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

Evan Archer∗124, Il Memming Park∗234, & Jonathan W. Pillow234
1. Institute for Computational and Engineering Sciences
2. Center for Perceptual Systems, 3. Dept. of Psychology,
4. Division of Statistics & Scientific Computation
The University of Texas at Austin

Abstract

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

1 Introduction

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

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

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

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

2 Entropy Estimation

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

\[
H(\pi) = -\sum_{i=1}^A \pi_i \log \pi_i,
\]

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

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

2.1 Bayesian entropy estimation

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

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

where \( p(\pi|x) \) is the posterior over \( \pi \) under some prior \( p(\pi) \) and categorical likelihood \( p(x|\pi) = \prod_j p(x_j|\pi) \),
where \( p(x_j = i) = \pi_i \). The conditional \( p(H|\pi) = \delta(H - \sum \pi_i \log \pi_i) \), since \( H \) is
deterministically related to \( \pi \). To the extent that \( p(\pi) \) 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 \( A \), the Dirichlet distribution provides an obvious
choice of prior due to its conjugacy to the discrete (or multinomial) likelihood. It takes the form
\( p(\pi) \propto \prod_{i=1}^A \pi_i^{\alpha-1} \), for \( \pi \) on the \( A \)-dimensional simplex (\( \pi_i \geq 1, \sum \pi_i = 1 \)), with concentration
parameter $\alpha$ [13]. Many previously proposed estimators can be viewed as Bayesian estimators with a particular fixed choice of $\alpha$. (See [14] for an overview).

### 2.2 Nemenman-Shafee-Bialek (NSB) estimator

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

$$p(\pi) = \int p_{\text{Dir}}(\pi | \alpha)p(\alpha) d\alpha,$$

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

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

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

$$\hat{H}_{\text{nsb}} = \mathbb{E}[H|x] = \int H(\pi)p(\pi|x, \alpha)p(\alpha|x) d\pi d\alpha = \int \mathbb{E}[H|x, \alpha] \frac{p(\pi|x, \alpha)p(\alpha)}{p(\pi)} d\alpha,$$

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

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

### 3 Stick-Breaking Priors

To construct a prior over countably infinite discrete distributions we employ a class of distributions from nonparametric Bayesian statistics known as stick-breaking processes [19]. In particular, we
focus on two well-known subclasses of stick-breaking processes: the Dirichlet Process (DP) and Pitman-Yor process (PYP). Both are stochastic processes whose samples are discrete probability distributions [7, 20]. A sample from a DP or PYP may be written as $\sum_{i=1}^{\infty} \pi_i \delta_{\phi_i}$, where $\pi = \{\pi_i\}$ denotes a countably infinite set of ‘weights’ on a set of atoms $\{\phi_i\}$ drawn from some base probability measure, where $\delta_{\phi_i}$ denotes a delta function on the atom $\phi_i$. The prior distribution over $\pi$ under the DP and PYP is technically called the GEM distribution or the two-parameter Poisson-Dirichlet distribution, but we will abuse terminology and refer to it more simply as script notation $\mathcal{DP}$ or $\mathcal{PY}$. The DP weight distribution $\mathcal{DP}(\alpha)$ may be described as a limit of the finite Dirichlet distributions where the alphabet size grows and concentration parameter shrinks, $A \to \infty$ and $\alpha' \to 0$, such that $\frac{\alpha'}{\alpha} \to \alpha$ [20]. The PYP generalizes the DP to allow power-law tails, and includes DP as a special case [7].

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

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

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

### 3.1 Expectations over DP and PY weight distributions

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

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

$$E_{(\pi|d,\alpha)} \left[ \sum_{i=1}^{\infty} f(\pi_i) \right] = E_{(\tilde{\pi}_1|d,\alpha)} \left[ \frac{f(\tilde{\pi}_1)}{\tilde{\pi}_1} \right], \quad (7)$$

$$E_{(\pi|d,\alpha)} \left[ \sum_{i,j \neq i} g(\pi_i, \pi_j) \right] = E_{(\tilde{\pi}_1, \tilde{\pi}_2|d,\alpha)} \left[ \frac{g(\tilde{\pi}_1, \tilde{\pi}_2)}{\tilde{\pi}_1 \tilde{\pi}_2} \right] (1 - \tilde{\pi}_1), \quad (8)$$

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

**Erratum:** Corrected eq. 8, May 2013

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

---

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

2 Note that the power-law exponent is given incorrectly in [9, 22].
3.2 Posterior distribution over weights

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

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

\[
\begin{align*}
(p_1, p_2, p_3, \ldots, p_K, p_\ast) &\sim \text{Dir}(n_1 - d, n_2 - d, \ldots, n_K - d, \alpha + Kd) \\
\pi &:= (\pi_1, \pi_2, \pi_3, \ldots) \sim \mathcal{PY}(d, \alpha + Kd).
\end{align*}
\]

4 Bayesian entropy inference with PY priors

4.1 Fixed PY priors

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

\[
\begin{align*}
\mathbb{E}[H(\pi)|d, \alpha] &= \psi_0(1 + \alpha) - \psi_0(1 - d), \\
\text{var}[H(\pi)|d, \alpha] &= \frac{\alpha + d}{(1 + \alpha)^2(1 - d)} + \frac{1 - d}{1 + \alpha} \psi_1(2 - d) - \psi_1(2 + \alpha),
\end{align*}
\]

where \( \psi_n \) is the polygamma of \( n \)-th order (i.e., \( \psi_0 \) is the digamma function). Fig. 3 shows these functions for a range of \( d \) and \( \alpha \) values. These reveal the same phenomenon that [6] observed for finite Dirichlet distributions: a \( \mathcal{PY} \) prior with fixed \( (d, \alpha) \) induces a narrow prior over \( H \). In the undersampled regime, Bayesian estimates under \( \mathcal{PY} \) priors will therefore be strongly determined by the choice of \( (d, \alpha) \), and posterior credible intervals will be unrealistically narrow.\(^3\)

4.2 Pitman-Yor process mixture (PYM) prior

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

\(^3\)The only exception is near the corner \( d \to 1 \) and \( \alpha \to -d \). There, one can obtain arbitrarily large prior variance over \( H \) for given mean. However, these such priors have very heavy tails and seem poorly suited to data with finite or exponential tails; we do not explore them further here.
Although in practice we parametrize our prior using the variables \(d, \alpha\) about tail behavior of \(\pi\), where \(\pi\) is any density on \([0, \infty]\). We refer to the resulting prior distribution over \(\pi\) as the Pitman-Yor mixture (PYM) prior. All results in the figures are generated using the prior \(q(\gamma) \propto \max(1 - \gamma, 0)\).

### 4.3 Posterior inference

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

\[
H_{\text{PYM}} = \mathbb{E}[H|x] = \int \mathbb{E}[H|x, d, \alpha] \frac{p(x|d, \alpha)p(\alpha, d)}{p(x)} \; d(\alpha, d),
\]

(14)

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

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

(15)

The evidence, \(p(x|d, \alpha)\), is given by

\[
p(x|d, \alpha) = \frac{\prod_{i=1}^{K} \Gamma(n_i - d)}{\Gamma(1 - d)^K } \frac{\Gamma(1 + \alpha)}{\Gamma(\alpha + N)}. \]

(16)
We can obtain confidence regions for $\hat{H}_{\text{PYM}}$ by computing the posterior variance $\mathbb{E}[(H - \hat{H}_{\text{PYM}})^2|x]$. The estimate takes the same form as eq. 14, except that we substitute $\text{var}[H|x,d,\alpha]$ for $\mathbb{E}[H|x,d,\alpha]$. Although $\text{var}[H|x,d,\alpha]$ has an analytic closed form that is fast to compute, it is a lengthy expression that we do not have space to reproduce here; we provide it in the Appendix.

### 4.4 Computation

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

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

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

### 4.5 Existence of posterior mean

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

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

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

### 5 Results

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

### 6 Discussion

In this paper we introduced PYM, a novel entropy estimator for distributions with unknown support. We derived analytic forms for the conditional mean and variance of entropy under a DP and PY prior for fixed parameters. Inspired by the work of [6], we defined a novel PY prior mixture prior, PYM, which implies an approximately flat prior on entropy. PYM addresses two major issues with NSB: its dependence on knowledge of $A$ and its inability (inherited from the Dirichlet distribution) to...
Figure 5: Convergence of entropy estimators with sample size, on two simulated and two real datasets. We write “MiMa” for “Miller-Maddow” and “NSB∞” for \( \hat{H}_{\text{nsb,}\infty} \). Note that DPM (“DP mixture”) is simply a PYM with \( \gamma = 0 \). Credible intervals are indicated by two standard deviation of the posterior for DPM and PYM. (A) Exponential distribution \( \pi_i \propto e^{-i} \). (B) Power-law distribution with exponent 2 \( (\pi_i \propto i^{-2}) \). (C) Word frequency from the novel Moby Dick. (D) Neural words from \( 8 \) simultaneously-recorded retinal ganglion cells. Note that for clarity \( \hat{H}_{\text{nsb,}\infty} \) has been cropped from B and D. All plots are average of 16 Monte Carlo runs.

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

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

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

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

References

Appendix: Bayesian estimation of discrete entropy with mixtures of stick-breaking priors
Evan Archer, Il Memming Park, & Jonathan W. Pillow

A Gamma and polygamma functional identities

As an aid to the reader, we provide a brief review of the gamma polygamma functions.

\[ \Gamma(x + 1) = x\Gamma(x), \quad \psi_0(x + 1) = \psi_0(x) + \frac{1}{x}, \quad \psi_1(x + 1) = \psi_1(x) - \frac{1}{x^2} \quad (18) \]

A.1 Beta integrals

If \( X \sim \text{Beta}(a, b) \), then \( (1 - X) \sim \text{Beta}(b, a) \).

\[ \int_0^1 x^{a-1}(1-x)^{b-1} \, dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = B(a, b) \]

\[ \int_0^1 x^p(1-x)^q\text{Beta}(x; a, b) \, dx = \frac{B(a+p, b+q)}{B(a, b)} = \frac{\Gamma(a+p)\Gamma(b+q)\Gamma(a+b)}{\Gamma(a+b+p+q)\Gamma(a)\Gamma(b)} \]

\[ \int_0^1 x^p(1-x)^q\log(x) \, dx = B(p, q)[\psi_0(p) - \psi_0(p + q)] \]

\[ \int_0^1 x^p(1-x)^q\log^2(x) \, dx = B(p, q)[(\psi_0(p) - \psi_0(p + q))^2 + \psi_1(p) - \psi_1(p + q)] \]

\[ \int_0^1 x^p(1-x)^q\log(x)\log(1-x) \, dx = \frac{p}{p+q+1}B(p, q + 1)[\psi_1(p + q + 2) + (\psi_0(p + 1) - \psi_0(p + q + 2))(\psi_0(q + 1) - \psi_0(p + q + 2))] \]

\[ \mathbb{E}[XH(X)] = \frac{a(a + 1)}{(a + b)(a + a + 1)} [\psi_0(a + b + 2) - \psi_0(a + 2)] \]

\[ + \frac{ab}{(a + b)(a + b + 1)} [\psi_0(a + b + 2) - \psi_0(b + 1)] \]

Although a matter of straightforward computation, as we will require it below and we did not find it in standard tables, we provide the derivation of \( \mathbb{E}[XH(X)] \) for \( X \sim \text{Beta}(a, b) \). Letting \( c = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \),

\[ -c\mathbb{E}[XH(X)] = \int_0^1 \left[ x^2 \log(x) + x(1-x) \log(1-x) \right] x^{a-1}(1-x)^{b-1} \, dx \]

\[ = \int_0^1 \log(x)x^{a+2-1}(1-x)^{b-1} \, dx + \int_0^1 \log(1-x)x^{a+1-1}(1-x)^{b+1-1} \, dx \]

\[ = \left( \frac{\Gamma(a + b + 2)}{\Gamma(a + 2)\Gamma(b)} \right)^{-1} \int_0^1 [\log(x)b(x; a + 2, b)] \, dx \]

\[ + \left( \frac{\Gamma(a + b + 2)}{\Gamma(a + 1)\Gamma(b + 1)} \right)^{-1} \int_0^1 [\log(x)b(x; b + 1, a + 1)] \, dx \]

\[ = \frac{(a)(a + 1)}{(a + b)(a + b + 1)} \left( \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \right)^{-1} [\psi_0(a + 2) - \psi_0(a + b + 2)] \]

\[ + \frac{ab}{(a + b)(a + b + 1)} \left( \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \right)^{-1} [\psi_0(b + 1) - \psi_0(a + b + 2)] \]

Thus,

\[ \mathbb{E}[XH(X)] = \frac{(a)(a + 1)}{(a + b)(a + b + 1)} [\psi_0(a + b + 2) - \psi_0(a + 2)] \]

\[ + \frac{ab}{(a + b)(a + b + 1)} [\psi_0(a + b + 2) - \psi_0(b + 1)] . \]
B Proof of Theorem 1

In this Appendix we give a proof for Theorem 1.

Proof. PYM is given by

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

where we write $H_{(d,\alpha)} := \mathbb{E}[H|d,\alpha,x]$ and $p(x|d,\alpha)$ is the evidence, given by eq. 16. We will assume $p(d,\alpha) = 1$ for all $\alpha$ 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:

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

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,

$$
H_{(d,\alpha)} \leq \psi_0(\alpha + N + 1) + \frac{\alpha + Kd}{\alpha + N} \left( \gamma + \frac{1}{1 - d} \right) - \frac{1}{\alpha + N} \left[ \sum_{i=1}^K (n_i - d) \psi_0(n_i - d + 1) \right]. \tag{19}
$$

Although second term is unbounded in $d$ notice that $\frac{\Gamma(n_i - d)}{\Gamma(1 - d)} = \prod_{i=1}^{n_i-1} (1 - d + i)$; thus, when $N - K \geq 1, H_{(d,\alpha)} p(x|d,\alpha)$ is integrable in $d$.

For the integral over alpha, it suffices to choose $\alpha_0 \gg N$ and consider the tail, $\int_{\alpha_0}^\infty H_{(d,\alpha)} p(x|d,\alpha) p(d,\alpha) \, d\alpha$. From eq. 19 and asymptotic expansion $\psi(x) = \log(x) - \frac{1}{2x} - \frac{1}{12x^2} + O(\frac{1}{x^3})$ as $x \to \infty$ to find 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 with respect to $\alpha$ that depends on $K$, $N$, and $d$. Thus, we have

$$
\int_{\alpha_0}^\infty H_{(d,\alpha)} p(x|d,\alpha) p(d,\alpha) \, d\alpha \leq \sum_{i=1}^K \frac{\Gamma(n_i - d)}{\Gamma(1 - d)} \int_{\alpha_0}^\infty \left( \log(\alpha + N + 2) + \frac{c}{\alpha} \right) \frac{1}{\alpha^{N-K}} \, d\alpha
$$

and so $H_{(d,\alpha)}$ is integrable in $\alpha$ when $N - K \geq 2$. \hfill \Box

C Finite Dirichlet distribution

For the aid of the reader, we provide a summary of well-known properties of the finite Dirichlet distribution which employ elsewhere. The Dirichlet distribution is a distribution over discrete probability distributions.

**Definition 1** (Dirichlet distribution). Given the concentration parameters $\alpha_1, \ldots, \alpha_K$,

$$
\Pr\{\pi_1, \ldots, \pi_K\} = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_i \pi_i^{\alpha_i - 1}, \tag{20}
$$

where $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} \, dt$ is the gamma function.

Note that since $\sum_i \pi_i = 1$, it has $K-1$ degrees of freedom. The space of such discrete probability distributions is a simplex denoted as $\Delta_K$, hence Dirichlet distribution is a (continuous) probability measure over the simplex. The mean of $\text{Dir}(\alpha)$ is $\frac{\alpha}{\sum_i \alpha_i}$.

**Lemma 2** (Aggregation (agglomerative) property). If,

$$(\pi_1, \pi_2, \ldots, \pi_K) \sim \text{Dir}(\alpha_1, \alpha_2, \ldots, \alpha_K).$$

Then,

$$(\pi_1 + \pi_2, \ldots, \pi_K) \sim \text{Dir}(\alpha_1 + \alpha_2, \alpha_3, \ldots, \alpha_K).$$


Lemma 3 (Decimative property). If,

\[(\pi_1, \pi_2, \ldots, \pi_K) \sim \text{Dir}(\alpha_1, \alpha_2, \ldots, \alpha_K)\]
\[\tau \sim \text{Beta}(\alpha_1 \beta, \alpha_1 (1 - \beta)).\]

Then,

\[(\pi_1 \tau, \pi_1 (1 - \tau), \pi_2, \ldots, \pi_K) = \text{Dir}(\alpha_1 \beta, \alpha_1 (1 - \beta), \alpha_2, \ldots, \alpha_K)\]

Lemma 4 (Neutrality property). If,

\[(\pi_1, \pi_2, \ldots, \pi_K) \sim \text{Dir}(\alpha_1, \alpha_2, \ldots, \alpha_K).\]

Then,

\[\pi_1 \perp (\pi_2 \frac{\pi_3}{1 - \pi_1}, \ldots, \frac{\pi_K}{1 - \pi_1}),\]

where \(X \perp Y\) denotes independence of two random variables \(X\) and \(Y\).

Dirichlet distribution is conjugate to multinomial distribution.

Definition 2 (Multinomial distribution). Given \(\pi_1, \pi_2, \ldots, \pi_K\) such that \(\pi_i \geq 0\) and \(\sum_i \pi_i = 1\), the multino-
mial probability of observing \(n_i\) number of samples corresponding to \(\pi_i\) is

\[
\Pr\{n_1, \ldots, n_K\} = \frac{\Gamma(N + 1)}{\prod_i \Gamma(n_i + 1)} \prod_i \pi_i^{n_i},
\]

where \(N = \sum_i n_i\) is the total number of samples.

Lemma 5 (Conjugacy of Dirichlet distribution). Given observation counts \(n_1, \ldots, n_K\), the posterior for multinomial distribution with Dirichlet prior \(\text{Dir}(\alpha_1, \ldots, \alpha_K)\) is \(\text{Dir}(\alpha_1 + n_1, \ldots, \alpha_K + n_K)\).

D Derivations of Dirichlet and \(\mathcal{P}Y\) moments

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

D.1 Mean entropy of finite Dirichlet

Proposition 2 (Replica trick for entropy [27]). For \(\pi \sim \text{Dir}(\alpha_1, \alpha_2, \ldots, \alpha_A)\), such that \(\sum_i \alpha_i = A\), and letting \(\vec{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_A)\), we have

\[
E[H(\pi) | \vec{\alpha}] = \psi_0(A + 1) - \sum_{i=1}^A \frac{\alpha_i}{A} \psi_0(\alpha_i + 1)
\] (21)
Proof. First, let \( c \) be the normalizer of Dirichlet, \( c = \frac{\prod \Gamma(\alpha_j)}{\Gamma(A)} \) and let \( \mathcal{L} \) denote the Laplace transform (on \( \pi \) to \( s \)). Now,

\[
\begin{align*}
\mathbb{E}[H|\vec{\alpha}] &= \int \left( -\sum_i \pi_i \log_2 \pi_i \right) \delta(\Sigma, \pi_i - 1) \prod_j \pi_j^{\alpha_j-1} d\pi \\
&= -\sum_i \int \left( \frac{d}{d(\alpha_i)} \pi_i \right) \delta(\Sigma, \pi_i - 1) \prod_j \pi_j^{\alpha_j-1} d\pi \\
&= -\sum_i \frac{d}{d(\alpha_i)} \prod_j \pi_j^{\alpha_j-1} d\pi \\
&= -\sum_i \frac{d}{d(\alpha_i)} \mathcal{L}^{-1} \left[ \mathcal{L}(\pi_i^{\alpha_i}) \prod_j \mathcal{L}(\pi_j^{\alpha_j-1}) \right] (1) \\
&= -\sum_i \frac{d}{d(\alpha_i)} \mathcal{L}^{-1} \left[ \frac{\Gamma(\alpha_i + 1) \prod_{j \neq i} \Gamma(\alpha_j)}{\Gamma(\sum_i \alpha_i + 1)} \right] \prod_j \Gamma(\alpha_j) \\
&= -\sum_i \left( \frac{\Gamma(\alpha_i + 1)}{\Gamma(\sum_i \alpha_i + 1)} \right) [\psi_0(\alpha_i + 1) - \psi_0(A + 1)] \prod_{j \neq i} \Gamma(\alpha_j) \\
&= \left[ \psi_0(A + 1) - \sum_{i=1}^A \frac{\alpha_i}{A} \psi_0(\alpha_i + 1) \right] \prod_j \Gamma(\alpha_j) / \Gamma(A),
\end{align*}
\]

\( \square \)

D.2 Variance of entropy of finite Dirichlet distribution

We derive \( \mathbb{E}[H^2(\pi)|\vec{\alpha}] \). The variance of entropy is then given by \( \text{var}[H(\pi)|\vec{\alpha}] = \mathbb{E}[H^2(\pi)|\vec{\alpha}] - \mathbb{E}[H(\pi)|\vec{\alpha}]^2 \).

Proposition 3. For \( \pi \sim \text{Dir}(\alpha_1, \alpha_2, \ldots, \alpha_A) \), such that \( \sum_{i=1}^A \alpha_i = A \), and letting \( \vec{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_A) \), we have

\[
\begin{align*}
\mathbb{E}[H^2(\pi)|\vec{\alpha}] &= \sum_{i \neq k} \left[ \frac{\alpha_i \alpha_k}{(A + 1)(A)} \right] [\psi_0(\alpha_k + 1) - \psi_0(A + 2)] (\psi_0(\alpha_i + 1) - \psi_0(A + 2) - \psi_1(A + 2)] \\
&\quad + \sum_i \left[ \frac{\alpha_i(\alpha_i + 1)}{(A + 1)(A)} \right] ([\psi_0(\alpha_i + 2) - \psi_0(A + 2)]^2 + \psi_1(\alpha_i + 2) - \psi_1(A + 2)]
\end{align*}
\]

Proof. We can evaluate the second moment in a manner similar to the mean entropy above. First, we will split 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, \( c = \prod_i \frac{\Gamma(\alpha_j)}{\Gamma(A)} \) we have,

\[
\begin{align*}
c\mathbb{E}[H^2|\vec{\alpha}] &= \int \left( -\sum_i \pi_i \log_2 \pi_i \right)^2 \delta(\Sigma, \pi_i - 1) \prod_j \pi_j^{\alpha_j-1} d\pi \\
&= \sum_i \int \pi_i^{\alpha_i+1} \log_2 \pi_i \delta(\Sigma, \pi_i - 1) \prod_j \pi_j^{\alpha_j-1} d\pi \\
&\quad + \sum_{i \neq k} \int (\pi_i^{\alpha_i} \log_2 \pi_i) (\pi_k^{\alpha_k} \log_2 \pi_k) \delta(\Sigma, \pi_i - 1) \prod_{j \neq (i,k)} \pi_j^{\alpha_j-1} d\pi \\
&= \sum_i \frac{d^2}{d(\alpha_i + 1)^2} \int \pi_i^{\alpha_i+1} \delta(\Sigma, \pi_i - 1) \prod_{j \neq i} \pi_j^{\alpha_j-1} d\pi \\
&\quad + \sum_{i \neq k} \frac{d}{d\alpha_i} \frac{d}{d\alpha_k} \int (\pi_i^{\alpha_i}) (\pi_k^{\alpha_k}) \delta(\Sigma, \pi_i - 1) \prod_{j \neq (i,k)} \pi_j^{\alpha_j-1} d\pi
\end{align*}
\]
Assuming \( i \neq k \), these will be the cross terms.

\[
\int (\pi_i \log_2 \pi_i)(\pi_k \log_2 \pi_k) \delta(\sum_i \pi_i - 1) \prod_j \pi_j^{\alpha_j - 1} d\pi
\]

\[
= \frac{d}{d\alpha_i} \frac{d}{d\alpha_k} \int (\pi_i^{\alpha_i})(\pi_k^{\alpha_k}) \delta(\sum_i \pi_i - 1) \prod_{j \neq (i,k)} \pi_j^{\alpha_j - 1} d\pi
\]

\[
= \frac{d}{d\alpha_i} \frac{d}{d\alpha_k} \left[ \frac{\Gamma(\alpha_i + 1) \Gamma(\alpha_k + 1)}{\Gamma(A + 2)} \right] \prod_{j \neq (i,k)} \Gamma(\alpha_j)
\]

\[
= \frac{d}{d\alpha_i} \left[ \frac{\alpha_i \Gamma(\alpha_i + 1)}{\Gamma(A + 2)} \psi_0(\alpha_i + 1) - \frac{\alpha_i \Gamma(\alpha_i + 1)}{\Gamma(A + 2)} \psi_0(\alpha_i + 1) \right] \prod_{j \neq i} \Gamma(\alpha_j)
\]

\[
= \frac{1}{(A + 2)} \left[ (\psi_0(\alpha_i + 1) - \psi_0(A + 2)) \psi_0(\alpha_i + 1) - \psi_0(A + 2) \right] \prod_{j \neq i} \Gamma(\alpha_j)
\]

\[
= \frac{1}{(A + 2)} \left[ (\psi_0(\alpha_i + 1) - \psi_0(A + 2)) (\psi_0(\alpha_i + 1) - \psi_0(A + 2)) - \psi_0(A + 2) \right] \prod_{j \neq i} \Gamma(\alpha_j)
\]

\[
= \frac{d^2}{d(\alpha_i + 1)^2} \int \pi_i^{\alpha_i + 1} \delta(\sum_i \pi_i - 1) \prod_j \pi_j^{\alpha_j - 1} d\pi = \frac{d^2}{d(\alpha_i + 1)^2} \left[ \frac{\Gamma(\alpha_i + 2)}{\Gamma(A + 2)} \right] \prod_{j \neq i} \Gamma(\alpha_j)
\]

\[
\sum \text{summing over all terms and adding the cross and square terms, we recover the desired expression for } E[H^2(\Pi)\mid \alpha]. \]

\[
\square \]

D.3 Prior entropy mean and variance under \([\mathcal{P}]\)

We derive the prior entropy mean and variance of a \([\mathcal{P}]\) distribution with fixed parameters \( \alpha \) and \( d \), \( E[\mathcal{H}(\pi) \mid d, \alpha] \) and \( \text{var}[\mathcal{H}(\pi) \mid d, \alpha] \). We first prove our Proposition 1 (mentioned in [7]), which allows us to compute expectations over \([\mathcal{P}]\) using the first size biased sample, \( \tilde{\pi}_1 \), with the identity \( E \left[ \sum_{i=1}^{\infty} f(\pi_i) \mid \alpha \right] = \int_0^1 f(x) p(\tilde{\pi}_1 \mid \alpha) dx \).

**Proof of Proposition 1.** First we validate eq. 7. Writing out the general form of the size-biased sample,

\[
p(\tilde{\pi}_1 = x \mid \pi) = \sum_{i=1}^{\infty} \pi_i \delta(x - \pi_i),
\]

we see that

\[
E_{\delta_1} \left[ \frac{f(\tilde{\pi}_1)}{\tilde{\pi}_1} \right] = \int_0^1 \frac{f(x)}{x} p(\tilde{\pi}_1 = x) dx = \int_0^1 E_{\hat{\pi}} \left[ \frac{f(x)}{x} p(\tilde{\pi}_1 = x \mid \pi) \right] dx
\]

\[
= \int_0^1 E_{\hat{\pi}} \left[ \sum_{i=1}^{\infty} \frac{f(x)}{x} \pi_i \delta(x - \pi_i) \right] dx = E_{\hat{\pi}} \left[ \sum_{i=1}^{\infty} \int_0^1 \frac{f(x)}{x} \pi_i \delta(x - \pi_i) dx \right]
\]

\[
= E_{\hat{\pi}} \left[ \sum_{i=1}^{\infty} f(\pi_i) \right].
\]

14
where the interchange of sums and integrals is justified by Fubini’s theorem.

A similar method validates eq. 8. 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 eq. 8,

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

(23)

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

\[
p(\tilde{\pi}_1 = \pi_i, \tilde{\pi}_2 = \pi_j) = p(\tilde{\pi}_1 = \pi_i)p(\tilde{\pi}_2 = \pi_j | \tilde{\pi}_1 = \pi_i) = \pi_i \cdot \frac{\pi_j}{1 - \pi_i}.
\]

(24)

We will compute the second moment explicitly, splitting \( \sum_{i} \sum_{j \neq i} \log(\pi_i) \) into square and cross terms,

\[
E[H(\pi)^2 | d, \alpha] = E \left[ \left( \sum_i \pi_i \log(\pi_i) \right)^2 \right]_{d, \alpha} = E \left[ \sum_i (\pi_i \log(\pi_i))^2 \right]_{d, \alpha} + E \left[ \sum_i \sum_{j \neq i} \pi_i \pi_j \log(\pi_i) \log(\pi_j) \right]_{d, \alpha} \]

(25)

The first term follows directly from eq. 7,

\[
E \left[ \sum_i (\pi_i \log(\pi_i))^2 \right]_{d, \alpha} = \int_0^1 x(-\log(x))^2 p(x | d, \alpha) \, dx = B^{-1}(1 - d, \alpha + d) \int_0^1 x \log^2(x)x^{1-d}(1-x)^{\alpha+d-1} \, dx = \frac{1 - d}{1 + \alpha} \left[ (\psi_0(2 - d) - \psi_0(2 + \alpha))^2 + \psi_1(2 - d) - \psi_1(2 + \alpha) \right]
\]

(26)

The second term of eq. 29, requires the first two size biased samples, and follows from eq. 8 with \( g(x, y) = \log(x) \log(y) \). For the PY prior, it is easier to integrate on \( V_1 \) and \( V_2 \), rather than the size biased samples. The
second term is then,
\[
\mathbb{E}[\mathbb{E} [\log(\tilde{\pi}) | \pi] | \alpha] = \mathbb{E}[\mathbb{E} [\log(V_1) \log((1 - V_1)V_2)(1 - V_1)|\pi] | \alpha]
\]
\[
= B^{-1}(1 - d, \alpha + d)B^{-1}(1 - d, \alpha + 2d) \int_0^1 \int_0^1 \log(v_1) \log((1 - v_1)v_2)(1 - v_1)v_1^{\alpha - d}(1 - v_1)^{\alpha + d - 1} v_2^{d}(1 - v_2)^{\alpha + 2d - 1} dv_1 dv_2
\]
\[
= B^{-1}(1 - d, \alpha + d) \left[ \int_0^1 \log(v_1) \log((1 - v_1)v_2)(1 - v_1)v_1^{\alpha - d}(1 - v_1)^{\alpha + d - 1} dv_1 
\right. 
+ B^{-1}(1 - d, \alpha + 2d) \int_0^1 \log(v_1) \log((1 - v_1)v_2)v_1^{\alpha - d}(1 - v_1)^{\alpha + d - 1} \int_0^1 \log(v_2)v_2^{d}(1 - v_2)^{\alpha + 2d - 1} dv_1 dv_2 \right]
\]
\[
= \frac{\alpha + d}{1 + \alpha} \left[ (\psi(0(1 - d) - (2 + \alpha)) - 1(2 + \alpha) \right]
\]

Finally combining the terms, the variance of the entropy under PYP prior is
\[
\text{var}[H(\pi)|d, \alpha] = \frac{1}{1 + \alpha} \left[ (\psi(0(2 - d) - (2 + \alpha)) \right. 
\] 
\[
+ \frac{\alpha + d}{1 + \alpha} (\psi(0(1 - d) - (2 + \alpha)) \right. 
\] 
\[
- (\psi(0(1 + d) - (2 + \alpha)) \right. 
\] 
\[
= \frac{\alpha + d}{1 + \alpha} \left[ (\psi(0(1 - d) - (2 + \alpha)) \right. 
\]

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, eq. 15.
\[
\mathbb{E}[H(\pi_{\text{post}}) | d, \alpha] = \mathbb{E} \left[ - \sum_{i=1}^{K} p_i \log p_i - p_* \sum_{i=1}^{\infty} \pi_i \log \pi_i \right] | d, \alpha]
\]
\[
= \mathbb{E} \left[ - (1 - p_*) \sum_{i=1}^{K} \frac{p_i}{1 - p_*} \log \left( \frac{p_i}{1 - p_*} \right) - p_* \sum_{i=1}^{\infty} \pi_i \log \pi_i + H(p_*) \right] | d, \alpha]
\]
\[
= \mathbb{E} \left[ - (1 - p_*) \sum_{i=1}^{K} \frac{p_i}{1 - p_*} \log \left( \frac{p_i}{1 - p_*} \right) - p_* \sum_{i=1}^{\infty} \pi_i \log \pi_i + H(p_*) \right] | d, \alpha]
\]
\[
= \mathbb{E} \left[ - (1 - p_*) \mathbb{E}_p [H(p) | d, \alpha] + p_* H(\pi) + H(p_*) \right] | d, \alpha]
\]
\[
= \mathbb{E}_{p_*} [ (1 - p_*) \mathbb{E}_p [H(p) | d, \alpha] + p_* \mathbb{E}_\pi [H(\pi) | d, \alpha] + H(p_*) ]
\]
using the formulae for $E_p[H(p)|d, \alpha]$, $E_\pi[H(\pi)|d, \alpha]$, and $E_{p^*}[H(p^*)]$ and rearranging terms, we obtain eq. 15,

$$
\mathbb{E}[H(\pi_{post})|d, \alpha] = \frac{A}{\alpha + N} \mathbb{E}_p[H(p)] + \frac{\alpha + K d}{\alpha + N} \mathbb{E}_\pi[H(\pi)] + E_{p^*}[H(p^*)]
$$

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

$$
= \psi_0(\alpha + N + 1) - \frac{\alpha + K d}{\alpha + N} \psi_0(1 - d) - \frac{1}{\alpha + N} \sum_{i=1}^{K} (n_i - d) \psi_0(n_i - d + 1)
$$

Derivation of posterior variance: We continue the notation from the subsection above. In order to exploit the independence properties of $\pi_{post}$ we first apply the law of total variance to obtain eq. 31,

$$
\text{var}[H(\pi_{post})|d, \alpha] = \text{var}_{\pi,p}[\mathbb{E}_p[H(\pi_{post})]|d, \alpha] + \mathbb{E}_p_p[\text{var}_{\pi,p}[H(\pi_{post})]|d, \alpha]
$$

(31)

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

Step 1: For the right-hand term of eq. 31, we use the independence properties of $\pi_{post}$ to express the variance in terms of PY, Dirichlet, and Beta variances,

$$
\mathbb{E}_{p^*}[\text{var}_{\pi,p}[H(\pi_{post})]|d, \alpha] = \mathbb{E}_{p^*}\left[ (1 - p^*)^2 \text{var}_p[H(p)] + p^*^2 \text{var}_\pi[H(\pi)] | d, \alpha \right]
$$

$$
= \frac{(N - K d)(N - K d + 1)}{(\alpha + N)(\alpha + N + 1)} \text{var}_p[H(p)] + \frac{(\alpha + K d)(\alpha + K d + 1)}{(\alpha + N)(\alpha + N + 1)} \text{var}_\pi[H(\pi)]
$$

(32)

Step 2: In the left-hand term of eq. 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,

$$
\text{var}_{p^*}[\mathbb{E}_{p,p}[H(\pi_{post})]|d, \alpha] = \text{var}_{p^*}\left[ (1 - p^*)\mathbb{E}_p[H(p)] + p^*\mathbb{E}_\pi[H(\pi)] + h(p^*) | d, \alpha \right]
$$

(33)

To evaluate this integral, we introduce some new notation,

$$
\mathcal{A} := \mathbb{E}_p[H(p)]
$$

$$
\mathcal{B} := \mathbb{E}_\pi[H(\pi)]
$$

$$
\Omega(p^*) := (1 - p^*) \mathbb{E}_p[H(p)] + p^* \mathbb{E}_\pi[H(\pi)] + h(p^*)
$$

$$
= (1 - p^*) \mathcal{A} + p^* \mathcal{B} + h(p^*)
$$

so that

$$
\Omega^2(p^*) = 2p^* h(p^*) [\mathcal{B} - \mathcal{A}] + 2Ah(p^*) + h^2(p^*) + p^2[\mathcal{B}^2 - 2A\mathcal{B}] + 2p^* A \mathcal{B} + (1 - p^*)^2 \mathcal{A}^2
$$

(34)

and we note that

$$
\text{var}_{p^*}[\mathbb{E}_{p,p}[H(\pi_{post})]|d, \alpha] = \mathbb{E}_{p^*}[\Omega^2(p^*)] - \mathbb{E}_{p^*}[\Omega(p^*)]^2
$$

(35)

In Appendix A we derive expressions for the components composing $\mathbb{E}_{p^*}[\Omega(p^*)]$, as well as each term of eq. 34. Although less elegant than the posterior mean, the expressions derived above permit us to compute eq. 31 numerically from its component expectations, without sampling.