On consistent and rate optimal estimation of the missing mass

Submitted to the Annales de l'Institut Henri Poincaré - Probabilités et Statistiques

On consistent and rate optimal estimation of the missing mass

Fadhel Ayed[a], Marco Battiston[b], Federico Camerlenghi[c,*] and Stefano Favaro[d,†]

[a] *Department of Statistics, University of Oxford*
OX1 3LB, Oxford, United Kingdom.
E-mail: fadhel.ayed@gmail.com

[b] *Department of Mathematics and Statistics, University of Lancaster*
LA1 4YF, Lancaster, United Kingdom.
E-mail: m.battiston@lancaster.ac.uk

[c] *Department of Economics, Management and Statistics, University of Milano – Bicocca*
20126, Milano, Italy.
E-mail: federico.camerlenghi@unimib.it

[d] *Department of Economics and Statistics, University of Torino and Collegio Carlo Alberto*
10134, Torino, Italy.
E-mail: stefano.favaro@unito.it

Abstract. Given n samples from a population of individuals belonging to different types with unknown proportions, how do we estimate the probability of discovering a new type at the $(n+1)$-th draw? This is a classical problem in statistics, commonly referred to as the missing mass estimation problem. Recent results have shown: i) the impossibility of estimating the missing mass without imposing further assumptions on type's proportions; ii) the consistency of the Good-Turing estimator of the missing mass under the assumption that the tail of type's proportions decays to zero as a regularly varying function with parameter $\alpha \in (0,1)$; ii) the rate of convergence $n^{-\alpha/2}$ for the Good-Turing estimator under the class of $\alpha \in (0,1)$ regularly varying P. In this paper we introduce an alternative, and remarkably shorter, proof of the impossibility of a distribution-free estimation of the missing mass. Beside being of independent interest, our alternative proof suggests a natural approach to strengthen, and expand, the recent results on the rate of convergence of the Good-Turing estimator under $\alpha \in (0,1)$ regularly varying type's proportions. In particular, we show that the convergence rate $n^{-\alpha/2}$ is the best rate that *any* estimator can achieve, up to a slowly varying function. Furthermore, we prove that a lower bound to the minimax estimation risk must scale at least as $n^{-\alpha/2}$, which leads to conjecture that the Good-Turing estimator is a rate optimal minimax estimator under regularly varying type proportions.

MSC 2010 subject classifications: 62G05, 62C20.
Keywords: Good-Turing estimator, Minimax rate, Missing mass, Optimal rate of convergence, Regular variation, Two-parameter Poisson-Dirichlet.

*Also affiliated to Collegio Carlo Alberto, Torino and BIDSA, Bocconi University, Milano.
†Also affiliated to IMATI-CNR "Enrico Magenes" (Milan, Italy).

1. Introduction

Given n samples from a population of individuals belonging to different types with unknown proportions, how do we estimate the probability of discovering a new type at the $(n+1)$-th draw? This is a classical problem in statistics, referred to as the missing mass estimation problem. It first appeared in ecology (e.g., Fisher et al. [9] and Good [15]), and its importance has grown in recent years driven by applications in biological and physical sciences (e.g., Kroes et al. [20], Gao et al. [11] and Ionita-Laza et al. [17]), machine learning and theoretical computer science (e.g., Motwani and Vassilvitskii [25] and Bubeck et al. [5]), and information theory (e.g., Orlitsky et al. [29] and Ben-Hamou et al. [3]). To introduce the missing mass, let $P = \sum_{j\geq 1} p_j \delta_{\theta_j}$ be an unknown discrete distribution, where $(\theta_j)_{j\geq 1}$ are atoms on a measurable space and $(p_j)_{j\geq 1}$ are the corresponding probability masses, i.e. $p_j \in [0,1]$ such that $\sum_{j\geq 1} p_j = 1$. If $\mathbf{X}_n = (X_1, \ldots, X_n)$ is a collection of independent and identically distributed random variables from P, then we define the missing mass as

$$M_n(P, \mathbf{X}_n) = \sum_{j\geq 1} p_j \mathbb{1}(\theta_j \notin \mathbf{X}_n), \tag{1.1}$$

where $\mathbb{1}(\cdot)$ is the indicator function. The Good-Turing estimator (Good [15]) is arguably the most popular nonparametric estimator of $M_n(P, \mathbf{X}_n)$. It has been the subject of numerous studies. These include, e.g., asymptotic normality and large deviations (Zhang and Zhang [38], Gao [10] and Grabchak and Zhang [16]), admissibility and concentration properties (McAllester and Ortiz, [21], Ohannessian and Dahleh [27] and Ben-Hamou et al. [2]), consistency and convergence rates (McAllester Schapire [22], Wagner et al. [37] and Mossel and Ohannessian [24]), optimality and minimax properties (Orlitsky et al. [28] and Rajaraman et al. [35]).

Let $\hat{M}_n(\mathbf{X}_n)$ denote an estimator of the missing mass $M_n(P, \mathbf{X}_n)$. Motivated by the recent works of Ohannessian and Dahleh [27], Mossel and Ohannessian [24], Ben-Hamou et al. [2] and Grabchak and Zhang [16], in this paper we consider the problem of consistent estimation of the missing mass under the multiplicative loss function

$$L(\hat{M}_n(\mathbf{X}_n), M_n(P, \mathbf{X}_n)) = \left| \frac{\hat{M}_n(\mathbf{X}_n)}{M_n(P, \mathbf{X}_n)} - 1 \right|. \tag{1.2}$$

As discussed in Ohannessian and Dahleh [27], the loss function (1.2) is adequate for estimating small value parameters, since it allows to achieve more informative results. The multiplicative loss function (1.2) has been already used in statistics, for instance in the estimation of small value probabilities using importance sampling (Chatterjee and Diaconis [6]) and in the estimation of tail probabilities in extreme value theory (Beirlant and Devroye [1]). Under the loss function (1.2), Ohannessian and Dahleh [27] showed that: i) the Good-Turing estimator may be inconsistent; ii) the Good-Turing estimator is strongly consistent if the tail of P decays to zero as a regularly varying function with parameter $\alpha \in (0,1)$ (e.g., Bingham et al. [4]). In particular, Grabchak and Zhang [16] showed that the convergence rate of the Good-Turing estimator is $n^{-\alpha/2}$, up to a slowly varying function. Mossel and Ohannessian [24] strengthened the inconsistency result of Ohannessian and Dahleh [27], showing the impossibility of estimating $M_n(P, \mathbf{X}_n)$ in a completely distribution-free fashion, that is without imposing further structural assumptions on P. In this paper we strengthen, and expand, the result of Grabchak and Zhang [16] on the rate of convergence of the Good-Turing estimator.

We start by introducing an alternative, and remarkably shorter, proof of the impossibility result of Mossel and Ohannessian [24]. Our proof relies on Bayesian nonparametric ideas, and in particular on the use a Ferguson-Dirichlet process prior (Ferguson [8]) for the unknown distribution P. The prior assumption on P allows us to avoid the winding constructive arguments of Mossel and Ohannessian [24], and then prove the impossibility of estimating $M_n(P, \mathbf{X}_n)$ by exploiting well-known properties of the posterior distribution of $M_n(P, \mathbf{X}_n)$. Beside being of independent interest, our alternative proof suggests a natural approach to study rates of convergence of the Good-Turing estimator under the class of $\alpha \in (0,1)$ regularly varying P. In particular, we make use of the two-parameter Poisson-Dirichlet process prior (Pitman and Yor [34]) for the

unknown distribution P, which is known to generate (almost surely) discrete distributions whose tails decay to zero as a regularly varying function with parameter $\alpha \in (0,1)$. See, e.g., Gnedin et al. [13] and references therein. Under this prior assumption on P, we exploit properties of the posterior distribution of $M_n(P, \mathbf{X}_n)$ to prove that $n^{-\alpha/2}$ is the best rate of convergence that any estimator of the missing mass $M_n(P, \mathbf{X}_n)$ can achieve, up to a slowly varying function. Our result strengthen the result of Grabchak and Zhang [16], showing the optimality, up to a slowly varying function, of the rate of convergence of the Good-Turing estimator. Finally, still relying on the two-parameter Poisson-Dirichlet process prior for P, we first study minimax rates for the Good-Turing estimator under the class of $\alpha \in (0,1)$ regularly varying P. In particular, we show that the minimax rate for estimating $M_n(P, \mathbf{X}_n)$ must be at least $n^{-\alpha/2}$, and we conjecture that the Good-Turing estimator is an asymptotically optimal minimax estimator under the class of $\alpha \in (0,1)$ regularly varying P.

The paper is structured as follows. In Section 2 we present the alternative proof of the main result of Mossel and Ohannessian [24]. In Section 3 we state and prove our main results on the rate of convergence and on the minimax rate of the Good-Turing estimator under $\alpha \in (0,1)$ regularly varying P. Section 4 contains a discussion of our results, and some open problems on the estimation of missing mass and generalizations thereof. Auxiliary results and technical lemmas are deferred to Appendix A. The following notation is adopted throughout the paper: $[0,1]$ is the unit interval, and $\mathcal{B}([0,1])$ denotes its Borel σ-algebra; $\mathcal{P}$ is the space of discrete distributions on $[0,1]$, endowed with the smallest σ-algebra making $P \mapsto P(A)$ measurable for every $A \in \mathcal{B}([0,1])$; P^n is the n-fold product of P on $[0,1]$, and $\mathbb{E}_P$ denotes the expectation with respect P; for ease of notation, we also use $\mathbb{E}_P$ to denote the expectation with respect to P^n; ℓ is a slowly varying function, i.e. a function satisfying $\ell(xc)/\ell(x) \to 1$ as $x \to \infty$ for every $c > 0$; C denotes a generic strictly positive constant; given a sequence of probabilities $(p_j)_{j \geq 1}$, we set $(p_{[j]})_{j \geq 1}$ to be its corresponding ordered sequence of probabilities, i.e. $p_{[1]} \geq p_{[2]} \geq \ldots$; given two functions f and g, $f \sim g$ stands for $\lim f/g = 1$, $f = \mathcal{O}(g)$ for $\limsup |f|/|g| < C$, $f = o(g)$ for $\lim f/g = 0$; $\mathcal{B}(a,b)$ denotes the Beta integral of parameters $a, b > 0$.

2. Impossibility of estimating $M_n(P, \mathbf{X}_n)$

Let $\mathbf{X}_n = (X_1, \ldots, X_n)$ be a collection of independent and identically distributed random variables from an unknown discrete distribution P. The actual values taken by the X_i's is not relevant for the missing mass estimation problem. In particular, without loss of generality, we assume that the X_i's takes values in the set $[0,1]$. Therefore, $P(\cdot) = \sum_j p_j \delta_{\theta_j}(\cdot)$ is assumed to be an unknown discrete distribution on the sample space $[0,1]$, given a sequence of atoms $\theta_j \in [0,1]$ and masses $p_j < 1$ such that $\sum_{j \geq 1} p_j = 1$. Both atoms and masses of the distribution P are supposed to be unknown. Given the sample $\mathbf{X}_n$, we are interested in estimating the missing $M_n(P, \mathbf{X}_n)$ defined in (1.1). The function $M_n(P, \mathbf{X}_n)$ is a jointly measurable function of P and $\mathbf{X}_n$, as proved in Proposition A.1. Given an estimator $\hat{M}_n(\mathbf{X}_n) : [0,1]^n \to [0,1]$ of $M_n(P, \mathbf{X}_n)$, we will measure its statistical performance by using the multiplicative loss function defined in (1.2). As we discussed in the introduction, this loss function is suitable to study theoretical properties of parameters or functionals taking small values, and it has already been used in previous works on missing mass estimation, e.g., Ohannessian and Dahleh [27], Mossel and Ohannessian [24], Ben-Hamou et al. [2] and Grabchak and Zhang [16].

A sequence of estimators $\hat{M}_n(\mathbf{X}_n)$ is said to be consistent for $M_n(P, \mathbf{X}_n)$, under parameter space $\mathcal{P}$ and the loss function L, if the loss incurred by the estimator converges in probability to zero under all points in the parameter space. Formally, $\hat{M}_n(\mathbf{X}_n)$ is consistent for $M_n(P, \mathbf{X}_n)$ if for all $P \in \mathcal{P}$ and for all $\epsilon > 0$ it holds true

$$P^n(L(\hat{M}_n(\mathbf{X}_n), M_n(P, \mathbf{X}_n)) > \epsilon) \to 0 \tag{2.1}$$

as $n \to \infty$. Furthermore, the estimator $\hat{M}_n(\mathbf{X}_n)$ is strongly consistent if (2.1) is replaced by almost sure convergence. Under this setting, Mossel and Ohannessian [24] proved Theorem 2.1 below. In this section we present an alternative, and remarkably shorter, proof of Theorem 2.1. Our proof relies on Bayesian nonparametric ideas, and in particular on the use the Ferguson-Dirichlet process prior for the unknown distribution P.

Theorem 2.1. *Let $\mathcal{P}$ be the set of all discrete distributions on $[0,1]$ and L be the loss function defined as (1.2). Then, there do not exist any consistent estimators for the missing mass $M_n(P, \boldsymbol{X}_n)$, i.e. there are no estimators $\hat{M}_n(\boldsymbol{X}_n)$ satisfying (2.1).*

Proof. We are going to show that for every estimator $\hat{M}_n(\mathbf{X}_n)$, there exists $\epsilon > 0$ such that

$$\sup_{P\in\mathcal{P}} \limsup_n P^n(L(\hat{M}_n(\mathbf{X}_n), M_n(P,\mathbf{X}_n)) > \epsilon) > 0 \tag{2.2}$$

and, therefore, there exists $P \in \mathcal{P}$ such that $L(\hat{M}_n(\mathbf{X}_n), M_n(P,\mathbf{X}_n))$ does not converge to zero in probability. First, note that for $\epsilon < 1/2$, Fact A.1 implies

$$P^n(L(\hat{M}_n(\mathbf{X}_n), M_n(P,\mathbf{X}_n)) > \epsilon) \geq P^n(L(M_n(P,\mathbf{X}_n), \hat{M}_n(\mathbf{X}_n)) > 2\epsilon),$$

hence it is sufficient to show that there exists $0 < \epsilon < 1$ such that for every estimator $\hat{M}_n(\mathbf{X}_n)$ the following is satisfied

$$\sup_{P\in\mathcal{P}} \limsup_n P^n(L(M_n(P,\mathbf{X}_n), \hat{M}_n(\mathbf{X}_n)) > \epsilon) > 0. \tag{2.3}$$

We prove that (2.3) holds for all $0 < \epsilon < 1/4$, and therefore (2.2) holds for any $0 < \epsilon < 1/8$, as well. Fix $\epsilon \in (0, 1/4)$ and let us denote by DP_γ the Ferguson-Dirichlet process on $\mathcal{P}$ (Ferguson [8]), with base measure γ on $[0,1]$. We choose γ uniform, i.e. $\gamma(d\theta) = \mathbb{1}(0 < d\theta < 1)$. We can now lower bound the supremum in (2.3) by averaging over $\mathcal{P}$ with respect to DP_γ and then use the Fubini theorem to get

$$\begin{aligned}
&\sup_{P\in\mathcal{P}} \limsup_n P^n(L(M_n(P,\mathbf{X}_n), \hat{M}_n(\mathbf{X}_n)) > \epsilon) \\
&\geq \int_{\mathcal{P}} \limsup_n P^n(L(M_n(P,\mathbf{X}_n), \hat{M}_n(\mathbf{X}_n)) > \epsilon)\mathrm{DP}_\gamma(dP) \\
&\geq \limsup_n \int_{\mathcal{P}} \int_{[0,1]^n} \mathbb{1}(L(M_n(P,\mathbf{X}_n), \hat{M}_n(\mathbf{X}_n)) > \epsilon) P^n(d\mathbf{X}_n)\mathrm{DP}_\gamma(dP) \\
&= \limsup_n \int_{[0,1]^n} \int_{\mathcal{P}} \mathbb{1}(L(M_n(P,\mathbf{X}_n), \hat{M}_n(\mathbf{X}_n)) > \epsilon)\mathrm{DP}_{\gamma+\sum_{i=1}^n \delta_{X_i}}(dP) P^n_{\mathrm{DP}_\gamma}(d\mathbf{X}_n) \\
&\geq \limsup_n \int_{[0,1]^n} \inf_{x\geq 0} \int_{\mathcal{P}} \mathbb{1}(L(M_n(P,\mathbf{X}_n), x) > \epsilon)\mathrm{DP}_{\gamma+\sum_{i=1}^n \delta_{X_i}}(dP) d\mathbf{X}_n,
\end{aligned}$$

where: i) the first inequality follows since we can lower bound the supremum by an average; ii) the second inequality follows from reverse Fatou's lemma; iii) the equality comes by swapping the marginal of P and conditional of $\mathbf{X}_n$ given P, on the one hand, with the marginal of $\mathbf{X}_n$ (denoted by $P^n_{\mathrm{DP}_\gamma}$) and the conditional of P given $\mathbf{X}_n$, on the other hand; iv) the last inequality follows by considering the infimum over all possible values of $\hat{M}_n(\mathbf{X}_n)$. We also recall that when P is distributed according to DP_γ, then the marginal of $\mathbf{X}_n$, $P^n_{\mathrm{DP}_\gamma}$, is a Generalized Polya urn, while the conditional of P given $\mathbf{X}_n$ is $\mathrm{DP}_{\gamma+\sum_{i=1}^n \delta_{X_i}}$ (see Theorem 4.6 and subsection 4.1.4 of Ghosal and Van der Vaart [12]). From Proposition A.2, $M_n(P,\mathbf{X}_n)$ under the posterior distribution $\mathrm{DP}_{\gamma+\sum_{i=1}^n \delta_{X_i}}$ is distributed according to a Beta random variable $\mathrm{Beta}(1,n)$. Therefore, we can write

$$\int_{\mathcal{P}} \mathbb{1}(L(M_n(P,\mathbf{X}_n), x) > \epsilon)\mathrm{DP}_{\gamma+\sum_{i=1}^n \delta_{X_i}}(dP) = \mathbb{P}\left(\left|\frac{Z}{x} - 1\right| > \epsilon\right) \tag{2.4}$$

where $Z \sim \mathrm{Beta}(1,n)$. From Lemma A.1, we have that $\mathbb{P}\left(\left|\frac{Z}{x} - 1\right| > \epsilon\right) \geq 1 - \frac{2\epsilon}{(1-\epsilon)}$ for all $x \geq 0$ and $n \geq 2$. Plugging this estimate in place of (2.4), we obtain

$$\sup_{P\in\mathcal{P}} \limsup_n P^n(L(M_n(P,\mathbf{X}_n), \hat{M}_n(\mathbf{X}_n)) > \epsilon) \geq 1 - \frac{2\epsilon}{(1-\epsilon)}$$

and the right hand side is strictly positive for all $0 < \epsilon < 1/4$. □

Mossel Ohannessian [24] proved Theorem 2.1 by exploiting a coupling of two generalized (dithered) geometric distributions. Here we presented an alternative proof of Theorem 2.1. While the proof of Mossel and Ohannessian [24] has the merit of being a constructive proof, our approach has the merit of being a direct approach, which exploits properties of the posterior distribution of the missing mass $M_n(P, \mathbf{X}_n)$ under a Ferguson-Dirichlet prior for P. Beside being of independent interest, our alternative proof suggests a natural approach to study rates of convergence of the Good-Turing estimator under the class of $\alpha \in (0, 1)$ regularly varying P. Indeed similar Bayesian arguments will be crucial in the next section to study the rate of convergence and the minimax rate ot the Good-Turing estimator under the class of $\alpha \in (0, 1)$ regularly varying P.

Theorem 2.1 shows that for any asymptotic result to hold uniformly over a set of possible distributions, the parametric space $\mathcal{P}$ must be restricted to a suitable subclass. In particular, the proof of Theorem 2.1 shows that there are no consistent estimators for the class of distributions sampled from a Ferguson-Dirichlet process. This suggests that conditions must be imposed on the tail decay of the elements of the parameter space. From Kingman [19] (Equation 65), we have that, if P is sampled from a Ferguson-Dirichlet process, its sequence of ordered masses behaves like $\log p_{[j]} \sim -jC$, as $j \to +\infty$. Therefore, the tail of P has approximately exponential form, resembling a geometric distribution and satisfying $p_{[j]} = o(j^{-\frac{1}{\alpha}})$ for every $\alpha \in (0, 1)$. Indeed, a geometric distribution was first used in Ohannessian and Dahleh [27] to prove that the Good-Turing estimator may be inconsistent. Theorem 2.1 shows that, under this very light regime, any estimator of the missing mass, not just the Good-Turing, fails to be consistent under multiplicative loss. This motivates us to consider, in the next section, the class of discrete distributions P's having heavy enough tails.

3. Consistent and minimax rate optimal estimation of $M_n(P, \mathbf{X}_n)$

We start by recalling the Good-Turing estimator (Good [15]) of the missing mass $M_n(P, \mathbf{X}_n)$, and then we investigate its convergence rate and minimax risk for regularly varying P. The definition of the Good-Turing estimator makes use of the proportion of unique values in the random sample $\mathbf{X}_n$ to estimate the missing mass. In particular, let $Y_{n,j}(\mathbf{X}_n)$ denote the number of times the value θ_j is observed in the sample $\mathbf{X}_n$, namely

$$Y_{n,j}(\mathbf{X}_n) = \sum_{i=1}^{n} \mathbb{1}(X_i = \theta_j).$$

Furthermore, let $K_{n,r}(\mathbf{X}_n)$ and $K_n(\mathbf{X}_n)$ be the number of values observed $1 \leq r \leq n$ times and the total number of distinct values, respectively, observed in $\mathbf{X}_n$, i.e.,

$$K_{n,r}(\mathbf{X}_n) = \sum_{j=1}^{\infty} \mathbb{1}(Y_{n,j} = r) \qquad K_n(\mathbf{X}_n) = \sum_{r=1}^{n} K_{n,r}(\mathbf{X}_n).$$

The Good-Turing estimator of $M_n(P, \mathbf{X}_n)$ is defined in terms of the statistic $K_{n,1}(\mathbf{X}_n)$, that is

$$\hat{GT}(\mathbf{X}_n) = \frac{K_{n,1}(\mathbf{X}_n)}{n}. \tag{3.1}$$

Ohannessian and Dahleh [27] first proved the inconsistency, as $n \to +\infty$, of $\hat{GT}(\mathbf{X}_n)$ under the choice of P being a geometric distribution. In the same paper, it is shown that under the assumption that the tail of P decays to zero as a regularly varying function with parameter $\alpha \in (0, 1)$, the Good-Turing estimator is strongly consistent. This latter result was generalized to the range $\alpha \in (0, 1]$ in the work of Ben-Hamou et al. [2].

The assumption of regularly varying distribution P provides with a generalization of the power law tail decay, adding some more flexibility by the introduction of the slowly varying function ℓ. Power-law distributions are observed in the empirical distributions of many quantities in different applied areas, and

their study have attracted a lot of interest in recent years. For extensive discussions of power laws in empirical data and their properties, the reader is referred to Mitzenmacher [23], Goldwater et al. [14], Newman [26], Clauset et al. [7] and Sornette [36]. Restricting the parameter space to probability distributions having regularly varying tail is not a mere technical assumption and, on the contrary, it represents a natural subset of the parameter space to consider, which we expect to contain the true data generating distribution for many different applications. To move into the concrete setting of regular variation (e.g., Bingham et al. [4]), for every $P \in \mathcal{P}$ we define a counting measure on $[0,1]$ as $\nu_P(dx) = \sum_j \delta_{p_j}(dx)$, with corresponding tail function defined as $\vec{\nu}_P(x) = \nu([x, +\infty))$ for all $x > 0$. Then a distribution $P \in \mathcal{P}$ is said to be regularly varying with parameter $\alpha \in (0,1)$ if

$$\vec{\nu}_P(x) \overset{x\downarrow 0}{\sim} x^{-\alpha}\ell(1/x), \tag{3.2}$$

where ℓ is a slowly varying function. From Lemma 22 and Proposition 23 of Gnedin et al. [13], (3.2) is equivalent to the more explicit condition in term of ordered masses of P

$$p_{[j]} \overset{j\uparrow\infty}{\sim} j^{-1/\alpha}\ell_*(j), \tag{3.3}$$

where ℓ_* is a slowly varying function depending on ℓ. We denote by $\mathcal{P}_{RV_\alpha} \subseteq \mathcal{P}$ the set of all regularly varying distribution on $[0,1]$ with parameter α. From (3.3) it is clear that such a class includes distributions having power-law tail decay, which correspond to the particular case of ℓ_* being a constant, which is equivalent to ℓ being constant. We denote the class of distributions having power law tail decay by $\mathcal{P}_{PL_\alpha} \subseteq \mathcal{P}_{RV_\alpha}$. In the following results, we will restrict our attention to the estimation problem under restricted parameter spaces $\mathcal{P}_{RV_\alpha}$ and $\mathcal{P}_{PL_\alpha}$.

Ohannessian and Dahleh [27] proved that the Good-Turing estimator is consistent under all distributions P in the class $\mathcal{P}_{RV_\alpha}$. Grabchak and Zhang [16] then proved that the convergence rate of the Good-Turing estimator is $n^{-\alpha/2}$, up to a slowly varying function. Hereafter we strengthen, and expand, the result of Grabchak and Zhang [16]. Recall that, under the loss function (1.2), a sequence $(r_n)_{n\in\mathbb{N}}$ is a convergence rate of an estimator $\hat{M}_n(\mathbf{X}_n)$ for the distribution $P \in \mathcal{P}$ if

$$\lim_n P^n(L(\hat{M}_n(\mathbf{X}_n), M_n(P, \mathbf{X}_n)) > T_n r_n) = 0$$

for all sequences $T_n \to \infty$. For the sake of clarity, in the next proposition we state the result of Grabchak and Zhang [16] on the rate of convergence of the Good-Turing estimator. An alternative proof of the next proposition follows from Proposition 3.2 below along with a straightforward application of Markov's inequality.

Proposition 3.1. *Let $\hat{GT}(\boldsymbol{X}_n)$ be the Good-Turing estimator, defined in (3.1). Then, for every $P \in \mathcal{P}_{RV_\alpha}$ and for all $T_n \to \infty$,*

$$\lim_n P^n(L(\hat{GT}(\boldsymbol{X}_n), M_n(P, \boldsymbol{X}_n)) > T_n n^{-\alpha/2}\ell^{-1/2}(n)) = 0, \tag{3.4}$$

where ℓ in (3.4) is the slowly varying function specific to P appearing in (3.2). Therefore, up to slowly varying functions, $n^{-\alpha/2}$ is a convergence rate for the Good-Turing estimator within the class $\mathcal{P}_{RV_\alpha}$.

The next theorem strengthen Proposition 3.1 on the rate of convergence of the Good-Turing estimator. Indeed it shows that the convergence rate achieved by the Good-Turing estimator is actually almost the best convergence rate any estimator of $M_n(\mathbf{X}_n, P)$ can achieve. For any other estimator of the missing mass, it is possible to find a point $P \in \mathcal{P}_{PL_\alpha}$ for which the rate of convergence is not faster than $n^{-\alpha/2}$.

Theorem 3.1. *For any estimator $\hat{M}_n(\boldsymbol{X}_n)$, there exists $P \in \mathcal{P}_{PL_\alpha} \subset \mathcal{P}_{RV_\alpha}$ such that for every $T_n \to 0$*

$$\liminf_n P^n(L(\hat{M}_n(\boldsymbol{X}_n), M_n(P, \boldsymbol{X}_n)) < T_n n^{-\alpha/2}) = 0.$$

Therefore the convergence rate of $\hat{M}_n(\boldsymbol{X}_n)$ cannot be faster than $n^{-\alpha/2}$.

Proof. Let $(T_n)_n$ any non-negative sequence converging to 0. We will show that for any estimator $\hat{M}_n(\mathbf{X}_n)$,

$$\inf_{P\in\mathcal{P}_{RV_\alpha}} \liminf_n P^n(L(\hat{M}_n(\mathbf{X}_n), M_n(P,\mathbf{X}_n)) < T_n n^{-\alpha/2}) = 0. \tag{3.5}$$

Let us denote by SP_α the law of a stable process on $[0,1]$ of parameter α. This a subordinator with Levy intensity, $\nu(d\omega) = \frac{\alpha}{\Gamma(1-\alpha)}\omega^{-1-\alpha}d\omega$. See Kingman [19] and Pitman [33] for details and additional references. Because of $\nu([x,\infty)) = \frac{x^{-\alpha}}{\Gamma(1-\alpha)}$, the stable process samples probability measures belonging to $\mathcal{P}_{PL_\alpha}$. Now we can upper bound the infimum in (3.5) by an average with respect to SP_α,

$$\begin{aligned}
&\inf_{P\in\mathcal{P}_{RV_\alpha}} \liminf_n P^n(L(\hat{M}_n(\mathbf{X}_n), M_n(P,\mathbf{X}_n)) < T_n n^{-\alpha/2}) \\
&\leq \int_{\mathcal{P}} \liminf_n P^n(L(\hat{M}_n(\mathbf{X}_n), M_n(P,\mathbf{X}_n)) < T_n n^{-\alpha/2}) SP_\alpha(dP) \\
&\leq \liminf_n \int_{\mathcal{P}} \int_{[0,1]^n} \mathbb{1}(L(\hat{M}_n(\mathbf{X}_n), M_n(P,\mathbf{X}_n)) < T_n n^{-\alpha/2}) P^n(d\mathbf{X}_n) SP_\alpha(dP)
\end{aligned}$$

where the last equality follows by applying Fatou's Lemma.

Take n large enough so that $T_n n^{-\alpha/2} < 1/2$. Let us denote by $P^n_{SP_\alpha}$ the marginal law of the observations under an α-stable process, when P is integrated out, i.e. the probability measure on $[0,1]^n$ defined as $P^n_{SP_\alpha}(A) = \int_{\mathcal{P}} P^n(A) SP_\alpha(dP)$ for all $A \in \mathcal{B}([0,1]^n)$. We swap the integration of the marginal of P and the conditional of $\mathbf{X}_n$ given P with the marginal of $\mathbf{X}_n$ and the conditional of P given $\mathbf{X}_n$ and then apply Fact A.1 to obtain

$$\begin{aligned}
&\int_{\mathcal{P}} \int_{[0,1]^n} \mathbb{1}(L(\hat{M}_n(\mathbf{X}_n), M_n(P,\mathbf{X}_n)) < T_n n^{-\alpha/2}) P^n(d\mathbf{X}_n) SP_\alpha(dP) \\
&= \int_{[0,1]^n} \int_{\mathcal{P}} \mathbb{1}(L(\hat{M}_n(\mathbf{X}_n), M_n(P,\mathbf{X}_n)) < T_n n^{-\alpha/2}) SP_\alpha|\mathbf{X}_n(dP) P^n_{SP_\alpha}(d\mathbf{X}_n) \\
&\leq \int_{[0,1]^n} \int_{\mathcal{P}} \mathbb{1}(L(M_n(P,\mathbf{X}_n), \hat{M}_n(\mathbf{X}_n))) < 2T_n n^{-\alpha/2}) SP_\alpha|\mathbf{X}_n(dP) P^n_{SP_\alpha}(d\mathbf{X}_n)
\end{aligned}$$

where $SP_\alpha|\mathbf{X}_n$ denotes the posterior distribution of P given the sample $\mathbf{X}_n$. Therefore, taking $s > 1$, we can upper bound the quantity appearing on the l.h.s. of (3.5) by

$$\begin{aligned}
&\limsup_n \int_{[0,1]^n} \mathbb{1}(K_n(\mathbf{X}_n) \in (n^\alpha/s, sn^\alpha)) \\
&\times \int_{\mathcal{P}} \mathbb{1}(L(M_n(P,\mathbf{X}_n), \hat{M}_n(\mathbf{X}_n))) < 2T_n n^{-\alpha/2}) SP_\alpha|\mathbf{X}_n(dP) P^n_{SP_\alpha}(d\mathbf{X}_n) \\
&\qquad + \limsup_n \; P^n_{SP_\alpha}(K_n(\mathbf{X}_n) \notin (n^\alpha/s, sn^\alpha)).
\end{aligned} \tag{3.6}$$

We will now upper bound the two terms of the sum in (3.6) independently. Let us focus on the first term of (3.6). Let n large enough so that $3 < \frac{\alpha n^\alpha}{s} < \alpha s n^\alpha < n-3$ and $T_n n^{-\alpha/2} < 1/4$. From Proposition A.2, under the posterior $SP_\alpha|\mathbf{X}_n$, $M_n(P,\mathbf{X}_n)$ is distributed according to a Beta random variable $\mathrm{Beta}(\alpha K_n(\mathbf{X}_n), n - \alpha K_n(\mathbf{X}_n))$. Let us denote $a(\mathbf{X}_n) = \alpha K_n(\mathbf{X}_n)$, $b(\mathbf{X}_n) = n - \alpha K_n(\mathbf{X}_n)$, and for easiness of notation we will simply write a and b in the following calculations. Moreover let $F_{a,b}$ be the cumulative distribution function of the Beta random variable $\mathrm{Beta}(a,b)$. Thanks to Proposition A.2 in the Appendix, we have that

$$\begin{aligned}
&\int_{\mathcal{P}} \mathbb{1}(L(M_n(P,\mathbf{X}_n), \hat{M}_n(\mathbf{X}_n))) < 2T_n n^{-\alpha/2}) SP_\alpha|\mathbf{X}_n(dP) \\
&\quad = F_{a,b}((1+2T_n n^{-\alpha/2})\hat{M}_n(\mathbf{X}_n)) - F_{a,b}((1-2T_n n^{-\alpha/2})\hat{M}_n(\mathbf{X}_n)) \\
&\quad \leq \sup_{x\in[0,1]} \left(F_{a,b}((1+2T_n n^{-\alpha/2})x) - F_{a,b}((1-2T_n n^{-\alpha/2})x) \right).
\end{aligned}$$

Consider the function $\psi : \mathbb{R}_+ \to [0,1]$ defined by

$$\psi(x) = F_{a,b}((1+2T_n n^{-\alpha/2})x) - F_{a,b}((1-2T_n n^{-\alpha/2})x).$$

Notice that $\psi \in \mathcal{C}^2$ and that $\psi(0) = \lim_{x\to\infty}\psi(x) = 0$. Therefore, ψ reaches its maximum in $x^*(a,b) \in \mathbb{R}_+$ (denoted as x^* for easiness of notation) satisfying

$$\begin{aligned}\psi'(x^*) = (1+&2T_n n^{-\alpha/2}) f_{a,b}((1+2T_n n^{-\alpha/2})x^*) \\ &- (1-2T_n n^{-\alpha/2}) f_{a,b}((1-2T_n n^{-\alpha/2})x^*) = 0,\end{aligned}$$

where $f_{a,b}$ is the density function of the Beta(a,b) distribution. On the event $K_n(\mathbf{X}_n) \in (n^\alpha/s, sn^\alpha)$, we have $a, b > 3$, thus $f_{a,b}$ is bell-shaped with second inflexion point given by

$$\kappa(a,b) = \frac{a-1}{a+b-2} + \frac{\sqrt{\frac{(a-1)(b-1)}{a+b-3}}}{a+b-2} \leq \frac{\alpha s n^\alpha}{n-2} + \frac{\sqrt{\alpha s n^\alpha}}{n-2} \leq \frac{2\alpha s n^\alpha}{n-2}.$$

Therefore, $f'_{a,b}$ is non decreasing on the interval $[\kappa(a,b), \infty)$ and, as a consequence, ψ'' is non negative on $[\frac{\kappa(a,b)}{(1-2T_n n^{-\alpha/2})}, \infty)$, from which we can deduce that ψ' is non decreasing on the same interval. Now, since $\lim_{x\to\infty}\psi'(x) = 0$, it follows that $\psi'(x) \leq 0$ on $[\frac{\kappa(a,b)}{(1-2T_n n^{-\alpha/2})}, \infty)$. Therefore $x^*(a,b)$ satisfies

$$x^*(a,b) \leq \frac{\kappa(a,b)}{(1-2T_n n^{-\alpha/2})} \leq \frac{2\alpha s n^\alpha}{(n-2)(1-2T_n n^{-\alpha/2})} \leq \frac{4\alpha s n^\alpha}{n-2}.$$

We can now upper bound $\sup_{x\geq 0}\psi(x)$ as follows:

$$\sup_{x\geq 0}\psi(x) = \psi(x^*) \leq 4T_n n^{-\alpha/2} x^* \sup_{x\geq 0} f_{a,b}(x) \leq 16 T_n n^{-\alpha/2} \frac{\alpha s n^\alpha}{n-2} \sup_{x\geq 0} f_{a,b}(x).$$

By Lemma A.3 in Appendix A it follows that, on the event $K_n \in (n^\alpha/s, sn^\alpha)$, for n large enough,

$$\begin{aligned}\sup_{x\geq 0}\psi(x) &\leq 128 T_n n^{-\alpha/2} \frac{\alpha s n^\alpha}{n-2} (a+b)^{3/2} a^{-1/2} b^{-1/2} \\ &\leq 128 T_n n^{-\alpha/2} \frac{\alpha s n^\alpha}{n-2} n^{3/2} (\alpha n^\alpha/s)^{-1/2} (n - s\alpha n^\alpha)^{-1/2} = T_n g(\alpha, s, n)\end{aligned}$$

where we also have $\limsup_{n\to+\infty} T_n g(\alpha, s, n) = 0$.

From all previous computations, we deduce that, on the event $K_n(\mathbf{X}_n) \in (1/sn^\alpha, sn^\alpha)$, there exists $n_0(\alpha, s)$, which does not depend on the value of K_n, such that for all $n \geq n_0(\alpha, s)$ the inequality

$$\int_{\mathcal{P}} \mathbb{1}(L(M_n(P, \mathbf{X}_n), \hat{M}_n(\mathbf{X}_n))) < 2T_n n^{-\alpha/2}) SP_\alpha|\mathbf{X}_n(dP) \leq T_n g(\alpha, s, n)$$

holds true, thus we get

$$\begin{aligned}\limsup_n \int_{[0,1]^n} &\mathbb{1}(K_n(\mathbf{X}_n) \in (1/sn^\alpha, sn^\alpha)) \\ &\times \int_{\mathcal{P}} \mathbb{1}(L(M_n(P, \mathbf{X}_n), \hat{M}_n(\mathbf{X}_n))) < 2T_n n^{-\alpha/2}) SP_\alpha|\mathbf{X}_n(dP) P^n_{SP_\alpha}(d\mathbf{X}_n) \\ &\leq \limsup_n T_n g(\alpha, s, n) = 0.\end{aligned}$$

Therefore the first term in (3.6) is equal to zero.

Let us now consider the second term in (3.6), namely

$$\limsup_n P^n_{SP_\alpha}(K_n(\mathbf{X}_n) \notin (n^\alpha/s, sn^\alpha)).$$

By virtue of Theorem 3.8 of Pitman [33], under the α-stable process, $\frac{K_n(\mathbf{X}_n)}{n^\alpha} \to S_\alpha$ almost surely, where S_α is a random variable on $\mathbb{R}_+$ distributed according to a Stable distribution of parameter α. As a consequence we have

$$\limsup_n P^n_{SP_\alpha}(K_n(\mathbf{X}_n) \notin (n^\alpha/s, sn^\alpha)) = \mathbb{P}(S_\alpha \notin (1/s, s)). \tag{3.7}$$

The r.h.s. of (3.7) converges to zero as $s \to \infty$, and then so does (3.6). □

Proposition 3.1 and Theorem 3.1 together show that the Good-Turing estimator achieves the best convergence rate up to a slowly varying function. In particular, if the distribution P has a power-law decay, i.e. $P \in \mathcal{P}_{PL_\alpha}$, the two rates match and the Good-Turing estimator achieves the best rate possible. In particular, because $\hat{GT}(\mathbf{X}_n)$ does not depend on α, it follows that the Good-Turing estimator is actually *rate adaptive* for the class of power law distributions, $\mathcal{P}_{PL} = \cup_{0<\alpha<1}\mathcal{P}_{PL_\alpha}$.

The next theorem considers the asymptotic minimax estimation risk for the missing mass under the loss function (1.2) and with parameter space $\mathcal{P}_{PL_\alpha}$. In particular Theorem 3.2 provides with a lower bound for the estimation risk of this statistical problem, showing that the minimax rate is not smaller than $n^{-\alpha/2}$.

Theorem 3.2. *Let $\mathcal{P}_{PL_\alpha}$ be the class of discrete distributions on $[0,1]$ with power law tail function and let L denote the multiplicative loss function (1.2). Then, there exists a positive constant $C > 0$ such that*

$$\liminf_n n^{\alpha/2} \inf_{\hat{M}_n(\boldsymbol{X}_n)} \sup_{P \in \mathcal{P}_{PL_\alpha}} \mathbb{E}_P\left(L(\hat{M}_n(\boldsymbol{X}_n), M_n(P, \boldsymbol{X}_n))\right) > C$$

where the infimum is taken over all possible estimators $\hat{M}_n(\boldsymbol{X}_n)$.

Proof. In the following, we make use of the generic notation C to refer to constants that can only depend on α (its value may change from a line to the other). As in the proof of Theorem 3.1, let SP_α denote the law of a stable process of parameter α and $P^n_{SP_\alpha}$ the marginal law of the observations under this prior. We can lower bound the minimax risk by the Bayesian risk with respect to the prior SP_α. Indeed one has

$$\begin{aligned}
&\inf_{\hat{M}_n(\mathbf{X}_n)} \sup_{P \in \mathcal{P}_{PL_\alpha}} \mathbb{E}_P\left(L(\hat{M}_n(\mathbf{X}_n), M_n(P, \mathbf{X}_n))\right) \\
&\geq \inf_{\hat{M}_n(\mathbf{X}_n)} \int_{\mathcal{P}} \mathbb{E}_P\left(L(\hat{M}_n(\mathbf{X}_n), M_n(P, \mathbf{X}_n))\right) SP_\alpha(dP) \\
&= \inf_{\hat{M}_n(\mathbf{X}_n)} \int_{[0,1]^n} \int_{\mathcal{P}} L(\hat{M}_n(\mathbf{X}_n), M_n(P, \mathbf{X}_n)) SP_\alpha|\mathbf{X}_n(dP) P^n_{SP_\alpha}(d\mathbf{X}_n) \\
&\geq \int_{[0,1]^n} \inf_{\hat{M}_n(\mathbf{X}_n)} \int_{\mathcal{P}} L(\hat{M}_n(\mathbf{X}_n), M_n(P, \mathbf{X}_n)) SP_\alpha|\mathbf{X}_n(dP) P^n_{SP_\alpha}(d\mathbf{X}_n).
\end{aligned} \tag{3.8}$$

From Proposition A.2 in Appendix A, the posterior distribution of missing mass $M_n(P, \mathbf{X}_n)$ under SP_α is distributed according to $\text{Beta}(\alpha K_n(\mathbf{X}_n), n - \alpha K_n(\mathbf{X}_n))$. Le $a(\mathbf{X}_n) = \alpha K_n(\mathbf{X}_n)$ and $b(\mathbf{X}_n) = n - \alpha K_n(\mathbf{X}_n)$, and for easiness of notation we will simply write a and b in the following calculations. The inner integral in (3.8) equals

$$\begin{aligned}
&\int_{\mathcal{P}} L(\hat{M}_n(\mathbf{X}_n), M_n(P, \mathbf{X}_n)) SP_\alpha|\mathbf{X}_n(dP) = \int_0^1 \frac{|\hat{M}_n(\mathbf{X}_n) - x|}{x} \frac{x^{a-1}(1-x)^{b-1}}{\mathcal{B}(a,b)} dx \\
&= \frac{\mathcal{B}(a-1,b)}{\mathcal{B}(a,b)} \int_0^1 |\hat{M}_n(\mathbf{X}_n) - x| \frac{x^{a-2}(1-x)^{b-1}}{\mathcal{B}(a-1,b)} dx = \frac{\mathcal{B}(a-1,b)}{\mathcal{B}(a,b)} \mathbb{E}_{M'}\left(|M' - \hat{M}_n(\mathbf{X}_n)|\right)
\end{aligned}$$

where M' is a random variable distributed according to Beta$(a-1,b)$. Plugging this quantity into (3.8), we find

$$\begin{aligned}
&\inf_{\hat{M}_n(\mathbf{X}_n)} \sup_{P\in\mathcal{P}_{PL_\alpha}} \mathbb{E}_P\left(L(\hat{M}_n(\mathbf{X}_n), M_n(P,\mathbf{X}_n))\right) \\
&\quad\geq \int_{[0,1]^n} \frac{\mathcal{B}(a-1,b)}{\mathcal{B}(a,b)} \inf_{\hat{M}_n(\mathbf{X}_n)} \mathbb{E}_{M'}\left(|M'-\hat{M}_n(\mathbf{X}_n)|\right) P^n_{SP_\alpha}(d\mathbf{X}_n) \\
&\quad= \int_{[0,1]^n} \frac{\mathcal{B}(a-1,b)}{\mathcal{B}(a,b)} \mathbb{E}_{M'}\left(|M'-\mathrm{med}(M')|\right) P^n_{SP_\alpha}(d\mathbf{X}_n) \qquad (3.9)
\end{aligned}$$

where $\mathrm{med}(M')$ denotes the median of M'. Now, let us denote by $f_{a,b}$ and $m_{a,b}$ the density function and the median of a Beta random variable with parameters a and b, respectively. From Lemma A.2, we can rewrite the inner expectation in (3.9) as

$$\frac{\mathcal{B}(a-1,b)}{\mathcal{B}(a,b)} \mathbb{E}_{M'}\left(|M'-m_{a-1,b}|\right) = 2\int_{m_{a-1,b}}^{m_{a,b}} f_{a,b}(x)dx,$$

therefore we have

$$\begin{aligned}
&\inf_{\hat{M}_n(\mathbf{X}_n)} \sup_{P\in\mathcal{P}_{PL_\alpha}} \mathbb{E}_P\left(L(\hat{M}_n(\mathbf{X}_n), M_n(P,\mathbf{X}_n))\right) \\
&\qquad\geq 2\int_{[0,1]^n} \left(\int_{m_{a-1,b}}^{m_{a,b}} f_{a,b}(x)dx\right) P^n_{SP_\alpha}(d\mathbf{X}_n), \qquad (3.10)
\end{aligned}$$

where we recall that $a=\alpha K_n(\mathbf{X}_n)$ and $b=n-\alpha K_n(\mathbf{X}_n)$. From Theorem 3.8 of Pitman [33], when P is distributed according to the α-stable process, $K_n(\mathbf{X}_n)/n^\alpha$ converges in distribution to a random variable S_α on $[0,\infty)$ with Stable distribution of parameter α. Therefore, there exist n_0 and two positive bounded values w_α and W_α such that for all $n>n_0$, $\mathbb{P}(\frac{K_n(\mathbf{X}_n)}{n^\alpha}\in[w_\alpha,W_\alpha])\geq\frac{1}{2}$ and $8<2\alpha w_\alpha n^\alpha\leq 2\alpha W_\alpha n^\alpha<n-1$.

We are now ready to lower bound (3.10). We will make use of some technical lemmas regarding the density and median of the Beta distribution, whose statements and proofs are deferred to Appendix A. The r.h.s. of (3.10) can be lower bounded by the following quantity

$$\begin{aligned}
&2\mathbb{P}(w_\alpha n^\alpha\leq K_n(\mathbf{X}_n)\leq W_\alpha n^\alpha) \\
&\qquad\times \mathbb{E}_{P^n_{SP_\alpha}}\left[\int_{m_{a-1,b}}^{m_{a,b}} f_{a,b}(x)dx \Big| w_\alpha n^\alpha\leq K_n(\mathbf{X}_n)\leq W_\alpha n^\alpha\right]. \qquad (3.11)
\end{aligned}$$

Given our choice of n_0, w_α and W_α, for $n>n_0$, we have that $\mathbb{P}(w_\alpha n^\alpha\leq K_n\leq W_\alpha n^\alpha)\geq\frac{1}{2}$. Recall now that $a=\alpha K_n(\mathbf{X}_n)$ and $b=n-\alpha K_n(\mathbf{X}_n)$. Moreover noticing that we are conditioning on the event $w_\alpha n^\alpha\leq K_n(\mathbf{X}_n)\leq W_\alpha n^\alpha$, one has $3<a<b$ and $a<b/2$; by applying Lemma A.4 we can lower bound (3.11) by

$$2\frac{1}{2}\mathbb{E}_{P^n_{SP_\alpha}}\left[\frac{C}{\sqrt{a}}|w_\alpha n^\alpha\leq K_n(\mathbf{X}_n)\leq W_\alpha n^\alpha\right],$$

for some strictly positive constant C. Ultimately this leads to

$$\begin{aligned}
&\inf_{\hat{M}_n(\mathbf{X}_n)} \sup_{P\in\mathcal{P}_{RV_\alpha}} \mathbb{E}_P\left(L(\hat{M}_n(\mathbf{X}_n), M_n(P,\mathbf{X}_n))\right) \\
&\qquad\geq C\mathbb{E}_{P^n_{SP_\alpha}}\left[a^{-1/2}\mid w_\alpha n^\alpha\leq K_n(\mathbf{X}_n)\leq W_\alpha n^\alpha\right] \\
&\qquad= C\alpha^{-1/2}\mathbb{E}_{P^n_{SP_\alpha}}\left[K_n^{-1/2}\mid w_\alpha n^\alpha\leq K_n(\mathbf{X}_n)\leq W_\alpha n^\alpha\right] \\
&\qquad\geq C(w_\alpha n^\alpha)^{-1/2} = Cn^{-\alpha/2},
\end{aligned}$$

which provides the lower bound rate for the minimax risk,

$$\liminf_n n^{\alpha/2} \inf_{\hat{M}_n(\mathbf{X}_n)} \sup_{P\in\mathcal{P}_{PL_\alpha}} \mathbb{E}_P\left(L(\hat{M}_n(\mathbf{X}_n), M_n(P,\mathbf{X}_n))\right) > C.$$

□

The lower bound of Theorem 3.2 can be used to derive the minimax rate, by matching it with appropriate upper bounds of specific estimators of the missing mass. This lower bound trivially still holds for any parametric set larger than $\mathcal{P}_{PL_\alpha}$ and, therefore Theorem 3.2 also provides with a lower bound of the estimation risk under the larger parameter space $\mathcal{P}_{RV_\alpha}$. In the next Proposition, we show that for a fixed distribution $P \in \mathcal{P}_{RV_\alpha}$, the Good-Turing estimator achieves the best possible rate of Theorem 3.2 up to a slowly varying term.

Proposition 3.2. *Let $\hat{GT}(\boldsymbol{X}_n)$ be the Good-Turing estimator and let $P \in \mathcal{P}_{RV_\alpha}$. Then, there exists a finite constant C such that for every n*

$$\mathbb{E}_P(L(\hat{GT}(\boldsymbol{X}_n), M_n(P,\boldsymbol{X}_n))) \leq Cn^{-\alpha/2}\ell^{-1/2}(n),$$

where ℓ is the slowly varying function specific to P appearing in (3.2).

Proof. Let $P \in \mathcal{P}_{RV_\alpha}$ and ℓ defined as in (3.2). In the following we use the generic notation C and C' to refer to constants that can only depend on P (their values may change from a line to the other). Here we study the convergence rate under the assumption of regular variation of the Good-Turing estimator $\hat{GT}(\mathbf{X}_n) = \frac{K_{n,1}(\mathbf{X}_n)}{n}$, proving that

$$\mathbb{E}_P\left(L(\hat{GT}(\mathbf{X}_n), M_n(P))\right) = O(n^{-\alpha/2}\ell(n)^{-1/2}).$$

Let us first notice that for $a \geq 0,\ b, c > 0$

$$L(a,c) = \left|\frac{a}{b}\left(\frac{b}{c} - 1\right) + \frac{a}{b} - 1\right| \leq L(a,b) + \frac{a}{b}L(b,c).$$

Therefore, we can upper bound the loss of $\hat{GT}(\mathbf{X}_n) = \frac{K_{n,1}(\mathbf{X}_n)}{n}$ by

$$\begin{aligned}
&L(\hat{GT}(\mathbf{X}_n), M_n(P,\mathbf{X}_n)) \\
&\quad\leq L(\hat{GT}(\mathbf{X}_n), \mathbb{E}_P(\hat{GT}(\mathbf{X}_n)) + \frac{\hat{GT}(\mathbf{X}_n)}{\mathbb{E}_P(\hat{GT}(\mathbf{X}_n))}L(\mathbb{E}_P(\hat{GT}(\mathbf{X}_n)), \mathbb{E}_P(M_n(P,\mathbf{X}_n)) \\
&\qquad + \frac{\hat{GT}(\mathbf{X}_n)}{\mathbb{E}_P(M_n(P,\mathbf{X}_n))}L(\mathbb{E}_P(M_n(P,\mathbf{X}_n)), M_n(P,\mathbf{X}_n)),
\end{aligned}$$

and consequently its risk by

$$\begin{aligned}
\mathbb{E}_P(L(\hat{GT}(\mathbf{X}_n), M_n(P,\mathbf{X}_n))) &\leq \mathbb{E}_P(L(\hat{GT}(\mathbf{X}_n), \mathbb{E}_P(\hat{GT}(\mathbf{X}_n))) \\
&\quad + L(\mathbb{E}_P(\hat{GT}(\mathbf{X}_n)), \mathbb{E}_P(M_n(P,\mathbf{X}_n)) \\
&\qquad + \mathbb{E}_P\left(\frac{\hat{GT}(\mathbf{X}_n)}{\mathbb{E}_P M_n(P,\mathbf{X}_n)}L(\mathbb{E}_P(M_n(P,\mathbf{X}_n)), M_n(P,\mathbf{X}_n))\right).
\end{aligned} \tag{3.12}$$

We will now separately upper bound the three components of the r.h.s. of the previous inequality. Let us first focus on the quantity $L(\mathbb{E}_P(\hat{GT}(\mathbf{X}_n)), \mathbb{E}_P(M_n(P,\mathbf{X}_n))$. From the work of Karlin [18] (see also Theorem 4.2 of Ben-Hamou et al. [2]), we know that

$$\mathbb{E}_P(\hat{GT}(\mathbf{X}_n)) \sim \alpha\Gamma(1-\alpha)n^{\alpha-1}\ell(n),$$

as $n \to \infty$ and since $0 \leq \mathbb{E}_P(\hat{GT}(\mathbf{X}_n)) - \mathbb{E}_P(M_n(P, \mathbf{X}_n)) \leq \frac{1}{n}$, we deduce that

$$L(\mathbb{E}_P(\hat{GT}(\mathbf{X}_n)), \mathbb{E}_P(M_n(P, \mathbf{X}_n)) \leq Cn^{-\alpha}\ell(n)^{-1}. \tag{3.13}$$

Let us now consider the first term in the r.h.s. of (3.12),

$$\mathbb{E}_P(L(\hat{GT}(\mathbf{X}_n), \mathbb{E}_P(\hat{GT}(\mathbf{X}_n))) = \mathbb{E}_P\left(\left|\frac{K_{n,1}(\mathbf{X}_n)}{\mathbb{E}_P(K_{n,1}(\mathbf{X}_n))} - 1\right|\right). \tag{3.14}$$

As a result of Ben-Hamou et al. [2] Proposition 3.5 (see also the proof of corollary 5.3 of the same paper), for every $\epsilon > 0$, we have

$$P^n(L(K_{n,1}(\mathbf{X}_n), \mathbb{E}_P(K_{n,1}(\mathbf{X}_n))) \geq \epsilon) \leq 4e^{-\epsilon^2 A_n^2},$$

where

$$A_n = \frac{\mathbb{E}(K_{n,1}(\mathbf{X}_n))}{\sqrt{8(\mathbb{E}(K_{n,1}(\mathbf{X}_n)) \vee 2\mathbb{E}(K_{n,2}(\mathbf{X}_n))) + 4/3}}.$$

Hence, we can now bound (3.14) as follows

$$\begin{aligned}
\mathbb{E}_P(L(\hat{GT}(\mathbf{X}_n), \mathbb{E}_P(\hat{GT}(\mathbf{X}_n))) &= \mathbb{E}_P(L(K_{n,1}(\mathbf{X}_n), \mathbb{E}_P(K_{n,1}(\mathbf{X}_n))) \\
&= \int_0^\infty P^n(L(K_{n,1}(\mathbf{X}_n), \mathbb{E}_P(K_{n,1}(\mathbf{X}_n))) \geq \epsilon) d\epsilon \\
&\leq 4\int_0^\infty e^{-\epsilon^2 A_n^2} d\epsilon = \frac{4}{A_n}\int_0^\infty e^{-y^2} dy = CA_n^{-1},
\end{aligned}$$

where we have used the change of variables $y = \epsilon A_n$. Therefore, from the asymptotic behaviors of $\mathbb{E}_P(K_{n,1}(\mathbf{X}_n))$ and $\mathbb{E}_P(K_{n,2}(\mathbf{X}_n))$ that are provided in the work of Karlin [18] (see also Theorem 4.2 of Ben-Hamou et al. [2]), we conclude that

$$\mathbb{E}_P(L(\hat{GT}(\mathbf{X}_n), \mathbb{E}_P(\hat{GT}(\mathbf{X}_n))) \leq Cn^{-\alpha/2}\ell(n)^{-1/2}. \tag{3.15}$$

Finally, let us look at the third term in (3.12), namely

$$\mathbb{E}_P\left(\frac{\hat{GT}(\mathbf{X}_n)}{\mathbb{E}_P M_n(P, \mathbf{X}_n)} L(\mathbb{E}_P(M_n(P, \mathbf{X}_n)), M_n(P, \mathbf{X}_n))\right). \tag{3.16}$$

Notice that (3.16) is equal to

$$\mathbb{E}_P\left(\frac{\hat{GT}(\mathbf{X}_n)}{M_n(P, \mathbf{X}_n)} L(M_n(P, \mathbf{X}_n), \mathbb{E}_P(M_n(P, \mathbf{X}_n)))\right), \tag{3.17}$$

and from the Cauchy-Schwarz inequality, we can upper bound (3.17) by

$$\sqrt{\mathbb{E}_P\left(\frac{\hat{GT}(\mathbf{X}_n)^2}{M_n(P, \mathbf{X}_n)^2}\right)} \sqrt{\mathbb{E}_P\left(L(M_n(P, \mathbf{X}_n), \mathbb{E}_P(M_n(P, \mathbf{X}_n)))^2\right)}. \tag{3.18}$$

We will first compute the asymptotic behavior of the second term in (3.18) and then show that the first term is asymptotically bounded. By applying Theorem 3.9 of Ben-Hamou et al. [2] and the asymptotic regimes of Karlin [18], we obtain that for every $\epsilon > 0$,

$$P^n(L(M_n(P, \mathbf{X}_n), \mathbb{E}_P(M_n(P, \mathbf{X}_n)) \geq \epsilon) \leq 2e^{-\epsilon^2 B_n},$$

where

$$B_n \leq Cn^\alpha \ell(n)$$

(see for example the proof of Corollary 5.3 of Ben-Hamou et al. [2]). Therefore, for all $\epsilon > 0$

$$P^n(L(M_n(P,\mathbf{X}_n),\mathbb{E}_P(M_n(P,\mathbf{X}_n)))^2 \geq \epsilon) \leq 2e^{-\epsilon B_n}$$

and, following the same reasoning used before, we obtain

$$\mathbb{E}_P\left(L(M_n(P,\mathbf{X}_n),\mathbb{E}_P(M_n(P,\mathbf{X}_n)))^2\right) \leq CB_n^{-1}$$

which leads to

$$\sqrt{\mathbb{E}_P\left(L(M_n(P,\mathbf{X}_n),\mathbb{E}_P(M_n(P,\mathbf{X}_n)))^2\right)} \leq Cn^{-\alpha/2}\ell(n)^{-1/2}. \tag{3.19}$$

It remains only to prove that the first term in (3.18) is bounded. First note that

$$\frac{\hat{GT}(\mathbf{X}_n)}{M_n(P,\mathbf{X}_n)} \leq \frac{K_n(\mathbf{X}_n)}{n\sum_{j>K_n(\mathbf{X}_n)} p_{[j]}}$$

For $t \geq 1$, let us define the function f by $f(t) = \frac{t}{\sum_{j>t} p_{[j]}}$. Noticing that $f(t) \geq 1 > 0$ we can write

$$\frac{\hat{GT}(\mathbf{X}_n)}{M_n(P,\mathbf{X}_n)} \leq \frac{f(K_n(\mathbf{X}_n))}{f(\mathbb{E}(K_n(\mathbf{X}_n)))}\frac{f(\mathbb{E}(K_n(\mathbf{X}_n)))}{n}. \tag{3.20}$$

Denoting by $\ell^{\frac{1}{\alpha}\#}$ the de Bruijn conjugate of $\ell^{\frac{1}{\alpha}}$ (see subsection 1.5.7 of Bingham et al. [4] for a definition), Proposition 23 of Gnedin et al. [13] implies the following

$$f(t) \sim Ct^{\frac{1}{\alpha}}\ell^{\frac{1}{\alpha}\#}(t^{\frac{1}{\alpha}}),$$

which in turns implies that f^2 is regularly varying with index $2/\alpha$. Since f is non decreasing, it is bounded on any set of the form $[1,T]$, we can apply Potter's Theorem (Theorem 1.5.6, Bingham et al. [4]) to obtain

$$\frac{f(K_n(\mathbf{X}_n))^2}{f(\mathbb{E}_P(K_n(\mathbf{X}_n)))^2} \leq C\left[\left(\frac{K_n(\mathbf{X}_n)}{\mathbb{E}_P(K_n(\mathbf{X}_n))}\right)^{\frac{1}{\alpha}} + \left(\frac{K_n(\mathbf{X}_n)}{\mathbb{E}_P(K_n(\mathbf{X}_n))}\right)^{\frac{3}{\alpha}}\right] + C'.$$

Following the same lines of reasoning that we used before, we can show that for all $\eta > 1$

$$\lim_{n\to+\infty} \mathbb{E}_P[L(K_n(\mathbf{X}_n),\mathbb{E}_P(K_n(\mathbf{X}_n)))^\eta] = 0, \tag{3.21}$$

and, thanks to the elementary inequality $x/|x-1| \leq 2$ for $x \geq 2$, it follows that, for all $\eta > 1$,

$$\left(\frac{K_n(\mathbf{X}_n)}{\mathbb{E}_P(K_n(\mathbf{X}_n))}\right)^\eta \leq 2^\eta + 2^\eta L(K_n(\mathbf{X}_n),\mathbb{E}_P(K_n(\mathbf{X}_n)))^\eta.$$

As a consequence of this last inequality, along with (3.21), for all $\eta > 1$ we obtain

$$\mathbb{E}_P\left(\left(\frac{K_n(\mathbf{X}_n)}{\mathbb{E}_P(K_n(\mathbf{X}_n))}\right)^\eta\right) \leq C$$

from which we get that

$$\mathbb{E}_P\left(\left(\frac{f(K_n(\mathbf{X}_n))}{f(\mathbb{E}_P(K_n(\mathbf{X}_n)))}\right)^\eta\right) \leq C. \tag{3.22}$$

Besides, since $\mathbb{E}_P(K_n(\mathbf{X}_n))^{\frac{1}{\alpha}} \sim n\ell^{\frac{1}{\alpha}}(n)$, which diverges to infinity, the uniform convergence theorem for slowly varying functions (Theorem 1.2.1, Bingham et al. [4]) gives that

$$\ell^{\frac{1}{\alpha}\#}(\mathbb{E}_P(K_n(\mathbf{X}_n))^{\frac{1}{\alpha}}) \sim \ell^{\frac{1}{\alpha}\#}(n\ell^{\frac{1}{\alpha}}(n)).$$

As a consequence of this and of the asymptotic properties of f, we obtain that

$$\frac{f(\mathbb{E}_P(K_n(\mathbf{X}_n)))}{n} \sim \ell^{\frac{1}{\alpha}}(n)\ell^{\frac{1}{\alpha}\#}(n\ell^{\frac{1}{\alpha}}(n)),$$

which, from the definition of the de Bruijn conjugate, in turn gives

$$\frac{f(\mathbb{E}_P(K_n(\mathbf{X}_n)))}{n} \sim 1,$$

and then

$$\frac{f(\mathbb{E}_P(K_n(\mathbf{X}_n)))^2}{n^2} \leq C. \tag{3.23}$$

From (3.20), (3.22) and (3.23) together, we finally obtain

$$\mathbb{E}_P\left(\frac{\hat{GT}(\mathbf{X}_n)^2}{M_n(P,\mathbf{X}_n)^2}\right) \leq C,$$

which together with (3.13), (3.15) and (3.19) concludes the proof. □

Extending Proposition 3.2 to hold uniformly over $\mathcal{P}_{RV_\alpha}$ is an open problem and probably requires a careful control over the size of $\mathcal{P}_{RV_\alpha}$. Indeed, the classes of distributions we are considering are defined through the asymptotic properties of their elements, while to obtain minimax results we need a control for each $n \in \mathbb{N}$. Even though Proposition 3.2 does not directly provide with the minimax rate of the Good-Turing estimator, it still provides with a sanity check for its asymptotic risk. Specifically, Proposition 3.2 implies that for every $P \in \mathcal{P}_{PL_\alpha}$,

$$\limsup_n n^{\alpha/2}\mathbb{E}_P(L(\hat{GT}(\mathbf{X}_n), M_n(P,\mathbf{X}_n))) < +\infty.$$

Moreover, from a minor change at the beginning of the proof of Theorem 3.2, we can also prove that for every estimator $\hat{M}_n(\mathbf{X}_n)$ and every sequence $(T_n)_n$ diverging to infinity, we can find an element $P \in \mathcal{P}_{PL_\alpha}$ such that one has $\limsup_n T_n n^{\alpha/2}\mathbb{E}_P(L(\hat{M}_n(\mathbf{X}_n), M_n(P,\mathbf{X}_n))) = +\infty$. This observation leads us to conjecture that the Good-Turing estimator provides with a rate optimal minimax estimator under the loss function (1.2).

4. Discussion

In this paper, we have considered the problem of consistent estimation of the missing mass under a multiplicative loss function. We have presented an alternative, and remarkably shorter, proof of the main result of Mossel and Ohannessian [24] on the impossibility of a distribution-free estimation of the missing mass. Our results relies on novel arguments from Bayesian nonparametric statistics, which are then exploited to study convergence rates and minimax rates of the Good-Turing estimator under the class of $\alpha \in (0,1)$ regularly varying P. In Proposition 3.1 and Theorem 3.1 it has been shown that, within the class $\mathcal{P}_{PL_\alpha}$, the Good-Turing estimator achieves the best convergence rate possible, while for the class, $\mathcal{P}_{RV_\alpha}$, this rate is the best up to a slowly varying function. An open problem is to understand weather this additional slowly varying term is intrinsic to the problem or our results can actually be improved to make the rate of the Good-Turing estimator matches the best possible rate also within the class of regularly varying distributions. Under the restricted parametric spaces, in Theorem 3.2 we have provided a lower bound for the asymptotic risk. This bound can be used to compare estimators from a minimax point of view, by finding suitable upper bounds matching the lower bound rate. In particular, in Proposition 3.2 we have shown that the asymptotic rate of the risk of the Good-Turing estimator matches the lower bound rate, up to a slowly varying function. However, the rate of Proposition 3.2 is a pointwise result, for a fixed $P \in \mathcal{P}_{RV_\alpha}$. An open problem is to extend Proposition 3.2 to the uniform case, when considering the supremum of the risk over all $P \in \mathcal{P}_{RV_\alpha}$. This extension probably requires a careful analysis and control of the size of this parameter space $\mathcal{P}_{RV_\alpha}$. Work on this is ongoing.

Appendix A

Fact A.1. *For $\epsilon < 1/2$ and $a, b \geq 0$, $L(a,b) \leq \epsilon$ implies $L(b,a) \leq 2\epsilon$, where L denotes the multiplicative loss function.*

Proof. Let a, b be positive real numbers, $\epsilon < 1/2$ and suppose $L(a,b) = |\frac{a}{b} - 1| \leq \epsilon$. Straightforwardly, we have that

$$-b\epsilon \leq a - b \leq b\epsilon \tag{A.1}$$

From the lower bound of (A.1), $a \geq (1-\epsilon)b \geq b/2$ and, therefore, $\frac{1}{a} \leq \frac{2}{b}$. Multiplying (A.1) by this last inequality, we conclude $L(b,a) \leq 2\epsilon$. □

Proposition A.1. *The missing mass $M_n(P, \boldsymbol{X}_n)$ is a jointly measurable map.*

Proof. Recall that $\mathcal{P}$ is endowed with the smallest σ-algebra making the mappings $P \mapsto P(A)$ measurable for every $A \in \mathcal{B}([0,1])$. This is also the Borel σ-algebra generated by the weak convergence topology, which can be induced by the *bounded Lipschitz metric* (see Appendix A of Ghosal and Van der Vaart [12]), defined as

$$d_{BL}(P,Q) = \sup_{\|f\|_{\mathcal{C}^1} \leq 1} \left| \int f dP - \int f dQ \right|$$

where the supremum is over all real functions satisfying $|f(x) - f(y)| \leq |x - y|$ for any $x, y \in [0,1]$. The product space $[0,1]^n$ is supposed to be endowed with the Euclidean topology, which can be induced by the ℓ_∞ norm.

Let us consider $O_n(P, \mathbf{X}_n) = 1 - M_n(P, \mathbf{X}_n)$ and define, for any $\eta > 0$, the function

$$f_{\eta, X_i}(x) := \max(0, \eta - |X_i - x|),$$

in addition put $f_{\eta, \mathbf{X}_n} := \max_i f_{\eta, X_i}$, which is 1−Lipschitz function, since all f_{η, X_i} are 1−Lipschitz. Now, let $O_{\eta,n}$ be defined as follows

$$O_{\eta,n}(P, \mathbf{X}_n) = \frac{1}{\eta} \sum_{j \geq 1} p_j f_{\eta, \mathbf{X}_n}(\theta_j) = \int \frac{1}{\eta} f_{\eta, \mathbf{X}_n} dP.$$

Consider a point $(P, \mathbf{X}_n) \in \mathcal{P} \times [0,1]^n$, we have that

$$\lim_{\eta \to 0} O_{\eta,n}(P, U_{1:n}) = O_n(P, U_{1:n}). \tag{A.2}$$

Indeed, for any $x \in [0,1]$ one has

$$\frac{f_{\eta, \mathbf{X}_n}(x)}{\eta} > 0 \Leftrightarrow \exists i \in \{1, .., n\}, \ |X_i - x| < \eta,$$

hence, if $x \notin \mathbf{X}_n$, $\lim_{\eta \to 0} \frac{f_{\eta, \mathbf{X}_n}(x)}{\eta} = 0$, whereas $\frac{f_{\eta, \mathbf{X}_n}(X_i)}{\eta} = 1$ for any η and i. Finally, since $\frac{f_{\eta, \mathbf{X}_n}(X_i)}{\eta} \leq 1$, the dominated convergence theorem gives (A.2). We prove that O_n is measurable, and so does M_n, by showing that $O_{\eta,n}$ are continuous functions.

Let $\epsilon > 0$, and consider $\mathbf{X}_n, \mathbf{Y}_n \in [0,1]^n$ such that $\|\mathbf{X}_n - \mathbf{Y}_n\|_\infty \leq \eta\epsilon/2$. Furthermore suppose that $P, Q \in \mathcal{P}$ with $d_{BL}(P,Q) \leq \eta\epsilon/2$. For any $x \in [0,1]$, one has

$$|f_{\eta, \mathbf{X}_n}(x) - f_{\eta, \mathbf{Y}_n}(x)| \leq \eta\epsilon/2.$$

Indeed, suppose for instance $f_{\eta,\mathbf{X}_n}(x) \geq f_{\eta,\mathbf{Y}_n}(x)$, with $f_{\eta,\mathbf{X}_n}(x) > 0$. Now consider X_i the closest point to x, we have

$$f_{\eta,\mathbf{X}_n}(x) = \eta - |X_i - x| \leq \eta - |Y_i - x| + \eta\epsilon/2 = f_{\eta,\mathbf{Y}_n}(x) + \eta\epsilon/2.$$

Finally, let us compute the distance between the two images,

$$\begin{aligned}
\eta\,|O_{\eta,n}(P,\mathbf{X}_n) - O_{\eta,n}(Q,\mathbf{Y}_n)| &= \left|\int f_{\eta,\mathbf{X}_n} dP - \int f_{\eta,\mathbf{Y}_n} dQ\right| \\
&\leq \left|\int f_{\eta,\mathbf{X}_n} dP - \int f_{\eta,\mathbf{X}_n} dQ\right| + \left|\int f_{\eta,\mathbf{X}_n} dQ - \int f_{\eta,\mathbf{Y}_n} dQ\right| \\
&\leq \left|\int f_{\eta,\mathbf{X}_n} dP - \int f_{\eta,\mathbf{X}_n} dQ\right| + \|f_{\eta,\mathbf{X}_n} - f_{\eta,\mathbf{Y}_n}\|_\infty \leq \eta\epsilon
\end{aligned}$$

which gives

$$|O_{\eta,n}(P,\mathbf{X}_n) - O_{\eta,n}(Q,\mathbf{Y}_n)| \leq \epsilon.$$

Therefore $O_{\eta,n}$ is continuous and hence measurable. Finally we conclude that M_n is measurable since it is the limit of the sum of measurable functions. □

Proposition A.2. *Let $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ be a sample such that $X_i|P \overset{iid}{\sim} P$ for all $1 \leq i \leq n$. The following distributional results hold true:*

i) if $P \sim DP(\gamma)$, where $\gamma(d\theta) = \mathbb{1}(0 < d\theta < 1)$, then $M_n(\boldsymbol{X}_n, P)|\boldsymbol{X}_n \sim \text{Beta}(1, n)$;
ii) if $P \sim SP_\alpha$, then $M_n(\boldsymbol{X}_n, P)|\boldsymbol{X}_n \sim \text{Beta}(\alpha K_n(\boldsymbol{X}_n), n - \alpha K_n(\boldsymbol{X}_n))$.

Proof. We are going the derive the posterior distribution of $M_n(\mathbf{X}_n, P)$ when P is distributed according to the law of a two-parameter Poisson-Dirichlet process (Pitman and Yor [34]), $P \sim \text{PY}(\eta, \alpha)$, with $\alpha < 1$ and $\eta > -\alpha$. Point i) in the statement is the particular case $\text{PY}(1, 0)$, while point ii) corresponds to $\text{PY}(0, \alpha)$.

From Corollary 20 of Pitman [32], the posterior distribution of P given $\mathbf{X}_n$ under the two-parameter Poisson-Dirichlet process satisfies the following distributional equality

$$P|\mathbf{X}_n \overset{d}{=} \sum_{i=1}^{K_n} w_i \delta_{X_i^*} + w_0 \tilde{P},$$

where $(X_1^*, \ldots, X_{K_n}^*)$ are K_n the distinct values in the sample $\mathbf{X}_n$ and having multiplicities $(n_1, \ldots, n_{K_n})$, $w = (w_0, w_1, \ldots, w_{K_n})$ is a random vector distributed according to a Dirichlet distribution $\text{Dir}(\eta + K_n\alpha, n_1 - \alpha, \ldots, n_{K_n} - \alpha)$ and $\tilde{P} \sim \text{PY}(\alpha, \eta + K_n\alpha)$ independent of w. Therefore,

$$M_n(\mathbf{X}_n, P) = P(\{\mathbf{X}_n\}^c)|\mathbf{X}_n \overset{d}{=} \sum_{i=1}^{K_n} w_i \delta_{X_i^*}(\{\mathbf{X}_n\}^c) + w_0 \tilde{P}(\{\mathbf{X}_n\}^c). \tag{A.3}$$

The point masses in (A.3) are all equal to zero, while $\tilde{P}(\{\mathbf{X}_n\}^c) = 1$ since the base measure of $\tilde{P}$ is diffuse. Therefore, $M_n(\mathbf{X}_n, P) \overset{d}{=} w_0$ and w_0 is distributed according to $\text{Beta}(\eta + K_n(\mathbf{X}_n)\alpha, n - \alpha K_n(\mathbf{X}_n))$ from the aggregation property of the Dirichlet distribution. □

Lemma A.1. *Let $Z \sim \text{Beta}(1, n)$ and $\epsilon > 0$. Then, for all $x \geq 0$ and $n \geq 2$, $\mathbb{P}\left(\left|\frac{Z}{x} - 1\right| > \epsilon\right) \geq 1 - \frac{2\epsilon}{(1-\epsilon)}$.*

Proof. First let us consider $x \in (0, \frac{1}{1+\epsilon}]$ and $n \geq 2$, that is

$$\begin{aligned}
\mathbb{P}\left(\left|\frac{Z}{x} - 1\right| > \epsilon\right) &= \mathbb{P}(Z > (1+\epsilon)x) + \mathbb{P}(Z < (1-\epsilon)x) \\
&= 1 + (1-(1+\epsilon)x)^n - (1-(1-\epsilon)x)^n \\
&= 1 - 2x\epsilon \sum_{k=0}^{n-1} (1-(1+\epsilon)x)^{n-1-k}(1-(1-\epsilon)x)^k \qquad \text{(A.4)} \\
&\geq 1 - 2x\epsilon \sum_{k=0}^{n-1} (1-(1-\epsilon)x)^{n-1} = 1 - 2x\epsilon n(1-(1-\epsilon)x)^{n-1} \\
&\geq 1 - 2\frac{\epsilon}{(1-\epsilon)} n(1-\epsilon)x(1-(1-\epsilon)x)^{n-1} \\
&\geq 1 - 2\frac{\epsilon}{(1-\epsilon)}(1-1/n)^{n-1} \qquad \text{(A.5)} \\
&\geq 1 - \frac{2\epsilon}{(1-\epsilon)}
\end{aligned}$$

where we have used $a^n - b^n = (a-b)\sum_{k=0}^{n-1} a^{n-1-k}b^k$ in (A.4) and that the maximum of the function $x \mapsto x(1-x)^{n-1}$ is achieved for $x = 1/n$ in (A.5). Now let $x > \frac{1}{1+\epsilon}$, noticing that $\frac{2\epsilon}{(1+\epsilon)} < 1$, we conclude that

$$\begin{aligned}
\mathbb{P}\left(\left|\frac{Z}{x} - 1\right| > \epsilon\right) &= \mathbb{P}\left(Z < (1-\epsilon)x\right) \geq \mathbb{P}\left(Z < \frac{1-\epsilon}{1+\epsilon}\right) = 1 - 2^n \frac{\epsilon^n}{(1+\epsilon)^n} \\
&\geq 1 - \frac{2\epsilon}{(1+\epsilon)} \geq 1 - \frac{2\epsilon}{(1-\epsilon)}
\end{aligned}$$

which proves the result. □

Lemma A.2. *Let M' be a random variable distributed according to* $\mathrm{Beta}(a-1, b)$ *and* $\mathrm{med}(M')$ *denote its median. Let $f_{a,b}$ and $m_{a,b}$ denote the density function and the median of a Beta distribution with parameters a and b, respectively. Then, the following equality holds true*

$$\frac{\mathcal{B}(a-1,b)}{\mathcal{B}(a,b)} \mathbb{E}_{M'}\left(|M' - m_{a-1,b}|\right) = 2\int_{m_{a-1,b}}^{m_{a,b}} f_{a,b}(x)dx. \qquad \text{(A.6)}$$

Proof. We start by computing the expected value on the l.h.s. of (A.6):

$$\begin{aligned}
\mathbb{E}_{M'}\left(|M' - m_{a-1,b}|\right) &= \int_0^1 |x - m_{a-1,b}| f_{a-1,b}(x)dx \\
&= \int_0^{m_{a-1,b}} |x - m_{a-1,b}| f_{a-1,b}(x)dx + \int_{m_{a-1,b}}^1 |x - m_{a-1,b}| f_{a-1,b}(x)dx \\
&= -\int_0^{m_{a-1,b}} (x - m_{a-1,b}) f_{a-1,b}(x)dx + \int_{m_{a-1,b}}^1 (x - m_{a-1,b}) f_{a-1,b}(x)dx \\
&= \int_{m_{a-1,b}}^1 x f_{a-1,b}(x)dx - \frac{1}{2}m_{a-1,b} - \int_0^{m_{a-1,b}} x f_{a-1,b}(x)dx + \frac{1}{2}m_{a-1,b} \\
&= \int_{m_{a-1,b}}^1 x f_{a-1,b}(x)dx - \int_0^{m_{a-1,b}} x f_{a-1,b}(x)dx.
\end{aligned}$$

Thanks to the previous chain of equalities we obtain

$$\begin{aligned}
&\frac{\mathcal{B}(a-1,b)}{\mathcal{B}(a,b)}\mathbb{E}_{M'}\left(|M'-m_{a-1,b}|\right)\\
&\quad=\frac{\mathcal{B}(a-1,b)}{\mathcal{B}(a,b)}\int_{m_{a-1,b}}^{1}xf_{a-1,b}(x)dx-\frac{\mathcal{B}(a-1,b)}{\mathcal{B}(a,b)}\int_{0}^{m_{a-1,b}}xf_{a-1,b}(x)dx\\
&\quad=\int_{m_{a-1,b}}^{1}f_{a,b}(x)dx-\int_{0}^{m_{a-1,b}}f_{a,b}(x)dx\\
&\quad=\int_{m_{a-1,b}}^{m_{a,b}}f_{a,b}(x)dx+\int_{m_{a,b}}^{1}f_{a,b}(x)dx-\int_{0}^{m_{a,b}}f_{a,b}(x)dx-\int_{m_{a,b}}^{m_{a-1,b}}f_{a,b}(x)dx\\
&\quad=\int_{m_{a-1,b}}^{m_{a,b}}f_{a,b}(x)dx+\frac{1}{2}-\frac{1}{2}-\int_{m_{a,b}}^{m_{a-1,b}}f_{a,b}(x)dx\\
&\quad=\int_{m_{a-1,b}}^{m_{a,b}}f_{a,b}(x)dx+\int_{m_{a-1,b}}^{m_{a,b}}f_{a,b}(x)dx=2\int_{m_{a-1,b}}^{m_{a,b}}f_{a,b}(x)dx
\end{aligned}$$

and the result now follows. □

Lemma A.3. *Let $f_{a,b}$ denote the density of the* Beta(a,b) *distribution. Then, there exists $n_0\in\mathbb{N}$ such that for all $b>a>n_0$, one has*

$$\sup_{x\in[0,1]}f_{a,b}(x)<8(a+b)^{3/2}a^{-1/2}b^{-1/2}. \tag{A.7}$$

Proof. Suppose that $a,b>2$. The mode of the Beta distribution is $\frac{a-1}{a+b-2}$, therefore

$$\sup_{x\in[0,1]}f_{a,b}(x)=f_{a,b}\left(\frac{a-1}{a+b-2}\right)\leq\frac{1}{\mathcal{B}(a,b)}\frac{a^{a-1}b^{b-1}}{(a+b-2)^{a+b-2}}. \tag{A.8}$$

From the Stirling's formula, there exists n_0 such that for $a,b>n_0$,

$$\frac{1}{\mathcal{B}(a,b)}\leq\frac{(a+b)^{a+b-1/2}}{a^{a-1/2}b^{b-1/2}}.$$

Exploiting the previous inequality to upper bound the r.h.s. in (A.8), we obtain

$$\sup_{x\in[0,1]}f_{a,b}(x)\leq a^{-1/2}b^{-1/2}(a+b)^{3/2}\frac{1}{(1-\frac{2}{a+b})^{a+b}}.$$

Finally, since for n_0 large enough $\frac{1}{(1-\frac{2}{a+b})^{a+b}}\leq 8$, we get the result (A.7). □

Lemma A.4. *Let $f_{a,b}$ denote the density of the* Beta(a,b) *distribution. Then, there exists a constant C such that for any $a,b>3$ with $a<b/2$,*

$$\int_{m_{a-1,b}}^{m_{a,b}}f_{a,b}(x)dx\geq\frac{C}{\sqrt{a}}.$$

Proof. In the sequel we use the generic notation C to refer to universal constants, the value of C may change from a line to another. To simplify the notations, let $n=a+b$ and denote by $I_x(a,b)$ the normalized incomplete Beta function which is defined as

$$I_x(a,b)=\int_0^x f_{a,b}(x)dx.$$

It is well-known that

$$I_x(a+1,b) = I_x(a,b) - \frac{x^a(1-x)^b}{a\mathcal{B}(a,b)},$$

as can be checked through the integration by parts formula. Successively, applying this result for $x = m_{a-1,b}$ and using the definition of the median, we can deduce that

$$\begin{aligned} I_{m_{a-1,b}}(a,b) &= I_{m_{a-1,b}}(a-1,b) - \frac{m_{a-1,b}^{a-1}(1-m_{a-1,b})^b}{(a-1)\mathcal{B}(a-1,b)} = 1/2 - \frac{m_{a-1,b}^{a-1}(1-m_{a-1,b})^b}{(a-1)\mathcal{B}(a-1,b)} \\ &= I_{m_{a,b}}(a,b) - \frac{m_{a-1,b}^{a-1}(1-m_{a-1,b})^b}{(a-1)\mathcal{B}(a-1,b)}, \end{aligned}$$

which in turn leads to

$$\int_{m_{a-1,b}}^{m_{a,b}} f_{a,b}(x)dx = I_{m_{a,b}}(a,b) - I_{m_{a-1,b}}(a,b) = \frac{m_{a-1,b}^{a-1}(1-m_{a-1,b})^b}{(a-1)\mathcal{B}(a-1,b)}. \tag{A.9}$$

Thanks to the Stirling formula applied to the Beta function, we know that for all $a, b > 1$,

$$\mathcal{B}(a,b) \leq C\frac{a^{a-1/2}b^{b-1/2}}{(a+b)^{a+b-1/2}} = C\frac{a^{a-1/2}b^{b-1/2}}{n^{n-1/2}}.$$

By plugging the previous formula in (A.9), we obtain

$$\int_{m_{a-1,b}}^{m_{a,b}} f_{a,b}(x)dx \geq C\frac{(n-1)^{n-3/2}}{(a-1)^{a-1/2}b^{b-1/2}} m_{a-1,b}^{a-1}(1-m_{a-1,b})^b. \tag{A.10}$$

Now, since $a - 1 < b$, the mode-median-mean inequality (see Payton et al. [31]) gives that

$$\frac{a-2}{n-3} \leq m_{a-1,b} \leq \frac{a-1}{n-1},$$

from which we deduce

$$m_{a-1,b}^{a-1} \geq \frac{(a-2)^{a-1}}{(n-3)^{a-1}} \geq \frac{(a-2)^{a-1}}{(n-1)^{a-1}}$$

and

$$(1-m_{a-1,b})^b \geq \frac{b^b}{(n-1)^b}.$$

Now, together with (A.10), the previous two inequalities yield

$$\int_{m_{a-1,b}}^{m_{a,b}} f_{a,b}(x)dx \geq C\frac{(n-1)^{n-3/2}}{(a-1)^{a-1/2}b^{b-1/2}}\frac{(a-2)^{a-1}b^b}{(n-1)^{n-1}} \geq C\sqrt{\frac{b}{an}}$$

where we used the fact that for $x > 2$, $x^x \geq (x-1)^x \geq Cx^x$. The result follows by noticing that $a < b/2$ implies $b/n < 2/3$. □

Acknowledgements

The authors thank an anonymous Referee and an Associate Editor for all her/his comments, corrections, and suggestions which remarkably improved the original version of the paper. Federico Camerlenghi and Stefano Favaro received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 817257. Federico Camerlenghi and 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] J. Beirlant and L. Devroye. On the impossibility of estimating densities in the extreme tail. *Statist. Probab. Lett.* **43**, (1999) 57–64.

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

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

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

[5] S. Bubeck, D. Ernst and A. Garivier. Optimal discovery with probabilistic expert advice: finite time analysis and macroscopic optimality. *J. Mach. Learn. Res.* **14**, (2013) 601–623.

[6] S. Chatterjee and P. Diaconis. The sample size required in importance sampling. *Ann. Appl. Probab.* **28**, (2018) 1099–1135.

[7] A. Clauset, C.R. Shalizi and M.E.J. Newman. Power-law Distributions in Empirical Data. *SIAM Rev.* **51**, (2009) 661–703.

[8] T.S. Ferguson. A Bayesian Analysis of Some Nonparametric Problems. *Ann. Statist.* **1**, (1973) 209–230.

[9] R.A. Fisher, A.S. Corbet and C.B. Williams. The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.* **12**, (1943) 42–58.

[10] F. Gao. Moderate deviations for a nonparametric estimator of sample coverage. *Ann. Statist.* **41**, (2013) 641–669.

[11] Z. Gao, C.H. Tseng, Z. Pei and M.J. Blaser. Molecular analysis of human forearm superficial skin bacterial biota. *Proc. Natl. Acad. Sci. USA* **104**, (2007) 2927–2932.

[12] S. Ghosal and A. Van der Vaart. *Fundamentals of nonparametric Bayesian inference.* Cambridge University Press, 2017.

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

[14] S. Goldwater, T. Griffiths and M. Johnson. Interpolating between types and tokens by estimating power-law generators. In *Proceedings of the Conference in Advances in Neural Information Processing Systems* (2006).

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

[16] M. Grabchak and Z. Zhang. Asymptotic properties of Turing's formula in relative error. *Mach. Learn.* **106**, (2017) 1771–1785.

[17] I. Ionita-Laza, C. Lange and N.M. Laird. Estimating the number of unseen variants in the human genome. *Proc. Natl. Acad. Sci. USA* **106**, (2009) 5008–5013.

[18] S. Karlin. Central limit theorems for certain infinite urn schemes. *J. Math. Mech.* **17**, (1967) 373–401.

[19] J.F.K. Kingman. Random Discrete Distributions. *J. Roy. Statist. Soc. Ser. B* **37**, (1975) 1–22.

[20] I. Kroes, P.W. Lepp and D.A. Relman. Bacterial diversity within the human subgingival crevice. *Proc. Natl. Acad. Sci. USA* **96**, (1999) 14547–14552.

[21] D. McAllester and L. Ortiz. Concentration inequalities for the missing mass and for histogram rule error. *J. Mach. Learn. Res.* **4**, (2003) 895–911.

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

[23] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics* **1**, (2004) 226–251.

[24] E. Mossel and M.I. Ohannessian. On the impossibility of learning the missing mass. *Entropy* **21**(1), 28 (2019).

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

[26] M.E.J. Newman. The Structure and Function of Complex Networks. *SIAM Rev.* **45**, (2003) 167–256.

[27] M.I. Ohannessian and M.A. Dahleh. Rare probability estimation under regularly varying heavy tails. *J. Mach. Learn. Res.* **23**, (2012) 1–24.

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

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

[30] A. Orlitsky, A.T. Suresh and Y. Wu. Optimal prediction of the number of unseen species. *Proc. Natl. Acad. Sci. USA* **113**, (2016) 13283–13288.

[31] M.E. Payton, L.J. Young and J.H. Young. Bounds for the difference between median and mean of beta and negative binomial distributions. *Metrika* **36**, (1989) 347–354.

[32] J. Pitman. Some developments of the Blackwell-MacQueen urn scheme. In *Statistics, Probability and Game Theory*, Institute of Mathematical Statistics (1996).

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

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

[35] N. Rajaraman, A. Thangaraj and A.T. Suresh. Minimax Risk for Missing Mass Estimation. In *Proceedings of the IEEE International Symposium on Information Theory* (2017).

[36] D. Sornette. *Critical Phenomena in Natural Sciences.* Springer (2006).

[37] B. Wagner, P. Viswanath and S.R. Kulkarni. Strong consistency of the Good-Turing estimator. In *Proceedings of the IEEE International Symposium on Information Theory* (2017).

[38] C.H. Zhang and Z. Zhang. Asymptotic normality of a nonparametric estimator of sample coverage. *Ann. Statist.* **37**, (2009) 2582–2595.