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.

$$\Gamma(x+1) = x\Gamma(x), \qquad \psi_0(x+1) = \psi_0(x) + \frac{1}{x}, \qquad \psi_1(x+1) = \psi_1(x) - \frac{1}{x^2} \tag{18}$$

A.1 Beta integrals

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

$$\begin{aligned}
\int_0^1 x^{a-1}(1-x)^{b-1}\,\mathrm{d}x &= \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = B(a,b) \\
\int_0^1 x^p(1-x)^q \mathrm{Beta}(x;a,b)\,\mathrm{d}x &= \frac{B(a+p,b+q)}{B(a,b)} = \frac{\Gamma(a+p)\Gamma(b+q)\Gamma(a+b)}{\Gamma(a+b+p+q)\Gamma(a)\Gamma(b)} \\
\int_0^1 x^p(1-x)^q \log(x)\,\mathrm{d}x &= B(p,q)[\psi_0(p) - \psi_0(p+q)] \\
\int_0^1 x^p(1-x)^q \log^2(x)\,\mathrm{d}x &= B(p,q)[(\psi_0(p) - \psi_0(p+q))^2 + \psi_1(p) - \psi_1(p+q)] \\
\int_0^1 x^p(1-x)^q \log(x)\log(1-x)\,\mathrm{d}x &= \frac{p}{p+q+1}B(p,q+1)[\psi_1(p+q+2) \\
&\quad + (\psi_0(p+1) - \psi_0(p+q+2))(\psi_0(q+1) - \psi_0(p+q+2))] \\
\mathbb{E}[XH(X)] &= \frac{a(a+1)}{(a+b)(a+b+1)}\left[\psi_0(a+b+2) - \psi_0(a+2)\right] \\
&\quad + \frac{ab}{(a+b)(a+b+1)}\left[\psi_0(a+b+2) - \psi_0(b+1)\right]
\end{aligned}$$

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 $X \sim \mathrm{Beta}(a,b)$. Letting $c = \left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\right)^{-1}$,

$$\begin{aligned}
-c\mathbb{E}[XH(X)] &= \int_0^1 \left[x^2\log(x) + x(1-x)\log(1-x)\right] x^{(a-1)}(1-x)^{(b-1)}dx \\
&= \int_0^1 \log(x)x^{(a+2-1)}(1-x)^{(b-1)}dx + \int_0^1 \log(1-x)x^{(a+1-1)}(1-x)^{(b+1-1)}dx \\
&= \left(\frac{\Gamma(a+b+2)}{\Gamma(a+2)\Gamma(b)}\right)^{-1}\int_0^1 [\log(x)b(x;a+2,b)]dx \\
&\quad + \left(\frac{\Gamma(a+b+2)}{\Gamma(a+1)\Gamma(b+1)}\right)^{-1}\int_0^1 [\log(x)b(x;b+1,a+1)]dx \\
&= \frac{(a)(a+1)}{(a+b)(a+b+1)}\left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\right)^{-1}[\psi_0(a+2) - \psi_0(a+b+2)] \\
&\quad + \frac{ab}{(a+b)(a+b+1)}\left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\right)^{-1}[\psi_0(b+1) - \psi_0(a+b+2)]
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{E}[XH(X)] &= \frac{(a)(a+1)}{(a+b)(a+b+1)}\left[\psi_0(a+b+2) - \psi_0(a+2)\right] \\
&\quad + \frac{ab}{(a+b)(a+b+1)}\left[\psi_0(a+b+2) - \psi_0(b+1)\right].
\end{aligned}$$

B Proof of Theorem 1

In this Appendix we give a proof for Theorem 1.

Proof. PYM is given by

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

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

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

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 $N-K \geq 1$, $H_{(\alpha,d)} p(\mathbf{x}|d,\alpha)$ is integrable in d.

For the integral over alpha, it suffices to choose $\alpha_0 \gg N$ and consider the tail, $\int_{\alpha_0}^\infty H_{(d,\alpha)} p(\mathbf{x}|d,\alpha) p(d,\alpha) \,\mathrm{d}\alpha$. From eq. 19 and asymptotic expansion $\psi(x) = \log(x) - \frac{1}{2x} - \frac{1}{12x^2} + O(\frac{1}{x^4})$ 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 α that depends on K, N, and d. Thus, we have

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

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

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 Δ_K, hence Dirichlet distribution is a (continuous) probability measure over the simplex. The mean of $\mathrm{Dir}(\vec{\alpha})$ is $\frac{\vec{\alpha}}{\sum_i \alpha_i}$.

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

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

Then,

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

Lemma 3 (Decimative property)**.** *If,*

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

Then,

$$(\pi_1\tau, \pi_1(1-\tau), \pi_2, \ldots, \pi_K) = \mathrm{Dir}(\alpha_1\beta, \alpha_1(1-\beta), \alpha_2, \ldots, \alpha_K)$$

Lemma 4 (Neuturality property)**.** *If,*

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

Then,

$$\pi_1 \perp \left(\frac{\pi_2}{1-\pi_1}, \frac{\pi_3}{1-\pi_1}, \ldots, \frac{\pi_K}{1-\pi_1}\right),$$

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 multinomial probability of observing n_i number of samples corresponding to π_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 $\mathrm{Dir}(\alpha_1, \ldots, \alpha_K)$ is $\mathrm{Dir}(\alpha_1 + n_1, \ldots, \alpha_K + n_K)$.*

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

$$\mathbb{E}[H(\boldsymbol{\pi})|\vec{\alpha}] = \psi_0(A+1) - \sum_{i=1}^{\mathcal{A}} \frac{\alpha_i}{A} \psi_0(\alpha_i + 1) \tag{21}$$

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

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

□

D.2 Variance of entropy of finite Dirichlet distribution

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

Proposition 3. *For $\boldsymbol{\pi} \sim \mathrm{Dir}(\alpha_1, \alpha_2, \ldots, \alpha_{\mathcal{A}})$, such that $\sum_{i=1}^{\mathcal{A}} \alpha_i = A$, and letting $\vec{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_{\mathcal{A}})$, we have*

$$\begin{aligned}
\mathbb{E}[H^2(\boldsymbol{\pi})|\vec{\alpha}] = &\sum_{i \neq k} \left[\frac{\alpha_i \alpha_k}{(A+1)(A)}\right] \left[(\psi_0(\alpha_k + 1) - \psi_0(A + 2))\,(\psi_0(\alpha_i + 1) - \psi_0(A + 2)) - \psi_1(A + 2)\right] \\
&+ \sum_i \left[\frac{\alpha_i(\alpha_i + 1)}{(A+1)(A)}\right] \left[(\psi_0(\alpha_i + 2) - \psi_0(A + 2))^2 + \psi_1(\alpha_i + 2) - \psi_1(A + 2)\right]
\end{aligned} \tag{22}$$

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, $c = \frac{\prod_j \Gamma(\alpha_j)}{\Gamma(A)}$ we have,

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

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

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

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

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

D.3 Prior entropy mean and variance under $\mathcal{PY}$

We derive the prior entropy mean and variance of a $\mathcal{PY}$ 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 [7]), which allows us to compute expectations over $\mathcal{PY}$ using the first size biased sample, $\tilde{\pi}_1$, with the identity $\mathbb{E}\left[\sum_{i=1}^{\infty} f(\pi_i)\middle|\alpha\right] = \int_0^1 \frac{f(\tilde{\pi}_1)}{\tilde{\pi}_1} p(\tilde{\pi}_1|\alpha) d\tilde{\pi}_1$.

Proof of Proposition 1. First we validate eq. 7. 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

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

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

A similar method validates eq. 8. We will need the second size-biased sample in addition to the first. We begin with the sum on the left side of eq. 8,

$$\begin{aligned}\sum_i \sum_{j\neq i} \pi_i \pi_j g(\pi_i, \pi_j) &= \frac{\sum_i \sum_{j\neq i} \pi_i \pi_j g(\pi_i, \pi_j)}{p(\tilde{\pi}_1 = \pi_i, \tilde{\pi}_2 = \pi_j)} p(\tilde{\pi}_1 = \pi_i, \tilde{\pi}_2 = \pi_j) \\ &= \sum_i \sum_{j\neq i} g(\pi_i, \pi_j)(1-\pi_i) p(\tilde{\pi}_1 = \pi_i, \tilde{\pi}_2 = \pi_j) \\ &= \mathbb{E}_{\tilde{\pi}_1, \tilde{\pi}_2}\left[g(\tilde{\pi}_1, \tilde{\pi}_2)(1-\pi_1)|\boldsymbol{\pi}\right]\end{aligned}$$

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

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

□

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

$$\tilde{\pi}_1 \sim \text{Beta}(1-d, \alpha+d) \tag{23}$$

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

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

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,

$$\tilde{\pi}_2 = (1-\tilde{\pi}_1)v_2, \quad v_2 \sim \text{Beta}(1-d, \alpha+2d) \tag{24}$$

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

$$\begin{aligned}\mathbb{E}[(H(\boldsymbol{\pi}))^2 | d,\alpha] &= \mathbb{E}\left[\left(-\sum_i \pi_i \log(\pi_i)\right)^2 \middle| d,\alpha\right] \\ &= \mathbb{E}\left[\sum_i (\pi_i \log(\pi_i))^2 \middle| d,\alpha\right] + \mathbb{E}\left[\sum_i \sum_{j\neq i} \pi_i \pi_j \log(\pi_i)\log(\pi_j) \middle| d,\alpha\right]\end{aligned} \tag{25}$$

The first term follows directly from eq. 7,

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

The second term of eq. 25, 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,

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

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

$$\begin{aligned}\operatorname{var}[H(\boldsymbol{\pi})|d,\alpha] &= \frac{1-d}{1+\alpha}\left[(\psi_0(2-d)-\psi_0(2+\alpha))^2+\psi_1(2-d)-\psi_1(2+\alpha)\right] \\ &\quad + \frac{\alpha+d}{1+\alpha}\left[(\psi_0(1-d)-\psi_0(2+\alpha))^2-\psi_1(2+\alpha)\right] \\ &\quad - (\psi_0(1+\alpha)-\psi_0(1-d))^2 \\ &= \frac{\alpha+d}{(1+\alpha)^2(1-d)}+\frac{1-d}{1+\alpha}\psi_1(2-d)-\psi_1(2+\alpha) \end{aligned} \tag{26}$$

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 $\Delta_{\mathcal{A}}$.

D.4 Posterior Moments of PY

First, we discuss the form of the $\mathcal{PY}$ posterior, and introduce independence properties that will be important in our derivation of the mean. We recall that the $\mathcal{PY}$ posterior, $\boldsymbol{\pi}_{\text{post}}$, of eq. 9 has three stochastically independent components: Bernoulli p_*, $\mathcal{PY}$ $\boldsymbol{\pi}$, and Dirichlet $\mathbf{p}$.

Component expectations: From the above derivations for expectations under the $\mathcal{PY}$ and Dirichlet distributions as well as the Beta integral identities of Appendix A, 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 $\mathbb{E}_{p_*}\left[H(p_*)\right]$.

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

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, eq. 15.

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

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 eq. 15,

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

Derivation of posterior variance: We continue the notation from the subsection above. In order to exploit the independence properties of π_{post} we first apply the law of total variance to obtain eq. 27,

$$\text{var}[H(\pi_{\text{post}})|d,\alpha] = \underset{p_*}{\text{var}}\left[\mathbb{E}_{\boldsymbol{\pi},\mathbf{p}}[H(\pi_{\text{post}})]\Big|d,\alpha\right] + \mathbb{E}_{p_*}\left[\underset{\boldsymbol{\pi},\mathbf{p}}{\text{var}}[H(\pi_{\text{post}})]\Big|d,\alpha\right] \tag{27}$$

We now seek expressions for each term in eq. 27 in terms of the expectations already derived.

Step 1: For the right-hand term of eq. 27, we use the independence properties of π_{post} to express the variance in terms of PY, Dirichlet, and Beta variances,

$$\begin{aligned}\mathbb{E}_{p_*}\left[\underset{\boldsymbol{\pi},\mathbf{p}}{\text{var}}[H(\pi_{\text{post}})|p_*]\Big|d,\alpha\right] &= \mathbb{E}_{p_*}\left[(1-p_*)^2\underset{\mathbf{p}}{\text{var}}[H(\mathbf{p})] + p_*^2\underset{\boldsymbol{\pi}}{\text{var}}[H(\boldsymbol{\pi})]\Big|d,\alpha\right] \\ &= \frac{(N-Kd)(N-Kd+1)}{(\alpha+N)(\alpha+N+1)}\underset{\mathbf{p}}{\text{var}}[H(\mathbf{p})] \\ &\quad + \frac{(\alpha+Kd)(\alpha+Kd+1)}{(\alpha+N)(\alpha+N+1)}\underset{\boldsymbol{\pi}}{\text{var}}[H(\boldsymbol{\pi})]\end{aligned} \tag{28}$$

Step 2: In the left-hand term of eq. 27 the variance is with respect to the Beta distribution, while the inner expectation is precisely the posterior mean we derived above. Expanding, we obtain,

$$\underset{p_*}{\text{var}}\left[\mathbb{E}_{\boldsymbol{\pi},\mathbf{p}}[H(\pi_{\text{post}})|p_*]\Big|d,\alpha\right] = \underset{p_*}{\text{var}}\left[(1-p_*)\mathbb{E}_{\mathbf{p}}\left[H(\mathbf{p})\right] + p_*\mathbb{E}_{\boldsymbol{\pi}}\left[H(\boldsymbol{\pi})|p_*\right] + h(p_*)\Big|d,\alpha\right] \tag{29}$$

To evaluate this integral, we introduce some new notation,

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

so that

$$\Omega^2(p_*) = 2p_*h(p_*)[\mathcal{B}-\mathcal{A}] + 2\mathcal{A}h(p_*) + h^2(p_*) + p_*^2[\mathcal{B}^2 - 2\mathcal{A}\mathcal{B}] + 2p_*\mathcal{A}\mathcal{B} + (1-p_*)^2\mathcal{A}^2 \tag{30}$$

and we note that

$$\underset{p_*}{\text{var}}\left[\mathbb{E}_{\boldsymbol{\pi},\mathbf{p}}[H(\pi_{\text{post}})|p_*]\Big|d,\alpha\right] = \mathbb{E}_{p_*}[\Omega^2(p_*)] - \mathbb{E}_{p_*}[\Omega(p_*)]^2 \tag{31}$$

In Appendix A we derive expressions for the components composing $\mathbb{E}_{p_*}[\Omega(p_*)]$, as well as each term of eq. 30. Although less elegant than the posterior mean, the expressions derived above permit us to compute eq. 27 numerically from its component expectations, without sampling.