Statistics and Probability Letters 82 (2012) 1102–1110

Contents lists available at SciVerse ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

The missing mass problem

Daniel Berend [a,b], Aryeh Kontorovich [a,*]

[a] Department of Computer Science, Ben-Gurion University, Beer Sheva, 84105, Israel

[b] Department of Mathematics, Ben-Gurion University, Beer Sheva, 84105, Israel

ARTICLE INFO

Article history:
Received 7 November 2011
Received in revised form 16 February 2012
Accepted 16 February 2012
Available online 23 February 2012

Keywords:
Missing mass
Probability estimate
Sampling

ABSTRACT

We give tight lower and upper bounds on the expected missing mass for distributions over finite and countably infinite spaces. An essential characterization of the extremal distributions is given. We also provide an extension to totally bounded metric spaces that may be of independent interest.

1. Introduction

1.1. Background

Let S be a countable set endowed with a probability measure P. Suppose that $X_1, \ldots, X_t$ are drawn independently from S according to P. Define the *missing mass* U_t as the following random variable:

$$U_t = P(S \setminus \{X_1, \ldots, X_t\}). \tag{1}$$

In words, U_t is the total probability mass of the elements of S that were not observed in the t samples.

The missing mass is a quantity of interest in almost any application involving sampling from a large discrete set, whether it be fish in a pond or words in a language corpus. (Famously, Alan Turing developed what became known as Good–Turing frequency estimators Good, 1953 as part of his work on cracking the Enigma cypher during WWII; see the account in Good (2000)). We note right away that

$$\mathbf{E}[U_t] = \sum_{s \in S} p_s (1 - p_s)^t$$

where $p_s = P(\{s\})$ and that $\mathbf{E}[U_t] \to 0$ as $t \to \infty$ (the latter follows from Lebesgue's Dominated Convergence Theorem). Observe also that $U_t \to 0$ almost surely as $t \to \infty$; one way of seeing this is to apply the deviation inequality

$$P[|U_t - \mathbf{E}[U_t]| \geq \varepsilon] \leq 2e^{-t\varepsilon^2} \tag{2}$$

of McAllester and Ortiz (2003, Theorem 16) together with the Borel–Cantelli Lemma.

* Corresponding author.
E-mail addresses: berend@cs.bgu.ac.il (D. Berend), karyeh@cs.bgu.ac.il (A. Kontorovich).

0167-7152/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2012.02.014

The topic of interest of this work is the rate of decay of $\mathbf{E}[U_t]$. For example, when S is finite, we have the trivial estimate

$$\mathbf{E}[U_t] \leq (1 - p_{\min})^t$$

where $p_{\min} = \min_{s \in S} p_s$ and we assume without loss of generality that P has full support. Of course, this bound is not distribution-free, since it depends on $p_{\min}$. Is a distribution-free estimate possible, at least for finite S? What about countable S? How about lower bounds on the decay rate of the missing mass? These and related questions are investigated in this work.

1.2. Related work

The missing mass problem is an unavoidable feature of density estimation, where non-zero density must be assigned to unobserved regions. Let us briefly recall the celebrated Good–Turing estimator for the missing mass. Given the sample $X = \{X_1, \ldots, X_t\}$, define $Y^{(1)} \subseteq X$ to consist of those X_i that occur exactly once. The Good–Turing missing mass estimator is given by

$$\hat{U}_t = \frac{1}{t} \left| Y^{(1)} \right|;$$

this is the proportion of frequency-1 elements. An attractive feature of this estimator is its diminishing bias:

$$\mathbf{E}[\hat{U}_t] - \mathbf{E}[U_t] = \frac{1}{t} \mathbf{E}[U_t^{(1)}] \tag{3}$$

where $U_t^{(1)} = P(Y_t^{(1)})$ is the random variable corresponding to the total mass of the frequency-1 items; this variant of Good's theorem is proved in McAllester and Schapire (2000, Theorem 1). Additionally, both the missing mass U_t and its estimate $\hat{U}_t$ are tightly concentrated about their expectations (inequality (2) establishes this for U_t; see McAllester and Schapire, 2000 for other deviation estimates).

More generally, studying the behavior of U_t falls into the framework of the theory of *large numbers of rare events* (LNRE), put forth by Khmaladze (1988). In particular, a special case of Khmaladze (2011, Theorem 2) provides precise asymptotics for U_t in the $t = \lambda n$ regime. The focus in our work is somewhat different, as n is fixed and our bounds hold for all $n, t \in \mathbb{N}$.

2. The main results

Although (2) and (3) provide a computationally efficient estimator of $\mathbf{E}U_t$, they do not yield any a priori information about the magnitude of the latter.

To state results, it will be convenient to define the *plateau length* ℓ of a probability distribution P over $\mathbb{N}$:

$$\ell(P) = \sup_{0 < \alpha < 1} |\{i \in \mathbb{N} : \alpha/2 \leq p_i < \alpha\}| \tag{4}$$

where $p_i = P(\{i\})$. (Note that $\ell(P) = \infty$ is possible.)

Our first two results deal with upper and lower bounds on $\mathbf{E}U_t$. We use the notation $[n] = \{1, 2, \ldots, n\}$ throughout the work.

Theorem 1. *The expected missing mass is bounded above as follows:*

(i) *For $n \in \mathbb{N}$ and $S = [n]$,*

$$\mathbf{E}U_t \leq \begin{cases} e^{-t/n}, & t \leq n, \\ \dfrac{n}{et}, & t > n. \end{cases}$$

(ii) *For $S = \mathbb{N}$,*

$$\mathbf{E}U_t \leq \frac{\ell(P)}{ct},$$

where c is a universal constant.

Remark 2. It is possible to somewhat (but not by much; see Proposition 3) improve the bound in (i) in some regimes of n and t; this will become apparent from our proofs. The bound $\frac{n}{et}$ holds everywhere, but is vacuous when $et < n$. A slightly better bound of $n(1 - 1/n)^n/t$ was obtained by Boppana in a very elegant way, in response to our question (Boppana, 2011). We took an entirely different route, which also basically characterizes the extremal distribution.

Proposition 3. *The estimates in Theorem 1 are essentially tight:*

(i) *For each $n \in \mathbb{N}$ and $t > n$, there is a distribution on $[n]$ such that*

$$\mathbf{E}U_t \geq c\frac{n-1}{t},$$

where c is an absolute constant.

(ii) *For each integer $a > 1$, there is a distribution P over $S = \mathbb{N}$ such that $\ell(P) = a$ and*

$$\mathbf{E}U_t \geq c\frac{a}{t}, \quad t > a,$$

where c is an absolute constant.

Furthermore, if we allow distributions with infinite plateau length, then no nontrivial uniform (or even pointwise) bound on $\mathbf{E}U_t$ is possible:

Proposition 4. *For any sequence $1 > r_1 > r_2 > \cdots$ decreasing to 0, there is a distribution on $S = \mathbb{N}$ such that $\mathbf{E}U_t > r_t$ for all $t \geq 1$.*

Next, we turn our attention to extremizing distributions for finite S. These turn out to exhibit a fairly regular behavior, with essentially a single phase transition. Since $\mathbf{E}U_t$ is a symmetric function of the $\{p_i\}$, we henceforth assume that $p_1 \leq p_2 \leq \cdots \leq p_n$. In the sequel, the vector $(p_1, \ldots, p_n)$ will be denoted by $\mathbf{p}$.

Theorem 5. *Let $|S| = n < \infty$. Then:*

(i) *Every local maximum $\mathbf{p}^*$ of $\mathbf{E}U_t$ is of the form*

$$p_1^* = p_2^* = p_3^* = \cdots = p_{n-1}^* \leq p_n^*$$

(that is, $\mathbf{p}^*$ consists of one "heavy" element and $n-1$ identical "light" ones, where the possibility of heavy = light is not excluded).

(ii) *There exists a threshold $\tau = \tau(n) > n$ such that:*

(a) *For $t < \tau$, there is a unique global maximum*

$$p_1^* = p_2^* = p_3^* = \cdots = p_{n-1}^* = p_n^* = \frac{1}{n}.$$

(b) *For $t > \tau$, there is a unique global maximum and it has the form*

$$p_1^* = p_2^* = p_3^* = \cdots = p_{n-1}^* < p_n^*.$$

(iii) *As $n \to \infty$,*

$$\tau = n + \sqrt{2n}(1 + o(1)).$$

(iv) *For $t \geq n + \sqrt{2n}$,*

$$\frac{1}{t+1} < p_2^* < \frac{1}{t+1} + e^{-\sqrt{n/2}}. \tag{5}$$

Remark 6. We have not excluded the possibility that for $t = \tau$, both of the distributions described in (ii) attain the maximum. (Of course, this seems highly improbable.)

3. Proofs

We begin with an elementary lemma, whose proof is omitted.

Lemma 7. *Consider the function $f(x) = x(1-x)^t$ on the interval $[0, 1]$ for an arbitrary fixed $t > 0$.*

(i) *For $t > 0$, f increases on $(0, 1/(t+1))$, decreases on $(1/(t+1), 1)$, and achieves its maximum at $x = 1/(t+1)$, where it is bounded above by $\frac{1}{et}$.*

(ii) *The derivative f' decreases on $(0, 2/(t+1))$ and increases on $(2/(t+1), 1)$.*

Proof of Theorem 1.

(i) It follows from (9) that, for $t \leq n$, the expected missing mass is maximized when $p_1^* = p_2^* = \cdots = p_n^* = 1/n$, yielding the bound

$$\mathbf{E}U_t = (1 - 1/n)^t \leq e^{-t/n}.$$

For general t, an application of Lemma 7 yields

$$\mathbf{E}U_t = \sum_{i=1}^{n} p_i(1-p_i)^t \leq \frac{n}{et}.$$

(ii) Let P be a distribution on $S = \mathbb{N}$. Then

$$\begin{aligned}\mathbf{E}U_t &= \sum_{i=1}^{\infty} p_i(1-p_i)^t \\ &= \sum_{i:p_i \geq 1/(t+1)} p_i(1-p_i)^t + \sum_{i:p_i < 1/(t+1)} p_i(1-p_i)^t \equiv E_1 + E_2.\end{aligned}$$

By Lemma 7,

$$\begin{aligned}E_1 &= \sum_{i:p_i \geq 1/(t+1)} p_i(1-p_i)^t \\ &= \sum_{j=0}^{\lfloor \log_2(t+1) \rfloor} \sum_{i:2^j/(t+1) \leq p_i < 2^{j+1}/(t+1)} p_i(1-p_i)^t \\ &\leq \sum_{j=0}^{\lfloor \log_2(t+1) \rfloor} \ell(P) \frac{2^j}{t+1} \left(1 - \frac{2^j}{t+1}\right)^t \\ &< c' \frac{\ell(P)}{t+1} \sum_{j=0}^{\infty} (2/e)^j \\ &\leq c'' \frac{\ell(P)}{t+1}, \end{aligned} \tag{6}$$

for appropriate absolute constants c', c''.

An analogous argument shows that

$$E_2 \leq c''' \frac{\ell(P)}{t+1}. \tag{7}$$

Combining (6) and (7), we obtain the claim. □

Proof of Proposition 3.

(i) Define the distribution $\mathbf{p}$ by

$$x = p_1 = p_2 = \cdots = p_{n-1} \leq p_n = 1 - (n-1)x,$$

where $x = 1/(t+1)$. Then

$$\mathbf{E}U_t > \frac{n-1}{t+1}\left(1 - \frac{1}{t+1}\right)^t = \frac{n-1}{t}\left(1 - \frac{1}{t+1}\right)^{t+1} \geq \frac{8(n-1)}{27t}.$$

(ii) For any $a \in \mathbb{N}$, define $\mathbf{p}$ as follows:

$$p_1 = p_2 = \cdots = p_a = \frac{1}{2a}; p_{a+1} = \cdots = p_{2a} = \frac{1}{4a}; \ldots; p_{ka+1} = \cdots = p_{(k+1)a} = \frac{1}{2^k a}; \ldots.$$

Then, defining $\kappa = \lceil \log_2(t/a) \rceil$, we have for $t > a$

$$\begin{aligned}\mathbf{E}U_t &= \sum_{k=1}^{\infty} \frac{1}{2^k}\left(1 - \frac{1}{2^k a}\right)^t > \frac{1}{2^\kappa}\left(1 - \frac{1}{2^\kappa a}\right)^t = \frac{1}{2} \cdot \frac{1}{2^{\kappa-1}}\left(1 - \frac{1}{2^\kappa a}\right)^t > \frac{a}{2t}\left(1 - \frac{1}{t}\right)^t \\ &\geq \frac{4a}{27t}. \quad \square\end{aligned}$$

Proof of Proposition 4. Let $1 > r_1 > r_2 > \cdots$ be a sequence decreasing to 0. Observe that

$$\mathbf{E}U_t = \sum_{i=1}^{n} p_i(1-p_i)^t \geq \sum_{i:p_i < 1/t^2} p_i\left(1 - \frac{1}{t^2}\right)^t = \left(1 - \frac{1}{t^2}\right)^t \sum_{i:p_i < 1/t^2} p_i.$$

Select $\tau > 10$ such that $r_\tau < 0.9$. Then we can choose (p_i) such that

$$\left(1 - \frac{1}{t^2}\right)^t \sum_{i:p_i < 1/t^2} p_i > r_t, \quad t \geq \tau. \tag{8}$$

Indeed, $(1-1/t^2)^t > 0.9$ for $t \geq \tau > 10$. Thus, for $t = \tau$, (8) is satisfied by any (p_i) with $p_i < 1/\tau^2$ for all $i \in \mathbb{N}$. For each $t > \tau$, choose a finite sequence (p_{it}) such that $p_{it} < 1/(t+1)^2$ for each i and

$$\sum p_{it} = r_t - r_{t+1}.$$

Let $\mathbf{p}$ be the distribution obtained by concatenating all the sequences $(p_{it})_{t>\tau}$ and the number $1 - r_\tau$. To prove the claim for $t < \tau$, let us define the following "doubling operator" on distributions:

$$D((p_1, p_2, \ldots)) = (p_1/2, p_1/2, p_2/2, p_2/2, \ldots).$$

It is straightforward to verify that for all distributions $\mathbf{p}$ and all $t \in \mathbb{N}$,

$$\mathbf{E}_{D^k\mathbf{p}} U_t \nearrow 1 \quad \text{as } k \to \infty$$

(where the subscript of $\mathbf{E}$ specifies the distribution under which the expectation is taken). Thus, if $\mathbf{p}$ is a distribution that satisfies (8) for all $t \geq \tau$, there is some finite k such that $D^k\mathbf{p}$ makes the proposition hold. □

Proof of Theorem 5.

(i) For $n, t \in \mathbb{N}$, define $F : [0,1]^n \to \mathbb{R}$ by

$$F(\mathbf{x}) = \sum_{i=1}^n x_i(1-x_i)^t = \sum_{i=1}^n f(x_i)$$

(where $f(x) = x(1-x)^t$). An elementary application of Lagrange multipliers shows that, under the constraint $\sum_{i=1}^n x_i = 1$, a necessary condition for an extremum is

$$\frac{\partial F}{\partial x_i} = \frac{\partial F}{\partial x_j}, \quad i, j \in [n].$$

Lemma 7 leaves two possibilities for an extreme point $\mathbf{p}^*$: either all the p_i^* take the value $1/n$ (we call such distributions *univalent*) or the p_i^* take two values $\pi < \bar{\pi}$, with $f'(\pi) = f'(\bar{\pi}) < 0$ (we call such distributions *bivalent*).
In the bivalent case, we have, without loss of generality,

$$p_1^* = p_2^* = \cdots = p_k^* = \pi < \bar{\pi} = p_{k+1}^* = p_{k+2}^* = \cdots = p_n^*$$

for some $1 < k < n$. Define the Lagrangian

$$L(\mathbf{x}, \lambda) = F(\mathbf{x}) + \lambda(g(\mathbf{x}) - 1),$$

where $g(\mathbf{x}) = \sum_{i=1}^n x_i$ and the associated $(n+1) \times (n+1)$ matrix $H = H(\mathbf{x}, \lambda)$, where

$$H_{ij} = \begin{cases} \dfrac{\partial^2 L}{\partial x_i \partial x_j} = t(x_i(t+1) - 2)(1-x_i)^{t-2}, & i = j \leq n, \\ \dfrac{\partial^2 L}{\partial x_i \partial x_j} = 0, & i \neq j \leq n, \\ \dfrac{\partial g}{\partial x_i} = 1, & i \leq n,\ j = n+1, \\ \dfrac{\partial g}{\partial x_j} = 1, & j \leq n,\ i = n+1, \\ 0, & i = j = n+1. \end{cases}$$

Suppose $k \leq n-2$ and consider the 3×3 lower right submatrix

$$B = B(\mathbf{x}) = \begin{pmatrix} t(x_{n-1}(t+1) - 2)(1-x_{n-1})^{t-2} & 0 & 1 \\ 0 & t(x_n(t+1) - 2)(1-x_n)^{t-2} & 1 \\ 1 & 1 & 0 \end{pmatrix};$$

note that our assumption on k forces $B_{11} = B_{22}$. The second-order necessary condition for $\mathbf{p}^*$ to be a local maximum is that a sequence of bordered Hessians, including $\det(B(\mathbf{p}^*))$, be nonnegative. Since $f'(\pi) = f'(\bar{\pi}) < 0$ and f' decreases on $(0, 2/(t+1))$ and increases on $(2/(t+1), 1)$, it follows that $\pi < 2/(t+1)$ and $\bar{\pi} > 2/(t+1)$. This implies that $B_{11} = B_{22} > 0$. Denoting this common value by b, we have

$$\det(B) = -2b < 0.$$

The necessary condition is thus violated, leaving two possibilities: the univalent case $p_i^* \equiv 1/n$ and the bivalent case with $k = n-1$:

$$\frac{1}{t+1} < p_1^* = p_2^* = \cdots = p_{n-1}^* < \frac{2}{t+1} < p_n. \tag{9}$$

(ii) We saw above that $\mathbf{E}U_t$ is always maximized by a distribution $\mathbf{p}$ of the form

$$x = p_1 = p_2 = \cdots = p_{n-1} \le p_n = 1 - (n-1)x.$$

For distributions of this form, we have

$$\mathbf{E}U_t = G_t(x) = (n-1)x(1-x)^t + (1-(n-1)x)((n-1)x)^t, \tag{10}$$

where G_t is defined on $[0, 1/n]$. Note that $x = 1/n$ corresponds to the univalent (uniform) distribution, while $x < 1/n$ corresponds to a bivalent distribution.

We claim the existence of a function $\tau : \mathbb{N} \to \mathbb{N}$ such that:

(a) for $t < \tau(n)$, G_t has the unique maximizer $x^* = 1/n$;

(b) for $t > \tau(n)$, G_t has the unique maximizer $x^* < 1/n$.

(In principle, it may be possible for G_t to have two distinct maxima on $[0, 1/n]$ for $t = \tau(n)$, but this is rather implausible.)

For $t \le n$, (9) implies that G_t has the unique maximizer $x^* = 1/n$; this shows that $\tau(n) > n$ (if the function τ described in (a) and (b) exists at all).

Now define the function $R_t(x) = G_t(x)/G_t(1/n)$. Then

$$R_t(x) = (n-1)x\left(\frac{1-x}{1-1/n}\right)^t + (1-(n-1)x)\left(\frac{(n-1)x}{1-1/n}\right)^t. \tag{11}$$

For $x < 1/n$, the first term on the right-hand side of (11) grows exponentially with t, and certainly $R_t(x) > 1$ is achieved for some finite t. But this means that G_t has a unique maximum at some $x < 1/n$ and so any function $\tau(n)$ satisfying (a) and (b) must be finite for all n.

Suppose that t is such that G_t achieves a maximum at $x < 1/n$. It follows from the uniqueness proof below and from (15) that the maximizer x^* of G_t is contained in the interval $I_t = (1/(t+1), 1/t)$. We claim that $R_t(x) \ge 1$ implies $R_{t+1}(x) > 1$ for all $x \in I_t$; from here, the existence of τ satisfying (a) and (b) follows immediately. Treating t as a continuous variable, we have

$$\frac{dR_t(x)}{dt} = (n-1)x\log\left(\frac{1-x}{1-\frac{1}{n}}\right)\left(\frac{1-x}{1-\frac{1}{n}}\right)^t + (1-(n-1)x)\log(nx)\,(nx)^t.$$

We establish the monotonicity claim by showing that

$$\frac{dR_t(x)}{dt} > 0, \quad t \ge n+1,\ x \in I_t. \tag{12}$$

Indeed, appealing to the inequalities $\xi^t \log\xi \ge -\frac{1}{et}$ for $0 < \xi < 1$ and $\xi^t \log\xi \ge \xi - 1$ for $\xi \ge 1$ (checked by elementary calculus), we see that the inequality

$$(n-1)x\left(\frac{1-x}{1-1/n} - 1\right) > \frac{1-(n-1)x}{et}.$$

is even stronger than (12). The latter will hold as long as

$$-entx^2 + (et + n - 1)x - 1 > 0,$$

and it suffices to verify the inequality at the endpoints $1/(t+1)$ and $1/t$ of I_t, which is straightforward. This proves the existence of τ as claimed in (a) and (b).

Uniqueness is established by noting (again, via elementary though rather tedious calculus) that G_t' vanishes at $x = 1/n$ and at not more than two points in the interval $[1/(t+1), 1/n)$, and is strictly positive on $[0, 1/(t+1))$.

(iii) For $n \in \mathbb{N}$ and $t \ne \tau(n)$, let x^* be the maximizer of the function G defined in (10), where we have dropped the subscript t. A sufficient condition for $G(x^*) > \left(1-\frac{1}{n}\right)^t$ to hold is $G\left(\frac{1}{t}\right) > \left(1-\frac{1}{n}\right)^t$. The latter, in turn, will hold as long as $(n-1)(1-1/t)^t/t > (1-1/n)^t$. We will show that the latter inequality holds for large n, if $t = n + \sqrt{2n}$. To this end, define the function

$$g(n) = \frac{n-1}{n+\sqrt{2n}}\left(1 - \frac{1}{n+\sqrt{2n}}\right)^{n+\sqrt{2n}} - \left(1-\frac{1}{n}\right)^{n+\sqrt{2n}}.$$

For $\nu \in (0, 1)$, define $\tilde{g}(\nu) = g(1/\nu)$ and expand it about $\nu = 0$:

$$\tilde{g}(\nu) = \frac{\sqrt{2}}{3e}\nu^{3/2} + O(\nu^2).$$

Since for this choice of t we have $G(1/t) - (1-1/n)^t = \Omega_+(n^{-3/2})$, it follows that

$$\tau(n) \le n + \sqrt{2n}, \quad n \gg 1. \tag{13}$$

To get a lower bound on τ, we estimate $G(x^*)$ from above:

$$G(x^*) \le \frac{n-1}{t+1}\left(1-\frac{1}{t+1}\right)^t + \left(1-\frac{n-1}{t}\right)\left(\frac{n-1}{t}\right)^t =: \bar{G}_t(n).$$

Now let

$$\begin{aligned} Q_t(n) &= \frac{\bar{G}_t(n)}{(1-1/n)^t} \\ &= \frac{n-1}{t+1}\left(\frac{1-1/(t+1)}{1-1/n}\right)^t + \left(1-\frac{n-1}{t}\right)\left(\frac{(n-1)/t}{1-1/n}\right)^t \end{aligned} \tag{14}$$

and note that $Q_t(n) < 1$ implies $t < \tau(n)$. Let us put $t = n + (1-\varepsilon)\sqrt{2n}$ for some $0 < \varepsilon < 1$, and observe that for this choice of t, the second term on the right-hand side of (14) is negligible:

$$\begin{aligned} \left(1-\frac{n-1}{t}\right)\left(\frac{(n-1)/t}{1-1/n}\right)^t &< \left(\frac{(n-1)/t}{1-1/n}\right)^t = \left(\frac{n}{t}\right)^t = \left(1-\frac{t-n}{t}\right)^t \\ &< \exp(-(t-n)) = \exp(-(1-\varepsilon)\sqrt{2n}). \end{aligned}$$

Let us now examine the asymptotic behavior of the first term in (14). To this end, define the functions

$$h(n) = \frac{n-1}{t+1}\left(\frac{1-1/(t+1)}{1-1/n}\right)^t$$

and $\tilde{h}(\nu) = h(1/\nu)$, and expand about $\nu = 0$:

$$\tilde{h}(\nu) = 1 - (2-\varepsilon)\varepsilon\nu + O(\nu^{3/2}).$$

Hence, for this choice of t, we have

$$\frac{G(x^*)}{(1-1/n)^t} \le 1 - (2-\varepsilon)\varepsilon n^{-1} + O(n^{-3/2}) + \exp(-(1-\varepsilon)\sqrt{2n}) < 1$$

for sufficiently large n. This, combined with (13), implies

$$\tau(n) = n + (1+o(1))\sqrt{2n}.$$

(iv) We claim that in the bivalent case, the maximizer x^* of G satisfies

$$\frac{1}{t+1} < x^* < \frac{1}{t}. \tag{15}$$

The first inequality follows from (9). To verify the second inequality it suffices to show that $G'(1/t) < 0$. Indeed,

$$\begin{aligned} \frac{G'(1/t)}{t(n-1)} &< -\left(\frac{1}{t}-\frac{1}{t-1}\right)\left(1-\frac{1}{t}\right)^{t-1} + \left(\frac{n-1}{t}\right)^{t-1} \\ &= -\frac{1}{t(t+1)}\left(1-\frac{1}{t}\right)^{t-1} + \left(\frac{n-1}{t}\right)^{t-1} \\ &< \left(\frac{n}{t}\right)^{t-1} - \frac{1}{t(t+1)(1-1/t)}\left(1-\frac{1}{t}\right)^t \\ &< \left(\frac{n}{t}\right)^{t-1} - \frac{1}{e(t+1)(t-1)} \le \left(\frac{n}{t}\right)^{t-1} - \frac{1}{et^2} \\ &\le \frac{n^2}{t^2}\left(\frac{n}{t}\right)^{t-3} - \frac{1}{et^2} \\ &\le \frac{1}{t^2}\left[\left(\frac{n}{n+\sqrt{2n}}\right)^{n-3} - \frac{1}{e}\right] < 0. \end{aligned}$$

To establish our claim, we seek a small $\delta > 0$ such that $G'(1/(t+1)+\delta) < 0$; any such δ will yield the bound $\frac{1}{t+1} < x^* < \frac{1}{t+1} + \delta$. Putting $p = \frac{1}{t+1} + \delta$, we have

$$\begin{aligned}\frac{G'(p)}{n-1} &= -\delta(t+1)(1-p)^{t-1} + (t-n+1-(n-1)\delta(t+1))((n-1)p)^{t-1} \\ &< -\delta(t+1)(1-p)^{t-1} + (t-n+1)((n-1)p)^{t-1} \\ &< -\delta t(1-p)^{t-1} + t((n-1)p)^{t-1}.\end{aligned}$$

Thus, to ascertain that $G'(p) < 0$ it suffices to show that

$$\delta > \left(\frac{(n-1)p}{1-p}\right)^{t-1}.$$

It follows from (15) that we may take $p < 1/t$, and hence

$$\left(\frac{(n-1)p}{1-p}\right)^{t-1} < \left(\frac{(n-1)/t}{1-1/(t+1)}\right)^{t-1} = \left(\frac{(n-1)(t+1)}{t^2}\right)^{t-1} < \left(\frac{n}{t}\right)^{t-1}.$$

Our assumption that $t \geq n + \sqrt{2n}$ implies

$$\left(\frac{n}{t}\right)^{t-1} \leq \left(\frac{n}{n+\sqrt{2n}}\right)^{t-1} \leq \exp(-\sqrt{n/2}).$$

Thus, we may take $\delta = e^{-\sqrt{n/2}}$. □

4. Application: missing mass in metric spaces

If P is a nondegenerate continuous distribution, then the missing mass as defined in (1) is trivially 1 for all $t \in \mathbb{N}$. To define a nontrivial extension of this notion to continuous spaces,[1] let us start with a metric probability space $(\mathcal{X}, P, d)$, whose σ-field is induced by the metric topology. For $x \in \mathcal{X}$, let $B_\varepsilon(x)$ be the ε-ball about x: $B_\varepsilon(x) = \{y \in \mathcal{X} : d(x,y) \leq \varepsilon\}$. For $S \subset \mathcal{X}$, define its ε-envelope, S_ε, to be

$$S_\varepsilon = \bigcup_{x \in S} B_\varepsilon(x).$$

For $\varepsilon > 0$, define the ε-covering number, $N(\varepsilon)$, of $\mathcal{X}$ as the minimal cardinality of a set $E \subset \mathcal{X}$ such that $\mathcal{X} = E_\varepsilon$. A space is *totally bounded* if $N(\varepsilon) < \infty$ for all $\varepsilon > 0$. Define the *ε-missing mass* of the sample $S = \{X_1, \ldots, X_t\}$ as the random variable

$$U_t(\varepsilon) = P(\mathcal{X} \setminus S_\varepsilon). \tag{16}$$

The expected ε-missing mass of totally bounded spaces is controlled via the covering numbers:

Theorem 8. *In a totally bounded metric probability space $(\mathcal{X}, P, d)$,*

$$\mathbf{E}U_t(\varepsilon) \leq \frac{N(\varepsilon)}{et}.$$

Proof. For a fixed $\varepsilon > 0$, let $\{e_1, e_2, \ldots, e_n\}$ be an ε-net for $\mathcal{X}$. For $i = 1, \ldots, n$, put $p_i = P(B_\varepsilon(e_i))$; note that $\sum p_i \geq 1$. Then, invoking Lemma 7, we have

$$\mathbf{E}U_t(\varepsilon) \leq \sum_{i=1}^{n} p_i(1-p_i)^t \leq \frac{n}{et}.$$ □

Acknowledgments

We thank Antonio Cuevas and Larry Wasserman for helpful correspondence. Many thanks also go to Roi Weiss and an anonymous referee for comments on and corrections to the manuscript.

References

Boppana, Ravi, 2011. Missing mass conjecture [answer]. Mathoverflow website. http://mathoverflow.net/questions/60722/missing-mass-conjecture.
Good, Irving J., 1953. The population frequencies of species and the estimation of population parameters. Biometrika 40, 237–264.
Good, Irving J., 2000. Turing's anticipation of empirical Bayes in connection with the cryptanalysis of the naval Enigma. J. Stat. Comput. Simul. 66 (2), 101–111.

[1] This problem is of interest in anomaly detection applications (Kontorovich et al., 2011).

Khmaladze, Estáte V., 1988. The statistical analysis of a large number of rare events. CWI Technical Report CWI. Department of Mathematical Statistics-R 8804, CWI.
Khmaladze, Estáte V., 2011. Convergence properties in certain occupancy problems including the Karlin–Rouault law. J. Appl. Probab. 48 (4), 1095–1113.
Kontorovich, Aryeh, Hendler, Danny, Menahem, Eitan, 2011. Metric anomaly detection via asymmetric risk minimization. Similarity-Based Pattern Analysis and Recognition (SIMBAD). pp. 17–30. doi:10.1007/978-3-642-24471-1_2.
McAllester, David A., Ortiz, Luis E., 2003. Concentration inequalities for the missing mass and for histogram rule error. J. Mach. Learn. Res. 4, 895–911.
McAllester, David A., Schapire, Robert E., 2000. On the convergence rate of Good–Turing estimators. In: Conference on Learning Theory, COLT, pp. 1–6.