Low probability states, data statistics, and entropy estimation

Damián G. Hernández [1,4,*], Ahmed Roman [1,*], and Ilya Nemenman[1,2,3]
[1] *Physics Department, Emory University, Atlanta, Georgia, USA*
[2] *Biology Department, Emory University, Atlanta, Georgia, USA*
[3] *Initiative for Theory and Modeling of Living Systems, Emory University, Atlanta, Georgia, USA*
[4] *Department of Medical Physics, Centro Atómico Bariloche and Instituto Balseiro, 8400 San Carlos de Bariloche, Argentina*
[*] *D.G.H. and A.R. contributed equally to this work.*

(Dated: July 5, 2022)

A fundamental problem in analysis of complex systems is getting a reliable estimate of entropy of their probability distributions over the state space. This is difficult because unsampled states can contribute substantially to the entropy, while they do not contribute to the Maximum Likelihood estimator of entropy, which replaces probabilities by the observed frequencies. Bayesian estimators overcome this obstacle by introducing a model of the low-probability tail of the probability distribution. Which statistical features of the observed data determine the model of the tail, and hence the output of such estimators, remains unclear. Here we show that well-known entropy estimators for probability distributions on discrete state spaces model the structure of the low probability tail based largely on few statistics of the data: the sample size, the Maximum Likelihood estimate, the number of coincidences among the samples, the dispersion of the coincidences. We derive approximate analytical entropy estimators for undersampled distributions based on these statistics, and we use the results to propose an intuitive understanding of how the Bayesian entropy estimators work.

I. INTRODUCTION

Estimating entropy – that is, the measure of uncertainty [1, 2] – of a random variable from its samples is often a key question in analysis of complex systems. This estimation from a finite (and often small) set of samples is a hard problem, especially for high dimensional systems, where the number of states that a variable can take quickly overwhelms the number of samples N. Then many of the states, hereafter called *low probability states*, have probability $< 1/N$. Collectively, we refer to all of these states as the *tail* of the probability distribution. While there may be a lot of samples in the tail, each low probability state will not be sampled typically, or will be sampled at most once. Because of the tail, the entropy estimator that replaces probabilities of states by their empirical frequencies (the so called *naive* or *Maximum Likelihood* estimator [3]) has a large sample size dependent bias [4]. Corrections have been derived to overcome this bias [5–7], but these tend to be valid only in the well-sampled regime. Outside of this regime, Bayesian [8–10] and some non-parametric [11–13] estimators may still result in low bias estimates by imposing *a priori* assumptions on the probabilities of the low-probability states.

Although these Bayesian and non-parametric estimators perform well on some data sets, it is known that no estimator can be universally unbiased in this regime [4, 14]. Thus it is crucial to understand how these estimators extract information about entropy from data, and hence when they will fail. Unfortunately, such theoretical understanding is missing for many estimators. Ma was the first to point out that estimation of entropy is possible for poorly-sampled uniform distributions by analysing a particular statistics of the data: *coincidences* [15]. Nemenman extended the theoretical idea that coincidences determine entropy to non-uniform distributions obeying some Bayesian priors [16]. However, a similar theoretical understanding is still missing in a broader context, and it remains unclear which statistics of data, in addition to the number of coincidences, may contribute to entropy estimation and why.

In this paper, we analytically investigate two Bayesian estimators: that of Nemenman, Shafee and Bialek [9, 17] and of Archer and Pillow [10]. We focus on the regime, which is arguably the most important for real life applications, where the number of states with at least one sample, K_1, is similar to the total number of samples, $K_1 \sim N \gg 1$, and yet $K_1 < N$, so that there are coincidences in the data. Outside of this regime, the probability distribution is either well-sampled (so that many different methods for entropy estimation would work), or there are no coincidences at all (so that entropy estimation is impossible). In our regime of interest, we show that the result of the estimation by the studied estimators depends on the Maximum Likelihood entropy estimate S_0, the number of coincidences, and also on two measures of *dispersion of coincidences*. The first of these, K_2, is the number of states with at least two samples. The second, which we call Q_1, characterizes the spread of coincidences over states with three or more samples.

We show that values of these statistics are related to the structure of the tails of the probability distribution that is assumed by the estimators. Specifically, a short, exponential, tail is more likely to be inferred by the estimators when there many coincidences or they are dispersed. If the number of coincidences is intermediate, and the coincidences are concentrated, then the estimators infer a long tail. In between these two regions, a mixed tail dominates. We show that the studied estimators correct Maximum Likelihood, and that the correction is larger when there are fewer coincidences and they are concentrated, which in turn happens with a large exponential tail or a slowly-decaying long tail. This understanding relates the observable data statistics to assumptions that Bayesian estimators make about the underlying probability distributions (see Fig. 1), and hence provides an intuitive explanation for how these estimators work and, crucially, when they fail.

II. OVERVIEW OF BAYESIAN ENTROPY ESTIMATION

Given a probability distribution $\{q_x\} = \boldsymbol{q}$ for a discrete one-dimensional random variable X, its entropy is defined as [1]

$$S(\boldsymbol{q}) = -\sum_x q_x \log q_x. \tag{1}$$

Note that we use the natural logarithm throughout this paper, and hence entropy is measured in *nats*. One is often faced with a problem when S must be estimated for unknown q_x from a set of N samples $\{x_1, \ldots, x_N\}$ from the probability distribution. The Maximum Likelihood estimator of entropy, S_0, is then defined by replacing the probabilities with frequencies $q_x \to \hat{q}_x = n_x/N$,

$$S_0 = S(\hat{\boldsymbol{q}}) = -\sum_x \frac{n_x}{N} \log \frac{n_x}{N}. \tag{2}$$

States with zero frequencies in the sample do not contribute to S_0 resulting typically in underestimation of the entropy [4]. In general, because of this low probability tail, estimation of entropy from data is very hard when the number of samples is smaller than the number of effective states of the variable, $N \ll \exp(S)$.

Bayesian estimators address the problem by imposing various *a priori* assumptions $p(\boldsymbol{q})$. One then uses Bayes theorem to infer the *a posteriori* distribution of $\boldsymbol{q}$, and finally integrates over $\boldsymbol{q}$ to get the *a posteriori* distribution or moments of entropy. Specifically, the mean posterior entropy $\hat{S} = \langle S|\boldsymbol{n}\rangle$ given the counts $\boldsymbol{n} = \{n_x\}$ of how many times state x was sampled is given by

$$\begin{aligned}\hat{S} = \langle S|\boldsymbol{n}\rangle &= \int S(\boldsymbol{q}) p(S|\boldsymbol{q}) p(\boldsymbol{q}|\boldsymbol{n}) d\boldsymbol{q} \\ &= \int S(\boldsymbol{q}) \delta\left(S + \sum_x q_x \log q_x\right) p(\boldsymbol{q}|\boldsymbol{n}) d\boldsymbol{q},\end{aligned} \tag{3}$$

where $p(\boldsymbol{q}|\boldsymbol{n})$ is the posterior over $\boldsymbol{q}$ under some prior $p(\boldsymbol{q})$,

$$p(\boldsymbol{q}|\boldsymbol{n}) = \frac{p(\boldsymbol{n}|\boldsymbol{q})p(\boldsymbol{q})}{p(\boldsymbol{n})} = \frac{\prod_x q_x^{n_x} p(\boldsymbol{q})}{p(\boldsymbol{n})}. \tag{4}$$

For distributions with known finite size $\mathcal{A}$ of the space of the possible outcomes (aka the *alphabet size*), the Dirichlet distribution is often chosen as a prior due to its conjugacy with the categorical distribution:

$$p(\boldsymbol{q}) = \mathrm{Dirichlet}(\boldsymbol{q}|\lambda) \propto \prod_{i=1}^{\mathcal{A}} q_i^{\lambda}, \tag{5}$$

where λ is known as the concentration parameter.

Note that any chosen prior $p(\boldsymbol{q})$ implicitly imposes assumptions on the structure of the low probability tail (and hence its contribution to the entropy) based on the observed statistics of the well-sampled part of the probability distribution. However, these implicit assumptions usually are not made explicit, and they remain mysterious even for most commonly used Bayesian estimators. Lifting this veil is the goal of this work.

A. The Nemenman-Shafee-Bialek (NSB) Estimator

Nemenman et al. [9] showed that, for variables with the finite alphabet size $\mathcal{A}$, Dirichlet priors on $\boldsymbol{q}$ with a fixed value for the concentration parameter λ correspond to highly concentrated *a priori* distribution on entropy, which persists for large sample sizes. This bias induces incorrect entropy estimates, which nonetheless have low variance and hence are certain about their outputs. To address this issue, Ref. [9] suggested a Dirichlet-mixture prior

$$p_{\mathrm{NSB}}(\boldsymbol{q}) = \int \mathrm{Dirichlet}(q|\lambda) p_{\mathrm{prior}}(\lambda) d\lambda, \tag{6}$$

where $p(\lambda)$ are the mixture weights determined by

$$p_{\mathrm{prior}}(\lambda) \propto \partial_\lambda \langle S|\lambda\rangle = \mathcal{A}\psi_1(\mathcal{A}\lambda + 1) - \psi_1(\lambda + 1), \tag{7}$$

and where $\langle S|\lambda\rangle$ is the *a priori* expected entropy under the Dirichlet$(\boldsymbol{q}|\lambda)$ prior, and $\psi_1(\cdot)$ is the tri-gamma function [18]. This choice of weights implies a nearly uniform *a priori* distribution for the entropy S on the interval $[0, \log \mathcal{A}]$. The resulting entropy estimate is then

$$\begin{aligned}\hat{S}_{\mathrm{NSB}} = \langle S|\boldsymbol{n}\rangle &= \int\int S(\boldsymbol{q}) p(\boldsymbol{q}|\boldsymbol{n}, \lambda) p(\lambda|\boldsymbol{n}) d\boldsymbol{q} d\lambda \\ &= \int \langle S|\boldsymbol{n}, \lambda\rangle \frac{p(\boldsymbol{n}|\lambda) p_{\mathrm{prior}}(\lambda)}{p(\boldsymbol{n})} d\lambda.\end{aligned} \tag{8}$$

Here $\langle S|\boldsymbol{n}, \lambda\rangle$ is the posterior mean entropy under the prior Dirichlet$(\boldsymbol{q}|\lambda)$, and $p(\boldsymbol{n}|\lambda)$ is the evidence (which has a Polya distribution) [19],

$$\begin{aligned}p(\boldsymbol{n}|\lambda) &= \int p(\boldsymbol{n}|\boldsymbol{q}) p(\boldsymbol{q}|\lambda) d\boldsymbol{q} \\ &= \frac{N!\Gamma(\mathcal{A}\lambda)}{\Gamma(\lambda)^{\mathcal{A}}\Gamma(N + \mathcal{A}\lambda)} \prod_{i=1}^{\mathcal{A}} \frac{\Gamma(n_i + \lambda)}{n_i!}\end{aligned} \tag{9}$$

where $\Gamma(\cdot)$ is the gamma function [18]. Using the analytical expressions for the first two moments of posterior mean entropy $\langle S|\boldsymbol{n}, \lambda\rangle$ (available from Refs. [8, 9]), one then uses one-dimensional numerical integration over λ to obtain $\hat{S}_{\mathrm{NSB}}$.

B. The Dirichlet and the Pitman-Yor Processes

When the size of the state space is unknown or infinite, the standard NSB construction does not work. Then

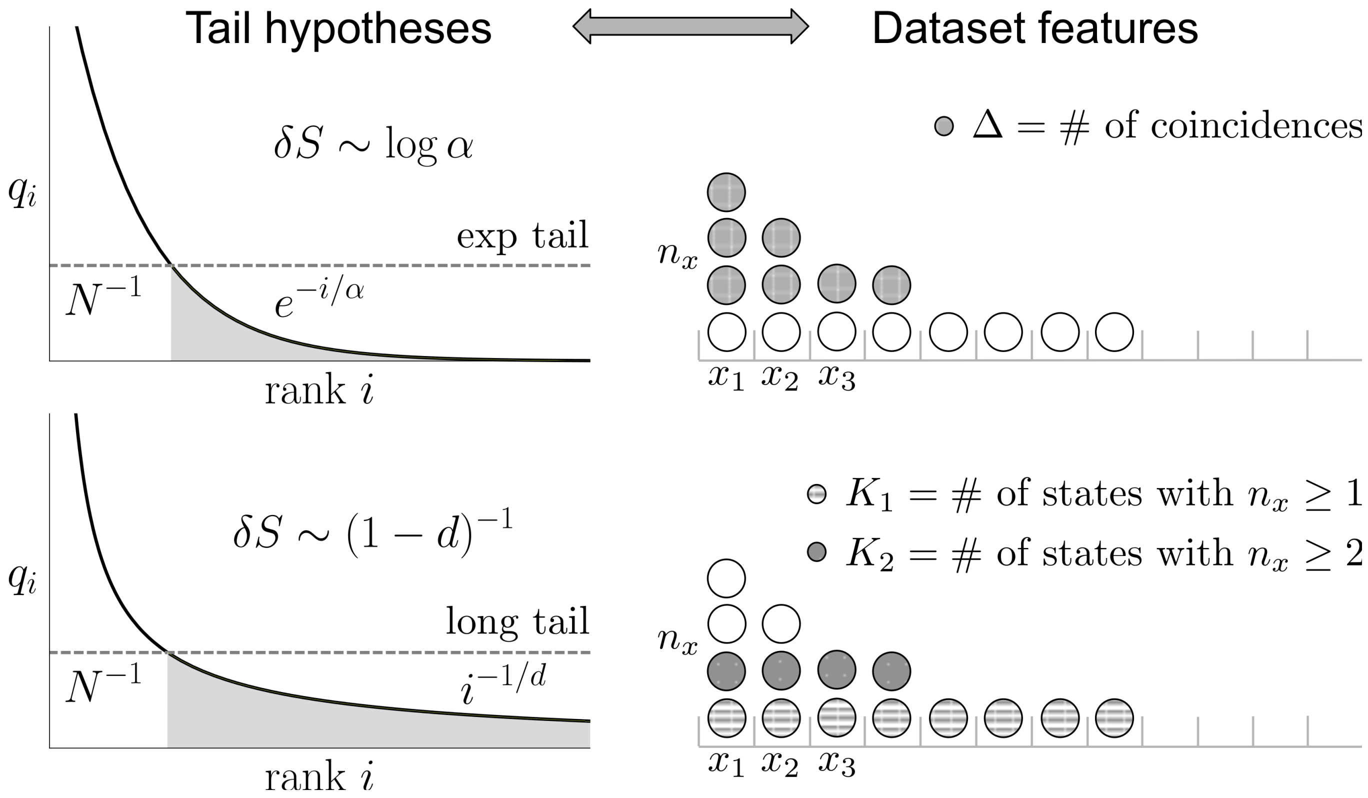

FIG. 1. Relation between assumptions about the tail structure and the statistics that determine entropy estimation. The set of unsampled states, $q_i \leq 1/N$, which we refer to as the *tail*, may contribute substantially to the entropy. However, the Maximum Likelihood estimation overlooks this contribution. If the rank ordered plot of the tail is exponential with the scale α (top panel), then the tail has effectively α states, which contribute $\delta S \sim \log\alpha$ to the entropy. While the tail cannot be observed directly, it pulls samples from the head of the distribution, so that the number of coincidences, Δ, in the head decreases as α grows. Thus one can estimate α and hence the entropy itself from Δ. Alternatively, if the rank-ordered plot of the tail has a power law structure with the exponent $-1/d$, then the tail does not have a finite effective size (bottom panels). Then its contribution to entropy depends on d as $\delta S \sim (1-d)^{-1}$. In this case, one can estimate d, and hence the entropy, from the dispersion of the coincidences, which depends, in part, on how many samples happen once or more, K_1, or twice or more, K_2, in the dataset.

one commonly uses one of the following two stochastic processes to construct a prior $p(\boldsymbol{q})$ over a countably infinite state space: the Pitman-Yor Process (PYP) [20] and its special case, the Dirichlet Process (DP) [21]. To specify these processes, one requires two inputs: a parameter vector and a base distribution. Parameters of the Pitman-Yor process are known as the discount parameter d, $0 \leq d < 1$, and the concentration parameter α. The parameters control the shape of typical distributions generated by the process. Specifically, d controls the structure of the low probability tail of $\boldsymbol{q}$, so that the tail typically decays as $q_x \propto x^{-1/d}$. The concentration parameter α control the probability mass near the head of the distribution. In the limit $d \to 0$, $\mathrm{PYP}(d,\alpha)$ becomes the Dirichlet Process, $\mathrm{DP}(\alpha)$. In other words, the Dirichlet Process generates distributions with short tails.

When the base distribution is the Beta distribution, one draws samples $q_x \sim \mathrm{PYP}(d,\alpha)$ via the so called *stick-breaking process* [22], which uses an infinite sequence of independent Beta-distributed random variables $\beta_x \sim \mathrm{Beta}(1-d, \alpha + xd)$, so that

$$\tilde{q}_x = \beta_x \prod_{y=1}^{x-1}(1-\beta_y). \tag{10}$$

Thus obtained $\tilde{\boldsymbol{q}}$ are not strictly decreasing with x, and so one obtains a strictly non-increasing distribution $\boldsymbol{q}$ from them by rank ordering.

C. Expectations over DP and PYP Posteriors

Previous studies [23] showed that PYP priors (for multinomial observations) yield a posterior $p(\boldsymbol{q}|\boldsymbol{n},\alpha,d)$, which consists of two parts: probability of K_1 states that exist in the sample with the counts of, at least, one, and probability of states that are not sampled. We will denote the set of states with nonzero counts as $\mathbb{K}$, and its cardinality is $K_1 = ||\mathbb{K}||$. Then the first term of the posterior is given by the Dirichlet distribution, $p(\boldsymbol{q} \in \mathbb{K}|\boldsymbol{\mu}) \propto \prod_x q_x^{\mu_x}$, where $\boldsymbol{\mu}$ is a concentration vector

$\boldsymbol{\mu} = (n_1 - d, \cdots, n_{K_1} - d, \alpha + K_1 d)$. This leaves the probability of $q_* = 1 - \sum_{x \in \mathbb{K}} q_x$ for the unobserved states. In other words, the states with nonzero counts contribute the following to the posterior:

$$\begin{aligned} p(\boldsymbol{q} \in \mathbb{K}|\boldsymbol{n}) &= p(q_1, \cdots, q_{K_1}, q_*|\boldsymbol{n}) \\ &= \mathrm{Dirichlet}(n_1 - d, \cdots, n_{K_1} - d, \alpha + K_1 d) \\ &\propto q_*^{\alpha + K_1 d} \prod_{i=1}^{K_1} q_i^{n_i - d}. \end{aligned} \tag{11}$$

For the states that have no samples, the posterior is equal to the prior. Thus their contribution to the posterior is the Pitman-Yor Process, normalized by their total probability being q^*:

$$p(\boldsymbol{q} \notin \mathbb{K}) = p(q_{K_1+1}, q_{K_1+2}, \cdots) = q^* \mathrm{PYP}(d, \alpha + K_1 d). \tag{12}$$

Overall, this yields a closed form solution for the posterior mean and variance of the entropy S. Specifically, the resulting posterior mean $\langle S|\boldsymbol{n}, \alpha, d\rangle$ is

$$\begin{aligned} \langle S|\boldsymbol{n}, \alpha, d\rangle &= \psi(\alpha + N + 1) - \frac{\alpha + K_1 d}{\alpha + N}\psi(1 - d) \\ &- \frac{1}{\alpha + N}\left(\sum_{x=1}^{K_1}(n_x - d)\psi(n_x - d + 1)\right), \end{aligned} \tag{13}$$

where $\psi(x) = \partial_x \log \Gamma(x)$ is the di-gamma function [18]. Unfortunately, this is usually not a good estimate of entropy since, for fixed α and d, the prior $\mathrm{PYP}(d, \alpha)$ on $\boldsymbol{q}$ corresponds to a highly concentrated *a priori* distribution on entropy, just like was noted before in the context of the NSB estimator. To counter this, Archer and Pillow [10] followed the NSB prescription and introduced a prior (mixture) over the parameters of $PYP(d, \alpha)$, $p_{\mathrm{prior}}(\alpha, d)$, which uniformized the induced prior over entropy (with the caveat that, for a distribution on a countable alphabet, the entropy may be infinite, and hence strict uniform distribution over entropy is impossible). Specifically, they used

$$p_{\mathrm{prior}}(\alpha, d) = p(\gamma) = e^{-10/(1-\gamma)}, \qquad \text{where} \tag{14}$$

$$\gamma = (\psi(1) - \psi(1 - d))/(\psi(\alpha + 1) - \psi(1 - d)), \tag{15}$$

and then they confirmed numerically that this choice of the prior leads to good estimates of entropy for various test data sets. In other words, they proposed a new estimate of entropy, the Pitman-Yor Mixture (PYM):

$$\begin{aligned} \hat{S}_{PYM} = \langle S|\boldsymbol{n}\rangle &= \int \langle S|\boldsymbol{n}, \alpha, d\rangle p_{\mathrm{posterior}}(\alpha, d|\boldsymbol{n}) d(\alpha, d) \\ &= \int \langle S|\boldsymbol{n}, \alpha, d\rangle \frac{p(\boldsymbol{n}|\alpha, d) p_{\mathrm{prior}}(\alpha, d)}{p(\boldsymbol{n})} d(\alpha, d), \end{aligned} \tag{16}$$

where $\langle S|\boldsymbol{n}, \alpha, d\rangle$ is given in Eq. (13). The evidence $p(\boldsymbol{n}|\alpha, d)$ is then given by (see Ref. [10] for a detailed derivation)

$$p(\boldsymbol{n}|\alpha, d) = \frac{\Gamma(1 + \alpha) \prod_{l=1}^{K_1}(\alpha + l d) \prod_{x=1}^{K_1} \Gamma(n_x - d)}{\Gamma(1 - d)^{K_1} \Gamma(\alpha + N)}. \tag{17}$$

Note that taking $d \to 0$ in Eqs. (16 and 17) and making the identification $\alpha = \mathcal{A}\lambda$ in the limits $\lambda \to 0$ and $\mathcal{A} \to \infty$ such that α is finite, result in a countably-infinite analogue of the NSB estimator.

III. DETERMINING DATA STATISTICS THAT DEFINE ENTROPY ESTIMATES

In the section, we approximate the likelihood function of the Pitman-Yor process, Eq. (17), analytically in terms of coincidence-based data statistics. We then numerically show that the resulting analytical entropy estimates are close to the exact Pitman-Yor Mixture estimator. We focus on the regime where the Maximum Likelihood entropy estimator fails dramatically. For this, we study random variables with many accessible states in the regime where the number of unique samples, K_1, is of the order of the total sample size N. This regime corresponds to $K_1 \lesssim N \leq \exp(S)$, where N is the number of samples and S is the true entropy.

We start by considering the log-likelihood function, which is the logarithm of the evidence $p(\boldsymbol{n}|\alpha, d)$ in Eq. (17):

$$\begin{aligned} \mathcal{L}(\mathbf{n}|\alpha, d) &= \log\Gamma(1+\alpha) - \log\Gamma(N+\alpha) + \log\Gamma\left(\frac{\alpha}{d} + K_1\right) \\ &- \log\Gamma\left(\frac{\alpha}{d} + 1\right) + \sum_{i=1}^{K_1} \log\Gamma(n_i - d) - K_1 \log\Gamma(1 - d). \end{aligned} \tag{18}$$

We now define K_m as the number of states with at least m counts in the total sample of size N, $K_m = \sum_{n_i \geq m} 1$. We denote by m_f the largest occupancy of any state in the sample. Further, we define $\mathcal{K}$ as the vector, whose mth element is K_m. We note that, for any function $f(n)$,

$$\sum_i f(n_i) = \sum_m (K_m - K_{m+1}) f(m). \tag{19}$$

Thus, in particular, the log-likelihood $\mathcal{L}(\mathbf{n}|\alpha, d)$ can be viewed as $\mathcal{L}(\mathcal{K}|\alpha, d)$. With this, we can expand Eq. (18) around $d = 0$ to get (see Appendix VI A for details):

$$\begin{aligned} \mathcal{L}(\mathbf{n}|\alpha, d) &\approx \mathcal{L}_a(\mathcal{K}|\alpha, d) \equiv \log\Gamma(1 + \alpha) \\ &- \log\Gamma(N + \alpha) + \log\Gamma\left(\frac{\alpha}{d} + K_1\right) - \log\Gamma\left(\frac{\alpha}{d} + 1\right) \\ &+ (K_1 - 1)\log d + K_2 \log(1 - d) - Q_1 d + \mathcal{O}(d^2), \end{aligned} \tag{20}$$

where

$$Q_1 = \sum_{m=3}^{m_f} \frac{K_m}{m - 1}, \tag{21}$$

and the subscript a denotes the $d \to 0$ asymptotic nature of the expression.

By rewriting the Maximum Likelihood estimate S_0 of Eq. (2) in terms of coincidences (see Appendix VI B),

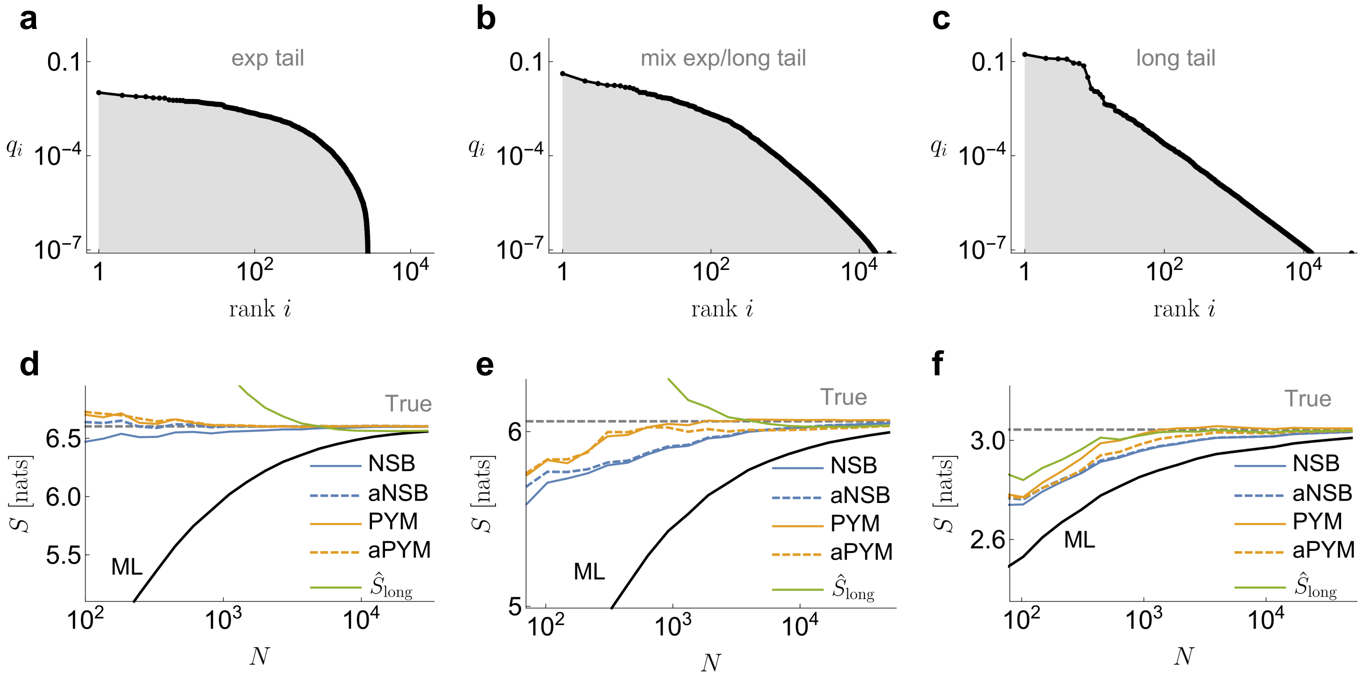

FIG. 2. Comparison between PYM and related estimators and their approximations for distributions with different tails. The upper panels (**a**-**c**) show the distributions, whose entropy is being estimated. The lower panels (**d**-**f**) show the corresponding entropy estimates as a function of the number of samples, averaged over ten sets of samples. The full estimators, PYM and NSB (with a large alphabet size $\mathcal{A} = 20K_1$), almost overlap with our approximations, aPYM and aNSB. In all panels, we show results for Maximum Likelihood (black), NSB (blue), aNSB (dashed blue), PYM (orange), aPYM (dashed orange), and $\hat{S}_{\text{long}}$ (green) estimators. The dashed gray line represents the true value of entropy for each of the studied distributions.

using the identity Eq. (19), and approximating certain terms that are finite in the limit $d \to 1$ via a Taylor expansion around $d \ll 1$, the mean posterior entropy, Eq. (13), results in (see Appendix VI C):

$$\begin{aligned}\langle S|\mathbf{n},\alpha,d\rangle &\approx \langle S|\mathcal{K},\alpha,d\rangle_a \equiv \psi(N+\alpha+1)\\ &-\left(\frac{\alpha+K_1}{\alpha+N}\right)\psi(1-d)+\frac{1}{\alpha+N}\Bigg[.N(S_0-\log N)-K_1\\ &\quad+K_2(\log 4-1-\psi(2-d))+Q_1 d\\ &\quad+\mathcal{O}\left(d^2,\sum_{m=3}\frac{K_m}{(m-1)^2}\right)\Bigg], \qquad (22)\end{aligned}$$

where $\mathcal{O}(d^2, \sum_{m=3} K_m/m^2)$ means that we kept terms that are at most linear in d and at most proportional to $\sum_{m=3}\frac{K_m}{(m-1)}$. Interestingly, within this approximation, the log-likelihood and the posterior mean entropy depend on the sample size N, the Maximum Likelihood entropy estimate S_0, and the three characteristics of the coincidence vector: K_1, K_2 and Q_1.

The final step in approximating the estimator $\hat{S}_{PYM}$, Eq. (16), is to integrate the expected entropy for fixed hyper-parameters $\langle S|\mathcal{K},\alpha,d\rangle_a$ over the posterior $p_{\text{posterior}}(\alpha,d|\mathbf{n}) \propto p(\mathbf{n}|\alpha,d)p_{\text{prior}}(\alpha,d)$ to form the Pitman-Yor mixture. Then the variance of the resulting estimator is dominated by the contribution from the uncertainty in the posterior distribution of the parameters α, d, which is about 80% of the total variance in our simulations.

This procedure of replacing $\langle S|\boldsymbol{n},\alpha,d\rangle$ with the asymptotic expression $\langle S|\mathcal{K},\alpha,d\rangle_a$ in Eq. (16) leads to a new estimator of entropy, which we call *approximate PYM* estimator, or aPYM. This estimator is fully determined by just few data statistics, N, S_0, K_1, K_2, and Q_1. There are also two limiting cases of this estimator. First, by taking $d \to 0$ in Eqs. (20, 22), we define the approximate version of the NSB limit of the PYM estimator on a countably infinite number of possible outcomes, which we denote as aNSB. At the other extreme, taking $\alpha \to 0$ in Eqs. (20, 22), corresponds to a prior that favors distributions with long tails. We denote the corresponding estimator as $\hat{S}_{\text{long}}$.

The above observation that, in the undersampled regime where $\exp(S/2) < N < \exp(S)$, the PYM entropy estimator and its relatives are determined approximately by just few statistics of the data, $\{N, S_0, K_1, K_2, Q_1\}$, is the main result of our paper. To corroborate this, we explore the quality of the approximation numerically for different distributions $\boldsymbol{q}$. Figure 2 presents results for three distributions with different structures of tails, generated from the Pitman-Yor Process: a distribution with an exponential tail (Fig. 2**a**, PYP($d = 0, \alpha = 400$) = DP(400)), one with a mixed tail (Fig. 2**b**:

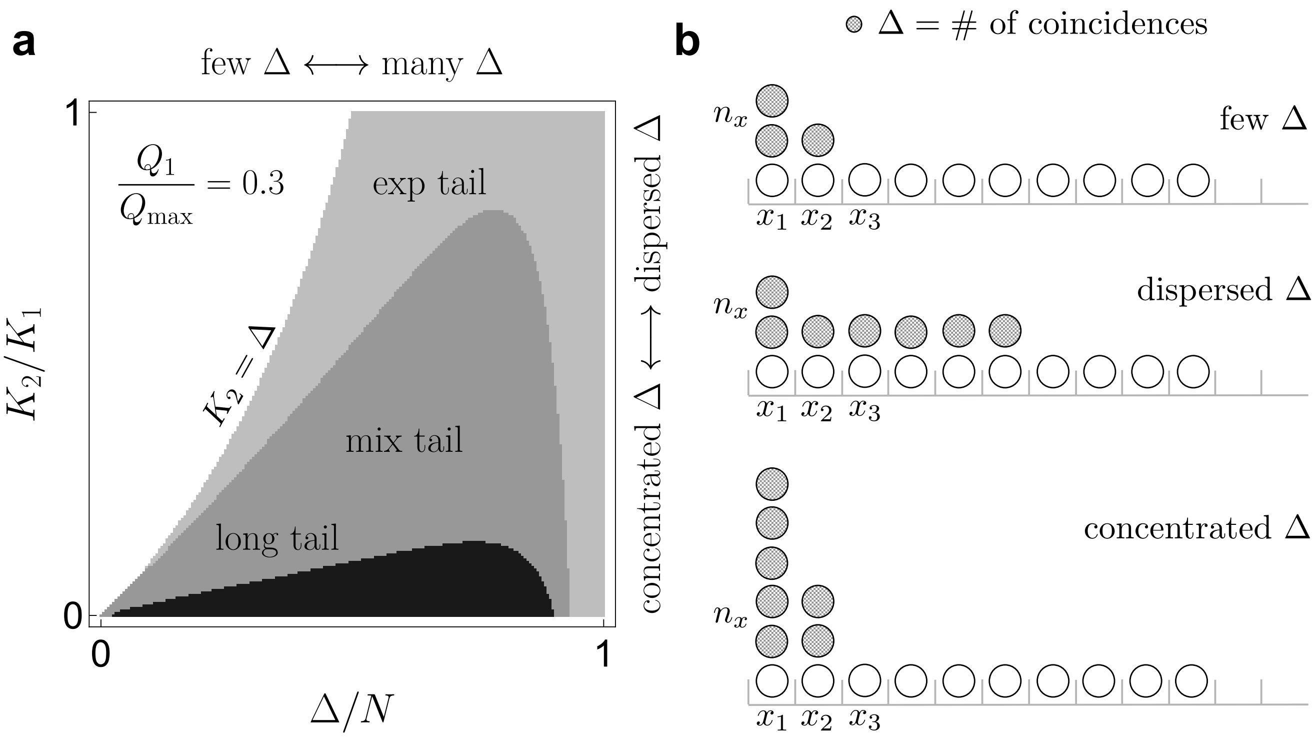

FIG. 3. **a**: Phase diagram of the dominant tail hypothesis selected by the PYM estimator as a function of various statistics of the data sample. The explored statistics are the fraction of coincidences in the sample, Δ/N, and dispersion of the coincidences, K_2/K_1. This diagram is evaluated at the third crucial data statistics set at $Q_1 = 0.3\,Q_{\max} = 0.3(\Delta - K_2)/2$. **b**: Schematic diagram that illustrates how sample sets with different Δ, K_1, and K_2 may look like. An empty or gray circle above a state x_i represent a single sample for that state. Gray circles denote coincidences.

PYP($d = 0.4, \alpha = 100$)), and one with a long tail (Fig. 2**c**: PYP($d = 0.6, \alpha = 0$)). In the lower panels we show the results of estimating entropy for different dataset sizes using the ML estimator, the PYM estimator, the NSB estimator with a large alphabet size $\mathcal{A} = 20K_1$, and the three approximations: aPYM, aNSB, and $\hat{S}_{\rm long}$. All results are averaged over ten sets of random samples. In all cases, the differences between NSB and aNSB on the one hand, and PYM and aPYM on the other are negligible, supporting the accuracy of the approximation. All four of these estimators produce high quality estimates for all sample sizes. Further, we also checked that the approximation of the posterior error of the estimators is close to that of the full versions (not shown). In contrast, $\hat{S}_{\rm long}$ only performs well when the distribution has a long tail, and the Maximum Likelihood never works well.

IV. TAIL-HYPOTHESIS AND ENTROPY ESTIMATION PHASE DIAGRAMS

The above discussion shows that the PYM estimator and its relatives work by first estimating the most likely α and d from the sampled data, and then using these estimated parameters to approximate the structure of the low probability tail (from short, to long) and hence of its contributions to the entropy. We further showed that, in the regime of interest, the log-likelihood of α and d is dominated by just few statistics: N, S_0, K_1, K_2, and Q_1. It is thus illustrative to understand, which combinations of these statistics select which hypothesis on the structure of the tail. Building the corresponding phase diagram of the selected tail structure as a function of the data statistics is the goal of this Section.

We will consider three classes of tails: exponential ($d = 0$ selected, denoted as hypothesis $H = 1$), long tail ($\alpha = 0$ selected, denoted as hypothesis $H = 2$), and a mixed tails (arbitrary α and d, denoted as $H = 3$). Our goal is then to evaluate which of the three tail hypotheses has a higher probability given the data. Long and short tail hypotheses have one parameter each, while the mixed tail hypothesis has two parameters and contains the other two hypotheses as special cases. Thus when evaluating the log-likelihoods of each of the hypotheses, we must penalize them for having a different number of parameters, which we do using Bayesian Information Criterion [24]. To do this, we evaluate the likelihoods

$$\mathcal{L}_H = \log p(\mathcal{K}|\hat{\alpha}, \hat{d}) + \log p_{\rm prior}(\hat{\alpha}, \hat{d}) - \frac{n_H}{2}\log N, \quad (23)$$

where $\hat{\alpha}$ and $\hat{d}$ are the maximum likelihood values of the parameters within each hypothesis, and n_H is the number of parameters for the hypothesis ($n_H = 2$ for $H = 3$, and $n_H = 1$ otherwise). We remind the reader that, by construction, $\hat{\alpha} = 0$ for the long tail hypothesis, $H = 2$, and $\hat{d} = 0$ for the short tailed hypothesis, $H = 1$.

We determine the regions of the N, S_0, K_1, K_2, Q_1 space, where one of the three $\mathcal{L}_H$ dominates, and plot the slice of this phase diagram in Fig. 3. Specifically, in the Figure, we vary the total number of *coincidences*, $\Delta = N - K_1$, and the number of *states with coincidences*, that is, the number of states with more than two counts, K_2. By sampling many distributions, we empirically observe that the value $Q_1 \sim 0.6(\Delta - K_2)/2$ is when the rest of the $\Delta - K_2$ counts are uniformly dispersed, and Q_1 tends to zero when the rest of the counts are concentrated in a single state. Note that the maximum value Q_1 can take is $Q_{\max} = \frac{\Delta - K_2}{2}$. For this reason, we choose the intermediate representative value $Q_1 = 0.3 Q_{\max} = 0.3 \frac{\Delta - K_2}{2}$.

To simplify the presentation, we plot the winning tail hypothesis as a function of Δ/N and K_2/K_1. Normalized in this way, the diagram is constrained to a square of size 1, as $0 \leq \Delta/N, K_2/K_1 \leq 1$. In addition, $K_2 \leq \Delta$, which means that the upper left corner is not accessible. The ratio Δ/N determines how common are the coincidences, and the ratio K_2/K_1 describes whether the coincidences in the data are concentrates in a few states, or dispersed over many states (see Figure 3b).

Figure 3**a** show that the exponential tail hypothesis dominates when there are many coincidences, $\Delta/N \sim 1$, or when the coincidences are dispersed, that is $K_2/K_1 \sim 1$ or $K_2/\Delta \sim 1$. Both cases can be explained as corresponding to distributions that are relatively uniform on some fixed number of states, and have zero probability elsewhere. A long tail only dominates when the fraction of coincidences has an intermediate value, but the coincidences are highly concentrated, $K_2/K_1 \ll 1$. In other words, in this case, there are dominant states, but a lot of samples still fall outside of them. For other values of Δ/N and K_2/K_1, the mixed tail hypothesis dominates.

Equipped with this picture of which tail hypothesis is selected by the PYM estimator as a function of data statistics, we now can calculate how the estimator corrects the ML entropy value S_0 for different data statistics. Integrating the mean posterior entropy $\langle S|\mathcal{K}, \alpha, d\rangle_a$, Eq. (22), over our approximation of the posterior, $p_a(\alpha, d|\mathcal{K})$, which we obtain by exponentiating Eq. (20), we get the approximate PYM estimator $\hat{S}_{PYM,a}$. The Maximum Likelihood estimate S_0 enters linearly in the posterior mean entropy, Eq. (22). Thus we write

$$\langle S|\mathcal{K}, \alpha, d\rangle_a = b_{\alpha,d}\, S_0 + \delta S_{\alpha,d}, \tag{24}$$

where $b_{\alpha,d}$ and $\delta S_{\alpha,d}$ can be read off from Eq. (22). Performing the integral over the approximate posterior, this becomes:

$$\hat{S} = \delta S + b\, S_0, \tag{25}$$

where δS and b are averages of the corresponding α- and d-dependent quantities. Thus *independent* of the Maximum Likelihood entropy value, within our approximation, the PYM estimator obtains the entropy estimate by decreasing the ML contribution from the well-sampled head of the distribution and adding an offset that comes from the low probability tail. This is similar to so-called partition-based entropy estimators, [12, 17, 25, 26], which divide the state space into sub-spaces, estimate entropy in each sub-space, and then add the estimates weighted by the probability of being in a corresponding sub-space. However, here this partitioning arises naturally from the Bayesian framework within our approximations.

Both the scale factor and the offset depend on the dominant α and d contributing to the estimator, and hence on the usual statistics of the data, Δ, K_1, K_2, and Q_1. Specifically, we numerically observe that the value of b obtained from Eq. (25) satisfies

$$b = \langle N/(\alpha + N)\rangle \leq 1, \tag{26}$$

where the average is over the product of the approximate posterior obtained by exponentiating Eq. (20) and the prior $p(\gamma) = e^{-7\gamma/100}$ with γ defined in Eq. 15. Note that α is a measure of how much probability is concentrated in the tail. Thus the ratio $N/(\alpha + N)$ approximates the overall weight of the the well-sampled head of the distribution, requiring to decrease the contribution to the entropy from the head by this factor. This matches our assertion that the aPYM estimator is a partition-based estimator, separating the head from the tail.

In Figure 4 we show results of numerical estimation of the offset δS and the scaling factor b as a function of the fraction of coincidences, Δ/N, and the dispersion of coincidences, K_2/K_1. As in the previous case, we keep $Q_1 = 0.3 Q_{\max}$. We also set $N = 10^4$. Figure 4(a) shows that the additive term grows when the fraction of coincidences Δ/N decreases, and when K_2/K_1 is small, so that coincidences are concentrated. Both of these cases correspond to a lot of mass in the tail (see corresponding long tail region in Figure 3(a). The largest values of δS occur along the boundary strip ($\Delta/N, K_2/K_1 \ll 1$) and the boundary $K_2 = \Delta$. Panel **b** shows that the scaling factor b is close to 1 in most areas, except near the boundary edge $K_2 = \Delta$. Along this boundary, the scaling factor becomes the largest when the number of coincidences decreases, $\Delta/N \ll 1$. Figure 4 clearly highlights when Bayesian corrections to the ML estimation of entropy are essential: regions of few and concentrated coincidences.

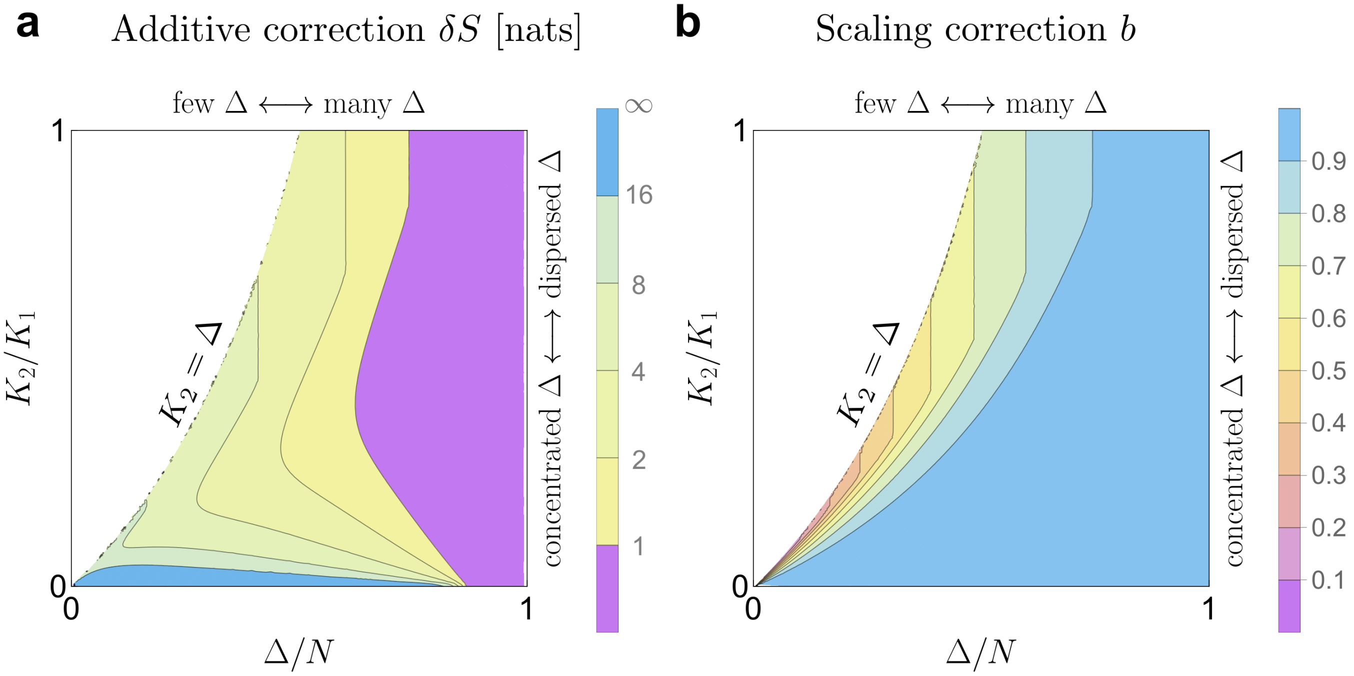

FIG. 4. Corrections to entropy estimation as a function of determining data statistics. We break down the final estimation for entropy in two parts, as $\hat{S} = \delta S({}^{\Delta}/{}_{N}, {}^{K_2}/{}_{K_1}) + b({}^{\Delta}/{}_{N}, {}^{K_2}/{}_{K_1})\, S_0$, where δS is the additive correction and b is scaling factor or weight for the Maximum Likelihood estimate. Well-sampled distributions are located in the upper-right corner where $\delta S = 0$ and $b = 1$. As in the previous plots, we leave $Q_1 = 0.3\,(\Delta - K_2)/2$. **a**: Additive correction to entropy. **b**: Scaling correction to entropy.

V. DISCUSSION

The major finding of this work is an excellent approximation for the PYM estimator, one of the best Bayesian entropy estimators, and its various relatives (such as NSB). The approximation simplifies the numerics considerably. Crucially, the approximation also shows that the output of the PYM entropy estimator depends on just a few statistics of the data, namely the maximum likelihood (ML) entropy estimate, the fraction of coincidences Δ/N, and the dispersion of coincidences K_1/K_2, and Q_1. We showed that that workflow of the estimator can be interpreted as first estimating the parameters d and α based on the aforementioned statistics, and with them the tail structure and the total weight of the tail. Then the estimator rescales the ML entropy estimate by the weight of the well-sampled head of the distribution, and adds to it the estimated entropy of the tail. The phase diagrams of which tail structure the estimator selects, Fig. 3, and how it corrects the ML estimate, Fig. 4, illustrate these points.

Early work of Ma [15] showed that when states are equiprobable, in the under-sampled regime, the coincidences in counts can help with the inference of the entropy of a system. Later Nemenman [16] showed that in the severely under-sampled regime (K_1 close to N), entropy estimation depends on the number of coincidences K_1. Further, he pointed out how reliable entropy estimates may be obtained by partitioning the overall state space of the variable into sub-spaces with similar sampling properties [26]. Here we extend these results to the whole regime where entropy estimation is challenging for multinomial observations, $\exp(S/2) < N < \exp(S)$, by approximating the more general PYM estimator. Our identification of the small set of statistics, which define the output of the estimator, lifts the veil from its inner workings, allowing for a simple, semi-analytical estimation procedure. In particular, this allows us to predict if a particular estimator will be biased simply by looking at the values of the select statistics of the data.

How to match *a priori* assumptions about the underlying distributions to the data to allow for an unbiased estimation of quantities of interest—such as entropy [10, 17] or the mutual information [27]— is an open problem [28]. It requires understanding the relation between the *a priori* assumptions and the data features that influence the inference. In this work, we build such a link for entropy estimation, and we hope that similar links might exist for other difficult estimation problems.

ACKNOWLEDGMENTS

This work was funded, in part, by the Simons Investigator award (IN), the Simons-Emory International Consortium (AR and IN), NSF Grant 1822677 (DGH and

IN), NIH Grant 2R01NS084844 (AR, IN), the International Physics of Living Systems Network (NSF Grant 1806833, AR). IN would like to acknowledge hospitality of the Aspen Center for Physics, funded in part by NSF Grant 1607611.

[1] C. E. Shannon, A mathematical theory of communication, Bell System Technical Journal **27**, 379 (1948).

[2] T. M. Cover and J. A. Thomas, *Elements of information theory* (John Wiley & Sons, 2012).

[3] S. P. Strong, R. Koberle, R. R. d. R. van Steveninck, and W. Bialek, Entropy and information in neural spike trains, Physical Review Letters **80**, 197 (1998).

[4] L. Paninski, Estimation of entropy and mutual information, Neural Computation **15**, 1191 (2003).

[5] G. Miller, Note on the bias of information estimates, Information theory in psychology: Problems and methods (1955).

[6] P. Grassberger, Entropy estimates from insufficient samplings, arXiv preprint physics/0307138 (2003).

[7] M. J. Berry II, G. Tkačik, J. Dubuis, O. Marre, and R. A. da Silveira, A simple method for estimating the entropy of neural activity, Journal of Statistical Mechanics: Theory and Experiment **2013**, P03015 (2013).

[8] D. H. Wolpert and D. R. Wolf, Estimating functions of probability distributions from a finite set of samples, Physical Review E **52**, 6841 (1995).

[9] I. Nemenman, F. Shafee, and W. Bialek, Entropy and inference, revisited, in *Advances in neural information processing systems* (2002) pp. 471–478.

[10] E. Archer, I. M. Park, and J. W. Pillow, Bayesian entropy estimation for countable discrete distributions, The Journal of Machine Learning Research **15**, 2833 (2014).

[11] A. Chao and T.-J. Shen, Nonparametric estimation of shannon's index of diversity when there are unseen species in sample, Environmental and ecological statistics **10**, 429 (2003).

[12] A. Chao, Y. Wang, and L. Jost, Entropy and the species accumulation curve: a novel entropy estimator via discovery rates of new species, Methods in Ecology and Evolution **4**, 1091 (2013).

[13] A. Cerquetti, Exact good-turing characterization of the two-parameter poisson-dirichlet superpopulation model, arXiv preprint arXiv:1901.09665 (2019).

[14] A. Antos and I. Kontoyiannis, Estimating the entropy of discrete distributions, in *Proceedings. 2001 IEEE International Symposium on Information Theory (IEEE Cat. No.01CH37252)*, p. 45.

[15] S.-k. Ma, Calculation of entropy from data of motion, Journal of Statistical Physics **26**, 221 (1981).

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

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

[18] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, ninth dover printing, tenth gpo printing ed. (Dover, New York City, 1964).

[19] T. Minka, Estimating a dirichlet distribution (2000).

[20] J. Pitman, M. Yor, *et al.*, The two-parameter poisson-dirichlet distribution derived from a stable subordinator, The Annals of Probability **25**, 855 (1997).

[21] T. S. Ferguson, A bayesian analysis of some nonparametric problems, The Annals of Statistics **1**, 209 (1973).

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

[23] H. Ishwaran and L. F. James, Generalized weighted chinese restaurant processes for species sampling mixture models, Statistica Sinica **13**, 1211 (2003).

[24] G. Schwarz, Estimating the dimension of a model, The Annals of Statistics **6**, 461 (1978).

[25] K. H. Srivastava, C. M. Holmes, M. Vellema, A. R. Pack, C. P. H. Elemans, I. Nemenman, and S. J. Sober, Motor control by precisely timed spike patterns, Proceedings of the National Academy of Sciences **114**, 1171 (2017).

[26] I. Nemenman, M. E. Wall, and C. E. Strauss, Of fishes and birthdays: Efficient estimation of polymer configurational entropies, arXiv preprint arXiv:1502.02364 (2015).

[27] D. G. Hernández and I. Samengo, Estimating the mutual information between two discrete, asymmetric variables with limited samples, Entropy **21**, 623 (2019).

[28] D. G. Hernández and I. Samengo, Inferring a property of a large system from a small number of samples, Entropy **24**, 125 (2022).

VI. APPENDIX

A. Marginal likelihood approximation for a Pitman-Yor process

In this Appendix we show how to approximate the marginal posterior of a Pitman-Yor process in the regime $K_1 \lesssim N \leq \exp(S)$. We start by manipulating each term in the logarithm of the evidence $\mathcal{L} = \log p(\mathbf{n}|\alpha, d)$ from Eq. (17),

$$\mathcal{L}(\mathbf{n}|\alpha, d) = \sum_{l=1}^{K_1-1} \log(\alpha + ld) + \sum_{i=1}^{K_1} \log\Gamma(n_i - d) - K_1 \log\Gamma(1-d) + \log\Gamma(1+\alpha) - \log\Gamma(N+\alpha). \quad (27)$$

To simplify the first term in Eq. (27), we rewrite it in terms of coincidences K_1 as follows:

$$\begin{aligned} I_1 &= \sum_{l=1}^{K_1-1} \log(\alpha + ld) \\ &= \sum_{l=1}^{K_1-1} \left[\log d + \log\left(\frac{\alpha}{d} + l\right)\right] = (K_1 - 1)\log d \\ &+ \sum_{l=1}^{K_1-1} \left[\log\Gamma\left(\frac{\alpha}{d} + l + 1\right) - \log\Gamma\left(\frac{\alpha}{d} + l\right)\right] \\ &= (K_1 - 1)\log d + \log\Gamma\left(\frac{\alpha}{d} + K_1\right) - \log\Gamma\left(\frac{\alpha}{d} + 1\right). \end{aligned} \tag{28}$$

In order to rewrite the rest of the terms of Eq. (27) in terms of various coincidence statistics, we use the identity Eq. (19). Joining the second and third terms in Eq. (27) and rewriting them in terms of count multiplicities yields

$$\begin{aligned} &\sum_{i=1}^{K_1} \log\Gamma(n_i - d) - K_1 \log\Gamma(1-d) \\ &= -K_2 \log\Gamma(1-d) + \sum_{m=2} (K_m - K_{m+1}) \log\Gamma(m-d) \\ &= \sum_{m=2} K_m \left[\log\Gamma(m-d) - \log\Gamma(m-1-d)\right] \\ &= \sum_{m=2} K_m \log(m-1-d) \\ &= K_2 \log(1-d) + Q(d), \end{aligned} \tag{29}$$

where

$$Q(d) = \sum_{m=3}^{m_f} K_m \log(m-1-d). \tag{30}$$

where m_f denotes the largest occupancy of any state in the sample. Since the domain of $0 \le d < 1$ is small, $Q(d)$ is approximately linearly varying with d, so that we can expand it around $d = 0$:

$$\begin{aligned} Q(d) = Q(0) - \sum_{j=1} \left[\sum_{m=3} \frac{K_m}{(m-1)^j}\right] \frac{d^j}{j} \\ \approx Q(0) - \left[\sum_{m=3} \frac{K_m}{m-1}\right] d \\ - \frac{1}{2}\left[\sum_{m=3} \frac{K_m}{(m-1)^2}\right] d^2 + \mathcal{O}(Q_3), \\ = Q_0 - Q_1 d - \frac{1}{2} Q_2 d^2 + \mathcal{O}(Q_3). \end{aligned}$$

where

$$Q_j = \sum_{m=3} \frac{K_m}{(m-1)^j} \tag{31}$$

for $j \ge 1$. As d approaches 1, the term $K_2 \log(1-d)$ goes to infinity, which renders any error in the Taylor expansion of $Q(d)$ irrelevant. This makes the approximations above useable even if we ignore $\mathcal{O}(d^2)$ terms.

Putting all of the approximations above together, the ensuing approximate logarithm of the evidence $\mathcal{L}(\mathbf{n}|\alpha, d)$ is

$$\begin{aligned} \mathcal{L}(\mathbf{n}|\alpha, d) &\approx (K_1 - 1)\log d + \log\Gamma\left(\frac{\alpha}{d} + K_1\right) \\ &- \log\Gamma\left(\frac{\alpha}{d} + 1\right) + \log\Gamma(1+\alpha) - \log\Gamma(N+\alpha) \\ &+ K_2 \log(1-d) - Q_1 d + \mathcal{O}\left(d^2 \sum_{m=3} \frac{K_m}{(m-1)^2}\right), \end{aligned} \tag{32}$$

up to an additive constant. This is Eq. (20) in the main text.

B. Maximum likelihood Entropy in terms of coincidences

To relate the conditional entropy, Eq. (13), to the Maximum Likelihood entropy estimator S_0, we need to rewrite the latter in terms coincidences. Utilizing the identity Eq. (19), we write

$$\begin{aligned} N\left[S_0 - \log N\right] &= -\sum_i n_i \log n_i = \\ &- \sum_{m=2} (K_m - K_{m+1}) m \log m \\ &= -K_2 (2\log 2) \\ &- \sum_{m=3} K_m \left[m\log m - (m-1)\log(m-1)\right]. \end{aligned} \tag{33}$$

Rewriting the expression in brackets as

$$m\log m - (m-1)\log(m-1) = 1 + \psi(m) + \mathcal{O}(m^{-2}). \tag{34}$$

and plugging this into Eq. (33), we finally obtain,

$$\begin{aligned} N\left[S_0 - \log N\right] &= -K_2 \log 4 - (N - K_1 - K_2) - \\ &\sum_{m=3} K_m \psi(m) + \mathcal{O}\left(\sum_m K_m/m^2\right). \end{aligned} \tag{35}$$

C. Mean posterior entropy approximation for the Pitman-Yor Process

Similar to Appendix VI A, here we approximate the posterior entropy, Eq. 13, in the limit of small d. To simplify the notation, we use the shorthand $S = \langle S|\boldsymbol{n}, \alpha, d\rangle$ in this Appendix. Rearranging Eq. (13), we obtain

$$\begin{aligned} &(\alpha + N)\left[S - \psi(N + \alpha + 1)\right] = \\ &-\alpha\,\psi(1-d) - K_1\,d\,\psi(1-d) - \sum_i (n_i - d)\psi(n_i + 1 - d). \end{aligned} \tag{36}$$

We now again use Eq. (19) and a Taylor expansion in small d to rewrite the last term on the right hand side of Eq. (36):

$$\begin{aligned}
&K_1\, d\, \psi(1-d) - \sum_i (n_i - d)\psi(n_i + 1 - d) \\
&= K_1\, d\, \psi(1-d) - \sum_{m=1} (K_m - K_{m+1})(m-d)\psi(m+1-d) \\
&= -\sum_{m=1} K_m \left[(m-d)\psi(m+1-d) - (m-1-d)\psi(m-d)\right] \\
&= -\sum_{m=1} K_m \left[1 + \psi(m-d)\right] \\
&= -\sum_{m=1} K_m - \sum_{m=1} K_m \psi(m-d) \\
&= -N - K_1\psi(1-d) - K_2\psi(2-d) - \sum_{m=3} K_m \psi(m-d).
\end{aligned} \tag{37}$$

where we used $\psi(m+1-d) = \left(\psi(m-d) + \frac{1}{m-d}\right)$.

Since $m \geq 3$, we can Taylor expand the sum in this last term around $d = 0$ to obtain

$$\begin{aligned}
\sum_{m=3} K_m \psi(m-d) \approx \sum_{m=3} K_m \psi(m) + d \sum_{m=3} K_m \psi'(m) \\
+ \mathcal{O}(d^2 \sum_m K_m \psi''(m)).
\end{aligned} \tag{38}$$

Now using the relations $\psi'(m) = \frac{1}{m-1} + \mathcal{O}(m^{-2})$ and the expression for $\sum_{m=3} K_m \psi(m)$ in Eq. (35), we rewrite Eq. (38) as

$$\begin{aligned}
&\sum_{m=3} K_m \psi(m-d) \\
&\quad \approx K_2 \log 4 + (N - K_1 - K_2) - N\left[S_0 - \log N\right] \\
&\qquad + d \sum_{m=3} \frac{K_m}{m-1} + \mathcal{O}(d^2, \sum_{m=3} K_m/m^2),
\end{aligned} \tag{39}$$

where $\mathcal{O}(d^2, \sum_{m=3} K_m/m^2)$ means that we kept terms that are at most linear in d and whose summands are at most proportional to $\sum_{m=3} K_m/m$. Plugging these approximation in Eq. (37) and noticing that $Q_1 = \sum_{m=3} \frac{K_m}{m-1}$, we obtain

$$\begin{aligned}
&(\alpha + N)\left[S - \psi(N + \alpha + 1)\right] = \\
&N(S_0 - \log N) - \alpha\, \psi(1-d) + K_1\left[-1 - \psi(1-d)\right] \\
&+ K_2\left[-1 - \psi(2-d) + \log 4\right] - Q_1\, d + \mathcal{O}(d^2, \sum_{m=3} K_m/m^2),
\end{aligned} \tag{40}$$

which after isolating S becomes Eq. (22) of the main text.