Journal of Machine Learning Research 25 (2024) 1-53 Submitted 8/23; Revised 5/24; Published 8/24

Random measure priors in Bayesian recovery from sketches

Mario Beraha MARIO.BERAHA@POLIMI.IT
Polytechnic University of Milan, Milan, Italy

Stefano Favaro STEFANO.FAVARO@UNITO.IT
University of Torino and Collegio Carlo Alberto, Torino, Italy

Matteo Sesia SESIA@MARSHALL.USC.EDU
University of Southern California, Los Angeles, California, United States

Editor: Debdeep Pati

Abstract

This paper introduces a Bayesian nonparametric approach to frequency recovery from lossy-compressed discrete data, leveraging all information contained in a sketch obtained through random hashing. By modeling the data points as random samples from an unknown discrete distribution endowed with a Poisson-Kingman prior, we derive the posterior distribution of a symbol's empirical frequency given the sketch. This leads to principled frequency estimates through mean functionals, e.g., the posterior mean, median and mode. We highlight applications of this general result to Dirichlet process and Pitman-Yor process priors. Notably, we prove that the former prior uniquely satisfies a sufficiency property that simplifies the posterior distribution, while the latter enables a convenient large-sample asymptotic approximation. Additionally, we extend our approach to the problem of cardinality recovery, estimating the number of distinct symbols in the sketched dataset. Our approach to frequency recovery also adapts to a more general "traits" setting, where each data point has integer levels of association with multiple symbols, typically referred to as "traits". By employing a generalized Indian buffet process, we compute the posterior distribution of a trait's frequency using both the Poisson and Bernoulli distributions for the trait association levels, respectively yielding exact and approximate posterior frequency distributions.

Keywords: Bayesian nonparametrics; cardinality recovery; frequency recovery; Poisson-Kingman prior; random hashing.

1. Introduction

1.1 Background and motivation

Information recovery about a large discrete dataset given a (lossy) compressed representation, or *sketch*, of that data is a classical problem at the crossroad of computer science and information theory (Misra and Gries, 1982; Alon et al., 1999; Manku and Motwani, 2002; Karp et al., 2003; Charikar et al., 2004; Cormode and Muthukrishnan, 2005; Indyk, 2006). Sketching is often driven by the need to manage memory constraints, as handling vast numbers of symbols can be computationally intensive, and by privacy concerns, especially when dealing with sensitive data (Blum et al., 2020; Cormode and Yi, 2020; Medjedovic et al., 2022). Two well-known challenges based on sketched data are *frequency recovery* and *cardinality recovery*. For concreteness, we begin by discussing frequency recovery.

In the typical "species" setting, where the original dataset comprises $n \geq 1$ points $(x_1, \ldots, x_n)$ with each x_i corresponding to a symbol or "species" label taking values in a set $\mathbb{S}$, the goal of *frequency recovery* is to estimate the number of occurrences of a new object x_{n+1} in $(x_1, \ldots, x_n)$, denoted as $f_{x_{n+1}}$. Formally, this can be written as:

$$f_{x_{n+1}} = \sum_{i=1}^{n} I(x_i = x_{n+1}),$$

with $I(\cdot)$ denoting the indicator function. This problem is relevant for many applications, including machine learning in high-dimensional feature spaces (Shi et al., 2009; Aggarwal and Yu, 2010), cybersecurity in tracking password popularity (Schechter et al., 2010), web and social network data analysis (Song et al., 2009; Cormode, 2017), natural language processing (Goyal et al., 2012), sequencing analysis in biological sciences (Zhang et al., 2014; Berger et al., 2016; Solomon and Kingsford, 2016; Marcais et al., 2019; Leo Elworth et al., 2020), and privacy-protecting data analysis (Dwork et al., 2010; Melis et al., 2016; Cormode, 2017; Cormode et al., 2018; Kockan et al., 2020).

The count-min sketch (CMS) is a popular algorithm for frequency recovery (Cormode and Muthukrishnan, 2005). It relies on a sketch of $(x_1, \ldots, x_n)$ obtained by $D \geq 1$ independent J-wide random hash functions $h_k : \mathbb{S} \to [J] := \{1, \ldots, J\}$, for $k \in [D] := \{1, \ldots, D\}$ and $J \geq 1$. Each hash function maps the x_i's into J buckets, defining the sketch $\mathbf{C}_{D,J} \in \mathbb{N}_0^{D \times J}$, with $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$, whose (k, j)-th element $C_{k,j}$ counts the data points mapped by the k-th hash function into the j-th bucket. Based on $\mathbf{C}_{D,J}$, the CMS bounds $f_{x_{n+1}}$ from above by taking the smallest count among the D buckets into which x_{n+1} is mapped, i.e.,

$$\hat{f}_{x_{n+1}} = \min\{C_{1,h_1(x_{n+1})}, \ldots, C_{D,h_D(x_{n+1})}\}. \tag{1}$$

We refer to Figure 1 in Section 2 for a schematic visualization of this procedure, focusing on the special case where the data are sketched using a single hash function. In general, while the CMS upper bound is remarkable for its simplicity and robustness, it becomes loose if hash collisions (different objects mapped into the same bucket) are frequent, due to the pessimistic assumption that data are fixed (Cormode and Yi, 2020, Chapter 3).

This challenge has recently motivated statistical approaches that treat the sketched data as random and rely on modeling assumptions to obtain more informative estimates. The first Bayesian nonparametric (BNP) approach to frequency recovery was introduced by Cai et al. (2018). They modeled the x_i's as a random sample $(X_1, \ldots, X_n)$ from an unknown discrete distribution endowed with a Dirichlet process (DP) prior (Ferguson, 1973) and obtained estimates of $f_{X_{n+1}}$ as mean functionals of the posterior distribution of $f_{X_{n+1}}$ given $(C_{1,h_1(X_{n+1})}, \ldots, C_{D,h_D(X_{n+1})})$, e.g., the mean, median and mode. In addition to enabling the inclusion of prior knowledge on the data distribution, the BNP approach allows assessing uncertainty using posterior distributions. See Dolera et al. (2021, 2023) for an extension of Cai et al. (2018) to the Pitman-Yor process (PYP) prior (Pitman and Yor, 1997).

As outlined in Section 1.2, this paper extends the foundations laid by previous research to establish a more versatile and comprehensive BNP framework. In particular, our novel framework allows: conditioning on all information contained in the sketched data, employing a wider array of prior distributions, and tackling other information retrieval challenges beyond frequency recovery in the "species" setting.

Specifically, we focus on analyzing a sketch $\mathbf{C}_J = (C_1, \ldots, C_J) \in \mathbb{N}_0^J$ obtained by a *single hash function*. The work of Cai et al. (2018) was motivated by the goal developing a learning-augmented version of the CMS, hence their interest in a posterior distribution with respect to the same information from a sketch $\mathbf{C}_{D,J}$ obtained from D distinct hash functions as in (1). By contrast, we develop a purely Bayesian approach, separate from the CMS, including in the posterior distribution all information from the sketch $\mathbf{C}_J$. From a Bayesian statistical perspective, our approach is arguably more natural than that of Cai et al. (2018) because: i) the practical usefulness of combining a statistical model for the x_i's with the sketch $\mathbf{C}_{D,J}$ from multiple hash functions would be unclear, being such a sketch designed for a (model-free) recovery algorithm; ii) the use of a posterior distribution with respect to the sole $C_{k,h_k(X_{n+1})}$'s may determine a loss of information, unless we can verify that the $C_{k,h_k(X_{n+1})}$'s are sufficient to estimate $f_{X_{n+1}}$.

1.2 Preview of our contributions

1.2.1 BNP frequency recovery

Our first contribution is a novel BNP approach to frequency recovery that utilize the *full* posterior distribution of $f_{X_{n+1}}$ given a sketch obtained with a single random hash function. This approach departs from that of Cai et al. (2018), which did not condition on the information contained in the sketch outside the bucket into which X_{n+1} is hashed. We compute the full posterior distribution of $f_{X_{n+1}}$ given $\mathbf{C}_J$ and the bucket in which X_{n+1} is hashed, denoted as $h(X_{n+1})$. We derive this posterior under a broad class of Poisson-Kingman (PK) priors (Pitman, 2003), encompassing both the DP and PYP priors. For the DP prior, we find that the posterior depends on $\mathbf{C}_J$ only through $C_{h(X_{n+1})}$; this is consistent with the approach of Cai et al. (2018), proving that it relies on a sufficient statistic. However, our findings also reveal that the DP prior is the only PK prior satisfying this sufficiency property. Additionally, we show that the posterior distribution is computationally intractable under the PYP prior if n is large. Thus, we develop a large-sample asymptotic approximation.

1.2.2 BNP cardinality recovery

Our second contribution extends the BNP approach to frequency recovery to address the *cardinality recovery* problem (Cormode and Yi, 2020, Chapter 2)—another classical problem in computer science. Here, the objective is to utilize the same sketch obtained with a single hash function to estimate the number of distinct symbols in $(x_1, \ldots, x_n)$, i.e.,

$$k_n = |x_1, \ldots, x_n|,$$

where $|\cdot|$ denotes the cardinality set-function. Prior works treated frequency and cardinality recovery as separate problems, each requiring a different sketching algorithm. For instance, the hyperloglog algorithm is widely used for cardinality recovery (Flajolet et al., 2007; Flajolet and Martin, 1983, 1985). See also Chassaing and Gerin (2006), Chen et al. (2011), Ting (2014, 2016), and Pettie and Wang (2021) for recent contributions, some of which rely on modeling assumptions for the data. We show that the BNP approach enables a direct connection between the frequency and cardinality recovery problems, yielding a posterior mean estimate of k_n from the sketch $\mathbf{C}_J$. To the best of our knowledge, this is the first

approach enabling both frequency and cardinality recovery from the same sketch $\mathbf{C}_J$. More generally, for any $l \in [n]$, we obtain a posterior mean estimate of the number $m_{l,n}$ of distinct symbols with frequency l in $(x_1, \ldots, x_n)$, commonly referred to as the *l-cardinality*. This enables the recovery of the partition structure of $(x_1, \ldots, x_n)$ from $\mathbf{C}_J$.

1.2.3 BNP frequency recovery in the "traits" setting

Our third contribution is a BNP approach to a new "traits" setting of frequency recovery, in which each data point may be associated with more than one symbol, called "traits", and exhibits levels of association with each of those traits (Campbell et al., 2018). Multi-trait data arise in many domains: single-cell data encompass multiple genes with their expression levels, members of social networks connect with multiple friends to whom they send messages, and documents encompass different topics with their words. We consider $n \geq 1$ data points $(x_1, \ldots, x_n)$, where each x_i takes on a value in a set $\mathbb{S}^\infty \times \mathbb{N}_0^\infty$ representing traits and their levels. We assume the x_i's to be modeled as a random sample $(X_1, \ldots, X_n)$ from the generalized Indian buffet process (James, 2017). In this setting, the X_i's are random counting measures, i.e. $X_i = \sum_{k \geq 1} A_{i,k} \delta_{w_k}$, where $A_{i,k} \in \mathbb{N}_0$ is the level of association of the trait $w_k \in \mathbb{S}$ for the i-th data point. The distribution of X_i is determined by: i) a distribution G_A for the level of association $A_{i,k}$, which depends on a parameter $J_k > 0$, for $k \geq 1$; ii) the law of a completely random measure (CRM) serving as a prior distribution for the J_k's. Based on a sketch $\mathbf{C}_J \in \mathbb{N}_0^J$ of $(X_1, \ldots, X_n)$ generated by a random hash function, we compute the posterior distribution of the empirical frequency level of a trait for a new X_{n+1}, given $\mathbf{C}_J$ and the bucket in which such a trait is hashed. It emerges that any CRM prior leads to a posterior distribution that depends on $\mathbf{C}_J$ solely through $C_{h(X_{n+1})}$, exhibiting a sufficiency property akin to that found in the "species" setting under the DP prior. Applications to Poisson and a Bernoulli level G_A are presented in details. Our findings illustrate how the BNP approach facilitates frequency recovery in the broader "traits" setting, for which we are unaware of any existing algorithms.

1.3 Related work

Several recent studies have studied from a statistical perspective the frequency recovery problem under the "species" setting. On the frequentist front, Ting (2018) introduced a bootstrap method tailored to the CMS algorithm, focusing on asymptotic guarantees. Sesia and Favaro (2022) and Sesia et al. (2023) proposed an alternative frequentist approach based on conformal inference ideas (Vovk, 2005), obtaining uncertainty estimates with finite-sample guarantees for any (possibly non-linear) sketch. It is worth noting that these different approaches can be viewed as complementary to each other. Specifically, conformal inference can be combined with both the bootstrap (Sesia and Favaro, 2022) and our Bayesian approach. This combination can, for example, yield "calibrated" Bayesian credible intervals with finite-sample frequentist properties (Sesia and Favaro, 2022). Recently, Beraha et al. (2023) proposed a "smoothed" BNP approach aimed at mitigating some of the computational limitations inherent in Bayesian methods, which we highlight in this paper. While the smoothed approach yields computationally simpler estimators compared to our full Bayesian approach, it is limited to the case of normalized random measures, thus excluding the Pitman-Yor process, and only addresses the "species" setting. In the case of

the DP prior, it turns our that the "smoothed" and BNP estimators coincide, yet the DP stands out as the only random measure for which this alignment occurs.

1.4 Organization of the paper

The paper is structured as follows. Section 2 considers the "species" setting for frequency recovery, developing our BNP approach under a general PK prior, and in the special cases of the DP and PYP priors. Also within the "species" setting, Section 3 presents a BNP approach to cardinality recovery under the DP and PYP priors, obtaining estimators for the l-cardinality and the cardinality of the dataset. In Section 4 we consider the "traits" setting for frequency recovery, developing our BNP approach under a general CRM prior and a general distribution for the level of association, as well as for the special cases of Poisson and Bernoulli level distribution. Section 5 contains an empirical validation of our methods on synthetic and real data, whereas Section 6 discusses some directions for future work. Proofs and other technical derivations are deferred to the Appendices. A software implementation of our methods is available at `https://github.com/mberaha/BNPSketching`.

2. BNP frequency recovery in the "species" setting

2.1 Problem statement and setup

For $n \geq 1$, consider data points $(x_1, \ldots, x_n)$, with each x_i representing a symbol (or "species") label from a dictionary $\mathbb{S}$. Consider having access only to a sketch of $(x_1, \ldots, x_n)$, obtained through random hashing (Mitzenmacher and Upfal, 2017, Chapter 5 and Chapter 15). For an integer $J \geq 1$, let h be a random hash function of width J, defined as a random mapping from $\mathbb{S}$ to $[J]$, chosen from a pairwise independent hash family $\mathcal{H}_J$. That is, $h : \mathbb{S} \to [J]$, and, for any $j_1, j_2 \in [J]$ and fixed $x_1, x_2 \in \mathbb{S}$ such that $x_1 \neq x_2$,

$$\Pr[h(x_1) = j_1,\, h(x_2) = j_2] = \frac{1}{J^2}.$$

The pairwise independence of $\mathcal{H}_J$, also known as strong universality, implies uniformity, meaning that $\Pr[h(x) = j] = J^{-1}$ for $j \in [J]$. Hashing $(x_1, \ldots, x_n)$ through h produces a random vector $\mathbf{C}_J = (C_1, \ldots, C_J) \in \mathbb{N}_0^J$, termed "sketch", whose j-th element (bucket) is

$$C_j = \sum_{i=1}^{n} I(h(x_i) = j),$$

so that $\sum_{1 \leq j \leq J} C_j = n$. This sketch generally has a smaller (physical) size than $(x_1, \ldots, x_n)$ due to the collisions of the x_i's induced by random hashing (Cormode and Yi, 2020, Chapter 3). The sketch $\mathbf{C}_J$ is a special version of the sketch $\mathbf{C}_{D,J} \in \mathbb{N}_0^{D \times J}$ at the basis of the CMS (Cormode and Muthukrishnan, 2005), which simultaneously applies a collection of $D \geq 1$ independent random hash functions from $\mathcal{H}_J$ (Cormode and Yi, 2020, Chapter 3). Based on a sketch $\mathbf{C}_J$ of $(x_1, \ldots, x_n)$, we study the BNP estimation of the empirical frequency $f_{x_{n+1}}$ of x_{n+1} in $(x_1, \ldots, x_n)$; see Figure 1 for a schematic visualization.

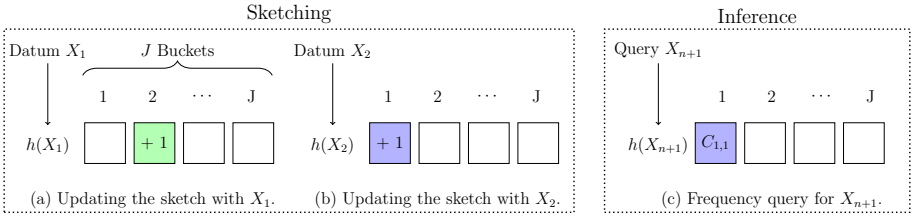


Figure 1: Data sketching in the "species" setting. Each object X_i, for $i \in \{1, \dots, n\}$, is assigned to one of J possible "buckets" (shown in different colors) by a random hash functions h, and the corresponding counters are incremented by one. One interesting problem is to estimate the empirical frequency of a "query" X_{n+1} based on the sketch and the bucket assignments for X_{n+1}.

2.2 The BNP model

Inspired by Cai et al. (2018) and Dolera et al. (2023), we rely on two modeling assumptions: i) the data $(x_1, \dots, x_n)$ are modeled as a random sample $\mathbf{X}_n = (X_1, \dots, X_n)$ from an unknown discrete distribution P endowed by a prior $\mathscr{P}$; ii) the hash family $\mathcal{H}_J$ is independent of P. Then, we write the BNP model as:

$$\begin{aligned} C_j &= \sum_{i=1}^{n} I(h(X_i) = j), \qquad j \in [J], \\ h &\sim \mathcal{H}_J, \\ X_1, \dots, X_n \,|\, P &\overset{\text{iid}}{\sim} P, \\ P &\sim \mathscr{P}. \end{aligned} \tag{2}$$

Under this model, the problem of recovering the empirical frequency $f_{X_{n+1}}$ from the sketch consists of computing the posterior distribution of $f_{X_{n+1}}$ given $\mathbf{C}_J$ and the bucket in which X_{n+1} is hashed, i.e., $h(X_{n+1})$. Below, we compute this posterior assuming $\mathscr{P}$ belongs to the broad class of PK priors, and then we specialize our result to DP and PYP priors.

2.3 Background on PK priors

To define a PK prior (Pitman, 2003), we consider a completely random measure (CRM) $\tilde{\mu}$ on $\mathbb{S}$, which is a random element with values on the space of bounded measures on $(\mathbb{S}, \mathcal{S})$, such that, for $k \geq 1$ and for disjoint Borel sets $A_1, \dots, A_k \in \mathcal{S}$, the random variables $\tilde{\mu}(A_1), \dots, \tilde{\mu}(A_k)$ are independent (Kingman, 1967). We focus on CRMs of the form $\tilde{\mu}(\cdot) = \int_{\mathbb{R}_+} s\tilde{N}(\mathrm{d}s, \cdot) = \sum_{k\geq 1} J_k \delta_{W_k}(\cdot)$, where $\tilde{N} = \sum_{k\geq 1} \delta_{(J_k, W_k)}$ is a Poisson random measure on $\mathbb{R}_+ \times \mathbb{S}$ with Lévy intensity $\nu(\mathrm{d}s, \mathrm{d}x)$, which characterizes the distribution of $\tilde{\mu}$ in terms its random jumps J_k's and random locations W_k's (Kingman, 1967, 1993). We focus on homogeneous Lévy intensities, in the form $\nu(\mathrm{d}s, \mathrm{d}x) = \theta\rho(s)\mathrm{d}s\, G_0(\mathrm{d}x)$, where $\theta > 0$ is a parameter, G_0 is a nonatomic probability measure on $\mathbb{S}$, and $\rho(s)\mathrm{d}s$ is a measure on $\mathbb{R}_+$ such that $\int_{\mathbb{R}_+} \rho(s)\mathrm{d}s = +\infty$ and

$$\psi(u) = \int_{\mathbb{R}_+} (1 - e^{-us})\rho(s)\mathrm{d}s < +\infty \tag{3}$$

for all $u > 0$, ensuring $0 < \tilde{\mu}(\mathbb{S}) < +\infty$ almost surely (Pitman, 2003; Regazzini et al., 2003). We write $\tilde{\mu} \sim \text{CRM}(\theta, \rho, G_0)$ to denote a homogeneous CRM on $\mathbb{S}$. A PK prior is the law of a "normalization" of a CRM with respect to its total mass (Pitman, 2006, Chapter 4).

Definition 1. *Let $\tilde{\mu} \sim CRM(\theta, \rho, G_0)$ with total mass $T = \tilde{\mu}(\mathbb{S}) \sim f_T$. Let $P_k := J_k/T$ be the normalized random jumps of $\tilde{\mu}$, and denote by $PK(\theta, \rho \,|\, T = t)$ the conditional distribution of $(P_k)_{k\geq 1}$ given $T = t$. If $g(\tilde{\mu}) \equiv g(T)$ such that $\mathbb{E}[g(T)] = 1$, then a PK prior with parameters $(\theta, \rho, gf_T, G_0)$ is the law of the (discrete) random probability measure $P(\cdot) = \sum_{k\geq 1} \tilde{P}_k \delta_{W_k}(\cdot)$, on $\mathbb{S}$, where $(\tilde{P}_k)_{k\geq 1}$ is distributed as the PK distribution $PK(\theta, \rho, gf_T) = \int_{\mathbb{R}_+} PK(\theta, \rho \,|\, T = t) g(t) f_T(t) \mathrm{d}t$, and the W_k's are independent and identically distributed as G_0.*

2.4 General posterior distribution for PK priors

Under the model in (2), with $\mathscr{P}$ being a PK prior, i.e., $P \sim \text{PK}(\theta, \rho, gf_T, G_0)$, the next theorem gives the posterior distribution of $f_{X_{n+1}}$ given $\mathbf{C}_J$ and $h(X_{n+1})$. This is result can be applied upon suitable specifications of the measure ρ and the function g. For instance, PK priors include the class of (homogeneous) normalized CRM priors, obtained by setting g as the identity function (James, 2002; Prünster, 2002; Pitman, 2003; Regazzini et al., 2003). It is sufficient to consider $g(t) \propto t^{-\gamma} e^{-\beta t}$ to recover from Definition 1 the most popular priors in BNPs. Common choices of $\tilde{\mu}$ are the Gamma CRM and the α-Stable CRM (Kingman, 1975), as they provide PK priors with a flexible tail behaviour, ranging from geometric tail to heavy power-law tails, respectively (Pitman, 2006, Chapter 3 and Chapter 4). Under the Gamma CRM, Definition 1 generalizes the DP prior, which is obtained by setting g as the identity function. Under the α-Stable CRM, Definition 1 provides a generalization of the normalized α-Stable prior (Kingman, 1975), which is obtained by setting g as the identity function, and it also includes the PYP prior and the normalized Gamma process prior (James, 2002; Lijoi et al., 2005, 2007). See (Pitman, 2006, Chapter 3) and Lijoi and Prünster (2010) for other examples of PK priors. The application of the next theorem to the DP and the PYP priors will be considered below, showing their peculiar features.

Theorem 2. *Let $\mathbf{C}_J$ be a sketch of $\mathbf{X}_n$ under (2), with $P \sim PK(\theta, \rho, gf_T, G_0)$ and $g(t) \propto t^{-\gamma} e^{-\beta t}$, and consider an additional (unobservable) X_{n+1}. With ψ defined as in (3),*

$$\begin{aligned}\Pr[f_{X_{n+1}} = l \,|\, \mathbf{C}_J = \mathbf{c}, h(X_{n+1}) = j] \\ = \frac{\theta}{J}\binom{c_j}{l} \frac{\int_{\mathbb{R}_+} u^{n+\gamma} \phi^{(c_j - l)}(u+\beta) \prod_{k\neq j} \phi^{(c_k)}(u+\beta)\, \kappa(u+\beta, l+1)\mathrm{d}u}{\int_{\mathbb{R}_+} u^{n+\gamma} \phi^{(c_j+1)}(u+\beta) \prod_{k\neq j} \phi^{(c_k)}(u+\beta)\mathrm{d}u}, \quad (4)\end{aligned}$$

for all $l \in \{0, 1, \ldots, c_j\}$, where $\phi^{(n)}(u) = (-1)^n \frac{\mathrm{d}^n}{\mathrm{d}u^n} e^{-\theta/J\psi(u)}$ and $\kappa(u, n) = \int_{\mathbb{R}_+} e^{-us} s^n \rho(s)\mathrm{d}s$.

See Appendix A.1 for the proof of Theorem 2. This result provides a BNP solution to the frequency recovery problem, as it leads to an estimator of $f_{X_{n+1}}$, with respect to a suitable loss function, as a mean functional of (2). Credible intervals may be also derived. The use of the squared loss leads to the posterior mean as an estimator of $f_{X_{n+1}}$, i.e.,

$$\hat{f}_{X_{n+1}} = \sum_{l=0}^{c_j} l \cdot \Pr[f_{X_{n+1}} = l \,|\, \mathbf{C}_J = \mathbf{c}, h(X_{n+1}) = j]. \quad (5)$$

Regarding the use of the sketch, Cai et al. (2018, Proposition 1) and Dolera et al. (2023, Theorems 1 and 2) provide posterior distributions of $f_{X_{n+1}}$ with respect to the sole bucket $C_{h(X_{n+1})}$ in which X_{n+1} is hashed. By contrast, Theorem 2 considers the entire sketch $\mathbf{C}_J$ and the hash bucket of X_{n+1}, denoted as $h(X_{n+1})$. Therefore, our posterior distribution provides a principled BNP estimator also in those situations where $C_{h(X_{n+1})}$ may not be a sufficient statistic, as elaborated below. Regarding the specification of the prior distribution, Cai et al. (2018) and Dolera et al. (2023) focused on the DP and PYP priors, obtaining posterior distributions through conjugacy or quasi-conjugacy properties. However, Theorem 2 considers a general PK prior. As discussed in Dolera et al. (2023), relying on conjugacy poses a limitation when aiming to consider more diverse prior distributions because the DP and PYP priors are the only quasi-conjugate PK priors. Our proof of Theorem 2 overcomes this limitation by avoiding the use of any form of conjugacy for the prior.

2.5 Results under the DP prior

We consider Definition 1 with $\tilde{\mu}$ being a Gamma CRM, i.e., $\rho(s) = s^{-1}\exp\{-s\}$, and g the identity function. Then, $P \sim \mathrm{PK}(\theta, \rho, gf_T, G_0)$ is a DP with mass $\theta > 0$ and base measure G_0 (Ferguson, 1973; Pitman, 2003); in short, $P \sim \mathrm{DP}(\theta, G_0)$. In this case, a simplified expression for the posterior of $f_{X_{n+1}}$ given $\mathbf{C}_J$ and $h(X_{n+1})$ is obtained from Theorem 2 as follows. Denote by $(a)_{(n)}$ the rising factorial of a of order n, i.e., $(a)_{(n)} = \prod_{0\le i\le n-1}(a+i)$ with the proviso $(a)_{(0)} := 1$ (Charalambides, 2005, Chapter 2). for any $l = 0, 1, \ldots, c_j$,

$$\Pr[f_{X_{n+1}} = l \,|\, \mathbf{C}_J = \mathbf{c}, h(X_{n+1}) = j] = \frac{\theta}{J}\frac{(c_j - l + 1)_{(l)}}{\left(\frac{\theta}{J} + c_j - l\right)_{(l+1)}}. \tag{6}$$

See Appendix A.3 for a proof of how (6) follows from Theorem 2, and Appendix A.4 for an alternative proof based on finite-dimensional properties of the DP. It is easy to see that (6) is a Beta-Binomial distribution, namely a Binomial distribution in which the probability of success at each of the c_j trials is a Beta random variable with parameter $(1, \theta/J)$, say $B_{1,\theta/J}$. That is, if $F_{X_{n+1}}$ is a random variable distributed as (6), then de Finetti's theorem implies that $c_j^{-1}F_{X_{n+1}}$ converges (weakly) to a $B_{1,\theta/J}$ as $c_j \to +\infty$. The posterior distribution in (6) depends on the sketch $\mathbf{C}_J$ only through $C_{h(X_{n+1})}$. Thus, under the DP prior, $C_{h(X_{n+1})}$ is a sufficient statistic for estimating $f_{X_{n+1}}$ from $\mathbf{C}_J$, making (6) equivalent to the posterior distribution in Cai et al. (2018, Proposition 1).

The next theorem, proved in Appendix A.2, characterizes the DP prior as the unique PK prior for which the posterior distribution of $f_{X_{n+1}}$ with respect to $C_{h(X_{n+1})}$ is equivalent to the posterior distribution with respect to $\mathbf{C}_J$ and $h(X_{n+1})$.

Theorem 3. *The DP prior is the sole PK prior for which* (4) *depends on* $\mathbf{C}_J$ *only through* $C_{h(X_{n+1})}$.

In practice, evaluating (6) requires estimating the unknown parameter $\theta > 0$ of the prior from the sketch $\mathbf{C}_J$. Cai et al. (2018) proposed an empirical Bayes approach to estimate θ, which relies on the following finite-dimensional projective property of $P \sim \mathrm{DP}(\theta, G_0)$: if $\{B_1, \ldots, B_k\}$ is a measurable k-partition of $\mathcal{S}$, for $k \ge 1$, then $(P(B_1), \ldots, P(B_k))$ follows a Dirichlet distribution with parameter $(\theta G_0(B_1), \ldots, \theta G_0(B_k))$ (Ferguson, 1973; Regazzini,

2001). Due to the finite-dimensional projective property and the assumption that $\mathcal{H}_J$ is independent of P, the sketch $\mathbf{C}_J$ is distributed as a Dirichlet-Multinomial distribution, i.e.,

$$\Pr[\mathbf{C}_J = \mathbf{c}] = \frac{n!}{(\theta)_{(n)}} \prod_{j=1}^{J} \frac{(\frac{\theta}{J})_{(c_j)}}{c_j!}. \tag{7}$$

This distribution enables estimating θ directly, by maximizing the (marginal) likelihood (7) over θ (Cai et al., 2018). The estimated value of θ can then be plugged into the posterior distribution (6). A fully Bayesian approach is also possible, by assigning a prior distribution to θ and assessing the resulting posterior distributions through Monte Carlo sampling.

2.6 Results under the PYP prior

Consider Definition 1 with $\theta = 1$, $\tilde{\mu}$ being an α-Stable CRM, i.e., $\rho(s) = (\alpha/\Gamma(1-\alpha))s^{-1-\alpha}$ for $\alpha \in (0,1)$ and $g(t) = (\Gamma(\gamma+1)/\Gamma(\gamma/\alpha+1)t^{-\gamma}$ for $\gamma > -\alpha$, where $\Gamma(\cdot)$ denotes the Gamma function, namely $\Gamma(x) = \int_{(0,+\infty)} z^{x-1}\mathrm{e}^{-z}\mathrm{d}z$ for $x > 0$. Then, $P \sim \mathrm{PK}(\theta, \rho, gf_T, G_0)$ is a PYP with discount α, mass γ and base measure G_0 (Pitman, 2003); in short, we write $P \sim \mathrm{PYP}(\alpha, \gamma, G_0)$. In this case, a simplified expression for the posterior of $f_{X_{n+1}}$ given $\mathbf{C}_J$ and $h(X_{n+1})$ is obtained from Theorem 2 as follows. For $n \geq 0$ and $0 \leq k \leq n$,

$$\mathscr{C}(n,k;\alpha) = \frac{1}{k!}\sum_{i=0}^{k}(-1)^i\binom{n}{i}(-i\alpha)_{(n)}$$

denotes the generalized factorial of n of order k, with $\mathscr{C}(0,0;\alpha) := 0$, and $\mathscr{C}(n,0;\alpha) := 1$ (Charalambides, 2005, Chapter 2). Then, from (4), for any $l = 0, 1, \ldots, c_j$,

$$\begin{aligned}&\Pr[f_{X_{n+1}} = l \,|\, \mathbf{C}_J = \mathbf{c}, h(X_{n+1}) = j] \\ &= \frac{\gamma}{J}\binom{c_j}{l}(1-\alpha)_{(l)} \frac{\sum_{\boldsymbol{i}\in\mathcal{S}(\boldsymbol{c},j,-l)} \frac{\Gamma\left(\frac{\gamma+\alpha}{\alpha}+|\boldsymbol{i}|\right)}{J^{|\boldsymbol{i}|}} \prod_{k=1}^{J} \mathscr{C}(c_k - l\delta_{k,j}, i_k; \alpha)}{\sum_{\boldsymbol{i}\in\mathcal{S}(\boldsymbol{c},j,1)} \frac{\Gamma\left(\frac{\gamma}{\alpha}+|\boldsymbol{i}|\right)}{J^{|\boldsymbol{i}|}} \prod_{k=1}^{J} \mathscr{C}(c_k + \delta_{k,j}, i_k; \alpha)},\end{aligned} \tag{8}$$

where $\mathcal{S}(\boldsymbol{c}, j, q)$ is the Cartesian product $\times_{1\leq k\leq J}\{0, \ldots, c_k + \delta_{k,j}q\}$, while $\delta_{k,j}$ is the Kronecker delta and $|\boldsymbol{i}| = \sum_{1\leq k\leq J} i_k$. See Appendix A.5 for the proof of (8). The posterior distribution (8) generalizes (6), which is recovered for $\gamma = \theta$ and $\alpha \to 0$; see Appendix A.6.

Unlike that in (6), the posterior distribution in (8) relies on the entirety of $\mathbf{C}_J$. Further, the size of $\mathcal{S}(\boldsymbol{c}, j, q)$ increases exponentially with J and n. For instance, if $J = 10$ and $c_j = 5$ for all j, then $|\mathcal{S}(\boldsymbol{c}, j, q)| \approx 60 \times 10^6$. Consequently, the evaluation of (8) is intractable even for moderately large n, as it necessitates summations over an exceedingly large number of generalized factorial coefficients, depending on J. See Appendix A.14 for some approaches to evaluate (8), which still lead to non-trivial computational obstacles. To overcome this challenge, we seek an approximation of (8) that depends on $\mathbf{C}_J$ only through $C_{h(X_{n+1})}$. The next theorem characterizes the large-sample behaviour of the posterior mean estimator (5).

Theorem 4. *Let $\mathbf{C}_J$ be a sketch of $\mathbf{X}_n$ under the model in (2), with $P \sim PYP(\alpha, \gamma, G_0)$ for $\alpha > 0$, and consider an additional (unobservable) X_{n+1}. Suppose $h(X_{n+1}) = j$, for*

some fixed $j \in \{1, \dots, J\}$. *Define* $\mathbf{c}_{-j} := (c_1, \dots, c_{j-1}, c_{j+1}, \dots, c_J)$. *Let* $\hat{f}_{X_{n+1}}$ *denote the posterior mean estimator as defined in* (5). *Then,*

$$\lim_{\mathbf{c}_j \to +\infty} \lim_{\mathbf{c}_{-j} \to +\infty} \frac{\hat{f}_{X_{n+1}}}{c_j} = \frac{\gamma}{\alpha} \cdot \frac{1-\alpha}{\gamma + J\alpha - \alpha + 1}.$$

See Appendix A.7 for the proof of Theorem 4. This result motivates approximating the posterior mean estimator $\hat{f}_{X_{n+1}}$, in those situations where all the c_i's are large, with:

$$\tilde{f}_{X_{n+1}} := c_j \cdot \frac{\gamma}{\alpha} \cdot \frac{1-\alpha}{\gamma + J\alpha - \alpha + 1}. \tag{9}$$

Applying either (8) or (9) requires estimating (α, γ) from the sketch $\mathbf{C}_J$. Under the PYP prior, the distribution of $\mathbf{C}_J$ has no simple closed-form expression, which impedes the use of empirical Bayes or fully Bayesian approaches to estimate (α, γ). To address a similar challenge, Dolera et al. (2023) adopted a likelihood-free approach based on the Wasserstein distance (Bernton et al., 2019). That involves sampling independent data sets $\mathbf{X}'_n$ from $P \sim \text{PYP}(\alpha', \gamma', G_0)$ and sketching them into $\mathbf{C}'_J$ using the same hash function h. Then, (α, γ) are estimated by minimizing (a suitable approximation of) the 1-Wasserstein distance between $\mathbf{C}'_J$ and $\mathbf{C}_J$. Since the objective function is not differentiable, Dolera et al. (2023) utilized Bayesian Optimization to estimate the parameters.

However, their approach is computationally demanding: the cost of evaluating the Wasserstein distance is $\mathcal{O}(J^3)$, while the cost of simulating $\mathbf{X}'_n$ scales super-linearly with n, depending on the parameters' values. Therefore, we propose an alternative approach: we consider the first $n' \ll n$ observations, sketch them through h, and estimate their frequencies using (9). Then, the mean absolute error of the frequency recovery can be easily minimized with respect to the unknown parameters using standard software packages since the loss function is differentiable almost everywhere.

2.7 Comparison of the posterior under the PYP and DP priors

We present a numerical illustration of the posterior distributions under the DP and PYP priors; see Section 5 for a more extensive simulation study. We set $n = 50$ and $J = 10$, considering four sketched datasets: $\mathbf{C}_J^{(1)}$, $\mathbf{C}_J^{(2)}$, $\mathbf{C}_J^{(3)}$, and $\mathbf{C}_J^{(4)}$, with values reported in Table 1. Specifically: i) in scenario $\mathbf{C}_J^{(1)}$, the values $C_j^{(1)}$'s are constant across j; ii) in scenario $\mathbf{C}_J^{(2)}$, the values $C_j^{(2)}$'s decay exponentially in j; iii) in scenario $\mathbf{C}_J^{(3)}$, the values $C_j^{(3)}$'s decay linearly in j; iv) in scenario $\mathbf{C}_J^{(4)}$, the values $C_j^{(4)}$'s are either nine, five, or one. Additionally, we assume X_{n+1} is mapped into bucket $h^{(i)}(X_{n+1})$ such that $C^{(\mathrm{i})}_{h^{(i)}(X_{n+1})} = 5$ for $i = 1, 2, 3, 4$. We consider a PYP prior with parameter $\gamma = 1$ and parameter $\alpha = 0, 0.1, 0.3, 0.5$; recall that the PYP prior with $\alpha = 0$ and $\gamma > 0$ coincides with the DP prior with parameter γ.

Figure 2 summarizes the posterior distribution of $f_{X_{n+1}}$ in each scenario. Given the sufficiency of $C_{h(X_{n+1})}$ under the DP prior, the corresponding posterior distributions remain unchanged across all scenarios. Moreover, the posterior distributions under the DP prior are the most concentrated on larger values. Across all the scenarios, increasing α pushes the posterior towards lower values. Although not clearly evident from the plots, there are slight differences under the PYP priors across scenarios. Particularly, for $\alpha = 0.3$ and $\alpha = 0.5$, the posterior mass assigned to 5 is larger in the second scenario compared to other settings.

$\mathbf{C}_J^{(1)}$	5	5	5	5	5	5	5	5	5	5
$\mathbf{C}_J^{(2)}$	14	10	7	5	4	3	2	2	2	1
$\mathbf{C}_J^{(3)}$	10	9	8	7	5	4	3	2	1	1
$\mathbf{C}_J^{(4)}$	9	9	9	5	5	5	5	1	1	1

Table 1: Empirical frequencies of the j-th bucket (by column) in 4 simulated scenarios.

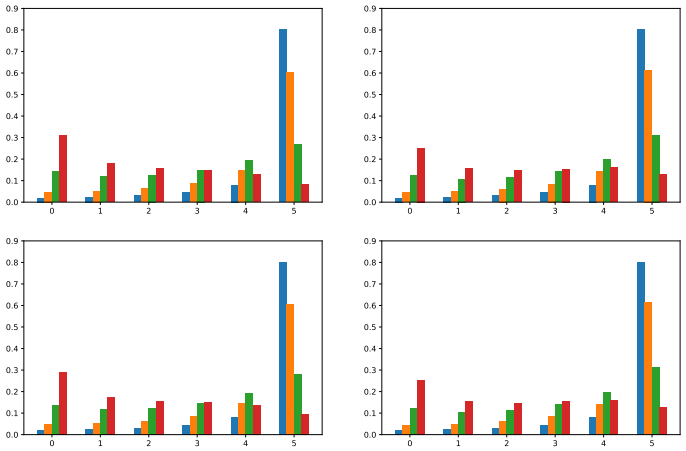


Figure 2: Posterior distribution of $f_{X_{n+1}}$ in the four scenarios of Table 1. The blue, orange, green, and red bars correspond to a PYP(α, γ) prior, with α equal to 0, 0.1, 0.3, and 0.5 respectively, and $\gamma = 1$. Since the DP posterior ($\alpha = 0$) concentrates more of its probability mass at 5, the results suggest the DP prior leads to a more accurate estimate across these 4 scenarios.

3. BNP cardinality recovery

3.1 Setup and overview

In the "species" setting described in Section 2, we tackle the problem of estimating the cardinality k_n of $(x_1, \ldots, x_n)$ and the l-cardinality $m_{l,n}$ of $(x_1, \ldots, x_n)$, for any $l \in [n]$, based on the data sketch $\mathbf{C}_J$. Estimating the $m_{l,n}$'s is a refinement of the problem of estimating k_n, as the estimates of the $m_{l,n}$'s imply an estimate of k_n, given by

$$k_n = \sum_{l=1}^{n} m_{l,n}, \qquad\qquad n = \sum_{l=1}^{n} l \cdot m_{l,n}.$$

Recovering the l-cardinalities implies recovering the partition structure of $(x_1, \ldots, x_n)$, which is the partition of $\{1, \ldots, n\}$ induced by the equivalence relation $i \sim j \iff x_i = x_j$; this is a sufficient statistic for the sample $(x_1, \ldots, x_n)$. Under the model (2), for the DP and the PYP priors, we demonstrate that recovering $m_{l,n}$ boils down to computing the posterior distribution of $f_{X_{n+1}}$ given $\mathbf{C}_J$. Thus, the BNP approach connects frequency and cardinality recovery, enabling the estimation of $f_{X_{n+1}}$ and k_n from the same sketch $\mathbf{C}_J$. Our approach relies on a distinctive feature of the DP and PYP priors, known as the "sufficiency" postulate (Bacallado et al., 2017), and does not extend to other PK priors. Below, we compute estimates of k_n and of the $m_{l,n}$'s in terms of posterior expectations, given $\mathbf{C}_J$.

3.2 Background on the PYP prior

Before stating our main results, we recall the distribution of a random sample $\mathbf{X}_n = (X_1, \ldots, X_n)$ from $P \sim \text{PYP}(\alpha, \gamma, G_0)$. Due to the (almost sure) discreteness of P, the random sample $\mathbf{X}_n$ induces a random partition of $[n]$ into $K_n = k \leq n$ blocks, denoted as $\{S_1^*, \ldots, S_{K_n}^*\}$, with frequencies $(N_{1,n}, \ldots, N_{K_n,n}) = (n_1, \ldots, n_k)$ such that $n_i > 0$ and $\sum_{1 \leq i \leq k} n_i = n$. Let $M_{l,n}$ be the number of blocks with frequency $l \in [n]$, i.e.,

$$M_{l,n} = \sum_{i=1}^{K_n} I(N_{i,n} = l),$$

so that $\sum_{l=1}^{n} M_{l,n} = K_n$ and $\sum_{l=1}^{n} l M_{l,n} = n$, and let $\mathbf{M}_n = (M_{1,n}, \ldots, M_{n,n})$. If we set

$$\mathcal{M}_n = \left\{ (m_1, \ldots, m_n) : m_i \geq 0, \sum_{l=1}^{n} m_l = k, \sum_{l=1}^{n} l m_l = n \text{ and } k \in [n] \right\},$$

then, for $\mathbf{m} \in \mathcal{M}_n$,

$$\Pr[\mathbf{M}_n = \mathbf{m}] = n! \frac{\left(\frac{\gamma}{\alpha}\right)_{(\sum_{l=1}^{n} m_l)}}{(\gamma)_{(n)}} \prod_{l=1}^{n} \left(\frac{\alpha(1-\alpha)_{(l-1)}}{l!} \right)^{m_l} \frac{1}{m_l!}. \tag{10}$$

The distribution in (10) first appeared in Pitman (1995, Proposition 9). Assuming $P \sim \text{DP}(\theta, G_0)$, the distribution of $\mathbf{M}_n$ is obtained from (10) by setting $\gamma = \theta$ and letting $\alpha \to 0$.

An application of (10) yields the conditional distribution of X_{n+1} given $\mathbf{X}_n$, known as the predictive distribution or generative scheme of the PYP prior, and also of the DP prior by letting $\alpha \to 0$ (Pitman, 2006, Chapter 3). In particular, for any $l \in [n]$, let

$$\mathcal{S}_0 = \mathbb{S} - \{S_1^*, \ldots, S_{K_n}^*\}, \qquad \mathcal{S}_l = \bigcup_{i=1}^{K_n} \{S_i^* \in \{S_1^*, \ldots, S_{K_n}^*\} : N_{i,n} = l\}.$$

Above, $\mathcal{S}_0$ is the set of "new" symbols not observed in $\mathbf{X}_n$, while $\mathcal{S}_r$ is the set of "old" symbols observed in $\mathbf{X}_n$ with frequency r. From Pitman (1995, Proposition 9),

$$\Pr[X_{n+1} \in \mathcal{S}_l \,|\, \mathbf{X}_n] = \begin{cases} \frac{\gamma + k\alpha}{\gamma + n}, & \text{if } l = 0, \\ \frac{m_l(l-\alpha)}{\gamma+n}, & \text{if } l \geq 1. \end{cases} \tag{11}$$

The PYP prior is the sole PK prior for which the probability that X_{n+1} is a "new" symbol depends on $\mathbf{X}_n$ only through K_n, and the probability that X_{n+1} is an "old" symbol with frequency l depends on $\mathbf{X}_n$ only through $M_{l,n}$. Further, the DP prior is the sole PK prior for which the probability that X_{n+1} is "new" depends on $\mathbf{X}_n$ only through n. These are known as the "sufficientness" postulates of the DP and PYP priors (Bacallado et al., 2017).

3.3 Main result

Our BNP approach integrates the results of Section 2, under both the DP and PYP priors, with the predictive distribution (11). We focus here on the PYP prior, noting that the

results for the DP follow by setting $\gamma = \theta$ and letting $\alpha \to 0$. Before presenting the main results, we outline the key arguments of our approach. In particular, for any $l \in [n]$,

$$\Pr[f_{X_{n+1}} = l \,|\, \mathbf{C}_J = \mathbf{c}] = \sum_{\mathbf{m} \in \mathcal{M}_n} \Pr[X_{n+1} \in \mathcal{S}_l \,|\, \mathbf{C}_J = \mathbf{c},\, \mathbf{M}_n = \mathbf{m}] \Pr[\mathbf{M}_n = \mathbf{m} \,|\, \mathbf{C}_J = \mathbf{c}], \tag{12}$$

where

$$\Pr[X_{n+1} \in \mathcal{S}_l \,|\, \mathbf{C}_J = \mathbf{c},\, \mathbf{M}_n = \mathbf{m}] = \Pr[X_{n+1} \in \mathcal{S}_l \,|\, \mathbf{M}_n = \mathbf{m}] = \frac{m_l(l-\alpha)}{\gamma + n}, \tag{13}$$

with the last identity in (13) derived from (11). By combining (12) with (13), we obtain:

$$\Pr[f_{X_{n+1}} = l \,|\, \mathbf{C}_J = \mathbf{c}] = \frac{l-\alpha}{\gamma+n} \mathbb{E}[M_{l,n} \,|\, \mathbf{C}_J = \mathbf{c}], \qquad l \in [n]. \tag{14}$$

This connects the posterior distribution of $f_{X_{n+1}}$, given $\mathbf{C}_J$, to the corresponding conditional expectation of $M_{l,n}$. This is the identity underpinning our approach to cardinality recovery, effectively linking it to the frequency recovery problem discussed in Section 2.

From (14), a BNP estimator for $m_{l,n}$ under a squared loss can be immediately derived. This is the conditional expectation of $M_{l,n}$, given $\mathbf{C}_J$, which is given for any $l \in [n]$ by

$$\hat{m}_{l,n} = \mathbb{E}[M_{l,n} \,|\, \mathbf{C}_J = \mathbf{c}] = \frac{\gamma+n}{l-\alpha} \Pr[f_{X_{n+1}} = l \,|\, \mathbf{C}_J = \mathbf{c}]. \tag{15}$$

Then, since $K_n = \sum_{l=1}^n M_{l,n}$, a BNP estimator for k_n under a squared loss is of the form

$$\hat{k}_n = \mathbb{E}[K_n \,|\, \mathbf{C}_J = \mathbf{c}] = \sum_{l=1}^n \frac{\gamma+n}{l-\alpha} \Pr[f_{X_{n+1}} = l \,|\, \mathbf{C}_J = \mathbf{c}]. \tag{16}$$

Both (15) and (16) are based on (14), which in turn depends on (13) relying only on $M_{l,n}$. Since the PYP prior is the sole PK prior for which (13) depends on $\mathbf{X}_n$ only through $M_{l,n}$, this approach cannot be extended to other PK priors. To conclude, we compute,

$$\Pr[f_{X_{n+1}} = l \,|\, \mathbf{C}_J = \mathbf{c}] = \sum_{j=1}^J \Pr[f_{X_{n+1}} = l \,|\, \mathbf{C}_J = \mathbf{c}, h(X_{n+1}) = j] \Pr[h(X_{n+1}) = j \,|\, \mathbf{C}_j = \mathbf{c}], \tag{17}$$

for any $l \in [n]$, where the distribution of $f_{X_{n+1}}$ conditional on $\mathbf{C}_j$ and $h(X_{n+1})$ is given in (8). We next provide an expression for (17) that simplifies the estimators (15) and (16).

Corollary 1. *Let $\mathbf{C}_J$ be a sketch of $\mathbf{X}_n$ under (2), with $P \sim PYP(\alpha, \gamma, G_0)$. For $l \in [n]$,*

$$\begin{aligned} &Pr[f_{X_{n+1}} = l \,|\, \mathbf{C}_J = \mathbf{c}] \\ &= \frac{\gamma}{J} \frac{(1-\alpha)_{(l)}}{\gamma+n} \sum_{j=1}^J \binom{c_j}{l} \frac{\sum_{\boldsymbol{i} \in \mathcal{S}(\mathbf{c},j,-l)} \frac{\Gamma\left(\frac{\gamma+\alpha}{\alpha} + |\boldsymbol{i}|\right)}{J^{|\boldsymbol{i}|}} \prod_{k=1}^J \mathscr{C}(c_k - l\delta_{k,j}, i_k; \alpha)}{\sum_{\boldsymbol{i} \in \mathcal{S}(\mathbf{c},j,0)} \frac{\Gamma\left(\frac{\gamma}{\alpha} + |\boldsymbol{i}|\right)}{J^{|\boldsymbol{i}|}} \prod_{k=1}^J \mathscr{C}(c_k, i_k; \alpha)}. \end{aligned} \tag{18}$$

See Appendix A.8 for the proof of Corollary 1. The BNP estimators for l-cardinality and cardinality are derived by combining (18) with (15) and (16), respectively. Note that, under the PYP prior, these estimators face the same computational challenges as those discussed in Section 2 for the frequency estimation problem.

By contrast, under the DP prior, the estimators simplify significantly. By setting $\gamma = \theta$ and letting $\alpha \to 0$ in (18), we prove in Appendix A.9 that, for any $l \in [n]$,

$$\Pr[f_{X_{n+1}} = l \,|\, \mathbf{C}_J = \mathbf{c}] = \frac{\frac{\theta}{J}}{\theta + n} \sum_{j=1}^{J} \frac{(c_j - l + 1)_{(l)}}{\left(\frac{\gamma}{J} + c_j - l\right)_{(l)}}. \tag{19}$$

By combining (19) with (15) and (16), with $\gamma = \theta$ and $\alpha = 0$, for any $l \in [n]$:

$$\hat{m}_{l,n} = \frac{\frac{\theta}{J}}{l} \sum_{j=1}^{J} \frac{(c_j - l + 1)_{(l)}}{\left(\frac{\gamma}{J} + c_j - l\right)_{(l)}} \qquad \hat{k}_n = -\theta\psi\left(1 - \frac{\theta}{J}\right) + \frac{\theta}{J} \sum_{j=1}^{J} \psi\left(1 - \frac{\theta}{J} - c_j\right),$$

where ψ is the digamma function, i.e. $\psi(x) = \frac{d}{dx} \log \Gamma(x)$. These estimates for k_n and $m_{l,n}$ depend on the prior parameter θ, which can be estimated from $\mathbf{C}_J$ as discussed in Section 2.

4. BNP frequency recovery in the "traits" setting

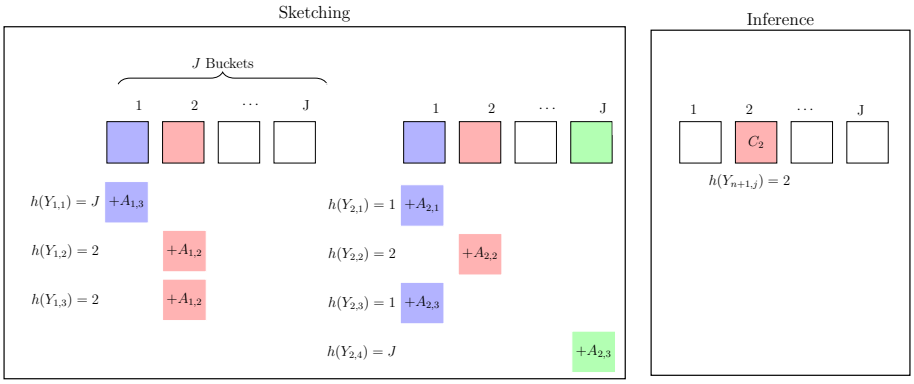
The "traits" setting for frequency recovery extends the "species" setting studied in Section 2 by allowing the data points to exhibit nonnegative integer levels of association with multiple symbols. We consider $n \geq 1$ data points $(x_1, \ldots, x_n)$ modeled as a random sample $\mathbf{X}_n = (X_1, \ldots, X_n)$, where each X_i is represented as $X_i = ((\tilde{Y}_{i,j}, \tilde{A}_{i,j}), j = 1, \ldots, K_i)$. Here, $\tilde{A}_{i,j} \in \mathbb{N}_0$ represents the level of association of the i-th sample with the trait $\tilde{Y}_{i,j} \in \mathbb{S}$. Sketching $\mathbf{X}_n$ is conducted at the trait level using a random hash function $h : \mathbb{S} \to \{1, \ldots, J\}$ from the hash family $\mathcal{H}_J$. This function assigns each $\tilde{Y}_{i,j}$ to a bucket, incrementing the bucket's counter $C_{h(\tilde{Y}_{i,j})}$ by $\tilde{A}_{i,j}$, see Figure 3 for a schematic visualization. Consequently, the sketch $\mathbf{C}_J = (C_1, \ldots, C_J)$ captures the total association levels for the traits hashed into each bucket. For a new data point X_{n+1}, we define the empirical frequency level of a trait $Y_{n+1,r}$ as:

$$f_{Y_{n+1,r}} = \sum_{i=1}^{n} \sum_{j=1}^{K_i} \tilde{A}_{i,j} I(\tilde{Y}_{i,j} = Y_{n+1,r}). \tag{20}$$

We make use of the sketch $\mathbf{C}_J$ to develop a BNP approach to estimate $f_{Y_{n+1,r}}$, assuming that $\mathbf{X}_n$ is sampled from a generalized Indian buffet process (James, 2017). Under this model, the frequency levels of traits are sufficient statistics, thus paralleling the role of species frequencies in our earlier BNP model for species data in Section 2.

Topic modeling is a prominent application of "traits" allocations, particularly through the use of a multinomial naive Bayes classifier to categorize documents into topics. Suppose each data point (X_i, T_i) consists of a document and a topic label $T_i \in \{1, \ldots, M\}$. If n_j is the number of documents in each topic, the naive Bayes classification rule is given by:

$$\Pr[T_i = m \,|\, X_i = \{(y_{i,j}, a_{i,j})\}_{j=1}^{K_i}] \propto \Pr[T_i = m] \prod_{j=1}^{K_i} (\pi_{y_{i,j}}^{m})^{a_{i,j}}, \tag{21}$$



(a) Sketch update with traits of X_1. (b) Sketch update with traits of X_2. (c) Frequency query for trait $Y_{n+1,j}$.

Figure 3: Data sketching in the "traits" setting. Each X_i, for $i \in \{1, \ldots, n\}$, is defined by K_i traits $(Y_{i,1}, \ldots, Y_{i,K_i})$ and exhibits an association level $A_{i,j}$ with each $Y_{i,j}$. A hash function maps each trait into one of J possible "buckets" (shown in different colors), and the corresponding counters are incremented by one. The recovery problem is to estimate the empirical frequency of a "trait" $X_{Y+1,j}$ based on the information contained in the sketch and the bucket assignments for $Y_{n+1,j}$.

where $\pi^m_{y_{i,j}}$ is the relative frequency of the trait $y_{i,j}$ in documents with topic m. For sketched data, $\pi^m_{y_{i,j}}$ can be replaced by the Bayesian estimator $\hat{f}^m_{y_{i,j}}/n_j$. Another application is in feature engineering, computing sketched "tf-idfs". In information retrieval, tf-idf is a preprocessing to adjust the $\tilde{A}_{i,k}$'s based on the frequency of documents with trait $\tilde{Y}_{i,k}$:

$$\frac{\tilde{A}_{i,k}}{\sum_{j=1}^{K_i} \tilde{A}_{i,j}} \log \frac{n}{\sum_{i=1}^{n} I(\tilde{Y}_{i,k} \in X_i)}. \tag{22}$$

It is known that tf-idfs can replace raw frequencies in (21) (Rennie et al., 2003). One may consider a sketched version of (22), where the counter is incremented by one instead of $\tilde{A}_{i,k}$, so that $f_{Y_{n+1},r}$ is the number of documents containing the word Y_{n+1}. Lastly, Zhou et al. (2016) combines elements from the naive Bayes approach with a BNP model for "traits" allocations, further illustrating the adaptability and potential of these methods.

4.1 BNP model and main result

We recall the definition of the generalized Indian buffet process; see also Broderick et al. (2015), Broderick et al. (2018).

Definition 5. *Let $\tilde{\mu} \sim CRM(\theta, \rho, G_0)$, that is $\tilde{\mu} = \sum_{k\geq 1} J_k \delta_{W_k}$, and let G_A be a probability mass function over $\mathbb{N}_0$. We say that a random variable X given $\tilde{\mu}$ is distributed as a generalized Indian buffet process with parameter G_A if $X = \sum_{k\geq 1} A_k \delta_{W_k}$, where $(A_k)_{k\geq 1}$ is independent of $(W_k)_{k\geq 1}$ and such that $A_k \,|\, J_k \sim G_A(J_k)$ for $k \geq 1$.*

We write $X \,|\, \tilde{\mu} \sim \text{IBP}(G_A \,|\, \tilde{\mu})$ to denote that X is distributed according to a generalized Indian buffet process with parameter G_A. While X represents an (infinite) random measure, it is assumed that only a finite number of the $A_{i,k}$'s are non-zero. The representation of X as a collection of displayed traits with their corresponding levels of association follows by

letting $\{(\tilde{Y}_j, \tilde{A}_j)\}_j = \{(W_k : A_k > 0, A_k)\}_k$. Accordingly, we write the BNP model as:

$$\begin{aligned} C_j &= \sum_{i=1}^{n} \sum_{k \geq 1} A_{i,k} I(h(W_k) = j), \qquad j \in [J], \\ h &\sim \mathcal{H}_J, \\ X_1, \ldots, X_n \,|\, \tilde{\mu} &\overset{\text{iid}}{\sim} \text{IBP}(G_A \,|\, \tilde{\mu}), \\ \tilde{\mu} &\sim \text{CRM}(\theta, \rho, G_0). \end{aligned} \tag{23}$$

Under this model (23),

$$f_{Y_{n+1,r}} = \sum_{i=1}^{n} X_i(Y_{n+1,r})$$

is the total weighted frequency from equation (20). The sketch $\mathbf{C}_J$ is derived through random hashing of $\mathbf{X}_n$, with each bucket j containing the hashed traits of $\mathbb{S}$ where $h(\omega) = j$, denoted as $D_j = h^{-1}(j) = \{\omega \in \mathbb{S} : h(\omega) = j\}$. For a new data point X_{n+1}, let $Y_{n+1,r}$ be the trait belonging to X_{n+1} whose frequency we are interested in, and let $B_j = X_{n+1}(D_j)$ be the increment to the j-th bucket of the sketch given by X_{n+1}, for $j \in [J]$. Let $\mathbf{B} = (B_1, \ldots, B_J)$.

In the "species" setting, the posterior distribution (4) represented the probability that X_{n+1} appeared $l+1$ times conditional on the augmented sketch $(C_1, \ldots, C_j + 1, \ldots, C_J)$ and $h(X_{n+1}) = j$. This has a natural counterpart in the "traits" setting, namely

$$\Pr\left[f_{Y_{n+1,r}} = l, X_{n+1}(Y_{n+1,r}) = a \,|\, h(Y_{n+1,r}) = j, \mathbf{C}_J = \mathbf{c}_J, \mathbf{B} = \mathbf{b}\right], \tag{24}$$

from which the posterior distribution of $f_{Y_{n+1,r}}$, given $\mathbf{C}_J$, $\mathbf{B}$, and $h(Y_{n+1,r})$, follows by Bayes' theorem. The next proposition establishes that the pair $(C_{h(Y_{n+1,r})}, B_{h(Y_{n+1,r})})$ is a sufficient statistic for the estimation of $f_{Y_{n+1,r}}$, with respect to $(h(Y_{n+1,r}), \mathbf{C}_J, \mathbf{B})$.

Proposition 6. *Let $\mathbf{C}_J$ be a sketch of $\mathbf{X}_n$ under the model (23). Assuming X_{n+1} is an additional, unobserved random sample, for $l = 0, \ldots, c$, the following holds:*

$$\begin{aligned} &\Pr\left[f_{Y_{n+1,r}} = l, X_{n+1}(Y_{n+1,r}) = a \,|\, h(Y_{n+1,r}) = j, \mathbf{C}_J = \mathbf{c}_J, \mathbf{B} = \mathbf{b}\right] \\ &\quad = \Pr\left[f_{Y_{n+1,r}} = l, X_{n+1}(Y_{n+1,r}) = a \,|\, h(Y_{n+1,r}) = j, C_j = c, B_j = b\right]. \end{aligned} \tag{25}$$

See Appendix A.10 for the proof of Proposition 6. The role of Proposition 6 is to highlight an interesting distinction between the "species" and "traits" settings in frequency recovery. In the "species" setting, Theorem 3 establishes that the DP prior is the sole prior for which the posterior distribution of $f_{X_{n+1}}$, given $\mathbf{C}_J$ and $h(X_{n+1})$, matches the distribution given just $C_{h(X_{n+1})}$. Conversely, in the "traits" setting, Proposition 6 demonstrates that for any CRM prior, the posterior distribution of $f_{Y_{n+1,r}}$, given $\mathbf{C}_J$, $\mathbf{B}$, and $h(Y_{n+1,r})$, aligns with that given $C_{h(Y_{n+1,r})}$ and $B_{h(Y_{n+1,r})}$. This interesting phenomenon indicates that all CRM priors in the "traits" setting satisfy a sufficiency property analogous to the DP prior in the "species" setting. The following theorem provides an expression for (25).

Theorem 7. *Let $\mathbf{C}_J$ represent a sketch of $\mathbf{X}_n$ under the model (23) with X_{n+1} being an additional (unobservable) random sample. For any $l \in \{0, \dots, c\}$,*

$$\begin{aligned}
&\Pr\left[f_{Y_{n+1,r}} = l, X_{n+1}(Y_{n+1,r}) = a \,|\, h(Y_{n+1,r}) = j, C_j = c, B_j = b\right] \\
&\quad = \frac{\theta}{J} \frac{\Pr\left[\sum_{k\geq 1}\sum_{i=1}^{n} A'_{i,k} = c - l, \sum_{k\geq 1} A'_{n+1,k} = b - a\right]}{\Pr\left[\sum_{k\geq 1}\sum_{i=1}^{n} A'_{i,k} = c, \sum_{k\geq 1} A'_{n+1,k} = b\right]} \\
&\quad \times \int_{\mathbb{R}_+} \Pr\left[\sum_{i=1}^{n} \tilde{A}_i = l, \tilde{A}_{n+1} = a \,|\, s\right] \rho(s)\mathrm{d}s,
\end{aligned} \tag{26}$$

where $\tilde{A}_1 \dots, \tilde{A}_{n+1} \,|\, s \overset{\text{iid}}{\sim} G_A(\cdot \,|\, s)$ and the A'_{ik}'s are the projection on the second coordinate of the points of $N' = \{(J'_k, (A'_{i,k})_{i=1}^{n+1})\}_{k\geq 1}$, which is a Poisson process on $\mathbb{R}_+ \times \mathbb{N}_0^{n+1}$ with Lévy intensity $(\theta/J)[\prod_{1\leq 1\leq n+1} G_A(\mathrm{d}a_i \,|\, s)]\rho(s)\mathrm{d}s$.

See Appendix A.11 for a proof of Theorem 7. The application of Theorem 7 necessitates defining the CRM prior via its Lévy intensity and choosing a distribution G_A for the trait association levels. Below, we apply Theorem 7 with G_A modeled as a Poisson distribution, which simplifies (26), and then we consider modeling G_A as a Bernoulli distribution.

4.2 Results under G_A Poisson

Let $G_A(\cdot \,|\, J_k)$ be the Poisson distribution with parameters λJ_k, for fixed $\lambda > 0$. Leveraging the closeness under convolution of Poisson distributions, we can simplify (26) considerably.

Proposition 8. *Let $\mathbf{C}_J$ be a sketch of $\mathbf{X}_n$ under the model (23) with $G_A(J_k)$ being the Poisson distribution with parameter rJ_k for a fixed $r > 0$, and let X_{n+1} be an additional (unobservable) random sample. Further, let ψ be the function defined in (3), and let $\phi^{(n)}(u) = (-1)^n \frac{\mathrm{d}^n}{\mathrm{d}u^n} e^{-\theta/J\psi(u)}$ and $\kappa(u, n) = \int_{\mathbb{R}_+} e^{-un} s^n \rho(s)\mathrm{d}s$. Then, for $l = 0, \dots, c$,*

$$\begin{aligned}
&\Pr\left[f_{Y_{n+1,r}} = l, X_{n+1}(Y_{n+1,r}) = a \,|\, h(Y_{n+1,r}) = j, C_j = c, B_j = b\right] \\
&\quad = \frac{\theta}{J}\binom{c}{l}\binom{b}{a} \frac{\phi^{(c-l+b-a)}((n+1)\lambda)}{\phi^{(c+b)}((n+1)\lambda)} \kappa(l + a, (n+1)\lambda).
\end{aligned} \tag{27}$$

See Appendix A.12 for a proof of Proposition 8. We further specialize Proposition 8 by assuming that $\tilde{\mu}$ is a Gamma CRM, i.e., $\rho(s) = s^{-1}\exp\{-s\}$. This simplifies (27) to:

$$\begin{aligned}
&\Pr\left[f_{Y_{n+1,r}} = l, X_{n+1}(Y_{n+1,r}) = a \,|\, h(Y_{n+1,r}) = j, C_j = c, B_j = b\right] \\
&\quad = \frac{\theta}{J}\binom{c}{l}\binom{b}{a}(l + a - 1)! \frac{\Gamma(\theta/J + c + b - l - a)}{\Gamma(\theta/J + c + b)}.
\end{aligned} \tag{28}$$

See Appendix B for the proof of (28). As a generalization of the Gamma CRM, we consider the generalized Gamma CRM, i.e., $\rho(s) = \alpha\Gamma(1-\alpha)^{-1}s^{-\alpha-1}e^{-\tau s}$ for $\alpha \in [0, 1)$ and $\tau > 0$ (Brix, 1999); see also Pitman (2003) for details. This simplifies (27) to:

$$\begin{aligned}
&\Pr\left[f_{Y_{n+1,r}} = l, X_{n+1}(Y_{n+1,r}) = a \,|\, h(Y_{n+1,r}) = j, C_j = c, B_j = b\right] \\
&\quad = \frac{\theta}{J}\binom{c}{l}\binom{b}{a} \frac{\alpha(1-\alpha)_{(l+a-1)}}{(\tau + (n+1)r)^{-\alpha+l+a}} \frac{\sum_{i=1}^{c-l+b-a}\left(\frac{\theta}{J}\right)^i \frac{\mathscr{C}(c-l+b-a,i;\alpha)}{(\tau+(n+1)\lambda)^{c-l+b-a-\alpha i}}}{\sum_{i=1}^{c+b}\left(\frac{\theta}{J}\right)^i \frac{\mathscr{C}(c+b,i;\alpha)}{(\tau+(n+1)\lambda)^{c+b-\alpha i}}}.
\end{aligned} \tag{29}$$

See Appendix B for the proof of (29). Both (28) and (29) are easy to evaluate.

As explained in Section 2, applying (28) or (29) necessitates estimating the unknown parameters λ and the prior parameters from the sketch $\mathbf{C}_J$. Given the convolution properties of the Poisson distribution, the distribution of $\mathbf{C}_j$ under the model (23), where $G_A(J_k)$ is a Poisson(λJ_k), corresponds to the distribution of the following hierarchical model:

$$\begin{aligned} C_j \,|\, T'_j &\overset{\text{ind}}{\sim} \text{Poi}(n\lambda T'_j), \\ T'_j &\overset{\text{ind}}{\sim} f_{D_j}(\theta, \rho, G_0), \end{aligned} \tag{30}$$

where $f_{D_j}(\theta, \rho, G_0)$ is the distribution of $\tilde{\mu}(D_j)$. If $\tilde{\mu}$ is modeled according to a Gamma CRM, then f_{D_j} follows a Gamma distribution with parameter $(\theta/J, 1)$. Consequently,

$$\Pr[\mathbf{C}_J = \mathbf{c}_J] = \frac{(n\lambda)^n}{(1+n\lambda)^{\theta+n}} \prod_{j=1}^{J} \frac{\left(\frac{\theta}{J}\right)_{(c_j)}}{c_j!}, \tag{31}$$

which allows estimating θ and λ by maximizing the marginal likelihood (31). If $\tilde{\mu}$ is a generalized Gamma CRM, the distribution of f_{D_j} is not conjugate to the Poisson distribution. This prevents obtaining a closed-form distribution for the sketch $\mathbf{C}_J$. However, we can apply a likelihood-free method similar to that discussed for the PYP prior in Section 2.

4.2.1 Comparisons under the Gamma and generalized Gamma process priors

We compare the posterior distributions of $f_{Y_{n+1,r}}$ under the Gamma CRM and generalized Gamma CRM priors, with $G_A(\cdot \,|\, J_k)$ modeled as a Poisson distribution with parameter λJ_k, for fixed $\lambda > 0$. Notably, when $a = b = 1$, the posterior distribution under the Gamma CRM prior coincides with the posterior distribution obtained under the DP prior, which is displayed in (6). In other cases, the posterior distribution of $f_{Y_{n+1,r}}$ incorporates additional information from $X_{n+1}(Y_r) = a$ and $X_{n+1}(D_{h(Y_{n+1,r})}) = b$, providing details on the relative frequency of trait $Y_{n+1,r}$ within its corresponding bucket. Figure 4 contrasts the posterior distributions under Gamma CRM and generalized Gamma CRM priors, illustrating significant differences influenced by the specific prior settings, especially for small values of $X_{n+1}(Y_{n+1,r})$. Moreover, for a given $X_{n+1}(D_{h(Y_{n+1,r})}) = b$, values of $X_{n+1}(Y_r)$ close to b push the posterior distribution to large values, reflecting the prevalence of the trait $Y_{n+1,r}$ within that bucket. Conversely, values of $X_{n+1}(Y_r)$ considerably smaller than b indicate the rarity of the trait within the bucket, driving the posterior of $f_{Y_{n+1,r}}$ towards lower values.

4.3 Results under G_A Bernoulli

Let $G_A(\cdot \,|\, J_k)$ be the Bernoulli distribution with parameter J_k. Under this assumption, the levels of associations of traits simply refer to their presence or absence. For this reason, traits are typically referred to as features. The conditional distribution of the random variable $S_n = \sum_{k\geq 1} \sum_{i=1}^{k} A'_{i,k}$ in (26), given $\tilde{\mu}$, is the distribution of the sum of independent Bernoulli random variables with parameters $J'_1, \ldots, J'_1, J'_2, \ldots, J'_2, \ldots$, where each J'_k appears exactly n times; this is the Poisson-Binomial distribution (Chen and Liu, 1997). Similarly, $Z = \sum_{k\geq 1} A'_{n+1,k}$ follows a Poisson-Binomial distribution with parameter $J'_1, J'_2, \ldots$. While the Poisson-Binomial distribution has a complicated expression, one may combine Le Cam's

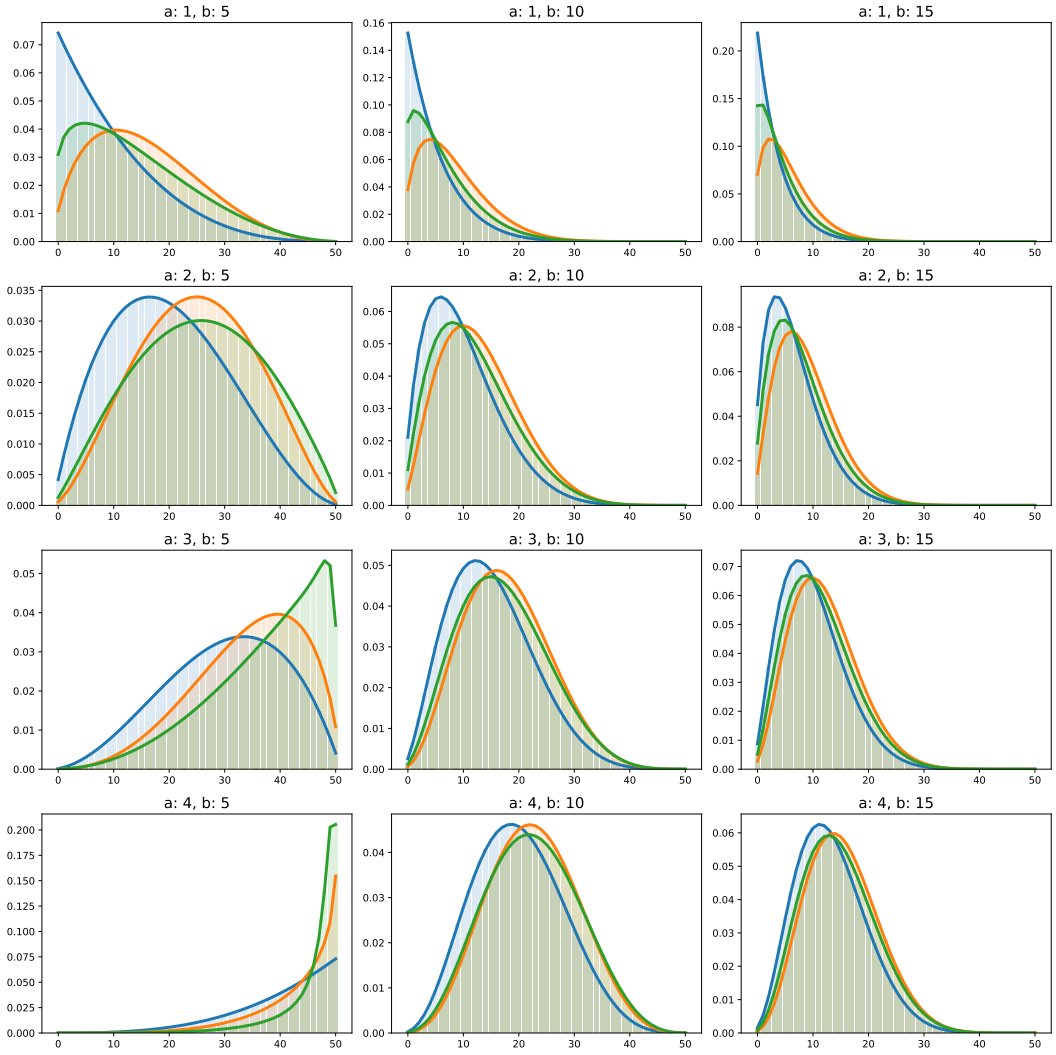


Figure 4: Posterior distribution of the empirical frequency level $f_{Y_{n+1,r}}$, assuming the Poisson setting with a Gamma CRM prior (blue line) and with a generalized Gamma CRM prior (orange and green lines). In all the panels, a and b vary, while we fix $c = 50$, $m = 1000$, $J = 50$, $\theta = 0.3$, $\tau = 1$, $\lambda = 1$, and $\alpha = 0.25, 0.75$ for the orange and green lines respectively.

theorem (Le Cam, 1960) with the results in Section 4.2 to approximate (26) in the Bernoulli setting. For notation's sake, considering S_n and Z as previously defined, we introduce $\tilde{S}_n$ and $\tilde{Z}$ such that, given $T' = \sum_{k\geq 1} J'_k$, $\tilde{S}_n \,|\, T' \sim \text{Poi}(nT')$ and $\tilde{Z} \,|\, T' \sim \text{Poi}(T')$. We first observe that the posterior distribution of $f_{Y_{n+1,r}}$ in (26) is proportional to

$$\begin{aligned}&\Pr\left[f_{Y_{n+1,r}} = l, X_{n+1}(Y_{n+1,r}) = 1 \,|\, h(Y_{n+1,r}) = j, C_j = c, B_j = b\right] \\ &\quad \propto \Pr\left[f_{Y_{n+1,r}} = l, X_{n+1}(Y_{n+1,r}), S_n = c - l, Z = b - 1\right].\end{aligned}$$

In the next theorem, we provide an approximation of the distribution of $f_{Y_{n+1,r}}$ by replacing S_n and Z with $\tilde{S}_n$ and $\tilde{Z}$ respectively. We also estimate the approximation error.

Theorem 9. *Let $\mathbf{C}_J$ be a sketch of $\mathbf{X}_n$ under the model (23) with $G_A(J_k)$ being the Bernoulli distribution with parameter J_k, and let X_{n+1} be an additional (unobservable) random sample. Furthermore, let ψ be the function defined in (3), and let $\phi^{(n)}(u) = (-1)^n \frac{\mathrm{d}}{\mathrm{d}u} e^{-\theta/J\psi(u)}$ and $\kappa(u,n) = \int_{\mathbb{R}_+} e^{-un} s^n \rho(s) \mathrm{d}s$. Then, the posterior distribution of $f_{Y_{n+1,r}}$ can be approximated by the distribution of the random variable $\tilde{f}_{Y_{n+1,r}}$ such that, for $l = 0, \dots, c$:*

$$\begin{aligned} &\Pr\left[\tilde{f}_{Y_{n+1,r}} = l, X_{n+1}(Y_{n+1,r}) = 1\right] \\ &\quad = \Pr\left[f_{Y_{n+1,r}} = l, X_{n+1}(Y_{n+1,r}) = 1 \,|\, \tilde{S}_n = c - l, \tilde{Z} = b - 1\right] \\ &\quad \propto (c - l + 1)_{(l)} \binom{n}{l} \phi^{(c+b-l-1)}(n+1) \int (s^{l+1} - s^{n+1}) \rho(s) \mathrm{d}s. \end{aligned} \tag{32}$$

Moreover, the total variation distance between the random vectors $[f_{Y_{n+1,r}}, X_{n+1}(Y_{n+1,r}), S_n, Z]$ and $[f_{Y_{n+1,r}}, X_{n+1}(Y_{n+1,r}), \tilde{S}_n, \tilde{Z}]$ is upper bounded by $(2\theta/J) \int_{\mathbb{R}_+} e^{-\psi(u)} \kappa(u, 2) \mathrm{d}u$.

5. Numerical illustrations

5.1 Frequency recovery

5.1.1 Limitations of the DP with heavy-tailed data distributions

To illustrate the limitations of the DP prior, we conducted two simulations with $n = 500,000$ data points, simulated either from a DP with parameters $\theta = 5, 10, 20, 100$ or from a Zipf distribution with tail parameters $c = 1.18, 1.54, 1.82, 2.22$. The Zipf distribution, applicable to infinitely many items, assigns the probability of the k-th item as $k^{-c}/\zeta(c)$, where $\zeta(\cdot)$ denotes the Riemann's zeta function, defined by $\zeta(c) = \sum_{k \geq 1} k^{-c}$.

The DP is known to produce distributions with geometric (light) tails (Teh and Jordan, 2010), which may not adequately capture data with heavier tails behaviours. In contrast, the Zipf distribution exhibits power-law tails, where the parameter c directly influences the decay rate of the tail: lower values of c indicate heavier tails, representing larger fractions of low-frequency items.

We assume model (2) in combination with DP prior for P, whose total mass parameter is estimated from the sketch as discussed in Section 2.5. We let J vary between 100 and $5,000$. As evaluation metric, we follow Dolera et al. (2023) and consider the mean absolute error (MAE) stratified by the true frequency of the tokens. That is, for all distinct symbols s appearing in the original data set, we compute

$$\mathrm{MAE}_m = \frac{1}{\sum_{s \in \mathbb{S}} I(f_s \in (l_m, u_m])} \sum_{s \in \mathbb{S}} |f_s - \hat{f}_s| \, I\left(f_s \in (l_m, u_m]\right),$$

where $f_s = \sum_{i=1}^n I(X_i = s)$ is empirical frequency, $\hat{f}_s$ its estimate, and $(l_m, u_m]_{m \geq 1}$ are non-overlapping frequency bins.

The top row of Figure 5 shows the MAEs when data are generated from a DP with parameter $\theta \in \{5, 10, 20, 100\}$. In this case, the MAEs for low and mid-frequency tokens decrease rapidly with J, especially if θ is small. This is expected since lower values of θ correspond to fewer distinct symbols in the sample, reducing the likelihood of hash collisions.

The results for Zipf-distributed data are displayed in the bottom row of Figure 5. In this case, the MAEs are considerably larger, especially when c is small. This is unsurprising since, as discussed above, smaller values of c correspond to a greater number of distinct symbols in the sample, increasing the likelihood of hash collisions. Moreover, the DP prior is not suited to modeling heavy-tailed data, which is clearly reflected in the very high MAEs associated with the low-frequency tokens.

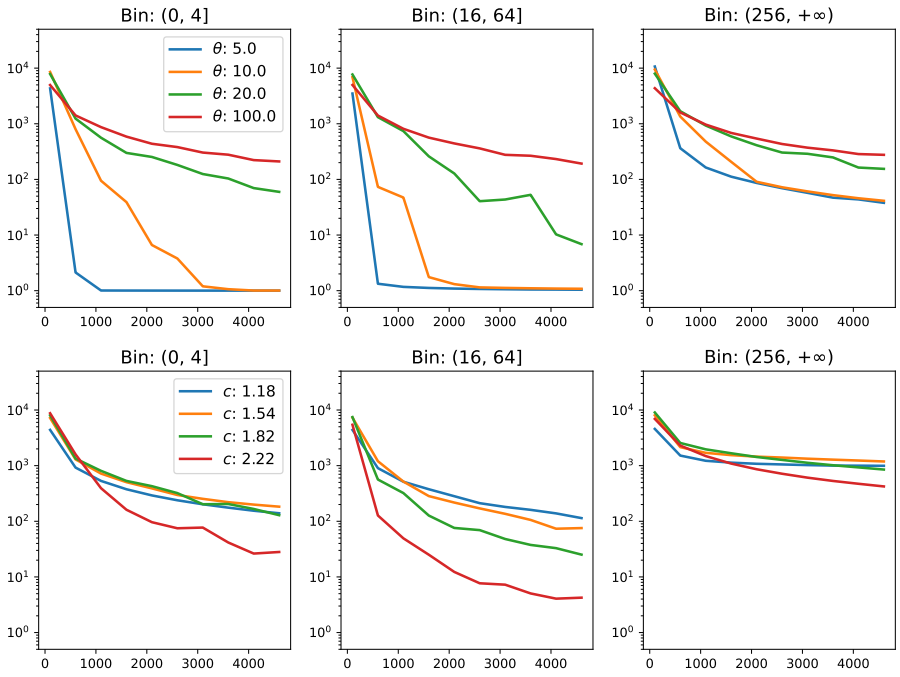


Figure 5: MAEs as a function of J for the frequency recovery problem under a DP prior. Top: data simulated from a DP with parameter θ. Bottom: data from a Zipf distribution with parameter c.

5.1.2 Frequency recovery with the approximate posterior under the PYP

We consider now the same synthetic datasets of Dolera et al. (2023), which consist of $n = 500,000$ samples from the Zipf distribution with parameter $c = 1.18, 1.54, 1.82, 2.22$. We make use of (9) to estimate the frequencies via the posterior expectation, and then we compare estimating the parameters via the likelihood-free approach in Dolera et al. (2023) (specifically, we refer to their estimated values) and our method based on minimizing the recovery error on the first 10,000 samples. The results are displayed in Figure 6.

Note that, to satisfy the asymptotic regime of Theorem 4, we set J to smaller values compared to the previous simulation. It is clear that the proposed estimation method leads to a significant enhancement for the frequency recovery problem. Additionally, employing a PYP prior yields markedly lower MAEs compared to a DP prior, particularly for low frequency tokens. Interestingly, while larger values of the parameter c correspond to better performance under the DP prior, the opposite holds true under the PYP.

This can be explained as follows: when the tail parameter of the Zipf distribution, c, is large, the data generating process exhibits lighter tails and fewer data points contribute

to most of the total mass, commonly referred to as "heavy hitters". When predicting the frequency of these heavy hitters, the asymptotic regime under which (9) is derived may not hold, as c_j may be comparable to n for some j while $c_k \approx 0$ for other $k \neq j$.

A similar scenario might arise if J is large. From (5), it is clear that our estimator applies linear shrinkage to c_j by a term inversely proportional to J. Therefore, if J is large and c_j is not sufficiently large, the estimator over-shrinks the frequency to zero. This becomes evident in our example when $c = 1.82$: in this case, the MAEs associated with low and mid-frequency tokens show an increase as J increases, which can be attributed to the invalidity of the assumptions underlying Theorem 4 for larger values of J.

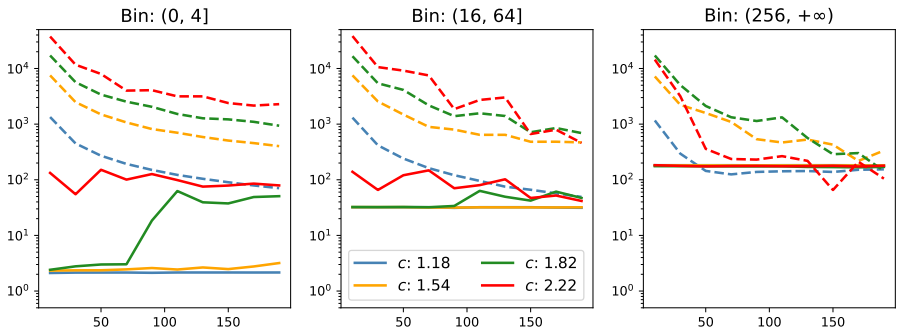


Figure 6: MAEs as a function of J for the frequency recovery problem of Zipf(c) data under a PYP prior. Solid lines correspond to prior's parameters (α, θ) estimated with our method; dashed lines correspond to estimates of (α, θ) obtained as in Dolera et al. (2023).

5.1.3 Application to the Gutenberg corpus

The second data set comprises 18 open-domain classic pieces of English literature from the Gutenberg Corpus (Project Gutenberg, 2022). These data are pre-processed with the same approach of Sesia and Favaro (2022): punctuation and unusual words are removed, retaining only words found in an English dictionary of size 25,487. Subsequently, 1,700,000 consecutive word pairs, or *2-grams*, are extracted. These 2-grams are then sketched using a random hash function, as usual. As shown in Figure 7 (left), these data exhibit a clear power law distribution characterized by numerous bigrams with low frequency.

We compare the frequency estimates obtained under the DP and PYP priors, with the latter utilizing the large-n approximation derived in Theorem 4. Given the large sample size, it is reasonable to expect that such an approximation should be quite accurate in this case. Figure 7 (right) displays the results, averaged over 20 independent hash functions. As anticipated, the PYP outperforms the DP significantly for low and mid-frequency tokens, while their performance is essentially equivalent for very high frequency ones (note that the number of tokens appearing more than 1024 times is less than 100).

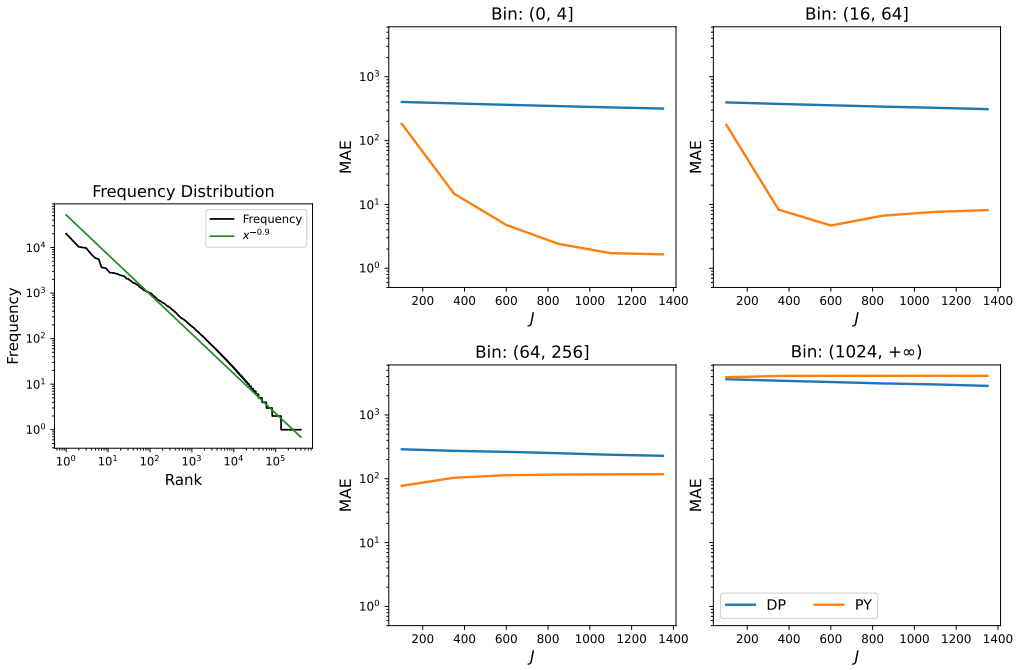


Figure 7: Left: frequency distribution for the Gutenberg corupus' bigrams. Right: MAEs as a function of J for the frequency recovery problem.

5.2 Cardinality recovery

5.2.1 Experiments with synthetic data

We investigate here the empirical performance on simulated data of the BNP cardinality estimator derived in Section 3, for the special cases of the DP prior of the general PYP prior. Synthetic data are generated from the BNP model in (2) using different values of the PYP parameters (α, θ), and then they are sketched using a hash function of width 128. Our BNP estimates are compared to the ground truth, which is available in these experiments because we know the prior parameters of the data-generating model and have access to the non-sketched data. All experiments are repeated 20 times and the results averaged, utilizing independent data sets and independent hash functions.

Figure 8 compares the true and estimated cardinality as a function of the sample size n, separately for data generated from DP prior models with $\alpha = 0$ and different values of θ. Figures 10 and 11 in Appendix C.1 consider instead data generated from PYP models with $\theta = 100$ and different values of α and from Zipf distributions with different tail parameters, respectively. To make the computations practical, all estimates are computed under the (possibly mis-specified) assumption that $\alpha = 0$, and estimating θ empirically via maximum marginal likelihood. As predicted by the theory, the results confirm our estimates are accurate when the DP prior is well-specified, while otherwise they tend to underestimate the number of distinct species, especially if n is large. Similar results are shown by Figure 11 in Appendix C, which reports on analogous experiments based on synthetic data generated from a Zipf distribution.

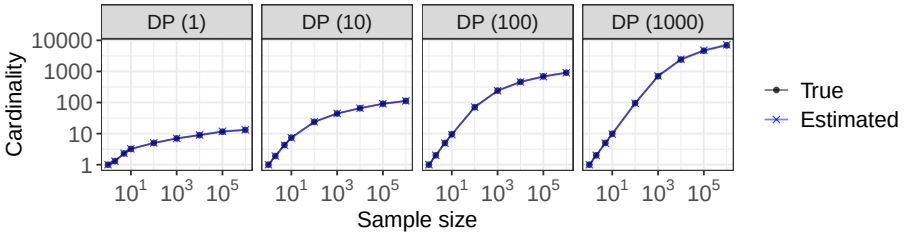


Figure 8: True and estimated cardinality in synthetic data from a DP prior model with different parameters, as a function of the sample size. The true and estimated cardinalities are almost indistinguishable.

5.2.2 Experiments with real data

In addition to the Gutenberg corpus' bigrams dataset discussed in Section 5.1.3, we consider here two additional datasets. The first one was made publicly available by the National Center for Biotechnology Information (Hatcher et al., 2017) and contains 43,196 sequences of approximately 30,000 nucleotides each, from SARS-CoV-2 viruses. For each sequence, we extract a list of all contiguous DNA sub-sequences of length 16 (i.e., *16-mers*), and then we sketch the resulting data set with the random hash function. The last data set is discussed in Rojas et al. (2018) and contains a list of 3,577,296 IP addresses, which we sketch directly without pre-processing; these data were made publicly available through the Kaggle machine-learning competition website. For all data sets, in these experiments we separately consider sketches obtained with two distinct hash functions, one with a width of 128 and another with a width of 4096.

Figure 9 compares the true and estimated cardinality for random subsets of the three aforementioned data sets, as a function of the sample size and for two different values of the hash width. Our BNP cardinality estimators are computed assuming $\alpha = 0$, and estimating θ empirically via maximum marginal likelihood. All results are averaged over 20 independent experiments with different hash functions. The results show the cardinality estimates obtained under the DP prior are relatively accurate for the DNA data set, which does not exhibit power-law tail behaviour (Sesia and Favaro, 2022), but tend to underestimate the true missing mass in the other cases, especially if the hash function width is small. In principle, we expect that using the PYP prior would yield more accurate estimators for the cardinality in all cases. However, computing (18) (for $\alpha \neq 0$) is computationally prohibitive even for moderate sample sizes.

6. Discussion

In the "species" setting, while Cai et al. (2018) and Dolera et al. (2021, 2023) employed BNPs to create learning-augmented versions of the CMS using DP and PYP priors, our approach diverges from the CMS framework as we condition on the information contained in the entire sketch. Moreover, our approach encompasses the broader class of PK priors and also facilitates cardinality recovery using the same data sketch. PK priors have been widely applied in BNPs, assuming access to true data, largely for their mathematical tractability, which leads to posterior inferences that are straightforward, computationally efficient,

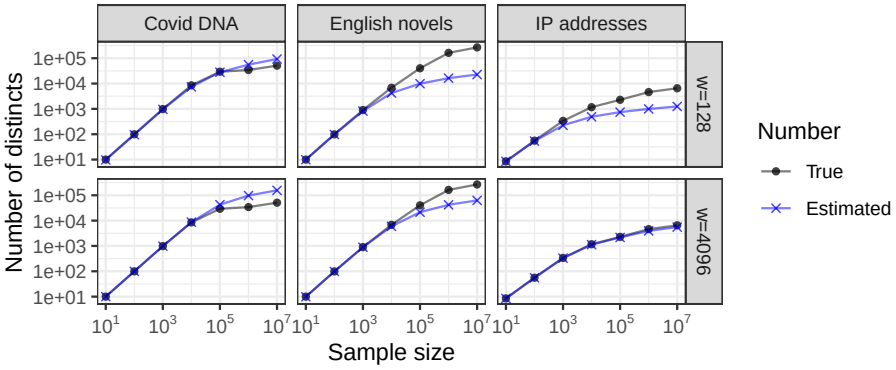


Figure 9: True and estimated cardinality in real data sets sketched with random hash functions of different width, as a function of the sample size.

and scalable to large datasets (Lijoi and Prünster, 2010). However, we have shown these advantages of PK priors do not necessarily extend to BNP inferences from sketched data, presenting unique challenges and constraints in selecting appropriate prior distributions.

This challenge creates opportunities for: (i) further exploration of both exact and approximate algorithms for the efficient numerical evaluation of the posterior under the PYP prior; (ii) continued investigation into large-sample approximations of the posterior under the PYP prior, potentially with reliable error bounds; and (iii) the consideration of alternative priors aimed at simplifying posterior distributions while providing more flexible tail behaviors than those offered by the DP.

The "traits" setting of the frequency recovery problem introduces another novel aspect that goes beyond the works of Cai et al. (2018) and Dolera et al. (2021, 2023), highlighting the adaptability of the BNP framework to diverse data structures. Unlike the "species" setting, the "traits" setting demonstrates greater flexibility in the choice of prior distributions, enabling the tractable evaluation of posterior distributions for large sample sizes, especially under the assumption of a Poisson distribution for the levels of trait associations. Such a desirable property stems from the Poisson process formulation of CRMs, which determines a posterior distribution that depends on the sketch $\mathbf{C}_J$ only through $C_{h(X_{n+1})}$.

In general, for both the "species" and "traits" settings, we argue that the BNP approach is also flexible with respect to the object of interest, due to the use of the posterior distribution as the main tool to obtain estimates. In the future, our BNP approach may also be extended to recover other quantities beyond those studied in this paper. Of notable interest in the CMS literature is the problem of recovering the (cumulative) empirical frequency of $s \geq 1$ new data points (Cormode and Yi, 2020, Chapter 3). We refer to Dolera et al. (2023) for a discussion in the context of learning-augmented versions of the CMS.

Acknowledgments

The authors thank the Action Editor, Professor Debdeep Pati, and three anonymous Referees for their helpful suggestions. M. B. and S. F. were funded by the European Research Council under the Horizon 2020 research and innovation programme, grant 817257.

M. B. and S. F. also gratefully acknowledge support from the Italian Ministry of Education, University and Research, "Dipartimenti di Eccellenza" grant 2023-2027.

Appendix A. Proofs

A.1 Proof of Theorem 2

By the definition of conditional probability, we have that

$$\Pr\left[f_{X_{n+1}} = l \,|\, \boldsymbol{C}_J = \boldsymbol{c}, h(X_{n+1}) = j\right] = \frac{\Pr\left[f_{X_{n+1}} = l, \boldsymbol{C}_J = \boldsymbol{c}, h(X_{n+1}) = j\right]}{\Pr\left[\boldsymbol{C}_J = \boldsymbol{c}, h(X_{n+1}) = j\right]}. \tag{33}$$

Let $g(t) = Z_T t^{-\gamma} e^{-\beta t}$, where Z_T is a normalizing constant.

Denominator. We have

$$\begin{aligned}
&\Pr[\boldsymbol{C}_J = \boldsymbol{c}, h(X_{n+1}) = j] \\
&\quad = \binom{n}{c_1, \ldots, c_J} \Pr\Big[X_1 \in D_j, \ldots, X_{c_j} \in D_j, X_{n+1} \in D_j, \\
&\qquad\qquad X_i \in D_1, i \in [c_j : c_j + c_1], \ldots, X_i \in D_j, i \in [\sum_{k=1}^{J-1} c_k, n]\Big] \\
&\quad = \binom{n}{c_1, \ldots, c_J} \mathbb{E}\left[P(D_j)^{c_j+1} \prod_{k \neq j} P(D_k)^{c_k}\right]. \qquad (34)
\end{aligned}$$

From (34), writing $P(\cdot) := \tilde{\mu}(\cdot)/\tilde{\mu}(\mathbb{S})$ and the expression of $g(t) = g(\tilde{\mu}(\mathbb{S}))$

$$\begin{aligned}
\Pr[\boldsymbol{C}_J = \boldsymbol{c}, h(X_{n+1}) = j] &= \binom{n}{c_1, \ldots, c_j} \mathbb{E}\left[P(D_j)^{c_j+1} \prod_{k \neq j} P(D_k)^{c_k}\right] \\
&= Z_T \binom{n}{c_1, \ldots, c_j} \mathbb{E}\left[\tilde{\mu}(\mathbb{S})^{-n-1-\gamma} e^{-\beta\tilde{\mu}(\mathbb{S})} \tilde{\mu}(D_j)^{c_j+1} \prod_{k \neq j} \tilde{\mu}(D_k)^{c_k}\right] \\
&= Z_T \binom{n}{c_1, \ldots, c_j} \int_{\mathbb{R}_+} \frac{u^{n+\gamma}}{\Gamma(n+\gamma+1)} \mathbb{E}\left[e^{-(u+\beta)\tilde{\mu}(D_j)} \tilde{\mu}(D_j)^{c_j+1}\right] \\
&\qquad \times \prod_{k \neq j} \mathbb{E}\left[e^{-(u+\beta)\tilde{\mu}(D_k)} \tilde{\mu}(D_k)^{c_k}\right] \\
&= Z_T \binom{n}{c_1, \ldots, c_j} \int_{\mathbb{R}_+} \frac{u^{n+\gamma}}{\Gamma(n+\gamma+1)} (-1)^{c+1} \frac{\mathrm{d}^{c_j+1}}{\mathrm{d}z^{c_j+1}} e^{-\theta\psi(z)/J}|_{(u+\beta)} \\
&\qquad \times \prod_{k \neq j} (-1)^{c_k} \frac{\mathrm{d}^{c_k}}{\mathrm{d}z^{c_k}} e^{-\theta\psi(z)/J}|_{(u+\beta)} \mathrm{d}u,
\end{aligned}$$

where the second equality follows from the definition of PK model and the third from the Gamma identity $\tilde{\mu}(\mathbb{S})^{-n-1-\gamma} = \int_{\mathbb{R}_+} u^{n+\gamma}/\Gamma(n+\gamma+1) e^{-u\tilde{\mu}(\mathbb{S})} \mathrm{d}x$, an application of Fubini's theorem, and the independence property of CRMs. The last equality follows from the properties of the exponential function and the definition of Laplace exponent of the CRM.

Numerator. Let B_ω denote a ball of radius ε around $\omega \in \mathbb{S}$. We have

$$\Pr\left[\sum_{i=1}^{n} \mathrm{I}_{(X_{n+1})}(X_i) = l, \boldsymbol{C}_J = \boldsymbol{c}, h(X_{n+1}) = j\right] = \lim_{\varepsilon \to 0} \int_{D_j} \Pr\left[\sum_{i=1}^{n} \mathrm{I}_{B_{\omega^*}}(X_i) = l, \boldsymbol{C}_J = \boldsymbol{c}, X_{n+1} \in B_{\omega^*}\right] \mathrm{d}\omega^*$$

We consider the integrand,

$$\begin{aligned}
&\Pr\left[\sum_{i=1}^{n} \mathrm{I}_{B_{\omega^*}}(X_i) = l, \sum_{i=1}^{n} \mathrm{I}_{h(\omega^*)}(h(X_i)) = c, X_{n+1} \in B_{\omega^*}\right] \\
&\quad = \binom{n}{l} \binom{n-l}{c_1, \ldots, c_j - l, \ldots c_J} \Pr\Big[X_1 \in B_{\omega^*} \ldots X_l \in B_{\omega^*}, X_{n+1} \in B_{\omega^*}, \\
&\qquad\qquad X_{l+1} \in D_j \setminus B_{\omega^*} \ldots, X_{c_j} \in D_j \setminus B_{\omega^*}, \\
&\qquad\qquad X_i \in D_1, i \in [c_j : c_j + c_1], \ldots, X_i \in D_j, i \in [\sum_{l=1}^{J-1} c_l, n]\Big] \\
&\quad = \binom{n}{l, c_1, \ldots, c_j - l, \ldots c_J} \mathbb{E}\left[P(B_{\omega^*})^{l+1} P(D_j \setminus B_{\omega^*})^{c_j - l} \prod_{k \neq j} P(D_k)^{c_k}\right]. \qquad (35)
\end{aligned}$$

To evaluate the expected value, we proceed as in the case of the denominator and write

$$\begin{aligned}
&\mathbb{E}\left[P(B_{\omega^*})^{l+1} P(D_j \setminus B_{\omega^*})^{c_j - l} \prod_{k \neq j} P(D_k)^{c_k}\right] \\
&\quad = Z_T \int_{\mathbb{R}_+} \frac{u^{n+\gamma}}{\Gamma(n+\gamma+1)} \mathbb{E}\left[e^{-(u+\beta)\tilde{\mu}(B_{\omega^*})} \tilde{\mu}(B_{\omega^*})^{l+1}\right] \\
&\qquad \times \mathbb{E}\left[e^{-(u+\beta)\tilde{\mu}(D_j \setminus B_{\omega^*})} \tilde{\mu}(D_j \setminus B_{\omega^*})^{c-l}\right] \prod_{k \neq j} \mathbb{E}\left[e^{-(u+\beta)\tilde{\mu}(D_k)} \tilde{\mu}(D_k)^{c_k}\right] \mathrm{d}u.
\end{aligned}$$

The two latter expected values above can be computed as in the denominator case. For the first expectation instead, letting $\varepsilon \to 0$ we have

$$\mathbb{E}\left[e^{-(u+\beta)\tilde{\mu}(B_{\omega^*})} \tilde{\mu}(B_{\omega^*})^{l+1}\right] \to \kappa(u+\beta, l+1)\theta G_0(\mathrm{d}\omega^*),$$

where $\kappa(u,l) = \int_{\mathbb{R}_+} s^l e^{-us} \rho(s) \mathrm{d}s$. This can be verified using, for instance, Lemma 1 in Camerlenghi et al. (2019). Hence, the numerator equals

$$\Pr\left[\sum_{i=1}^{n} \mathrm{I}_{(X_{n+1})}(X_i) = l, \boldsymbol{C}_J = \boldsymbol{c}, h(X_{n+1}) = j\right] \frac{1}{l!} \binom{n}{c_1, \ldots, c_j - l, \ldots c_J} \frac{Z_T}{\Gamma(n+\gamma+1)}$$

$$\times \int_{D_j} \int_{\mathbb{R}_+} u^{n+\gamma} (-1)^{c_j - l} \left(\frac{\mathrm{d}^{c_j - l}}{\mathrm{d}(u+\beta)^{c_j - l}} e^{-\theta/J \psi(u+\beta)} \right) \times$$
$$\times \prod_{k \neq j} (-1)^{c_k} \frac{\mathrm{d}^{c_k}}{\mathrm{d}z^{c_k}} e^{-\theta \psi(z)/J}|_{(u+\beta)} \kappa(u+\beta, l+1) \theta \mathrm{d}u G_0(\mathrm{d}\omega^*),$$

where we can further integrate with respect to $\mathrm{d}\omega^*$ and observe $\int_{D_j} G_0(\mathrm{d}\omega^*) = 1/J$ thanks to the universality assumption on the hash function h. Combining numerator and denominator, and using the definition of $\phi^{(n)}(u)$, yields the result.

A.2 Proof of Theorem 3

Recalling the definition of $\psi^{(n)}(u)$, the claim of the theorem entails that

$$\frac{\int_{\mathbb{R}_+} u^{n+\gamma} \frac{\mathrm{d}^{c_j - l}}{\mathrm{d}(z)^{c_j - l}} e^{-\theta/J \psi(z)}|_{u+\beta} \prod_{k \neq j} \frac{\mathrm{d}^{c_k}}{\mathrm{d}z^{c_k}} e^{-\theta \psi(z)/J}|_{(u+\beta)} \kappa(u+\beta, l+1) \mathrm{d}u}{\int_{\mathbb{R}_+} u^{n+\gamma} \frac{\mathrm{d}^{c_j+1}}{\mathrm{d}z^{c_j+1}} e^{-\theta \psi(z)/J}|_{(u+\beta)} \prod_{k \neq j} \frac{\mathrm{d}^{c_k}}{\mathrm{d}z^{c_k}} e^{-\theta \psi(z)/J}|_{(u+\beta)} \mathrm{d}u} = f(n, c_j, l),$$

where f is an unknown function which cannot depend on $c_k, k \neq j$. Then

$$\int_{\mathbb{R}_+} u^{n+\gamma} \frac{\mathrm{d}^{c_j - l}}{\mathrm{d}(z)^{c_j - l}} e^{-\theta/J \psi(z)}|_{u+\beta} \prod_{k \neq j} \frac{\mathrm{d}^{c_k}}{\mathrm{d}z^{c_k}} e^{-\theta \psi(z)/J}|_{(u+\beta)} \kappa(u+\beta, l+1) \mathrm{d}u =$$
$$f(n, c_j, l) \int_{\mathbb{R}_+} u^{n+\gamma} \frac{\mathrm{d}^{c_j+1}}{\mathrm{d}z^{c_j+1}} e^{-\theta \psi(z)/J}|_{(u+\beta)} \prod_{k \neq j} \frac{\mathrm{d}^{c_k}}{\mathrm{d}z^{c_k}} e^{-\theta \psi(z)/J}|_{(u+\beta)} \mathrm{d}u,$$

or, equivalently,

$$\int_{\mathbb{R}_+} u^{n+\gamma} \prod_{k \neq j} \frac{\mathrm{d}^{c_k}}{\mathrm{d}z^{c_k}} e^{-\theta \psi(z)/J}|_{(u+\beta)} \times$$
$$\left(\frac{\mathrm{d}^{c_j - l}}{\mathrm{d}(z)^{c_j - l}} e^{-\theta/J \psi(z)}|_{u+\beta} \kappa(u+\beta, l+1) - f(n, c_j, l) \frac{\mathrm{d}^{c_j+1}}{\mathrm{d}z^{c_j+1}} e^{-\theta \psi(z)/J}|_{(u+\beta)} \right) \mathrm{d}u = 0.$$

Since the above equality must hold true for all values of $c_k, k \neq j$, it follows that, $\forall u \geq 0$,

$$\frac{\mathrm{d}^{c_j - l}}{\mathrm{d}(z)^{c_j - l}} e^{-\theta/J \psi(z)}|_{u+\beta} \kappa(u+\beta, l+1) - f(n, c_j, l) \frac{\mathrm{d}^{c_j+1}}{\mathrm{d}z^{c_j+1}} e^{-\theta \psi(z)/J}|_{(u+\beta)} \equiv 0. \tag{36}$$

The simplest nontrivial case is when $c_j = 1, l = 0$ (indeed note that if $c_j = 0$, $l = 0$ by definition and we get $f(n, c_j, l) = 1$). Now, note that

$$\frac{\mathrm{d}}{\mathrm{d}z} e^{-\theta/J \psi(z)} = e^{-\theta/J \psi(z)} \left(-\frac{\theta}{J} \right) \frac{\mathrm{d}}{\mathrm{d}z} \psi(z),$$
$$\frac{\mathrm{d}^2}{\mathrm{d}z^2} e^{-\theta/J \psi(z)} = e^{-\theta/J \psi(z)} \left[\left(-\frac{\theta}{J} \right) \frac{\mathrm{d}}{\mathrm{d}z} \psi(z) \right]^2 + e^{-\theta/J \psi(z)} \left(-\frac{\theta}{J} \right) \frac{\mathrm{d}^2}{\mathrm{d}z^2} \psi(z).$$

Plugging these into (36) we get

$$e^{-\theta/J\psi(u+\beta)}\Bigg\{ -\frac{\theta}{J}\frac{\mathrm{d}}{\mathrm{d}z}\psi(z)|_{u+\beta}\kappa(u+\beta,1)+ \\ - f(n,c_j,l)\left(\left[\left(-\frac{\theta}{J}\right)\frac{\mathrm{d}}{\mathrm{d}z}\psi(z)\right]^2_{u+\beta}\right) - \frac{\theta}{J}\frac{\mathrm{d}^2}{\mathrm{d}z^2}\psi(z)|_{u+\beta}\Bigg\} = 0.$$

Given the positivity of the exponential function, we can set the term in the curly brackets equal to zero. Since $\mathrm{d}/\mathrm{d}z\,\psi(z) = -\kappa(z,1)$, the term in the curly brackets above reduces to

$$\frac{\theta}{J}\kappa(u+\beta,1)^2 - f(n,c_j,l)\left(\frac{\theta^2}{J^2}\kappa(u+\beta,1)^2 + \frac{\theta}{J}\frac{\mathrm{d}}{\mathrm{d}z}\kappa(z,1)|_{u+\beta}\right) = 0,$$

which entails

$$\frac{\mathrm{d}}{\mathrm{d}z}\kappa(z,1) = -\left(\frac{\theta}{J} - \frac{1}{f(m,c_j,l)}\right)\kappa(z,1)^2.$$

Letting $w := \left(\frac{\theta}{J} - \frac{1}{f(n,c_j,l)}\right)^{-1}$, the differential equation above can be seen to have solution

$$\kappa(z,1) = \frac{1}{c + \frac{z}{w}} = \frac{w}{\tau + z},$$

where c is an arbitrary constant and $\tau = cw$.

Let now $c_j = l = 0$ and recall that in this case $f \equiv 1$. Plugging these into (36) we get

$$e^{-\theta/J\psi(u+\beta)}\kappa(u+\beta,1) - \frac{\mathrm{d}}{\mathrm{d}z}e^{-\theta/J\psi(z)} = 0,$$

which leads to

$$e^{-\theta/J\psi(u+\beta)}\left(\frac{w}{\tau+z} + \frac{\theta}{J}\frac{\mathrm{d}}{\mathrm{d}z}\psi(z)\right)_{|u+\beta} = 0.$$

Setting the term in the parentheses equal to zero, we obtain that

$$\psi(z) = K\log(\tau+z). \tag{37}$$

Hence, we have shown that if $\Pr\left[f_{X_{n+1}} = l \,|\, \boldsymbol{C}_J = \boldsymbol{c}, h(X_{n+1}) = j\right]$ does not depend on c_k, $k \neq j$, the CRM $\tilde{\mu}$ must have Lévy exponent (37).

Let now $\tilde{\mu}'$ be a CRM with Lévy exponent ψ as above. We first note that, without loss of generality, we can set $K = 1$ since setting $K \neq 1$ the Laplace transform

$$\mathbb{E}e^{-z\tilde{\mu}'(A)} = e^{-K\log(\tau+z)\theta G_0(A)} = (\tau+z)^{-K\theta G_0(A)} =: F'(z)$$

simply amounts to rescaling the total mass parameter.

We now show that τ must necessarily be equal to one. Note that if $\tilde{\mu}_G$ is a Gamma process, then its Lévy exponent is $\log(1+z)$, which is (37) but shifted of a term $(1-\tau)$:

$$\mathbb{E}e^{-z\tilde{\mu}_G(A)} = (1+z)^{-K\theta G_0(A)} = F'(z+(1-\tau)). \tag{38}$$

Let f_A be the probability density function of $\tilde{\mu}_G(A)$, and f'_A be the probability density function of $\tilde{\mu}'(A)$. By the properties of the Laplace transform, (38) is equivalent to

$$f_A(t) = e^{(\tau-1)t}f'_A(t), \quad \text{for all } t,$$

which is clearly impossible if $\tau \neq 1$ since we must have that both f_A and f'_A must integrate to one since they are probability density functions. Hence, we have shown that if $\Pr\left[f_{X_{n+1}} = l \,|\, \boldsymbol{C}_J = \boldsymbol{c}, h(X_{n+1}) = j\right]$ does not depend on $c_k, k \neq j$, the underlying CRM must be a Gamma process.

We are left with two degrees of freedom: namely the parameters γ and β defining the change of measure in the Poisson-Kingman model. However, note that if $\tilde{\mu}$ is a Gamma process, the resulting PK model with $g(t) \propto t^{-\gamma} e^{-\beta t}$ is still a DP. This can be checked, for instance, starting from Eq. (177) in James (2002). This concludes the proof.

A.3 Proof of Equation (6) as a special case of Theorem 2

The DP is obtained by normalizing a Gamma process, whose Lévy intensity is $\theta s^{-1} e^{-s} \mathrm{d}s G_0(\mathrm{d}x)$. Hence,

$$\begin{aligned}
\psi(u) &= \int_{\mathbb{R}_+} (1 - e^{-us}) \rho(s) \mathrm{d}s = \int_{\mathbb{R}_+} (1 - e^{-us}) s^{-1} e^{-s} \mathrm{d}s \\
&= \int_{\mathbb{R}_+} \int_{\mathbb{R}_+} (1 - e^{-us}) e^{-ts} e^{-s} \mathrm{d}t \, \mathrm{d}s \\
&= \int_{\mathbb{R}_+} \frac{u}{(t+1)(t+u+1)} \mathrm{d}t = \log(1+u),
\end{aligned}$$

where the third equality follows from the identity $s^{-1} = \int_{\mathbb{R}_+} e^{-ts} \mathrm{d}t$ and the fourth one by an application of Fubini's theorem. Then it follows

$$(-1)^n \frac{\mathrm{d}^u}{\mathrm{d}u^n} e^{-\theta \psi(u)} = \frac{\Gamma(\theta + n)}{\Gamma(\theta)} (1+u)^{-\theta - n},$$

Moreover, $k(l, u) = \Gamma(l)/(u+1)^l$. Hence, the integral at the numerator of (4) can be evaluated as

$$\begin{aligned}
&\frac{\Gamma(\theta/J + c_j - l)}{\Gamma(\theta/J)} \prod_{k \neq j} \frac{\Gamma(\theta/J + c_k)}{\Gamma(\theta/J)} \Gamma(l+1) \int_{\mathbb{R}_+} u^n + (1+u)^{-\theta - n - 1} \\
&\quad = \frac{\Gamma(\theta/J + c_j - l)}{\Gamma(\theta/J)} \prod_{k \neq j} \frac{\Gamma(\theta/J + c_k)}{\Gamma(\theta/J)} \Gamma(l+1) \frac{\Gamma(n+1)}{\Gamma(\theta + n + 1)}.
\end{aligned}$$

Similarly, the integral at the denominator of (4) equals

$$\begin{aligned}
&\frac{\Gamma(\theta/J + c_j + 1)}{\Gamma(\theta/J)} \prod_{k \neq j} \frac{\Gamma(\theta/J + c_k)}{\Gamma(\theta/J)} \int_{\mathbb{R}_+} u^n + (1+u)^{-\theta - n - 1} \\
&\quad = \frac{\Gamma(\theta/J + c_j + 1)}{\Gamma(\theta/J)} \prod_{k \neq j} \frac{\Gamma(\theta/J + c_k)}{\Gamma(\theta/J)} \frac{\Gamma(n+1)}{\Gamma(\theta + n + 1)}.
\end{aligned}$$

Combining these expressions together, we have

$$\Pr\left[f_{X_{n+1}} = l \,|\, \boldsymbol{C}_J = \boldsymbol{c}, h(X_{n+1}) = j\right] = \frac{\theta}{J} \binom{c_j}{l} l! \frac{\Gamma(\theta/J + c_j - l)}{\Gamma(\theta/J + c_j + 1)}.$$

A.4 Proof of Equation (6) from the finite dimensional laws of the DP prior

In this proof, we exploit only the original characterization of the DP in terms of its finite-dimensional distributions. That is, given $\theta > 0$ and G_0 a probability measure on $(\mathbb{S})$, P is a DP with mean measure θG_0 if and only if, for any $n > 0$ and any n measurable partition $A_1, \dots, A_n$ of $\mathbb{X}$,

$$(P(A_1), \dots, P(A_n)) \sim \mathrm{Dir}_n(\theta G_0(A_1), \dots, \theta G_0(A_n)), \tag{39}$$

where Dir_n denotes the $n-1$ dimensional Dirichlet distribution.

We argue as in Appendix A.1. To compute the denominator in (33), from (34), (39)

$$\begin{aligned}
&\Pr[\boldsymbol{C}_J = \boldsymbol{c}, h(X_{n+1}) = j] \\
&\quad = \binom{n}{c_1, \dots, c_J} \frac{\Gamma(\theta)}{\Gamma(\theta + n + 1)} \frac{\Gamma(\theta/J + c_j + 1)}{\Gamma(\theta/J)} \prod_{k \neq j} \frac{\Gamma(\theta/J + c_k)}{\Gamma(\theta/J)},
\end{aligned}$$

where we also exploited the e uniformity of the hash function which ensures that $G_0(D_\ell) = G_0(h^{-1}(\{\ell\})) = 1/J$ for any $\ell = 1, \dots, J$.

Similarly, to compute the numerator, consider (35). Since P is a DP,

$$\begin{aligned}
&\mathbb{E}\left[P(B_{\omega^*})^{l+1} P(D_j \setminus B_{\omega^*})^{c_j - l} \prod_{k \neq j} P(D_k)^{c_k} \right] \\
&\quad = \frac{1}{l!} \binom{n}{c_1, \dots, c_j - l, \dots c_J} \frac{\Gamma(\theta)}{\Gamma(\theta + n + 1)} \frac{\Gamma(\alpha G_0(B_{\omega^*}) + l + 1)}{\Gamma(\theta G_0(B_{\omega^*}))} \\
&\qquad \times \frac{\Gamma(\theta G_0(D_j \setminus B_{\omega^*}) + c_j - l)}{\Gamma(\theta G_0(D_j \setminus B_{\omega^*}))} \prod_{k \neq j} \frac{\Gamma(\theta/J + c_k)}{\Gamma(\theta/J)}.
\end{aligned}$$

We now let $\varepsilon \to 0$. First note that, of course

$$\frac{\Gamma(\theta G_0(D_j \setminus B_{\omega^*}) + c_j - l)}{\Gamma(\theta G_0(D_j \setminus B_{\omega^*}))} \to \frac{\Gamma(\theta G_0(D_j) + c_j - l)}{\Gamma(\theta G_0(D_j))} = \frac{\Gamma(\theta/J + c_j - l)}{\Gamma(\theta/J)}.$$

To evaluate the limit of $\Gamma(\theta G_0(B_{\omega^*}) + l + 1)/\Gamma(\theta G_0(B_{\omega^*}))$, we first unroll the numerator using the recurrence relation $\Gamma(z+1) = z\Gamma(z)$ l times so that

$$\frac{\Gamma(\theta G_0(B_{\omega^*}) + l + 1)}{\Gamma(\theta G_0(B_{\omega^*}))} = (\theta G_0(B_{\omega^*}) + l) \cdots (\theta G_0(B_{\omega^*})) = \theta \Gamma(l+1) G_0(B_{\omega^*}) + o(G_0(B_{\omega^*})).$$

Letting now $\varepsilon \to 0$, we can ignore higher order infinitesimals and get that

$$\frac{\Gamma(\theta G_0(B_{\omega^*}) + l + 1)}{\Gamma(\theta G_0(B_{\omega^*}))} \to \theta \Gamma(l+1) G_0(\mathrm{d}\omega^*),$$

which leads to

$$\Pr\left[\sum_{i=1}^{n} \mathrm{I}_{(X_{n+1})}(X_i) = l, \boldsymbol{C}_J = \boldsymbol{c}, h(X_{n+1}) = j \right]$$

$$
\begin{aligned}
&= \int_{D_j} \Pr\left[\sum_{i=1}^{n} \mathrm{I}_{X_{n+1}}(X_i) = l, \boldsymbol{C}_J = \boldsymbol{c}, X_{n+1} \in \mathrm{d}\omega^*\right] \\
&= \frac{1}{l!}\binom{n}{c_1, \ldots, c_j - l, \ldots c_J} \frac{\Gamma(\theta)}{\Gamma(\theta+n+1)} \theta \Gamma(l+1) G_0(B_{\omega^*}) \\
&\quad \times \frac{\Gamma(\theta/J + c_j - l)}{\Gamma(\theta/J)} \prod_{k \neq j} \frac{\Gamma(\theta/J + c_k)}{\Gamma(\theta/J)} \int_{D_j} G_0(\mathrm{d}\omega^*).
\end{aligned}
$$

Integration with respect to $\mathrm{d}\omega^*$ is now straightforward and $\int_{D_j} G_0(\mathrm{d}\omega^*) = 1/J$.

Combining numerator and denominator yields the proof.

A.5 Proof of (8) as a special case of Theorem 2

In the case of a PYP, we have $\beta = 0$, $\psi(u) = u^\alpha$ and

$$
\kappa(u, l) = \alpha u^{\alpha - l} \frac{\Gamma(l - \alpha)}{\Gamma(1 - \alpha)} = \alpha u^{\alpha - l} (1 - \alpha)_{(l-1)}.
$$

Using the Faa di Bruno formula, we have (see, e.g., Lemma 1 in Camerlenghi et al., 2019):

$$
\begin{aligned}
(-1)^c \frac{\mathrm{d}^c}{\mathrm{d}u^c} e^{-u^\alpha/J} &= e^{-u^\alpha/J} \sum_{i=0}^{c} \left(\frac{1}{J}\right)^i \sum_{(*)} \frac{1}{i!} \binom{c}{k_1 \cdots k_i} \prod_{j=1}^{i} \kappa(u, k_j) \\
&= e^{-u^\alpha/J} \sum_{i=0}^{c} \left(\frac{1}{J}\right)^i \frac{\alpha^i}{u^{c - \alpha i} i!} \sum_{(*)} \binom{c}{k_1 \cdots k_i} \prod_{j=1}^{i} (1 - \alpha)_{k_j - 1} \\
&= e^{-u^\alpha/J} \sum_{i=0}^{c} \left(\frac{1}{J}\right)^i \frac{1}{u^{c - \alpha i}} \mathscr{C}(c, i; \alpha),
\end{aligned}
$$

where $\mathscr{C}(c, i; \alpha)$ is the generalized factorial coefficient and $(*)$ denotes the summation over positive integers $(k_1, \ldots, k_i)$ such that $\sum_{j=1}^{i} k_j = c$.

Recall the definition of the multi-index set $S(\boldsymbol{c}, j, q)$ as in the main text below (8). Then, the integral at the numerator of (4) equals

$$
\begin{aligned}
&\int_{\mathbb{R}_+} u^{b+\gamma} e^{-u^\alpha/J} \sum_{i_j=0}^{c_j - l} J^{-i_j} \frac{\mathscr{C}(c_j - l, i_j; \alpha)}{u^{c_j - l - \alpha i_j}} \\
&\qquad \times \left\{ \prod_{k \neq j} e^{-u^\alpha/J} \sum_{i_k=0}^{c_k} J^{-i_k} \frac{\mathscr{C}(c_k, i_k; \alpha)}{u^{c_k - \alpha i_k}} \right\} \alpha u^{\alpha - l - 1} (1 - \alpha)_{(l)} \mathrm{d}u \\
&= \sum_{i_1=0}^{c_1} \cdots \sum_{i_j=0}^{c_j - l} \cdots \sum_{i_J=0}^{c_J} J^{-\sum_k i_k} \mathscr{C}(c_j - l, i_j; \alpha) \\
&\qquad \times \prod_{k \neq j} \mathscr{C}(c_k, i_k; \alpha) \alpha (1 - \alpha)_{(l)} \int_{\mathbb{R}_+} e^{-u^\alpha} u^{\gamma - 1 + \alpha \sum_k i_k + \alpha} \mathrm{d}u \\
&= \sum_{i_1=0}^{c_1} \cdots \sum_{i_j=0}^{c_j - l} \cdots \sum_{i_J=0}^{c_J} J^{-\sum_k i_k} \mathscr{C}(c_j - l, i_j; \alpha)
\end{aligned}
$$

$$\times \prod_{k\neq j} \mathscr{C}(c_k, i_k; \alpha)(1-\alpha)_{(l)} \Gamma\left(\frac{\gamma+\alpha}{\alpha} + \sum_k i_k\right)$$

$$= (1-\alpha)_{(l)} \sum_{\boldsymbol{i}\in\mathcal{S}(\boldsymbol{c},j,-l)} \Gamma\left(\frac{\gamma+\alpha}{\alpha} + |\boldsymbol{i}|\right) J^{-|\boldsymbol{i}|} \prod_{k=1}^{J} \mathscr{C}(c_k - l\delta_{k,j}, i_k; \alpha).$$

Similarly, the integral at the denominator of (4) equals

$$\sum_{i_1=0}^{c_1} \cdots \sum_{i_j=0}^{c_j+1} \cdots \sum_{i_J=0}^{c_J} J^{-\sum_k i_k} \mathscr{C}(c_j+1, i_j; \alpha) \prod_{k\neq j} \mathscr{C}(c_k, i_k; \alpha) \Gamma\left(\frac{\gamma}{\alpha} + \sum_k i_k\right)$$

$$= \sum_{\boldsymbol{i}\in\mathcal{S}(\boldsymbol{c},j,1)} \Gamma\left(\gamma/\alpha + |\boldsymbol{i}|\right) J^{-|\boldsymbol{i}|} \prod_{k=1}^{J} \mathscr{C}(c_k + \delta_{k,j}, i_k; \alpha).$$

Combining these gives (8).

A.6 Proof of Equation (6) from Equation (8)

The proof relies on Charalambides (2005, Theorem 2.16), which characterizes the behaviour of generalized factorial coefficients as $\alpha \to 0$. For $n \geq 0$ and $0 \leq k \leq n$ it holds that

$$\lim_{\alpha\to+\infty} \frac{\mathscr{C}(n,k;\alpha)}{\alpha^k} = |s(n,k)|, \tag{40}$$

where $|s(n,k)|$ is the signless Stirling number of the first type; see Charalambides (2005, Chapter 2) for details. Now, we apply (40) to the posterior in (8). For $l = 0, 1, \ldots, c_j$:

$$\begin{aligned}
&\lim_{\alpha\to 0} \Pr[f_{X_{n+1}} = l \,|\, \mathbf{C}_J = \mathbf{c}, h(X_{n+1}) = j] \\
&= \lim_{\alpha\to 0} \frac{\gamma}{J} \binom{c_j}{l} (1-\alpha)_{(l)} \frac{\sum_{\boldsymbol{i}\in\mathcal{S}(\boldsymbol{c},j,-l)} \frac{\Gamma\left(\frac{\gamma+\alpha}{\alpha}+|\boldsymbol{i}|\right)}{J^{|\boldsymbol{i}|}} \prod_{k=1}^{J} \mathscr{C}(c_k - l\delta_{k,j}, i_k; \alpha)}{\sum_{\boldsymbol{i}\in\mathcal{S}(\boldsymbol{c},j,1)} \frac{\Gamma\left(\frac{\gamma}{\alpha}+|\boldsymbol{i}|\right)}{J^{|\boldsymbol{i}|}} \prod_{k=1}^{J} \mathscr{C}(c_k + \delta_{k,j}, i_k; \alpha)} \\
&= \lim_{\alpha\to 0} \frac{\gamma}{J} \binom{c_j}{l} (1-\alpha)_{(l)} \frac{\sum_{\boldsymbol{i}\in\mathcal{S}(\boldsymbol{c},j,-l)} \frac{\prod_{t=0}^{|\boldsymbol{i}|-1}(\gamma+\alpha+\alpha t)}{J^{|\boldsymbol{i}|}} \prod_{k=1}^{J} \frac{\mathscr{C}(c_k - l\delta_{k,j}, i_k; \alpha)}{\alpha^{i_k}}}{\sum_{\boldsymbol{i}\in\mathcal{S}(\boldsymbol{c},j,1)} \frac{\prod_{t=0}^{|\boldsymbol{i}|-1}(\gamma+\alpha t)}{J^{|\boldsymbol{i}|}} \prod_{k=1}^{J} \frac{\mathscr{C}(c_k + \delta_{k,j}, i_k; \alpha)}{\alpha^{i_k}}} \\
&= \frac{\gamma}{J} \binom{c_j}{l} l! \frac{\sum_{\boldsymbol{i}\in\mathcal{S}(\boldsymbol{c},j,-l)} \left(\frac{\gamma}{J}\right)^{|\boldsymbol{i}|} \prod_{k=1}^{J} |s(c_k - l\delta_{k,j}, i_k)|}{\sum_{\boldsymbol{i}\in\mathcal{S}(\boldsymbol{c},j,1)} \left(\frac{\gamma}{J}\right)^{\boldsymbol{i}} \prod_{k=1}^{J} |s(c_k + \delta_{k,j}, i_k)|}.
\end{aligned}$$

Recall that signless Stirling numbers of the first type are the coefficients of the series expansion of a rising factorial (Charalambides, 2005, Equation 2.4), i.e. for $t > 0$, $(t)_{(n)} = \sum_{0\leq k\leq n}^{n} |s(n,k)| t^k$. Then,

$$\lim_{\alpha\to 0} \Pr[f_{X_{n+1}} = l \,|\, \mathbf{C}_J = \mathbf{c}, h(X_{n+1}) = j] = \frac{\gamma}{J} \binom{c_j}{l} l! \frac{\prod_{k=1}^{J} \left(\frac{\gamma}{J}\right)_{(c_k - l\delta_{k,j})}}{\prod_{k=1}^{J} \left(\frac{\gamma}{J}\right)_{(c_k+\delta_{k,j})}}$$

$$= \frac{\gamma}{J} \frac{(c_j - l + 1)_{(l)}}{\left(\frac{\gamma}{J} + c_j - l\right)_{(l+1)}},$$

which coincides with the posterior in (6) by setting $\theta = \gamma$.

A.7 Proof of Theorem 4

The first step, proved below, consists of establishing that, with c_j fixed,

$$\lim_{\mathbf{c}_{-j} \to +\infty} \Pr[f_{X_{n+1}} = l \,|\, \mathbf{C}_J = \mathbf{c}, h(X_{n+1}) = j] \tag{41}$$
$$= \gamma \binom{c_j}{l} (1-\alpha)_{(l)} \frac{\Gamma\left(\frac{\gamma}{\alpha} + J\right) (\gamma + J\alpha)_{(c_j - l)}}{\Gamma\left(\frac{\gamma}{\alpha} + J - 1\right) (\gamma + J\alpha - \alpha)_{(c_j+1)}}.$$

Then, it follows by the Chu-Vandermonde identity (Charalambides, 2005, Chapter 2) that

$$\begin{aligned}
&\sum_{l=0}^{c_j} l\gamma \binom{c_j}{l} (1-\alpha)_{(l)} \frac{\Gamma\left(\frac{\gamma}{\alpha} + J\right) (\gamma + J\alpha)_{(c_j - l)}}{\Gamma\left(\frac{\gamma}{\alpha} + J - 1\right) (\gamma + J\alpha - \alpha)_{(c_j+1)}} \\
&= \gamma \frac{\Gamma\left(\frac{\gamma}{\alpha} + J\right)}{\Gamma\left(\frac{\gamma}{\alpha} + J - 1\right) (\gamma + J\alpha - \alpha)_{(c_j+1)}} \sum_{l=0}^{c_j} l \binom{c_j}{l} (1-\alpha)_{(l)} (\gamma + J\alpha)_{(c_j - l)} \\
&= \gamma \frac{\Gamma\left(\frac{\gamma}{\alpha} + J\right)}{\Gamma\left(\frac{\gamma}{\alpha} + J - 1\right) (\gamma + J\alpha - \alpha)_{(c_j+1)}} \sum_{l=0}^{c_j - 1} c_j \binom{c_j - 1}{l} \frac{\Gamma(1-\alpha+1+l)}{\Gamma(1-\alpha)} \frac{\Gamma(1-\alpha+1)}{\Gamma(1-\alpha+1)} (\gamma + J\alpha)_{c_j - 1 - l} \\
&= \gamma \frac{\Gamma\left(\frac{\gamma}{\alpha} + J\right)}{\Gamma\left(\frac{\gamma}{\alpha} + J - 1\right) (\gamma + J\alpha - \alpha)_{(c_j+1)}} \frac{1}{\alpha} c_j (1-\alpha)(2 + \gamma + J\alpha - \alpha)_{(c_j - 1)} \\
&= \frac{\gamma}{\alpha} c_j \frac{1-\alpha}{\gamma + J\alpha - \alpha + 1}.
\end{aligned}$$

Recalling the definition of $\hat{f}_{X_{n+1}}$ in (5), this directly leads to the desired result.

To prove (41), we rely on an asymptotic property of (normalized) generalized factorial coefficients. That is, from Dolera and Favaro (2020, Lemma 2), for any $z > 0$ it holds that:

$$\lim_{n \to +\infty} \frac{z^k \mathscr{C}(n, k; s)}{\sum_{k=1}^{n} z^k \mathscr{C}(n, k; s)} = \mathrm{e}^{-z} \frac{z^{k-1}}{(k-1)!}. \tag{42}$$

We rewrite numerator and the denominator of (8). For the numerator of (8),

$$\begin{aligned}
&\frac{\gamma}{J} \binom{c_j}{l} (1-\alpha)_{(l)} \\
&\quad \times \sum_{i_1=1}^{c_1} J^{-i_1} \mathscr{C}(c_1, i_1; \alpha) \cdots \sum_{i_j=1}^{c_j - l} J^{-i_j} \mathscr{C}(c_j - l, i_j; \alpha) \cdots \sum_{i_J=1}^{c_J} J^{-i_J} \mathscr{C}(c_J, i_J; \alpha) \\
&\quad \times \Gamma\left(\frac{\gamma + \alpha}{\alpha} + i_1 + \cdots + i_j + \cdots + i_J\right) \\
&= \frac{\gamma}{J} \binom{c_j}{l} (1-\alpha)_{(l)} \Gamma\left(\frac{\gamma + \alpha}{\alpha}\right) \left(\prod_{1 \le s \ne j \le J} \sum_{i_s=1}^{c_s} \frac{1}{J^{i_s}} \mathscr{C}(c_s, i_s; \alpha) \right)
\end{aligned} \tag{43}$$

$$\times \sum_{i_j=1}^{c_j-l} \left(\frac{1}{J}\right)^{i_j} \mathscr{C}(c_j-l, i_j; \alpha) \sum_{i_1=1}^{c_1} \left(\frac{\gamma+\alpha}{\alpha}\right)_{(i_1)} \frac{\left(\frac{1}{J}\right)^{i_1} \mathscr{C}(c_1, i_1; \alpha)}{\sum_{i_1=1}^{c_1} \left(\frac{1}{J}\right)^{i_1} \mathscr{C}(c_1, i_1; \alpha)} \cdots$$
$$\cdots \times \left(\frac{\gamma+\alpha}{\alpha} + i_1 + \cdots + i_{j-1}\right)_{(i_j)} \cdots$$
$$\cdots \times \sum_{i_J=1}^{c_J} \left(\frac{\gamma+\alpha}{\alpha} + i_1 + \cdots + i_{J-1}\right)_{(i_J)} \frac{\left(\frac{1}{J}\right)^{i_J} \mathscr{C}(c_J, i_J; \alpha)}{\sum_{i_J=1}^{c_J} \left(\frac{1}{J}\right)^{i_J} \mathscr{C}(c_J, i_J; \alpha)},$$

and for the denominator of (8)

$$\sum_{i_1=1}^{c_1} J^{-i_1} \mathscr{C}(c_1, i_1; \alpha) \cdots \sum_{i_j=1}^{c_j+1} J^{-i_j} \mathscr{C}(c_j+1, i_j; \alpha) \cdots \sum_{i_J=1}^{c_J} J^{-i_J} \mathscr{C}(c_J, i_J; \alpha) \tag{44}$$
$$\times \Gamma\left(\frac{\gamma}{\alpha} + i_1 + \cdots + i_j + \cdots + i_J\right)$$
$$= \Gamma\left(\frac{\gamma}{\alpha}\right) \left(\prod_{1 \le s \ne j \le J} \sum_{i_s=1}^{c_s} \frac{1}{J^{i_s}} \mathscr{C}(c_s, i_s; \alpha)\right)$$
$$\times \sum_{i_j=1}^{c_j+1} \left(\frac{1}{J}\right)^{i_j} \mathscr{C}(c_j+1, i_j; \alpha) \sum_{i_1=1}^{c_1} \left(\frac{\gamma}{\alpha}\right)_{(i_1)} \frac{\left(\frac{1}{J}\right)^{i_1} \mathscr{C}(c_1, i_1; \alpha)}{\sum_{i_1=1}^{c_1} \left(\frac{1}{J}\right)^{i_1} \mathscr{C}(c_1, i_1; \alpha)} \cdots$$
$$\cdots \times \left(\frac{\gamma}{\alpha} + i_1 + \cdots + i_{j-1}\right)_{(i_j)} \cdots$$
$$\cdots \times \sum_{i_J=1}^{c_J} \left(\frac{\gamma}{\alpha} + i_1 + \cdots + i_{J-1}\right)_{(i_J)} \frac{\left(\frac{1}{J}\right)^{i_J} \mathscr{C}(c_J, i_J; \alpha)}{\sum_{i_J=1}^{c_J} \left(\frac{1}{J}\right)^{i_J} \mathscr{C}(c_J, i_J; \alpha)}.$$

By combining (43) with (44), we write the posterior distribution (8) as

$$\Pr[f_{X_{n+1}} = l \,|\, \mathbf{C}_J = \mathbf{c}, h(X_{n+1}) = j] = \frac{\gamma}{J} \binom{c_j}{l} (1-\alpha)_{(l)} \frac{\Gamma\left(\frac{\gamma+\alpha}{\alpha}\right)}{\Gamma\left(\frac{\gamma}{\alpha}\right)} \frac{N_{\gamma,\alpha,J}(l; c_1, \ldots, c_J)}{D_{\gamma,\alpha,J}(c_1, \ldots, c_J)}, \tag{45}$$

where

$$N_{\gamma,\alpha,J}(l; c_1, \ldots, c_J) \tag{46}$$
$$= \sum_{i_j=1}^{c_j-l} \left(\frac{1}{J}\right)^{i_j} \mathscr{C}(c_j-l, i_j; \alpha) \sum_{i_1=1}^{c_1} \left(\frac{\gamma+\alpha}{\alpha}\right)_{(i_1)} \frac{\left(\frac{1}{J}\right)^{i_1} \mathscr{C}(c_1, i_1; \alpha)}{\sum_{i_1=1}^{c_1} \left(\frac{1}{J}\right)^{i_1} \mathscr{C}(c_1, i_1; \alpha)} \cdots$$
$$\cdots \times \left(\frac{\gamma+\alpha}{\alpha} + i_1 + \cdots + i_{j-1}\right)_{(i_j)} \cdots$$
$$\cdots \times \sum_{i_J=1}^{c_J} \left(\frac{\gamma+\alpha}{\alpha} + i_1 + \cdots + i_{J-1}\right)_{(i_J)} \frac{\left(\frac{1}{J}\right)^{i_J} \mathscr{C}(c_J, i_J; \alpha)}{\sum_{i_J=1}^{c_J} \left(\frac{1}{J}\right)^{i_J} \mathscr{C}(c_J, i_J; \alpha)}$$

and

$$D_{\gamma,\alpha,J}(c_1, \ldots, c_J) \tag{47}$$

$$= \sum_{i_j=1}^{c_j+1} \left(\frac{1}{J}\right)^{i_j} \mathscr{C}(c_j+1, i_j; \alpha) \sum_{i_1=1}^{c_1} \left(\frac{\gamma}{\alpha}\right)_{(i_1)} \frac{\left(\frac{1}{J}\right)^{i_1} \mathscr{C}(c_1, i_1; \alpha)}{\sum_{i_1=1}^{c_1} \left(\frac{1}{J}\right)^{i_1} \mathscr{C}(c_1, i_1; \alpha)} \cdots$$
$$\cdots \times \left(\frac{\gamma}{\alpha} + i_1 + \cdots + i_{j-1}\right)_{(i_j)} \cdots$$
$$\cdots \times \sum_{i_J=1}^{c_J} \left(\frac{\gamma}{\alpha} + i_1 + \cdots + i_{J-1}\right)_{(i_J)} \frac{\left(\frac{1}{J}\right)^{i_J} \mathscr{C}(c_J, i_J; \alpha)}{\sum_{i_J=1}^{c_J} \left(\frac{1}{J}\right)^{i_J} \mathscr{C}(c_J, i_J; \alpha)}.$$

Now, we apply repeatedly (42) to both (46) and (47), starting from the last terms, indexed by i_J. In particular, for the sum in $i_J = 1, \ldots, c_J$ of (46), we have that

$$\lim_{c_J \to +\infty} \sum_{i_J=1}^{c_J} \left(\frac{\gamma+\alpha}{\alpha} + i_1 + \cdots + i_{J-1}\right)_{(i_J)} \frac{\left(\frac{1}{J}\right)^{i_J} \mathscr{C}(c_J, i_J; \alpha)}{\sum_{i_J=1}^{c_J} \left(\frac{1}{J}\right)^{i_J} \mathscr{C}(c_J, i_J; \alpha)} \tag{48}$$
$$= \sum_{i_J \geq 1} \left(\frac{\gamma+\alpha}{\alpha} + i_1 + \cdots + i_{J-1}\right)_{(i_J)} \mathrm{e}^{-\frac{1}{J}} \frac{\left(\frac{1}{J}\right)^{i_J-1}}{(i_J-1)!}$$
$$= \mathrm{e}^{-\frac{1}{J}} \frac{\frac{\gamma+\alpha}{\alpha} + i_1 + \cdots + i_{J-1}}{\left(1-\frac{1}{J}\right)^{\frac{\gamma+\alpha}{\alpha}+i_1+\cdots+i_{J-1}+1}},$$

and similarly, for the sum in $i_J = 1, \ldots, c_J$ of (47), we have:

$$\lim_{c_J \to +\infty} \sum_{i_J=1}^{c_J} \left(\frac{\gamma}{\alpha} + i_1 + \cdots + i_{J-1}\right)_{(i_J)} \frac{\left(\frac{1}{J}\right)^{i_J} \mathscr{C}(c_J, i_J; \alpha)}{\sum_{i_J=1}^{c_J} \left(\frac{1}{J}\right)^{i_J} \mathscr{C}(c_J, i_J; \alpha)} \tag{49}$$
$$= \mathrm{e}^{-\frac{1}{J}} \frac{\frac{\gamma}{\alpha} + i_1 + \cdots + i_{J-1}}{\left(1-\frac{1}{J}\right)^{\frac{\gamma}{\alpha}+i_1+\cdots+i_{J-1}+1}}.$$

Now, consider the sum in $i_{J-1} = 1, \ldots, c_{J-1}$ of (46), combined with (48). That is,

$$\lim_{c_{J-1} \to +\infty} \sum_{i_{J-1}=1}^{c_{J-1}} \left(\frac{\gamma+\alpha}{\alpha} + i_1 + \cdots + i_{J-2}\right)_{(i_{J-1})} \mathrm{e}^{-\frac{1}{J}} \frac{\frac{\gamma+\alpha}{\alpha} + i_1 + \cdots + i_{J-1}}{\left(1-\frac{1}{J}\right)^{\frac{\gamma+\alpha}{\alpha}+i_1+\cdots+i_{J-1}+1}}$$
$$\times \frac{\left(\frac{1}{J}\right)^{i_{J-1}} \mathscr{C}(c_{J-1}, i_{J-1}; \alpha)}{\sum_{i_{J-1}=1}^{c_{J-1}} \left(\frac{1}{J}\right)^{i_{J-1}} \mathscr{C}(c_{J-1}, i_{J-1}; \alpha)}$$
$$= \sum_{i_{J-1} \geq 1} \left(\frac{\gamma+\alpha}{\alpha} + i_1 + \cdots + i_{J-2}\right)_{(i_{J-1})} \mathrm{e}^{-\frac{1}{J}} \frac{\frac{\gamma+\alpha}{\alpha} + i_1 + \cdots + i_{J-1}}{\left(1-\frac{1}{J}\right)^{\frac{\gamma+\alpha}{\alpha}+i_1+\cdots+i_{J-1}+1}} \mathrm{e}^{-\frac{1}{J}} \frac{\left(\frac{1}{J}\right)^{i_{J-1}-1}}{(i_{J-1}-1)!}$$
$$= \frac{\mathrm{e}^{-\frac{2}{J}}}{\left(1-\frac{1}{J}\right)^{\frac{\gamma+\alpha}{\alpha}+i_1+\cdots+i_{J-2}+2}} \sum_{i_{J-1} \geq 1} \left(\frac{\gamma+\alpha}{\alpha} + i_1 + \cdots + i_{J-2}\right)_{(i_{J-1}+1)} \frac{\left(\frac{1}{J-1}\right)^{i_{J-1}-1}}{(i_{J-1}-1)!}$$
$$= \frac{\mathrm{e}^{-\frac{2}{J}}}{\left(1-\frac{1}{J}\right)^{\frac{\gamma+\alpha}{\alpha}+i_1+\cdots+i_{J-2}+2}} \frac{\left(\frac{\gamma+\alpha}{\alpha} + i_1 + \cdots + i_{J-2}\right)_{(2)}}{\left(\frac{J-2}{J-1}\right)^{\frac{\gamma+\alpha}{\alpha}+i_1+\cdots+i_{J-2}+2}}$$

$$= \mathrm{e}^{-\frac{2}{J}} \frac{\left(\frac{\gamma+\alpha}{\alpha} + i_1 + \cdots + i_{J-2}\right)_{(2)}}{\left(1-\frac{2}{J}\right)^{\frac{\gamma+\alpha}{\alpha}+i_1+\cdots+i_{J-2}+2}}.$$

Similarly, consider the sum in $i_{J-1} = 1, \ldots, c_{J-1}$ of (47), together with (49), which leads to

$$\begin{aligned}
&\lim_{c_{J-1}\to+\infty} \sum_{i_{J-1}=1}^{c_{J-1}} \left(\frac{\gamma}{\alpha} + i_1 + \cdots + i_{J-2}\right)_{(i_{J-1})} \mathrm{e}^{-\frac{1}{J}} \frac{\frac{\gamma}{\alpha} + i_1 + \cdots + i_{J-1}}{\left(1-\frac{1}{J}\right)^{\frac{\gamma}{\alpha}+i_1+\cdots+i_{J-1}+1}} \\
&\quad \times \frac{\left(\frac{1}{J}\right)^{i_{J-1}} \mathscr{C}(c_{J-1}, i_{J-1}; \alpha)}{\sum_{i_{J-1}=1}^{c_{J-1}} \left(\frac{1}{J}\right)^{i_{J-1}} \mathscr{C}(c_{J-1}, i_{J-1}; \alpha)} \\
&\quad = \mathrm{e}^{-\frac{2}{J}} \frac{\left(\frac{\gamma}{\alpha} + i_1 + \cdots + i_{J-2}\right)_{(2)}}{\left(1-\frac{2}{J}\right)^{\frac{\gamma}{\alpha}+i_1+\cdots+i_{J-2}+2}}.
\end{aligned}$$

By proceeding recursively, for the sum in $i_{j+1} = 1, \ldots, c_{j+1}$ of (46), we find:

$$\lim_{c_{j+1}\to+\infty} \sum_{i_{j+1}=1}^{c_{j+1}} \left(\frac{\gamma+\alpha}{\alpha} + i_1 + \cdots + i_j\right)_{(i_{j+1})} \mathrm{e}^{-\frac{J-j-1}{J}} \frac{\left(\frac{\gamma+\alpha}{\alpha} + i_1 + \cdots + i_{j+1}\right)_{(J-j-1)}}{\left(1-\frac{J-j-1}{J}\right)^{\frac{\gamma+\alpha}{\alpha}+i_1+\cdots+i_{j+1}+J-j-1}} \tag{50}$$

$$\begin{aligned}
&\quad \times \frac{\left(\frac{1}{J}\right)^{i_{j+1}} \mathscr{C}(c_{j+1}, i_{j+1}; \alpha)}{\sum_{i_{j+1}=1}^{c_{j+1}} \left(\frac{1}{J}\right)^{i_{j+1}} \mathscr{C}(c_{j+1}, i_{j+1}; \alpha)} \\
&= \sum_{c_{j+1}\geq 1} \left(\frac{\gamma+\alpha}{\alpha} + i_1 + \cdots + i_j\right)_{(i_{j+1})} \mathrm{e}^{-\frac{J-j-1}{J}} \frac{\left(\frac{\gamma+\alpha}{\alpha} + i_1 + \cdots + i_{j+1}\right)_{(J-j-1)}}{\left(1-\frac{J-j-1}{J}\right)^{\frac{\gamma+\alpha}{\alpha}+i_1+\cdots+i_{j+1}+J-j-1}} \mathrm{e}^{-\frac{1}{J}} \frac{\left(\frac{1}{J}\right)^{i_{j+1}-1}}{(i_{j+1}-1)!} \\
&= \frac{\mathrm{e}^{-\frac{J-j}{J}}}{\left(1-\frac{J-j-1}{J}\right)^{\frac{\gamma+\alpha}{\alpha}+i_1+\cdots+i_j+J-j}} \sum_{i_{j+1}\geq 1} \left(\frac{\gamma+\alpha}{\alpha} + i_1 + \cdots + i_j\right)_{(i_{j+1}+J-j-1)} \frac{\left(\frac{1}{j+1}\right)^{i_{j+1}-1}}{(i_{j+1}-1)!} \\
&= \frac{\mathrm{e}^{-\frac{J-j}{J}}}{\left(1-\frac{J-j-1}{J}\right)^{\frac{\gamma+\alpha}{\alpha}+i_1+\cdots+i_j+J-j}} \frac{\left(\frac{\gamma+\alpha}{\alpha} + i_1 + \cdots + i_j\right)_{(J-j)}}{\left(\frac{j}{j+1}\right)^{\frac{\gamma+\alpha}{\alpha}+i_1+\cdots+i_j+J-j}} \\
&= \mathrm{e}^{-\frac{J-j}{J}} \frac{\left(\frac{\gamma+\alpha}{\alpha} + i_1 + \cdots + i_j\right)_{(J-j)}}{\left(1-\frac{J-j}{J}\right)^{\frac{\gamma+\alpha}{\alpha}+i_1+\cdots+i_j+J-j}},
\end{aligned}$$

and, similarly, the sum in $i_{j+1} = 1, \ldots, c_{j+1}$ of (47) it leads to the following:

$$\lim_{c_{j+1}\to+\infty} \sum_{i_{j+1}=1}^{c_{j+1}} \left(\frac{\gamma}{\alpha} + i_1 + \cdots + i_j\right)_{(i_{j+1})} \mathrm{e}^{-\frac{J-j-1}{J}} \frac{\left(\frac{\gamma}{\alpha} + i_1 + \cdots + i_{j+1}\right)_{(J-j-1)}}{\left(1-\frac{J-j-1}{J}\right)^{\frac{\gamma}{\alpha}+i_1+\cdots+i_{j+1}+J-j-1}} \tag{51}$$

$$\times \frac{\left(\frac{1}{J}\right)^{i_{j+1}} \mathscr{C}(c_{j+1}, i_{j+1}; \alpha)}{\sum_{i_{j+1}=1}^{c_{j+1}} \left(\frac{1}{J}\right)^{i_{j+1}} \mathscr{C}(c_{j+1}, i_{j+1}; \alpha)}$$

$$= \mathrm{e}^{-\frac{J-j}{J}} \frac{\left(\frac{\gamma}{\alpha}+i_1+\cdots+i_j\right)_{(J-j)}}{\left(1-\frac{J-j}{J}\right)^{\frac{\gamma}{\alpha}+i_1+\cdots+i_j+J-j}}.$$

Now, we consider the sum in $i_{j-1}=1,\ldots,c_{j-1}$ of (46), together with (50). That is,

$$\begin{aligned}
&\lim_{c_{j-1}\to+\infty} \sum_{i_{j-1}=1}^{c_{j-1}} \left(\frac{\gamma+\alpha}{\alpha}+i_1+\cdots+i_{j-2}\right)_{(i_{j-1})} \left(\frac{\gamma+\alpha}{\alpha}+i_1+\cdots+i_{j-1}\right)_{(i_j)} \\
&\quad\times \mathrm{e}^{-\frac{J-j}{J}} \frac{\left(\frac{\gamma+\alpha}{\alpha}+i_1+\cdots+i_j\right)_{(J-j)}}{\left(1-\frac{J-j}{J}\right)^{\frac{\gamma+\alpha}{\alpha}+i_1+\cdots+i_j+J-j}} \frac{\left(\frac{1}{J}\right)^{i_{j-1}} \mathscr{C}(c_{j-1},i_{j-1};\alpha)}{\sum_{i_{j-1}=1}^{c_{j-1}} \left(\frac{1}{J}\right)^{i_{j-1}} \mathscr{C}(c_{j-1},i_{j-1};\alpha)} \\
&= \sum_{i_{j-1}\geq 1} \left(\frac{\gamma+\alpha}{\alpha}+i_1+\cdots+i_{j-2}\right)_{(i_{j-1})} \left(\frac{\gamma+\alpha}{\alpha}+i_1+\cdots+i_{j-1}\right)_{(i_j)} \\
&\quad\times \mathrm{e}^{-\frac{J-j}{J}} \frac{\left(\frac{\gamma+\alpha}{\alpha}+i_1+\cdots+i_j\right)_{(J-j)}}{\left(1-\frac{J-j}{J}\right)^{\frac{\gamma+\alpha}{\alpha}+i_1+\cdots+i_j+J-j}} \mathrm{e}^{-\frac{1}{J}} \frac{\left(\frac{1}{J}\right)^{i_{j-1}-1}}{(i_{j-1}-1)!} \\
&= \frac{\mathrm{e}^{-\frac{J-j+1}{J}}}{\left(1-\frac{J-j}{J}\right)^{\frac{\gamma+\alpha}{\alpha}+i_j+i_1+\cdots+i_{j-2}+J-j+1}} \sum_{i_{j-1}\geq 1} \left(\frac{\gamma+\alpha}{\alpha}+i_1+\cdots+i_{j-2}\right)_{(i_{j-1}+i_j+J-j)} \frac{\left(\frac{1}{j}\right)^{i_{j-1}-1}}{(i_{j-1}-1)!} \\
&= \frac{\mathrm{e}^{-\frac{J-j+1}{J}}}{\left(1-\frac{J-j}{J}\right)^{\frac{\gamma+\alpha}{\alpha}+i_j+i_1+\cdots+i_{j-2}+J-j+1}} \frac{\left(\frac{\gamma+\alpha}{\alpha}+i_1+\cdots+i_{j-2}\right)_{(i_j+J-j+1)}}{\left(\frac{j-1}{j}\right)^{\frac{\gamma+\alpha}{\gamma}+i_j+i_1+\cdots+i_{j-2}+J-j+1}} \\
&= \mathrm{e}^{-\frac{J-j+1}{J}} \frac{\left(\frac{\gamma+\alpha}{\alpha}+i_1+\cdots+i_{j-2}\right)_{(i_j+J-j+1)}}{\left(\frac{j-1}{J}\right)^{\frac{\gamma+\alpha}{\gamma}+i_j+i_1+\cdots+i_{j-2}+J-j+1}}.
\end{aligned}$$

Similarly, consider the sum in $i_{j-1}=1,\ldots,c_{j-1}$ of (47), together with (51), which leads to

$$\begin{aligned}
&\lim_{c_{j-1}\to+\infty} \sum_{i_{j-1}=1}^{c_{j-1}} \left(\frac{\gamma}{\alpha}+i_1+\cdots+i_{j-2}\right)_{(i_{j-1})} \left(\frac{\gamma}{\alpha}+i_1+\cdots+i_{j-1}\right)_{(i_j)} \\
&\quad\times \mathrm{e}^{-\frac{J-j}{J}} \frac{\left(\frac{\gamma}{\alpha}+i_1+\cdots+i_j\right)_{(J-j)}}{\left(1-\frac{J-j}{J}\right)^{\frac{\gamma}{\alpha}+i_1+\cdots+i_j+J-j}} \frac{\left(\frac{1}{J}\right)^{i_{j-1}} \mathscr{C}(c_{j-1},i_{j-1};\alpha)}{\sum_{i_{j-1}=1}^{c_{j-1}} \left(\frac{1}{J}\right)^{i_{j-1}} \mathscr{C}(c_{j-1},i_{j-1};\alpha)} \\
&= \mathrm{e}^{-\frac{J-j+1}{J}} \frac{\left(\frac{\gamma}{\alpha}+i_1+\cdots+i_{j-2}\right)_{(i_j+J-j+1)}}{\left(\frac{j-1}{J}\right)^{\frac{\gamma}{\gamma}+i_j+i_1+\cdots+i_{j-2}+J-j+1}}.
\end{aligned}$$

By proceeding recursively, we arrive to the sum in $i_1=1,\ldots,c_1$ of (46). That is,

$$\lim_{c_1\to+\infty} \sum_{i_1=1}^{c_1} \left(\frac{\gamma+\alpha}{\alpha}\right)_{(i_1)} \mathrm{e}^{-\frac{J-2}{J}} \frac{\left(\frac{\gamma+\alpha}{\alpha}+i_1\right)_{(i_j+J-2)}}{\left(\frac{2}{J}\right)^{\frac{\gamma+\alpha}{\alpha}+i_j+i_1+J-2}} \frac{\left(\frac{1}{J}\right)^{i_1} \mathscr{C}(c_1,i_1;\alpha)}{\sum_{i_1=1}^{c_1} \left(\frac{1}{J}\right)^{i_1} \mathscr{C}(c_1,i_1;\alpha)} \tag{52}$$

$$
\begin{aligned}
&= \sum_{i_1 \geq 1} \binom{\gamma+\alpha}{\alpha}_{(i_1)} \mathrm{e}^{-\frac{J-2}{J}} \frac{\left(\frac{\gamma+\alpha}{\alpha}+i_1\right)_{(i_j+J-2)}}{\left(\frac{2}{J}\right)^{\frac{\gamma+\alpha}{\alpha}+i_j+i_1+J-2}} \mathrm{e}^{-\frac{1}{J}} \frac{\left(\frac{1}{J}\right)^{i_1-1}}{(i_1-1)!} \\
&= \frac{\mathrm{e}^{-\frac{J-1}{J}}}{\left(\frac{2}{J}\right)^{\frac{\gamma+\alpha}{\alpha}+i_j+J-1}} \sum_{i_1 \geq 1} \binom{\gamma+\alpha}{\alpha}_{(i_1+i_j+J-2)} \frac{\left(\frac{1}{2}\right)^{i_1-1}}{(i_1-1)!} \\
&= \frac{\mathrm{e}^{-\frac{J-1}{J}}}{\left(\frac{2}{J}\right)^{\frac{\gamma+\alpha}{\alpha}+i_j+J-1}} \frac{\left(\frac{\gamma+\alpha}{\alpha}\right)_{(i_j+J-1)}}{\left(\frac{1}{2}\right)^{\frac{\gamma+\alpha}{\alpha}+i_j+J-1}} \\
&= \mathrm{e}^{-\frac{J-1}{J}} \frac{\left(\frac{\gamma+\alpha}{\alpha}\right)_{(i_j+J-1)}}{\left(\frac{1}{J}\right)^{\frac{\gamma+\alpha}{\alpha}+i_j+J-1}}
\end{aligned}
$$

and, similarly, we arrive to the sum in $i_1 = 1, \ldots, c_1$ of (47), which leads to

$$
\begin{aligned}
&\lim_{c_1 \to +\infty} \sum_{i_1=1}^{c_1} \binom{\gamma}{\alpha}_{(i_1)} \mathrm{e}^{-\frac{J-2}{J}} \frac{\left(\frac{\gamma}{\alpha}+i_1\right)_{(i_j+J-2)}}{\left(\frac{2}{J}\right)^{\frac{\gamma}{\alpha}+i_j+i_1+J-2}} \frac{\left(\frac{1}{J}\right)^{i_1} \mathscr{C}(c_1, i_1; \alpha)}{\sum_{i_1=1}^{c_1} \left(\frac{1}{J}\right)^{i_1} \mathscr{C}(c_1, i_1; \alpha)} \\
&\quad = \mathrm{e}^{-\frac{J-1}{J}} \frac{\left(\frac{\gamma}{\alpha}\right)_{(i_j+J-1)}}{\left(\frac{1}{J}\right)^{\frac{\gamma}{\alpha}+i_j+J-1}}.
\end{aligned} \tag{53}
$$

From (52) and (53), and (45),

$$
\begin{aligned}
&\lim_{\mathbf{c}_{-j} \to +\infty} \frac{\gamma}{J} \binom{c_j}{l} (1-\alpha)_{(l)} \frac{\Gamma\left(\frac{\gamma+\alpha}{\alpha}\right)}{\Gamma\left(\frac{\gamma}{\alpha}\right)} \frac{N_{\gamma,\alpha,J}(l; c_1, \ldots, c_J)}{D_{\gamma,\alpha,J}(c_1, \ldots, c_J)} \\
&= \frac{\gamma}{J} \binom{c_j}{l} (1-\alpha)_{(l)} \frac{\Gamma\left(\frac{\gamma+\alpha}{\alpha}\right)}{\Gamma\left(\frac{\gamma}{\alpha}\right)} \frac{\sum_{i_j=1}^{c_j-l} \left(\frac{1}{J}\right)^{i_j} \mathscr{C}(c_j-l, i_j; \alpha) \mathrm{e}^{-\frac{J-1}{J}} \frac{\left(\frac{\gamma+\alpha}{\alpha}\right)_{(i_j+J-1)}}{\left(\frac{1}{J}\right)^{\frac{\gamma+\alpha}{\alpha}+i_j+J-1}}}{\sum_{i_j=1}^{c_j+1} \left(\frac{1}{J}\right)^{i_j} \mathscr{C}(c_j+1, i_j; \alpha) \mathrm{e}^{-\frac{J-1}{J}} \frac{\left(\frac{\gamma}{\alpha}\right)_{(i_j+J-1)}}{\left(\frac{1}{J}\right)^{\frac{\gamma}{\alpha}+i_j+J-1}}} \\
&= \gamma \binom{c_j}{l} (1-\alpha)_{(l)} \frac{\Gamma\left(\frac{\gamma+\alpha}{\alpha}+J-1\right)}{\Gamma\left(\frac{\gamma}{\alpha}+J-1\right)} \frac{\sum_{i_j=1}^{c_j-l} \mathscr{C}(c_j-l, i_j; \alpha) \left(\frac{\gamma+\alpha}{\alpha}+J-1\right)_{(i_j)}}{\sum_{i_j=1}^{c_j+1} \mathscr{C}(c_j+1, i_j; \alpha) \left(\frac{\gamma}{\alpha}+J-1\right)_{(i_j)}} \\
&= \gamma \binom{c_j}{l} (1-\alpha)_{(l)} \frac{\Gamma\left(\frac{\gamma+\alpha}{\alpha}+J-1\right)}{\Gamma\left(\frac{\gamma}{\alpha}+J-1\right)} \frac{(\gamma+\alpha+J\alpha-\alpha)_{(c_j-l)}}{(\gamma+J\alpha-\alpha)_{(c_j+1)}},
\end{aligned}
$$

where the sums over i_j follows from Charalambides (2005, Equation 2.49).

A.8 Proof of Corollary 1

The proof follows the same arguments developed in Appendix A.5. According to (17), it is sufficient to compute the conditional probability of $h(X_{n+1})$, given $\mathbf{C}_J$, as the conditional probability of $f_{X_{n+1}}$, given $\mathbf{C}_J$ and $h(X_{n+1})$ is available from (8). For a PK prior,

$$
\Pr[h(X_{n+1}) = j \,|\, \mathbf{C}_J = \mathbf{c}] = \frac{\Pr[\mathbf{C}_J = \mathbf{c}, h(X_{n+1}) = j]}{\Pr[\mathbf{C}_J = \mathbf{c}]}
$$

for all $j \in [J]$, where, from (34),

$$\Pr[\mathbf{C}_J = \mathbf{c}, h(X_{n+1}) = j] = \binom{n}{c_1, \ldots, c_J} \mathbb{E}\left[P(D_j)^{c_j+1} \prod_{k \neq j} P(D_k)^{c_k} \right] \tag{54}$$
$$= Z_T \binom{n}{c_1, \ldots, c_j} \int_{\mathbb{R}_+} \frac{u^{n+\gamma}}{\Gamma(n+\gamma+1)} (-1)^{c+1} \frac{\mathrm{d}^{c_j+1}}{\mathrm{d}z^{c_j+1}} e^{-\theta\psi(z)/J}|_{(u+\beta)}$$
$$\times \prod_{k \neq j} (-1)^{c_k} \frac{\mathrm{d}^{c_k}}{\mathrm{d}z^{c_k}} e^{-\theta\psi(z)/J}|_{(u+\beta)} \mathrm{d}u$$

and

$$\Pr[\mathbf{C}_J = \mathbf{c}] = \binom{n}{c_1, \ldots, c_J} \mathbb{E}\left[P(D_j)^{c_j} \prod_{k \neq j} P(D_k)^{c_k} \right] \tag{55}$$
$$= Z_T \binom{n}{c_1, \ldots, c_j} \int_{\mathbb{R}_+} \frac{u^{n+\gamma}}{\Gamma(n+\gamma+1)} \prod_{k=1}^{J} (-1)^{c_k} \frac{\mathrm{d}^{c_k}}{\mathrm{d}z^{c_k}} e^{-\theta\psi(z)/J}|_{(u+\beta)} \mathrm{d}u.$$

In the case of a PYP, we have $\beta = 0$, $\psi(u) = u^\alpha$ and

$$\kappa(u, l) = \alpha u^{\alpha-l} \frac{\Gamma(l-\alpha)}{\Gamma(1-\alpha)} = \alpha u^{\alpha-l} (1-\alpha)_{(l-1)}.$$

Then, by applying to (54) and (55) the same arguments of Appendix A.5, we obtain:

$$\Pr[\mathbf{C}_J = \mathbf{c}, h(X_{n+1}) = j] = \binom{n}{c_1, \ldots, c_J} \sum_{\boldsymbol{i} \in \mathcal{S}(\mathbf{c}, j, 1)} \frac{\frac{(\frac{\gamma}{\alpha})_{(|\boldsymbol{i}|)}}{J^{|\boldsymbol{i}|}}}{(\gamma)_{(n+1)}} \prod_{k=1}^{J} \mathscr{C}(c_k + \delta_{k,j}, i_k; \alpha)$$

and

$$\Pr[\mathbf{C}_J = \mathbf{c}] = \binom{n}{c_1, \ldots, c_J} \sum_{\boldsymbol{i} \in \mathcal{S}(\mathbf{c}, j, 0)} \frac{\frac{(\frac{\gamma}{\alpha})_{(|\boldsymbol{i}|)}}{J^{|\boldsymbol{i}|}}}{(\gamma)_{(n)}} \prod_{k=1}^{J} \mathscr{C}(c_k, i_k; \alpha).$$

Then,

$$\Pr[h(X_{n+1}) = j \,|\, \mathbf{C}_J = \mathbf{c}] = \frac{\binom{n}{c_1, \ldots, c_J} \sum_{\boldsymbol{i} \in \mathcal{S}(\mathbf{c}, j, 1)} \frac{\frac{(\frac{\gamma}{\alpha})_{(|\boldsymbol{i}|)}}{J^{|\boldsymbol{i}|}}}{(\gamma)_{(n+1)}} \prod_{k=1}^{J} \mathscr{C}(c_k + \delta_{k,j}, i_k; \alpha)}{\binom{n}{c_1, \ldots, c_J} \sum_{\boldsymbol{i} \in \mathcal{S}(\mathbf{c}, j, 0)} \frac{\frac{(\frac{\gamma}{\alpha})_{(|\boldsymbol{i}|)}}{J^{|\boldsymbol{i}|}}}{(\gamma)_{(n)}} \prod_{k=1}^{J} \mathscr{C}(c_k, i_k; \alpha)} \tag{56}$$
$$= \frac{1}{\gamma+n} \frac{\sum_{\boldsymbol{i} \in \mathcal{S}(\mathbf{c}, j, 1)} \frac{(\frac{\gamma}{\alpha})_{(|\boldsymbol{i}|)}}{J^{|\boldsymbol{i}|}} \prod_{k=1}^{J} \mathscr{C}(c_k + \delta_{k,j}, i_k; \alpha)}{\sum_{\boldsymbol{i} \in \mathcal{S}(\mathbf{c}, j, 0)} \frac{(\frac{\gamma}{\alpha})_{(|\boldsymbol{i}|)}}{J^{|\boldsymbol{i}|}} \prod_{k=1}^{J} \mathscr{C}(c_k, i_k; \alpha)}.$$

Finally, we combine (17) with (8) and (56). In particular, for $l \in [n]$,

$$\Pr[f_{X_{n+1}} = l \,|\, \mathbf{C}_J = \mathbf{c}]$$

$$= \sum_{j=1}^{J} \frac{\gamma}{J} \binom{c_j}{l} (1-\alpha)_{(l)} \frac{\sum_{\boldsymbol{i} \in \mathcal{S}(\boldsymbol{c},j,-l)} \frac{\Gamma\left(\frac{\gamma+\alpha}{\alpha}+|\boldsymbol{i}|\right)}{J^{|\boldsymbol{i}|}} \prod_{k=1}^{J} \mathscr{C}(c_k - l\delta_{k,j}, i_k; \alpha)}{\sum_{\boldsymbol{i} \in \mathcal{S}(\boldsymbol{c},j,1)} \frac{\Gamma\left(\frac{\gamma}{\alpha}+|\boldsymbol{i}|\right)}{J^{|\boldsymbol{i}|}} \prod_{k=1}^{J} \mathscr{C}(c_k + \delta_{k,j}, i_k; \alpha)}$$

$$\times \frac{1}{\gamma+n} \frac{\sum_{\boldsymbol{i} \in \mathcal{S}(\mathbf{c},j,1)} \frac{\left(\frac{\gamma}{\alpha}\right)_{(|\boldsymbol{i}|)}}{J^{|\boldsymbol{i}|}} \prod_{k=1}^{J} \mathscr{C}(c_k + \delta_{k,j}, i_k; \alpha)}{\sum_{\boldsymbol{i} \in \mathcal{S}(\mathbf{c},j,0)} \frac{\left(\frac{\gamma}{\alpha}\right)_{(|\boldsymbol{i}|)}}{J^{|\boldsymbol{i}|}} \prod_{k=1}^{J} \mathscr{C}(c_k, i_k; \alpha)}$$

$$= \frac{\frac{\gamma}{J}}{\gamma+n} (1-\alpha)_{(l)} \sum_{j=1}^{J} \binom{c_j}{l} \frac{\sum_{\boldsymbol{i} \in \mathcal{S}(\boldsymbol{c},j,-l)} \frac{\Gamma\left(\frac{\gamma+\alpha}{\alpha}+|\boldsymbol{i}|\right)}{J^{|\boldsymbol{i}|}} \prod_{k=1}^{J} \mathscr{C}(c_k - l\delta_{k,j}, i_k; \alpha)}{\sum_{\boldsymbol{i} \in \mathcal{S}(\mathbf{c},j,0)} \frac{\Gamma\left(\frac{\gamma}{\alpha}+|\boldsymbol{i}|\right)}{J^{|\boldsymbol{i}|}} \prod_{k=1}^{J} \mathscr{C}(c_k, i_k; \alpha)}.$$

A.9 Proof of Equation (19)

The proof follows the same arguments developed in Appendix A.6, exploiting the behaviour of generalized factorial coefficients as $\alpha \to 0$. From (18), we write that

$$\lim_{\alpha \to 0} \Pr[f_{X_{n+1}} = l \,|\, \mathbf{C}_J = \mathbf{c}]$$

$$= \lim_{\alpha \to 0} \frac{\frac{\gamma}{J}}{\gamma+n} (1-\alpha)_{(l)} \sum_{j=1}^{J} \binom{c_j}{l} \frac{\sum_{\boldsymbol{i} \in \mathcal{S}(\boldsymbol{c},j,-l)} \frac{\Gamma\left(\frac{\gamma+\alpha}{\alpha}+|\boldsymbol{i}|\right)}{J^{|\boldsymbol{i}|}} \prod_{k=1}^{J} \mathscr{C}(c_k - l\delta_{k,j}, i_k; \alpha)}{\sum_{\boldsymbol{i} \in \mathcal{S}(\boldsymbol{c},j,0)} \frac{\Gamma\left(\frac{\gamma}{\alpha}+|\boldsymbol{i}|\right)}{J^{|\boldsymbol{i}|}} \prod_{k=1}^{J} \mathscr{C}(c_k, i_k; \alpha)}$$

$$= \lim_{\alpha \to 0} \frac{\frac{\gamma}{J}}{\gamma+n} (1-\alpha)_{(l)} \sum_{j=1}^{J} \binom{c_j}{l} \frac{\sum_{\boldsymbol{i} \in \mathcal{S}(\boldsymbol{c},j,-l)} \frac{\prod_{t=0}^{|\boldsymbol{i}-1|}(\gamma+\alpha+\alpha t)}{J^{|\boldsymbol{i}|}} \prod_{k=1}^{J} \frac{\mathscr{C}(c_k - l\delta_{k,j}, i_k; \alpha)}{\alpha^{i_k}}}{\sum_{\boldsymbol{i} \in \mathcal{S}(\boldsymbol{c},j,0)} \frac{\prod_{t=0}^{|\boldsymbol{i}-1|}(\gamma+\alpha t)}{J^{|\boldsymbol{i}|}} \prod_{k=1}^{J} \frac{\mathscr{C}(c_k, i_k; \alpha)}{\alpha^{i_k}}}$$

$$= \frac{\frac{\gamma}{J}}{\gamma+n} \sum_{j=1}^{J} \binom{c_j}{l} l! \frac{\sum_{\boldsymbol{i} \in \mathcal{S}(\boldsymbol{c},j,-l)} \left(\frac{\gamma}{J}\right)^{|\boldsymbol{i}|} \prod_{k=1}^{J} |s(c_k - l\delta_{k,j}, i_k)|}{\sum_{\boldsymbol{i} \in \mathcal{S}(\boldsymbol{c},j,0)} \left(\frac{\gamma}{J}\right)^{\boldsymbol{i}} \prod_{k=1}^{J} |s(c_k, i_k)|}.$$

Exploiting the definition of signless Stirling numbers of the first type as the coefficients of the series expansion of a rising factorial (Charalambides, 2005, Equation 2.4), we obtain

$$\lim_{\alpha \to 0} \Pr[f_{X_{n+1}} = l \,|\, \mathbf{C}_J = \mathbf{c}] = \frac{\frac{\gamma}{J}}{\gamma+n} \sum_{j=1}^{J} \binom{c_j}{l} l! \frac{\prod_{k=1}^{J} \left(\frac{\gamma}{J}\right)_{(c_k - l\delta_{k,j})}}{\prod_{k=1}^{J} \left(\frac{\gamma}{J}\right)_{(c_k)}}$$

$$= \frac{\frac{\gamma}{J}}{\gamma+n} \sum_{l=1}^{J} \frac{(c_j - l + 1)_{(l)}}{\left(\frac{\gamma}{J} + c_j - l\right)_{(l)}},$$

which coincides with the posterior in (19) by setting $\theta = \gamma$.

A.10 Proof of Proposition 6

Without loss of generality, assume $A_{i,k} \in \mathbb{N}_0 := \{0, 1, \ldots\}$. From the Poisson process representation of CRMs and the marking theorem (Kingman, 1993), $\tilde{N} := \{(\omega_k, J_k, (A_{i,k})_{i=1}^{n+1})\}_{k \geq 1}$ is a Poisson process on $\mathbb{S} \times \mathbb{R}_+ \times \mathbb{N}_0^{n+1}$. Consider now thinned processes $\tilde{N}_j$, $j = 1, \ldots, J$ obtained from $\tilde{N}$ by taking only those points for which $\omega_k \in D_j$. By the coloring theorem (Kingman, 1993), the $\tilde{N}_j$'s are independent Poisson processes.

Now observe that the random variables $f_{Y_{n+1,r}}$, $X_{n+1}(Y_{n+1,r})$ depend only on $\tilde{N}_{h(Y_{n+1,r})}$. Similarly, each (C_j, Bj) depends only on $\tilde{N}_j$ and the independence is preserved also when marginalizing the $\tilde{N}_j$'s. Hence, $(f_{Y_{n+1,r}}, X_{n+1}(Y_r), C_{h(Y_{n+1,r})}, B_{h(Y_r)})$ are independent of all other C_k's $(k \neq h(Y_r))$ and all of B_k $(k \neq h(Y_r))$, which yields the proof.

A.11 Proof of Theorem 7

We evaluate (25) by writing

$$\begin{aligned}\Pr\left[f_{Y_{n+1,r}} = l, X_{n+1}(Y_{n+1,r}) = a \,|\, h(Y_{n+1,r}) = j, C_j = c, B_j = b\right] \\ = \frac{\Pr\left[f_{Y_{n+1,r}} = l, X_{n+1}(Y_{n+1,r}) = a, C_j = c, B_j = b \,|\, h(Y_{n+1,r}) = j\right]}{\Pr\left[C_j = c, B_j = b \,|\, h(Y_{n+1,r}) = j\right]}\end{aligned}$$

and computing the numerator and denominator separately. In the following, we will denote by D_ω the preimage of $h(\omega)$, for $\omega \in \mathbb{S}$.

Denominator. From the Poisson process representation of CRMs and the marking theorem, $N := \left\{(\omega_k, S_k, \{A_{i,k}\}_{i=1}^{n+1}\right\}_{k\geq 1}$ is a Poisson process on $\mathbb{S} \times \mathbb{R}_+ \times \mathbb{N}^{n+1}$ with intensity

$$\theta G_0(\mathrm{d}w)\left[\prod_{i=1}^{n+1} G_A(\mathrm{d}a_i \,|\, s)\right]\rho(s)\mathrm{d}s.$$

Then, by the colouring theorem (see, e.g., Chapter 5 in Kingman, 1993), we have that selecting from N only those points for which $\omega_k \in D_j$ (i.e., the points whose features' hashes coincide with the hash of $Y_{n+1,r}$), leads to a point process $N' := \left\{(\omega'_k, S'_k, \{A'_{i,k}\}_{i=1}^{n+1}\right\}_{k\geq 1}$ which is Poisson on $D_j \times \mathbb{R}_+ \times \mathbb{N}^{n+1}$ with intensity

$$\frac{\theta}{J}\bar{G}_j(\mathrm{d}w)\left[\prod_{i=1}^{n+1} G_A(\mathrm{d}a_i \,|\, s)\right]\rho(s)\mathrm{d}s, \tag{57}$$

where $\bar{G}_j(\mathrm{d}w)$ is G_0 truncated on D_{ω^*} and re-normalized. Then,

$$\begin{aligned}\Pr\left[C_j = c, B_j = b\right] &= \Pr\left[\sum_{k\geq 1}\mathrm{I}[\omega_k \in D_j]\sum_{i=1}^{n} A_{i,k} = c, \sum_{k\geq 1}\mathrm{I}[\omega_k \in D_j]A_{i,n+1} = b\right] \\ &= \Pr\left[\sum_{k\geq 1}\sum_{i=1}^{m} A'_{i,k} = c, \sum_{k\geq 1} A'_{n+1,k} = b\right].\end{aligned}$$

Observe that ω_k is not involved in the last probability, so we can marginalize with respect to it and consider the point process $\{J'_k, (A'_{i,k})_{i=1}^{n+1}\}_{k\geq 1}$ on $\mathbb{R}_+ \times \mathbb{N}_0^{n+1}$ with intensity $\theta J^{-1}\left[\prod_{i=1}^{n+1} G_A(\mathrm{d}a_i \,|\, s)\right]\rho(s)\mathrm{d}s$.

Numerator. We start by considering $Y_{n+1,r} = \omega^*$ fixed. Let B_{ω^*} be a ball of radius ε centered in ω^*. We compute

$$\Pr\left[\sum_{i=1}^{n} X_i(B_{\omega^*}) = l, X_{n+1}(B_{\omega^*}) = a, C_j = c, B_j = b \,|\, h(\omega^*) = j\right],$$

which converges to the numerator in (25) as $\varepsilon \to 0$. By the definition of B_j and C_j we have

$$\begin{aligned}\Pr&\left[\sum_{i=1}^{n} X_i(B_{\omega^*}) = l, X_{n+1}(B_{\omega^*}) = a, C_j = c, B_j = b\right] \\ &= \Pr\Big[\sum_{i=1}^{n} X_i(B_{\omega^*}) = l, X_{n+1}(B_{\omega^*}) = a, \\ &\quad \sum_{i=1}^{n} X_i(D_j \setminus B_{\omega^*}) = (c-l), X_{n+1}(D_j \setminus B_{\omega^*}) = (b-a)\Big].\end{aligned} \tag{58}$$

Further, $(X_i(B_{\omega^*}))_{i\geq 1}$ is a collection of random variables independent of $(X_i(D_j \setminus B_{\omega^*}))_{i\geq 1}$. Therefore, we can consider the first two events and the last two events separately.

For the last two, arguing as in the denominator case, we have that as $\varepsilon \to 0$

$$\begin{aligned}\Pr\Big[\sum_{i=1}^{n} X_i(D_j \setminus B_{\omega^*}) = (c-l), X_{n+1}(D_j \setminus B_{\omega^*}) = (b-a)\Big] \\ \longrightarrow \Pr\left[\sum_{k\geq 1}\sum_{i=1}^{n} A'_{i,k} = c-l, \sum_{k\geq 1} A'_{n+1,k} = b-a\right],\end{aligned}$$

where the A'_{ik}'s come from the Poisson process $\{J'_k, (A'_{i,k})_{i=1}^{n+1}\}_{k\geq 1}$ on $\mathbb{R}_+ \times \mathbb{N}_0^{n+1}$ with intensity $\theta J^{-1}\left[\prod_{i=1}^{n+1} G_A(\mathrm{d}a_i \,|\, s)\right]\rho(s)\mathrm{d}s$.

For the first two events instead, an application of Campbell's theorem yields

$$\begin{aligned}\Pr&\Big[\sum_{i=1}^{n} X_i(B_{\omega^*}) = l, X_{n+1}(B_{\omega^*}) = a\Big] \\ &= \mathbb{E}\left[\int I\left[\left\{\sum_{i=1}^{n} a_i\right\} = l\right] I[a_{n+1} = a] I[\omega \in B_{\omega^*}] N(\mathrm{d}s\,\mathrm{d}\boldsymbol{a}\,\mathrm{d}\omega)\right] \\ &= \theta G_0(B_{\omega^*}) \int \Pr\left[\sum_{i=1}^{n} \tilde{A}_i = l, \tilde{A}_{n+1} = a \,|\, s\right]\rho(s)\mathrm{d}s.\end{aligned}$$

The proof concludes by letting $\varepsilon \to 0$ and noting that $\int_{D_j} G_0(\mathrm{d}\omega^*) = J^{-1}$.

A.12 Proof of Proposition 8

The closeness of the Poisson distribution under convolution entails that $\sum_{i=1}^{n} \tilde{A}_i \,|\, s \sim \mathrm{Poi}(n\lambda s)$. Similarly, $\sum_{k\geq 1}\sum_{i=1}^{n} A'_{i,k} \sim \mathrm{Poi}(n\lambda T')$ where $T' = \sum_{k\geq 1} J'_k$. Then,

$$\int_{\mathbb{R}_+} \Pr\left[\sum_{i=1}^{n} \tilde{A}_i = l, \tilde{A}_{n+1} = a \,|\, s\right]\rho(s)\mathrm{d}s = \int_{\mathbb{R}_+} \frac{1}{l!}\frac{1}{a!}(nrs)^l e^{-nrs}(\lambda s)^a e^{-\lambda s}\rho(s)\mathrm{d}s$$

$$= \frac{n^l \lambda^{l+a}}{l!a!} \kappa(l+a, (n+1)\lambda).$$

Moreover

$$\begin{aligned}
\Pr\left[\sum_{k\geq 1}\sum_{i=1}^{n} A'_{i,k} = c, \sum_{k\geq 1} A'_{n+1,k} = b\right] &= \frac{n^c \lambda^{c+b}}{c!b!} \mathbb{E}_{T'}\left[e^{-(n+1)\lambda T'} (T')^{c+b}\right] \\
&= \frac{n^c \lambda^{c+b}}{c!b!} (-1)^{c+b} \frac{\mathrm{d}^{c+b}}{\mathrm{d}((n+1)\lambda)^{c+b}} \mathbb{E}_{\mu'}[e^{-(n+1)\lambda T'}] \\
&= \frac{n^c \lambda^{c+b}}{c!b!} (-1)^{c+b} \frac{\mathrm{d}^{c+b}}{\mathrm{d}z^{c+b}} \exp\left(-\theta/J\psi(z)\right)\big|_{(n+1)\lambda}.
\end{aligned}$$

In a similar fashion,

$$\begin{aligned}
&\Pr\left[\sum_{k\geq 1}\sum_{i=1}^{n} A'_{i,k} = c-l, \sum_{k\geq 1} A'_{n+1,k} = b-a\right] \\
&\qquad \frac{n^{c-l} \lambda^{c-l+b-a}}{(c-l)!(b-a)!} (-1)^{c-l+b-a} \frac{\mathrm{d}^{c-l+b-a}}{\mathrm{d}z^{c-l+b-a}} \exp\left(-\theta/J\psi(z)\right)\big|_{(n+1)\lambda}.
\end{aligned} \tag{59}$$

Combining these expressions together yields the proof.

A.13 Proof of Theorem 9

Observe that $(J'_k)_{k\geq 1}$ in Theorem 7 is a Poisson process on $\mathbb{R}_+$ with intensity $\theta/J\rho(s)\mathrm{d}s$. Consider now the numerator in (26). Let $S_n := \sum_{k\geq 1}\sum_{i=1}^{n} A'_{i,k}$ and $Z := \sum_{k\geq 1} A'_{n+1,k}$. Conditional on $(J'_k)_{k\geq 1}$, $S_{n,k}$ is Poisson-binomial with parameters $J'_1, \ldots, J'_1, J'_2, \ldots, J'_2, \ldots$, where each J'_k appears exactly n times. Similarly, $Z \,|\, (J'_k)_{k\geq 1}$ is Poisson-binomial with parameters $J_1, J_2, \ldots$. Let $\tilde{S}_n \,|\, (J'_k)_{k\geq 1} \sim \mathrm{Poi}(nT')$ and $\tilde{Z} \,|\, (J'_k)_{k\geq 1} \sim \mathrm{Poi}(T')$ where $T' = \sum_k J'_k$. Then, by Le Cam (1960), conditionally to $(J'_k)_{k\geq 1}$, $S_n \approx \tilde{S}_n$, $Z \approx \tilde{Z}$. Hence

$$\Pr[S_n = c-l, Z = b-1] \approx \Pr[\tilde{S}_n = c-l, \tilde{Z} = b-1].$$

Then, from (59) we get

$$\begin{aligned}
&\Pr[\tilde{S}_n = c-l, \tilde{Z} = b-1] = \\
&\qquad \frac{n^{c-l} \lambda^{c-l+b-1}}{(c-l)!(b-1)!} (-1)^{c-l+b-1} \frac{\mathrm{d}^{c-l+b-1}}{\mathrm{d}z^{c-l+b-1}} \exp\left(-\theta/J\psi(z)\right)\big|_{(n+1)\lambda}.
\end{aligned}$$

Moreover,

$$\begin{aligned}
\Pr\left[\sum_{i=1}^{n} \tilde{A}_i = l, \tilde{A}_{n+1} = 1 \,|\, s\right] &= \binom{n}{l} \Pr[\tilde{A}_1 = 1, \ldots, \tilde{A}_l = 1, \tilde{A}_{n+1} = 1, \tilde{A}_{l+1} = 0, \ldots, \tilde{A}_n = 0] \\
&= \binom{n}{l} s^{n+1} (1-s)^{m-l} = \binom{n}{l} \left(s^{l+1} - s^{n+1}\right).
\end{aligned}$$

Integrating this with respect to $\rho(s)\mathrm{d}s$ and ignoring multiplicative terms yields (26).

To prove the error bound, for ease of notation, let $f_Y \equiv f_{Y_{n+1,r}}$, $X \equiv X_{n+1}(Y_{n+1,r})$ and let $M' = (J_k)_{k\geq 1}$. Further, note that (f_Y, X) is independent of (S_n, Z) and $(\tilde{S}_n, \tilde{Z})$. Then,

$$\begin{aligned}
&TV\left(\left[f_Y, X, S_n, Z\right], \left[f_Y, X, \tilde{S}_n, \tilde{Z}\right]\right) \\
&= \sum_{l,x,s,z} \left| \Pr\left[f_Y = l, X = x, S_n = s, Z = z\right] - \Pr\left[f_Y = l, X = x, \tilde{S}_n = s, \tilde{Z} = z\right] \right| \\
&= \sum_{l,x,s,z} \Pr\left[f_Y = l, X = x\right] \left| \Pr\left[S_n = s, Z = z\right] - \Pr\left[\tilde{S}_n = s, \tilde{Z} = z\right] \right| \\
&\leq \sum_{s,z} \left| \mathbb{E}\Pr\left[S_n = s, Z = z \,|\, M'\right] - \mathbb{E}\Pr\left[\tilde{S}_n = s, \tilde{Z} = z \,|\, M'\right] \right| \\
&\leq \mathbb{E} \sum_{s,z} \left| \Pr\left[S_n = s, Z = z \,|\, M'\right] - \Pr\left[\tilde{S}_n = s, \tilde{Z} = z \,|\, M'\right] \right|,
\end{aligned}$$

where the last inequality follows from Jensen's inequality and an application of the Fubini theorem. Then by the conditional independence between S_n and Z, and $\tilde{S}_n$ and $\tilde{Z}$ we get

$$\begin{aligned}
&\mathbb{E} \sum_{s,z} \left| \Pr\left[S_n = s, Z = z \,|\, M'\right] - \Pr\left[\tilde{S}_n = s, \tilde{Z} = z \,|\, M'\right] \right| \\
&\qquad = \mathbb{E}\left[TV(S_n \,|\, M', \tilde{S}_n \,|\, M') + TV(Z \,|\, M', \tilde{Z} \,|\, M') \right].
\end{aligned}$$

To upper bound the error, we start from Eq. (5.5) in Steele (1994) so that

$$\begin{aligned}
TV(S_n \,|\, M', \tilde{S}_n \,|\, M') &\leq \frac{(1 - e^{-nT'})}{nT'} n \sum_{k\geq 1} J_k'^2 \\
TV(Z \,|\, M', \tilde{Z} \,|\, M') &\leq \frac{(1 - e^{-T'})}{T'} \sum_{k\geq 1} J_k'^2,
\end{aligned}$$

so that

$$\begin{aligned}
&\mathbb{E}\left[TV(S_n \,|\, M', \tilde{S}_n \,|\, M') + TV(Z \,|\, M', \tilde{Z} \,|\, M') \right] \\
&\qquad \leq \mathbb{E}\left[\frac{2}{T'} \sum_{k\geq 1} J_k'^2 \right] = 2 \int_{\mathbb{R}_+} \mathbb{E}\left[e^{-uT'} \sum_{k\geq 1} J_k'^2 \right] \mathrm{d}u \\
&\qquad = 2 \int_{\mathbb{R}_+} \mathbb{E}\left[\int_{\mathbb{R}_+} e^{-u \int_{\mathbb{R}_+} zM'(\mathrm{d}z)} s^2 M'(\mathrm{d}z) \right] \mathrm{d}u,
\end{aligned}$$

where the last equality follows by identifying M' with the random counting measure $\sum_{k\geq 1} \delta_{J_k'}$. Then, Mecke's equation (Last and Penrose, 2018, Theorem 4.1) leads to:

$$2 \int_{\mathbb{R}_+} \mathbb{E}\left[\int_{\mathbb{R}_+} e^{-u \int_{\mathbb{R}_+} zM'(\mathrm{d}z)} s^2 M'(\mathrm{d}z) \right] \mathrm{d}u = \frac{2\theta}{J} \int_{\mathbb{R}_+} \int_{\mathbb{R}_+} \mathbb{E}[e^{-uT'}] e^{-us} s^2 \rho(s) \mathrm{d}s.$$

The proof follows by noticing that, by the Lévy-Kintchine representation, $\mathbb{E}[e^{-uT'}] = e^{-\psi(u)}$ and from the definition of $\kappa(u, n)$.

A.14 Monte Carlo Estimation of (8)

As in Dolera et al. (2023), we note that (8) can be equivalently expressed as:

$$\Pr\left[f_{X_{n+1}} = l \,|\, \boldsymbol{C}_J = \boldsymbol{c}, h(X_{n+1}) = j\right] = \frac{\gamma}{J}\binom{c_j}{l}(1-\alpha)_{(l)}\frac{(\gamma)_{(c_j-l)}}{(\gamma)_{(c_j+1)}}\frac{\mathbb{E}\left[\frac{\left(\frac{\gamma+\alpha}{\gamma}\right)_{K_\cdot^{\mathcal{S}(\boldsymbol{c},j,-l)}}}{J^{K_\cdot^{\mathcal{S}(\boldsymbol{c},j,-l)}}\prod_{k=1}^J\left(\frac{\gamma}{\alpha}\right)_{K_{c_k-l\delta_{k,j}}}}\right]}{\mathbb{E}\left[\frac{\left(\frac{\gamma}{\gamma}\right)_{K_\cdot^{\mathcal{S}(\boldsymbol{c},j,1)}}}{J^{K_\cdot^{\mathcal{S}(\boldsymbol{c},j,1)}}\prod_{k=1}^J\left(\frac{\gamma}{\alpha}\right)_{K_{c_k+\delta_{k,j}}}}\right]}, \quad (60)$$

where K_m is number of distinct values in a sample of size m from a PYP and $K_\cdot^{\mathcal{S}(\boldsymbol{c},j,-l)} := \sum_{1\le k\le J} K_{c_k-l\delta_{k,j}}$. To prove (60), recall that the number of distinct elements K_m in a sample of size m from a PYP has distribution

$$\Pr\left[K_m = k\right] = \frac{\left(\frac{\gamma}{\alpha}\right)_{(k)}}{(\gamma)_{(m)}}\mathscr{C}(m,k;\alpha).$$

Consider now (8) and note that $\Gamma\left(a+b\right) = (a)_{(b)}\Gamma(a)$. Then

$$\begin{aligned}
&\Pr\left[f_{X_{n+1}} = l \,|\, \boldsymbol{C}_J = \boldsymbol{c}, h(X_{n+1}) = j\right] = \\
&= \frac{\gamma}{J}\binom{c_j}{l}(1-\alpha)_{(l)}\frac{(\gamma)_{(c_j-l)}}{(\gamma)_{(c_j+1)}}\frac{\sum_{\boldsymbol{i}\in\mathcal{S}(\boldsymbol{c},j,-l)}\frac{\left(\frac{\gamma+\alpha}{\alpha}\right)_{(|\boldsymbol{i}|)}}{J^{|\boldsymbol{i}|}\prod_{h=1}^J\left(\frac{\gamma}{\alpha}\right)_{(i_h)}}\prod_{k=1}^J\frac{\left(\frac{\gamma}{\alpha}\right)_{(i_k)}}{(\gamma)_{(c_k-l\delta_{k,j})}}\mathscr{C}(c_k-l\delta_{k,j},i_k;\alpha)}{\sum_{\boldsymbol{i}\in\mathcal{S}(\boldsymbol{c},j,1)}\frac{\left(\frac{\gamma}{\alpha}\right)_{(|\boldsymbol{i}|)}}{J^{|\boldsymbol{i}|}\prod_{h=1}^J\left(\frac{\gamma}{\alpha}\right)_{(i_h)}}\prod_{k=1}^J\frac{\left(\frac{\gamma}{\alpha}\right)_{(i_k)}}{(\gamma)_{(c_k+\delta_{k,j})}}\mathscr{C}(c_k-l\delta_{k,j},i_k;\alpha)} \\
&= \frac{\gamma}{J}\binom{c_j}{l}(1-\alpha)_{(l)}\frac{(\gamma)_{(c_j-l)}}{(\gamma)_{(c_j+1)}}\frac{\sum_{\boldsymbol{i}\in\mathcal{S}(\boldsymbol{c},j,-l)}\frac{\left(\frac{\gamma+\alpha}{\alpha}\right)_{(|\boldsymbol{i}|)}}{J^{|\boldsymbol{i}|}\prod_{h=1}^J\left(\frac{\gamma}{\alpha}\right)_{(i_h)}}\prod_{k=1}^J\Pr\left[K_{c_k-l\delta_{k,j}}=i_k\right]}{\sum_{\boldsymbol{i}\in\mathcal{S}(\boldsymbol{c},j,1)}\frac{\left(\frac{\gamma}{\alpha}\right)_{(|\boldsymbol{i}|)}}{J^{|\boldsymbol{i}|}\prod_{h=1}^J\left(\frac{\gamma}{\alpha}\right)_{(i_h)}}\prod_{k=1}^J\Pr\left[K_{c_k+\delta_{k,j}}=i_k\right]} \\
&= \frac{\gamma}{J}\binom{c_j}{l}(1-\alpha)_{(l)}\frac{(\gamma)_{(c_j-l)}}{(\gamma)_{(c_j+1)}}\frac{\mathbb{E}\left[\frac{\left(\frac{\gamma+\alpha}{\gamma}\right)_{K_\cdot^{\mathcal{S}(\boldsymbol{c},j,-l)}}}{J^{K_\cdot^{\mathcal{S}(\boldsymbol{c},j,-l)}}\prod_{k=1}^J\left(\frac{\gamma}{\alpha}\right)_{K_{c_k-l\delta_{k,j}}}}\right]}{\mathbb{E}\left[\frac{\left(\frac{\gamma}{\gamma}\right)_{K_\cdot^{\mathcal{S}(\boldsymbol{c},j,1)}}}{J^{K_\cdot^{\mathcal{S}(\boldsymbol{c},j,1)}}\prod_{k=1}^J\left(\frac{\gamma}{\alpha}\right)_{K_{c_k+\delta_{k,j}}}}\right]}.
\end{aligned}$$

Appendix B. Details about the Poisson-IBP

Proof of Equation (28) It follows from $\psi(u) = \log(1+u)$ and $\kappa(u,n) = (n-1)!/(u+1)^n$.

Proof of (29) We have

$$\psi(u) = \frac{\alpha}{\Gamma(1-\alpha)}\int_{\mathbb{R}_+}(1-e^{-us})\rho(s)\mathrm{d}s = \frac{\alpha}{\Gamma(1-\alpha)}\int_{\mathbb{R}_+}(1-e^{-us})s^{-1-\alpha}e^{-\tau s}\mathrm{d}s$$

$$
\begin{aligned}
&= \frac{\alpha}{\Gamma(1-\alpha)} \int_{\mathbb{R}_+} \int_{\mathbb{R}_+} (1 - e^{-us}) \frac{t^\alpha e^{-ts}}{\Gamma(\alpha+1)} e^{-s} \mathrm{d}t \, \mathrm{d}s \\
&= \frac{\alpha}{\Gamma(1-\alpha)} \int_{\mathbb{R}_+} \frac{t^\alpha u}{(t+1)(t+u+1)} \mathrm{d}t = \frac{\alpha \Gamma(\alpha)\Gamma(1-\alpha)}{\Gamma(1-\alpha)\Gamma(\alpha+1)} \left[(\tau+u)^\alpha - \tau^\alpha\right] \\
&= \left[(\tau+u)^\alpha - \tau^\alpha\right],
\end{aligned}
$$

and

$$
\kappa(l,u) = \alpha \frac{\Gamma(l-\alpha)}{\Gamma(1-\alpha)} (\tau+u)^{\alpha-l} = \alpha (1-\alpha)_{(l-1)} (\tau+u)^{\alpha-l}.
$$

Denoting by $(*)$ the summation over positive integers $(k_1, \ldots, k_i)$ such that $\sum_{j=1}^i k_j = n$,

$$
\begin{aligned}
\frac{\mathrm{d}^n}{\mathrm{d}u^n} e^{-\theta\psi(u)/J} &= e^{\theta/J\tau^\alpha} \frac{\mathrm{d}^n}{\mathrm{d}u^n} e^{-\theta/J(\tau+u)^\alpha} \\
&= e^{-\theta/J\psi(u)} \sum_{i=1}^n \left(\frac{\theta}{J}\right)^i \sum_{(*)} \frac{1}{i!} \binom{n}{k_1, \ldots, k_i} \prod_{j=1}^i \alpha (\tau+u)^{\alpha-k_j} (1-\alpha)_{k_j-1} \\
&= e^{-\theta/J\psi(u)} \sum_{i=1}^n \left(\frac{\theta}{J}\right)^i \frac{\mathscr{C}(n,i;\alpha)}{(\tau+u)^{n-\alpha i}}.
\end{aligned}
$$

Plugging these in the expression of Proposition 8 leads to (29).

Appendix C. Additional Numerical Results

C.1 Cardinality recovery

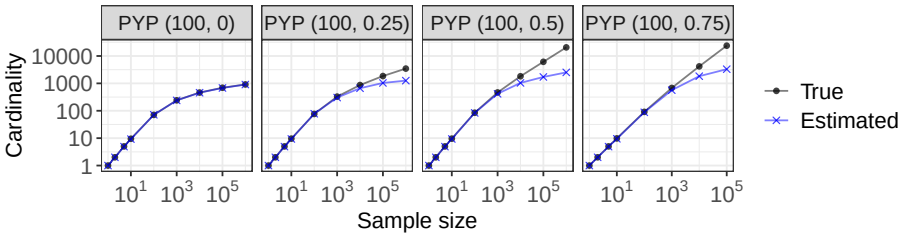


Figure 10: True and estimated cardinality in synthetic data from PYP prior models, as a function of the sample size. The estimates assume a mis-specified DP prior fitted via maximum marginal likelihood. Other details are as in Figure 8.

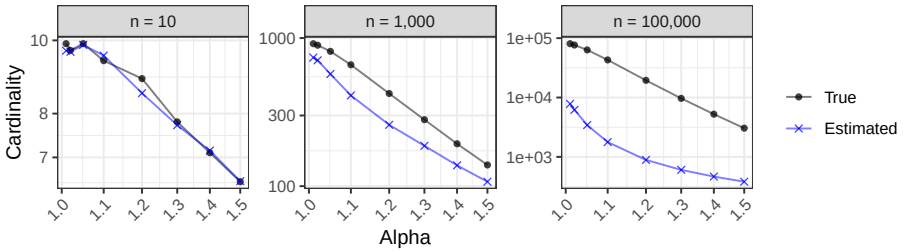


Figure 11: True and estimated cardinality for data sampled from a Zipf distribution, as a function of the tail parameter α. The estimates assume a mis-specified DP prior fitted via maximum marginal likelihood. Other details are as in Figure 8.

References

Aggarwal, C. and Yu, P. (2010). On classification of high-cardinality data streams. In *SIAM International Conference on Data Mining*.

Alon, N., Matias, Y. and Szegedy, M. (1999). The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences* **58**, 137–147.

Bacallado, S., Battiston, M., Favaro, S. and Trippa, L. (2017). "Sufficientness" postulates for Gibbs-type priors and hierarchial generalizations. *Statistical Science* **32**, 487–500.

Beraha, M., Favaro, S., and Sesia, M. (2023). Frequency and cardinality recovery from sketched data: a novel approach bridging Bayesian and frequentist views. *Preprint arXiv:2309.15408*.

Berger, B., Daniels, N.M., and Yu, Y.W. (2016). Computational biology in the 21st century: scaling with compressive algorithms. *Communication of the ACM* **59**, 72–80.

Bernton, E., Jacob, P.E., Gerber, M. and Robert, C.P. (2019). On parameter estimation with the Wasserstein distance. *Information and Inference: a journal of the IMA* **8**, 657–676.

Blum, A., Hopcroft, J. and Kannan, R. (2020). *Foundations of data science*. Cambridge University Press.

Brix, A. (1999). Generalized gamma measures and shot-noise Cox processes. *Advances in Applied Probability* **31**, 929–953.

Broderick, T., Mackey, L., Paisley, J., and Jordan, M.I. (2015). Combinatorial clustering and the beta negative binomial process. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **37**, 290–306.

Broderick, T., Wilson, A.C, and Jordan, M.I. (2018). Posteriors, conjugacy, and exponential families for completely random measures. *Bernoulli* **24**, 3181–3221.

Cai, D., Mitzenmacher, M., and Adams, R. P. (2018). A Bayesian nonparametric view on count-min sketch. In *Advances in Neural Information Processing Systems*.

Camerlenghi, F., A. Lijoi, P. Orbanz, and I. Prünster (2019). Distribution theory for hierarchical processes. *The Annals of Statistics* **47**, 67–92.

Campbell, T., Cai, D. and Broderick. T. (2018). Exchangeable trait allocations. *Electronic Journal of Statistics* **12**, 2290–2322.

Charalambides, C. (2005). *Combinatorial methods in discrete distributions*. Wiley.

Charikar, M., Chen, K. and Farach-Colton, M. (2002). Finding frequent items in data streams. *Theoretical Computer Science* **312**, 3–15.

Chassaing, P. and Gerin, L. (2006). Efficient estimation of the cardinality of large data sets. In *Colloquium on Mathematics and Computer Science Algorithms, Trees, Combinatorics and Probabilities*.

Chen, A., Cao, J., Shepp, L. and Nguyen, T. (2011). Distinct counting with a self-learning bitmap. *Journal of the American Statistical Association* **106**, 879–890.

Chen, S. X. and Liu, J. S. (1997). Statistical applications of the Poisson-Binomial and conditional Bernoulli distributions. *Statistica Sinica* **7**, 875–892.

Cormode, G. (2017). Data sketching. *Communication of the ACM* **60**, 48–55.

Cormode, G., Jha, S., Kulkarni, T., Li, N., Srivastava D. and Wang, T. (2018). Privacy at scale: local differential privacy in practice. In *International Conference on Management*.

Cormode, G. and Muthukrishnan, S. (2005). An improved data stream summary: the count-min sketch and its applications. *Journal of Algorithms* **55**, 58–75.

Cormode, G. and Yi, K. (2020). *Small summaries for big data*. Cambridge University Press.

Dolera, E. and Favaro, S. (2020). A Berry-Esseen theorem for Pitman's α-diversity. *The Annals of Applied Probability* **30**, 847–869.

Dolera, E., Favaro, S. and Peluchetti, S. (2021). A Bayesian nonparametric approach to count-min sketch under power-law data stream. In *International Conference on Artificial Intelligence and Statistics*.

Dolera, E., Favaro, S. and Peluchetti, S. (2023). Learning-augmented count-min sketches via Bayesian nonparametrics. *Journal of Machine Learning Research* **24**, 1–60.

Dwork, C. and Naor, M. and Pitassi, T. and Rothblum, G. and Yekhanin, S. (2010). Pan-private streaming algorithms. In *Symposium on Innovations in Computer Science*.

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

Flajolet, P., Fusy, E., Gandouet, O. and Meunier, F. (2007). Hyperloglog: the analysis of a near-optimal cardinality estimation algorithm. In *Analysis of Algorithms*.

Flajolet, P. and Martin, G.N. (1983). Probabilistic counting. In *IEEE Conference on Foundations of Computer Science*.

Flajolet, P. and Martin, G.N. (1985). Probabilistic counting algorithms for database applications. *Journal of Computer and System Sciences* **31**, 182–209.

Goyal, A., Daumé, H. and Cormode, G. (2012). Sketch algorithms for estimating point queries in NLP. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.

Hatcher, E.L., Zhdanov, S.A., Bao, Y., Blinkova, O., Nawrocki, E.P., Ostapchuck, Y., Schäffer, A.A., and Brister, J.R. (2017). Virus variation resource-improved response to emergent viral outbreaks. *Nucleic Acids Research* **45**, D482–D490.

Indyk, P. (2006). Stable distributions, pseudorandom generators, embeddings and data stream computation. *Journal of the ACM* **53**, 307–325.

James, L.F. (2002). Poisson process partition calculus with applications to exchangeable models and Bayesian nonparametrics. *Preprint arXiv:math/0205093*.

James, L.F. (2017). Bayesian Poisson calculus for latent feature modeling via generalized Indian buffet process priors *The Annals of Statistics* **45**, 2016–2045.

Karp, R., Shenker, S. and Papadimitriou, C.H. (2003). A simple algorithm for finding frequent elements in streams and bags. *ACM Transactions on Database Systems* **28**, 51–53.

Kingman, J.F.C (1967). Completely random measures. *Pacific Journal of Mathematics* **21**, 59–78.

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

Kingman, J.F.C (1993). *Poisson processes*. Oxford University Press.

Kockan, C., Zhu, K., Dokmai, N., Karpov, N., Kulekci, M.O., Woodruff, D.P, and Sahinalp, S.C. (2020). Sketching algorithms for genomic data analysis and querying in a secure enclave. *Nature methods* **17**, 295–301.

Last, G. and Penrose, M. (2018). Lectures on the Poisson process. *Cambridge University Press*.

Le Cam, L. (1960). An approximation theorem for the Poisson-Binomial distribution. *Pacific Journal of Mathematics* **10**, 1181–1197.

Leo Elworth, L.A., Wang, Q., Kota, P.K., Barberan, C.J., Coleman, B., Balaji, A., Gupta, G., Baraniuk, R.G., Shrivastava, A. and Treangen, T.J. (2020). To petabytes and beyond: recent advances in probabilistic and signal processing algorithms and their application to metagenomics. *Nucleic Acids Research* **48** 5217–5234.

Lijoi, A. and Prünster, I. (2010). Models beyond the Dirichlet process. In *Bayesian Nonparametrics*, Hjort, N.L., Holmes, C.C. Müller, P. and Walker, S.G. Eds. Cambridge University Press.

Lijoi, A., Mena, R. H., and Prünster, I. (2005). Hierarchical mixture modeling with normalized inverse-Gaussian priors. *Journal of the American Statistical Association* **100**, 1278–1291.

Lijoi, A., Mena, R.H. and Prünster, I. (2007). Controlling the reinforcement in Bayesian nonparametric mixture models. *Journal of the Royal Statistical Society Series B* **69**, 715-740.

Manku, G.S. and Motwani, R. (2002). Approximate frequency counts over data streams. In *International Conference on Very Large Data Bases*.

Marcais, G., Solomon, B., Patro, R. and Kingsford, C (2019). Sketching and sublinear data structures in genomics. *Annual Review of Biomedical Data Science* **89**, 669–682

Medjedovic, D., Tahirovic, E. and Dedovic, I. (2022). *Algorithms and data structures for massive datasets*. Manning.

Melis, L., Danezis, G., and Cristofaro, E.D. (1982). Efficient private statistics with succinct sketches. In *Network and Distributed System Security Symposium Symposium*.

Misra, J. and Gries D. (1982). Finding repeated elements. *Science of computer programming* **2**, 143–152.

Mitzenmacher, M. and Upfal, E. (2017). *Probability and computing: randomization and probabilistic techniques in algorithms and data analysis*. Cambridge University Press.

Pettie, S. and Wang, D. (2021). Information theoretic limits of cardinality estimation: Fisher meets Shannon. In *Symposium on Theory of Computing*.

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields* **102**, 145–158.

Pitman, J. (2003). Poisson-Kingman partitions. In *Science and Statistics: A Festschrift for Terry Speed*, Goldstein, D.R. Eds. Institute of Mathematical Statistics.

Pitman, J. (2006). *Combinatorial Stochastic Processes.* Ecole d'Eté de Probabilités de Saint-Flour XXXII. Lecture notes in mathematics, Springer.

Pitman, J. and Yor, M. (1997). The two parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability* **25**, 855–900.

Project Gutenberg (accessed: 2022-02-05). www.gutenberg.org.

Prünster, I. (2002). *Random probability measures derived from increasing additive processes and their application to Bayesian statistics*. PhD Thesis, University of Pavia.

Regazzini, E. (2001). *Foundations of Bayesian statistics and some theory of Bayesian nonparametric methods.* Lecture Notes, Stanford University.

Regazzini, E., Lijoi, A. and Prünster, I. (2003). Distributional results for means of normalized random measures with independent increments. *Annals of Statistics* **31**, 560–585.

Rennie, J. D. and Shih, L. and Teevan, J. and Karger, D. R. (2003). Tackling the poor assumptions of naive Bayes text classifiers. In *International Conference on Machine Learning*.

Rojas, J.S., Gallón, A.R. and Corrales, J.C. (2018). Personalized service degradation policies on OTT applications based on the consumption behavior of users. In *International Conference on Computational Science and Its Applications*, 543–557.

Schechter, S., Herley, C. and Mitzenmacher, M. (2010). Popularity is everything: a new approach to protecting passwords from Statistical-Guessing attacks. *USENIX Workshop on Hot Topics in Security* **10**.

Sesia, M. and Favaro, S. (2022). Conformalized frequency estimation from sketched data. In *Advances in Neural Information Processing Systems*.

Sesia, M., Favaro, S. and Dobriban, E. (2023). *Conformal Frequency Estimation using Discrete Sketched Data with Coverage for Distinct Queries*. In *Journal of Machine Learning Research* **24**, 1–80.

Shi, Q., Petterson, J., Dror, G., Langford, J., Smola, A. and Vishwanathan, S. (2009). Hash kernels for structured data. *Journal of Machine Learning Research* **10**, 2615–2637.

Solomon, B. and Kingsford, C. (2016). Fast search of thousands of short-read sequencing experiments. *Nature Biotechnology* **34**, 300–302.

Song, H.H., Cho, T.W., Dave, V., Zhang, Y. and Qiu, L. (2009). Scalable proximity estimation and link prediction in online social networks. In *ACM SIGCOMM Conference on Internet Measurement*.

Steele J. M. (1994). Le Cam's inequality and Poisson approximations. *The American Mathematical Monthly* **101**, 48–54.

Teh, Y.W. and Jordan, M. I. (2010). *Hierarchical Bayesian nonparametric models with applications.* In *Bayesian nonparametrics*.

Ting, D. (2014). Streamed approximate counting of distinct elements: beating optimal batch methods. In *International Conference of Knowledge Discovery and Data Mining*.

Ting, D. (2016). Towards Optimal Cardinality Estimation of Unions and Intersections with Sketches. In *International Conference of Knowledge Discovery and Data Mining*.

Ting, D. (2018). Count-min: Optimal estimation and tight error bounds using empirical error distributions. In *International Conference of Knowledge Discovery and Data Mining*.

Vovk, V., Gammerman, A., and Shafer, G. (2005) Algorithmic learning in a random world. Springer.

Zhang, Q., Pell, J., Canino-Koning, R., Howe, A.C. and Brown, C.T. (2014). These are not the k-mers you are looking for: efficient online k-mer counting using a probabilistic data structure. *PloS one* **9**.

Zhou, M. and Padilla, O. H. M and Scott, J. G. (2016). Priors for random count matrices derived from a family of negative binomial processes. *Journal of the American Statistical Association* **111**, 1144–1156.