Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uasa20

A smoothed-Bayesian approach to frequency recovery from sketched data*

Mario Beraha, Stefano Favaro & Matteo Sesia

To cite this article: Mario Beraha, Stefano Favaro & Matteo Sesia (21 Apr 2025): A smoothed-Bayesian approach to frequency recovery from sketched data*, Journal of the American Statistical Association, DOI: 10.1080/01621459.2025.2490302

To link to this article: https://doi.org/10.1080/01621459.2025.2490302

A smoothed-Bayesian approach to frequency recovery from sketched data*

Mario Beraha[a], Stefano Favaro[b], and Matteo Sesia[c,d,#]

[a]Department of Economics, Management and Statistics, University of Milano-Bicocca, Milan, Italy

[b]Department of Economics and Statistics, University of Torino and Collegio Carlo Alberto, Torino, Italy

[c]Department of Data Sciences and Operations, University of Southern California, Los Angeles, CA, USA

[d]Department of Computer Science, University of Southern California, Los Angeles, CA, USA

*This article is dedicated to the memory of Luca Trevisan.

[#]sesia@marshall.usc.edu

Abstract

We provide a novel statistical perspective on a classical problem at the intersection of computer science and information theory: recovering the empirical frequency of a symbol in a large discrete dataset using only a compressed representation, or sketch, obtained via random hashing. Departing from traditional algorithmic approaches, recent works have proposed Bayesian nonparametric (BNP) methods that can provide more informative frequency estimates by leveraging modeling assumptions about the distribution of the sketched data. In this paper, we propose an alternative *smoothed-Bayesian* approach, inspired by existing BNP methods but designed to overcome their computational limitations when dealing with large-scale data from realistic distributions, including those with power-law tail behaviors. For sketches obtained with a single hash function, our approach is supported by precise theoretical guarantees, including unbiasedness and optimality under a Bayesian framework within an intuitive class of linear estimators. For sketches with multiple hash functions, we introduce an approach based on *multi-view* learning to construct computationally efficient frequency estimators. We validate our method on synthetic and real data, comparing its performance to that of existing alternatives.

Keywords: *nonparametric estimation; normalized random measures; random hashing; smoothed estimation; worst-case analysis.*

1 Introduction

1.1 Background and motivation

An interesting statistical problem born at the intersection of computer science and information theory is to recover the empirical frequency of an object in a large discrete dataset using a lossy compressed representation, or "sketch". Sketches are central to many data science applications involving memory or privacy constraints, including real-time analysis, fast query processing, and scalable machine learning (Cormode and Yi, 2020). The count-min sketch (CMS) by Cormode and Muthukrishnan (2005), reviewed in Appendix A1.1, is a popular algorithm that uses random hash functions to create a data sketch, providing a deterministic upper bound for any object's frequency. Additionally, it allows computing confidence intervals for an object's frequency by leveraging concentration inequalities that utilize the randomness in the hash functions while treating the data as fixed.

A limitation of the CMS is that, as an algorithmic approach that treats the data as fixed, it may not lead to the most informative estimates when the data are random samples from a population (Ting, 2018). This issue has motivated the development of *learning-augmented* versions of the CMS, which apply optimization and machine learning techniques to improve sketching algorithms, as well as frequentist and Bayesian statistical approaches to extract more informative estimates from sketches, leveraging modeling assumptions about the data distribution (Cai et al., 2018; Ting, 2018; Hsu et al., 2019; Aamand et al., 2019; Bertsimas and Digalakis, 2021; Sesia and Favaro, 2022; Aamand et al., 2024; Cao et al., 2024).

Within the Bayesian framework, Cai et al. (2018) pioneered a Bayesian nonparametric (BNP) approach to frequency recovery, assuming a Dirichlet process prior (Ferguson, 1973) for the data distribution and computing the posterior distribution of a new object's empirical frequency conditional on the sketch. Subsequently, Dolera et al. (2021, 2023) and Beraha et al. (2024) extended this approach to more general prior distributions. While the BNP approach is effective under the Dirichlet process prior, it has two limitations. Firstly, it becomes very computationally expensive with more general prior distributions needed to describe important patterns often found in real data, such as power-law tails (Ferrer i Cancho and Solé, 2001; Zipf, 2016). Secondly, it is challenging to apply to sketches obtained from multiple hash functions (Cai et al., 2018; Dolera et al., 2023). See Appendix A1.2 for a more detailed review of the BNP approach and its computational challenges.

The current limitations of the BNP approach motivate us to develop a novel statistical method for frequency recovery from sketched data, by a single or multiple hash functions. Inspired by the BNP approach but integrating frequentist ideas, our method provides greater flexibility in modeling realistic data while enabling efficient large-scale applications. Additionally, it enjoys desirable theoretical guarantees of unbiasedness and optimality.

1.2 Problem statement

To describe the frequency recovery problem, consider a data set $(x_1,\ldots,x_n)$, for $n \geq 1$, with the x_i's taking values in a (possibly infinite) alphabet $\mathbb{S}$ of symbols. Only a sketch of these data can be observed, obtained through *random hashing* (Mitzenmacher and Upfal, 2017, Chapter 15). For simplicity, we begin by focusing on a single hash function.

A hash function $h:\mathbb{S}\to[J]:=\{1,\ldots,J\}$ with width $J \geq 1$ maps each symbol into one of *J*buckets. This function is assumed to be random and distributed as a pairwise independent hash family $\mathcal{H}_J$. That is, $h:\mathbb{S}\to[J]$ such that $\Pr[h(x_1)=j_1, h(x_2)=j_2]=J^{-2}$ for any $j_1, j_2 \in [J]$ and fixed $x_1, x_2 \in \mathbb{S}$ such that $x_1 \neq x_2$. The pairwise independence of $\mathcal{H}_J$, known as strong universality, implies uniformity, meaning that $\Pr[h(x)=j]=J^{-1}$ for all $j \in [J]$. Strong universality is a common and mathematically convenient assumption, although different settings may be also considered (Chung et al., 2013).

Hashing the data through *h* produces a random vector $\mathbf{C}_J \in \mathbb{N}_0^J$, referred to as the sketch, whose *j*-th element C_j is the number of x_i's mapped into the *j*-th bucket:

$$C_j = \sum_{i=1}^{n} I[h(x_i)=j], \qquad \text{for all } j \in [J],$$

where $I[\cdot]$ is the indicator function. Therefore, $\sum_{1\leq j\leq J} C_j = n$. By setting *J* to be (much) smaller than the anticipated number of distinct symbols in $(x_1,\ldots,x_n)$, the sketch $\mathbf{C}_J$ is intended to have a (much) smaller memory footprint compared to the original sample.

The frequency recovery problem consists of estimating the number of occurrences of a symbol x_{n+1} (the *query*) in the sample, i.e.,

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

using only the information in $\mathbf{C}_J$, without looking at $(x_1,\ldots,x_n)$. This task is challenging because distinct symbols may be mapped into the same bucket, an event known as a *hash collision*. When the sample size *n* and the cardinality of $\mathbb{S}$ are both larger than *J*, as is typical in practice, hash collisions become numerous, making exact recovery of $f_{x_{n+1}}$ impossible. Fortunately, it is possible to obtain useful estimates of $f_{x_{n+1}}$ from $\mathbf{C}_J$, especially if we assume the data are a random sample from some discrete distribution *P* on $\mathbb{S}$.

1.3 Related works

BNP approaches to the frequency recovery problem were pioneered by Cai et al. (2018) and extended by Dolera et al. (2021, 2023) and Beraha et al. (2024). These works noted that BNP approaches become computationally expensive beyond the Dirichlet process prior (Dolera et al., 2023; Beraha et al., 2024), which is often too rigid to describe real data. Moreover, BNP

approaches are challenging to extend to sketches obtained from multiple hash functions (Cai et al., 2018; Dolera et al., 2023), further limiting their applicability.

This paper draws inspiration from the aforementioned works while aiming to overcome their limitations by developing a practical and scalable *smoothed-Bayesian* method. Our approach addresses the computational challenges of existing BNP methods and provides greater flexibility to model realistic data. Additionally, this paper complements the recent works of Sesia and Favaro (2022) and Sesia et al. (2023), who proposed a *distribution-free* frequentist approach based on conformal inference (Vovk et al., 2005). Their focus is on providing confidence intervals with finite-sample coverage guarantees for frequency estimates computed by any *black-box* model or algorithm. Combining their approach with ours can endow our smoothed-Bayesian method with frequentist uncertainty estimates.

1.4 Main contributions

Our first contribution is to introduce a class of smoothed-Bayesian estimators of the empirical frequency $f_{X_{n+1}}$. The terminology "smoothed" stems from the seminal work of Good (1953), which studied the problem of estimating population frequencies by first producing an *oracle* estimator based on knowledge of the data-generating distribution, namely *P*, and then specifying a parametric family for that unknown (discrete) distribution.

In a similar spirit, we first consider the frequency recovery problem conditional on the distribution *P*, deriving an oracle estimator for $f_{X_{n+1}}$ denoted by $\varepsilon_f(P)$. To relax the assumption that *P* is known, we smooth this estimator by computing the expected value of $\varepsilon_f(P)$ with respect to a distribution $P \sim \wp$, which can be interpreted as a prior distribution. We consider a broad class of normalized random measures (James, 2002; Prünster, 2002; Regazzini et al., 2003), for which the resulting smoothed estimator can be expressed as the expected value of a mixture of Binomial distributions. Concretely, we focus on $\wp$ corresponding to the law of the Dirichlet process and the normalized generalized Gamma process (James, 2002; Pitman, 2003; Lijoi et al., 2007)—a flexible solution that allows the tail behaviour of *P* to range from exponential tails to heavy power-law tails.

Our smoothed-Bayesian approach yields computationally efficient estimators that depend linearly on $C_{h(X_{n+1})}$. This is a key advantage over the BNP estimators reviewed in Section 2.3, which are typically intractable. Our approach also enjoys desirable theoretical properties, including *unbiasedness* and *optimality* among linear estimators. Specifically, it minimizes mean squared error relative to the Bayes estimator, which, while optimal under a Bayesian framework, is generally intractable. Furthermore, the smoothed-Bayesian and BNP estimators coincide when $\wp$ is the law of the Dirichlet process, ensuring consistency with prior work in that computationally feasible special case. Crucially, however, our estimator does not require posterior evaluation, making it scalable beyond this setting.

Our second contribution extends the smoothed-Bayesian approach to deal with data sketched by multiple hash functions, in analogy with the CMS. This is a challenging problem, for which we propose a solution based on *multi-view learning* (Xu et al., 2013; Shankar et al., 2018; Li et al., 2018), leveraging by the fact that the smoothed-Bayesian framework leads not only to a principled *point estimate* but also to a *probability distribution* for $f_{X_{n+1}}$. As we shall

see, our multi-view learning method can recover the BNP approach of Cai et al. (2018) in the special case of P being a Dirichlet process, but it is more generally applicable.

1.5 Organization of the paper

Section 2.1 introduces our framework and derives a closed-form expression for $\varepsilon_f(P)$ as a function of P. Section 2.2 presents a minimax analysis for the estimation of $f_{X_{n+1}}$, underscoring the need for prior (or smoothing) assumptions. Section 2.3 reviews the difficulties associated with the BNP approach. Section 3 presents our smoothed-Bayesian approach for a single hash function, and Section 4 extends it to multiple hash functions. Section 5 validates our method empirically, comparing it to existing algorithmic and BNP solutions. Section 6 concludes with a discussion. The Appendices in the Supplementary Material contain proofs, additional methodological details, further comparisons, and numerical results.

2 Preliminary results

2.1 Statistical framework and problem setup

We rely on the following assumptions. Firstly, $(x_1,\ldots,x_n)$ are modeled as a random sample $\mathbf{X}_n=(X_1,\ldots,X_n)$ from an unknown discrete distribution $P=\sum_{s\in\mathbb{S}} p_s\delta_s$ on $\mathbb{S}$, where p_s is the (unknown) probability of $s\in\mathbb{S}$. Secondly, the strong universal hash family $\mathcal{H}_J$, from which h is drawn, is independent of $\mathbf{X}_n$. Formally, for any $n\geq 1$, we can thus write:

$$\begin{aligned} X_1,\ldots,X_n &\overset{\text{iid}}{\sim} P, \\ h &\overset{\text{ind}}{\sim} \mathrm{H}_J, \\ C_j &= \sum_{i=1}^{n} I(h(X_i)=j), \quad \forall j\in[J]. \end{aligned} \tag{1}$$

The next theorem provides the conditional distribution of $f_{X_{n+1}}$, given the sketch $\mathbf{C}_J$, the bucket $h(X_{n+1})$ in which X_{n+1} is hashed, and h. This is a key component of our approach.

Theorem 1 .

For $n\geq 1$, suppose $(x_1,\ldots,x_n)$ is a random sample $\mathbf{X}_n$ from (1) with corresponding sketch $\mathbf{C}_J=\mathbf{c}$, obtained using a fixed hash function h. If $\mathbb{S}_j:=\{s\in\mathbb{S}: h(s)=j\}$ and $q_j:=\Pr[h(X_i)=j\mid h]$ for any $j\in[J]$ and $i\in[n]:=\{1,\ldots,n\}$, then the X_i's hashed into the j-th bucket, i.e., $\{X_i: h(X_i)=j\}$, are i.i.d. as $P_j=\sum_{s\in\mathbb{S}_j}\frac{p_s}{q_j}\delta_s$. Further, for $j\in[J]$ and $r\in\{0,1,\ldots,c_j\}$,

$$\pi_j(r;P,h) \coloneqq \Pr[\, f_{X_{n+1}} = r \mid \mathbf{C}_J = \mathbf{c}, h(X_{n+1}) = j, h] = \binom{c_j}{r} \sum_{s \in \mathbb{S}_j} \left(\frac{p_s}{q_j}\right)^{r+1} \left(1 - \frac{p_s}{q_j}\right)^{c_j - r} . \quad (2)$$

See Appendix A2.1.1 for the proof of Theorem 1. Note that $\mathbb{S}_j$, q_j, and P_j all implicitly depend on the hash function *h*, which is treated as fixed here. This result gives us the optimal "oracle" estimator $\varepsilon_f(P,h)$ of $f_{X_{n+1}}$, under squared error loss, for given *P* and *h*:

$$\varepsilon_f(P,h) \coloneqq \mathbb{E}\Big[\, f_{X_{n+1}} \mid \mathbf{C}_J = \mathbf{c}, h(X_{n+1}) = j, h\Big] = \sum_{r=0}^{c_j} r\pi_j(r;P,h) = c_j \sum_{s \in \mathbb{S}_j} \left(\frac{p_s}{q_j}\right)^2 . \quad (3)$$

This depends on the unknown distribution P_j, induced by *P* through hashing via *h*. From (3), it follows that the size C_j of the bucket in which X_{n+1} is hashed is a sufficient statistic to estimate $f_{X_{n+1}}$. Note that $\varepsilon_f(P,h)$ may be interpreted as measuring the diversity in the composition of the *j*-th bucket, because $\sum_{s \in \mathbb{S}_j} (p_s / q_j)^2$ represents the probability that two randomly chosen elements from the *j*-th bucket correspond to the same symbol.

2.2 The necessity of prior assumptions

To demonstrate the necessity of introducing modeling assumptions about the data distribution *P* for obtaining an informative estimate of $f_{X_{n+1}}$, we begin by studying the frequency recovery problem from a frequentist minimax perspective. This study is inspired by Painsky (2022, 2023, 2024), which conducted similar worst-case minimax analyses in the different context of missing mass estimation.

Consider a class of *linear* estimators $\hat{f}_\beta = c_j \beta_j$, for $\beta_j \geq 0$. This includes the oracle estimator $\varepsilon_f(P,h)$ in (3) for $\beta_j = \sum_{s \in \mathbb{S}_j} (p_s / q_j)^2$. The corresponding quadratic risk is:

$$R(\hat{f}_\beta; P, h) = \mathbb{E}_P[(\beta_{h(X_{n+1})} C_{h(X_{n+1})} - f_{X_{n+1}})^2 \mid h], \quad (4)$$

which depends on the unknown data distribution *P*.

In this analysis, we aim to minimize an upper bound $\tilde{R}(\hat{f}_\beta, h) \geq R(\hat{f}_\beta; P, h)$ for all *P* within a suitable family $\mathcal{P}$. This leads to a *worst-case optimal* estimator β that solves $\min_\beta \tilde{R}(\hat{f}_\beta, h)$. We focus on a broad class of distributions with a bounded number of support points, defined as $\mathcal{P}_L \coloneqq \{P : P \text{ has at most } L \text{ support points}\}$. While the technical details of this analysis are involved and deferred to Appendix A2.2, the result is easy to understand.

Theorem 2 (Informal statement).

For $n \geq 1$, suppose $(x_1, \ldots, x_n)$ is a random sample $\mathbf{X}_n$ from (1), with corresponding sketch $\mathbf{C}_J = \mathbf{c}$. For a fixed hash-function h, if L is large enough, the worst-case optimal estimator of $f_{X_{n+1}}$, over the class $\mathcal{P}_L$ for $\tilde{R}$, is $\hat{f}_\beta \equiv C_{h(X_{n+1})}$. Moreover, the upper bound $\tilde{R}(\hat{f}_\beta, h)$ is tight and it is achieved by $P \equiv P^*$, where P^* is a degenerate distribution that places all probability mass on a single symbol $s^* \in \mathbb{S}$.

A rigorous statement of Theorem 2 is presented in Appendix A2.2, along with its proof. At first sight, this result may seem disappointing because the worst-case optimal estimator is identical to the original CMS upper bound (see Appendix A1.1), and that is often inaccurate because it does not explicitly account for possible hash collisions. However, Theorem 2 is interesting because the tightness of the analysis highlights the inherent complexity of our frequency recovery problem. In fact, this result tells us that it is impossible to obtain a worst-case estimator that is more informative than the classical CMS upper bound.

Similar conclusions can also be obtained from an alternative (more classical) minimax analysis that consists of solving $\inf_\beta \sup_{P \in \mathcal{P}} R(\hat{f}_\beta; P, h)$, for a given *h*, where $\mathcal{P}$ is an appropriate family of discrete distributions that will be specified below. In this case, analytical computations are not feasible but Appendix A2.3 presents a numerical investigation that leads to an equally unsatisfactory estimator. This minimax estimator is generally uninformative because it always tends to 0 as the latent support size *L* of the data distribution grows, irrespective of the information contained in the data sketch.

In conclusion, the minimax analyses demonstrate that to obtain informative estimates of $f_{X_{n+1}}$, some assumptions about the data distribution *P* are needed. This motivates the BNP approaches developed by prior works, which, however, have their own limitations.

2.3 The limitations of BNP approaches

In the BNP framework of Beraha et al. (2024), the model in (1) is complemented with a prior for *P*, i.e., $P \sim \wp$, an then inference is carried out through the *full-sketch* posterior distribution of $f_{X_{n+1}}$ given $\mathbf{C}_J$ and $h(X_{n+1})$, i.e.,

$$\pi_j^F(r) = \frac{\mathbb{E}_{P \sim \wp, h \sim \mathcal{H}_J}\left[\Pr[f_{X_{n+1}} = r, \mathbf{C}_j = \mathbf{c}, h(X_{n+1}) = j \mid P, h]\right]}{\mathbb{E}_{P \sim \wp, h \sim \mathcal{H}_J}\left[\Pr[\mathbf{C}_j = \mathbf{c}, h(X_{n+1}) = j \mid P, h]\right]}. \quad (5)$$

Theorem 2.2 in Beraha et al. (2024) provides an explicit expression for (5) for a broad class of priors $\wp$. However, beyond the Dirichlet process prior, the computational cost of evaluating $\pi_j^F(r)$ scales exponentially with *J* and *n*, highlighting a fundamental challenge in BNP estimation. For further discussion on why these BNP estimators are unfeasible in their exact form and difficult to approximate using Monte Carlo methods, see Appendix A1.2.

The earlier BNP approaches of Cai et al. (2018) and Dolera et al. (2021, 2023) are somewhat simpler, as they rely on a *single-bucket* posterior distribution, $\pi_j^S(r)$, which conditions only on $C_{h(X_{n+1})}$ rather than the full sketch $\mathbf{C}_J$. This aligns more closely with the original CMS,

which estimates $f_{X_{n+1}}$ using only the information in $C_{h(X_{n+1})}$. As shown by Beraha et al. (2024), π^F and π^S coincide if and only if P is the Dirichlet process. While π^S is computationally simpler than π^F, it remains impractical for large-scale applications. For instance, under a Pitman-Yor process prior, the cost of computing π^S scales quadratically with *n* (Dolera et al., 2023). Appendix A1.2.2 provides a more detailed discussion of computational costs. Notably, computing π^S is several orders of magnitude slower than our smoothed estimators, while computing π^F is infeasible in any realistic setting. This fundamental lack of scalability is the primary limitation of BNP approaches.

The second limitation of BNP approaches is that they struggle to deal with multiple hash functions. Consider that each datum is passed through *M* independent hash functions $h_1,\dots,h_M$, resulting in a sketch $\mathbf{C}_{M,J} \in \mathbb{N}_0^{M\times J}$, a matrix with entries $C_{m,j} = \sum_{i=1}^{n} \mathbf{1}[h_m(x_i) = j]$. In this case, Cai et al. (2018) assume that the posterior distribution of $f_{X_{n+1}}$ given $(C_{1,h_1(X_{n+1})},\dots,C_{M,h_M(X_{n+1})})$ is proportional to the product of the single-bucket posteriors computed using each row of the sketch separately, motivated by the independence of the hash functions. However, this assumption relies not just on the independence of the hash functions but on the independence of the different rows of $\mathbf{C}_{M,J}$. Unfortunately, this independence property does not hold in general, as explained in Section 5.3.

These limitations of BNP approaches, combined with the necessity of using prior assumptions on the data distribution *P* highlighted in Section 2.2, motivate our proposal of the novel *smoothed-Bayesian* approach presented in the next section.

3 Smoothed frequency estimation

3.1 Outline of our approach

The smoothed-Bayesian estimators presented in this section are designed to incorporate prior information naturally, similar to Bayesian approaches, but without the same computational issues. The ideas underlying our approach date back to Good (1953), which first introduced the idea of *smoothing an oracle estimator*, such as that in (3). In the context of frequency recovery from sketched data, smoothed estimation can lead to informative approximations of $f_{X_{n+1}}$ by leveraging assumptions about the distribution *P* on $\mathbb{S}$. The main difficulty, however, is that the oracle estimator $\varepsilon_f(P,h)$ in (3) depends on an intrinsically unobservable quantity, namely the distribution P_j on $\mathbb{S}_j$ induced by the hashing procedure.

Therefore, our approach requires specifying *P* carefully to ensure that the bucket-restricted distributions P_j induced from *P* are well-defined and mathematically tractable. This problem is challenging to address with parametric assumptions on *P* directly, but it can be solved using suitable nonparametric assumptions, as discussed next.

Our solution, inspired by existing BNP approaches, is to treat *P* as a *random* element from the space of discrete probability distributions on $\mathbb{S}$, modeling its distribution P through a low-dimensional smoothing parameter. Specifically, we consider the broad class of

normalized random measures (NRMs) (James, 2002; Prünster, 2002; Pitman, 2003; Regazzini et al., 2003). This choice is inspired by prior work on BNP estimators for $f_{X_{n+1}}$ (Cai et al., 2018; Dolera et al., 2021, 2023), and it leads to an explicit expression for the expected value of $\varepsilon_f(P)$ with respect to the smoothing distribution $\mathcal{P}$, which gives us a practical estimator of $f_{X_{n+1}}$. Further, NRMs allow for the empirical estimation of the smoothing parameters from the sketch $\mathbf{C}_J$, which, similar to the prior's parameters in BNP approaches, can significantly impact posterior inferences (Giordano et al., 2023); see Appendix A2.6 for details on the estimation of the smoothing parameters. Until then, we will assume that the smoothing parameters are known.

3.2 Background on normalized random measures

Consider a purely atomic completely random measure (CRM) $\tilde{\mu}$ on $\mathbb{S}$, $\tilde{\mu}(\cdot) = \sum_{k\geq 1} J_k \delta_{S_k}(\cdot)$ with Lévy intensity measure $\nu(\mathrm{d}x, \mathrm{d}s)$ on $\mathbb{R}_+ \times \mathbb{S}$ (Kingman, 1967, 1993). We consider Lévy intensities of the form $\nu(\mathrm{d}x,\mathrm{d}s) = \theta G_0(\mathrm{d}s)\rho(x)\mathrm{d}x$ where $\theta > 0$ is a parameter, G_0 is a non-atomic probability measure on $\mathbb{S}$, and $\rho(x)\mathrm{d}x$ is a measure on $\mathbb{R}_+$ such that $\int_{\mathbb{R}_+} \rho(x)\mathrm{d}x = +\infty$ and $\psi(u) = \int_{\mathbb{R}_+} (1-e^{-ux})\rho(x)\mathrm{d}x < +\infty$ for all $u > 0$, ensuring the total mass $T = \tilde{\mu}(\mathbb{S})$ of $\tilde{\mu}$ is positive and almost-surely finite (Pitman, 2003; Regazzini et al., 2003). Then, we say that a distribution P on $\mathbb{S}$ is an NRM with parameter (θ, G_0, ρ), in short $P \sim \mathrm{NRM}(\theta, G_0, \rho)$, if $P(\cdot) = \tilde{\mu}(\cdot)/T = \sum_{k\geq 1} (J_k/T)\delta_{S_k}(\cdot)$, where the distribution of the (random) probabilities $(J_k/T)_{k\geq 1}$ is directed by ρ, and the locations $(S_k)_{k\geq 1}$ are i.i.d. from G_0, independent of $(J_k/T)_{k\geq 1}$.

3.3 Smoothed estimation with normalized random measures

We leverage smoothing assumptions for the distribution P as follows. We consider $P \sim \mathrm{NRM}(\theta, G_0, \rho)$, and estimate $f_{X_{n+1}}$ by taking the expected value of $\varepsilon_f(P,h)$ with respect to the law of P and the distribution of h. It follows from (3) that, to compute the expectation of $\varepsilon_f(P,h)$, it suffices to evaluate

$$\pi_j(r) := \mathbb{E}_{P\sim \mathrm{NRM}(\theta,G_0,\rho),h\sim \mathcal{H}_J}[\pi_j(r;P,h)], \quad (6)$$

for all $j \in [J]$. Each $\pi_j(r)$ can be computed by exploiting the Poisson process representation of NRMs and the Poisson coloring theorem (Kingman, 1993, Chapter 5). This is achieved by the following *restriction property* of NRMs.

Suppose P is a NRM and, for any Borel set $A \in \mathcal{S}$, let P_A denote the random probability measure on A induced by $P \sim \mathrm{NRM}(\theta, G_0, \rho)$; i.e., the renormalized restriction of P to the set A. Then, $P_A \sim \mathrm{NRM}(\theta G_0(A), G_{0,A}/G_0(A), \rho)$, where $G_{0,A}$ is the restriction of the probability measure G_0 to A. This property of NRMs is critical to compute $\pi_j(r)$.

Theorem 3 .

For $n \geq 1$, *let* $(x_1, \ldots, x_n)$ *be a random sample* $\mathbf{X}_n$ *from (1), and let* $\mathbf{C}_J = \mathbf{c}$ *be the corresponding sketch. If* $P \sim \text{NRM}(\theta, G_0, \rho)$, *then for any* $j \in [J]$ *and* $r \in \{0,1,\ldots,c_j\}$,

$$\pi_j(r) = \int_0^1 \text{Binomial}(r; c_j, v)\, f_{V_j}(v)\text{d}v, \quad (7)$$

where

$$f_{V_j}(v) = \frac{\theta}{J} v \int_0^{+\infty} t\rho(tv)\, f_{T_j}(t(1-v))\text{d}t. \quad (8)$$

Above, f_{T_j} *denotes the density function of the total mass of a CRM with Lévy intensity* $\theta G_{0,\mathbb{S}_j}(\text{d}s)\rho(x)\text{d}x$.

See Appendix A2.4.1 for the proof of Theorem 3. By the tower property of the expectation, we get that, if $h(X_{n+1}) = j$, the smoothed version of (3) is

$$\hat{f}^{\text{NRM}}_{X_{n+1}} = \mathbb{E}_{P \sim \text{NRM}, h \sim \mathbb{H}_J}[\varepsilon_f(P, h)] = c_j \mathbb{E}[V_j], \quad (9)$$

where V_j is as in (8). Intuitively, $\mathbb{E}[V_j]$ is equal to the probability that two symbols sampled independently at random from $P_{\mathbb{S}_j}$ are equal (see Equation (2.25) in Pitman, 2006). Therefore, the lower $\mathbb{E}[V_j]$, the higher the number of distinct symbols in the sample $\mathbf{X}_n$, leading to more hash collisions. This, in turn, inflates c_j relative to $f_{X_{n+1}}$. Specific examples will be presented in Section 3.4, where it will also become clear that smoothed estimators are more practical and have much lower computational costs compared to their BNP counterparts.

The following result establishes an optimality property of $\hat{f}^{\text{NRM}}_{X_{n+1}}$.

Theorem 4 .

Let $\hat{f}_\beta = c_j \beta$ *for* $\beta \geq 0$ *and consider the quadratic risk associated with* β *conditional on the bucket into which observation* $n+1$ *is hashed, namely* $\text{cMSE}(\beta) = \mathbb{E}\left[(\beta C_j - f_{X_{n+1}})^2 \mid h(X_{n+1}) = j\right]$. *If* $X_1, \ldots, X_{n+1} \mid P \overset{\text{iid}}{\sim} P$ *and* $P \sim \text{NRM}(\theta, G_0, \rho)$, *then* $\beta = \mathbb{E}[V_j]$ *achieves the minimum risk. Moreover, this estimator is unbiased.*

See Appendix A2.4.2 for the proof. Theorem 4 has a clear Bayesian interpretation. Under the hierarchical model $X_1, \ldots, X_{n+1} \mid P \overset{\text{iid}}{\sim} P$ and $P \sim \text{NRM}(\theta, G_0, \rho)$, we know that the optimal estimator is the Bayes estimator: $\mathbb{E}[f_{X_{n+1}} \mid \mathbf{C}_j, h(X_{n+1}) = j]$, which can be expressed as

$\sum_{r=0}^{c_j} r\pi_j^F(r)$, where π_j^F denotes the full-sketch posterior distribution given in (5). However, as discussed in Section 2.3, this posterior is generally intractable. Theorem 4 thus says that our smoothed estimator is the closest, among all linear estimators, to this optimal but generally intractable Bayes estimator.

Moreover, Theorem 4 can guide us in estimating the smoothing parameter $\theta > 0$ and any hyper-parameters in the expression of the Lévy intensity ρ. In fact, the main assumption of Theorem 4 is that the data are randomly sampled from the random probability measure $P \sim \text{NRM}$. Therefore, a sensitive approach for choosing the smoothing parameters is to (approximately) maximize the marginal likelihood of the sketch. See Appendix A2.6 for further details on how to estimate the smoothing parameters efficiently, by leveraging a subsampling shortcut that allows this component of our method to maintain a computational cost that is constant with respect to the size of the sketched data.

3.4 Examples of smoothed estimators

As interesting examples of NRM smoothing distributions, we consider the Dirichlet (DP, Ferguson, 1973) and normalized generalized Gamma (NGGP, James, 2002; Prünster, 2002; Pitman, 2003; Lijoi et al., 2007) processes. These are NRMs with Lévy intensities $\rho(x) = \mathrm{e}^{-x}x^{-1}$ and $\rho(x) = \Gamma(1-\alpha)^{-1}x^{-1-\alpha}\mathrm{e}^{-\tau x}$, respectively. They lead to simple smoothed estimators that can be directly related to existing BNP approaches. The NGGP, in particular, allows modeling *P* with a flexible tail behaviour, ranging from geometric to heavy power-law tails. The details of the derivations summarized below are in Appendix A2.5.

Smoothing with the Dirichlet process.

Specializing (8) to the Lévy intensity of a DP with parameter $\theta > 0$, it is possible to see that $V_j \sim \text{Beta}(1, \theta/J)$. Then, for any $j \in [J]$ and $r \in \{0,1,\ldots,c_j\}$, $\pi_j(r)$ in (7) becomes

$$\pi_j(r) = \binom{c_j}{r} \frac{\theta}{J} \frac{\Gamma(r+1)\Gamma(\theta/J + c_j - r)}{\Gamma(\theta/J + c_j + 1)}, \quad (10)$$

which is the Beta-Binomial distribution with parameter $(c_j, 1, \theta/J)$. Then, (9) reduces to:

$$\hat{f}^{\text{DP}}_{X_{n+1}} := \mathbb{E}_{P\sim\text{DP},h\sim\mathcal{H}_J}[\varepsilon_f(P,h)] = c_j \frac{J}{\theta + J}, \quad (11)$$

Interestingly, this estimator coincides with the BNP estimator under the DP prior (Cai et al., 2018). However, the DP is the sole NRM for which the smoothed-Bayesian and BNP approaches lead to the same estimator for $f_{X_{n+1}}$.

Theorem 5 .

Let $\hat{f}^{\mathrm{NRM}}_{X_{n+1}}$ *be the smoothed estimator under the model (1) obtained with* $P \sim \mathrm{NRM}(\theta, G_0, \rho)$. *This estimator coincides with the BNP estimator obtained by considering the expectation of* π^F *(cf. Section 2.3) if and only if the NRM is the DP.*

See Appendix A.2.5.3 for a proof of Theorem 5. We also refer to Appendix A1.2 for a more careful discussion of the relation between smoothed estimation and the BNP approach.

The estimator in (11) applies linear shrinkage to c_j and is very fast to compute for any value of θ. Furthermore, the smoothing parameter θ can be efficiently estimated by maximizing the marginal likelihood of the data sketch, as detailed in Appendix A2.6.

It is also interesting to note that the estimator in (11) converges to the CMS estimator as $\theta \to 0$, consistent with the worst-case analysis from Section 2.2. Indeed, as $\theta \to 0$, $P \sim DP(\theta, G_0)$ approaches a degenerate distribution with only one support point; see Theorem 2. In general, (11) shrinks c_j by a weight inversely proportional to the parameter θ, which is intuitive since larger values of θ lead to a sample with more distinct symbols.

Smoothing with the normalized generalized Gamma process.

We now consider smoothing with an NGGP with parameters (θ, α, τ), which does not have a practical counterpart in the BNP framework due to the intractability of the full posterior distribution for a general NGGP prior. In this case, (9) can be evaluated by noting that

$$V_j = V_{\theta,\alpha,\tau} := B_{1-\alpha,\alpha}\left(1 - \left(\frac{\frac{\theta\tau^\alpha}{J\alpha}}{\frac{\theta\tau^\alpha}{J\alpha} + E}\right)^{1/\alpha}\right), \quad (12)$$

where $B_{1-\alpha,\alpha}$ is a Beta random variable with parameter $(1-\alpha, \alpha)$, and E is an independent negative Exponential random variable with parameter 1. Therefore, for any $j \in [J]$ and $r \in \{0, 1, \ldots, c_j\}$, $\pi_j(r)$ in (7) becomes

$$\pi_j(r) = \int_0^1 \binom{c_j}{r} v^r (1-v)^{c_j - r} f_{V_{\theta,\alpha,\tau}}(v)\mathrm{d}v, \quad (13)$$

which is a generalization of the Beta-Binomial distribution introduced in (10). Then, the estimator in (9) reduces to

$$\hat{f}^{\tiny{(\mathrm{NGGP})}}_{X_{n+1}} = c_j(1-\alpha)\left(1 - \frac{\theta\tau^\alpha}{J\alpha}\mathrm{e}^{\frac{\theta\tau^\alpha}{J\alpha}} E_{1/\alpha}\left(\frac{\theta\tau^\alpha}{J\alpha}\right)\right), \quad (14)$$

where E denotes the exponential integral function: $E_a(z) := \int_1^{+\infty} x^{-a}\exp\{-zx\}\mathrm{d}x$. See Appendix A2.5.5 for the proof of (14).

Similar to the estimator under DP smoothing, (14) also applies linear shrinkage to c_j and is straightforward to evaluate for any choice of (θ, α, τ), requiring only the computation of the exponential integral function, which is easily evaluated numerically. The smoothing parameters (θ, α, τ) can be estimated by (approximately) maximizing either the marginal likelihood of the sketch, or the marginal likelihood of a (small) subset of the data. Although this process is somewhat more involved than estimating the parameter θ in the DP case, due to the absence of a closed-form marginal likelihood expression for the general NGGP, it remains practical. See Appendix A2.6 for further details.

Note that it is intuitive why the smoothed estimator in (14) is decreasing in all three parameters θ, τ, and α: the expected number of distinct symbols in $\mathbf{X}_n$ is an increasing function of these parameters (Lijoi et al., 2007).

In the special case of $\alpha = 0$ and $\tau = 1$, (14) reduces to (11), as $\text{NGGP}(\theta, 0, 1)$ reduces to $\text{DP}(\theta)$. This equivalence also can be seen by looking at the random variable $V_{\theta,\alpha,\tau}$ in (12), since $\lim_{\alpha \to 0} V_{\theta,\alpha,1} = 1 - \mathrm{e}^{-\frac{J}{\theta}E} \overset{\mathrm{d}}{=} B_{1,\theta/J}$. A more interesting case is that of $\alpha = 1/2$, which is special NGGP known as the normalized inverse Gaussian process (NIGP) (Lijoi et al., 2005; Favaro et al., 2012). The smoothing parameter $\alpha \in [0,1)$ controls the tail behaviour of the distribution *P*: the larger α, the heavier the tail (Lijoi et al., 2007). Therefore, the NGGP can capture flexible tail behaviours for the data distribution, ranging from the geometric tails of the DP to the heavy power-law tails, which are often observed in applications.

Beyond the normalized generalized Gamma process.

Given the generality of Theorem 3, it is natural to wonder whether it might be worth considering alternative smoothing distributions (Lijoi and Prünster, 2010). To address this question, we point out that the use of other NRMs may lead to less explicit distributions for V_j in Theorem 3, which is likely to complicate the estimation of $f_{X_{n+1}}$. Fortunately, though, the NGGP already offers a reasonable trade-off between the flexibility of the smoothing assumption, in terms of enabling a flexible tail behaviour, and its tractability, in terms of leading to a distribution for V_j that can be evaluated easily. To highlight the advantages of the NGGP, we plot in Figure 1 the probability $\pi_j(r)$ computed under different choices of θ and α, assuming $\tau = 1$. The power of the NGGP can be evinced by comparing the relatively flexible behaviour of $\pi_j(r)$, for different choices of θ and α, to the much more constrained behaviour corresponding to the DP, which limits $\pi_j(r)$ to be either monotonically increasing or decreasing.

We conclude with some remarks about the connection between the smoothed frequency estimators obtained in under the DP and NGGP smoothing distributions, and the estimators obtained with the BNP approach (Beraha et al., 2024). As discussed in Section 2.3 the use of prior distributions other than the DP, including for example other NRMs, leads to computationally challenging BNP estimators for the empirical frequency $f_{X_{n+1}}$. By contrast, the advantage of the approach of smoothed estimation is that it allows one to leverage much

more flexible models from the broader class of NRMs and still obtain tractable estimators that can be written in terms of expectations of mixtures of binomial distributions.

4 Sketches with multiple hash functions

4.1 Problem statement and setup

We discuss how to extend our smoothed-Bayesian approach to combine information from sketches produced by multiple independent hash functions, as in the general setting of the CMS of Cormode and Muthukrishnan (2005). This strengthens the connection between our approach and the algorithmic approach of the CMS, reviewed in Appendix A1.1. In fact, the latter is typically applied using multiple independent hash functions, to make the data compression more efficient (Cormode and Yi, 2020, Chapter 2).

To introduce the more general sketch, for $M \geq 1$, let $h_1, \ldots, h_M$ be J-wide hash functions, with $h_l : \mathbb{S} \to [J]$ for $l \in [M]$, that are i.i.d. from a pairwise independent hash family $\mathcal{H}_J$. Hashing $(X_1, \ldots, X_n)$ through $h_1, \ldots, h_M$ produces a random matrix $\mathbf{C}_{M,J} \in \mathbb{N}_0^{M \times J}$, referred to as the sketch. Therefore, the natural extension of the setup of Section 2.1 is:

$$\begin{aligned} X_1, \ldots, X_n &\overset{\text{iid}}{\sim} P, \\ h_1, \ldots, h_M &\overset{\text{iid}}{\sim} \mathrm{H}_J, \\ C_{l,j} &= \sum_{i=1}^{n} I(h_l(X_i) = j), \qquad \forall l \in [M], \forall j \in [J]. \end{aligned} \tag{15}$$

Following the same strategy as in Section 2.1, one may think of investigating an oracle estimator for $f_{X_{n+1}}$, by assuming full knowledge of the distribution P. However, even this preliminary step becomes extremely challenging in the context of multiple hash functions. In particular, the resulting estimator involves intractable combinatorial coefficients that reduce to the multinomial coefficient only if $M = 1$. See Appendix A4 for further details.

To circumvent the computational challenges associated with direct estimation in the case of multiple hash functions, we propose an approach inspired by *multi-view* (or multimodal) learning (Xu et al., 2013; Shankar et al., 2018; Li et al., 2018). In general, multi-view learning is concerned with aggregating distinct inferences obtained from different views or representations of the same data set. Here, we focus on aggregating the separate inferences corresponding to the distinct sketches produced by each hash function.

4.2 Multi-view learning

We consider the sketch $\mathbf{C}_{M,J}$ as a collection of M distinct sketches, say $\mathbf{C}_{1,J}, \ldots, \mathbf{C}_{M,J}$, where each $\mathbf{C}_{\ell,J}$ for $\ell \in [M]$ is obtained by applying the corresponding hash function h_ℓ to the sample $(x_1, \ldots, x_n)$. Following the general multi-view learning approach, we can interpret the sketches $\mathbf{C}_{1,J}, \ldots, \mathbf{C}_{M,J}$ as M different views of the data set, and then apply Theorem 1 to each of them separately. This leads to a distinct conditional distribution for $f_{X_{n+1}}$, given the sketch

$\mathbf{C}_{\ell,J}$ and the bucket $h_\ell(X_{n+1})$ in which X_{n+1} is hashed, for each $\ell \in [M]$. That is, for each $r \in \{0,1,\ldots,c_{\ell,j}\}$, we compute the probability

$$\pi_{j_\ell}(r;P) := \Pr[\, f_{X_{n+1}} = r \,|\, \mathbf{C}_{l,J} = \mathbf{c}_\ell, h_\ell(X_{n+1}) = j_l\,]$$
$$= \binom{c_{\ell,j}}{r} \sum_{s \in \mathcal{S}_{j_\ell}} \left(\frac{p_s}{q_{j_\ell}}\right)^{r+1} \left(1 - \frac{p_s}{q_{j_\ell}}\right)^{c_{\ell,j}-r} . \qquad (16)$$

Then, we consider two practical and intuitive multi-view rules to aggregate the probabilities $\pi_{j_\ell}(r;P)$ in (16), as explained below, although other options are also possible.

The first multi-view rule that we consider is known as the *product of experts* (Hinton, 2002). For each $r \in \{0,1,\ldots,n\}$, it defines the probability mass function as follows:

$$\tilde{\pi}_{\mathbf{j}}(r;P) := \frac{1}{Z} \prod_{l=1}^{M} \pi_{j_\ell}(r,P), \quad (17)$$

where *Z* is the normalizing constant. Intuitively, the distribution function (17) assigns high mass to values of *r* for which none of the individual "single-view" distributions $\pi_{j_\ell}(r)$ is too small and are well supported by most of these distributions. This can be interpreted as seeking a consensus among "experts" (i.e., distributions) that agree on a region of values. For further intuition on the product of experts, we refer to Hinton (2002).

The second multi-view rule is called *minimum of experts*, and it is inspired by the CMS. It considers the distribution of the minimum of *M* independent random variables distributed according to (16). That is, if $\Pi_{j_\ell}(\cdot;P)$ denotes the cumulative distribution function corresponding to (16), for each $r \in \{0,1,\ldots,n\}$ we define a probability mass function $\overset{\circ}{\pi}_{\mathbf{j}}(r;P)$ associated with a cumulative distribution function given by

$$\overset{\circ}{\Pi}_{\mathbf{j}}(r;P) := 1 - \prod_{\ell=1}^{M} (1 - \Pi_{j_\ell}(r,P)). \quad (18)$$

Based on the approximations in (17) and (18) of the "true" (intractable) conditional distribution of $f_{X_{n+1}}$, we propose to utilize the corresponding expected values as practical estimators of the empirical frequency $f_{X_{n+1}}$ from the sketch $\mathbf{C}_{M,J}$, following the same arguments developed in Sections 2 and 3.

Under the (fully) nonparametric approach of Section 2.2, it can be seen that the application of the worst-case analysis to each single view leads to recovering the classical CMS algorithm from (18). Indeed, from Theorem 2 it is straightforward to see that $\pi_{j_\ell}(r) = \delta_{c_{\ell,j_\ell}}(r)$ and (18) reduces to $\min\{c_{1,j_1},\ldots,c_{M,j_M}\}$. Under the approach of smoothed estimation of Section 3, the DP smoothing assumption on (17) leads to an expression for $\tilde{\pi}_{\mathbf{j}}$ which coincides with the posterior distribution of $f_{X_{n+1}}$, given $\mathbf{C}_{M,J}$ and $\mathbf{h}(X_{n+1})$, under a DP prior, obtained in Cai et

al. (2018, Theorem 2). See also Equation (8) in Dolera et al. (2023). Therefore, the smoothed estimator $\tilde{\hat{f}}^{\mathrm{DP}}_{X_{n+1}}$ coincides with the BNP estimator proposed by Cai et al. (2018) as the posterior mean. See also Dolera et al. (2023) for further details.

Along the same lines, one may also consider further generalizations by applying the NGGP smoothing assumption to (17). These smoothing assumptions could also be applied to (18), leading to estimators that have not been previously considered in the literature.

5 Numerical illustrations

5.1 Setup and performance metrics

We present several simulation studies to assess the performance of our methods. We consider two generative models for $X_1,\ldots,X_n$: i) the process (generalized Pólya urn) induced by the Pitman-Yor process, with parameters γ (strength) and σ (discount) to be specified case-by-case, in short $\text{PYP}(\gamma,\sigma)$; ii) a Zipfian distribution with tail parameter *c*. Both display a power law behaviour. We evaluate the results based on mean absolute error between true and estimated frequencies, stratified by true frequency (Dolera et al., 2023). That is,

$$\text{MAE}_j = \frac{1}{\sum_{s\in\mathbb{S}} I(f_s \in (l_j,u_j])} \sum_{s\in\mathbb{S}} |f_s - \hat{f}_s| I\left(f_s \in (l_j,u_j]\right),$$

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

5.2 Synthetic data: single hash function

We consider the two smoothing distributions discussed in Section 3: the Dirichlet process and the NGGP. To estimate the parameters in the DP, we maximize the integrated likelihood of the observed sketch as discussed in Appendix A2.6. While considering the NGGP, we store in memory the first *m* observations $X_1,\ldots,X_m$, with $m \approx n/20$, and choose the parameters that maximize the integrated likelihood of $X_1,\ldots,X_m$. See Appendix A2.6 for further details on the associated algorithms. As far as the computational cost is concerned, optimizing under the DP smoothing is almost instantaneous (< 0.1 seconds) while for the NGGP the optimization takes 1–5 seconds in all experiments. Note that the optimization is done only once per experiment and does not need to be repeated for each query.

We generate $n =$ 100,000 observations from $\text{PYP}(\gamma,\sigma)$ for $\gamma \in \{1,10,100,1000\}$ and $\sigma \in \{0, 0.25, 0.5, 0.75\}$ and from a Zipfian distribution with parameter $c = 1.3, 1.6, 1.9, 2.2$. The data are hashed by a random hash function of with $J = 128$. Figure 2 reports the MAEs of the frequency estimators stratified across 3 frequency bins when data are generated from $\text{PYP}(\gamma, 0.75)$ and from the Zipf distribution, while Table in the Appendix reports all the remaining numerical values, while in Appendix A5.2, we discuss the impact of the parameter *J*. Results are averaged across 50 independent repetitions. It is clear that the NGGP smoothing outperforms the DP smoothing for both data generating processes, and across all values of their parameters. Henceforth, we will present results only for the NGGP smoothing. The results for the DP smoothing are available in the appendix, where it is clear that the NGGP outperforms the DP also in the other experiments.

5.3 Synthetic data: multiple hash functions

We now move to the case of multiple hash functions. In the following, we fix a total memory budget for the sketch $\mathbf{C}_{M,J}$ so that $M \times J = 1{,}000$. We generate data from a $\text{PYP}(\gamma, \sigma)$, letting $\gamma \in \{10, 100, 1000\}$ and $\sigma \in \{0.25, 0.75\}$, and simulate $n = 500{,}000$ observations for each possible combination of the parameters. In analogy with the multiview literature, we analyze each view separately (in parallel) and estimate one set of parameters for each view.

We consider frequency estimators based on the product of expert (PoE) aggregation (17) and the minimum of expert (min) aggregation (18) rules, under NGG smoothing. We also compare their performance with the original count-min sketch algorithm and the "debiased" count-min from Ting (2018), which we call D-CMS. Figure 3 reports the MAEs for two choices of the parameters of the data generating process, and Tables - in the Appendix report all the remaining values including the case of the DP smoothing.

Several insights emerge from these experiments. First, although it may not be immediately evident in Figure 3, the min aggregation rule generally outperforms the PoE rule, as shown by Tables – in the Appendix. The CMS and D-CMS demonstrate competitive performance with the smoothed estimator when the data exhibit a moderate power-law behavior (i.e., when the discount parameter of the PYP generating the data is $\sigma = 0.25$). However, with heavy power-law tails, the CMS and D-CMS incur larger errors, particularly at low frequencies. Across all datasets, the smoothed estimators perform well with a moderate (e.g., 10) number of hash functions. The CMS also benefits from using a moderate number of hash functions with moderate power laws, but with heavier power laws, the CMS achieves better performance with only one or two hash functions. In contrast, the D-CMS performs poorly when using only a single hash function.

5.4 Calibration of multi-hash aggregation rules

Our methodology can produce prediction *intervals* for $f_{X_{n+1}}$. For instance, a 95% prediction interval can be constructed using the 2.5% and 97.5% quantile of (17) or (18). It is then crucial to evaluate both the calibration of these intervals and their width. Sesia and Favaro (2022) empirically demonstrated that the BNP posterior distribution for $f_{X_{n+1}}$ under a DP prior—originally obtained in Cai et al. (2018) and coinciding with our PoE rule with DP smoothing—is often miscalibrated and results in excessively wide intervals.

The following empirical results will show that miscalibration is not unique to the Bayesian setting, but it is rather a common problem, one that affects both of the aggregation rules that we proposed. We will call the intervals obtained using (17) and (18) the "smoothed-PoE" and "smoothed-min" intervals, respectively. We propose to overcome this pitfall using the conformal inference approach proposed by Sesia and Favaro (2022). In particular, we show that using the point estimators considered previously (i.e., obtained as the expected value of (17) or (18)) to produce intervals as in Sesia and Favaro (2022) leads to shorter prediction intervals with valid coverage. We will call such intervals the "conformal-PoE" and "conformal-min" intervals, respectively. Observe that both the smoothed and conformal intervals depend on the choice of smoothing distribution.

We generate data from a PYP with parameters $(10, 0.25)$ and $(100, 0.25)$, simulating $n = 250,000$ observations. For the "conformalized" approach, we follow Sesia and Favaro (2022), using their default tuning values. Specifically, we reserve the first $m = 25,000$ observations for calibration and assess coverage and interval length on an additional 2,500 data points. Figure 4 presents results averaged over 50 independent replications. All intervals exceed the nominal 0.9 coverage level due to the discreteness of the data, which includes many repeated observations of the same symbol. Achieving exact 90% marginal coverage would require randomly perturbing the intervals corresponding to repeated queries, but we do not introduce such randomness here for simplicity.

The average length of the intervals varies significantly across settings. The top plots in Figure 4, for $\mathrm{PYP}(10, 0.25)$, show that smoothed-PoE and smoothed-min intervals are relatively large, while conformal-PoE and conformal-min intervals are much shorter and of similar length. In contrast, the bottom plots, for $\mathrm{PYP}(100, 0.25)$, indicate that smoothed and conformal intervals have comparable average length, though the conformal calibration still provides a notable improvement.

5.5 Real data sets

We consider two real data sets, displaying remarkably different behaviours in terms of their frequency distribution, that were also analyzed in Sesia and Favaro (2022). The first data set consists of the texts of 18 open-domain classic pieces of English literature from the Gutenberg Corpus. The frequency distribution here has a clear power-law behaviour. We subsample 600,000 bigrams from the corpus, displaying approximately 120,000 distinct combinations. The second data set contains nucleotide sequences from SARS-CoV-2 viruses, made publicly available by the National Center for Biotechnology Information (Hatcher et al., 2017). These data include 43,196 sequences, each consisting of approximately 30,000 nucleotides. The goal is to estimate the empirical frequency of each possible 16-mer, a distinct sequence of 16 DNA bases found in contiguous nucleotides. Since each nucleotide has one of 4 bases, there are $4^{16} \approx 4.3$ billion possible 16-mers. The frequency distribution of the DNA sequences is rather unusual: there are no "common" sequences, and most sequences appear approximately 1,000 times in the data set. There are also a few common sequences that appear 2000 times and some rare sequences that appear only a few times. We subsample 2,000,000 data points, displaying approximately 20,000 unique sequences.

We follow the same setup of Section 5.3 and fix a memory budget of $M \times J =$ 10,000, letting $M = 10, 4, 2, 1$. We sketch the data and evaluate the estimators based on the NGGP smoothing combined via the product of experts and the "min" rule. Figure 5 shows the MAEs stratified by frequency. For the both the English bigrams and the Covid-DNA data, the behaviour is as expected: the NGGP achieves the best performance. For comparison, in both datasets, the errors of the CMS algorithm are at least an order of magnitude larger than those of our smoothed estimators, which also outperform the D-CMS.

6 Discussion

Although it was not discussed in the paper, our smoothed-Bayesian approach also lends itself to address different recovery problems, such as the estimation of the total number of distinct symbols in the sketched data. This is the *cardinality recovery* problem, for which there exist algorithmic solutions (Flajolet and Martin, 1985; Flajolet et al., 2007) relying on different sketching algorithms, as well as solutions that rely on modeling assumptions for the data (Chassaing and Gerin, 2006; Chen et al., 2011; Ting, 2019; Pettie and Wang, 2021). Appendix A3 extends our results to address cardinality recovery using the same sketch obtained via random hashing. These additional results include both a worst-case theoretical analysis and a novel class of smoothed-Bayesian estimators. We believe these extensions may be of interest especially in the context of privacy-preserving analyses.

This work opens several opportunities for future research. For example, one may conduct an in-depth analysis for the problem of endowing our method with uncertainty estimates, going beyond current conformal inference approaches (Sesia et al., 2023). Further, one may consider the more general setting in which data belong to more than one symbol, referred to as traits, and exhibit nonnegative integer levels of associations with each trait; e.g., single cell data containing multiple genes with their expression levels, or documents containing different topics with their words. In the trait allocation setting, a BNP approach to frequency recovery has been developed in Cai et al. (2018) and Beraha et al. (2024), showing some computational issues in the evaluation of posterior distributions. Extending smoothed estimation to the trait allocation setting may lead to more computationally efficient estimators compared to the BNP approach, under flexible modeling assumptions.

Another research direction could involve combining smoothed estimation with "learning-based" hashing algorithms (Aamand et al., 2024), which leverage additional data features (not considered in this paper) to identify the most common symbols and hash them separately from the rest of the dataset. Further, one could explore smoothed estimation in large-scale streaming and distributed settings. This may require smoothing distributions that can describe non-exchangeable data (Airoldi et al., 2014), adapting our method to allow for a sequential estimation of smoothing parameters, and extending our multi-view formulation to account for the cost of aggregating inference across different servers under communication constraints. Moreover, our framework may be extended to problems involving more complex data such as graphs (Cormode and Yi, 2020, Chapter 7), for which several BNP models have been proposed (Caron and Fox, 2017; Ricci et al., 2022).

Software Availability

The code implementing the methods described in this paper, as well as scripts to reproduce the numerical experiments and data analysis, is publicly available at: https://github.com/mberaha/SmoothedSketching.

Acknowledgements

M.B. and S.F. were supported by the European Research Council (Horizon 2020, grant 817257). M.B. was also supported by MUR, grant Dipartimento di Eccellenza 2023-2027, and S.F. by the Italian Ministry of Education, University and Research (MIUR), "Dipartimenti di Eccellenza" grant 2018-2022. S.F. is also affiliated with IMATI-CNR "Enrico Magenes" (Milan, Italy). M.S. was supported by NSF grant DMS 2210637 and by an Amazon Research Award. The authors are grateful to the editors and anonymous referees for their valuable suggestions, which helped improve an earlier version of this manuscript.

Generative AI Use: GPT-4o was used for language improvement.

Disclosure Statement

The authors report there are no competing interests to declare.

References

Aamand, A., J. Chen, H. Nguyen, S. Silwal, and A. Vakilian (2024). Improved frequency estimation algorithms with and without predictions. *Adv. Neural Inf. Process. Syst.36*.

Aamand, A., P. Indyk, and A. Vakilian (2019). Frequency estimation algorithms under Zipfian distribution. *Preprint arXiv:1908.05198*.

Airoldi, E. M., T. Costa, F. Bassetti, F. Leisen, and M. Guindani (2014). Generalized species sampling priors with latent beta reinforcements. *J. Am. Stat. Assoc.109*(508), 1466–1480.

Beraha, M., S. Favaro, and M. Sesia (2024). Random measure priors in Bayesian recovery from sketches. *J. Mach. Learn. Res.25*(249), 1–53.

Bertsimas, D. and V. Digalakis (2021). Frequency estimation in data streams: Learning the optimal hashing scheme. *IEEE Trans. Knowl. Data Eng.35*(2), 1541–1553.

Cai, D., M. Mitzenmacher, and R. P. Adams (2018). A Bayesian nonparametric view on count-min sketch. In *Adv. Neural Inf. Process. Syst.*, Volume 31.

Cao, Y., Y. Feng, H. Wang, X. Xie, and S. K. Zhou (2024). Learning to sketch: A neural approach to item frequency estimation in streaming data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Caron, F. and E. B. Fox (2017). Sparse graphs using exchangeable random measures. *J. Royal Stat. Soc. B79*, 1295–1366.

Chassaing, P. and L. Gerin (2006). Efficient estimation of the cardinality of large data sets. *Discrete Mathematics & Theoretical Computer Science*.

Chen, A., J. Cao, L. Shepp, and T. Nguyen (2011). Distinct counting with a self-learning bitmap. *J. Am. Stat. Assoc.106*(495), 879–890.

Chung, K., M. Mitzenmacher, and S. P. Vadhan (2013). Why simple hash functions work: exploiting the entropy in a data stream. *Theory of Computing9*, 897–945.

Cormode, G. and S. Muthukrishnan (2005). An improved data stream summary: the count-min sketch and its applications. *Journal of Algorithms55*, 58–75.

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

Dolera, E., S. Favaro, and S. Peluchetti (2021). A Bayesian nonparametric approach to count-min sketch under power-law data stream. In *AISTATS*.

Dolera, E., S. Favaro, and S. Peluchetti (2023). Learning-augmented count-min sketches via Bayesian nonparametrics. *J. Mach. Learn. Res.24*, 1–60.

Favaro, S., Lijoi, and Prünster (2012). On the stick-breaking representation of normalized inverse Gaussian priors. *Biometrika99*, 663–674.

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Stat.1*, 209–230.

Ferrer i Cancho, R. and R. V. Solé (2001). Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited. *J. Quant. Linguistics8*(3), 165–173.

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

Flajolet, P. and G. Martin (1985). Probabilistic counting algorithms for database applications. *Journal of Computer and System Sciences31*, 182–209.

Giordano, R., R. Liu, M. I. Jordan, and T. Broderick (2023). Evaluating sensitivity to the stick-breaking prior in Bayesian nonparametrics. *Bayesian Analysis18*(1), 287–366.

Good, I. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika40*, 237–264.

Hatcher, E. L., S. A. Zhdanov, Y. Bao, O. Blinkova, E. P. Nawrocki, Y. Ostapchuck, A. Schaffer, and J. R. Brister (2017). Virus Variation Resource–improved response to emergent viral outbreaks. *Nucleic acids research45*, D482–D490.

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Computing14*, 1771–1800.

Hsu, C., P. Indyk, D. Katabi, and A. Vakilian (2019). Learning-based frequency estimation algorithms. In *International Conference on Learning Representations*.

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

Kingman, J. (1967). Completely random measures. *Pac. J. Math.21*, 59–78.

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

Li, Y., M. Yang, and Z. Zhang (2018). A survey of multi-view representation learning. *IEEE Trans. Knowl. Data Eng.31*, 1863–1883.

Lijoi, A., R. H. Mena, and I. Prünster (2005). Hierarchical mixture modeling with normalized inverse-Gaussian priors. *J. Am. Stat. Assoc.100*, 1278–1291.

Lijoi, A., R. H. Mena, and I. Prünster (2007). Controlling the reinforcement in Bayesian non-parametric mixture models. *J. Royal Stat. Soc. B69*, 715–740.

Lijoi, A. and I. Prünster (2010). Models beyond the Dirichlet process. In *Bayesian Nonparametrics*. Cambridge University Press.

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

Painsky, A. (2022). Convergence guarantees for the Good-Turing estimator. *J. Mach. Learn. Res.23*, 1–37.

Painsky, A. (2023). Generalized Good-Turing improves missing mass estimation. *J. Am. Stat. Assoc.118*(543), 1890–1899.

Painsky, A. (2024). Confidence intervals for parameters of unobserved events. *J. Am. Stat. Assoc.* (in press), 1–20.

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

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

Pitman, J. (2006). *Combinatorial stochastic processes: Ecole d'eté de probabilités de Saint-Flour XXXII-2002*. Springer.

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

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

Ricci, F. Z., M. Guindani, and E. Sudderth (2022). Thinned random measures for sparse graphs with overlapping communities. In *Adv. Neural Inf. Process. Syst.*

Sesia, M. and S. Favaro (2022). Conformalized frequency estimation from sketched data. In *Adv. Neural Inf. Process. Syst.*

Sesia, M., S. Favaro, and E. Dobriban (2023). Conformal frequency estimation using discrete sketched data with coverage for distinct queries. *J. Mach. Learn. Res.24*(348).

Shankar, S., L. P. Prieto, M. J. Rodriguez-Triana, and A. Ruiz-Calleja (2018). A review of multimodal learning analytics architectures. In *Int. Conf. Adv. Learn. Technol.*

Ting, D. (2018). Count-min: Optimal estimation and tight error bounds using empirical error distributions. In *Proc. Int. Conf. Knowl. Discov. Data Min.*

Ting, D. (2019). Approximate distinct counts for billions of datasets. In *Proceedings of the 2019 International Conference on Management of Data*, pp. 69–86.

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

Xu, C., D. Tao, and C. Xu (2013). A survey on multi-view learning. *Preprint arXiv:1304.5634*.

Zipf, G. K. (2016). *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio Books.

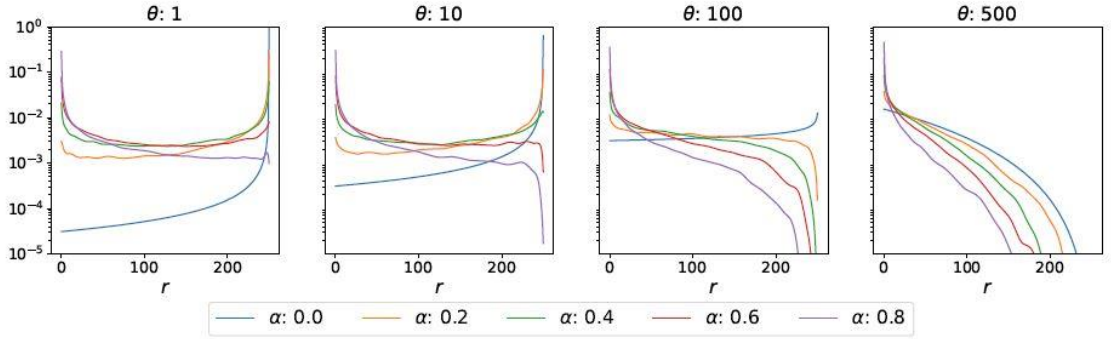

Figure 1: Visualization of the modelling flexibility of the NGGP. The probabilities $\pi_j(r)$ are plotted as a function of r for $c_j = 250$, for different values of the NGGP smoothing parameters. Different panels focus on different values of θ, while the curves drawn in different colors correspond to alternative values of α. In all cases, $\tau = 1$.

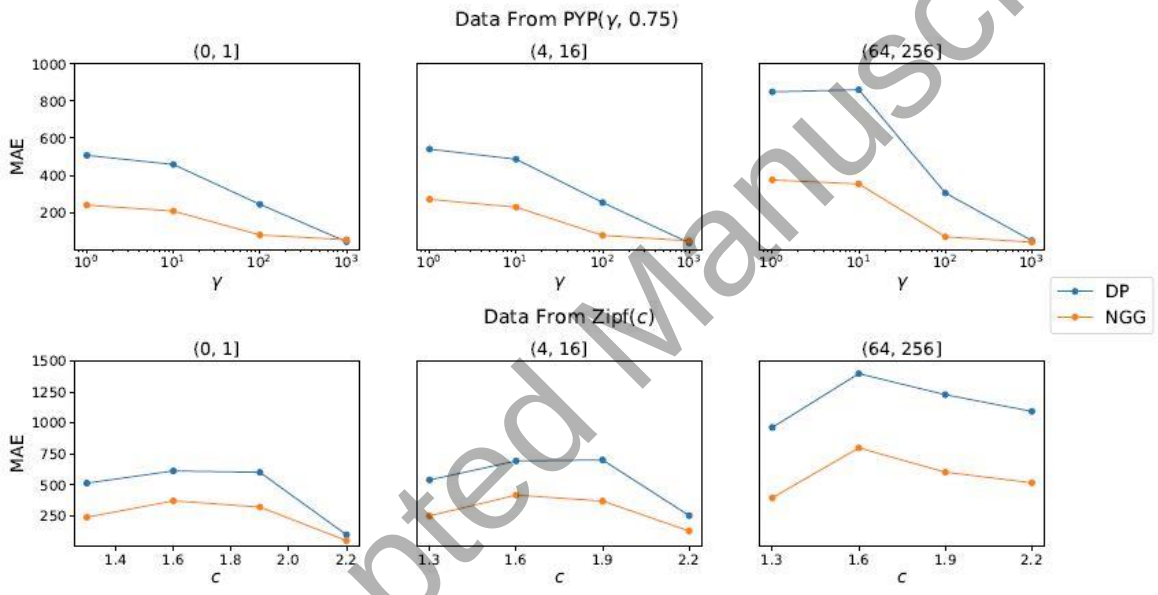

Figure 2: MAEs for the frequency estimators with DP and NGG smoothing, in experiments on synthetic data from a Pitman-Yor process with parameters γ (varies across the *x*-axis) and $\sigma = 0.75$ (see Section 5.2). Different plots correspond to different frequency bins.

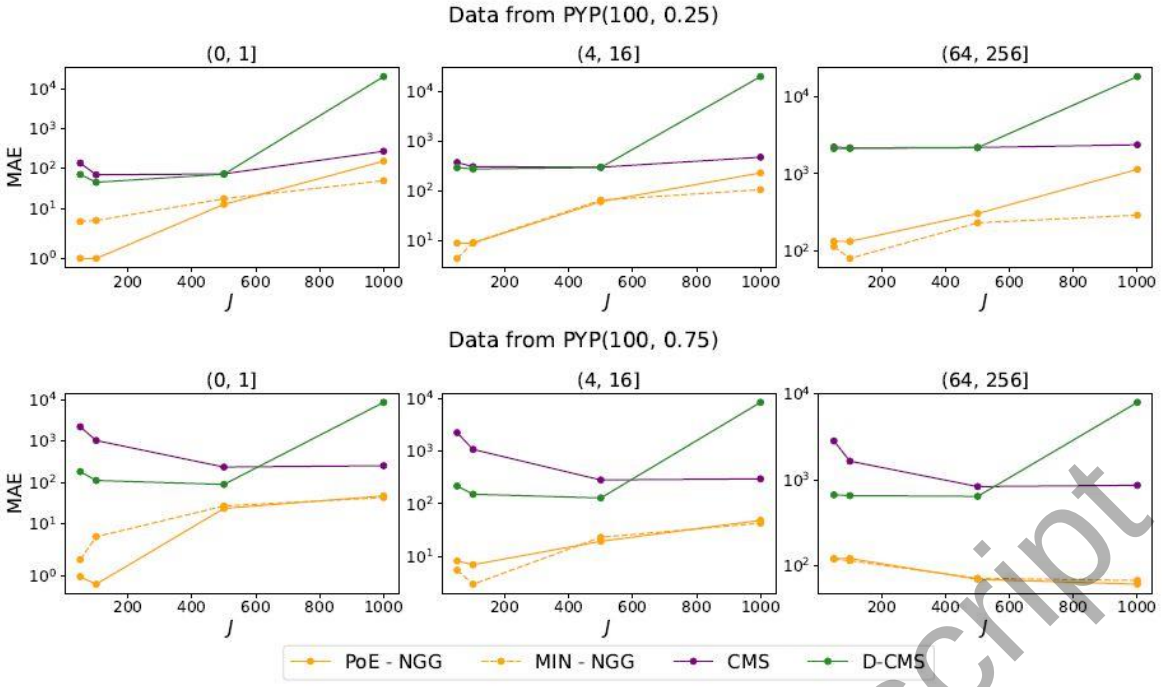

Figure 3: MAEs for the frequency estimators in Section 5.3, stratified by true frequency bins. Top row, data generated from a Pitman-Yor process with parameters $(100, 0.25)$. Bottom row, data from a Pitman-Yor process with parameters $(100, 0.75)$. In the bottom row, the PoE-NGG and MIN-NGG lines are overlapping.

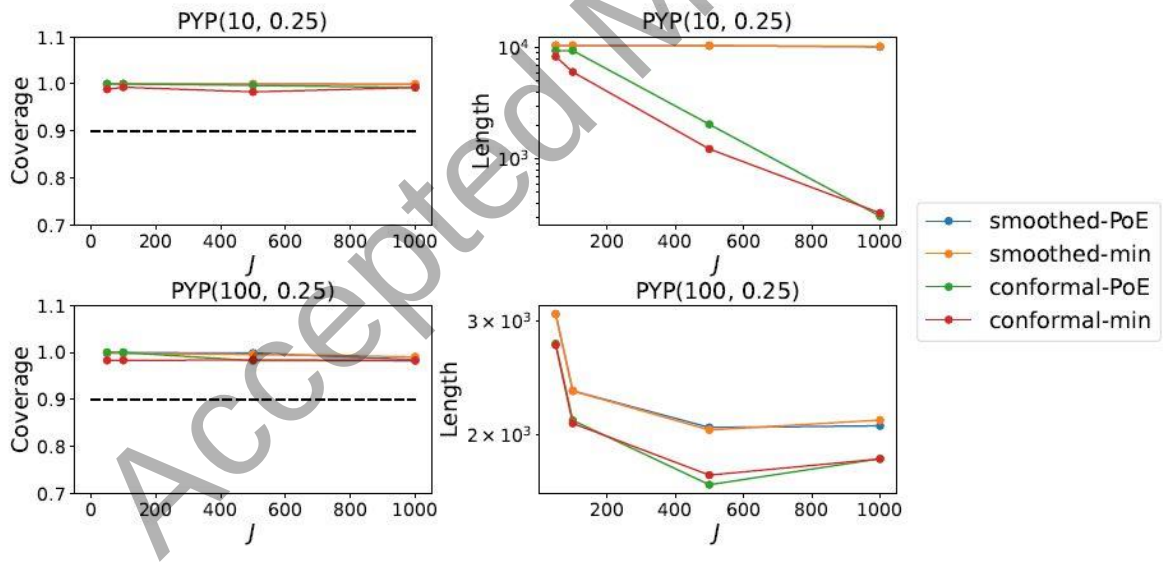

Figure 4: Calibration and average lengths of the intervals derived from the product of expert and the min aggregation rule using an NGGP smoothing for two different data generating processes, in experiments involving multiple independent hash functions.The results are shown as a function of the hash width, while the total memory budget is fixed.

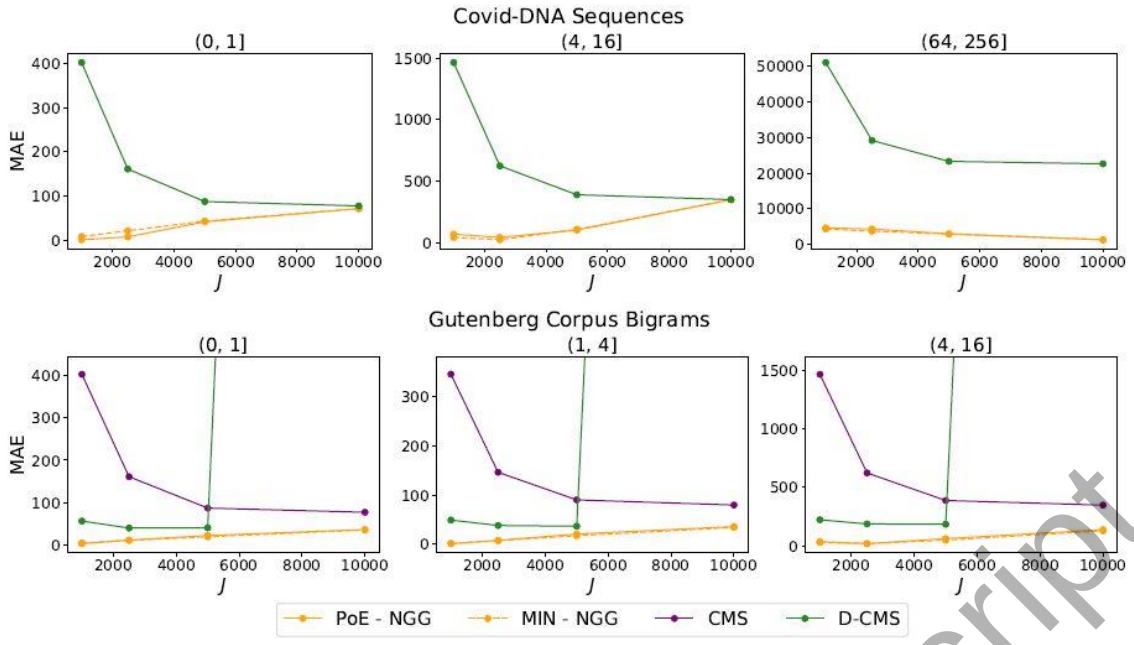

Figure 5: MAEs for the frequency estimators on the Covid-DNA sequences (top row) and Gutenberg corpus bigrams (bottom row), stratified by true frequency bins. MAEs of the CMS estimator are not reported for the Covid-DNA sequences as they are much higher.