_i = O(i^b)$ for some $b > 1$ (again π_i is assumed non-increasing).

While consistency is an important property for any estimator, we emphasize that PYM is designed to address the under-sampled regime. Indeed, since $\hat{H}_{\mathrm{plugin}}$ is consistent and has an optimal rate of convergence for a large class of distributions (Vu et al., 2007; Antos and Kontoyiannis, 2001; Zhang, 2012), asymptotic properties provide little reason to use $\hat{H}_{\mathrm{PYM}}$. Nevertheless, notice that Theorem 4 makes very weak assumptions about $p(d, \alpha)$. In particular, the result is not dependent 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. Simulation Results

We compare $\hat{H}_{\mathrm{PYM}}$ to other proposed entropy estimators using several example data sets. Each plot in Figures 7, 8, 9, and 10 shows convergence as well as small sample performance. We compare

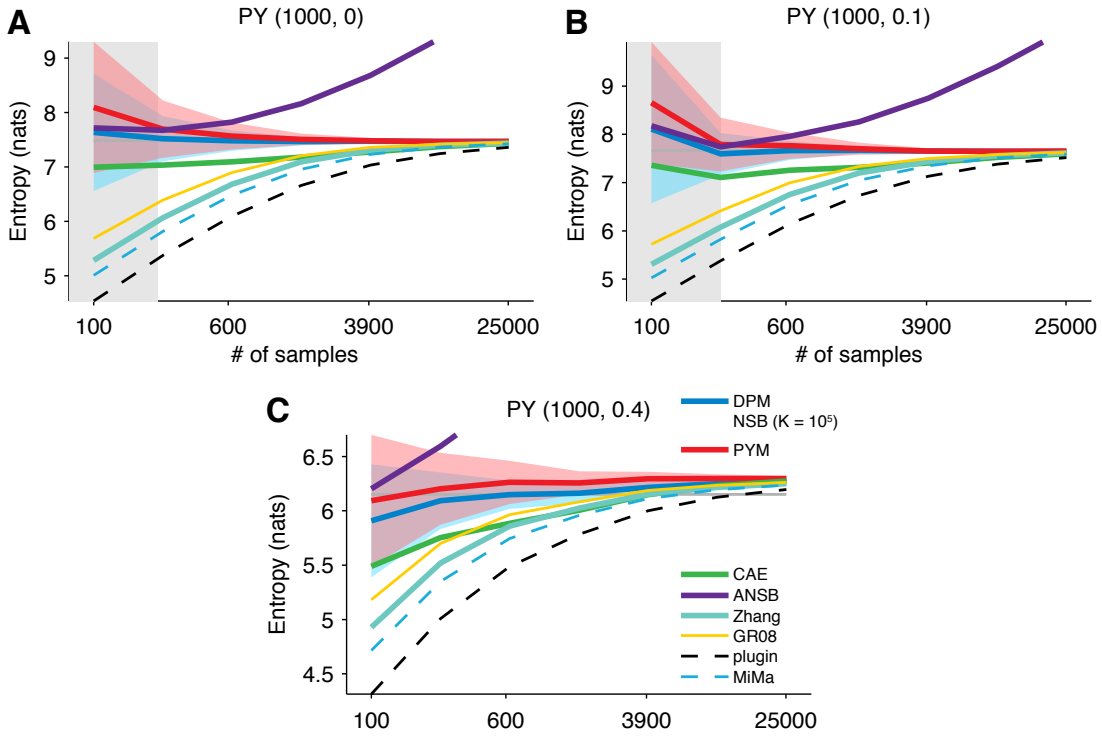

Figure 7: Comparison of estimators on stick-breaking distributions. Poisson-Dirichlet distribution with $(d = 0, \alpha = 1000)$ (**A**), $(d = 0.1, \alpha = 1000)$ (**B**), $(d = 0.4, \alpha = 100)$ (**C**). Recall that the Dirichlet Process is the Pitman-Yor Process with $d = 0$. We compare our estimators (DPM, PYM) with other enumerable support estimators (CAE, ANSB, Zhang, GR08), and finite support estimators (plugin, MiMa). Note that in these examples, the DPM estimator performs indistinguishably from NSB with alphabet size $\mathcal{A}$ fixed to a large value ($\mathcal{A} = 10^5$). For the experiments, we first sample a single $\boldsymbol{\pi} \sim \mathrm{PY}(d, \alpha)$ using the stick-breaking procedure of (9). For each N (x-axis), we apply all estimators to each of 10 sample data sets drawn randomly from $\boldsymbol{\pi}$. Solid lines are averages over all 10 realizations. Colored shaded area represent 95% credible intervals averaged over all 10 realizations. Gray shaded area represents the ANSB approximation regime defined as expected number of unique symbols being more than 90% the total number of samples.

our estimators, DPM ($d = 0$ only) and PYM ($\hat{H}_{\mathrm{PYM}}$), 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), Grassberger's asymptotic bias correction (GR08) (Grassberger, 2008), and Good-Turing estimator (Zhang, 2012). Note that similar to ANSB, DPM is an asymptotic (Poisson-Dirichlet) limit of NSB, and hence in practice behaves identically to NSB with large but finite K. We also compare with plugin (3) and a standard Miller-Maddow (MiMa) bias correction method with a conservative assumption that the number of uniquely observed symbols is K (Miller, 1955). To make comparisons more straightforward, we do not apply additional bias correction methods (e.g., jackknife) to any of the estimators.

In each simulation, we draw 10 sample distributions $\boldsymbol{\pi}$. From each $\boldsymbol{\pi}$ we draw a data set of N iid samples. In each figure we show the performance of all estimators averaged across the 10 sampled data sets.

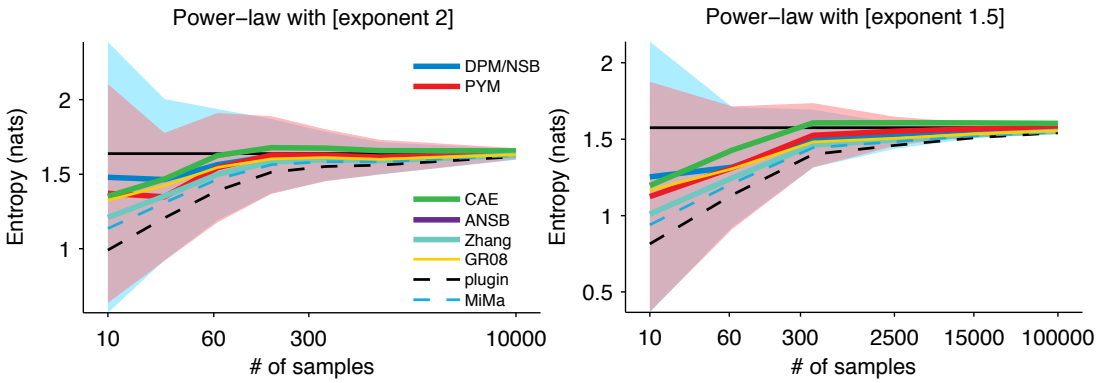

Figure 8: Comparison of estimators on power-law distributions. The highest probabilities in these power-law distributions were large enough that they were never effectively under-sampled.

The experiments of Figure 7 show performance on a single $\boldsymbol{\pi} \sim \mathrm{PY}(d, \alpha)$ drawn using the stick-breaking procedure of (9). We draw π_i according to (9) in blocks of size 10^3 until $1 - \sum_{N_s} \pi_i < 10^{-3}$, where N_s is the number of sticks. Unsurprisingly, PYM performs well when the data are truly generated by a Pitman-Yor process (Figure 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; Figure 7 top), DPM shows slightly improved performance. When the tail has a strong power-law decay, (Figure 7 bottom), PYM performs better than DPM. Most of the other estimators are consistently biased down, with the exception of ANSB.

The shaded gray area indicates the ANSB approximation regime, where the approximation used to define the ANSB estimator is approximately valid. Although this region has no definitive boundary, it corresponds to a regime where where the average number of coincidences is small relative to the number of samples. Following Nemenman (2011), we define the under-sampled regime to be the region where $E[K_N]/N > 0.9$, where K_N is the number of unique symbols appearing in a sample of size N. Note that only 3 out of 10 results in Figures. 7, 8, 9, 10 have shaded area; the ANSB approximation regime is not large enough to appear in the plots. This regime appears to be designed for a relatively broad distribution (close to uniform distribution). In cases where a single symbol has high probability, the ANSB approximation regime is essentially never valid. In our example distributions, this is the case with for power-law distributions and $\mathcal{PY}$ distributions with large d. For example, Figure 8 is already outside of the ANSB approximation regime with only 4 samples.

Although Pitman-Yor process $\mathrm{PY}(d, \alpha)$ has a power-law tail controlled by d, the high probability portion is modulated by α and does not strictly follow a power-law distribution as a whole. In Figure 8, we evaluate the performance for $p_i \propto i^{-2}$ and $p_i \propto i^{-1.5}$. PYM and DPM have slight negative biases, but the credible interval covers the true entropy for all sample sizes. For small sample sizes, most estimators are negatively biased, again except for 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 Figure 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). We tokenized the novel into individual words using the Python library NLTK.[10] Punctuation is disregarded. These real-world data sets have heavy, approximately power-law tails[11] as pointed out

10. Further information about the Natural Language Toolkit (NLTK) may be obtained at the project's website, `http://www.nltk.org/`.
11. 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 data sets in questions, and neither do we rely upon the properties of power-law distributions in our analyses.

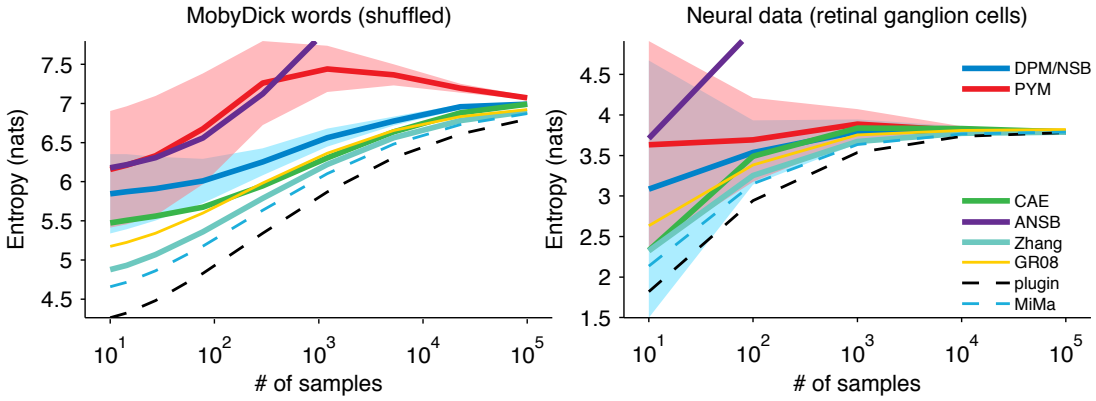

Figure 9: Comparison of estimators on real data sets.

earlier in Figure 2. For Moby Dick, PYM slightly overestimates, while DPM slightly underestimates, yet both methods are 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 data set with both PYM and DPM bracketing the asymptotic value with their credible intervals.

Finally, we applied the denumerable-support estimators to finite-support distributions (Figure 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 (Figure 10 top). Initially PYM shows the least bias, but DPM provides a better estimate for increasing sample size. However, notice 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 (Figure 10 middle). For Poisson distribution, a theoretically enumerable-support distribution on the natural numbers, 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 ANSB, work quite well.

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.

The computation time for our estimators is $O(LG)$, where L number symbols with distinct frequencies (bounded above by the quantity M defined in Section 4.3.1) and G is the number of grid points used to compute the numerical integral. Hence, computation time as a function of sample size depends upon how L scales with samples size N, always sublinearly, and $O(N^{1/2})$ in the worse case. In our examples, computation times for 10^5 samples are in the order of 0.1 seconds (Figure 11). Hence in practice, for the examples we have shown, more time is spent building the histogram from the data than computing the entropy estimate.

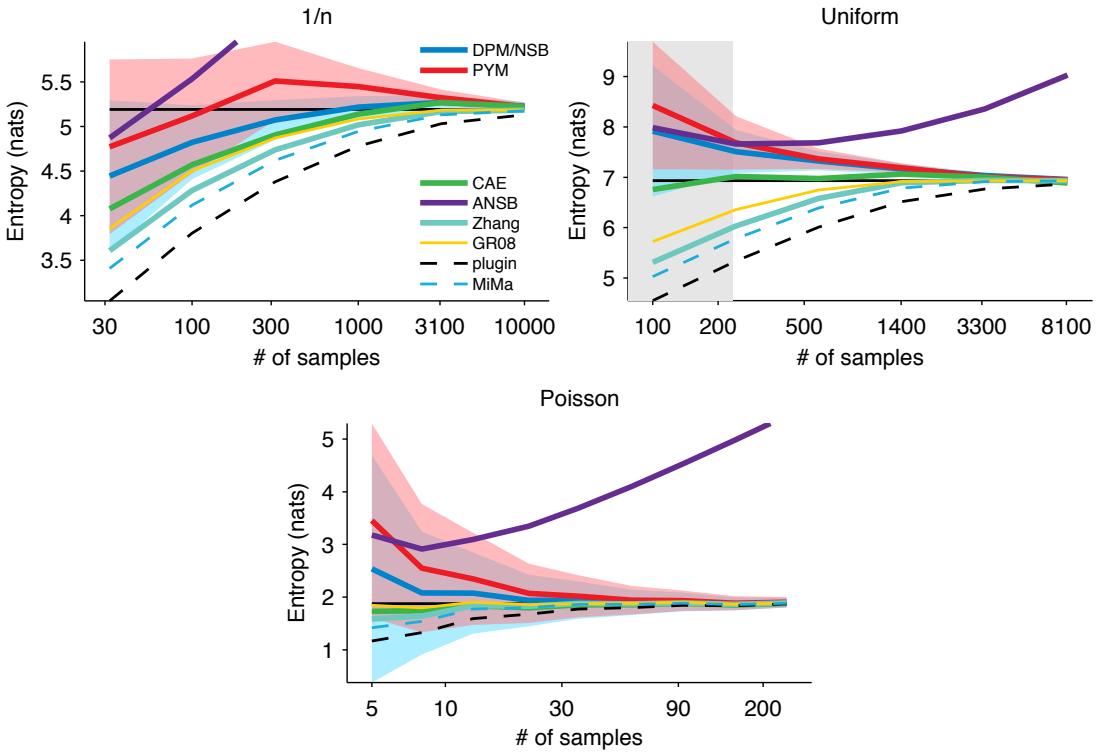

Figure 10: Comparison of estimators on finite support distributions. Black solid line indicates the true entropy. Poisson distribution ($\lambda = e$) has a countably infinite (but very thin) tail. All probability mass was concentrated on 26 symbols, within machine precision.

7. Conclusion

In this paper we introduced PYM, a new 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.

Further experiments on diverse data sets might reveal ways to improve PYM, such as new tactics or theory for selecting the prior on tail behavior, $q(\gamma)$. It may also prove fruitful to investigate ways to tailor PYM to a specific application, for instance by combining it with with more structured priors such as those employed by Archer et al. (2013). Further, while we have shown that PYM is consistent for any prior, an expanded theory might investigate the convergence rate, perhaps in relation to the choice of prior.

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`.

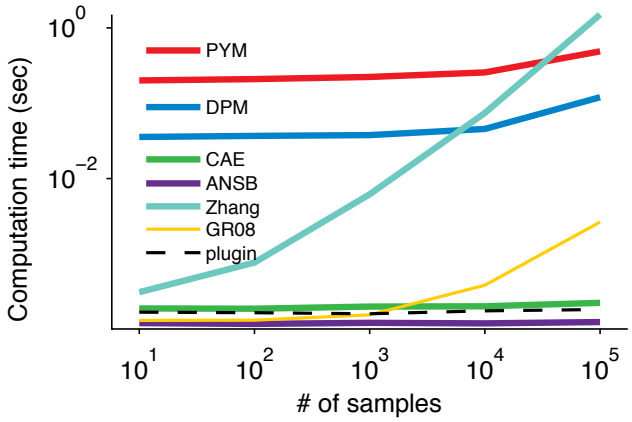

Figure 11: Median computation time to estimate entropy for the neural data. The computation time excludes the preprocessing required to build the histogram and convert to multiplicity representation. Note that for DPM and PYM this time also includes estimating the posterior variance.

Acknowledgments

We thank E. J. Chichilnisky, A. M. Litke, A. Sher and J. Shlens for retinal data, and Y. W. Teh and A. Cerquetti 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.

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.

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*

$$\mathbb{E}[H(\boldsymbol{\pi})|\vec{\alpha}] \quad = \quad \psi_0(A+1) - \sum_{i=1}^{\mathcal{A}} \frac{\alpha_i}{A} \psi_0(\alpha_i + 1) \tag{18}$$

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 we have that

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

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

■

A.2 Variance of Entropy for Finite Dirichlet

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

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

$$
\begin{aligned}
\mathbb{E}[H^2(\boldsymbol{\pi})|\vec{\alpha}] &= \sum_{i\neq k}\frac{\alpha_i\alpha_k}{(A+1)(A)}I_{ik} + \sum_i \frac{\alpha_i(\alpha_i+1)}{(A+1)(A)}J_i \qquad (19)\\
I_{ik} &= (\psi_0(\alpha_k+1) - \psi_0(A+2))\,(\psi_0(\alpha_i+1) \\
&\quad -\psi_0(A+2)) - \psi_1(A+2) \\
J_i &= (\psi_0(\alpha_i+2) - \psi_0(A+2))^2 + \psi_1(\alpha_i+2) \\
&\quad - \psi_1(A+2)
\end{aligned}
$$

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

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

$$
\begin{aligned}
&= \sum_i \int \pi_i^{\alpha_i+1} {\log_2}^2 \pi_i \delta(\textstyle\sum_i \pi_i - 1) \prod_{j\neq i} \pi_j^{\alpha_j-1} d\boldsymbol{\pi} \\
&+ \sum_{i\neq k} \int (\pi_i^{\alpha_i} \log_2 \pi_i)\,(\pi_k^{\alpha_k} \log_2 \pi_k)\, \delta(\textstyle\sum_i \pi_i - 1) \prod_{j\notin\{i,k\}} \pi_j^{\alpha_j-1} d\boldsymbol{\pi} \\
&= \sum_i \frac{d^2}{d(\alpha_i+1)^2} \int \pi_i^{\alpha_i+1} \delta(\textstyle\sum_i \pi_i - 1) \prod_{j\neq i} \pi_j^{\alpha_j-1} d\boldsymbol{\pi} \\
&+ \sum_{i\neq k} \frac{d}{d\alpha_i}\frac{d}{d\alpha_k} \int (\pi_i^{\alpha_i})\,(\pi_k^{\alpha_k})\, \delta(\textstyle\sum_i \pi_i - 1) \prod_{j\notin\{i,k\}} \pi_j^{\alpha_j-1} d\boldsymbol{\pi}
\end{aligned}
$$

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

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

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

$$
\begin{aligned}
&= \frac{\Gamma(1+z)}{\Gamma(c+z)}\left[(\psi_0(1+z)-\right. \\
&\qquad \left.\psi_0(c+z))^2+\psi_1(1+z)-\psi_1(c+z)\right]\prod_{j\neq i}\Gamma(\alpha_j) \\
&= \frac{z(z-1)}{(c+z-1)(c+z-2)}\left[(\psi_0(1+z)-\psi_0(c+z))^2\right. \\
&\qquad \left.+\psi_1(1+z)-\psi_1(c+z)\right]\frac{\prod_j\Gamma(\alpha_j)}{\Gamma(A)} \\
&= \frac{(\alpha_i+1)(\alpha_i)}{(A+1)(A)}\left[(\psi_0(\alpha_i+2)-\psi_0(A+2))^2+\psi_1(\alpha_i+2)\right. \\
&\qquad \left.-\psi_1(A+2)\right]\frac{\prod_j\Gamma(\alpha_j)}{\Gamma(A)}
\end{aligned}
$$

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 $\operatorname{var}_{\boldsymbol{\pi}}[H(\boldsymbol{\pi})|d,\alpha]$. 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 (10). 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

$$
\begin{aligned}
\mathbb{E}_{\tilde{\pi}_1}\left[\frac{f(\tilde{\pi}_1)}{\tilde{\pi}_1}\right] &= \int_0^1 \frac{f(x)}{x}p(\tilde{\pi}_1 = x)dx \\
&= \int_0^1 \mathbb{E}_{\boldsymbol{\pi}}\left[\frac{f(x)}{x}p(\tilde{\pi}_1 = x|\boldsymbol{\pi})\right]dx \\
&= \int_0^1 \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{i=1}^{\infty}\frac{f(x)}{x}\pi_i\delta(x-\pi_i)\right]dx \\
&= \mathbb{E}_{\boldsymbol{\pi}}\left[\int_0^1\sum_{i=1}^{\infty}\frac{f(x)}{x}\pi_i\delta(x-\pi_i)dx\right] \\
&= \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{i=1}^{\infty}\int_0^1\frac{f(x)}{x}\pi_i\delta(x-\pi_i)dx\right] \\
&= \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{i=1}^{\infty}f(\pi_i)\right],
\end{aligned}
$$

where the interchange of sums and integrals is justified by Fubini's theorem.

A similar method validates (11). 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 (11)

$$\sum_i \sum_{j \neq i} g(\pi_i, \pi_j) \tag{20}$$

$$= \frac{\sum_i \sum_{j \neq i} g(\pi_i, \pi_j)}{p(\tilde{\pi}_1 = \pi_i, \tilde{\pi}_2 = \pi_j)} p(\tilde{\pi}_1 = \pi_i, \tilde{\pi}_2 = \pi_j) \tag{21}$$

$$= \sum_i \sum_{j \neq i} \frac{g(\pi_i, \pi_j)}{\pi_i \pi_j} (1 - \pi_i) p(\tilde{\pi}_1 = \pi_i, \tilde{\pi}_2 = \pi_j) \tag{22}$$

$$= \mathbb{E}_{\tilde{\pi}_1, \tilde{\pi}_2} \left[\frac{g(\tilde{\pi}_1, \tilde{\pi}_2)}{\tilde{\pi}_1 \tilde{\pi}_2} (1 - \tilde{\pi}_1) \Big| \boldsymbol{\pi} \right] \tag{23}$$

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

$$\tilde{\pi}_1 \sim \text{Beta}(1 - d, \alpha + d) \tag{24}$$

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,

$$\tilde{\pi}_2 = (1 - \tilde{\pi}_1) v_2, \quad v_2 \sim \text{Beta}(1 - d, \alpha + 2d) \tag{25}$$

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

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

$$= \mathbb{E}\left[\sum_i (\pi_i \log(\pi_i))^2 \Big| d, \alpha \right] \tag{26}$$

$$+ \mathbb{E}\left[\sum_i \sum_{j \neq i} \pi_i \pi_j \log(\pi_i) \log(\pi_j) \Big| d, \alpha \right] \tag{27}$$

The first term follows directly from (10)

$$\begin{aligned}
&\mathbb{E}\left[\sum_i (\pi_i \log(\pi_i))^2 \middle| d, \alpha\right] = \int_0^1 x(-\log(x))^2 p(x|d,\alpha)\,\mathrm{d}x \\
&= B^{-1}(1-d, \alpha+d)\int_0^1 x\log^2(x)x^{1-d}(1-x)^{\alpha+d-1}\,\mathrm{d}x \\
&= \frac{1-d}{\alpha+1}\left[(\psi_0(2-d)-\psi_0(2+\alpha))^2 + \psi_1(2-d) - \psi_1(2+\alpha)\right] \qquad (28)
\end{aligned}$$

The second term of (27), requires the first two size biased samples, and follows from (11) 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. Letting $\gamma = B^{-1}(1-d, \alpha+2d)$ and $\zeta = B^{-1}(1-d, \alpha+d)$, the second term is then,

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

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

$$\begin{aligned}
&\mathrm{var}[H(\boldsymbol{\pi})|d,\alpha] = \qquad (29) \\
&\frac{1-d}{\alpha+1}\left[(\psi_0(2-d)-\psi_0(2+\alpha))^2 + \psi_1(2-d) - \psi_1(2+\alpha)\right] \\
&\quad + \frac{\alpha+d}{\alpha+1}\left[(\psi_0(1-d)-\psi_0(2+\alpha))^2 - \psi_1(2+\alpha)\right] \\
&\quad - (\psi_0(1+\alpha)-\psi_0(1-d))^2 \\
&= \frac{\alpha+d}{(\alpha+1)^2(1-d)} + \frac{1-d}{\alpha+1}\psi_1(2-d) - \psi_1(2+\alpha) \qquad (30)
\end{aligned}$$

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, $\boldsymbol{\pi}_{\mathrm{post}}$, of (14) 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, (15).

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

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 (15)

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

$$= \psi_0(\alpha + N + 1) - \frac{\alpha + Kd}{\alpha + N}\psi_0(1-d) - \frac{A}{\alpha + N}\left[\sum_{i=1}^{K}\frac{\alpha_i}{A}\psi_0(\alpha_i + 1)\right]$$
$$= \psi_0(\alpha + N + 1) - \frac{\alpha + Kd}{\alpha + N}\psi_0(1-d) - \frac{1}{\alpha + N}\left[\sum_{i=1}^{K}(n_i - d)\psi_0(n_i - d + 1)\right].$$

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 (31)

$$\operatorname{var}[H(\pi_{\text{post}})|d,\alpha] = \operatorname*{var}_{p_*}\left[\mathbb{E}_{\boldsymbol{\pi},\mathbf{p}}[H(\pi_{\text{post}})]\middle|d,\alpha\right] + \mathbb{E}_{p_*}\left[\operatorname*{var}_{\boldsymbol{\pi},\mathbf{p}}[H(\pi_{\text{post}})]\middle|d,\alpha\right] \tag{31}$$

We now seek expressions for each term in (31) in terms of the expectations already derived.

Step 1: For the right-hand term of (31), we use the independence properties of π_{post} to express the variance in terms of PYP, Dirichlet, and Beta variances

$$\mathbb{E}_{p_*}\left[\operatorname*{var}_{\boldsymbol{\pi},\mathbf{p}}[H(\pi_{\text{post}})|p_*]\middle|d,\alpha\right] \tag{32}$$
$$= \mathbb{E}_{p_*}\left[(1-p_*)^2\operatorname*{var}_{\mathbf{p}}[H(\mathbf{p})] + p_*^2\operatorname*{var}_{\boldsymbol{\pi}}[H(\boldsymbol{\pi})]\middle|d,\alpha\right]$$
$$= \frac{(N-Kd)(N-Kd+1)}{(\alpha+N)(\alpha+N+1)}\operatorname*{var}_{\mathbf{p}}[H(\mathbf{p})] + \frac{(\alpha+Kd)(\alpha+Kd+1)}{(\alpha+N)(\alpha+N+1)}\operatorname*{var}_{\boldsymbol{\pi}}[H(\boldsymbol{\pi})] \tag{33}$$

Step 2: In the left-hand term of (31) the variance is with respect to the Beta distribution, while the inner expectation is precisely the posterior mean we derived above. Expanding, we obtain

$$\operatorname*{var}_{p_*}\left[\mathbb{E}_{\boldsymbol{\pi},\mathbf{p}}[H(\pi_{\text{post}})|p_*]\middle|d,\alpha\right]$$
$$= \operatorname*{var}_{p_*}\left[(1-p_*)\mathbb{E}_{\mathbf{p}}\left[H(\mathbf{p})\right] + p_*\mathbb{E}_{\boldsymbol{\pi}}\left[H(\boldsymbol{\pi})|p_*\right] + H(p_*)\middle|d,\alpha\right] \tag{34}$$

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

$$\Omega^2(p_*) = 2p_*H(p_*)[\mathbf{B}-\mathbf{A}] + 2\mathbf{A}H(p_*) + h^2(p_*) + p_*^2[\mathbf{B}^2 - 2\mathbf{A}\mathbf{B}] + 2p_*\mathbf{A}\mathbf{B} + (1-p_*)^2\mathbf{A}^2 \tag{35}$$

and we note that

$$\operatorname*{var}_{p_*}\left[\mathbb{E}_{\boldsymbol{\pi},\mathbf{p}}[H(\pi_{\text{post}})|p_*]\middle|d,\alpha\right] = \mathbb{E}_{p_*}[\Omega^2(p_*)] - \mathbb{E}_{p_*}[\Omega(p_*)]^2 \tag{36}$$

The components composing $\mathbb{E}_{p_*}[\Omega(p_*)]$, as well as each term of (35) are derived by Archer et al. (2012). Although less elegant than the posterior mean, the expressions derived above permit us to compute (31) numerically from its component expectations, 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 (17). 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

$$H_{(d,\alpha)} \leq \psi_0(\alpha+N+1) + \frac{\alpha+Kd}{\alpha+N}\left(\gamma + \frac{1}{1-d}\right) \tag{37}$$

$$- \frac{1}{\alpha+N}\left[\sum_{i=1}^K (n_i-d)\psi_0(n_i-d+1)\right]. \tag{38}$$

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 α, it suffices to choose $\alpha_0 \gg N$. Consider the tail $\int_{\alpha_0}^\infty H_{(d,\alpha)}p(\mathbf{x}|d,\alpha)p(d,\alpha)\,\mathrm{d}\alpha$. From (37) 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

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

As a result we have that

$$\begin{aligned}\mathbb{E}[H|K_N+c,d,\mathbf{x}_N] &= \psi_0(K_N+c+N+1) \\ &- \frac{(1+d)K_N+c}{K_N+N+c}\psi_0(1-d) \\ &- \frac{1}{K_N+c+N}\left(\sum_{i=1}^{K_N}(n_i-d)\psi_0(n_i-d+1)\right)\end{aligned} \tag{39}$$

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 (39). 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

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

As a result

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

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 $\exp(-H_{(\alpha,d,N)})\le 1$, 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]

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

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

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

Therefore, it remains to show that

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

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

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

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 $\gamma(d) = \prod_{i=1}^{K_N} \Gamma(n_i - d)$).

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

Fortunately, the first integral on d will cancel with a term from the lower bound of $p(\mathbf{x}_N)$. Using[12] $\frac{(N+\alpha-1)\Gamma(\alpha+K_N)}{\Gamma(\alpha+N)} = \frac{\mathrm{Beta}(\alpha+K_N, N-K-1)}{\Gamma(N-K-1)}$,

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

Lower Bound: Again, we first integrate d

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

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

12. Note that in the argument for the inequalities we use K rather than K_N for clarity of notation.

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

where we've used the fact that $\log(\frac{1}{t})^{-1} \geq \frac{t}{1-t}$. Finally, we obtain 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, $K_N = o(N^{1-1/C})$. ■

Appendix D. Results on Unimodality of Evidence

Theorem 8 (Unimodal evidence on d) *The evidence $p(\mathbf{x}|d,\alpha)$ given by (17) has only one local maximum (unimodal) for a fixed $\alpha > 0$.*

Proof Equivalently, we show that the log evidence is unimodal.

$$\begin{aligned}L &= \log p(\mathbf{x}|d,\alpha) \\ &= \sum_{l=1}^{K-1} \log(\alpha + ld) + \sum_{i=1}^{K} \log\Gamma(n_i - d) + \log\Gamma(1+\alpha) - K\log\Gamma(1-d) - \log\Gamma(\alpha+N)\end{aligned}$$

It is sufficient to show that the partial derivative w.r.t. d has at most one positive root.

$$\begin{aligned}\frac{\partial L}{\partial d} &= \sum_{l=1}^{K-1} \frac{l}{\alpha + ld} - \sum_{i=1}^{K} \left(\psi_0(n_i - d) - \psi_0(1-d)\right) \\ &= \sum_{l=1}^{K-1} \frac{l}{\alpha + ld} + \sum_{i=1}^{K} \sum_{j=1}^{n_i - 1} \frac{1}{d-j}\end{aligned}$$

Note that as $d \to 1$, the derivative tends to $-\infty$. Combined with the observation that it is a linear combination of convex functions, there is at most one root for $\frac{\partial L}{\partial d} = 0$. ■

Theorem 9 (Unimodal evidence on α) *The evidence $p(\mathbf{x}|d,\alpha)$ given by (17) has only one local maximum (unimodal), on the region $\alpha > 0$, for a fixed d.*

Proof Similar to Theorem 8, it is sufficient to show that the partial derivative w.r.t. α has at most one positive root.

$$\begin{aligned}\frac{\partial L}{\partial \alpha} &= \sum_{l=1}^{K-1} \frac{1}{\alpha + ld} + \psi_0(1+\alpha) - \psi_0(\alpha+N) \\ &= \sum_{l=1}^{K-1} \frac{1}{\alpha + ld} - \sum_{j=1}^{N-1} \frac{1}{j+\alpha}\end{aligned}$$

Let $\alpha = \frac{1}{x}$ be a root, then,

$$\sum_{i=1}^{K-1} \frac{1}{1+xid} = \sum_{j=1}^{N-1} \frac{1}{1+xj}. \tag{40}$$

Note that since $xid < xi$, $\frac{1}{1+xid} > \frac{1}{1+xi}$ for $1 \leq i \leq K-1$. Therefore, we can split the equality as follows:

$$f_i(x) = a_i \frac{1}{1+xid} = \frac{1}{1+xj} = g_i(x) \qquad \text{for } i \leq K-1 \tag{41}$$

$$f_{ij}(x) = b_{ij} \frac{1}{1+xid} = c_{ij} \frac{1}{1+xj} = g_{ij}(x) \qquad \text{for } i \leq K-1 \text{ and } K < j < N \tag{42}$$

where $0 \leq a_i, b_{ij}, c_{ij} \leq 1$, $\forall i < K$, $a_i + \sum_j b_{ij} = 1$, and $\forall j < N$, $\sum_i c_{ij} = 1$. Fix a_i, b_{ij}, c_{ij}'s, and now suppose $\frac{1}{y} > \frac{1}{x} > 0$ is another positive root. Then, we observe the following strict inequalities due to $0 \leq d < 1$,

$$\frac{f_i(x)}{f_i(y)} = \frac{1+yid}{1+xid} < \frac{1+yj}{1+xj} = \frac{g_i(x)}{g_i(y)} \qquad \text{for } i \leq K-1 \tag{43}$$

$$\frac{f_{ij}(x)}{f_{ij}(y)} = \frac{1+yid}{1+xid} < \frac{1+yj}{1+xj} = \frac{g_{ij}(x)}{g_{ij}(y)} \qquad \text{for } i \leq K-1 \text{ and } K < j < N \tag{44}$$

Using Lemma 10 to put the sum back together, we obtain,

$$\sum_{i=1}^{K-1} \frac{1}{1+yid} > \sum_{j=1}^{N-1} \frac{1}{1+yj}. \tag{45}$$

which is a contradiction to our assumption that $\frac{1}{y}$ is a positive root. ■

Lemma 10 *If $f_j, g_j > 0$, $f_j(x) = g_j(x)$ and $\frac{f_j(y)}{f_j(x)} > \frac{g_j(y)}{g_j(x)}$ for all j, then $\sum_j f_j(y) > \sum_j g_j(y)$.*

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*, pages 2024–2032. MIT Press, Cambridge, MA, 2012.

E. Archer, I. M. Park, and J. W. Pillow. Bayesian entropy estimation for binary spike train data using parametric prior knowledge. In *Advances in Neural Information Processing Systems*, 2013.

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. *IEEE Transactions on Information Theory*, 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.

W. J. Ewens. Population genetics theory-the past and the future. In *Mathematical and Statistical Developments of Evolutionary Theory*, pages 177–227. Springer, 1990.

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 *ACM-SIAM Symposium on Discrete Algorithms*, pages 48–57. Society for Industrial and Applied Mathematics, 1995.

S. Favaro, A. Lijoi, R. H. Mena, and I. Prünster. Bayesian non-parametric inference for species variety with a two-parameter Poisson-Dirichlet process prior. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):993–1008, November 2009.

T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2): 209–230, 1973.

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. In *Advances in Neural Information Processing Systems*, page 459. MIT Press, Cambridge, MA, 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. In *Advances in Neural Information Processing Systems*, pages 399–406. MIT Press, Cambridge, MA, 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.

K. C. Knudson and J. W. Pillow. Spike train entropy-rate estimation using hierarchical dirichlet process priors. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, pages 2076–2084. Curran Associates, Inc., 2013.

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. In *Advances in Neural Information Processing Systems*, pages 471–478. MIT Press, Cambridge, MA, 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 Computation*, 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. H. Wolpert and S. DeDeo. Estimating functions of distributions defined over spaces of unknown size. *Entropy*, 15(11):4668–4699, 2013.

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. K. Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley Press, 1949.