Optimal estimation of high-order missing masses, and the rare-type match problem

Stefano Favaro, Zacharie Naulet

Optimal estimation of high-order missing masses, and the rare-type match problem

Stefano Favaro
University of Torino and Collegio Carlo Alberto
10134 Torino, Italy
stefano.favaro@unito.it

Zacharie Naulet
Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay
91405, Orsay, France
zacharie.naulet@universite-paris-saclay.fr

Abstract

For $n \geq 1$, consider a random sample $(X_1, \ldots, X_n)$ from an unknown discrete distribution $P = \sum_{j \geq 1} p_j \delta_{s_j}$ on a countable alphabet of symbols $\mathbb{S}$, and let $(Y_{n,j})_{j \geq 1}$ be the empirical frequencies of distinct symbols s_j's in the sample. In this paper, we consider the problem of estimating the r-order missing mass, which, for any $r \geq 1$, is a discrete functional of P defined as

$$\theta_r(P; \mathbf{X}_n) = \sum_{j \geq 1} p_j^r I(Y_{n,j} = 0).$$

This is generalization of the missing mass, or 1-order missing mas, whose estimation is a classical problem in statistics, being the subject of numerous studies both in theory and methods. First, we introduce a nonparametric estimator of $\theta_r(P; \mathbf{X}_n)$ and a corresponding non-asymptotic confidence interval through concentration properties of $\theta_r(P; \mathbf{X}_n)$. Then, we investigate minimax estimation of $\theta_r(P; \mathbf{X}_n)$ under a multiplicative or relative loss function, which is the main contribution of our work. We show that minimax estimation is not feasible over the class of all discrete distributions on $\mathbb{S}$, and not even for distributions with regularly varying tails, which only guarantee that our estimator is consistent for $\theta_r(P; \mathbf{X}_n)$. This leads to introduce the stronger assumption of second-order regular variation for the tail behaviour of P, which is proved to be sufficient for minimax estimation of $\theta_r(P; \mathbf{X}_n)$, making the proposed estimator an optimal minimax estimator of $\theta_r(P; \mathbf{X}_n)$. Our interest in the r-order missing mass arises from forensic statistics, where the estimation of the 2-order missing mass appears in connection to the estimation of the likelihood ratio $T(P, \mathbf{X}_n) = \theta_1(P; \mathbf{X}_n)/\theta_2(P; \mathbf{X}_n)$, known as the rare-type match problem or the "fundamental problem of forensic mathematics". We apply our results to the rare-type match problem, presenting the first theoretical guarantees to nonparametric estimation of $T(P, \mathbf{X}_n)$.

1 Introduction

The estimation of the missing mass is a classical problem in statistics, dating back to the work of Alan M. Turing and Irving J. Good at Bletchley Park in 1940s [25]. Consider a population of units taking values in a (possibly infinite) universe $\mathbb{S}$ of symbols, i.e. a countable alphabet, and consider $n \geq 1$ observable units from such a population. In its most common formulation, the problem of estimating the missing mass assumes that observable units are modeled as a random sample $\mathbf{X}_n = (X_1, \ldots, X_n)$ from an unknown distribution $P = \sum_{j \geq 1} p_j \delta_{s_j}$, with p_j being the probability of the symbol $s_j \in \mathbb{S}$, for $j \geq 1$. Denoting by $(Y_{n,j})_{j \geq 1}$ the empirical frequencies of

distinct symbols in the sample, i.e. $Y_{n,j} = \sum_{1\leq i\leq n} I(X_i = j)$, the missing mass is defined as follows:

$$\theta(P;\mathbf{X}_n) = \sum_{j\geq 1} p_j I(Y_{n,j} = 0), \tag{1}$$

namely the total probability mass of symbols not observed in the sample $\mathbf{X}_n$. The interest in the estimation of $\theta(P;\mathbf{X}_n)$ has grown over the past three decades, primarily driven by biological and physical applications [28, 22, 26, 13, 14]. In biological sciences, the missing mass mostly appears as the probability of detecting unobserved genetic variants in new (unobservable) samples, which is a critical quantity to determine how many additional genomes must be sequenced in order to explain a certain proportion of genetic variation. See [15], and references therein, for an up-to-date overview on applications of the missing mass in biology. Other applications of the missing mass, as well as generalizations thereof, can be found in, e.g., statistical machine learning and information theory [8, 35, 3], theoretical computer science [32, 9], empirical linguistics and natural language processing [20] and in forensic DNA analysis [11].

The Good-Turing estimator is arguably the most popular estimator of the missing mass [25, 38, 40]. If $M_{n,r}$ denotes the number of distinct symbols with frequency $r \geq 1$ in the random sample $\mathbf{X}_n$, i.e. $M_{n,r} = \sum_{j\geq 1} I(Y_{n,j} = r)$, then the Good-Turing estimator is

$$\hat{\theta}^{(\mathrm{GT})}(\mathbf{X}_n) = \frac{M_{n,1}}{n}.$$

This is a nonparametric estimator of $\theta(P;\mathbf{X}_n)$, as its derivation does not rely on any assumption on the distribution P. In particular, $\hat{\theta}^{(\mathrm{GT})}(\mathbf{X}_n)$ is obtained through a moment-based approach that compares the expected values of $\theta(P;\mathbf{X}_n)$ and $M_{n,1}$ [25]. It also admits a nonparametric empirical Bayes derivation in sense of [38, 39], that is $\hat{\theta}(\mathbf{X}_n)$ may be viewed as a posterior expectation with respect to an empirical (nonparametric) prior distribution. The Good-Turing estimator has been the subject of numerous studies, which led to comprehensive analysis of the problem of estimating the missing mass. These studies include, e.g., consistent and minimax estimation of $\theta(P;\mathbf{X}_n)$ with respect to both a quadratic loss and a multiplicative loss function [30, 34, 33, 31, 1, 2], large sample asymptotic properties of $\hat{\theta}^{(\mathrm{GT})}(\mathbf{X}_n)$ in terms of central limit theorems, local limit theorem and sharp large deviations [16, 17, 42, 21], and non-asymptotic concentration properties of $\theta(P;\mathbf{X}_n)$ with respect $\hat{\theta}^{(\mathrm{GT})}(\mathbf{X}_n)$ [40, 29, 33, 4, 2].

1.1 Our contributions

In this paper, we consider the problem of estimating high-order missing masses, which generalize the missing mass $\theta(P;\mathbf{X}_n)$ by taking a power function of order $r \geq 1$ for the probabilities p_j's in (1). Formally, for $r \geq 1$, we define the r-order missing mass as

$$\theta_r(P;\mathbf{X}_n) = \sum_{j\geq 1} p_j^r I(Y_{n,j} = 0). \tag{2}$$

Clearly, the missing mass in (1) is recovered from (2) by setting $r = 1$, i.e. the 1-order missing mass. We introduce a nonparametric estimator of $\theta_r(P;\mathbf{X}_n)$, which is of the form

$$\hat{\theta}_r^{(\mathrm{GT})}(\mathbf{X}_n) = \frac{M_{n,r}}{\binom{n+r}{r}},$$

and we obtain some (non-asymptotic) concentration properties of $\theta_r(P;\mathbf{X}_n)$ with respect to $\mathbb{E}_P[\theta_r(P;\mathbf{X}_n)]$. Confidence intervals for $\theta_r(P;\mathbf{X}_n)$, with respect to the estimator $\hat{\theta}_r^{(\mathrm{GT})}(\mathbf{X}_n)$, then follow as a corollary of our concentration inequalities. Our results do not rely on any assumption on the distribution P, and they generalize to the r-order missing mass some well-known concentration inequalities for the missing mass [29, 33, 4]. Such a generalization is straightforward with

respect to the left tail inequality, since it exploits the fact that $\theta_r(P;\mathbf{X}_n)$ has a sub-Gaussian left tail for any $r \geq 1$, as proved in [29, 4] for the missing mass $\theta(P;\mathbf{X}_n)$. With respect to right tail, from [4] it is known that $\theta(P;\mathbf{X}_n)$ has a sub-Gamma right tail, and our result leads to conjecture that such a tail behaviour is not preserved for $\theta_r(P;\mathbf{X}_n)$ with $r \geq 2$. For $r = 1$, our concentration inequalities may be sharper than the corresponding inequalities in [4].

Then, we investigate minimax estimation of $\theta_r(P;\mathbf{X}_n)$. Inspired by recent works on consistent estimation of the missing mass [33, 31, 2], we consider a multiplicative or relative loss function. That is, if $\hat{\theta}_r(\mathbf{X}_n)$ is an estimator of $\theta_r(P;\mathbf{X}_n)$, we consider the loss

$$\ell(\hat{\theta}_r(\mathbf{X}_n), \theta_r(P;\mathbf{X}_n)) = \left| \frac{\hat{\theta}_r(\mathbf{X}_n)}{\theta_r(P;\mathbf{X}_n)} - 1 \right|. \tag{3}$$

The use of (3) is motivated by the fact that $\theta_r(P;\mathbf{X}_n)$ is a small-valued parameter, which makes more meaningful to measure errors in a relative loss rather than in an absolute distance loss [33]. With respect to the loss (3), we show that minimax estimation of $\theta_r(P;\mathbf{X}_n)$ over the class $\mathcal{P}$ of all discrete distributions on $\mathbb{S}$ is not feasible, i.e.

$$\inf_{\hat{\theta}_r(\mathbf{X}_n)} \sup_{P \in \mathcal{P}} \mathbb{P}_P(\ell(\hat{\theta}_r(\mathbf{X}_n), \theta_r(P;\mathbf{X}_n)) \geq 1) = 1.$$

This result leads to study conditions on P to guarantee minimax estimation of $\theta_r(P;\mathbf{X}_n)$. From [33, 2] it is known that $\hat{\theta}^{(\mathrm{GT})}(\mathbf{X}_n)$ is a consistent estimator of $\theta(P;\mathbf{X}_n)$ if P has regularly varying tails [6, 24], and here we show that an analogous result holds true for $\hat{\theta}_r^{(\mathrm{GT})}(\mathbf{X}_n)$, for any $r \geq 1$. In contrast, we prove that minimax estimation of $\theta_r(P;\mathbf{X}_n)$ is not feasible under the assumption of regular variation, showing that the minimax rate over the class $\mathcal{P}$ is the same as the minimax rate over the class of all discrete distributions on $\mathbb{S}$ with regularly varying tails. Then, we introduce the stronger assumption of second-order regular variation for the tail behaviour of P, and we show that it is sufficient for minimax estimation of $\theta_r(P;\mathbf{X}_n)$, making $\hat{\theta}_r^{(\mathrm{GT})}(\mathbf{X}_n)$ an optimal estimator of $\theta_r(P;\mathbf{X}_n)$. Such a result provides a solution to the open problem of optimality in minimax estimation of the missing mass $\theta(P;\mathbf{X}_n)$, which was first discussed in [2]. Our proofs rely on novel Bayesian arguments, making use of suitable nonparametric priors that generate (almost surely) discrete distributions with regularly varying tails [36, 24].

Although the estimation of the r-order missing mass is of independent interest, our motivation to study such a problem arises from forensic statistics, where the estimation of the 2-order missing mass appears in the rare-type match problem, also known as the "fundamental problem of forensic mathematics" [7, 11, 12]. The problem refers to situations where there is a match between the characteristics of some control material and the corresponding characteristics of the recovered material, with these characteristics being rare in the sense that they are not observed in any existing database of reference. The most popular example deals with a database of $n \geq 1$ DNA profiles, say $\mathbf{X}_n$, and considers the following "matching event": the suspect's DNA profile X_{n+1} does not belong to the database $\mathbf{X}_n$ and X_{n+1} matches the crime stain's DNA profile X_{n+2}. Assuming $\mathbf{X}_n$ to be a random sample from an unknown (discrete) distribution P, the estimation of the probability of the "matching event" is applied to discriminate between the "prosecution" hypothesis that the crime stain's profile comes from the suspect, i.e. X_{n+1} is a random sample from P and X_{n+2} is equal to X_{n+1} with probability one, and the "defense" hypothesis that the crime stain's profile comes from an unknown donor, i.e. X_{n+1} and X_{n+2} are random samples from P. In particular, a largely accepted method consists in the estimation of the likelihood ratio or evidence

$$T(P;\mathbf{X}_n) = \frac{\mathbb{P}_P(\{X_{n+1} \notin \{X_1,\ldots,X_n\}\} \cap \{X_{n+1} = X_{n+2}\};\ \text{"prosecution"}, \mathbf{X}_n)}{\mathbb{P}_P(\{X_{n+1} \notin \{X_1,\ldots,X_n\}\} \cap \{X_{n+1} = X_{n+2}\};\ \text{"defense"}, \mathbf{X}_n)},$$

where: i) the numerator of $T(P;\mathbf{X}_n)$ is the probability of the "matching event" under the "prosecution" hypothesis, namely the missing mass $\theta(P;\mathbf{X}_n)$; ii) the denominator of $T(P;\mathbf{X}_n)$ is the

probability of the "matching event" under the "defense" hypothesis, namely the 2-order missing mass $\theta_2(P;\mathbf{X}_n)$. We apply our results to the rare-type match problem, presenting the first theoretical guarantees to nonparametric estimation of $T(P,\mathbf{X}_n)$.

1.2 Organization of the paper

The paper is structured as follows. In Section 2 we introduce the estimator of the r-order missing mass $\hat{\theta}_r^{(\mathrm{GT})}(\mathbf{X}_n)$ of $\theta_r(P;\mathbf{X}_n)$, present a concentration inequality for $\theta_r(P;\mathbf{X}_n)$, and apply this inequality to obtain a confidence interval for $\theta_r(P;\mathbf{X}_n)$, with respect to $\hat{\theta}_r^{(\mathrm{GT})}(\mathbf{X}_n)$. Section 3 contains a minimax analysis for the problem of estimating $\theta_r(P;\mathbf{X}_n)$, showing that, with respect to a multiplicative loss function, $\hat{\theta}_r^{(\mathrm{GT})}(\mathbf{X}_n)$ is an optimal estimator of $\theta_r(P;\mathbf{X}_n)$ if P has second-order regularly varying tails. In Section 4 we apply our results to the problem of estimating the likelihood ratio $T(P;\mathbf{X}_n)$.

2 Estimation of $\theta_r(P;\mathbf{X}_n)$ and confidence intervals

For $n \geq 1$ let $\mathbf{X}_n = (X_1,\ldots,X_n)$ be a random sample from an unknown distribution $P = \sum_{j\geq 1} p_j \delta_{s_j}$ on $\mathbb{S}$, with both the probability masses p_j's and the $\mathbb{S}$-valued atoms s_j's being unknown. The actual values taken by the X_i's is not relevant for the estimation of $\theta_r(P;\mathbf{X}_n)$, and hence $\mathbb{S}$ is an arbitrary space, e.g. the set $[0,1]$. We denote by $(Y_{n,j})_{j\geq 1}$ the empirical frequencies of distinct symbols in $\mathbf{X}_n$, i.e. $Y_{n,j} = \sum_{i=1}^n I(X_i = s_j)$ with $\sum_{j\geq 1} Y_{n,j} = n$, and by $M_{n,r}$ the number of distinct symbols with frequency $r \geq 1$ in $\mathbf{X}_n$, i.e.

$$M_{n,r} = \sum_{j\geq 1} I(Y_{n,j} = r)$$

with $\sum_{1\leq r\leq n} r M_{n,r} = n$. Moreover, let $C_{n,r} = \sum_{j\geq 1} I(Y_{n,j} \geq r)$, such that $C_{n,1}$ is the number of distinct symbols in $\mathbf{X}_n$. For any sequences $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$, write $a_n \simeq b_n$ to mean that $a_n/b_n \to 1$ as $n \to +\infty$. An estimator of the r-order missing mass $\theta_r(P;\mathbf{X}_n)$ can be obtained through a moment-based approach. Specifically, we write

$$\mathbb{E}_P(\theta_r(P;\mathbf{X}_n)) = \sum_{j\geq 1} p_j^r (1-p_j)^n = \frac{1}{\binom{n+r}{r}} \sum_{j\geq 1} \binom{n+r}{r} p_j^r (1-p_j)^n \simeq \frac{\mathbb{E}_P(M_{n,r})}{\binom{n+r}{r}},$$

and set

$$\hat{\theta}_r^{(\mathrm{GT})}(\mathbf{X}_n) = \frac{M_{n,r}}{\binom{n+r}{r}}$$

as an estimator of $\theta_r(P;\mathbf{X}_n)$, for any $n \geq 1$. The estimator $\hat{\theta}_r^{(\mathrm{GT})}(\mathbf{X}_n)$ is nonparametric, in the sense that the above moment-based derivation does not rely on any assumption on P.

Now, we obtain non-asymptotic concentration inequalities for $\theta_r(P;\mathbf{X}_n)$ with respect to $\mathbb{E}_P[\theta_r(P;\mathbf{X}_n)]$. Confidence intervals for $\theta_r(P;\mathbf{X}_n)$ then follows as a corollary.

Proposition 1. *Let $v_n(P) := \sum_{j\geq 1} p_j^{2r}(1-p_j)^n$. Then, for all $r \geq 1$, $n \geq 0$ and $x > 0$,*

$$\mathbb{P}_P\left(\theta_r(P;\mathbf{X}_n) - \mathbb{E}_P[\theta_r(P;\mathbf{X}_n)] \leq -\sqrt{2v_n(P)x}\right) \leq e^{-x}, \tag{4}$$

and,

$$\mathbb{P}_P\left(\theta_r(P;\mathbf{X}_n) - \mathbb{E}_P[\theta_r(P;\mathbf{X}_n)] \geq \sqrt{2v_n(P)x} + \left[\frac{r\log(n)}{n}\right]^r \frac{2x}{3} + \frac{e^x}{n^r}\right) \leq 2e^{-x}. \tag{5}$$

See Appendix A.1 for the proof of Proposition 1. The concentration inequalities of Proposition 1 do not rely on any assumption on the unknown distribution P, and they generalize to the r-order missing mass some concentration inequalities for the missing mass obtained in [29, 33, 4]. Regarding the left tail inequality (4), it follows by generalizing the proof of [4, Proposition 3.7] to an arbitrary $r \geq 1$. This is because $\theta_r(P; \mathbf{X}_n)$ is sub-Gaussian on the left tail, which allows to derive (4) by adapting to the case $r \geq 2$ the arguments applied in the proof of [4, Proposition 3.7]. Regarding the right tail inequality (5), [4, Theorem 3.9] established that the missing mass $\theta(P; \mathbf{X}_n)$ is sub-Gamma on the right tail, and then they obtained a concentration inequality through some standard arguments. See [5, Chapter 2 and Chapter 3] and references therein. For $r \geq 2$ it is not clear whether $\theta_r(P; \mathbf{X}_n)$ is sub-Gamma on the right tail, and most likely it is not. The inequality (5) can not be obtained by adapting to the case $r \geq 2$ the arguments applied in the proof of [4, Theorem 3.9], and then it is obtained through a suitable truncation argument. For $r = 1$, it is of interest to compare the results of Proposition 1 with the corresponding results in [4, Proposition 3.7] and [4, Theorem 3.9]. In particular, for the small $x \geq 0$ regime, our bounds have some advantages since the leading term $\sqrt{2v_n(P)x}$ is improved over [4]: it matches the left tail, and is the correct order of the variance of $\theta_r(P; \mathbf{X}_n)$. For the large $x \geq 0$ regime, however, we pay an extra $\log(n)$ factor that makes the bound of [4] bound better than our bound.

The next proposition exploits Proposition 1 in order to build a non asymptotic confidence interval for $\theta_r(P; \mathbf{X}_n)$, providing a way to quantify its uncertainty of the estimator $\hat{\theta}_r^{(\mathrm{GT})}(\mathbf{X}_n)$. We define the lower bound of our confidence intervals as, for all $x > 0$,

$$\mathfrak{L}_{n,r}(x) \coloneqq \frac{\binom{n+r}{r}}{\binom{n}{r}} \hat{\theta}_r^{(\mathrm{GT})}(\mathbf{X}_n) - \frac{a_1(n,r)}{n^r}\sqrt{C_{n,r}x} - \frac{a_2(n,r)}{n^r}x - \frac{a_3(n,r)}{n^{1+r}}C_{n,r}, \tag{6}$$

where

$$a_1(n,r) \coloneqq n^r \left[\sqrt{\frac{2}{\binom{n}{2r}}} + \frac{2\sqrt{2}}{\binom{n+r}{r}} + \frac{4\sqrt{2}r(r+1)}{(n-r)\binom{n+r}{r}}\right]\frac{\binom{n+r}{r}}{\binom{n}{r}},$$

$$a_2(n,r) \coloneqq n^r \left[\frac{4}{\sqrt{\binom{n}{2r}}} + \frac{8}{\binom{n+r}{r}} + \frac{2}{3\binom{n+r}{r}} + \frac{8r(r+1)}{(n-r)\binom{n+r}{r}}\right]\frac{\binom{n+r}{r}}{\binom{n}{r}},$$

$$a_3(n,r) \coloneqq n^{1+r}\frac{r(r+1)}{(n-r)\binom{n+r}{r}}\frac{\binom{n+r}{r}}{\binom{n}{r}}.$$

Along similar lines, we define the upper bound of our confidence intervals as, for all $x > 0$,

$$\mathfrak{U}_{n,r}(x) \coloneqq \frac{\binom{n+r}{r}}{\binom{n}{r}} \hat{\theta}_r^{(\mathrm{GT})}(\mathbf{X}_n) + \frac{b_1(n,r)}{n^r}\sqrt{C_{n,r}x} + b_2(n,r)\Big(\frac{r\log(n)}{n}\Big)^r x + \frac{\binom{n+r}{r}}{\binom{n}{r}}\frac{e^x}{n^r}, \tag{7}$$

where

$$b_1(n,r) \coloneqq n^r \left[\sqrt{\frac{2}{\binom{n}{2r}}} + \frac{2\sqrt{2}}{\binom{n+r}{r}}\right]\frac{\binom{n+r}{r}}{\binom{n}{r}}$$

$$b_2(n,r) \coloneqq \Big(\frac{n}{r\log(n)}\Big)^r \left[\frac{4}{\binom{n}{2r}} + \frac{8}{\binom{n+r}{r}} + \frac{2}{3\binom{n+r}{r}} + \frac{2}{3}\Big[\frac{r\log(n)}{n}\Big]^r\right].$$

The next proposition applies the lower bound (6) and the upper bound (7) to provide a non asymptotic confidence intervals for $\theta_r(P; \mathbf{X}_n)$. Again, we stress the fact that such a confidence interval does not rely on any assumption on the unknown distribution P.

Proposition 2. *If $n > 2r$, for any $r \geq 1$, then under P with probability at least $1 - 6e^{-x}$*

$$\theta_r(P; \mathbf{X}_n) \geq \max\Big(0,\, \mathfrak{L}_{n,r}(x)\Big)$$

and with probability at least $1 - 7e^{-x}$

$$\theta_r(P; \mathbf{X}_n) \leq \min\Big(\mathfrak{U}_{n,r}(x),\, 1\Big).$$

See Appendix A.2 for the proof of Proposition 2. The numbers $a_1(n,r)$, $a_2(n,r)$, $a_3(n,r)$, $b_1(n,r)$ and $b_2(n,r)$ can also be traded by their asymptotic limits for more convenience. In particular, by means of Stirling's formula, a direct computation shows that

$$\lim_{n\to\infty} a_1(n,r) = \lim_{n\to\infty} b_1(n,r) = \sqrt{2}\Big(2r! + \sqrt{(2r)!}\Big),$$

and

$$\lim_{n\to\infty} a_2(n,r) = \frac{26}{3}r! + 4\sqrt{(2r)!}, \quad \lim_{n\to\infty} a_3(n,r) = r(r+1)!, \quad \lim_{n\to\infty} b_2(n,r) = \frac{2}{3}.$$

The constants $a_j(n,r)$, $j = 1,2,3$, and $b_j(n,r)$, $j = 1,2$ are slightly over pessimistic, as our bounds are obtained by controlling separately $\theta_r(P;\mathbf{X}_n) - \mathbb{E}_P(\theta_r(P;\mathbf{X}_n))$ and $\hat{\theta}_r^{(\mathrm{GT})}(\mathbf{X}_n) - \mathbb{E}_P[\hat{\theta}_r^{(\mathrm{GT})}(\mathbf{X}_n)]$. In contrast, for $r = 1$ [4, Proposition 5.5.] builds a confidence interval by directly controlling $\theta_1(P;\mathbf{X}_n) - \hat{\theta}^{(\mathrm{GT})}(\mathbf{X}_n)$ through a Poisson embedding argument [24], i.e. assuming n to be a Poisson random variable. The resulting interval is tight, and although it is claimed that a similar interval can be obtained without poissonization, its tightness is not obvious. One can show that $\theta_1(P;\mathbf{X}_n) - \hat{\theta}^{(\mathrm{GT})}(\mathbf{X}_n) = \sum_{j\geq 1} f_j(Y_{n,j})$ for some functions $(f_j)_{j\geq 1}$. Under the Poisson embedding the $Y_{n,i}$'s are independent random variables, enabling for sharp concentration inequalities for $\theta_1(P;\mathbf{X}_n) - \hat{\theta}^{(\mathrm{GT})}(\mathbf{X}_n)$. Without the Poisson embedding, the $Y_{n,i}$'s are not independent random variables, and concentration inequalities can be obtained by relying on negative association. This requires to decompose $f_j(Y_{n,j}) = g_j(Y_{n,j}) + h_j(Y_{n,j})$ where $(g_j(Y_{n,j}))_{j=1}^n$ are negatively associated (respectively $(h_j(Y_{n,j}))_{j=1}^n$), and then controlling separately $\sum_{j\geq 1} g_j(Y_{n,j})$ and $\sum_{j\geq 1} h_j(Y_{n,j})$. Here there is no obvious decomposition that would improve over controlling separately $\theta_r(P;\mathbf{X}_n) - \mathbb{E}_P(\theta_r(P;\mathbf{X}_n))$ and $\hat{\theta}_r^{(\mathrm{GT})}(\mathbf{X}_n) - \mathbb{E}_P[\hat{\theta}_r^{(\mathrm{GT})}(\mathbf{X}_n)]$, even when $r = 1$. Another issue is that we do not have access to tight bounds on the variance of $M_{n,r}$, and hence of $\theta_r(P;\mathbf{X}_n)$. These issues disappear under the Poisson embedding, so that tighter intervals can be obtained.

3 Optimal estimation of $\theta_r(P;\mathbf{X}_n)$

We consider the problem of minimax estimation of the r-order missing mass $\theta_r(P;\mathbf{X}_n)$. Inspired by some recent works on consistent estimation of the missing mass [31, 2], for an estimator $\hat{\theta}_r(\mathbf{X}_n)$ of $\theta_r(P;\mathbf{X}_n)$, we consider the multiplicative or relative loss function

$$\ell(\hat{\theta}_r(\mathbf{X}_n), \theta_r(P;\mathbf{X}_n)) := \begin{cases} \Big|\frac{\hat{\theta}_r(\mathbf{X}_n)}{\theta_r(P;\mathbf{X}_n)} - 1\Big| & \text{if} \quad \theta_r(P;\mathbf{X}_n) > 0, \\ h(\hat{\theta}_r(\mathbf{X}_n)) & \text{otherwise,} \end{cases} \tag{8}$$

where h is any real-valued measurable function. The small values of $\theta_r(P;\mathbf{X}_n)$ makes more meaningful to measure the error of $\hat{\theta}_r(\mathbf{X}_n)$, through a loss function based on a relative distance, as (8), rather than through a loss function based on the absolute distance $|\hat{\theta}_r(\mathbf{X}_n) - \theta_r(P;\mathbf{X}_n)|$. See [33, 31] for a detailed discussion. To see why, one may consider the trivial estimator $\hat{\theta}_r(\mathbf{X}_n) := 0$ for which $|\hat{\theta}_r(\mathbf{X}_n) - \theta_r(P;\mathbf{X}_n)| = \theta_r(P;\mathbf{X}_n) = o_P(1)$ as $n \to \infty$. In general, there may be several unreasonable estimators of $\theta_r(P;\mathbf{X}_n)$ that are consistent in absolute distance loss. The next theorem shows that minimax estimation of $\theta_r(P;\mathbf{X}_n)$ over the class $\mathcal{P}$ of all discrete distributions on $\mathbb{S}$ is not feasible.

Theorem 1. *For every* $1 \leq r \leq n$

$$\inf_{\hat{\theta}_r(\mathbf{X}_n)} \sup_{P\in\mathcal{P}} \mathbb{P}_P\Big(\ell(\hat{\theta}_r(\mathbf{X}_n), \theta_r(P;\mathbf{X}_n)) \geq 1\Big) = 1.$$

Theorem 1 is critical for our work, and we offer two proofs of it. The first proof in Appendix A.3 uses the same ideas as [2] relying on lower bounding the minimax risk by the Bayes risk relative to a Dirichlet Process prior on $\mathcal{P}$. The second proof, in Appendix A.4, indeed establishes a weaker result, but appears to be new in this context, and its simplicity makes it in our opinion quite enlightening. The rationale is the following. Consider two distributions $P_j = (1-\omega_j)\delta_\heartsuit + \omega_j \delta_\diamond$, ω_j being small positive numbers, $j = 1, 2$. By taking ω_1, ω_2 very small the event $E := \{X_1 = \cdots = X_n = \heartsuit\}$ has probability ≈ 1 under both P_1 and P_2. But on the event E, $\theta_r(P_j; \mathbf{X}_n) = \omega_j^r > 0$ under P_j, $j = 1, 2$. So if $\frac{\omega_1}{\omega_2} \approx 0$ any estimator $\hat{\theta}_r(\mathbf{X}_n)$ that has small loss under P_1 will have large loss under P_2, and vice-versa. However, by choosing $0 < \omega_1 < \omega_2$ accordingly we can make the total variation between P_1 and P_2 arbitrarily small, which by an argument à la Le Cam implies (a slightly weaker version of) the result of Theorem 1. As a corollary of Theorem 1, for any estimator $\hat{\theta}_r(\mathbf{X}_n)$ we can find a probability distribution $P \in \mathcal{P}$ such that $\mathbb{P}_P\left(\ell(\hat{\theta}_r(\mathbf{X}_n), \theta_r(P; \mathbf{X}_n)) \geq 1\right)$ is larger than $1/2$.

With respect to the multiplicative loss function (8), the main result of [33] shows that the Good-Turing estimator $\hat{\theta}^{(\mathrm{GT})}(\mathbf{X}_n)$ is a consistent estimator of the missing mass $\theta(P; \mathbf{X}_n)$ if P has regularly varying tails. See also [31]. More precisely, a sufficient condition to enable estimation is to assume that the function $\bar{F}_P : [0, 1] \to \mathbb{Z}_+$ such that

$$\bar{F}_P(x) := \sum_{j \geq 1} I(p_j > x)$$

has regularly varying tails, namely there exists a tail-index $\alpha \in (0, 1)$ and a slowly-varying function at infinity L, i.e. $L : \mathbb{R}_+ \to \mathbb{R}_+$ with $\lim_{t\to\infty} \frac{L(tz)}{L(t)} = 1$ for any $z > 0$, such that

$$\bar{F}_P(x) \sim x^{-\alpha} L(1/x) \tag{9}$$

as $x \to \infty$ [6, 24]. We denote by $\Sigma(\alpha, L)$ the class of all discrete distributions on $\mathbb{S}$ that satisfy (9), i.e. the class of distributions with regularly varying tails. The next theorem generalizes the main result of [33] to the r-order missing mass. It shows that if P belongs to $\Sigma(\alpha, L)$ then $\hat{\theta}_r^{(\mathrm{GT})}(\mathbf{X}_n)$ is a consistent estimator of $\theta_r(P; \mathbf{X}_n)$ for any $r \geq 1$.

Proposition 3. *If $P \in \Sigma(\alpha, L)$, for some $\alpha \in (0, 1)$ and slowly-varying at infinity function L, then*

$$\sqrt{n^\alpha L(n)}\,\ell(\hat{\theta}_r^{(GT)}(\mathbf{X}_n), \theta_r(P; \mathbf{X}_n)) = O_P(1)$$

as $n \to \infty$.

See Appendix A.5 for the proof of Proposition 3. The proof of Proposition 3 is based on the application of some well-known universal limit theorems for $M_{n,r}$ and $\theta_r(P; \mathbf{X}_n)$ under $P \in \Sigma(\alpha, L)$ then both $M_{n,r}$ [27, 24]. These theorems guarantee that for any $P \in \Sigma(\alpha, L)$ there exist n_0 sufficiently large such that $M_{n,r} \simeq \mathbb{E}_P(M_{n,r})$ and $\theta_r(P; \mathbf{X}_n) \simeq \mathbb{E}_P(\theta_r(P; \mathbf{X}_n))$ for all $n \geq n_0$, with the latter expectation being determined by α and L. From Proposition 3 one may ask whether there exists n_0 not depending on P such that it is possible to estimate $\theta_r(P; \mathbf{X}_n)$ uniformly over the class $\Sigma(\alpha, L)$ at a given accuracy. The next theorem provides a negative answer to such a question, showing that the minimax rate of estimating $\theta_r(P; \mathbf{X}_n)$ over the class $\Sigma(\alpha, L)$ is the same as the minimax rate over the whole $\mathcal{P}$ when $0 < \alpha < 1$. That is, in the next theorem we show that minimax estimation of $\theta_r(P; \mathbf{X}_n)$ over the class $\Sigma(\alpha, L)$ is not feasible.

Theorem 2. *For all $0 < \alpha < 1$ and $L > 0$*

$$\inf_{\hat{\theta}_r(\mathbf{X}_n)} \sup_{P \in \Sigma(\alpha, L)} \mathbb{P}_P\left(\ell(\hat{\theta}_r(\mathbf{X}_n), \theta_r(P; \mathbf{X}_n)) \geq 1\right) = \inf_{\hat{\theta}_r(\mathbf{X}_n)} \sup_{P \in \mathcal{P}} \mathbb{P}_P\left(\ell(\hat{\theta}_r(\mathbf{X}_n), \theta_r(P; \mathbf{X}_n)) \geq 1\right) = 1.$$

See Appendix A.6 for the proof of Theorem 2. The proof of the Theorem 2 consists in showing that the class $\Sigma(\alpha, L)$ is dense in the class $\mathcal{P}$ for the topology induced by the total variation metric. This result is stated in the next proposition, and it may be of independent interest in the study of distributions with regularly varying tails.

Proposition 4. *For all $P \in \mathcal{P}$, for all $\alpha \in (0,1)$, for all $\varepsilon > 0$, and for all $L > 0$ there exists $Q \in \Sigma(\alpha, L)$ such that $\|P - Q\|_{\mathrm{TV}} \leq \varepsilon$.*

In view of Theorem 2, minimax estimation of $\theta_r(P; \mathbf{X}_n)$ requires to assume more than regularly varying tails for $\bar{F}_P$. As in the work of [18], it is natural to consider classes of discrete distributions for which the variations of $\bar{F}_P$ near zero are well-controlled. Here, to make the analysis simpler, we shall restrict to distributions for which

$$L_\alpha(P) \coloneqq \lim_{x \to 0} x^\alpha \bar{F}_P(x)$$

exists and is non-zero. See [18] for analogous tail assumptions on P. For $\alpha \in (0,1)$, $\beta > 0$, and $C > 0$, the next theorem establishes minimax estimation of $\theta_r(P; \mathbf{X}_n)$ over a class of discrete distributions on $\mathbb{S}$ with second-order regularly varying tails, i.e. the class

$$\Sigma(\alpha, \beta, C) \coloneqq \left\{ P \in \mathcal{P} : \sup_{x \in (0,1)} x^{-\beta} \left| \frac{\bar{F}_P(x)}{L_\alpha(P) x^{-\alpha}} - 1 \right| \leq C,\ 0 < L_\alpha(P) < \infty \right\}.$$

Of course, it is not necessary to assume that P belongs to $\Sigma(\alpha, \beta, C)$ to establish minimax estimation of $\theta_r(P; \mathbf{X}_n)$, though it turns out to be a convenient sufficient condition.

Theorem 3. *For all $\varepsilon > 0$ there exist $n_0 := n_0(\alpha, \beta, C, r)$ and $A > 0$ such that for all $n \geq n_0$*

$$\sup_{P \in \Sigma(\alpha,\beta,C)} \mathbb{P}_P \left(\sqrt{n^\alpha L_\alpha(P)} \ell(\hat{\theta}_r^{(GT)}(\mathbf{X}_n), \theta_r(P; \mathbf{X}_n)) > A \right) \leq \varepsilon.$$

Moreover, for all $\alpha \in (0,1)$, for all $0 < \beta < \frac{\alpha}{2}$ there exists n_0, $C > 0$, and $A' > 0$ such that

$$\inf_{\hat{\theta}_r(\mathbf{X}_n)} \sup_{P \in \Sigma(\alpha,\beta,C)} \mathbb{P}_P \left(\sqrt{n^\alpha L_\alpha(P)} \ell(\hat{\theta}_r(\mathbf{X}_n); \theta_r(P; \mathbf{X}_n)) > A' \right) \geq \frac{1}{2}.$$

See Appendix A.8 for the proof of Theorem 3. The proof of Theorem 3 does not rely on the use of Le Cam's two point method to establish the minimax lower bound. In fact, the use of Le Cam's argument requires a sharp bound on the total variation distance between the law of two random partitions with parameters $P_1, P_2 \in \Sigma(\alpha, \beta, C)$. See [18, Proposition 1] for details. It is easy to get such bound when one wants to consider the worst case with $P_1, P_2 \in \mathcal{P}$, certainly much more challenging when one wants to add the restriction that $P_1, P_2 \in \Sigma(\alpha, \beta, C)$. Instead, we lower bound the minimax risk by the Bayes risk for a suitable choice of prior distribution over $\Sigma(\alpha, \beta, C)$. The problem of optimality in minimax estimation of the missing mass $\theta(P; \mathbf{X}_n)$ was first discussed [2], where it was deferred to future work. For $r = 1$, Theorem 3 provides a solution to such a problem.

4 The rare-type match problem

In forensic statistics, the rare-type match problem refers to the situation in which the suspect's DNA profile, matching the DNA profile that is found at the crime scene, is not in the database of reference, that is the DNA profile is considered to be a rare profile. Intuitively, the rarer the suspect's DNA profile the more guilty the suspect is. See [7], and references therein, for a detailed account on the rare-type match problem and generalizations thereof. Within such a context, the crime stain's DNA profile serves as a piece of evidence to discriminate between the following mutually exclusive hypotheses:

i) the "prosecution" hypothesis that the crime stain's profile comes from the suspect; in other terms, the piece of evidence found on the crime scene is used against the suspect;

ii) the "defense" hypothesis that the crime stain's profile comes from an unknown donor; in other terms, the piece evidence found on the crime scene is used in favor of the suspect.

Both the "prosecution" hypothesis and the "defense" hypothesis require to evaluate a matching probability that is defined as the probability that a DNA profile that is not in the database of reference is found on the crime scene and on the suspect. As the population of the DNA profiles is unknown, a common approach to evaluate the matching probability consists in estimating it from the available database [7, 11, 12]. The resulting estimates are then applied to evaluate the weight of the evidence with respect to the "prosecution" hypothesis and the "defense" hypothesis, leading to a decision.

Following [7], the database of DNA profiles is modeled as a random sample $\mathbf{X}_n = (X_1, \ldots, X_n)$ from an unknown distribution $P = \sum_{j \geq 1} p_j \delta_{s_j}$, with p_j being the probability of the DNA profile s_j, for $j \geq 1$. Then, the "matching event" can be defined as follows: the suspect's DNA profile X_{n+1} does not belong to the database $\mathbf{X}_n$ and X_{n+1} matches the crime stain's DNA profile X_{n+2}. Under "prosecution" hypothesis, the probability of the "matching event" is evaluated assuming that X_{n+1} is a random sample from P and X_{n+2} is equal to X_{n+1} with probability one, which implies that

$$\mathbb{P}_P(\{X_{n+1} \notin \{X_1, \ldots, X_n\}\} \cap \{X_{n+1} = X_{n+2}\};\ \text{“prosecution”}, \mathbf{X}_n) = \theta_1(P; \mathbf{X}_n). \tag{10}$$

Instead, under the "defense" hypothesis, the probability of the "matching event" is evaluated assuming that both X_{n+1} and X_{n+2} are random samples from P, which implies that

$$\mathbb{P}_P(\{X_{n+1} \notin \{X_1, \ldots, X_n\}\} \cap \{X_{n+1} = X_{n+2}\};\ \text{“defense”}, \mathbf{X}_n) = \theta_2(P; \mathbf{X}_n). \tag{11}$$

As pointed by [7], it is assumed that the database is representative of all innocent suspects, which may fail to be true in many situations. Yet, the problem described above is fundamental to understand because if we fail to give a proper analysis of DNA evidence under the simplest assumptions, it is impossible to give a proper analysis in any situation.

The estimation of the probabilities (10) and (11) is applied to discriminate between the "prosecution" hypothesis and the "defense" hypothesis. See [7, 11, 12] and references therein. In particular, a largely accepted method consists in the estimation of the likelihood ratio

$$T(P; \mathbf{X}_n) = \frac{\theta_1(P; \mathbf{X}_n)}{\theta_2(P; \mathbf{X}_n)}.$$

Hereafter, we show how the results of Section 2 and Section 3 apply to the estimation of $T(P; \mathbf{X}_n)$. In particular, it follows that a natural nonparametric estimator of $T(P; \mathbf{X}_n)$ is

$$\hat{T}(\mathbf{X}_n) \coloneqq \frac{\hat{\theta}_1^{(\mathrm{GT})}(\mathbf{X}_n)}{\hat{\theta}_2^{(\mathrm{GT})}(\mathbf{X}_n)}.$$

Indeed, Proposition 3, in combination with Slutsky's lemma, is enough to guarantee that

$$\frac{\hat{T}(\mathbf{X}_n)}{T(P; \mathbf{X}_n)} = 1 + O_p\Big(\frac{1}{\sqrt{n^\alpha L(n)}}\Big)$$

if $P \in \Sigma(\alpha, L)$, which is an assumption considered plausible by [11, 12]. Also of interest, our Proposition 2 allows for building confidence intervals for $T(P; \cdot)$ which are non-asymptotic and valid for all $P \in \mathcal{P}$. Recalling the definitions of $\mathfrak{L}_{n,r}$ and $\mathfrak{U}_{n,r}$ from equations (6) and (7), it is immediately deduce that Proposition 2 that the intervals

$$\mathfrak{T}_n(x) \coloneqq \begin{cases} [\frac{\max(0, \mathfrak{L}_{n,1}(x)}{\min(\mathfrak{U}_{n,2}(x), 1)}, \frac{\min(\mathfrak{U}_{n,1}(x), 1)}{\mathfrak{L}_{n,2}(x)}] & \text{if } \mathfrak{L}_{n,2}(x) > 0, \\ [\frac{\max(0, \mathfrak{L}_{n,1}(x)}{\min(\mathfrak{U}_{n,2}(x), 1)}, \infty) & \text{otherwise,} \end{cases}$$

have coverage probability at least $1 - 26e^{-x}$ under P, for all $P \in \mathcal{P}$. Although $\mathfrak{T}_n(x)$ might be vacuous if P is not reasonable, this is not the case when P has regularly varying tails.

We conclude by establishing a minimax result analogous to Theorem 3, but for the ratio $T(P; \mathbf{X}_n)$. This result shows that second-order regular variation is sufficient for minimax estimation of $T(P; \mathbf{X}_n)$, making $\hat{T}(\mathbf{X}_n)$ an optimal minimax estimator of $T(P; \mathbf{X}_n)$.

Theorem 4. *For all $\varepsilon > 0$ there exist $n_0 := n_0(\alpha, \beta, C, r)$ and $A > 0$ such that for all $n \geq n_0$*

$$\sup_{P \in \Sigma(\alpha,\beta,C)} \mathbb{P}_P \left(\sqrt{n^\alpha L_\alpha(P)} \ell(\hat{T}(\mathbf{X}_n), T(P; \mathbf{X}_n)) > A \right) \leq \varepsilon.$$

Moreover, for all $\alpha \in (0, 1)$, for all $0 < \beta < \frac{\alpha}{2}$ there exists n_0, $C > 0$, and $A' > 0$ such that

$$\inf_{\hat{\theta}_r(\mathbf{X}_n)} \sup_{P \in \Sigma(\alpha,\beta,C)} \mathbb{P}_P \left(\sqrt{n^\alpha L_\alpha(P)} \ell(\hat{T}(\mathbf{X}_n); T(P; \mathbf{X}_n)) > A' \right) \geq \frac{1}{4}.$$

A Proofs

A.1 Proof of Proposition 1

A.1.1 Proof of the left tail inequality

Proceeding as in the proof of Proposition 3.7 in [4], we have for all $\lambda \in \mathbb{R}$ that

$$\log \mathbb{E}_P[e^{\lambda[\theta_r(P;\mathbf{X}_n) - \mathbb{E}_P(\theta_r(;\mathbf{X}_n))]}] \leq \sum_{j \geq 1} (1 - P_j)^n \phi(\lambda P_j^r)$$

with $\phi(x) \coloneqq e^x - 1 - x$. In particular, when $x \leq 0$ it is famous that $\phi(x) \leq x^2/2$. Therefore when $\lambda < 0$ it is the case that $\log \mathbb{E}_P[e^{\lambda[\theta_r(P;\mathbf{X}_n) - \mathbb{E}_P(\theta_r(;\mathbf{X}_n))]}] \leq \frac{\lambda^2}{2} v_n(P)$. The conclusion follows using Chernoff's bound.

A.1.2 Proof of the right tail inequality

We let $\gamma > 0$ to be chosen accordingly later. We split $\theta_r(P; \mathbf{X}_n) - \mathbb{E}_P[\theta_r(P; \mathbf{X}_n)]$ into two random variables, $Z_1 \coloneqq \sum_{j \geq 1} P_j^r [I(X_1 \neq s_j, \ldots, X_n \neq s_j) - (1 - P_j)^n] I(P_j^r \leq \gamma)$, and $Z_2 \coloneqq \sum_{j \geq 1} P_j^r [I(X_1 \neq s_j, \ldots, X_n \neq s_j) - (1 - P_j)^n] I(P_j^r > \gamma)$. We first obtain a right-tail inequality for Z_1. Proceeding as in the proof of Proposition 3.7 in [4] and via the classical argument that $x \mapsto \frac{e^x - x - 1}{x^2}$ is monotone increasing for $x > 0$, we have for all $\lambda > 0$

$$\begin{aligned}
\log \mathbb{E}_P[e^{\lambda Z_1}] &\leq \sum_{j \geq 1} (1 - P_j)^n (\lambda P_j^r)^2 \frac{e^{\lambda P_j^r} - \lambda P_j^r - 1}{(\lambda P_j^r)^2} I(P_j^r \leq \gamma) \\
&\leq \left(e^{\lambda\gamma} - \lambda\gamma - 1\right) \sum_{j \geq 1} P_j^{2r} (1 - P_j)^n.
\end{aligned}$$

Thus, Z_1 satisfies the following Bernstein inequality (see [5])

$$\mathbb{P}_P \left(Z_1 > \sqrt{2 v_n(P) x} + \frac{2}{3} \gamma x \right) \leq e^{-x}.$$

Now regarding Z_2, we see that almost-surely

$$Z_2 \leq \sum_{j \geq 1} P_j^r I(X_1 \neq s_j, \ldots, X_n \neq s_j) I(P_j^r > \gamma)$$

so that by Markov's inequality for all $t > 0$

$$\begin{aligned}\mathbb{P}_P\big(Z_2 \geq t\big) &\leq \frac{1}{t}\mathbb{E}_P(Z_2) \\ &= \frac{1}{t}\sum_{j\geq 1} P_j^r(1-P_j)^n I(P_j^r > \gamma) \\ &\leq \frac{1}{t}\sum_{j\geq 1} P_j^r e^{-nP_j} I(P_j^r \gamma) \\ &\leq \frac{1}{t} e^{-n\gamma^{1/r}}.\end{aligned}$$

The conclusion follows by choosing $t = e^u e^{-n\gamma^{1/r}}$ and choosing $n\gamma^{1/r} = r\log(n)$.

A.2 Proof of Proposition 2

A.2.1 Intermediate useful results

We recall the following result taken from Proposition 3.5 in [4].

Proposition 5. *Let define* $w_n(P) \coloneqq \min(\max_{k\in\{r,r+1\}} k\mathbb{E}_P(M_{n,k}), \mathbb{E}_P(C_{n,r}))$. *For all* $x \geq 0$ *and all* $n > r$*:*

$$\mathbb{P}_P\Big(|M_{n,r} - \mathbb{E}_P(M_{n,r})| \geq \sqrt{8w_n(P)x} + \frac{2x}{3}\Big) \leq 4e^{-x}.$$

The following result can also be trivially obtained from Proposition 3.4 in [4].

Proposition 6. *For all* $x \geq 0$ *and* $n > r$*:*

$$\mathbb{P}_P\Big(C_{n,r} \leq \mathbb{E}_P(C_{n,r}) - \sqrt{8\mathbb{E}_P(C_{n,r})x}\Big) \leq e^{-x},$$

and

$$\mathbb{P}_P\Big(C_{n,r} \geq \mathbb{E}_P(C_{n,r}) + \sqrt{8\mathbb{E}_P(C_{n,r})x} + \frac{4}{3}x\Big) \leq e^{-x}.$$

Before proving the bounds in Proposition 2, we state a certain number of intermediate results. First, Rearranging and manipulating the events involved in Proposition 6, we see that for all $x \geq 0$ and $n > r$

$$\mathbb{P}_P\Big(\sqrt{\mathbb{E}_P(C_{n,r})} \leq \sqrt{C_{n,r}} + \sqrt{8x}\Big) \geq 1 - e^{-x}. \tag{12}$$

From equation (18), we also get whenever $n \geq 2r$

$$v_n(P) \leq \sum_{j\geq 1} P_j^{2r}(1-P_j)^n \leq \sum_{j\geq 1} P_j^{2r}(1-P_j)^{n-2r} = \frac{\mathbb{E}_P(M_{n,2r})}{\binom{n}{2r}} \leq \frac{\mathbb{E}_P(C_{n,r})}{\binom{n}{2r}}. \tag{13}$$

Also,

$$\mathbb{E}_P(\hat{\theta}_r^{(\mathrm{GT})}) - \frac{\binom{n}{r}}{\binom{n+r}{r}}\mathbb{E}_P(\theta_r(P)) = \frac{\binom{n}{r}}{\binom{n+r}{r}}\sum_{j\geq 1} P_j^r(1-P_j)^{n-r}[1-(1-P_j)^r]$$

and then

$$\begin{aligned}0 \leq \mathbb{E}_P(\hat{\theta}_r^{(\mathrm{GT})}) - \frac{\binom{n}{r}}{\binom{n+r}{r}}\mathbb{E}_P(\theta_r(P)) &\leq \frac{r\binom{n}{r}}{\binom{n+r}{r}}\sum_{j\geq 1} P_j^{r+1}(1-P_j)^{n-r} \\ &\leq \frac{r(r+1)}{(n-r)\binom{n+r}{r}}\mathbb{E}_P(M_{n,r+1})\end{aligned}$$

from which we obtain

$$0 \leq \mathbb{E}_P(\hat{\theta}_r^{(\mathrm{GT})}) - \frac{\binom{n}{r}}{\binom{n+r}{r}}\mathbb{E}_P(\theta_r(P)) \leq \frac{r(r+1)}{(n-r)\binom{n+r}{r}}\mathbb{E}_P(C_{n,r}). \tag{14}$$

A.2.2 Proof of right side bound

Because $w_n(P) \leq \mathbb{E}_P(C_{n,r})$, the Proposition 5 and the equation (14) imply that we have with probability at least $1 - 4e^{-x}$

$$\begin{aligned}\hat{\theta}_r^{(\mathrm{GT})} &\leq \mathbb{E}_P(\hat{\theta}_r^{(\mathrm{GT})}) + \frac{1}{\binom{n+r}{r}}\Big(\sqrt{8\mathbb{E}_P(C_{n,r})x} + \frac{2x}{3}\Big) \\ &\leq \frac{\binom{n}{r}}{\binom{n+r}{r}}\mathbb{E}_P(\theta_r(P)) + \frac{r(r+1)}{(n-r)\binom{n+r}{r}}\mathbb{E}_P(C_{n,r}) + \frac{1}{\binom{n+r}{r}}\Big(\sqrt{8\mathbb{E}_P(C_{n,r})x} + \frac{2x}{3}\Big).\end{aligned}$$

So by Proposition 1 and equation (13), with probability at least $1 - 5e^{-x}$

$$\begin{aligned}\hat{\theta}_r^{(\mathrm{GT})} &\leq \frac{\binom{n}{r}}{\binom{n+r}{r}}\theta_r(P) + \sqrt{2v_n(P)x} + \frac{r(r+1)}{(n-r)\binom{n+r}{r}}\mathbb{E}_P(C_{n,r}) + \frac{1}{\binom{n+r}{r}}\Big(\sqrt{8\mathbb{E}_P(C_{n,r})x} + \frac{2x}{3}\Big) \\ &\leq \frac{\binom{n}{r}}{\binom{n+r}{r}}\theta_r(P) + \left[\sqrt{\frac{2}{\binom{n}{2r}}} + \frac{\sqrt{8}}{\binom{n+r}{r}}\right]\sqrt{\mathbb{E}_P(C_{n,r})x} + \frac{r(r+1)}{(n-r)\binom{n+r}{r}}\mathbb{E}_P(C_{n,r}) + \frac{1}{\binom{n+r}{r}}\frac{2x}{3}\end{aligned}$$

Then the conclusion follows from the equation (12) and direct algebraic manipulations.

A.2.3 Proof of left side bound

Because $w_n(P) \leq \mathbb{E}_P(C_{n,r})$, the Proposition 5 and the equation (14) imply that we have with probability at least $1 - 4e^{-x}$

$$\begin{aligned}\hat{\theta}_r^{(\mathrm{GT})} &\geq \mathbb{E}_P(\hat{\theta}_r^{(\mathrm{GT})}) - \frac{1}{\binom{n+r}{r}}\Big(\sqrt{8\mathbb{E}_P(C_{n,r})x} + \frac{2x}{3}\Big) \\ &\geq \frac{\binom{n}{r}}{\binom{n+r}{r}}\mathbb{E}_P(\theta_r(P)) - \frac{1}{\binom{n+r}{r}}\Big(\sqrt{8\mathbb{E}_P(C_{n,r})x} + \frac{2x}{3}\Big).\end{aligned}$$

So by Proposition 1 and equation (13), with probability at least $1 - 6e^{-x}$

$$\begin{aligned}\hat{\theta}_r^{(\mathrm{GT})} &\geq \frac{\binom{n}{r}}{\binom{n+r}{r}}\Big(\theta_r(P) - \sqrt{2v_n(P)x} - \Big[\frac{r\log(n)}{n}\Big]^r\frac{2x}{3} - \frac{e^x}{n^r}\Big) - \frac{1}{\binom{n+r}{r}}\Big(\sqrt{8\mathbb{E}_P(C_{n,r})x} + \frac{2x}{3}\Big) \\ &\geq \frac{\binom{n}{r}}{\binom{n+r}{r}}\left(\theta_r(P) - \sqrt{\frac{2\mathbb{E}_P(C_{n,r})x}{\binom{n}{2r}}} - \Big[\frac{r\log(n)}{n}\Big]^r\frac{2x}{3} - \frac{e^x}{n^r}\right) - \frac{1}{\binom{n+r}{r}}\Big(\sqrt{8\mathbb{E}_P(C_{n,r})x} + \frac{2x}{3}\Big).\end{aligned}$$

Then the conclusion follows from the equation (12) and direct algebraic manipulations.

A.3 Proof of Theorem 1 : First proof

We use the same ideas as [2] relying on lower bounding the minimax risk by the Bayes risk relative to a Dirichlet Process prior on $\mathcal{P}$. Indeed, for the sake of simplifying future proofs, we derive all the computations under the *Pitman-Yor* prior with parameters (α, d, G), $\alpha \in [0,1]$, $d > -\alpha$ and G a non-atomic center measure [which for the special case $\alpha = 0$ coincides with the Dirichlet Process prior; we refer to [23, Section 14.4] for more details and definitions].

We consider a Pitman-Yor process with parameters (α, d, G) as prior distribution, denoted in the sequel $\mathrm{PY}_{\alpha,d,G}$. We will choose $d > -\alpha$ accordingly at the end of the proof. The choice of G is irrelevant for our purpose. Then, for any estimator $\hat{\theta}_r$ and any $\varepsilon > 0$

$$\sup_{P\in\mathcal{P}} \mathbb{P}_P\Big(\ell(\hat{\theta}_r, \theta_r) > \varepsilon\Big) \geq \int_{\mathcal{P}} \mathbb{P}_P\Big(\ell(\hat{\theta}_r, \theta_r) > \varepsilon\Big)\mathrm{PY}_{\alpha,d,G}(\mathrm{d}P).$$

We now bound the rhs of the last display. Overall we assume that $0 < \varepsilon < 1$. We write Π the joint distribution of $(\mathbf{X}_n, P)$ such that $\mathbf{X}_n \mid P \sim P^{\otimes n}$ with $P \sim \mathrm{PY}_{\alpha,d,G}$. It follows that,

$$\begin{aligned}
&\int_{\mathcal{P}} \mathbb{P}_P\Big(\ell(\hat{\theta}_r(\mathbf{X}_n), \theta_r(P;\mathbf{X}_n)) > \varepsilon\Big)\mathrm{PY}_{\alpha,d,G}(\mathrm{d}P) \\
&\qquad = \mathbb{E}_\Pi\big(I(\ell(\hat{\theta}_r(\mathbf{X}_n), \theta_r(P;\mathbf{X}_n)) > \varepsilon)\big) \\
&\qquad = \mathbb{E}_\Pi\big[\mathbb{E}_\Pi\big(I(\ell(\hat{\theta}_r(\mathbf{X}_n), \theta_r(P;\mathbf{X}_n)) > \varepsilon) \mid \mathbf{X}_n\big)\big] \\
&\qquad \geq \mathbb{E}_\Pi\big[\inf_{t\in\mathbb{R}_+} \mathbb{E}_\Pi\big(I(\ell(t, \theta_r(P;\mathbf{X}_n)) > \varepsilon) \mid \mathbf{X}_n\big)\big]. \qquad (15)
\end{aligned}$$

But by Lemma 1, conditional on $\mathbf{X}_n$ it is the case that $\theta_r(P;\mathbf{X}_n)$ has the law of YZ where $Y = W_0^r$ and $Z = \sum_{j\geq 1} Q_j^r$ with $W_0 \mid \mathbf{X}_n \sim \mathrm{Beta}(d+\alpha K_n, d+n)$ is independent of $Q \mid \mathbf{X}_n \sim \mathrm{PY}_{\alpha,d+\alpha K_n,G}$. In particular $\theta_r(P;\mathbf{X}_n)$ is almost-surely non-zero, so $\ell(t, \theta_r(P;\mathbf{X}_n)) = \big|\frac{t}{\theta_r} - 1\big|$ almost-surely too. It follows for all $t \in \mathbb{R}_+$

$$\begin{aligned}
&\mathbb{E}_\Pi\big(I(\ell(t, \theta_r(P;\mathbf{X}_n)) > \varepsilon) \mid \mathbf{X}_n\big) \\
&\qquad = \mathbb{E}_\Pi\Big(I\Big(\Big|\frac{t/Z}{Y} - 1\Big| > \varepsilon\Big) \mid \mathbf{X}_n\Big) \\
&\qquad = \mathbb{E}_\Pi\Big[\mathbb{E}_\Pi\Big(I\Big(\Big|\frac{t/Z}{Y} - 1\Big| > \varepsilon\Big) \mid \mathbf{X}_n, Z\Big) \mid \mathbf{X}_n\Big] \\
&\qquad \geq \mathbb{E}_\Pi\Big[\inf_{z\in\mathbb{R}_+} \mathbb{E}_\Pi\Big(I\Big(\Big|\frac{t/Z}{Y} - 1\Big| > \varepsilon\Big) \mid \mathbf{X}_n, Z = z\Big) \mid \mathbf{X}_n\Big] \\
&\qquad = \inf_{z\in\mathbb{R}_+} \mathbb{E}_\Pi\Big[I\Big(\Big|\frac{t/z}{Y} - 1\Big| > \varepsilon\Big) \mid \mathbf{X}_n\Big] \qquad (16)
\end{aligned}$$

where the last line follows because Y and Z are independent conditional on $\mathbf{X}_n$. It follows by plugging (16) into (15)

$$\begin{aligned}
&\int_{\mathcal{P}} \mathbb{P}_P\Big(\ell(\hat{\theta}_r(\mathbf{X}_n), \theta_r(P;\mathbf{X}_n)) > \varepsilon\Big)\mathrm{PY}_{\alpha,d,G}(\mathrm{d}P) \\
&\qquad\qquad \geq \mathbb{E}_\Pi\Big[\inf_{t\in\mathbb{R}_+} \mathbb{E}_\Pi\Big(I\Big(\Big|\frac{t}{Y} - 1\Big| > \varepsilon\Big) \mid \mathbf{X}_n\Big)\Big].
\end{aligned}$$

Remark that if $t \neq 0$ then

$$\begin{aligned}
\Big|\frac{t}{Y} - 1\Big| \leq \varepsilon \iff 1 - \varepsilon \leq \frac{t}{W_0^r} \leq 1 + \varepsilon \\
\iff \frac{1}{(1+\varepsilon)^{1/r}} \leq \frac{W_0}{t^{1/r}} \leq \frac{1}{(1-\varepsilon)^{1/r}} \\
\implies \Big|\frac{W_0}{t^{1/r}} - 1\Big| \leq \max\Big(\frac{(1+\varepsilon)^{1/r} - 1}{(1+\varepsilon)^{1/r}}, \frac{1 - (1-\varepsilon)^{1/r}}{(1-\varepsilon)^{1/r}}\Big) \\
\implies \Big|\frac{W_0}{t^{1/r}} - 1\Big| \leq \max\Big(\frac{(\varepsilon/r)e^{\varepsilon/r}}{(1+\varepsilon)^{1/r}}, \frac{\varepsilon/r}{(1-\varepsilon)^{1/r}}\Big) \\
\implies \Big|\frac{W_0}{t^{1/r}} - 1\Big| \leq \frac{\varepsilon/r}{1-\varepsilon}.
\end{aligned}$$

So letting $\varepsilon' := \frac{\varepsilon/r}{1-\varepsilon}$, we deduce that

$$\int_{\mathcal{P}} \mathbb{P}_P\Big(\ell(\hat{\theta}_r(\mathbf{X}_n), \theta_r(P; \mathbf{X}_n)) > \varepsilon\Big) \mathrm{PY}_{\alpha,d,G}(\mathrm{d}P) \\ \geq \mathbb{E}_\Pi\Big[\inf_{t\in\mathbb{R}_+^*} \mathbb{E}_\Pi\Big(I\Big(\Big|\frac{W_0}{t} - 1\Big| > \varepsilon'\Big) \mid \mathbf{X}_n\Big)\Big].$$

Next apply the Lemma 2 with $\alpha \equiv d + \alpha K_n$ and $\beta \equiv d + n$ to find that by choosing $d \geq 1$ there exists a universal constant $C > 0$ such that for all $n \geq 1$ [recall $1 \leq K_n \leq n$ and $\alpha \in (0, 1)$]

$$\mathbb{E}_\Pi\Big[\inf_{t\in\mathbb{R}_+^*} \mathbb{E}_\Pi\Big(I\Big(\Big|\frac{W_0}{t} - 1\Big| > \varepsilon'\Big) \mid \mathbf{X}_n\Big)\Big] \geq 1 - C\sqrt{d + \alpha K_n}\varepsilon'. \tag{17}$$

So the conclusion follows by choosing $\alpha = 0$ and by letting $d \to -\alpha = 0$.

Lemma 1. *If $P \sim \mathrm{PY}_{\alpha,d,G}$ for $\alpha \in [0, 1)$ and $d > -\alpha$, then the posterior distribution of P based on observations $X_1, \ldots, X_n \mid P \overset{iid}{\sim} P$ is the distribution of the random measure*

$$R_n \sum_{j=1}^{K_n} W_j \delta_{\tilde{X}_j} + (1 - R_n) Q_n$$

where $R_n \sim \mathrm{Beta}(n - \alpha K_n, d + \alpha K_n)$, $(W_1, \ldots, W_{K_n}) \sim \mathrm{Dirichlet}(K_n; N_{1,n} - \alpha, \ldots, N_{K_n,n} - \alpha)$, and $Q_n \sim \mathrm{PY}_{\alpha,d+\alpha K_n,G}$), all independently distributed. Here $\tilde{X}_1, \ldots, \tilde{X}_{K_n}$ are the distinct values of $X_1, \ldots, X_n$ and $N_{1,n}, \ldots, N_{K_n,n}$ their multiplicities.

Proof. See [23, Theorem 14.37]. □

Lemma 2. *Let $X \sim \mathrm{Beta}(\alpha, \beta)$ with $\alpha, \beta > 1$. Let $0 < \delta < 1$. Then,*

$$\inf_{t\in\mathbb{R}_+^*} P\Big(\Big|\frac{X}{t} - 1\Big| > \delta\Big) \geq 1 - \frac{\delta\alpha\sqrt{\alpha + \beta - 1}\sqrt{2/\pi}}{\sqrt{(\alpha - 1)(\beta - 1)}} \exp\Big(\frac{\alpha + \beta - 1}{\beta - 1}\delta + \frac{1}{12(\alpha + \beta) - 2}\Big).$$

Proof. Let g denote the density of the $\mathrm{Beta}(\alpha, \beta)$ distribution. Then,

$$P(|X/t - 1| > \delta) = 1 - P\big(t(1 - \delta) \leq X \leq t(1 + \delta)\big) = 1 - \int_{t(1-\delta)}^{t(1+\delta)} g(x)\mathrm{d}x.$$

The function $t \mapsto \int_{t(1-\delta)}^{t(1+\delta)} g(x)\mathrm{d}x$ is maximized when $(1 + \delta)g(t(1 + \delta)) = (1 - \delta)g(t(1 - \delta))$; ie when

$$(1 + \delta)^\alpha [1 - t(1 + \delta)]^{\beta-1} = (1 - \delta)^\alpha [1 - t(1 - \delta)]^{\beta-1}$$

whose solution is given by

$$t_* = \frac{1}{1 + \delta} \cdot \frac{1 - (1 - \frac{2\delta}{1+\delta})^{\alpha/(\beta-1)}}{1 - (1 - \frac{2\delta}{1+\delta})^{(\alpha+\beta-1)/(\beta-1)}}.$$

Deduce that,

$$\inf_{t\in\mathbb{R}_+^*} P(|X/t - 1| > \delta) \geq 1 - 2t_*\delta\|g\|_\infty.$$

A lower bound on t_* is easily obtained using that $uxe^{-ux} \leq 1 - (1 - x)^u \leq ux$ for all $x \in [0, 1]$ and all $u > 1$. Namely,

$$t_* \leq \frac{\alpha}{\alpha + \beta - 1} \exp\Big(\frac{\alpha + \beta - 1}{\beta - 1}\delta\Big).$$

To obtain an upper bound on $\|g\|_\infty$, we note that the mode of the Beta(α,β) distribution is located at $x_* = \frac{\alpha-1}{\alpha+\beta-2}$. Then, using the famous relation [41, Section 3.6]

$$z\log(z) - z + \frac{1}{2}\log\frac{2\pi}{z} \leq \log\Gamma(z) \leq z\log(z) - z + \frac{1}{2}\log\frac{2\pi}{z} + \frac{1}{12z}$$

we obtain

$$\begin{aligned}
\|g\|_\infty &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x_*^{\alpha-1}(1-x_*)^{\beta-1} \\
&= \frac{(\alpha+\beta-2)(\alpha+\beta-1)\Gamma(\alpha+\beta-2)}{(\alpha-1)\Gamma(\alpha-1)(\beta-1)\Gamma(\beta-1)} \cdot \frac{(\alpha-1)^{\alpha-1}(\beta-1)^{\beta-1}}{(\alpha+\beta-2)^{\alpha+\beta-2}} \\
&\leq \frac{(\alpha+\beta-1)\sqrt{\alpha+\beta-2}}{\sqrt{(\alpha-1)(\beta-1)}} \cdot \frac{1}{\sqrt{2\pi}} e^{\frac{1}{12(\alpha+\beta-2)}} \\
&\leq \frac{(\alpha+\beta-1)^{3/2}}{\sqrt{(\alpha-1)(\beta-1)}} \cdot \frac{1}{\sqrt{2\pi}} e^{\frac{1}{12(\alpha+\beta-2)}}.
\end{aligned}$$

In the end, we have found that

$$\begin{aligned}
&\inf_{t\in\mathbb{R}_+^*} P(|X/t - 1| > \delta) \\
&\quad \geq 1 - \frac{\delta\alpha\sqrt{\alpha+\beta-1}\sqrt{2/\pi}}{\sqrt{(\alpha-1)(\beta-1)}} \exp\Big(\frac{\alpha+\beta-1}{\beta-1}\delta + \frac{1}{12(\alpha+\beta)-2}\Big)
\end{aligned}$$

This concludes the proof. □

A.4 Proof of Theorem 1 : Second proof

For $\omega_j \in (0,1)$ we let $P_j = (1-\omega_j)\delta_\heartsuit + \omega_j\delta_\diamond$, $j = 1,2$. By Le Cam's two point method, for all $A \in (0,1)$ and for all estimators $\hat{\theta}_r$

$$\begin{aligned}
&\sup_{P\in\mathcal{P}} \mathbb{P}_P\Big(\ell(\hat{\theta}_r(\mathbf{X}_n), \theta_r(P;\mathbf{X}_n)) \geq A\Big) \\
&\qquad \geq \frac{1}{2}\mathbb{P}_{P_1}\Big(\ell(\hat{\theta}_r(\mathbf{X}_n), \theta_r(P_1;\mathbf{X}_n)) \geq A\Big) + \frac{1}{2}\mathbb{P}_{P_2}\Big(\ell(\hat{\theta}_r(\mathbf{X}_n), \theta_r(P_2;\mathbf{X}_n)) \geq A\Big).
\end{aligned}$$

Conditional on the event $E := \{X_1 \neq \diamond, \ldots, X_n \neq \diamond\}$, it is the case that $\theta_r(P_j;\mathbf{X}_n) = \omega_j^r$ almost-surely under P_j, $j = 1,2$. Therefore,

$$\mathbb{P}_{P_j}\Big(\ell(\hat{\theta}_r(\mathbf{X}_n), \theta_r(P_j;\mathbf{X}_n)) \geq A \bigcap E\Big) = \mathbb{P}_{P_j}\Big(\Big|\frac{\hat{\theta}_r(\mathbf{X}_n)}{\omega_j^r} - 1\Big| \geq A \bigcap E\Big), \quad j = 1,2.$$

Now we make the choice that

$$\omega_1 = \omega_2\Big(\frac{1-A}{1+A}\Big)^{1/r}.$$

With this choice, observe that

$$\begin{aligned}
\Big|\frac{\hat{\theta}_r(\mathbf{X}_n)}{\omega_1^r} - 1\Big| < A &\iff \omega_1^r(1-A) < \hat{\theta}_r(\mathbf{X}_n) < \omega_1^r(1+A) \\
&\iff \omega_2^r\frac{(1-A)^2}{1+A} < \hat{\theta}_r(\mathbf{X}_n) < \omega_2^r(1-A) \\
&\implies \Big|\frac{\hat{\theta}_r(\mathbf{X}_n)}{\omega_2^r} - 1\Big| \geq A.
\end{aligned}$$

Letting $F \coloneqq \{|\theta_r(\hat{\mathbf{X}}_n)/\omega_1^r - 1| \geq A\}$, it follows that

$$\begin{aligned}
&\sup_{P\in\mathcal{P}} \mathbb{P}_P\Big(\ell(\hat{\theta}_r(\mathbf{X}_n), \theta_r(P;\mathbf{X}_n)) \geq A\Big) \\
&\quad\geq \frac{1}{2}\mathbb{P}_{P_1}(F\cap E) + \frac{1}{2}\mathbb{P}_{P_2}(F^c\cap E) \\
&\quad\geq \frac{1}{2}\mathbb{P}_{P_1}(F\cap E) + \frac{1}{2}\mathbb{P}_{P_2}((F\cap E)^c) - \frac{1}{2}\mathbb{P}_{P_2}(E^c) \\
&\quad\geq \frac{1}{2}\Big(1 - \|\mathbb{P}_{P_1} - \mathbb{P}_{P_2}\|_{\mathrm{TV}}\Big) - \frac{n\omega_2}{2}
\end{aligned}$$

where $\|\mathbb{P}_{P_1} - \mathbb{P}_{P_2}\|_{\mathrm{TV}}$ denotes the total variation distance between $\mathbb{P}_{P_1}$ and $\mathbb{P}_{P_2}$. Note that $\|P_1 - P_2\|_{\mathrm{TV}} = \omega_2 - \omega_1 \leq \omega_2$ which goes to zero as $\omega_2 \to 0$. This in particular implies that $\|\mathbb{P}_{P_1} - \mathbb{P}_{P_2}\|_{\mathrm{TV}} \to 0$ as well when $\omega_2 \to 0$. Since the last display was true for all $\omega_2 \in (0,1)$ and all $A \in (0,1)$, we deduce that

$$\sup_{P\in\mathcal{P}} \mathbb{P}_P\Big(\ell(\hat{\theta}_r(\mathbf{X}_n), \theta_r(P;\mathbf{X}_n)) \geq 1\Big) \geq \frac{1}{2}.$$

The previous display is a weaker statement than the one obtained using the first proof of Theorem 1, yet its proof is simpler and shed some lights on the impossibility of estimating the missing mass using a relative loss function.

A.5 Proof of Proposition 3

The proof is nearly identical to the proofs in [4]. Nevertheless, we recall here the main steps as they will be useful later on to prove our Theorem 3. We first remark that when $P \in \Sigma(\alpha, L)$, then it must be that $\ell(\hat{\theta}_r^{(\mathrm{GT})}, \theta_r) = \Big|\frac{\hat{\theta}_r^{(\mathrm{GT})}(\mathbf{X}_n)}{\theta_r(P;\mathbf{X}_n)} - 1\Big|$ almost-surely. Then, it can be seen that on the event $\{|\hat{\theta}_r^{(\mathrm{GT})} - \mathbb{E}_P(\hat{\theta}_r^{(\mathrm{GT})})| \leq t\mathbb{E}_P(\hat{\theta}_r^{(\mathrm{GT})})\} \bigcap \{|\theta_r(P;\mathbf{X}_n) - \mathbb{E}_P(\theta_r(P;\mathbf{X}_n))| \leq t\mathbb{E}_P(\theta_r(P;\mathbf{X}_n))\}$ it must be that $\ell(\hat{\theta}_r^{(\mathrm{GT})}, \theta_r) \leq \frac{2t}{1-t}$. Deduce that

$$\begin{aligned}
&\mathbb{P}_P\Big(\ell(\hat{\theta}_r^{(\mathrm{GT})}, \theta_r) > \frac{2t}{1-t}\Big) \\
&\quad\leq \mathbb{P}_P\Big(\Big|\frac{\hat{\theta}_r^{(\mathrm{GT})}(\mathbf{X}_n)}{\mathbb{E}_P(\hat{\theta}_r^{(\mathrm{GT})}(\mathbf{X}_n))} - 1\Big| > t\Big) + \mathbb{P}_P\Big(\Big|\frac{\theta_r(P;\mathbf{X}_n)}{\mathbb{E}_P(\theta_r(P;\mathbf{X}_n))} - 1\Big| > t\Big) \\
&\quad\leq \frac{\mathrm{var}_P(\hat{\theta}_r^{(\mathrm{GT})}(\mathbf{X}_n))}{t^2\mathbb{E}_P(\hat{\theta}_r^{(\mathrm{GT})}(\mathbf{X}_n))^2} + \frac{\mathrm{var}_P(\theta_r(P;\mathbf{X}_n)}{t^2\mathbb{E}_P(\theta_r(P;\mathbf{X}_n))} \\
&\quad= \frac{\mathrm{var}_P(M_{n+r,r})}{t^2\mathbb{E}_P(M_{n+r,r})^2} + \frac{\mathrm{var}_P(\theta_r(P;\mathbf{X}_n)}{t^2\mathbb{E}_P(\theta_r(P;\mathbf{X}_n))^2}
\end{aligned}$$

by Chebychev's inequality. By the Proposition 3.3 in [4] (see also Lemma A.2 in [18]), we have the bound

$$\mathrm{var}_P(M_{n+r,r}) \leq 2\max\Big(r\mathbb{E}_P(M_{n+r,r}), (r+1)\mathbb{E}_P(M_{n+r,r+1})\Big).$$

Also, by the same negative association argument as in the proof of Proposition 3.6 in [4], we find that

$$\mathrm{var}_P(\theta_r(P;\mathbf{X}_n)) \leq \sum_{j=1}^{\infty} P_j^{2r}(1-P_j)^n = \frac{\mathbb{E}_P(M_{n+2r,2r})}{\binom{n+2r}{2r}}. \tag{18}$$

Since $\mathbb{E}_P(\theta_r(P;\mathbf{X}_n)) = \mathbb{E}_P(M_{n+r,r})/\binom{n+r}{r}$, we deduce the bound

$$\mathbb{P}_P\Big(\ell(\hat{\theta}_r^{(\mathrm{GT})},\theta_r) > \frac{2t}{1-t}\Big) \leq \frac{2\max\big(r\mathbb{E}_P(M_{n+r,r}),(r+1)\mathbb{E}_P(M_{n+r,r+1})\big)}{t^2\mathbb{E}_P(M_{n+r,r})^2} + \frac{\binom{n+r}{r}^2\mathbb{E}_P(M_{n+2r,2r})}{t^2\binom{n+2r}{2r}\mathbb{E}_P(M_{n+r,r})^2}. \quad (19)$$

The proof follows using Stirling's formula and the fact that $\mathbb{E}_P(M_{n,k}) \sim \frac{\alpha\Gamma(k-\alpha)}{k!}n^\alpha L(n)$ as $n\to\infty$ if $P\in\Sigma(\alpha,L)$; a fact that which can be found for instance in Theorem 4.2 in [4], or in [27], or in [24], or in Appendix E in [18]. In particular, the last rhs is $O\big(\frac{1}{t^2 n^\alpha L(n)}\big)$ as $n\to\infty$.

A.6 Proof of Theorem 2

Using a density argument, we establish that the minimax risk over $\Sigma(\alpha,L)$ is the same as the risk over $\mathcal{P}$, then the result follows from the Theorem 1. In particular, in Proposition 4 below, we prove that $\Sigma(\alpha,L)$ is dense in $\mathcal{P}$ for the topology induced by the distance $(P,Q)\mapsto\|P^{\otimes n}-Q^{\otimes n}\|_{\mathrm{TV}}$ [it is a well-known fact that this topology is equivalent to the topology induced by the distance $(P,Q)\mapsto\|P-Q\|_{\mathrm{TV}}$]. Consequently, for all estimator $\hat{\theta}_r$ and all $\varepsilon>0$, and for all $P\in\mathcal{P}$, we can find $Q\in\Sigma(\alpha,L)$ such that

$$\begin{aligned}\mathbb{P}_P\big(\ell(\hat{\theta}_r,\theta_r)\geq 1\big) &\leq \mathbb{P}_Q\big(\ell(\hat{\theta}_r,\theta_r)\geq 1\big) + \|P^{\otimes n}-Q^{\otimes n}\|_{\mathrm{TV}} \\ &\leq \sup_{Q\in\Sigma(\alpha,L)}\mathbb{P}_Q\big(\ell(\hat{\theta}_r,\theta_r)\geq 1\big)+\varepsilon.\end{aligned}$$

Hence for all estimator $\hat{\theta}_r$

$$\sup_{P\in\mathcal{P}}\mathbb{P}_P\big(\ell(\hat{\theta}_r,\theta_r)\geq 1\big) \leq \sup_{Q\in\Sigma(\alpha,L)}\mathbb{P}_Q\big(\ell(\hat{\theta}_r,\theta_r)\geq 1\big). \quad (20)$$

On the other hand, it is obvious that for all estimator $\hat{\theta}_r$

$$\sup_{P\in\mathcal{P}}\mathbb{P}_P\big(\ell(\hat{\theta}_r,\theta_r)\geq 1\big) \geq \sup_{Q\in\Sigma(\alpha,L)}\mathbb{P}_Q\big(\ell(\hat{\theta}_r,\theta_r)\geq 1\big). \quad (21)$$

Combining equations (20) and (21), we obtain that

$$\inf_{\hat{\theta}_r}\sup_{P\in\mathcal{P}}\mathbb{P}_P\big(\ell(\hat{\theta}_r,\theta_r)\geq 1\big) = \inf_{\hat{\theta}_r}\sup_{Q\in\Sigma(\alpha,L)}\mathbb{P}_Q\big(\ell(\hat{\theta}_r,\theta_r)\geq 1\big).$$

A.7 Proof of Proposition 4

Let us write $\mathfrak{P}$ the support of P, and $P=\sum_{s\in\mathfrak{P}}P_s\delta_s$. Without loss of generality we can assume that $\mathfrak{P}$ is finite and $\min_s P_s\geq\delta$ for some $\delta\in(0,1)$. This is because the set of measures with finite support and with masses bounded from below is dense in $\mathcal{P}$ for the total variation metric. Now we let $M=\sum_{t\in\mathfrak{M}}M_t\delta_t$ be a probability measure on the set of symbols $\mathfrak{M}$ such that $\mathfrak{M}\cap\mathfrak{P}=\varnothing$ and $\lim_{x\to 0}x^\alpha\bar{F}_M(x)=L\eta^{-\alpha}$ for $\eta>0$ small. Such a measure M always exists by the Lemma 3 below. Now build the measure $Q=(1-\eta)P+\eta M$. Whenever $x<(1-\eta)\delta$, it is the case that

$$\bar{F}_Q(x) = |\mathfrak{P}| + \sum_{t\in\mathfrak{M}} I(\eta M_t > x).$$

Since $|\mathfrak{P}|<\infty$ and $\alpha\in(0,1)$, deduce that $\lim_{x\to\infty}x^\alpha\bar{F}_Q(x)=\lim_{x\to 0}x^\alpha\bar{F}_M(x/\eta)=L$ by construction. Hence $Q\in\Sigma(\alpha,L)$ and

$$\|P-Q\|_{\mathrm{TV}}=\eta.$$

Since η was taken arbitrarily small, this concludes the proof.

Lemma 3. *For all $L > 0$ and all $\alpha \in (0,1)$ there exists a probability distribution P on $\mathbb{N}$ such that $\lim_{x\to 0} x^\alpha \bar{F}_P(x) = L$.*

Proof. We will prove the lemma by constructing such a distribution. We take $c > 0$ arbitrary for now and $b \geq 1$ integer also arbitrary for now; their value will be chosen accordingly at the end of the day. We consider any $Q \equiv (Q_1, Q_2, \dots)$ [we use the convention $\mathbb{P}_Q(X = i) = Q_i$] such that $|\{i \in \mathbb{N} : Q_i = c2^{-k/\alpha}\}| = b2^k$ for all $k \geq 0$ integer. We pick the value of $c > 0$ (as a function of b) such that Q is a proper probability distribution. We see immediately that it must be that [remark that $b2^k$ is integer]

$$1 = \sum_{i\geq 1} Q_i = \sum_{k\geq 0} c2^{-k/\alpha} |\{i \in \mathbb{N} : Q_i = c2^{-k/\alpha}\}| = cb \sum_{k\geq 0} 2^{-k(1/\alpha - 1)} = cb \cdot \frac{2^{1/\alpha}}{2^{1/\alpha} - 2}.$$

This establishes that we must have

$$c = \frac{2^{1/\alpha} - 2}{b2^{1/\alpha}}.$$

On the other hand we have,

$$\begin{aligned}
\bar{F}_Q(x) &= \sum_{i\geq 1} I(Q_i > x) \\
&= \sum_{k\geq 0} |\{i \in \mathbb{N} : Q_i = c2^{-k/\alpha}\}| I(c2^{-k/\alpha} > x) \\
&= b \sum_{k\geq 0} 2^k I(2^k < (x/c)^{-\alpha}) \\
&= b(2^{k_*+1} - 1)
\end{aligned}$$

where k_* is the only integer satisfying $(x/c)^{-\alpha} \leq 2^{k_*} \leq (x/c)^{-\alpha} + 1$. It follows immediately that

$$\lim_{x\to 0} x^\alpha \bar{F}_Q(x) = 2bc^\alpha = b^{1-\alpha}\big(2^{1/\alpha} - 2\big)^\alpha$$

Therefore if we choose [recall that b must be an integer!]

$$b = \left\lceil \Big(\frac{L}{(2^{1/\alpha} - 2)^\alpha} \Big)^{1/(1-\alpha)} \right\rceil.$$

we have that $\lim_{x\to 0} x^\alpha \bar{F}_Q(x) = L'$ for a $L' \geq L$. We can then obtain the desired P by taking $P = (1-\omega)Q + \omega\delta_1$ for an appropriate choice of $\omega \in [0,1]$. Indeed for such P

$$\begin{aligned}
&\lim_{x\to 0} x^\alpha \bar{F}_P(x) \\
&\quad = \lim_{x\to 0} x^\alpha \left(\sum_{j\geq 1} I((1-\omega)Q_j > x) + I(\omega > x) \right) \\
&\quad = \lim_{x\to 0} x^\alpha \bar{F}_Q\Big(\frac{x}{1-\omega}\Big) \\
&\quad = (1-\omega)^\alpha L'.
\end{aligned}$$

So by taking $1 - \omega = (L/L')^{1/\alpha}$ (which is between 0 and 1) we are done. ∎

A.8 Proof of Theorem 3

A.8.1 Proof of the upper bound

The starting point of the proof is the equation (19) which is valid for all $P \in \mathcal{P}$ and all $n \geq 2$. The only thing we need to conclude and that differs from the Proposition 3 is a uniform control over $\mathbb{E}_P(M_{n,k})$ when $P \in \Sigma(\alpha, \beta, C)$. Such uniform control is given by the following Lemma.

Lemma 4. *Let $C_{n,k} \coloneqq \sum_{j \geq k} M_{n,j}$. Then, there exists universal constants $B > 0$ and $n_0 > 0$ such that for all $\alpha \in (0,1)$, all $\beta > 0$ all $C > 0$, all $n \geq n_0$ and all $0 \leq k \leq n$*

$$\sup_{P \in \Sigma(\alpha,\beta,C)} \left| \frac{\mathbb{E}_P(C_{n,k})}{L_\alpha(P)} - n^\alpha \frac{\Gamma(k-\alpha)}{\Gamma(k)} \right| \leq \frac{BC\Gamma(k-\alpha+\beta)n^{\alpha-\beta}}{\Gamma(k)}.$$

Proof. The proof is similar to Lemma E.3 in [18], we only sketch the main arguments. It can be established that (see Lemma E.2 in [18]) that

$$\mathbb{E}_P(C_{n,k}) = k\binom{n}{k}\int_0^1 \bar{F}_P(x)x^{k-1}(1-x)^{n-k}\mathrm{d}x$$

and by straightforward computations that

$$\int_0^1 x^{-\alpha} \cdot x^{k-1}(1-x)^{n-k}\mathrm{d}x = \frac{(n-k)!\Gamma(k-\alpha)}{\Gamma(n-\alpha+1)}.$$

Therefore,

$$\begin{aligned}
&\left|\mathbb{E}_P(C_{n,k}) - L_\alpha(P)\frac{\Gamma(k-\alpha)}{\Gamma(k)}\frac{\Gamma(n+1)}{\Gamma(n-\alpha+1)}\right| \\
&\quad \leq k\binom{n}{k}\int_0^1 |\bar{F}_P(x) - L_\alpha(P)x^{-\alpha}|x^{k-1}(1-x)^{n-k}\mathrm{d}x \\
&\quad \leq CL_\alpha(P)k\binom{n}{k}\int_0^1 x^{k+\beta-\alpha-1}(1-x)^{n-k}\mathrm{d}x \\
&\quad = CL_\alpha(P)\frac{\Gamma(k-\alpha+\beta)}{\Gamma(k)}\frac{\Gamma(n+1)}{\Gamma(n-\alpha+\beta+1)}
\end{aligned}$$

The conclusion of the proof follows from standard estimates on the ratio of Gamma functions, see for instance [18, Lemma E.3] or [41]).

□

A.8.2 Proof of the lower bound

We use the traditional approach that the minimax risk is always larger than the Bayes risk relative to any choice of prior. We consider a Pitman-Yor process [23, Section 14.4] with parameters (α, d, G) as prior distribution, denoted in the sequel $\mathrm{PY}_{\alpha,d,G}$. We will chose $d > -\alpha$ accordingly at the end of the proof and G is any nonatomic probability measure. Then, for any estimator $\hat{\theta}_r$, by letting $S \coloneqq \{P : \infty > L_\alpha(P) \geq B^2\}$ for a constant $B > 0$ to be chosen accordingly

$$\begin{aligned}
&\sup_{P \in \Sigma(\alpha,\beta,C)} \mathbb{P}_P\Big(\sqrt{n^\alpha L_\alpha(P)}\ell(\hat{\theta}_r, \theta_r) > AB\Big) \\
&\quad \geq \int_{\Sigma(\alpha,\beta,C)\bigcap S} \mathbb{P}_P\Big(\sqrt{n^\alpha L_\alpha(P)}\ell(\hat{\theta}_r, \theta_r) > AB\Big)\mathrm{PY}_{\alpha,d,G}(\mathrm{d}P) \\
&\quad \geq \int_{\mathcal{P}} \mathbb{P}_P\Big(n^{\alpha/2}\ell(\hat{\theta}_r, \theta_r) > A\Big)\mathrm{PY}_{\alpha,d,G}(\mathrm{d}P) - \mathrm{PY}_{\alpha,d,G}\big(S^c\big) - \mathrm{PY}_{\alpha,d,G}\big(\Sigma(\alpha,\beta,C)^c\big) \qquad (22)
\end{aligned}$$

The rest of this section is dedicated to bounding each of the term involved in the last rhs.

Control of the first term of (22) The first term is bounded exactly as in Section A.3 by setting $\varepsilon = An^{-\alpha/2}$ all along the proof. The difference is only that we don't take $\alpha = 0$ in (17). Instead, by Cauchy-Schwarz's inequality, it is true that $\mathbb{E}_\Pi(\sqrt{d + \alpha K_n}) \leq \sqrt{d + \alpha\mathbb{E}_\Pi(K_n)}$, ant it is well known that $\mathbb{E}_\Pi(K_n)$ is a multiple constant (depending solely on α and d) times n^α; see

for instance [37, Section 3.3]. Consequently, there is a constant $C(\alpha, d) > 0$ such that for all estimators $\hat{\theta}_r$

$$\int_{\mathcal{P}} \mathbb{P}_P\Big(\ell(\hat{\theta}_r(\mathbf{X}_n), \theta_r(P; \mathbf{X}_n)) > \varepsilon\Big) \mathrm{PY}_{\alpha,d,G}(\mathrm{d}P) \geq 1 - C(\alpha, d) n^{\alpha/2} \varepsilon'. \tag{23}$$

Recalling that $\varepsilon' = \frac{\varepsilon/r}{1-\varepsilon}$, the conclusion follows.

Control of the two last terms of (22) Here we bound the terms $\mathrm{PY}_{\alpha,d,G}(S^c)$ and $\mathrm{PY}_{\alpha,d,G}(\Sigma(\alpha, \beta, C)^c)$. To do so, we recall the following helpful construction of the Pitman-Yor process from [23, Example 14.47]. Let $(\Omega, \mathcal{A}, \mathbb{P}_\alpha)$ be a probability space on which is defined a Poisson process with with mean intensity measure $\rho_\alpha(\mathrm{d}x) = \frac{\alpha x^{-\alpha-1} \mathrm{d}x}{\Gamma(1-\alpha)}$ [the *so-called* stable process] and a collection $(S_i)_{i \geq 1}$ of iid random variables with marginal distribution G. Also let $J_1 \geq J_2 \geq \ldots$ the ordered jumps of the Poisson process and $T \coloneqq \sum_{i \geq 1} J_i$. Also let $g_{\alpha,d}(t) \coloneqq \frac{\Gamma(d+1)}{\Gamma(d/\alpha+1)} t^{-d} f_\alpha(t)$ where f_α is the density of a stable distribution, *i.e.* $\int_0^\infty e^{-\lambda t} f_\alpha(t) \mathrm{d}t = e^{-\lambda^\alpha}$ for all $\lambda \geq 0$. Then, for any $d > -\alpha$

$$\mathrm{PY}_{\alpha,d,G}(E) = \int_0^\infty \mathbb{P}_\alpha\Big(\sum_{i \geq 1} \frac{J_i}{T} \delta_{S_i} \in E \mid T = t\Big) g_{\alpha,d}(t) \mathrm{d}t.$$

Lemma 5. *For all $\alpha \in (0,1)$, for all $d > -\alpha$, and for all $\varepsilon > 0$, there exists $B > 0$ such that $\mathrm{PY}_{\alpha,d,G}(S^c) \leq \varepsilon$.*

Proof. Using the construction of the Pitman-Yor process described above, we see that

$$\begin{aligned}
\mathrm{PY}_{\alpha,d,G}(S^c) &= \int_0^\infty \mathbb{P}_\alpha\Big(\lim_{x \to 0} x^\alpha \bar{F}_P(x) \in [0, B) \cup \{+\infty\} \mid T = t\Big) g_{\alpha,d}(t) \mathrm{d}t \\
&= \int_0^\infty \mathbb{P}_\alpha\Big(\lim_{x \to 0} x^\alpha \sum_{i \geq 1} I(J_i > Tx) \in [0, B) \cup \{+\infty\} \mid T = t\Big) g_{\alpha,d}(t) \mathrm{d}t \\
&= \int_0^\infty \mathbb{P}_\alpha\Big(T^{-\alpha} \lim_{y \to 0} y^\alpha \sum_{i \geq 1} I(J_i > y) \in [0, B) \cup \{+\infty\} \mid T = t\Big) g_{\alpha,d}(t) \mathrm{d}t.
\end{aligned}$$

Now it is a well-known fact that $\lim_{y \to 0} y^\alpha \sum_{j \geq 1} I(J_j > y) = \frac{1}{\Gamma(1-\alpha)}$ $\mathbb{P}_\alpha$-almost-surely. Consequently,

$$\mathrm{PY}_{\alpha,d,G}(S^c) = \int_{1/[B\Gamma(1-\alpha)]^{1/\alpha}}^\infty g_{\alpha,d}(t) \mathrm{d}t$$

which goes to zero as $B \to 0$. ∎

Lemma 6. *For all $\alpha \in (0,1)$, for all $d > -\alpha$, for all $0 < \beta < \frac{\alpha}{2}$, and for all $\varepsilon > 0$ there exists $n_0 > 0$ and $C > 0$ such that for all $n \geq n_0$*

$$\mathrm{PY}_{\alpha,d,G}\big(\Sigma(\alpha, \beta, C)^c\big) \leq \varepsilon.$$

Proof. It is convenient to first reduce the problem to controlling expectations under the (unconditional) α-stable process rather than mixture of conditioned processes. To do so, we remark that

by Cauchy-Schwarz' inequality, for any event E, and whenever $d > -\alpha/2$

$$\begin{aligned}
\mathrm{PY}_{\alpha,d,G}(E) &= \int_0^\infty \mathbb{P}_\alpha\Big(\sum_{i\geq 1}\frac{J_i}{T}\delta_{S_i} \in E \mid T = t\Big)\frac{\Gamma(d+1)}{\Gamma(d/\alpha+1)}t^{-d}f_\alpha(t)\mathrm{d}t \\
&\leq \frac{\Gamma(d+1)}{\Gamma(d/\alpha+1)}\bigg(\int_0^\infty t^{-2d}f_\alpha(t)\mathrm{d}t\bigg)^{1/2}\bigg(\int_0^\infty \mathbb{P}_\alpha\Big(\sum_{i\geq 1}\frac{J_i}{T}\delta_{S_i} \in E \mid T = t\Big)^2 f_\alpha(t)\mathrm{d}t\bigg)^{1/2} \\
&\leq \frac{\Gamma(d+1)}{\Gamma(d/\alpha+1)}\bigg(\int_0^\infty t^{-2d}f_\alpha(t)\mathrm{d}t\bigg)^{1/2}\mathrm{PY}_{\alpha,0,G}(E)^{1/2} \\
&= \frac{\Gamma(d+1)}{\Gamma(d/\alpha+1)}\sqrt{\frac{\Gamma(2d/\alpha+1)}{\Gamma(2d+1)}}\sqrt{\mathrm{PY}_{\alpha,0,G}(E)}
\end{aligned}$$

Then it is enough to prove that $\mathrm{PY}_{\alpha,0,G}\big(\Sigma(\alpha,\beta,C)^c\big) \leq \eta$ for $\eta > 0$ as small as needed. Observe that for $d = 0$ we do have

$$\mathrm{PY}_{\alpha,0,G}(E) = \mathbb{P}_\alpha\Big(\sum_{i\geq 1}\frac{J_i}{Y}\delta_{S_i} \in E\Big).$$

With $P = \sum_{i\geq 1}\frac{J_i}{T}\delta_{S_i}$ and the same arguments as in the proof of Lemma 5, we have that $\bar{F}_P(x) = \sum_{i\geq 1} I(J_i > Tx)$ and $L_\alpha(P) = \frac{T^{-\alpha}}{\Gamma(1-\alpha)}$ almost-surely under $\mathbb{P}_\alpha$. We deduce that under $\mathbb{P}_\alpha$

$$\begin{aligned}
\sup_{x\in(0,1)} x^{-\beta}\Big|\frac{\bar{F}_P(x)}{L_\alpha(P)x^{-\alpha}} - 1\Big| &\overset{d}{=} T^\beta\Gamma(1-\alpha)\sup_{x\in(0,1)}(Tx)^{-\beta+\alpha}\Big|\sum_{i\geq 1} I(J_i > Tx) - \frac{(Tx)^{-\alpha}}{\Gamma(1-\alpha)}\Big| \\
&= T^\beta\Gamma(1-\alpha)\sup_{x\in(0,T)} x^{-\beta+\alpha}\Big|\sum_{i\geq 1} I(J_i > x) - \frac{x^{-\alpha}}{\Gamma(1-\alpha)}\Big|.
\end{aligned}$$

By a famous result of [19], The process $\{J_1, J_2, \ldots\}$ is equal in law to the process $\{\bar{\rho}^{-1}(\Gamma_1), \bar{\rho}^{-1}(\Gamma_2), \ldots\}$ where $\bar{\rho}^{-1}(x) \coloneqq \big(\Gamma(1-\alpha)y\big)^{-1/\alpha}$ and $\{\Gamma_1, \Gamma_2, \ldots\}$ are the jumps of a standard homogeneous Poisson process on the half real-line. Deduce that

$$\begin{aligned}
\sup_{x\in(0,1)} x^{-\beta}\Big|\frac{\bar{F}_P(x)}{L_\alpha(P)x^{-\alpha}} - 1\Big| &\overset{d}{=} T^\beta\Gamma(1-\alpha)^{\beta/\alpha}\sup_{x\in(\bar{\rho}(T),\infty)} x^{-1+\beta/\alpha}\Big|\sum_{j\geq 1} I(\Gamma_j \leq x) - x\Big| \\
&\leq T^\beta\Gamma(1-\alpha)^{\beta/\alpha}\sup_{x\in(0,\infty)} x^{-1+\beta/\alpha}\Big|\sum_{j\geq 1} I(\Gamma_j \leq x) - x\Big|.
\end{aligned}$$

Since T follows an α-stable distribution under $\mathbb{P}_\alpha$, for all $\varepsilon > 0$ we can find $M > 0$ such that $T^\beta \leq M$ with probability $\geq 1 - \varepsilon$. Then the result follows from Lemma 7 below. □

Lemma 7. *Let $\{\Gamma_1, \Gamma_2, \ldots\}$ be the jumps of a standard homogeneous Poisson process on the half real-line and let $0 < \delta < 1/2$. Then for all $\varepsilon > 0$ there exists a constant $B > 0$ such that with probability more than $1 - \varepsilon$*

$$\sup_{x\in(0,\infty)} x^{-(1/2+\delta)}\Big|\sum_{j\geq 1} I(\Gamma_j \leq x) - x\Big| \leq B.$$

Proof. We only sketch the proof as it is a trivial adaptation of the proof in [18, Lemma B.1].

Defining $\Gamma_0 = 0$

$$\begin{aligned}
&\sup_{x\in(0,\infty)} x^{-(1/2+\delta)}\Big|\sum_{j\geq 1} I(\Gamma_j \leq x) - x\Big| \\
&\quad= \sup_{k\geq 0}\sup_{x\in(\Gamma_k,\Gamma_{k+1}]} x^{-(1/2+\delta)}\Big|\sum_{j\geq 1} I(\Gamma_j \leq x) - x\Big| \\
&\quad= \sup_{k\geq 0}\sup_{x\in(\Gamma_k,\Gamma_{k+1}]} x^{-(1/2+\delta)}|k - x| \\
&\quad\leq \max\Big(\Gamma_1^{1/2-\delta},\ \sup_{k\geq 1}\Gamma_k^{-(1/2+\delta)}\max(|\Gamma_k - k|,\ |\Gamma_{k+1} - k|)\Big) \\
&\quad\leq \max\Big(\Gamma_1^{1/2-\delta},\ \sup_{k\geq 1}\Gamma_k^{-(1/2+\delta)}(|\Gamma_k - k| + \xi_{k+1})\Big).
\end{aligned}$$

That is,

$$\begin{aligned}
&\sup_{x\in(0,\infty)} x^{-(1/2+\delta)}\Big|\sum_{j\geq 1} I(\Gamma_j \leq x) - x\Big| \\
&\qquad\qquad\leq \max\left(\Gamma_1^{1/2-\delta},\ \sup_{k\geq 1}\Big(\frac{\Gamma_k}{k}\Big)^{-(1/2+\delta)}\cdot\sup_{k\geq 1}\frac{|\Gamma_k - k| + \xi_{k+1}}{k^{1/2+\delta}}\right).
\end{aligned}$$

Then it suffices to show that the last rhs is bounded with high probability. Details can be found in [18, Lemma B.1]. A quick heuristic (non rigorous though) show that this must be true since $\lim_k \frac{\Gamma_k}{k} = 1$ almost-surely as $k\to\infty$ by the law of large numbers and $\limsup_k \frac{|\Gamma_k - k|}{\sqrt{2k\log(\log(k))}} = 1$ by the law of iterated logarithm. □

A.9 Proof of Theorem 4

A.9.1 Proof of the upper bound

Remark that if $P\in\Sigma(\alpha,\beta,C)$ then $\theta_r(P;\mathbf{X}_n)\neq 0$ P-as. Consequently,

$$\begin{aligned}
\ell(\hat{T},T) &= \Big|\frac{\hat{\theta}_1^{(\mathrm{GT})}}{\hat{\theta}_2^{(\mathrm{GT})}}\frac{\theta_2}{\theta_1} - 1\Big| \\
&\leq \Big|\frac{\hat{\theta}_1^{(\mathrm{GT})}}{\theta_1} - 1\Big| + \frac{\hat{T}}{T}\Big|\frac{\hat{\theta}_2^{(\mathrm{GT})}}{\theta_2} - 1\Big| \\
&\leq \ell(\hat{\theta}_1^{(\mathrm{GT})},\theta_1) + \ell(\hat{\theta}_2^{(\mathrm{GT})},\theta_2)\Big(1 + \ell(\hat{T},T)\Big).
\end{aligned}$$

Deduce that P-as,

$$\ell(\hat{T},T) \leq \frac{\ell(\hat{\theta}_1^{(\mathrm{GT})},\theta_1) + \ell(\hat{\theta}_2^{(\mathrm{GT})},\theta_2)}{1 - \ell(\hat{\theta}_2^{(\mathrm{GT})},\theta_2)}.$$

From here, the conclusion follows immediately by Theorem 3.

A.9.2 Proof of the lower bound

Mimicking the proof of Theorem 3 up to equation (22), it is found that

$$\begin{aligned}
&\sup_{P\in\Sigma(\alpha,\beta,C)} \mathbb{P}_P\Big(\sqrt{n^\alpha L_\alpha(P)}\ell(\hat{T},T(P;\mathbf{X}_n)) > AB\Big) \\
&\quad\geq \int_{\mathcal{P}} \mathbb{P}_P\Big(n^{\alpha/2}\ell(\hat{T},T(P;\mathbf{X}_n)) > A\Big)\mathrm{PY}_{\alpha,d,G}(\mathrm{d}P) - \mathrm{PY}_{\alpha,d,G}\big(S^c\big) - \mathrm{PY}_{\alpha,d,G}\big(\Sigma(\alpha,\beta,C)^c\big).
\end{aligned}$$

The terms $\mathrm{PY}_{\alpha,d,G}(S^c)$ and $\mathrm{PY}_{\alpha,d,G}(\Sigma(\alpha,\beta,C)^c)$ have been taken care of in the proof of Theorem 3. In particular it has been demonstrated that they can be made arbitrarily small by choosing B and C accordingly. To deal with the first term, we mimick the proof of Theorem 1 up to equation (15) to find that for all $\varepsilon > 0$

$$\int_{\mathcal{P}} \mathbb{P}_P\Big(\ell(\hat{T}, T(P;\mathbf{X}_n)) > \varepsilon\Big)\mathrm{PY}_{\alpha,d,G}(\mathrm{d}P) \geq \mathbb{E}_\Pi\Big[\inf_{t\in\mathbb{R}_+} \mathbb{E}_\Pi\big(I(\ell(t, T(P;\mathbf{X}_n)) > \varepsilon) \mid \mathbf{X}_n\big)\Big].$$

Here the proof needs some modification compared to Theorem 1. But, using Lemma 1, conditional on $\mathbf{X}_n$ it is the case that $(\theta_1(P;\mathbf{X}_n), \theta_2(P;\mathbf{X}_n))$ has the law of $(W_0, W_0^2\sum_{j\geq 1} Q_j^2)$ with $W_0 \mid \mathbf{X}_n \sim \mathrm{Beta}(d+\alpha K_n, d+n)$ is independent of $Q \mid \mathbf{X}_n \sim \mathrm{PY}_{\alpha,d+\alpha K_n}$. Consequently, $T(P;\mathbf{X}_n)$ is equal in law to $\frac{1}{W_0\sum_{j\geq 1} Q_j^2}$. In particular $T(P;\mathbf{X}_n)$ is almost-surely non-zero, so $\ell(t, T(P;\mathbf{X}_n)) = \big|\frac{t}{T(P;\mathbf{X}_n)} - 1\big|$ almost-surely too. Letting $Z = \sum_{j\geq 1} Q_j^2$, and following the steps up to equation (16), it is found that for all $t \in \mathbb{R}_+$

$$\mathbb{E}_\Pi\big(I(\ell(t, T(P;\mathbf{X}_n)) > \varepsilon\})\mid \mathbf{X}_n\big) \geq \inf_{z\in\mathbb{R}_+} \mathbb{E}_\Pi\big(I(\{|tzW_0 - 1| > \varepsilon\})\mid \mathbf{X}_n\big).$$

Therefore,

$$\begin{aligned}
&\int_{\mathcal{P}} \mathbb{P}_P\Big(\ell(\hat{T}, T(P;\mathbf{X}_n)) > \varepsilon\Big)\mathrm{PY}_{\alpha,d,G}(\mathrm{d}P) \\
&\quad\geq \mathbb{E}_\Pi\Big[\inf_{t\in\mathbb{R}_+} \mathbb{E}_\Pi\Big(I(\{|tW_0 - 1| > \varepsilon\})\mid \mathbf{X}_n\Big)\Big] \\
&\quad\geq 1 - C(\alpha, d)n^{\alpha/2}\varepsilon
\end{aligned}$$

where the last line follows by the exact same steps that lead us to (17) in the proof of Theorem 1, and then the steps that lead us to (23) in the proof of Theorem 3.

Acknowledgements

The authors are grateful to Giulia Cereda and Richard D. Gill for stimulating conversations on the rare type match problem and its interplay with the estimation of the missing mass. Stefano Favaro and Zacharie Naulet received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 817257. Stefano Favaro gratefully acknowledge the financial support from the Italian Ministry of Education, University and Research (MIUR), "Dipartimenti di Eccellenza" grant 2018-2022.

References

[1] Acharya, J., Bao, Y., Kang, Y. and Sun, Z. (2018). Improved bounds on minimax risk of estimating missing mass. In *IEEE International Symposium on Information Theory.*

[2] Ayed, F., Battiston, M., Camerlenghi, F. and Favaro, S. (2018). On consistent and rate optimal estimation of the missing mass. *Annales de l'Institut Henri Poincaré - Probabilités et Statistques* **57**, 1476–1494.

[3] Ben-Hamou, A., Boucheron, S. and Gassiat, E. (2018). Pattern coding meets censoring: (almost) adaptive coding on countable alphabets. *Preprint arXiv:1608.08367*

[4] Ben-Hamou, A., Boucheron, S. and Ohannessian, M.I. (2017). Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications. *Bernoulli* **23**, 249–287.

[5] Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration inequalities: a nonasymptotic theory of independence*. Oxford University Press.

[6] Bingham, N.H., Goldie, C.M. and Teugels, J.L. (1987). *Regular Variation.* Cambridge University Press.

[7] Brenner, C.H. (2010). Fundamental problem of forensic mathematics - the evidential value of a rare haplotype. *Forensic Science International* **4**, 281–291

[8] Bubeck, S., Ernst, D., and Garivier, A. (2013). Optimal discovery with probabilistic expert advice: finite time analysis and macroscopic optimality. *Journal of Machine Learning Research* **14**, 601–623.

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

[10] Cereda, G. (2017) Bayesian approach to LR assessment in case of rare type match. *Statistica Neerlandica* **71**, 141–164.

[11] Cereda, G. (2017) Impact of model choice on LR assessment in case of rare haplotype match (frequentist approach). *Scandinavian Journal of Statistics* **44**, 230–248.

[12] Cereda, G. and Gill, R.D. (2020) A nonparametric Bayesian approach to the rare type match problem. *Entropy* **22**, 439.

[13] Daley, T. and Smith, A.D. (2013). Predicting the molecular complexity of sequencing libraries. *Nature Methods* **10**, 325–327.

[14] Daley, T. and Smith, A.D. (2014). Modeling genome coverage in single-cell sequencing. *Bioinformatics* **30**, 3159–3165.

[15] Deng, C. Daley, T., De Sena Brandine, G. and Smith, A.D. (2019). Molecular heterogeneity in large-scale biological data: techniques and applications. *Annual Review of Biomedical Data Science* **2**, 39–67.

[16] Esty, W.W. (1982). Confidence intervals for the coverage of low coverage samples. *The Annals of Statistics* **10**, 190–196.

[17] Esty, W.W. (1983). A normal limit law for a nonparametric estimator of the coverage of a random sample. *The Annals of Statistics* **11**, 905–912.

[18] Favaro, S. and Naulet, Z. (2022). Near-optimal estimation of the unseen under regularly varying tail populations. *Preprint arXiv:2104.03251.*

[19] Ferguson, T.S. and Klass, M.J. (1972). A representation of independent increment processes without Gaussian components. *The Annals of Mathematical Statistics* **43**, 1634–1643.

[20] Gale, W.A. and Sampson, G. (1995). Good-Turing frequency estimation without tears. *Journal of Quantitative Linguistics* **2**, 217–237.

[21] Gao, F. (2013). Moderate deviations for a nonparametric estimator of sample coverage. *The Annals of Statistics* **41**, 641–669.

[22] Gao, Z., Tseng, C.H., Pei, Z. an Blaser, M.J. (2007). Molecular analysis of human forearm superficial skin bacterial biota. *Proceedings of the National Academy of Sciences of USA* **104**, 2927–2932.

[23] Ghosal, S. and Van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics.

[24] Gnedin, A., Hansen, B. and Pitman, J. (2007). Notes on the occupancy problems with infinitely many boxes: general asymptotics and power law. *Probab. Surv.* **4**, 146–171.

[25] Good, I.J.(1953). The population frequencies of species and the estimation of population parameters. *Biometrika* **40**, 237–264.

[26] Ionita-Laza, I., Lange, C. and Laird, N.M. (2009). Estimating the number of unseen variants in the human genome. *Proceedings of the National Academy of Sciences of USA* **106**, 5008–5013.

[27] Karlin, S. (1967). Central limit theorems for certain infinite urn schemes. *Journal of Mathematics and Mechanics* **17**, 373–401

[28] Mao, C.X. and Lindsay, B.G. (2004). A Poisson model for the coverage problem with a genomic application. *Biometrika* **89**, 669–682.

[29] McAllester, D. and Ortiz, L. (2003). Concentration inequalities for the missing mass and for histogram rule error. *Journal of Machine Learning Research* **4**, 895–911.

[30] McAllester, D. and Schapire, R.E. (2000). On the convergence rate of Good-Turing estimators. *Proceedings of the Conference on Computational Learning Theory.*

[31] Mossel, E. and Ohannessian, M.I. (2019). On the impossibility of learning the missing mass. *Entropy* **21**, 28.

[32] Motwani, S. and Vassilvitskii, S. (2006) Distinct value estimators in power law distributions. In *Proceedings of the Workshop on Analytic Algorithms and Combinatorics.*

[33] Ohannessian, M.I. and Dahleh, M.A. (2012). Rare probability estimation under regularly varying heavy tails. In *Proceedings of the Conference on Learning Theory.*

[34] Orlitsky, A., Santhanam, N.P. and Zhang, J. (2003). Always Good-Turing: asymptotically optimal probability estimation. *Science* **302**, 427–431.

[35] Orlitsky, A., Santhanam, N.P. and Zhang, J. (2004). Universal compression of memoryless sources over unknown alphabets. *IEEE Transaction on Information Theory* **50**, 1469–1481.

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

[37] Pitman, J. (2006). *Combinatorial Stochastic Processes: Ecole d'eté de probabilités de Saint-Flour xxxii-2002.* Springer.

[38] Robbins, H.E. (1956). An empirical Bayes approach to statistics. *Proceedings of the Berkeley Symposium* **1**, 157–163.

[39] Robbins, H.E. (1964). The empirical Bayes approach to statistical decision problems. *The Annals of Mathematical Statistics* **35**, 1–20.

[40] Robbins, H.E. (1968). Estimating the total probability of the unobserved outcomes of an experiment. *The Annals of Mathematical Statistics* **39**, 256–257.

[41] Temme, N. M. (1996). *Special functions: An introduction to the classical functions of mathematical physics.* John Wiley & Sons.

[42] Zhang, C.H. and Zhang, Z. (2009). Asymptotic normality of a nonparametric estimator of sample coverage. *The Annals of Statistics* **37**, 2582–2595.