Bernoulli **24**(3), 2018, 1910–1941
https://doi.org/10.3150/16-BEJ915

Finite sample properties of the mean occupancy counts and probabilities

GEOFFREY DECROUEZ[1], MICHAEL GRABCHAK[2] and QUENTIN PARIS[3]

[1]*Higher School of Economics, National Research University, Moscow, Russia. E-mail: ggdecrouez@hse.ru*
[2]*Department of Mathematics and Statistics, University of North Carolina, Charlotte, NC, USA. E-mail: mgrabcha@uncc.edu*
[3]*Higher School of Economics & Laboratory of Stochastic Analysis and its Applications, National Research University, Moscow, Russia. E-mail: qparis@hse.ru*

For a probability distribution P on an at most countable alphabet $\mathcal{A}$, this article gives finite sample bounds for the expected occupancy counts $\mathbb{E}K_{n,r}$ and probabilities $\mathbb{E}M_{n,r}$. Both upper and lower bounds are given in terms of the counting function ν of P. Special attention is given to the case where ν is bounded by a regularly varying function. In this case, it is shown that our general results lead to an optimal-rate control of the expected occupancy counts and probabilities with explicit constants. Our results are also put in perspective with Turing's formula and recent concentration bounds to deduce bounds in probability. At the end of the paper, we discuss an extension of the occupancy problem to arbitrary distributions in a metric space.

Keywords: counting measure; finite sample bounds; occupancy problem; regular variation; Turing's formula; urn scheme

1. Introduction

The occupancy problem

From a general point of view, the occupancy problem – also referred to as the urn scheme – is to describe the spread of a random sample drawn from a probability distribution supported by an at most countable alphabet. In the literature, this task is usually carried out by studying the so-called occupancy counts and occupancy probabilities – also known as rare probabilities – defined below. Interest for the occupancy problem arises in many practical situations such as Ecology (Good and Toulmin [16], Chao [5]), Genomics (Mao and Lindsay [25]), Language Processing (Chen and Goodman [9]), Authorship Attribution (Efron and Thisted [10], Thisted and Efron [34], Zhang and Huang [38]), Information Theory (Orlitsky, Santhanam and Zhang [30]) and Computer Science (Zhang [35]).

Consider an at most countable alphabet $\mathcal{A}$ with an associated probability distribution $P = \{p_a : a \in \mathcal{A}\}$, where $p_a \in [0, 1]$ and $\sum_{a\in\mathcal{A}} p_a = 1$. Let $\mathcal{S} = \{a \in \mathcal{A} : p_a > 0\}$ denote the support of P, and let $X_1, \ldots, X_n$ be independent and identically distributed $\mathcal{A}$-valued random variables,

defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with distribution P. For all $a \in \mathcal{A}$, we set

$$\xi_n(a) = \sum_{i=1}^{n} \mathbf{1}\{X_i = a\}, \tag{1.1}$$

where the notation $\mathbf{1}\{\cdots\}$ stands for the indicator function of the event $\{\cdots\}$. For all integers $0 \le r \le n$, the occupancy counts $K_{n,r}$ and occupancy probabilities $M_{n,r}$ are defined, respectively, by

$$K_{n,r} = \sum_{a \in \mathcal{A}} \mathbf{1}\big\{\xi_n(a) = r\big\} \quad \text{and} \quad M_{n,r} = \sum_{a \in \mathcal{A}} p_a \mathbf{1}\big\{\xi_n(a) = r\big\}. \tag{1.2}$$

For any integer $0 \le r \le n$, the random variable $K_{n,r}$ stands for the number of points in $\mathcal{A}$ represented exactly r times in the sample. A clear interpretation of the occupancy probabilities is given by the following equivalent representation. Introducing a generic $\mathcal{A}$-valued random variable X, independent of the sample and distributed according to P, we have, almost surely,

$$M_{n,r} = \mathbb{P}\big(\xi_n(X) = r | X_1, \ldots, X_n\big).$$

Hence, for any integer $0 \le r \le n$, $M_{n,r}$ stands for the (conditional) probability that, given the first n observations, the next one will be of a letter that is already represented r times in the sample. The quantity $M_{n,0}$ is particularly important. In the literature it is usually called the missing mass, and has attracted a lot of attention due to its practical interpretation as the probability of novelty. The goal of this paper is to understand the finite sample properties of $\mathbb{E}K_{n,r}$ and $\mathbb{E}M_{n,r}$.

Related work

Following the pioneering work of Karlin [23], it is understood that the asymptotic behavior of the occupancy counts $K_{n,r}$ is strongly connected to the behavior of the tail of the counting measure $\boldsymbol{\nu}$ of P, which is defined on $[0, 1]$ by

$$\boldsymbol{\nu}(\mathrm{d}x) = \sum_{a \in \mathcal{A}} \delta_{p_a}(\mathrm{d}x). \tag{1.3}$$

The function $\nu : [0, 1] \to \mathbb{N}$, defined by

$$\nu(\varepsilon) = \boldsymbol{\nu}\big([\varepsilon, 1]\big), \tag{1.4}$$

is usually referred to as the counting function of P. A short account of some of its basic properties is given in Appendix A. We now illustrate the relationship between the behavior of ν and that of $K_{n,r}$. Toward this end, we recall some terminology from Karlin [23]. We say that a function $f : [0, +\infty) \to \mathbb{R}$ is regularly varying at $x_0 \in \{0, \infty\}$ with exponent $\alpha \in \mathbb{R}$, and we write $f \in \mathrm{rv}_{x_0}^{\alpha}$, if

$$\forall c > 0, \qquad \lim_{x \to x_0} \frac{f(cx)}{f(x)} = c^{\alpha}.$$

If $\alpha = 0$, we say that f is slowly varying at x_0. Note that $f \in \mathrm{rv}_0^\alpha$ if and only if there exists $\ell \in \mathrm{rv}_\infty^0$ such that, for all $\varepsilon > 0$,

$$f(\varepsilon) = \varepsilon^{-\alpha}\ell(1/\varepsilon). \tag{1.5}$$

It is well known that the counting function ν, defined in (1.4), satisfies $\nu(\varepsilon) = \varepsilon^{-\alpha}\ell(1/\varepsilon)$, for some $\alpha \in (0, 1)$ and some $\ell \in \mathrm{rv}_\infty^0$, if and only if

$$\forall r \geq 1: \qquad K_{n,r} \underset{\text{a.s.}}{\sim} \mathbb{E}K_{n,r} \sim \frac{\alpha\Gamma(r-\alpha)}{r!} n^\alpha \ell(n), \tag{1.6}$$

as $n \to +\infty$. (Here and throughout, for any two real-valued functions g and h and any $x_0 \in [0, +\infty]$, we write $h(x) \sim g(x)$ as $x \to x_0$ if and only if $h(x)/g(x) \to 1$ as $x \to x_0$.) For a detailed exposition and developments on this topic, we refer the reader to the classic text by Johnson and Kotz [21] or the more recent, and very complete, survey by Gnedin, Hansen and Pitman [14], which, in particular, studies extensions of (1.6) to the case $\alpha \in \{0, 1\}$ under additional care.

In the same spirit, Ohannessian and Dahleh [29] extended (1.6) to the case of occupancy probabilities proving that, if $\nu(\varepsilon) = \varepsilon^{-\alpha}\ell(1/\varepsilon)$ for some $\alpha \in (0, 1)$ and some $\ell \in \mathrm{rv}_\infty^0$, then

$$\forall r \geq 0: \qquad M_{n,r} \underset{\text{a.s.}}{\sim} \mathbb{E}M_{n,r} \sim \frac{\alpha\Gamma(1+r-\alpha)}{r!} n^{\alpha-1} \ell(n), \tag{1.7}$$

as $n \to +\infty$. While the second asymptotic equivalence in (1.7) is, as mentioned by the authors, easily derived from (1.6) and the relation

$$\mathbb{E}M_{n,r} = \left(\frac{1+r}{1+n}\right)\mathbb{E}K_{n+1,r+1}, \tag{1.8}$$

the first asymptotic equivalence in (1.7) is established by Ohannessian and Dahleh [29] by proving more powerful concentration properties of $M_{n,r}$ around its expectation.

Some of the first concentration properties in this context were established by McAllester and Schapire [27] for the missing mass $M_{n,0}$. The concentration properties of the missing mass have since been investigated by McAllester and Ortiz [26], Ohannessian and Dahleh [28], Berend and Kontorovich [4] and Khanloo and Haffari [24]. Many extensions and new results concerning the concentration properties of the occupancy counts $K_{n,r}$ and occupancy probabilities $M_{n,r}$ can be found in Ohannessian and Dahleh [29] and Ben-Hamou, Boucheron and Ohannessian [2].

Establishing concentration properties is a fundamental step toward understanding the finite sample behavior of the occupancy counts and probabilities. However, a full understanding of the finite sample properties of $K_{n,r}$ and $M_{n,r}$ requires finite sample bounds for their expectations. We are only aware of two contributions in this direction, namely Ohannessian and Dahleh [28] and Berend and Kontorovich [3], which both focus on the missing mass. In particular, Ohannessian and Dahleh [28] introduce the accrual function $F(x) = P(\{a : p_a \leq x\})$, and show that

$$\sup_{0\leq\varepsilon\leq 1}\left\{(1-\varepsilon)^n F(\varepsilon)\right\} \leq \mathbb{E}M_{n,0} \leq \inf_{0\leq\varepsilon\leq 1}\left\{(1-\varepsilon)^n + F(\varepsilon)\right\}. \tag{1.9}$$

It should be noted that, as described in Appendix B, this result yields, in many cases, explicit bounds with almost optimal rates of convergence. In Berend and Kontorovich [3], the authors show that, in the finite support case,

$$\forall n \leq |\mathcal{S}|: \qquad \mathbb{E}M_{n,0} \leq e^{-n/|\mathcal{S}|} \quad \text{and} \quad \forall n > |\mathcal{S}|: \qquad \mathbb{E}M_{n,0} \leq \frac{|\mathcal{S}|}{ne},$$

while in the infinite support case, there exists a universal constant $c > 0$ such that

$$\mathbb{E}M_{n,0} \leq \frac{L(P)}{cn}, \qquad \text{where } L(P) = \sup_{0<\varepsilon<1} \big\{\nu(\varepsilon/2) - \nu(\varepsilon)\big\}.$$

In addition, the authors prove that, for any integer $a > 1$, there exists a distribution P for which $L(P) = a$ and $\mathbb{E}M_{n,0} \geq c'a/n$, where $c' > 0$ denotes a universal constant. Hence, their bound is shown to be sharp for a certain class of probability distributions. Unfortunately, $L(P) = +\infty$ in many interesting cases, including when $\mathcal{A} = \{1, 2, \dots\}$ and, for some $\alpha \in (0, 1)$, the distribution P has masses $p_k = Ck^{-1/\alpha}, k \in \mathcal{A}$.

Concerning lower bounds, there are interesting results from a somewhat different perspective given in Lemma 4.1 of Almudevar, Bhattacharya and Sastri [1] and Lemma 1 of Zhang [37]. These are discussed, in detail, in Appendix C.

Contribution and organisation of the paper

Building on the previous work from Ohannessian and Dahleh [28] and Berend and Kontorovich [3], this paper establishes finite sample upper and lower bounds for the expected occupancy counts $\mathbb{E}K_{n,r}$ and the expected occupancy probabilities $\mathbb{E}M_{n,r}$ for arbitrary $n \geq 1$ and arbitrary $0 \leq r \leq n$. For simplicity of exposition, focus is put on the expected occupancy probabilities $\mathbb{E}M_{n,r}$ knowing that relation (1.8) immediately implies similar bounds for the expected occupancy counts. Section 2 is devoted to our main results. We first give general bounds in terms of the counting function ν, which make no assumptions about the underlying distribution. Additional assumptions on ν are used to derive more explicit bounds. In particular, when the counting function is regularly varying, the bounds are shown to be consistent with (1.6) and (1.7), and are thus rate optimal. Section 3 presents some applications and extensions. Specifically, Section 3.1 discusses the relationship between our results and Turing's formula, while Section 3.2 shows how we can combine our results with recent concentration bounds to derive bounds in probability for $M_{n,r}$ and $K_{n,r}$. Further, in Section 3.3, we present an extension to the case of a random number of observations modelled by a non-homogeneous Poisson process, and in Section 3.4 we discuss an interesting perspective for future research in the context of arbitrary probability measures – i.e. not necessarily discrete – in a metric space. Proofs are postponed to Section 4. Finally, Appendix A collects a few basic properties of the counting function, Appendix B investigates the performance of bounds given in terms of the accrual function, and Appendix C discusses the lower bounds from Almudevar, Bhattacharya and Sastri [1] and Zhang [37].

Notation

Throughout, the notation $\mathbf{1}\{\cdots\}$ stands for the indicator function of the event $\{\cdots\}$. For any set B, we write $|B|$ to denote the number (possibly infinite) of elements in B. For any $t > 0$ and any $x \geq 0$, we denote by

$$\gamma(t, x) = \int_0^x u^{t-1} e^{-u} \, \mathrm{d}u \tag{1.10}$$

the lower incomplete Gamma function. Note that the Gamma function is given by $\Gamma(t) = \gamma(t, +\infty)$.

2. Main results

In this section, we give upper and lower bounds for the expected occupancy probabilities $\mathbb{E}M_{n,r}$. From (1.8) it follows that all of the results in this section can be immediately adapted to the expected occupancy counts $\mathbb{E}K_{n,r}$. However, for ease of exposition, we only report results in terms of the occupancy probabilities.

2.1. Upper bounds

Let $P = \{p_a : a \in \mathcal{A}\}$ be a probability measure on the countable alphabet $\mathcal{A}$. Its counting function ν defined in (1.4), can be equivalently written as

$$\nu(\varepsilon) = \big|\{a \in \mathcal{A} : p_a \geq \varepsilon\}\big|, \qquad 0 \leq \varepsilon \leq 1. \tag{2.1}$$

A short account of the basic properties of ν is given in Appendix A. Our first result provides a general upper bound in terms of ν. In the sequel, we denote

$$c(r) = \begin{cases} e^{-1} & \text{if } r = 0, \\ e(1+r)/\sqrt{\pi} & \text{if } r \geq 1. \end{cases} \tag{2.2}$$

Theorem 2.1. *For any $n \geq 1$ and any $0 \leq r \leq n-1$, we have*

$$\mathbb{E}M_{n,r} \leq \inf_{0 \leq \varepsilon \leq 1} \big\{\varphi^+_{n,r}(\varepsilon) + \psi^+_{n,r}(\varepsilon)\big\}, \tag{2.3}$$

where

$$\varphi^+_{n,r}(\varepsilon) = \frac{c(r)\nu(\varepsilon)}{n},$$

$$\psi^+_{n,r}(\varepsilon) = 2^{1+r}\binom{n}{r}\int_0^\varepsilon \nu\left(\frac{u}{2}\right) u^r \left(1 - \frac{u}{2}\right)^{n-r} \mathrm{d}u.$$

Further, for any $n \geq 1$,

$$\mathbb{E}M_{n,n} \leq \inf_{0 \leq \varepsilon \leq 1} \left\{ p_\star^{n+1} \nu(\varepsilon) + \varepsilon^n \right\}, \tag{2.4}$$

where $p_\star = \max\{p_a : a \in \mathcal{A}\} \in (0, 1]$.

Remark 2.1. The proof of Theorem 2.1 consists in studying separately, and for all $\varepsilon \in [0, 1]$, the contributions of large (i.e., larger then ε) and small (i.e., smaller then ε) probabilities. These contributions are bounded, respectively, by $\varphi_{n,r}^+(\varepsilon)$ and $\psi_{n,r}^+(\varepsilon)$. Details in the proof reveal that the term $\psi_{n,r}^+(\varepsilon)$ can in fact be replaced by the quantity

$$\frac{b^{1+r}}{b-1} \binom{n}{r} \int_0^\varepsilon \nu\left(\frac{u}{b}\right) u^r \left(1 - \frac{u}{b}\right)^{n-r} \mathrm{d}u,$$

for any $b > 1$. In principle, the value of b may be optimized, but for the sake of simplicity, we choose $b = 2$. Note that, since ν is bounded on intervals away from 0, this should not affect the bound in a substantial way.

Observe that, in (2.3), the two terms $\varphi_{n,r}^+(\varepsilon)$ and $\psi_{n,r}^+(\varepsilon)$ have opposite monotonic behaviours in ε. In full generality, the value of ε leading to the optimal tradeoff is not obvious. However, in many interesting cases, a relevant choice of ε yields explicit and, as far as we know, new bounds.

Corollary 2.1. *Suppose that* $\mathcal{S}$ *is finite. Then, for all* $n \geq 1$ *and all* $0 \leq r \leq n-1$,

$$\mathbb{E}M_{n,r} \leq \frac{c(r)|\mathcal{S}|}{n} \quad \text{and} \quad \mathbb{E}M_{n,n} \leq p_\star^{n+1}|\mathcal{S}|,$$

where $c(r)$ *is as in* (2.2).

The proof of Corollary 2.1 simply involves taking $\varepsilon = 0$ in Theorem 2.1 and is therefore omitted. Note that, when we take $r = 0$ in Corollary 2.1, we recover the bound $\mathbb{E}M_{n,0} \leq |\mathcal{S}|/(ne)$ for the expected missing mass provided by Berend and Kontorovich [3]. Next, we study several situations, where P has an infinite support.

Corollary 2.2. *Suppose that* $\mathcal{S}$ *is infinite. Assume that, for* $\alpha \in [0, 1]$ *and* $\ell \in \mathrm{rv}_\infty^0$, *we have* $\nu(\varepsilon) \leq \varepsilon^{-\alpha}\ell(1/\varepsilon)$ *for all* $0 < \varepsilon \leq 1$. *Suppose, in addition, that* ℓ *is non-increasing. Then, for all* $n \geq 2$ *and all* $0 \leq r \leq n-1$, *we have*

$$\mathbb{E}M_{n,r} \leq c_1(\alpha, r) n^{\alpha-1} \ell(n),$$

where

$$c_1(\alpha, r) = c(r) + \frac{4^{1+r}}{r!}(1+r)^{1+r-\alpha} \gamma\left(1 + r - \alpha, \frac{1}{2}\right),$$

$c(r)$ *is as in* (2.2), *and* $\gamma(\cdot, \cdot)$ *denotes the lower incomplete Gamma function defined in* (1.10).

According to (1.7), the bound of Corollary 2.2 is rate optimal in terms of n. In order for the bound to be even more explicit, note that for all $t > 0$ and all $x \geq 0$, the constant $\gamma(t, x)$ may be roughly upper bounded by $t^{-1}x^t$. Observe, finally, that when $\alpha = 1$ and $r = 0$ the bound in Corollary 2.2 is trivial since $\gamma(0, \frac{1}{2}) = +\infty$.

The next corollary studies the case of an arbitrary $\ell \in \mathrm{rv}_\infty^0$. First, let $\ell \in \mathrm{rv}_\infty^0$ and denote, for all $\beta \in (0, 1)$ and all $x \geq 1$,

$$\ell_\beta^\circ(x) = \sqrt{\int_{2x}^{+\infty} \frac{\ell(u)^2}{u^{2-\beta}}\,\mathrm{d}u}. \tag{2.5}$$

Then, one may deduce that $\ell_\beta^\circ \in \mathrm{rv}_\infty^{-(1-\beta)/2}$ and satisfies,

$$\ell_\beta^\circ(x) \sim \frac{\ell(x)}{(2x)^{\frac{1-\beta}{2}}\sqrt{1-\beta}}, \tag{2.6}$$

as $x \to +\infty$, by an application of Karamata's theorem (Karamata [22]). We are now in position to state our next result.

Corollary 2.3. *Suppose that $\mathcal{S}$ is infinite. Assume that, for $\alpha \in [0, 1]$ and $\ell \in \mathrm{rv}_\infty^0$, we have $v(\varepsilon) \leq \varepsilon^{-\alpha}\ell(1/\varepsilon)$ for all $0 < \varepsilon \leq 1$. Then, for all $n \geq 2$, all $0 \leq r \leq n-1$ and all $\beta \in (0, 1)$ with $\beta > 2(\alpha - r) - 1$, we have*

$$\mathbb{E}M_{n,r} \leq c(r)n^{\alpha-1}\ell(n) + c_2(\alpha, \beta, r)n^{\alpha-\frac{1+\beta}{2}}\ell_\beta^\circ(n),$$

where

$$c_2(\alpha, \beta, r) = \frac{4^{1+r}}{r!}\left(\frac{1+r}{2}\right)^{\frac{1+\beta}{2}+r-\alpha}\sqrt{\gamma\big(1+\beta+2(r-\alpha), 1\big)},$$

$c(r)$ is as in (2.2), and $\gamma(\cdot, \cdot)$ denotes the lower incomplete Gamma function defined in (1.10).

Observe that, for every $\beta \in (0, 1)$ with $\beta > 2(\alpha - r) - 1$, this bound is rate optimal according to (1.7) since, using (2.6), we have

$$n^{\alpha-\frac{1+\beta}{2}}\ell_\beta^\circ(n) \sim \frac{n^{\alpha-1}\ell(n)}{2^{\frac{1-\beta}{2}}\sqrt{1-\beta}}, \tag{2.7}$$

as $n \to +\infty$. Hence, the result in Corollary 2.3 differs from that of Corollary 2.2 mainly at the level of constants.

Next, we present an additional result in the spirit of Theorem 2.1, which will shed an interesting light on the lower bounds presented further. This result is less explicit than Theorem 2.1, but allows for tighter upper bounds in certain cases, including when the counting function is regularly varying with exponent $\alpha \in (0, 1]$. First, we introduce the function κ_+ defined, for all

$\varepsilon \in (0, 1]$, by

$$\kappa_+(\varepsilon) = \sup_{0<u\leq\varepsilon} \frac{\nu(u/2)}{\nu(u)}. \tag{2.8}$$

Note that κ_+ is non-decreasing by construction. Also, given that ν is non-increasing, we have $\kappa_+(\varepsilon) \geq 1$ for all $0 < \varepsilon \leq 1$. For simplicity of notation, we write

$$\kappa_+^0 = \lim_{\varepsilon\to 0} \kappa_+(\varepsilon). \tag{2.9}$$

Theorem 2.2. *For any $n \geq 1$ and any $0 \leq r \leq n-1$, we have*

$$\mathbb{E}M_{n,r} \leq \inf_{0<\varepsilon\leq 1/2} \big\{\varphi_{n,r}^+(\varepsilon) + \theta_{n,r}^+(\varepsilon)\big\},$$

where $\varphi_{n,r}^+(\varepsilon)$ is defined in Theorem 2.1 and

$$\theta_{n,r}^+(\varepsilon) = 2^{1+r}\binom{n}{r}\int_0^\varepsilon \big(\kappa_+(2u)-1\big)\nu(u)u^r\left(1-\frac{u}{2}\right)^{n-r} \mathrm{d}u.$$

The monotonicity of ν leads, immediately, to the fact that $(\kappa_+(2u) - 1)\nu(u) \leq (\kappa_+(2u) - 1)\nu(u/2)$, for all $0 < u \leq 1$. As a result, by monotonicity of κ_+, $\theta_{n,r}^+(\varepsilon) \leq \psi_{n,r}^+(\varepsilon)$ for $0 < \varepsilon \leq 1/2$ provided $\kappa_+(2\varepsilon) \leq 2$. The next statement shows that when $\nu \in \mathrm{rv}_0^\alpha$ this condition is always satisfied for ε small enough leading to a potentially tighter bound than Theorem 2.1.

Proposition 2.1. *Suppose that $\nu \in \mathrm{rv}_0^\alpha$, for $\alpha \in [0, 1]$. Then $\kappa_+^0 = 2^\alpha$.*

The proof of Proposition 2.1 follows, almost immediately, from the definition of slowly varying functions at $+\infty$, and is therefore omitted. We end this subsection by the following corollary.

Corollary 2.4. *Suppose that $\mathcal{S}$ is infinite. Assume that $\kappa_+^0 \in (1, 2]$ and that, for $\alpha \in [0, 1]$ and $\ell \in \mathrm{rv}_\infty^0$, we have $\nu(\varepsilon) \leq \varepsilon^{-\alpha}\ell(1/\varepsilon)$ for all $0 < \varepsilon \leq 1$. Then, for all $n \geq 2$ large enough so that*

$$\kappa_+\left(\frac{2}{n}\right) \leq 2\kappa_+^0 - 1, \tag{2.10}$$

all $0 \leq r \leq n-1$ and all $\beta \in (0, 1)$ with $\beta > 2(\alpha - r) - 1$, we have

$$\mathbb{E}M_{n,r} \leq c(r)n^{\alpha-1}\ell(n) + \big(\kappa_+^0 - 1\big)c_2(\alpha, \beta, r)\left(\frac{n}{2}\right)^{\alpha-\frac{1+\beta}{2}} \ell_\beta^\circ\left(\frac{n}{2}\right),$$

where $c(r)$ is as in (2.2), $c_2(\alpha, \beta, r)$ is as given in Corollary 2.3, and $\gamma(\cdot, \cdot)$ denotes the lower incomplete Gamma function defined in (1.10).

According to Proposition 2.1, the assumptions of Corollary 2.4 are satisfied when $\nu \in \mathrm{rv}_0^\alpha$ for $\alpha \in (0, 1]$, in which case $(\kappa_+^0 - 1) \in (0, 1]$ and therefore $(\kappa_+^0 - 1)c_2(\alpha, \beta, r) \leq c_2(\alpha, \beta, r)$. As in

Corollary 2.3, note that the bound is rate optimal, for all $\beta \in (0,1)$ with $\beta > 2(\alpha - r) - 1$, thanks to (2.7). Lastly, observe that, with additional information on ℓ, the range of n for which (2.10) applies can be made explicit.

2.2. Lower bounds

In this subsection, we tackle the problem of finding non-asymptotic lower bounds for the expectation of the occupancy probabilities. For this purpose, we introduce the function κ_- defined, for all $\varepsilon \in (0,1]$, by

$$\kappa_-(\varepsilon) = \sup_{0<u\leq\varepsilon} \frac{\nu(u)}{\nu(u/2)}. \tag{2.11}$$

Note that κ_- is non-decreasing and satisfies $\kappa_-(\varepsilon) \leq 1$. We further define

$$\kappa_-^0 = \lim_{\epsilon\to 0} \kappa_-(\epsilon). \tag{2.12}$$

The following result is in the spirit of Theorem 2.2.

Theorem 2.3. *For any $n \geq 1$ and any $0 \leq r \leq n-1$, we have*

$$\mathbb{E}M_{n,r} \geq \sup_{0<\varepsilon\leq 1/2} \left\{\varphi_{n,r}^-(\varepsilon) + \theta_{n,r}^-(\varepsilon)\right\},$$

where

$$\varphi_{n,r}^-(\varepsilon) = \binom{n}{r} \nu(\varepsilon)\varepsilon^{r+1}(1-p_\star)^{n-r},$$

$$\theta_{n,r}^-(\varepsilon) = 2^{-r}\binom{n}{r} \int_0^\varepsilon \big(1-\kappa_-(2u)\big)\nu(u)u^r(1-2u)^{n-r}\,\mathrm{d}u,$$

and $p_\star = \max\{p_a : a \in \mathcal{A}\}$.

In order for the term $\theta_{n,r}^-(\varepsilon)$ to be strictly positive, there needs to be at least one $\varepsilon \in (0, 1/2]$ with $\kappa_-(2\varepsilon) < 1$. The next proposition indicates that this requirement holds when $\nu \in \mathrm{rv}_0^\alpha$ with $\alpha \in (0,1]$.

Proposition 2.2. *Suppose that $\nu \in \mathrm{rv}_0^\alpha$, for $\alpha \in [0,1]$. Then $\kappa_-^0 = 2^{-\alpha}$.*

The proof of Proposition 2.2 follows, almost immediately, from the definition of slowly varying functions at $+\infty$ and is thus omitted.

Given the monotonicity of κ_- and the fact that $\kappa_-(\varepsilon) \leq 1$, Proposition 2.2 implies that, for $\alpha = 0$, κ_- is identically equal to 1 and the second term of the bound is therefore equal to 0. The reader may easily check that this last observation also holds when $\mathcal{S}$ is finite since, obviously, $\kappa_-(\varepsilon) \to 1$ as $\varepsilon \to 0$ in this case. Thus, the term $\theta_{n,r}^-(\varepsilon)$ contributes to the bound when $\nu \in \mathrm{rv}_0^\alpha$

with $\alpha \in (0, 1]$ and is identically 0 when the support $\mathcal{S}$ is finite or $\nu \in \mathrm{rv}_0^0$. Note, however, that the lower bound is attained for uniform distributions. Indeed, suppose that $2 \leq |\mathcal{S}| < +\infty$ and that P is uniform. Setting $\varepsilon_0 = 1/|\mathcal{S}|$, we have $\nu(\varepsilon_0) = |\mathcal{S}|$, $\kappa_-(\varepsilon_0) = 1$, and

$$\begin{aligned}\mathbb{E}M_{n,r} &= \binom{n}{r} \sum_{k=1}^{|S|} \left(\frac{1}{|S|}\right)^{r+1} \left(1 - \frac{1}{|S|}\right)^{n-r} \\ &= \binom{n}{r} \nu(\varepsilon_0) \varepsilon_0^{r+1} (1 - p_\star)^{n-r} = \sup_{0<\varepsilon\leq 1/2} \varphi_{n,r}^-(\varepsilon).\end{aligned}$$

We end this section with a corollary similar in nature to Corollary 2.4. First recall that, for any $t > 0$ and any $x \geq 0$, we have

$$\left(1 - \frac{x}{n}\right)^n \to e^{-x} \quad \text{and} \quad \int_0^x u^{t-1} \left(1 - \frac{u}{n}\right)^n \mathrm{d}u \to \gamma(t, x),$$

as $n \to +\infty$, where the second limit follows by dominated convergence.

Corollary 2.5. *Suppose that $\kappa_-^0 < 1$ and that, for $\alpha \in [0, 1]$ and $\ell \in \mathrm{rv}_\infty^0$, we have $\nu(\varepsilon) \geq \varepsilon^{-\alpha}\ell(1/\varepsilon)$ for all $0 < \varepsilon \leq 1$. Assume, in addition, that ℓ is non-decreasing. Fix $r \geq 0$ and let n_0 be the smallest $n \geq \max\{2, 1 + r\}$ satisfying the conditions*

$$\text{(a)} \quad \kappa_-\left(\frac{2}{n}\right) \leq \frac{1 + \kappa_-^0}{2},$$

$$\text{(b)} \quad \left(1 - \frac{r}{n}\right)^n \geq \frac{e^{-r}}{2},$$

$$\text{(c)} \quad \int_0^2 u^{r-\alpha} \left(1 - \frac{u}{n}\right)^n \mathrm{d}u \geq \frac{\gamma(1 + r - \alpha, 2)}{2}.$$

Then, for all $n \geq n_0$, we have

$$\mathbb{E}M_{n,r} \geq \frac{e^{-r}}{2r!} \left[(1 - p_\star)^n + \frac{(1 - \kappa_-^0)\gamma(1 + r - \alpha, 2)}{2^{1-\alpha} 4^{1+r}}\right] \frac{\ell(n)}{n^{1-\alpha}},$$

where $p_\star = \max\{p_a : a \in \mathcal{A}\}$.

According to Proposition 2.2, all assumptions of Corollary 2.5 are satisfied if $\nu(\varepsilon) = \varepsilon^{-\alpha}\ell(1/\varepsilon)$ with $\alpha \in (0, 1]$ and $\ell \in \mathrm{rv}_\infty^0$ is non-decreasing. We do not know whether a similar bound holds for arbitrary ℓ. However, note that, for all $\ell \in \mathrm{rv}_\infty^0$, one may use the fact that, for $\alpha > 0$ and $0 < \eta < \alpha$, there exists $0 < \varepsilon_\eta < 1$ and $C_\eta > 0$ such that $\varepsilon^{-\alpha}\ell(1/\varepsilon) \geq C_\eta \varepsilon^{\eta-\alpha}$ for all $0 < \varepsilon \leq \varepsilon_\eta$. Unfortunately, this approach yields suboptimal rates of convergence. Finally note that, for practical purposes, one can use the crude lower bound $\gamma(t, x) \geq (te^x)^{-1} x^t$.

3. Applications and extensions

3.1. Turing's formula

In many practical applications, one needs to estimate the occupancy probabilities, $M_{n,r}$. Perhaps the most famous estimator of this quantity is Turing's formula, which was introduced by Good [15], where the ideas were primarily credited to Alan M. Turing. For this reason, the estimator has come to be called Turing's formula or the Good–Turing formula. It is given by

$$T_{n,r} = \frac{(1+r)K_{n,1+r}}{n}.$$

A heuristic justification for Turing's formula may be obtained as follows. Denote $\hat{p}_a = n^{-1}\xi_n(a)$ the natural estimator of p_a, where ξ_n is defined by (1.1). Then one has

$$\begin{aligned} T_{n,r} &= \frac{1+r}{n}\sum_{a\in\mathcal{A}} \mathbf{1}\left\{\hat{p}_a = \frac{1+r}{n}\right\} \\ &= \sum_{a\in\mathcal{A}} \hat{p}_a \mathbf{1}\left\{\hat{p}_a = \frac{1+r}{n}\right\} \\ &\approx \sum_{a\in\mathcal{A}} p_a \mathbf{1}\left\{\hat{p}_a = \frac{r}{n}\right\} = M_{n,r}. \end{aligned}$$

Many properties of this estimator, including bias, consistency, and asymptotic normality have been studied, see, for example, Harris [18,19], Robbins [32], Starr [33], Holst [20], Esty [11], Chao [6], Chao and Lee [7], McAllester and Schapire [27], Gandolfi and Sastri [13], Zhang [35], Zhang and Huang [39], Zhang and Zhang [36], Ohannessian and Dahleh [28,29], Grabchak and Cosme [17], and the references therein. Noting that $\mathbb{E}T_{n,r} = \mathbb{E}M_{n-1,r}$, the bias of Turing's formula is given by

$$\mathbb{E}[M_{n,r} - T_{n,r}] = \mathbb{E}M_{n,r} - \mathbb{E}M_{n-1,r}.$$

Thus the results of this paper provide upper and lower bounds on the bias of Turing's formula.

Further, they provide bounds for the bias of certain modifications of Turing's formula. In particular, for the important case $r = 0$, a class of modified Turing formulas was introduced in Chao, Lee and Chen [8] (see also Zhang and Huang [38]). The motivation comes from the fact that for all $s = 1, 2, \ldots, n$

$$\begin{aligned} \mathbb{E}M_{n,0} &= \sum_{k\geq 1} p_k(1-p_k)^n \\ &= \sum_{i=1}^{s}(-1)^{i+1}\sum_{k\geq 1} p_k^i(1-p_k)^{n-i} + (-1)^s \sum_{k\geq 1} p_k^{s+1}(1-p_k)^{n-s} \\ &= \sum_{i=1}^{s}(-1)^{i+1}\frac{\mathbb{E}K_{n,i}}{\binom{n}{i}} + (-1)^s \frac{\mathbb{E}M_{n,s}}{\binom{n}{s}}. \end{aligned}$$

This suggests the family of estimators

$$T_{n,0}^{(s)} = \sum_{i=1}^{s} (-1)^{i+1} \frac{K_{n,i}}{\binom{n}{i}}, \qquad s = 1, 2, \ldots, n,$$

each with bias

$$\mathbb{E}\big[M_{n,0} - T_{n,0}^{(s)}\big] = (-1)^s \frac{\mathbb{E}M_{n,s}}{\binom{n}{s}}.$$

Note that $T_{n,0}^{(1)} = T_{n,0}$ is just Turing's formula. In Chao, Lee and Chen [8] it was shown that, so long as $p_\star < 0.5$, we have

$$\big|\mathbb{E}\big[M_{n,0} - T_{n,0}^{(1)}\big]\big| \geq \big|\mathbb{E}\big[M_{n,0} - T_{n,0}^{(2)}\big]\big| \geq \cdots \geq \big|\mathbb{E}\big[M_{n,0} - T_{n,0}^{(n)}\big]\big|.$$

Thus, these modifications reduce the bias, and the amount of bias remaining can be bounded using the results of this paper. Note, however, that controlling the bias of Turing's formula can only be of interest if this bias is shown to be of smaller order than the rate of decay of $M_{n,r}$ itself. The following subsection provides insights in this direction.

3.2. Bounds in probability

A natural application of the bounds provided in this article is to combine them with concentration bounds for the occupancy counts and probabilities in order to derive bounds in probability. State of the art concentration results for $K_{n,r}$ and $M_{n,r}$ may be found in Ohannessian and Dahleh [29] and Ben-Hamou, Boucheron and Ohannessian [2]. For instance, defining

$$K_{n,\bar{r}} = \sum_{s \geq r} K_{n,s}$$

and setting

$$v_{n,r} = 2 \min\big\{\mathbb{E}K_{n,\bar{r}}, \max\big\{r\mathbb{E}K_{n,r}, (1+r)\mathbb{E}K_{n,1+r}\big\}\big\},$$

Proposition 3.5 in Ben-Hamou, Boucheron and Ohannessian [2] states that, for all $t \geq 0$,

$$|K_{n,r} - \mathbb{E}K_{n,r}| < \sqrt{4v_{n,r}t} + \frac{2t}{3},$$

with probability at least $1 - 4e^{-t}$. The results of Section 2 may be applied to deduce explicit lower and upper bounds for $\mathbb{E}K_{n,r}$, denoted, respectively, by $k_{n,r}^-$ and $k_{n,r}^+$, as well as an explicit upper bound $v_{n,r}^+$ for $v_{n,r}$. Combining these bounds with the above results implies that, for all $t > 0$,

$$\max\left\{0, k_{n,r}^- - \sqrt{4v_{n,r}^+ t} - \frac{2t}{3}\right\} \leq K_{n,r} \leq k_{n,r}^+ + \sqrt{4v_{n,r}^+ t} + \frac{2t}{3},$$

with probability at least $1-4e^{-t}$. For ease of exposition, we avoid explicit formulas in this case. Instead, we present explicit bounds for the missing mass using the results of McAllester and Ortiz [26], which states that, for all $t>0$, the inequalities

$$M_{n,0} \leq \mathbb{E}M_{n,0} + \sqrt{\frac{t}{n}} \quad \text{and} \quad M_{n,0} \geq \mathbb{E}M_{n,0} - \sqrt{\frac{2t}{ne}} \tag{3.1}$$

each hold with probability at least $1-e^{-t}$. The following upper bound follows immediately from (3.1) and Corollary 2.2.

Corollary 3.1. *Suppose that $\mathcal{S}$ is infinite. Assume that, for $\alpha \in [0,1]$ and a non-increasing function $\ell \in \mathrm{rv}_\infty^0$, we have $\nu(\varepsilon) \leq \varepsilon^{-\alpha}\ell(1/\varepsilon)$, for all $0<\varepsilon\leq 1$. Then, for all $n\geq 2$ and all $t>0$,*

$$M_{n,0} \leq \left(e^{-1} + 4\gamma\left(1-\alpha, \frac{1}{2}\right)\right)\frac{\ell(n)}{n^{1-\alpha}} + \sqrt{\frac{t}{n}},$$

with probability at least $1-e^{-t}$.

The reader may deduce a similar result by using Corollary 2.3 instead of Corollary 2.2. Similarly, the reader may deduce a lower bound in probability by using Corollary 2.5. There is an important case where we can combine Corollaries 2.2 and 2.5 to get upper and lower bounds in probability that hold simultaneously. Specifically, assume that $\mathcal{A}=\{1,2,\dots\}$ and that $P=\{p_k : k\geq 1\}$ is such that, for some $\alpha\in(0,1)$,

$$p_k = \frac{k^{-1/\alpha}}{\zeta(1/\alpha)},$$

where $\zeta(1/\alpha)=\sum_{k\geq 1}k^{-1/\alpha}$ is the Riemann zeta function at $1/\alpha$. In this case the counting function is regularly varying and we can get the following.

Corollary 3.2. *Let P be as above. If $n\geq \max\{2, 2^{1/\alpha}\zeta(1/\alpha)\}$ is such that*

$$\kappa_-\left(\frac{2}{n}\right) \leq \frac{2^\alpha+1}{2^{\alpha+1}} \quad \text{and} \quad \int_0^2 u^{-\alpha}\left(1-\frac{u}{n}\right)^n \mathrm{d}u \geq \frac{\gamma(1-\alpha,2)}{2},$$

then, for all $t>0$,

$$\mathbb{P}\big(m_{n,0}^-(t,\alpha) \leq M_{n,0} \leq m_{n,0}^+(t,\alpha)\big) \geq 1-2e^{-t},$$

where we have denoted

$$m_{n,0}^-(t,\alpha) = \frac{(2^\alpha-1)\gamma(1-\alpha,2)}{32}\frac{\zeta(1/\alpha)^{-\alpha}}{n^{1-\alpha}} - \sqrt{\frac{2t}{ne}},$$
$$m_{n,0}^+(t,\alpha) = \left(e^{-1}+4\gamma\left(1-\alpha,\frac{1}{2}\right)\right)\frac{\zeta(1/\alpha)^{-\alpha}}{n^{1-\alpha}} + \sqrt{\frac{t}{n}}.$$

3.3. Random number of observations

In this subsection, we study extensions of our main results to the case where the number of observations is random and modelled by a Poisson distribution. This case corresponds to the practical situation in which the time, t, during which the observations are collected is fixed, but the number of observations is not. For this purpose, let $(n_t)_{t\geq 0}$ be a non-homogeneous Poisson process with intensity function $\lambda : \mathbb{R}_+ \to \mathbb{R}_+$, independent of the observations $(X_i)_{i\geq 1}$. Here, it is understood that n_t stands for the number of observations collected by time t. Defining

$$\xi_t(a) = \sum_{i=1}^{n_t} \mathbf{1}\{X_i = a\},$$

the occupancy counts and occupancy probabilities at time t are, respectively, defined, for all $r \geq 0$, by

$$K_r(t) = \sum_{a\in\mathcal{A}} \mathbf{1}\big\{\xi_t(a) = r\big\} \quad \text{and} \quad M_r(t) = \sum_{a\in\mathcal{A}} p_a \mathbf{1}\big\{\xi_t(a) = r\big\}.$$

Then, provided

$$\Lambda_t = \int_0^t \lambda(u)\,\mathrm{d}u \to +\infty,$$

as $t \to \infty$, a slight modification of the proof of (1.7) reveals that, if $v(\varepsilon) = \varepsilon^{-\alpha}\ell(1/\varepsilon)$ with $\alpha \in (0, 1)$ and $\ell \in \mathrm{rv}_\infty^0$, then, for all $r \geq 1$,

$$\mathbb{E}M_r(t) \sim \frac{\alpha\Gamma(1 + r - \alpha)}{r!}\Lambda_t^{\alpha-1}\ell(\Lambda_t), \tag{3.2}$$

as $t \to +\infty$. An analogous result for $K_r(t)$ also holds, but, for simplicity, we focus on $M_r(t)$. As in the case of a fixed number of observations, this result sets a benchmark for finite sample bounds. Based on the observation that

$$\mathbb{E}M_r(t) = \frac{\Lambda_t^r}{r!}\sum_{a\in\mathcal{A}} p_a^{1+r} e^{-\Lambda_t p_a},$$

the proofs of Theorems 2.1, 2.2, and 2.3 may be easily adapted to this case. Modifying Theorem 2.1, for instance, shows that for any $t \geq 0$ and any $r \geq 0$, we have

$$\mathbb{E}M_r(t) \leq \inf_{0\leq\varepsilon\leq 1}\big\{\varphi_r^+(t,\varepsilon) + \psi_r^+(t,\varepsilon)\big\}, \tag{3.3}$$

where

$$\varphi_r^+(t,\varepsilon) = \frac{\bar{c}(r)v(\varepsilon)}{\Lambda_t},$$

$$\psi_r^+(t,\varepsilon) = \frac{2^{1+r}\Lambda_t^r}{r!}\int_0^\varepsilon v\left(\frac{u}{2}\right)u^r e^{-\frac{\Lambda_t u}{2}}\,\mathrm{d}u,$$

and $\bar{c}(r) = (1+r)^{1+r}/(r!e^{1+r})$ for all integers $r \geq 0$. For the sake of brevity, we avoid explicitly stating the respective analogs of Theorems 2.2 and 2.3 involving the functions κ_+ and κ_-. As with the results of Section 2, these bounds lead to optimal-rate upper and lower bounds in terms of t. For instance, following the lines of the proof of Corollary 2.2, and under the same assumptions on the counting function, considering $\varepsilon = \Lambda_t^{-1}$ in (3.3) yields

$$\mathbb{E}M_r(t) \leq \left[\bar{c}(r) + \frac{4^{1+r}\gamma(1+r-\alpha, \frac{1}{2})}{r!}\right]\Lambda_t^{\alpha-1}\ell(\Lambda_t), \tag{3.4}$$

for any $r \geq 0$ and any $t > 0$ with $\Lambda_t^{-1} \leq 1$. One may easily deduce bounds in the spirit of Corollaries 2.3 and 2.4 under related assumptions on ν.

3.4. Arbitrary distributions in a metric space

So far in this article, the distribution, P, of our observations has been supported on an arbitrary and at most countable alphabet $\mathcal{A}$. In this subsection, we briefly investigate a generalization of the notion of occupancy probabilities to the context of an arbitrary distribution, P, on a metric space E.

Let (E, d) be a metric space and let P be any probability distribution on E equipped with its Borel σ-field. Suppose that we are given independent and identically distributed E-valued random variables $X_1, \ldots, X_n$ with common distribution P. Since P may not be discrete, a natural analog of the occupancy probabilities $M_{n,r}$ may be defined as follows. First, for $\delta > 0$ and $x \in E$, we let

$$\xi_n^{(\delta)}(x) = \sum_{i=1}^{n} \mathbf{1}\{x \in B_{X_i,\delta}\}, \tag{3.5}$$

where, for $u \in E$, $B_{u,\delta} = \{x \in E : d(x,u) < \delta\}$. In other words, $\xi_n^{(\delta)}(x)$ is the numbers of sample points from which x is at a distance strictly less than δ. Now, let X be an E-valued random variable independent of the sample and having distribution P. For any integer $0 \leq r \leq n$, we set

$$M_{n,r}^{(\delta)} = \mathbb{P}\big(\xi_n^{(\delta)}(X) = r | X_1, \ldots, X_n\big) = \int_E \mathbf{1}\big\{\xi_n^{(\delta)}(x) = r\big\} P(\mathrm{d}x). \tag{3.6}$$

The random variable $M_{n,r}^{(\delta)}$ represents the (conditional) probability that, given the first n observations, the next one will fall into the δ-neighbourhood of exactly r of them. A similar extension of the missing mass was studied in Section 4 of Berend and Kontorovich [3]. In our context, a slight generalisation of Theorem 8 in Berend and Kontorovich [3] can be written as follows. For $A \subset E$, we denote $N(A, \delta)$ the δ-covering number of A, that is, the minimal number of balls $B_{u,\delta}$ needed to cover A.

Theorem 3.1. *For all $x \in E$ and $t > 0$, let $\tau_x(t) = 1 - P(B_{x,t})$ and $N_x(t, \varrho) = N(B_{x,t}, \varrho)$. Then, for all $n \geq 1$,*

$$\mathbb{E}M_{n,0}^{(\delta)} \leq \inf_{x,t,\varrho}\left\{\tau_x(t) + \frac{N_x(t,\varrho)}{ne}\right\},$$

where the infimum is taken over all $x \in E$, *all* $t > 0$ *and all* $0 < \varrho \leq \delta/2$.

This result involves, in an interesting way, the geometry of the support of P. Suppose, for instance, that the support $\mathcal{S}$ of P is totally bounded. Then, as noted by Berend and Kontorovich [3], taking x in $\mathcal{S}$, t larger than the diameter of $\mathcal{S}$ and $\varrho = \delta/2$, leads to

$$\mathbb{E}M_{n,0}^{(\delta)} \leq \frac{N_x(t, \delta/2)}{ne},$$

which is a natural analog of their result in the discrete case.

In the sequel, we develop an alternative approach. First, we introduce an analog of the counting measure $\boldsymbol{\nu}$. For all $\delta > 0$, let $\boldsymbol{\nu}_\delta$ be the measure on $[0, 1]$ defined by

$$\int_0^1 f(u)\boldsymbol{\nu}_\delta(\mathrm{d}u) = \int_E \frac{f(P(B_{x,\delta}))}{P(B_{x,\delta})} P(\mathrm{d}x), \tag{3.7}$$

for all measurable $f : [0, 1] \to \mathbb{R}_+$. Then, denoting $\mathcal{L}_\delta(\varepsilon) = \{x \in E : P(B_{x,\delta}) \geq \varepsilon\}$, we introduce the function ν_δ defined, for all $\varepsilon \in [0, 1]$, by

$$\nu_\delta(\varepsilon) = \boldsymbol{\nu}_\delta\big([\varepsilon, 1]\big) = \int_{\mathcal{L}_\delta(\varepsilon)} P(B_{x,\delta})^{-1} P(\mathrm{d}x). \tag{3.8}$$

The function ν_δ is a natural analog of the counting function ν defined for discrete probability measures. Indeed, an easy application of the Dominated Convergence theorem shows that, if P is discrete and if for some $c > 0$ the distance between any two points in its support is lower bounded by c, then for any $\varepsilon \in (0, 1]$

$$\lim_{\delta \to 0} \nu_\delta(\varepsilon) = \nu(\varepsilon). \tag{3.9}$$

The next result is in the spirit of (1.7). Using the notation introduced in Section 3.3, we denote

$$M_r^{(\delta)}(t) = M_{n_t,r}^{(\delta)},$$

where (n_t) stands for a non-homogeneous Poisson process with intensity function $\lambda : \mathbb{R}_+ \to \mathbb{R}_+$.

Theorem 3.2. *Fix* $\delta > 0$. *Suppose that for some* $\alpha \in [0, 1]$ *and some* $\ell \in \mathrm{rv}_\infty^0$, *possibly depending on* δ, *we have* $\nu_\delta(\varepsilon) = \varepsilon^{-\alpha}\ell(1/\varepsilon)$. *Then, for all* $r \geq 0$,

$$\mathbb{E}M_r^{(\delta)}(t) \sim \frac{\alpha\Gamma(1 + r - \alpha)}{r!} \Lambda_t^{\alpha-1} \ell(\Lambda_t), \tag{3.10}$$

as $t \to +\infty$, *provided* $\Lambda_t \to +\infty$ *as* $t \to +\infty$.

To keep the proof simple, we present this result in the context where the number of observations, n, follows a non-homogeneous Poisson process. However, this proof can be modified to give an analogous result in the case where n is fixed. Finally, denoting

$$\kappa_-^{(\delta)}(\varepsilon) = \sup_{0<u\leq\varepsilon} \frac{\nu_\delta(u)}{\nu_\delta(u/2)} \quad \text{and} \quad \kappa_+^{(\delta)}(\varepsilon) = \sup_{0<u\leq\varepsilon} \frac{\nu_\delta(u/2)}{\nu_\delta(u)},$$

for $0 < \varepsilon \leq 1$, the reader may easily check that, in the context of this subsection, Theorems 2.1, 2.2 and 2.3 hold exactly provided ν, κ_-, κ_+ and $p_\star$ are replaced, respectively, by

$$\nu_\delta, \qquad \kappa_-^{(\delta)}, \qquad \kappa_+^{(\delta)} \quad \text{and} \quad p_\star^{(\delta)} = \sup\big\{P(B_{x,\delta}) : x \in E\big\}.$$

The results of this subsection could find interesting applications in the context of a continuous time stochastic process $X = (X_t)_{0\leq t\leq T}$, considering P to be the distribution of the whole path $X = (X_t)_{0\leq t\leq T}$ or of X_t for some $0 \leq t \leq T$. In order to have relevant information on the generalized counting function ν_δ, one needs explicit upper and lower bounds on the probabilities of balls $P(B_{x,\delta})$, $x \in E$. Results in this direction have been widely studied and may be related to large deviations theory and density estimates for stochastic partial differential equations. Finally, an interesting question is whether the work of Ben-Hamou, Boucheron and Ohannessian [2] on concentration inequalities can be adapted to this general case. This is left for future research.

4. Proofs

Proof of Theorem 2.1. For all $n \geq 1$ and all $0 \leq r \leq n$,

$$\mathbb{E}M_{n,r} = \sum_{a\in\mathcal{A}} p_a \mathbb{P}\big(\xi_n(a) = r\big). \tag{4.1}$$

For any $a \in \mathcal{A}$, the variables $\mathbf{1}\{X_i = a\}$, $i = 1, \ldots, n$, are independent and have the same Bernoulli distribution with parameter p_a. This implies that

$$\mathbb{P}\big(\xi_n(a) = r\big) = \binom{n}{r} p_a^r (1 - p_a)^{n-r}. \tag{4.2}$$

As a result, we deduce from (4.1) and (4.2) that, for all $n \geq 1$ and all $0 \leq r \leq n$,

$$\mathbb{E}M_{n,r} = \binom{n}{r} \sum_{a\in\mathcal{A}} p_a^{r+1} (1 - p_a)^{n-r}. \tag{4.3}$$

Now, suppose that $0 \leq \varepsilon \leq 1$, $n \geq 1$ and $0 \leq r \leq n - 1$ are fixed. Note that, from (4.3), we can write

$$\begin{aligned}
\mathbb{E}M_{n,r} &= \binom{n}{r} \sum_{a\in\mathcal{A}} p_a^{r+1} (1 - p_a)^{n-r} \\
&= \binom{n}{r} \sum_{a:p_a\geq\varepsilon} p_a^{r+1} (1 - p_a)^{n-r} + \binom{n}{r} \sum_{a:p_a<\varepsilon} p_a^{r+1} (1 - p_a)^{n-r} \qquad (4.4) \\
&=: \binom{n}{r} (S_1 + S_2).
\end{aligned}$$

To bound the first term observe that from the definition of the counting function ν introduced in (1.4), we obtain

$$\begin{aligned}\binom{n}{r} S_1 &\leq \binom{n}{r} \nu(\varepsilon) \sup_{u\in[0,1]} u^{r+1}(1-u)^{n-r} \\ &= \binom{n}{r} \nu(\varepsilon) \frac{(1+r)^{1+r}(n-r)^{n-r}}{(1+n)^{1+n}}.\end{aligned} \tag{4.5}$$

In the case where $r = 0$, the upper bound (4.5) becomes

$$\begin{aligned}\binom{n}{r} S_1 &\leq \nu(\varepsilon) \frac{n^n}{(1+n)^{1+n}} \\ &= \frac{\nu(\varepsilon)}{n} \left(1 - \frac{1}{1+n}\right)^{1+n} \\ &\leq \frac{\nu(\varepsilon)}{ne},\end{aligned} \tag{4.6}$$

where, in (4.6), we have used the fact that $\forall u \in [0,1]$: $(1-u) \leq e^{-u}$. In the case $1 \leq r \leq n-1$, developing the binomial coefficient in (4.5), we need to evaluate the term

$$\frac{n!}{r!(n-r)!} \frac{(1+r)^{1+r}(n-r)^{n-r}}{(1+n)^{1+n}}. \tag{4.7}$$

Using the Stirling type bound (see Robbins [31])

$$\sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n+\frac{1}{12n+1}} < n! < \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n+\frac{1}{12n}},$$

valid for all $n \geq 1$, we deduce in particular that for all $1 \leq r \leq n-1$, we have the following inequalities

$$\begin{aligned}n! &< \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n+\frac{1}{12n}} \\ &\leq \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n+\frac{1}{12(1+r)}} \\ &< \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n+\frac{1}{12r+1}},\end{aligned} \tag{4.8}$$

$$r! > \sqrt{2\pi} r^{r+\frac{1}{2}} e^{-r+\frac{1}{12r+1}}, \tag{4.9}$$

$$(n-r)! > \sqrt{2\pi} (n-r)^{n-r+\frac{1}{2}} e^{-n+r}. \tag{4.10}$$

Using inequalities (4.8), (4.9) and (4.10), we obtain

$$\frac{n!}{r!(n-r)!} \leq \frac{1}{\sqrt{2\pi}} \frac{n^{n+\frac{1}{2}}}{r^{r+\frac{1}{2}}(n-r)^{n-r+\frac{1}{2}}},$$

implying that the expression in (4.7) can be upper bounded by

$$\frac{1}{\sqrt{2\pi}}\frac{(1+r)^{1+r}}{r^{r+\frac{1}{2}}}\frac{n^{n+\frac{1}{2}}}{(1+n)^{1+n}}\frac{1}{(n-r)^{\frac{1}{2}}}. \tag{4.11}$$

Using the inequality $1/(n-r) \leq (1+r)/n$, valid for all $0 \leq r \leq n-1$, we obtain

$$\frac{n^{n+\frac{1}{2}}}{(1+n)^{1+n}}\frac{1}{(n-r)^{\frac{1}{2}}} \leq (1+r)^{\frac{1}{2}}\left(\frac{n}{1+n}\right)^{1+n}\frac{1}{n} \leq \frac{(1+r)^{\frac{1}{2}}}{n}. \tag{4.12}$$

Combining (4.11) and (4.12), the term (4.7) is therefore upper bounded by

$$\begin{aligned}\frac{1}{n\sqrt{2\pi}}\left(1+\frac{1}{r}\right)^{r+\frac{1}{2}}(1+r) &\leq \frac{1}{n\sqrt{\pi}}\left(1+\frac{1}{r}\right)^{r}(1+r)\\ &\leq \frac{e(1+r)}{n\sqrt{\pi}},\end{aligned} \tag{4.13}$$

where we have used that $(1+1/r)^r \leq e$ for all $r \geq 1$. Combining (4.5) and (4.13) brings finally

$$\binom{n}{r}S_1 \leq \frac{e(1+r)}{\sqrt{\pi}}\frac{\nu(\varepsilon)}{n}, \tag{4.14}$$

for all $1 \leq r \leq n-1$. Combining (4.6) and (4.14), we have therefore established that

$$\binom{n}{r}S_1 \leq \frac{c(r)\nu(\varepsilon)}{n}, \tag{4.15}$$

for all $0 \leq r \leq n-1$ where $c(r)$ is as in (2.2). We now focus on bounding the second term in (4.4). Toward this end, we choose $b > 1$ and write

$$\begin{aligned}S_2 &= \sum_{j=0}^{+\infty}\sum_{a:p_a<\varepsilon}\mathbf{1}\left\{\frac{\varepsilon}{b^{j+1}} \leq p_a < \frac{\varepsilon}{b^j}\right\}p_a^{r+1}(1-p_a)^{n-r} \qquad (4.16)\\ &\leq \sum_{j=0}^{+\infty}\left[\nu\left(\frac{\varepsilon}{b^{j+1}}\right)-\nu\left(\frac{\varepsilon}{b^j}\right)\right]\left(\frac{\varepsilon}{b^j}\right)^{r+1}\left(1-\frac{\varepsilon}{b^{j+1}}\right)^{n-r}\\ &\leq \sum_{j=0}^{+\infty}\nu\left(\frac{\varepsilon}{b^{j+1}}\right)\left(\frac{\varepsilon}{b^j}\right)^{r+1}\left(1-\frac{\varepsilon}{b^{j+1}}\right)^{n-r}\\ &= \frac{b}{b-1}\sum_{j=0}^{+\infty}\left(\frac{\varepsilon}{b^j}-\frac{\varepsilon}{b^{j+1}}\right)\nu\left(\frac{\varepsilon}{b^{j+1}}\right)\left(\frac{\varepsilon}{b^j}\right)^{r}\left(1-\frac{\varepsilon}{b^{j+1}}\right)^{n-r}\end{aligned}$$

$$\leq \frac{b^{1+r}}{b-1}\sum_{j=0}^{+\infty}\int_{\frac{\varepsilon}{b^{j+1}}}^{\frac{\varepsilon}{b^j}} v\left(\frac{u}{b}\right)u^r\left(1-\frac{u}{b}\right)^{n-r}\mathrm{d}u \tag{4.17}$$

$$= \frac{b^{1+r}}{b-1}\int_0^{\varepsilon} v\left(\frac{u}{b}\right)u^r\left(1-\frac{u}{b}\right)^{n-r}\mathrm{d}u. \tag{4.18}$$

For inequality (4.17) we have used the fact that the functions $u \mapsto u^r$ and $u \mapsto v(u)(1-u)^{n-r}$ are respectively, non-decreasing and non-increasing so that, for all $u \in [\varepsilon b^{-j-1}, \varepsilon b^{-j}]$, we have

$$v\left(\frac{\varepsilon}{b^{j+1}}\right)\left(\frac{\varepsilon}{b^j}\right)^r\left(1-\frac{\varepsilon}{b^{j+1}}\right)^{n-r} \leq b^r v\left(\frac{u}{b}\right)u^r\left(1-\frac{u}{b}\right)^{n-r}.$$

Hence, equation (4.18) implies that

$$\binom{n}{r}S_2 \leq \frac{b^{1+r}}{b-1}\binom{n}{r}\int_0^{\varepsilon} v\left(\frac{u}{b}\right)u^r\left(1-\frac{u}{b}\right)^{n-r}\mathrm{d}u,$$

which, along which equation (4.14) and the choice of $b = 2$, proves the first claim in Theorem 2.1. We next turn to the inequality (2.4). Again, suppose that $\varepsilon \in [0, 1]$ is fixed and note that, for $r = n$, (4.4) becomes

$$\mathbb{E}M_{n,n} = \sum_{a:p_a \geq \varepsilon} p_a^{n+1} + \sum_{k:p_a < \varepsilon} p_a^{n+1}. \tag{4.19}$$

Bounding each p_a by $p_\star$ in the first sum and by ε in the second, we obtain

$$\mathbb{E}M_{n,n} \leq p_\star^{n+1} v(\varepsilon) + \varepsilon^n \sum_{a:p_a \leq \varepsilon} p_a \leq p_\star^{n+1} v(\varepsilon) + \varepsilon^n, \tag{4.20}$$

which completes the proof. □

Proof of Corollary 2.2. Let $n \geq 1$ and $0 \leq r \leq n-1$ be fixed. Theorem 2.1 implies, in particular, that $\mathbb{E}M_{n,r} \leq \varphi_{n,r}^+(1/n) + \psi_{n,r}^+(1/n)$. Given the assumption on the counting function, we have

$$\varphi_{n,r}^+(1/n) \leq \frac{c(r)\ell(n)}{n^{1-\alpha}}. \tag{4.21}$$

To bound the second term, note that

$$\psi_{n,r}^+(1/n) = 2^{1+r}\binom{n}{r}I_n \qquad \text{where } I_n = \int_0^{\frac{1}{n}} v\left(\frac{u}{2}\right)u^r\left(1-\frac{u}{2}\right)^{n-r}\mathrm{d}u.$$

Since $v(\varepsilon) \leq \varepsilon^{-\alpha}\ell(1/\varepsilon)$, we deduce that

$$I_n \leq 2^\alpha \int_0^{\frac{1}{n}} \ell\left(\frac{2}{u}\right)u^{r-\alpha}\left(1-\frac{u}{2}\right)^{n-r}\mathrm{d}u$$

$$= 2^{1+r} \int_0^{\frac{1}{2n}} \ell\left(\frac{1}{u}\right) u^{r-\alpha}(1-u)^{n-r}\,\mathrm{d}u \tag{4.22}$$

$$\leq 2^{1+r}\ell(n) \int_0^{\frac{1}{2n}} u^{r-\alpha}(1-u)^{n-r}\,\mathrm{d}u, \tag{4.23}$$

where (4.22) follows from a change of variables and (4.23) uses the fact that ℓ is non-increasing. Then, since $(1-u) \leq e^{-u}$ for $0 \leq u \leq 1$, we have

$$\begin{aligned} I_n &\leq 2^{1+r}\ell(n) \int_0^{\frac{1}{2n}} u^{r-\alpha} e^{-(n-r)u}\,\mathrm{d}u \\ &= \frac{2^{1+r}\ell(n)}{(n-r)^{1+r-\alpha}} \int_0^{\frac{n-r}{2n}} u^{r-\alpha} e^{-u}\,\mathrm{d}u \\ &\leq \frac{2^{1+r}\ell(n)}{(n-r)^{1+r-\alpha}} \int_0^{\frac{1}{2}} u^{r-\alpha} e^{-u}\,\mathrm{d}u \\ &\leq \frac{2^{1+r}\ell(n)(1+r)^{1+r-\alpha}}{n^{1+r-\alpha}} \int_0^{\frac{1}{2}} u^{r-\alpha} e^{-u}\,\mathrm{d}u \qquad (4.24) \\ &= \frac{2^{1+r}\ell(n)(1+r)^{1+r-\alpha}}{n^{1+r-\alpha}} \gamma\left(1+r-\alpha, \frac{1}{2}\right), \end{aligned}$$

where, in (4.24), we used the fact that $1/(n-r) \leq (1+r)/n$ for $r \leq n-1$. Finally, using the fact that $\binom{n}{r} \leq n^r/r!$, we obtain

$$\psi^+_{n,r}(1/n) \leq \frac{4^{1+r}}{r!}(1+r)^{1+r-\alpha}\gamma\left(1+r-\alpha, \frac{1}{2}\right)\frac{\ell(n)}{n^{1-\alpha}}. \tag{4.25}$$

Combining (4.21) and (4.25) gives the result. □

Proof of Corollary 2.3. The proof of Corollary 2.3 follows along the same lines as the proof of Corollary 2.2 up to (4.22). Then, applying Cauchy–Schwarz's inequality, we obtain for all $\beta \in (0,1)$ such that $\beta > 2(\alpha - r) - 1$,

$$\begin{aligned} I_n &\leq 2^{1+r} \int_0^{\frac{1}{2n}} \ell\left(\frac{1}{u}\right) u^{r-\alpha}(1-u)^{n-r}\,\mathrm{d}u \\ &= 2^{1+r} \int_0^{\frac{1}{2n}} u^{-\frac{\beta}{2}} \ell\left(\frac{1}{u}\right) u^{r-\alpha+\frac{\beta}{2}}(1-u)^{n-r}\,\mathrm{d}u \\ &\leq 2^{1+r} \sqrt{\int_0^{\frac{1}{2n}} u^{-\beta}\ell\left(\frac{1}{u}\right)^2 \mathrm{d}u} \sqrt{\int_0^{\frac{1}{2n}} u^{2(r-\alpha)+\beta}(1-u)^{2(n-r)}\,\mathrm{d}u} \end{aligned} \tag{4.26}$$

$$\leq 2^{1+r}\ell_\beta^\circ(n)\sqrt{\int_0^{\frac{1}{2n}} u^{2(r-\alpha)+\beta}(1-u)^{2(n-r)}\,\mathrm{d}u},$$

where (4.26) follows from (2.5) and a change of variables. Then, from similar arguments as in the proof of Corollary 2.2, we deduce

$$\begin{aligned}
\psi_{n,r}^+(1/n) &= 2^{1+r}\binom{n}{r} I_n \\
&\leq \frac{4^{1+r}}{r!} n^r \ell_\beta^\circ(n)\sqrt{\int_0^{\frac{1}{2n}} u^{2(r-\alpha)+\beta}(1-u)^{2(n-r)}\,\mathrm{d}u} \\
&\leq \frac{4^{1+r}}{r!} n^r \ell_\beta^\circ(n)\sqrt{\int_0^{\frac{1}{2n}} u^{2(r-\alpha)+\beta} e^{-2u(n-r)}\,\mathrm{d}u} \\
&= \frac{4^{1+r}}{r!} n^r \ell_\beta^\circ(n)\sqrt{\int_0^{1-\frac{r}{n}} u^{2(r-\alpha)+\beta} e^{-u}\,\mathrm{d}u}\left(\frac{1}{2(n-r)}\right)^{r-\alpha+\frac{\beta+1}{2}} \\
&\leq c_2(\alpha,\beta,r) n^{\alpha-\frac{1+\beta}{2}} \ell_\beta^\circ(n),
\end{aligned}$$

which completes the proof. □

Proof of Theorem 2.2. The proof of Theorem 2.2 follows the same lines as the proof of Theorem 2.1 up to (4.16), where we take $b = 2$. Then, we write

$$\begin{aligned}
S_2 &= \sum_{j=0}^{+\infty}\sum_{a:\,p_a<\varepsilon}^{+\infty} \mathbf{1}\left\{\frac{\varepsilon}{2^{j+1}} \leq p_a < \frac{\varepsilon}{2^j}\right\} p_a^{r+1}(1-p_a)^{n-r} \\
&\leq \sum_{j=0}^{+\infty}\left[\nu\left(\frac{\varepsilon}{2^{j+1}}\right) - \nu\left(\frac{\varepsilon}{2^j}\right)\right]\left(\frac{\varepsilon}{2^j}\right)^{r+1}\left(1-\frac{\varepsilon}{2^{j+1}}\right)^{n-r} \\
&= \sum_{j=0}^{+\infty}\left[\frac{\nu(\frac{\varepsilon}{2^{j+1}})}{\nu(\frac{\varepsilon}{2^j})} - 1\right]\nu\left(\frac{\varepsilon}{2^j}\right)\left(\frac{\varepsilon}{2^j}\right)^{r+1}\left(1-\frac{\varepsilon}{2^{j+1}}\right)^{n-r} \\
&\leq \sum_{j=0}^{+\infty}\left[\kappa_+\left(\frac{\varepsilon}{2^j}\right) - 1\right]\nu\left(\frac{\varepsilon}{2^j}\right)\left(\frac{\varepsilon}{2^j}\right)^{r+1}\left(1-\frac{\varepsilon}{2^{j+1}}\right)^{n-r} \\
&= 2\sum_{j=0}^{+\infty}\left(\frac{\varepsilon}{2^j} - \frac{\varepsilon}{2^{j+1}}\right)\left[\kappa_+\left(\frac{\varepsilon}{2^j}\right) - 1\right]\nu\left(\frac{\varepsilon}{2^j}\right)\left(\frac{\varepsilon}{2^j}\right)^{r}\left(1-\frac{\varepsilon}{2^{j+1}}\right)^{n-r} \\
&\leq 2^{1+r}\sum_{j=0}^{+\infty}\int_{\frac{\varepsilon}{2^{j+1}}}^{\frac{\varepsilon}{2^j}} \big(\kappa_+(2u)-1\big)\nu(u)u^r\left(1-\frac{u}{2}\right)^{n-r}\mathrm{d}u \qquad (4.27)
\end{aligned}$$

$$= 2^{1+r} \int_0^{\varepsilon} \big(\kappa_+(2u) - 1\big) \nu(u) u^r \left(1 - \frac{u}{2}\right)^{n-r} \mathrm{d}u, \tag{4.28}$$

where, in (4.27), we use the monotonicity of both $u \mapsto (\kappa_+(u) - 1)u^r$ and $u \mapsto \nu(u)(1-u)^{n-r}$. This completes the proof. □

Proof of Corollary 2.4. Using the same arguments as in the beginning of the proof of Corollary 2.2, but $\theta^+_{n,r}(1/n)$ in place of $\psi^+_{n,r}(1/n)$, we obtain

$$\mathbb{E}M_{n,r} \leq \frac{c(r)\ell(n)}{n^{1-\alpha}} + 2^{1+r} \binom{n}{r} J_n, \tag{4.29}$$

where

$$J_n = \int_0^{\frac{1}{n}} \big(\kappa_+(2u) - 1\big) \nu(u) u^r \left(1 - \frac{u}{2}\right)^{n-r} \mathrm{d}u.$$

Note that

$$J_n \leq \int_0^{\frac{1}{n}} \big(\kappa_+(2u) - 1\big) \ell\left(\frac{1}{u}\right) u^{r-\alpha} \left(1 - \frac{u}{2}\right)^{n-r} \mathrm{d}u.$$

Given assumption (2.10), and the fact that κ_+ is non-decreasing, we deduce that

$$\begin{aligned} J_n &\leq \big(\kappa_+(2/n) - 1\big) \int_0^{\frac{1}{n}} \ell\left(\frac{1}{u}\right) u^{r-\alpha} \left(1 - \frac{u}{2}\right)^{n-r} \mathrm{d}u \\ &\leq 2\big(\kappa^0_+ - 1\big) \int_0^{\frac{1}{n}} \ell\left(\frac{1}{u}\right) u^{r-\alpha} \left(1 - \frac{u}{2}\right)^{n-r} \mathrm{d}u. \end{aligned} \tag{4.30}$$

Proceeding now as in the proof of Corollary 2.3 and applying Cauchy–Schwarz's inequality in (4.30) leads, for all $\beta > 2(\alpha - r) - 1$, to

$$\begin{aligned} J_n &\leq 2\big(\kappa^0_+ - 1\big) \int_0^{\frac{1}{n}} u^{-\frac{\beta}{2}} \ell\left(\frac{1}{u}\right) u^{r-\alpha+\frac{\beta}{2}} \left(1 - \frac{u}{2}\right)^{n-r} \mathrm{d}u \\ &\leq 2\big(\kappa^0_+ - 1\big) \sqrt{\int_0^{\frac{1}{n}} u^{-\beta} \ell\left(\frac{1}{u}\right)^2 \mathrm{d}u} \sqrt{\int_0^{\frac{1}{n}} u^{2(r-\alpha)+\beta} \left(1 - \frac{u}{2}\right)^{2(n-r)} \mathrm{d}u} \\ &= 2\big(\kappa^0_+ - 1\big) \ell^\circ_\beta\left(\frac{n}{2}\right) \sqrt{\int_0^{\frac{1}{n}} u^{2(r-\alpha)+\beta} \left(1 - \frac{u}{2}\right)^{2(n-r)} \mathrm{d}u}, \end{aligned} \tag{4.31}$$

where (4.31) follows from (2.5) and a change of variables. Using, as in the proofs of Corollary 2.2 and Corollary 2.3, the fact that $(1-u) \leq e^{-u}$ for $0 \leq u \leq 1$ and the observation that $1/(n-r) \leq$

$(1+r)/n$ for $r \le n-1$, the reader may easily check that the square root term in (4.31) is upper bounded by

$$\left(\frac{1+r}{n}\right)^{\frac{1+\beta}{2}+r-\alpha} \sqrt{\gamma\big(1+\beta+2(r-\alpha),1\big)}.$$

Finally, combining this last observation with (4.31), and the fact that $\binom{n}{r} \le n^r/r!$, we deduce that the second term on the right-hand side of (4.29) is at most

$$\begin{aligned}
&2^{2+r}\binom{n}{r}\big(\kappa_+^0-1\big)\ell_\beta^\circ\left(\frac{n}{2}\right)\left(\frac{1+r}{n}\right)^{\frac{1+\beta}{2}+r-\alpha}\sqrt{\gamma\big(1+\beta+2(r-\alpha),1\big)}\\
&\quad\le \frac{2^{2+r}}{r!}\big(\kappa_+^0-1\big)\ell_\beta^\circ\left(\frac{n}{2}\right)n^{\alpha-\frac{1+\beta}{2}}(1+r)^{\frac{1+\beta}{2}+r-\alpha}\sqrt{\gamma\big(1+\beta+2(r-\alpha),1\big)}\\
&\quad= \big(\kappa_+^0-1\big)c_2(\alpha,\beta,r)\left(\frac{n}{2}\right)^{\alpha-\frac{1+\beta}{2}}\ell_\beta^\circ\left(\frac{n}{2}\right),
\end{aligned}$$

where $c_2(\alpha,\beta,r)$ is as in Corollary 2.3. This completes the proof. □

Proof of Theorem 2.3. Fix $n \ge 1$, $0 \le r \le n-1$ and $\varepsilon \in (0, 1/2]$. As in the proof of Theorem 2.1, we write

$$\begin{aligned}
\mathbb{E}M_{n,r} &= \binom{n}{r}\sum_{a:p_a\ge\varepsilon} p_a^{r+1}(1-p_a)^{n-r} + \binom{n}{r}\sum_{a:p_a<\varepsilon} p_a^{r+1}(1-p_a)^{n-r}\\
&=: \binom{n}{r}(S_1+S_2).
\end{aligned}$$

From the definition of ν, it is clear that

$$S_1 \ge \nu(\varepsilon)\varepsilon^{r+1}(1-p_\star)^{n-r}.$$

To bound S_2, we write

$$\begin{aligned}
S_2 &= \sum_{j=0}^{\infty}\sum_{a:p_a<\varepsilon}\mathbf{1}\left\{\frac{\varepsilon}{2^{j+1}} \le p_a < \frac{\varepsilon}{2^j}\right\}p_a^{r+1}(1-p_a)^{n-r}\\
&\ge \sum_{j=0}^{\infty}\left[\nu\left(\frac{\varepsilon}{2^{j+1}}\right)-\nu\left(\frac{\varepsilon}{2^j}\right)\right]\left(\frac{\varepsilon}{2^{j+1}}\right)^{r+1}\left(1-\frac{\varepsilon}{2^j}\right)^{n-r}\\
&= \sum_{j=0}^{\infty}\left(\frac{\varepsilon}{2^j}-\frac{\varepsilon}{2^{j+1}}\right)\left[1-\frac{\nu(\frac{\varepsilon}{2^j})}{\nu(\frac{\varepsilon}{2^{j+1}})}\right]\nu\left(\frac{\varepsilon}{2^{j+1}}\right)\left(\frac{\varepsilon}{2^{j+1}}\right)^{r}\left(1-\frac{\varepsilon}{2^j}\right)^{n-r}
\end{aligned}$$

$$\geq \sum_{j=0}^{\infty}\left(\frac{\varepsilon}{2^j}-\frac{\varepsilon}{2^{j+1}}\right)\left[1-\kappa_-\left(\frac{\varepsilon}{2^j}\right)\right]\nu\left(\frac{\varepsilon}{2^{j+1}}\right)\left(\frac{\varepsilon}{2^{j+1}}\right)^r\left(1-\frac{\varepsilon}{2^j}\right)^{n-r}$$

$$\geq 2^{-r}\sum_{j=0}^{\infty}\int_{\frac{\varepsilon}{2^{j+1}}}^{\frac{\varepsilon}{2^j}}\big(1-\kappa_-(2u)\big)\nu(u)u^r(1-2u)^{n-r}\,\mathrm{d}u \tag{4.32}$$

$$= 2^{-r}\int_0^{\varepsilon}\big(1-\kappa_-(2u)\big)\nu(u)u^r(1-2u)^{n-r}\,\mathrm{d}u$$

where in (4.32), we used the monotonicity of both $u\mapsto u^r$ and $u\mapsto(1-\kappa_-(2u))\nu(u)(1-2u)^{n-r}$. This completes the proof. □

Proof of Corollary 2.5. Fix $n\geq n_0$. The bound in Corollary 2.5 is obtained by taking $\varepsilon=n^{-1}$ in Theorem 2.3. First, using the assumption on ν, note that

$$\varphi^-_{n,r}(1/n)\geq\frac{\binom{n}{r}}{n^r}(1-p_\star)^n\frac{\ell(n)}{n^{1-\alpha}}$$

$$\geq\frac{1}{r!}\left(1-\frac{r}{n}\right)^n(1-p_\star)^n\frac{\ell(n)}{n^{1-\alpha}} \tag{4.33}$$

$$\geq\frac{e^{-r}}{2r!}(1-p_\star)^n\frac{\ell(n)}{n^{1-\alpha}}, \tag{4.34}$$

where (4.33) is due to the fact that

$$\frac{\binom{n}{r}}{n^r}\geq\frac{1}{r!}\left(1-\frac{r}{n}\right)^r\geq\frac{1}{r!}\left(1-\frac{r}{n}\right)^n, \tag{4.35}$$

and (4.34) follows from condition (b). Next, denote

$$J_n=\int_0^{\frac{1}{n}}\big(1-\kappa_-(2u)\big)\nu(u)u^r(1-2u)^{n-r}\,\mathrm{d}u.$$

Using condition (a) and the assumption on ν, it may be easily checked that

$$\begin{aligned}J_n&\geq\frac{(1-\kappa_-^0)\ell(n)}{2}\int_0^{\frac{1}{n}}u^{r-\alpha}(1-2u)^{n-r}\,\mathrm{d}u\\&=\frac{(1-\kappa_-^0)\ell(n)}{2(2n)^{1+r-\alpha}}\int_0^2u^{r-\alpha}\left(1-\frac{u}{n}\right)^{n-r}\mathrm{d}u\\&\geq\frac{(1-\kappa_-^0)\ell(n)}{2(2n)^{1+r-\alpha}}\int_0^2u^{r-\alpha}\left(1-\frac{u}{n}\right)^{n}\mathrm{d}u\\&\geq\frac{(1-\kappa_-^0)\ell(n)}{4(2n)^{1+r-\alpha}}\gamma(1+r-\alpha,2),\end{aligned} \tag{4.36}$$

where (4.36) follows from condition (c). By rearranging the terms and using (4.35) once again along with condition (b), we finally deduce that

$$\theta^{-}_{n,r}(1/n) = 2^{-r}\binom{n}{r} J_n \geq \frac{e^{-r}}{2r!}\big(1-\kappa^0_-\big)\frac{\gamma(1+r-\alpha,2)}{2^{1-\alpha}4^{1+r}}\frac{\ell(n)}{n^{1-\alpha}}. \tag{4.37}$$

The result follows from (4.34) and (4.37). □

Proof of Corollary 3.2. Let us denote $z = \zeta(1/\alpha)$ for brevity. First, it may be easily verified that the counting function ν of the distribution considered satisfies, for all $0 < \varepsilon < 1$, $\nu(\varepsilon) = \lfloor (z\varepsilon)^{-\alpha} \rfloor$, where $\lfloor \cdot \rfloor$ is the floor function. As a result, $\varepsilon^\alpha \nu(\varepsilon) \to z^{-\alpha}$ as $\varepsilon \to 0$, and thus $\nu \in \mathrm{rv}_0^\alpha$. From here, Proposition 2.2 implies that $\kappa^0_- = 2^{-\alpha}$. Noticing that $p_\star = z^{-1}$ and that, for $\varepsilon \leq (2^{1/\alpha}z)^{-1}$, we have $\nu(\varepsilon) \geq (z\varepsilon)^{-\alpha} - 1 \geq (z\varepsilon)^{-\alpha}/2$ (i.e. for $x \geq 2^{1/\alpha}z$ we can take $\ell(x) = z^{-\alpha}/2$), Corollary 2.5 implies that, provided $n \geq \max\{2, 2^{1/\alpha}z\}$ and the conditions

$$\kappa_-\left(\frac{2}{n}\right) \leq \frac{2^\alpha + 1}{2^{\alpha+1}} \quad \text{and} \quad \int_0^2 u^{-\alpha}\left(1-\frac{u}{n}\right)^n \mathrm{d}u \geq \frac{\gamma(1-\alpha,2)}{2}$$

are satisfied (since $r = 0$, condition (b) in Corollary 2.5 automatically holds), we obtain

$$\begin{aligned}
\mathbb{E}M_{n,0} &\geq \frac{1}{4}\left[\big(1-z^{-1}\big)^n + \frac{(1-2^{-\alpha})\gamma(1-\alpha,2)}{2^{1-\alpha}4}\right]\frac{z^{-\alpha}}{n^{1-\alpha}} \\
&= \frac{1}{4}\left[\big(1-z^{-1}\big)^n + \frac{(2^\alpha-1)\gamma(1-\alpha,2)}{8}\right]\frac{z^{-\alpha}}{n^{1-\alpha}} \\
&\geq \frac{(2^\alpha-1)\gamma(1-\alpha,2)}{32}\frac{z^{-\alpha}}{n^{1-\alpha}},
\end{aligned}$$

where the last inequality follows from the fact that $z > 1$. Now note that $\nu(\varepsilon) = \lfloor (z\varepsilon)^{-\alpha} \rfloor \leq (z\varepsilon)^{-\alpha}$. Thus applying Corollary 2.2 with ℓ constant and equal to $z^{-\alpha}$ gives an upper bound on $\mathbb{E}M_{n,0}$. Combining the upper and lower bounds with the concentration bound in (3.1) yields the desired result. □

Proof of Theorem 3.1. Fix $x \in E$, $t > 0$ and $0 < \varrho \leq \delta/2$. Then, observe that,

$$\begin{aligned}
\mathbb{E}M^{(\delta)}_{n,0} &= \int_E \mathbb{P}\big(\xi^{(\delta)}_n(u) = 0\big)P(\mathrm{d}u) \\
&= \int_E \big(1 - P(B_{u,\delta})\big)^n P(\mathrm{d}u) \qquad (4.38) \\
&\leq \tau_x(t) + \int_{B_{x,t}} \big(1 - P(B_{u,\delta})\big)^n P(\mathrm{d}u). \qquad (4.39)
\end{aligned}$$

Here, (4.38) follows from the fact that, for all $u \in E$, $\xi^{(\delta)}_n(u)$ has a Binomial distribution with parameters n and $P(B_{u,\delta})$, and (4.39) follows from the fact that $(1 - P(B_{u,\delta}))^n \leq 1$. Next, let

$N = N(B_{x,t}, \varrho)$ and let $B_1, \dots, B_N \subset E$ be balls with radius ϱ satisfying $B_{x,t} \subset B_1 \cup \cdots \cup B_N$. Then, we obtain

$$\begin{aligned}\int_{B_{x,t}} \big(1 - P(B_{u,\delta})\big)^n P(\mathrm{d}u) &\le \sum_{i=1}^{N} \int_{B_i} \big(1 - P(B_{u,\delta})\big)^n P(\mathrm{d}u) \\ &\le \sum_{i=1}^{N} P(B_i)\big(1 - P(B_i)\big)^n,\end{aligned} \tag{4.40}$$

where (4.40) follows from the fact that, since $\varrho \le \delta/2$, if $u \in B_i$ then necessarily $B_i \subset B_{u,\delta}$. Now exactly as in (4.6), one may deduce that

$$\begin{aligned}\sum_{i=1}^{N} P(B_i)\big(1 - P(B_i)\big)^n &\le N \sup_{0 \le p \le 1} p(1-p)^n \\ &\le \frac{N}{ne}.\end{aligned} \tag{4.41}$$

The result follows by combining (4.39), (4.40), (4.41) and taking the infimum over $x \in E$, $t > 0$ and $\varrho \le \delta/2$. □

Proof of Theorem 3.2. First, note that

$$\mathbb{E}M_r^{(\delta)}(t) = \int_E \mathbb{P}\big(\xi_{n_t}^{(\delta)}(x) = r\big) P(\mathrm{d}x). \tag{4.42}$$

For all $x \in E$, the variable $\xi_{n_t}^{(\delta)}(x)$ follows a Poisson distribution with parameter $\Lambda_t P(B_{x,\delta})$. Combining this with (3.7) gives

$$\begin{aligned}\mathbb{E}M_r^{(\delta)}(t) &= \frac{\Lambda_t^r}{r!} \int_E P(B_{x,\delta})^r e^{-\Lambda_t P(B_{x,\delta})} P(\mathrm{d}x) \\ &= \frac{\Lambda_t^r}{r!} \int_0^1 u^{1+r} e^{-\Lambda_t u} \boldsymbol{\nu}_\delta(\mathrm{d}u) \\ &= \frac{\Lambda_t^r}{r!} \mathfrak{L}_{1+r}^{(\delta)}(\Lambda_t),\end{aligned} \tag{4.43}$$

where $\mathfrak{L}_{1+r}^{(\delta)}(\cdot)$ stands for the Laplace transform of the measure $\boldsymbol{\nu}_\delta^{1+r}(\mathrm{d}u) = u^{1+r}\boldsymbol{\nu}_\delta(\mathrm{d}u)$. Now according to equality (A.1) of Proposition A.1 from Appendix A, we know that for all $0 < \varepsilon < 1$,

$$\boldsymbol{\nu}_\delta^{1+r}\big([0,\varepsilon]\big) = -\varepsilon^{1+r} \nu_\delta(\varepsilon) + (1+r) \int_0^\varepsilon u^r \nu_\delta(u)\, \mathrm{d}u. \tag{4.44}$$

The assumption on the function ν_δ implies that

$$\varepsilon^{1+r} \nu_\delta(\varepsilon) = \varepsilon^{1+r-\alpha} \ell(1/\varepsilon) \tag{4.45}$$

and

$$\int_0^{\varepsilon} u^r \nu_\delta(u)\,\mathrm{d}u = \int_0^{\varepsilon} u^{r-\alpha}\ell(1/u)\,\mathrm{d}u \sim \frac{\varepsilon^{1+r-\alpha}\ell(1/\varepsilon)}{1+r-\alpha}, \tag{4.46}$$

as $\varepsilon \to 0$, where the equivalent follows from Karamata's theorem (Karamata [22]). Combining (4.44), (4.45) and (4.46) leads to

$$\boldsymbol{\nu}_\delta^{1+r}\big([0,\varepsilon]\big) \sim \frac{\alpha}{1+r-\alpha}\varepsilon^{1+r-\alpha}\ell(1/\varepsilon), \tag{4.47}$$

as $\varepsilon \to 0$. Finally, applying the Tauberian theorem (see, e.g., Theorem 2, Section 5, Chapter 13 in Feller [12]) and using the fact that $\Gamma(2+r-\alpha) = (1+r-\alpha)\Gamma(1+r-\alpha)$, we deduce that

$$\mathfrak{L}_{1+r}^{(\delta)}(t) \sim \alpha\Gamma(1+r-\alpha)t^{-(1+r-\alpha)}\ell(t),$$

as $t \to +\infty$. From here, the result follows from the fact that $\Lambda_t \to +\infty$ as $t \to +\infty$ and identity (4.43). □

Appendix A: Basic properties of ν

The counting function ν, defined in (1.4), is non-increasing by definition. As ε tends to 0, $\nu(\varepsilon)$ increases towards $|\mathcal{S}|$, the cardinality of the support of P, which may, of course, be infinite. Since the masses p_a sum to 1, it may be easily observed that, for all $0 < \varepsilon \le 1$,

$$\nu(\varepsilon) \le \varepsilon^{-1}.$$

Next, we recall general integration by parts formulas from which we will deduce additional properties of ν.

Proposition A.1. *Let μ be any positive measure on* $[0,1]$. *Then, for all $\tau \ge 1$ and all $0 < \varepsilon < 1$, we have the two identities*

$$\int_{[0,\varepsilon]} x^\tau \mu(\mathrm{d}x) = -\varepsilon^\tau \mu\big([\varepsilon,1]\big) + \tau \int_{[0,\varepsilon]} x^{\tau-1}\mu\big([x,1]\big)\,\mathrm{d}x, \tag{A.1}$$

$$\int_{[\varepsilon,1]} x^\tau \mu(\mathrm{d}x) = +\varepsilon^\tau \mu\big([\varepsilon,1]\big) + \tau \int_{[\varepsilon,1]} x^{\tau-1}\mu\big([x,1]\big)\,\mathrm{d}x. \tag{A.2}$$

The proof is a standard application of Fubini's theorem and is thus omitted. Note that, in the above, we did not assume finiteness of the integrals. When an integral on one side is infinite, the above should be interpreted to mean that the integral on the other side is infinite as well. We now reproduce an argument presented at the end of Section 3 in Gnedin, Hansen and Pitman [14]. Letting $\boldsymbol{\nu}$ be the counting measure defined in (1.3), it may be easily seen that, for all $\tau \ge 1$ and all $0 < \varepsilon < 1$, by (A.2) we have

$$\varepsilon^\tau \nu(\varepsilon) + \tau \int_{[\varepsilon,1]} x^{\tau-1}\nu(x)\,\mathrm{d}x = \int_{[\varepsilon,1]} x^\tau \boldsymbol{\nu}(\mathrm{d}x) = \sum_{a:p_a \ge \varepsilon} p_a^\tau \le 1.$$

Taking the limit as $\varepsilon \to 0$ and applying dominated convergence gives

$$\lim_{\varepsilon\to 0}\left(\varepsilon^{\tau}\nu(\varepsilon)+\tau\int_{[\varepsilon,1]}x^{\tau-1}\nu(x)\,\mathrm{d}x\right)=\lim_{\varepsilon\to 0}\sum_{a:p_a\geq\varepsilon}p_a^{\tau}=\sum_a p_a^{\tau}\leq 1, \tag{A.3}$$

with equality holding if and only if $\tau = 1$. Monotonicity guarantees the convergence of the integral in (A.3), which, together with the result of (A.3), implies that $\varepsilon^{\tau}\nu(\varepsilon)$ has a limit as $\varepsilon \to 0$. Finally, if this limit was $c > 0$, this would contradict the convergence of the integral in (A.3) since we would have $x^{\tau-1}\nu(x) \sim c/x$ as $x \to 0$, which, in turn, would imply the integrability of $1/x$ at 0. Taking $\tau = 1$ gives the following.

Corollary A.1.

$$\int_0^1 \nu(x)\,\mathrm{d}x = 1 \quad \textit{and} \quad \lim_{\varepsilon\to 0}\varepsilon\nu(\varepsilon)=0.$$

Appendix B: On the accrual function

In this appendix, we briefly discuss some properties of the accrual function F, which is defined for all $0 \leq \varepsilon \leq 1$ by

$$F(\varepsilon)=\int_{[0,\varepsilon]}x\boldsymbol{\nu}(\mathrm{d}x),$$

and of the bound (1.9) provided by Ohannessian and Dahleh [28]. First, note that (A.1) establishes that the counting and accrual functions are related through the formula

$$F(\varepsilon)=-\varepsilon\nu(\varepsilon)+\int_{[0,\varepsilon]}\nu(x)\,\mathrm{d}x, \tag{B.1}$$

for all $0 < \varepsilon < 1$. The next result shows that a straight-forward application of (1.9) provides, at least in the pure power setting, rate optimal bounds up to a log term. Recall that, in the regularly varying setting, rate optimal bounds are ones that, asymptotically, behave as in (1.7).

Proposition B.1. *Suppose that, for some constants $0 < C_- < C_+ < +\infty$ and some $\alpha \in (0, 1)$, the counting function ν satisfies $C_-\varepsilon^{-\alpha} \leq \nu(\varepsilon) \leq C_+\varepsilon^{-\alpha}$, for all $0 < \varepsilon < 1$. Then, for all $n \geq 1$, the expected missing mass satisfies*

$$\left(1-\frac{1}{n}\right)^n\frac{C_\alpha^-}{n^{1-\alpha}}\leq\mathbb{E}M_{n,0}\leq\frac{1+C_\alpha^+(\log n)^{1-\alpha}}{n^{1-\alpha}},$$

where $C_\alpha^- = \max\{0, C_-/(1-\alpha) - C_+\}$ and $C_\alpha^+ = (1-\alpha)^{1-\alpha}\{C_+/(1-\alpha) - C_-\}$.

The lower bound follows easily from (B.1), the bounds on ν, and the choice $\varepsilon = 1/n$ in the lower bound of (1.9). Similarly, the upper bound follows easily from (B.1), the fact that $(1-\varepsilon)^n \leq e^{-n\varepsilon}$, the bounds on ν, and the choice $\varepsilon = (1-\alpha)(\log n)/n$ in the upper bound of (1.9).

Supposing that the constant $C_\alpha^- > 0$, it follows that, since $(1 - n^{-1})^n \geq e^{-1}/2$ for large enough n, the lower bound in Proposition B.1 is rate optimal.

Appendix C: On lower bounds

This appendix gives an interesting and known result, versions of which can be found in for example, Lemma 4.1 of Almudevar, Bhattacharya and Sastri [1] or Lemma 1 of Zhang [37].

Theorem C.1. *Suppose that* $|\mathcal{S}| = \infty$. *Then there exists a sequence* (k_n) *of positive integers, with* $k_n \to \infty$ *as* $n \to \infty$, *such that, for any* $r \geq 0$,

$$\liminf_{n\to\infty} \frac{k_n^{r+1}}{\binom{n}{r}} \mathbb{E}M_{n,r} \geq e^{-1}$$

and

$$\liminf_{n\to\infty} \frac{k_n^{r+1}}{\binom{k_n}{r}} \mathbb{E}M_{k_n,r} \geq e^{-1}.$$

Proof. We only prove the first inequality as the proof of the second is similar. Let a_n be an element of $\mathcal{A}$ with $n \leq 1/p_{a_n}$ and let $k_n = \lfloor 1/p_{a_n} \rfloor$, where $\lfloor \cdot \rfloor$ denotes the floor function. Note that $n \leq k_n$, $(1 - p_{a_n}) \leq k_n p_{a_n} \leq 1$, and $p_{a_n} \to 0$ as $n \to \infty$. For $n > r$ we therefore have

$$\begin{aligned}
\frac{k_n^{r+1}}{\binom{n}{r}} \mathbb{E}M_{n,r} &\geq k_n^{r+1} p_{a_n}^{r+1} (1 - p_{a_n})^{n-r} \\
&\geq (1 - p_{a_n})^{n+1} \\
&\geq (1 - p_{a_n}) \left(1 - \frac{1}{k_n}\right)^n \\
&\geq (1 - p_{a_n}) \left(1 - \frac{1}{k_n}\right)^{k_n}.
\end{aligned}$$

The result follows by observing that, for a fixed r, the term on the right-hand side of the last inequality tends to e^{-1} as $n \to \infty$. □

Acknowledgements

The authors wish to thank the Editor and the two anonymous referees whose comments led to improvements in the presentation of this paper. In particular, we thank them for showing us a nicer form for $c(r)$. The work of G. Decrouez and Q. Paris was supported by the Russian Academic Excellence Project 5-100.

References

[1] Almudevar, A., Bhattacharya, R.N. and Sastri, C.C.A. (2000). Estimating the probability mass of unobserved support in random sampling. *J. Statist. Plann. Inference* **91** 91–105. MR1792366

[2] 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. MR3556773

[3] Berend, D. and Kontorovich, A. (2012). The missing mass problem. *Statist. Probab. Lett.* **82** 1102–1110. MR2915075

[4] Berend, D. and Kontorovich, A. (2013). On the concentration of the missing mass. *Electron. Commun. Probab.* **18** no. 3, 7. MR3011530

[5] Chao, A. (1981). On estimating the probability of discovering a new species. *Ann. Statist.* **9** 1339–1342. MR0630117

[6] Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scand. J. Stat.* **11** 265–270. MR0793175

[7] Chao, A. and Lee, S.-M. (1992). Estimating the number of classes via sample coverage. *J. Amer. Statist. Assoc.* **87** 210–217. MR1158639

[8] Chao, A., Lee, S.-M. and Chen, T.-C. (1988). A generalized Good's nonparametric coverage estimator. *Chinese J. Math.* **16** 189–199. MR0993611

[9] Chen, S.F. and Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Comput. Speech Lang.* **13** 359–394.

[10] Efron, B. and Thisted, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika* **63** 435–447.

[11] Esty, W.W. (1983). A normal limit law for a nonparametric estimator of the coverage of a random sample. *Ann. Statist.* **11** 905–912. MR0707940

[12] Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*, *Vol. II*, 2nd ed. New York–London–Sydney: Wiley. MR0270403

[13] Gandolfi, A. and Sastri, C.C.A. (2004). Nonparametric estimations about species not observed in a random sample. *Milan J. Math.* **72** 81–105. MR2099128

[14] Gnedin, A., Hansen, B. and Pitman, J. (2007). Notes on the occupancy problem with infinitely many boxes: General asymptotics and power laws. *Probab. Surv.* **4** 146–171. MR2318403

[15] Good, I.J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika* **40** 237–264. MR0061330

[16] Good, I.J. and Toulmin, G.H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika* **43** 45–63. MR0077039

[17] Grabchak, M. and Cosme, V. (2016). On the performance of turing's formula: A simulation study. *Comm. Statist. Simulation Comput.* To appear.

[18] Harris, B. (1959). Determining bounds on integrals with applications to cataloging problems. *Ann. Math. Stat.* **30** 521–548. MR0102876

[19] Harris, B. (1968). Statistical inference in the classical occupancy problem unbiased estimation of the number of classes. *J. Amer. Statist. Assoc.* **63** 837–847. MR0231480

[20] Holst, L. (1981). Some asymptotic results for incomplete multinomial or Poisson samples. *Scand. J. Stat.* **8** 243–246. MR0642805

[21] Johnson, N.L. and Kotz, S. (1977). *Urn Models and Their Application*: *An Approach to Modern Discrete Probability Theory*. *Wiley Series in Probability and Mathematical Statistics*. New York–London–Sydney: John Wiley & Sons. MR0488211

[22] Karamata, J. (1933). Sur un mode de croissance régulière. Théorèmes fondamentaux. *Bull. Soc. Math. France* **61** 55–62. MR1504998

[23] Karlin, S. (1967). Central limit theorems for certain infinite urn schemes. *J. Math. Mech.* **17** 373–401. MR0216548
[24] Khanloo, B.Y.S. and Haffari, G. (2015). Novel Bernstein-like concentration inequalities for the missing mass. In *Proceedings of the* 31*st Conference on Uncertainty in Artificial Intelligence*.
[25] Mao, C.X. and Lindsay, B.G. (2002). A Poisson model for the coverage problem with a genomic application. *Biometrika* **89** 669–681. MR1929171
[26] McAllester, D. and Ortiz, L. (2004). Concentration inequalities for the missing mass and for histogram rule error. *J. Mach. Learn. Res.* **4** 895–911. MR2076001
[27] McAllester, D.A. and Schapire, R. (2000). On the convergence rate of Good–Turing estimators. In *Proceedings of the* 13*th Annual Conference on Computational Learning Theory* 1–6.
[28] Ohannessian, M.I. and Dahleh, M.A. (2010). Distribution-dependent performance of the Good–Turing estimator for the missing mass. In *Proceedings of the* 19*th International Symposium on Mathematical Theory of Networks and Systems* 679–682.
[29] Ohannessian, M.I. and Dahleh, M.A. (2012). Rare probability estimation under regularly varying heavy tails. In *Proceedings of the* 25*th Annual Conference on Learning Theory* **23** 21.1–21.24.
[30] Orlitsky, A., Santhanam, N.P. and Zhang, J. (2004). Universal compression of memoryless sources over unknown alphabets. *IEEE Trans. Inform. Theory* **50** 1469–1481. MR2095850
[31] Robbins, H. (1955). A remark on Stirling's formula. *Amer. Math. Monthly* **62** 26–29. MR0069328
[32] Robbins, H.E. (1968). Estimating the total probability of the unobserved outcomes of an experiment. *Ann. Math. Stat.* **39** 256–257. MR0221695
[33] Starr, N. (1979). Linear estimation of the probability of discovering a new species. *Ann. Statist.* **7** 644–652. MR0527498
[34] Thisted, R. and Efron, B. (1987). Did Shakespeare write a newly-discovered poem? *Biometrika* **74** 445–455. MR0909350
[35] Zhang, C.-H. (2005). Estimation of sums of random variables: Examples and information bounds. *Ann. Statist.* **33** 2022–2041. MR2211078
[36] Zhang, C.-H. and Zhang, Z. (2009). Asymptotic normality of a nonparametric estimator of sample coverage. *Ann. Statist.* **37** 2582–2595. MR2543704
[37] Zhang, Z. (2016). Domains of attraction on countable alphabets. *Bernoulli*. To appear.
[38] Zhang, Z. and Huang, H. (2007). Turing's formula revisited. *J. Quant. Linguist.* **14** 222–241.
[39] Zhang, Z. and Huang, H. (2008). A sufficient normality condition for Turing's formula. *J. Nonparametr. Stat.* **20** 431–446. MR2424251

Received February 2016 and revised October 2016