Mean-Squared Accuracy of Good-Turing Estimator

Maciej Skorski
University of Luxembourg

***Abstract*—The briliant method due to Good and Turing allows for estimating objects missing in a sample. The problem, known under names "sample coverage" or "missing mass" goes back to their cryptographic work during WWII, but over years has found has many applications, including language modeling, inference in ecology, estimation of entropy and distribution properties.**

This work characterizes the maximal mean-squared error of the Good-Turing estimator, for any sample *and* alphabet size.

***Index Terms*—Good-Turing Estimator, Mean-Squared Error, Missing Mass, Sample Coverage, Non-linear Programming**

I. Introduction

A. Background

Let $X_1,\ldots,X_n \sim^{IID} p$ be a sample from a distribution $(p_s)_{s\in S}$ on a countable alphabet S, and $f_s = \#\{i : X_i = s\}$ be the empirical (observed) frequencies. The missing mass

$$M_0 = \sum_s p_s \mathbb{I}(f_s = 0), \tag{1}$$

which quantifies how much of the population is not covered by the sample, is of interest to statistics [1] and several applied disciplines such as ecology [2]–[5], quantitative linguistic [6]–[8], archeology [9] network design [10], [11], information theory [12], [13], and bio-molecular modeling [14], [15]. The most popular estimator due to Good-Turing [16] is given by:

$$\widehat{M} = \frac{1}{n}\sum_s \mathbb{I}(f_s = 1). \tag{2}$$

In this paper, the focus is on the maximal mean-squared error

$$MSE = \mathbf{E}\left[(\widehat{M} - M_0)^2\right], \tag{3}$$

under the constrained alphabet (upper-bounded support)

$$\#S = m. \tag{4}$$

B. Related Work

No prior work has studied the MSE under alphabet constraints; we thus review the closest (in spirit) results of [17] and [18] obtained for the unconstrained case $m = +\infty$.

The work [17] expressed the mean-squared error (3) in terms of *occupancy numbers*. More precisely, let

$$N_k = \sum_s \mathbb{I}(f_s = k) \tag{5}$$

(the number of elements observed exactly k times). Then [17]:

$$MSE = \frac{\mathbf{E}\left[\frac{2N_2}{n} + \frac{N_1}{n}\left(1 - \frac{N_1}{n}\right)\right]}{n} + O(n^{-2}), \tag{6}$$

and moreover $\max_{(p_s)\in\mathbb{P}(S)} MSE = \Theta(n^{-1})$ when $\#S = +\infty$, where $\mathbb{P}(S)$ is the set of probability measures on S.

The work [18] made an attempt to improve upon [17] and establish a sharp constant (still with no constraint on alphabets); building on the formulas from [17] and some further simplifications, it suggests to apply the method of Lagrange multipliers to prove that the maximum is achieved by a uniform distribution; noticeably, the proof was not given[1].

Beyond the scope of our problem are works on consistent estimation of missing mass [19], [20], concentration [21]–[26], expectation under constraints [27], [28], applications to distribution estimation [29]–[31], and others [32]–[34].

C. Our Contribution

- We determine the worst MSE given the sample size n and the alphabet size m; the maximizer is a Dirac-Uniform mixture with *phase transition* depending on the ratio $\frac{m}{n}$. Studying the alphabet constraint is natural, but also practically motivated (prior support bound); it well fits other research on missing mass under constraints [27], [28]; finally, it demonstrates useful optimization techniques.
- For $m = +\infty$, we obtain a *rigorous* proof of the result stated in [18]. The method of Lagrange multipliers cannot be applied that easily; the main issue is the unbounded dimension, another is that maximizers for finite dimensions are more complicated than uniform distributions.
- Python implementation with examples, on GitHub [35].

II. Results

A. Convenient Mean-Squared Error of Good-Turing Estimator

Below we give a formula which involves *first moments of occupancy numbers*, rather than variances as in prior works [17], [18]. Our expression is thus simpler to analyze.

Theorem 1. *For any distribution (p_s) we have:*

$$MSE = \frac{\frac{2\mathbf{E}[N_2]}{n} + \frac{\mathbf{E}[N_1]}{n}\left(1 - \frac{\mathbf{E}[N_1]}{n}\right)}{n} + O(n^{-2}), \tag{7}$$

The constant in $O(n^{-2})$ is independent of (p_s) and n.

B. Exponential Approximation / Poissonization

Moments of occupancy numbers are often approximated with *Poisson-like* expressions [36]–[38]. We develop such an approximation, and use later for constrained optimization.

Theorem 2. *For any distribution (p_s) we have:*

$$MSE = \frac{n\sum_s p_s^2 \mathrm{e}^{-np_s} + \sum_s p_s \mathrm{e}^{-np_s} - \left(\sum_s p_s \mathrm{e}^{-np_s}\right)^2}{n} + O(n^{-2}). \tag{8}$$

[1] The proof is omitted in both proceedings and public version.

C. Extreme Mean-Squared Error Behavior

Using non-linear programming (beyond Lagrange multipliers) we characterize the maximal MSE with respect to $\#S$.

Theorem 3. *For any $n \geqslant 2$ and $m = \#S \geqslant 2$, consider*

$$\begin{aligned} \max \quad & \alpha(w,c) = w(1+c)\mathrm{e}^{-c} - (w\mathrm{e}^{-c})^2 \\ \text{s.t.} \quad & 0 \leqslant w \leqslant 1, \quad w \leqslant \frac{m}{n}c. \end{aligned} \tag{9}$$

Let α be the optimal value, and c, w the optimal solution. Then it holds (also for $m = +\infty$) that:

$$\max_{(p_s)\in\mathbb{P}(S)} MSE = \frac{\alpha^*}{n} + O(n^{-2}), \tag{10}$$

and this value is realized when (p_s) is the mixture of the distribution uniform on $\max\{\lfloor wn/c - 1\rfloor, 1\}$ elements and the Dirac mass, with the weights respectively w and $1-w$.

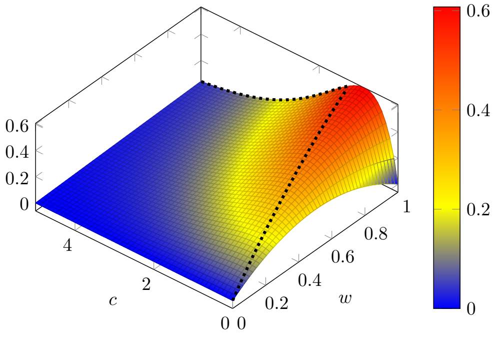

Fig. 1. The optimization landscape of the program in Theorem 3.

The optimal value can be computed explicitly, as shown below; by $W(\cdot)$ we denote the Lambert-W function [39].

Corollary 1. *Under the setup of Theorem 3:*

$$\alpha^{\mathrm{GT}} = \begin{cases} \frac{W(2)^2+2W(2)}{4} = 0.608... & \frac{m}{n} \geqslant \frac{1}{W(2)} \\ \max\limits_{0\leqslant c\leqslant \frac{n}{m}} \frac{mc(1+c)\mathrm{e}^{-c}}{n} - \left(\frac{mc\mathrm{e}^{-c}}{n}\right)^2 & \frac{m}{n} \leqslant \frac{1}{W(2)}. \end{cases} \tag{11}$$

This is illustrated in Figure 2; note the phase transition in the maximizer, not anticipated in prior works! Since $\alpha^* = \Theta(\max\{\frac{m}{n}, 1\})$, the error $O(n^{-2})$ is negligible when $m \gg 1$.

III. Preliminaries

In our estimations we need the following fact [40], [41]

Lemma 1 (Mode of Beta-Distribution)**.** *The expression $x^a(1-x)^b$ for $x \in [0,1]$, when $a, b > 0$, is maximized at $x = \frac{a}{a+b}$.*

Below we study in detail the function which particularly often comes up in our calculations (see also Figure 3).

Lemma 2 (Exponential-Quadratic Function)**.** *Define $g_b(u) \triangleq (u^2+bu)\mathrm{e}^{-u}$ for $u \geqslant 0$, with parameter $b \in \mathbb{R}$. Then:*

- *g has two local extremes $u = \frac{2-b\pm\sqrt{b^2+4}}{2}$ when $b < 0$ and one extreme at $u = \frac{2-b+\sqrt{b^2+4}}{2}$ when $b \geqslant 0$,*

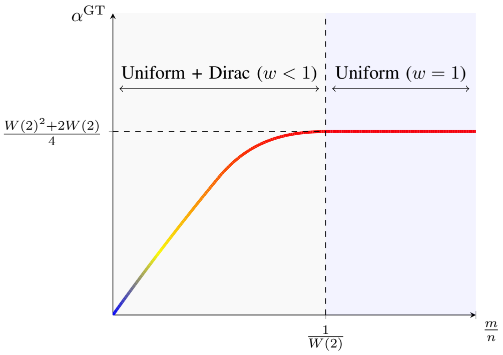

Fig. 2. The dependency of the max MSE (Corollary 1) on $m = \#S$. The phase transition (the maximizer becomes uniform) occurs at $m = \frac{n}{W(2)}$.

- *when $b < 1$, g is concave in $\left(\frac{4-b-\sqrt{b^2+8}}{2}, \frac{4-b+\sqrt{b^2+8}}{2}\right)$ and convex in $\left(0, \frac{4-b-\sqrt{b^2+8}}{2}\right) \cup \left(\frac{4-b+\sqrt{b^2+8}}{2}, +\infty\right)$,*
- *whenanswer $b \geqslant 1$, g is concave in $\left(0, \frac{4-b+\sqrt{b^2+8}}{2}\right)$ and convex in $\left(\frac{4-b+\sqrt{b^2+8}}{2}, +\infty\right)$.*

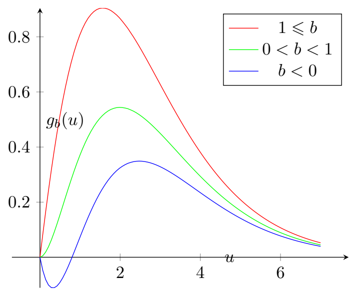

Fig. 3. The auxiliary function $g_b(u) = (u^2+bu)\mathrm{e}^{-u}$, from Lemma 2.

Our proofs rely on non-linear optimization, particularly on the Karush–Kuhn–Tucker (first-order) conditions. For a detailed discussion we refer to optimization books [42]–[44].

Lemma 3 (KKT conditions)**.** *Consider the program*

$$\begin{aligned} \max \quad & f(x) \\ \text{s.t.} \quad & \begin{cases} h_i(x) = 0, & i \in I \\ g_j(x) \leqslant 0, & i \in J \end{cases} \end{aligned} \tag{12}$$

with differentiable real functions $f, (h_i)_{i\in I}, (g_j)_{j\in J}$ in variables $x = x_1, \ldots, x_d$. If the maximum occurs at x, then:

$$\frac{\partial}{\partial x} f(x) = \sum_i \lambda_i \frac{\partial}{\partial x} h_i(x) + \sum_j \mu_j \frac{\partial}{\partial x} g_j(x) \tag{13}$$

where $\frac{\partial}{\partial x} = (\frac{\partial}{\partial x_1} \ldots \frac{\partial}{\partial x_d})$, for $\lambda_i \in \mathbb{R}$, $\mu_j \geqslant 0$ such that

$$\lambda_i \in \mathbb{R}, \; \mu_j \geqslant 0, \; \mu_j g_j(x) = 0, \tag{14}$$

provided that regularity conditions hold at x.

We briefly remind the optimization terminology. The function f is called objective, and any $x \in \mathbb{R}^d$ satisfying the constraints is called feasible. The optimal value is also called the program value. The constraint h_i respectively g_j is called active at x, when $h_i(x) = 0$, respectively $g_j(x) = 0$.

Remark 1 **(LICQ Constraints Qualification).** *The KKT conditions hold for optimal x when the gradients of the constraints active at x are linearly independent.*

IV. Proofs

A. Proof of Theorem 1

Denote $\xi_s = \mathbb{I}(f_s = 1)$. We have

$$\sum_{s\neq s'} \mathbf{E}[\xi_s \xi_{s'}] = n(n-1)\sum_{s\neq s'} p_s p_{s'} (1-p_s-p_{s'})^{n-2}. \tag{15}$$

Our goal is to estimate this expression up to $O(n)$.

We can assume that $p_s \leqslant \frac{1}{3}$. Indeed, since (p_s) is a probability distribution, there are at most two values of s such that $p_s > \frac{1}{3}$. The total contribution from all such s to the right-hand side of the equation is at most $O(n^2 2^{-n}) \leqslant O(1)$.

We use the following bound, valid for $x \in [0,1]$ and $n \geqslant 1$:

$$(1-x)^n = 1 - O(nx), \tag{16}$$

to $x = 1 - \frac{1-p_s-p_{s'}}{(1-p_s)(1-p_{s'})} = \frac{p_s p_{s'}}{(1-p_s)(1-p_{s'})}$, and obtain:

$$\frac{(1-p_s-p_{s'})^{n-2}}{(1-p_s)^{n-2}(1-p_{s'})^{n-2}} = 1 - O(np_s p_{s'}). \tag{17}$$

This implies:

$$\sum_{s\neq s'} \mathbf{E}[\xi_s \xi_{s'}] = n(n-1)\sum_{s\neq s'} p_s p_{s'} (1-p_s)^{n-2}(1-p_{s'})^{n-2} + O(n), \tag{18}$$

where we used $O(n\cdot n(n-1))\sum_{s\neq s'} p_s^2 p_{s'}^2 (1-p_s)^{n-2}(1-p_{s'})^{n-2} = O(n^3)(\sum_s p_s^2(1-p_s)^{n-2})^2 = O(n)$; the last step follows by Lemma 1 with $a=1, b=n-2$, and $\sum_s p_s = 1$.

Furthermore, we have $\sum_{s=s'} p_s p_{s'}(1-p_s)^{n-2}(1-p_{s'})^{n-2} = O(n^{-1})$ by Lemma 1 applied to $a = 1, b = 2\cdot(n-2)$ and the condition $\sum_s p_s = 1$; thus, we obtain:

$$\sum_{s\neq s'} p_s p_{s'} (1-p_s)^{n-2}(1-p_{s'})^{n-2} = \\ = \left(\sum_s p_s (1-p_s)^{n-2}\right)^2 + O(n^{-1}). \tag{19}$$

Using the last bound we conclude that:

$$\sum_{s\neq s'} \mathbf{E}[\xi_s \xi_{s'}] = n^2 \left(\sum_s p_s(1-p_s)^{n-2}\right)^2 + O(n). \tag{20}$$

In terms of the occupancy numbers N_k, we have shown that:

$$\mathbf{E}[N_1^2 - N_1] = n^2\left(\sum_s p_s (1-p_s)^{n-2}\right)^2 + O(n), \tag{21}$$

because $N_1 = \sum_s \xi_s$ and $\sum_{s\neq s'} \xi_s \xi_{s'} = N_1^2 - N_1$. Finally, $p_s(1-p_s)^{n-2} = p_s(1-p_s)^{n-1}\cdot(1+O(p_s))$ because $\frac{1}{1-p_s} = 1 + O(p_s)$ for $p_s \leqslant \frac{1}{3}$; since $\sum_s p_s^2(1-p_s)^{n-1} = O(n^{-1})$, by Lemma 1 applied to $a = 1, b = n-1$ and $\sum_s p_s = 1$:

$$\mathbf{E}[N_1^2 - N_1] = n^2\left(\sum_s p_s(1-p_s)^{n-1}\right)^2 + O(n), \tag{22}$$

and, since $n\sum_s p_s(1-p_s)^{n-1} = \mathbf{E}[N_1]$, we finally obtain:

$$\mathbf{E}[N_1^2 - N_1] = \mathbf{E}[N_1]^2 + O(n). \tag{23}$$

Combining this with Equation (6) and the fact that $\mathbf{E}N_1 = n\sum_s p_s(1-p_s)^{n-1}) = O(n)$ finishes the proof.

B. Proof of Theorem 2

Since we have $\mathbf{E}[N_1] = n\sum_s p_s(1-p_s)^{n-1}$ and $\mathbf{E}[N_2] = \binom{n}{2}\sum_s p_s^2(1-p_s)^{n-2}$, and $\sum_s p_s^2 (1-p_s)^{n-2} = O(n^{-1})$ by Lemma 1 applied to $a = 1, b = n-2$, it suffices to show that:

$$\begin{aligned} \sum_s p_s(1-p_s)^{n-1} &= \sum_s p_s \mathrm{e}^{-np_s} + O(n^{-1}) \\ \sum_s p_s^2(1-p_s)^{n-2} &= \sum_s p_s^2 \mathrm{e}^{-np_s} + O(n^{-2}). \end{aligned} \tag{24}$$

To prove this, we first notice that we can assume $p_s \leqslant \frac{1}{3}$ (the same argument as in the proof of Theorem 1).

Note that $(1-p)^n = \mathrm{e}^{-np}(1-x)^n$, $x = 1 - \frac{1-p}{\mathrm{e}^{-p}}$; $\frac{1}{2} \leqslant \frac{x}{p^2} \leqslant 1$ implies $(1-x)^n = 1 - O(nx) = 1 - O(np^2)$, thus:

$$(1-p)^n = \mathrm{e}^{-np}(1 - O(np^2)). \tag{25}$$

Moreover, we have the series of bounds:

$$\sum_s p_s^k \mathrm{e}^{-np_s} = O(n^{1-k}), \quad k = 2,3,4 \tag{26}$$

obtained by introducing $q_s = 1 - \mathrm{e}^{-p_s}$ so that $p_s^3 \mathrm{e}^{-np_s} = O(1) q_s^3 (1-q_s)^n$, using Lemma 1 with $a = 1, b = n$, $a = 2, b = n$ or $a = 3, b = n$, and $\sum_s q_s = O(\sum_s p_s) = O(1)$.

These bounds finally give us:

$$\begin{aligned} \sum_s p_s(1-p_s)^{n-1} &= \sum_s p_s \mathrm{e}^{-(n-1)p_s} + O(n^{-1}) \\ \sum_s p_s^2(1-p_s)^{n-2} &= \sum_s p_s^2 \mathrm{e}^{-(n-2)p_s} + O(n^{-2}). \end{aligned} \tag{27}$$

It remains to notice that $e^{-(n-1)p} = \mathrm{e}^{-np}\cdot \mathrm{e}^{p} = \mathrm{e}^{-np}(1 + O(p))$, and similarly $e^{-(n-2)p} = \mathrm{e}^{-np}\mathrm{e}^{2p} = \mathrm{e}^{-np}(1+O(p))$. This means that replacing $n-1$ and $n-2$ above by n we make the error of respectively $O(\sum_s p_s^2 \mathrm{e}^{-np_s}) = O(n^{-1})$ and $O(\sum_s p_s^3 \mathrm{e}^{-np_s}) = O(n^{-2})$. This completes the proof.

C. Proof of Theorem 3

1) Non-Linear Programming : In view of Theorem 2:

$$\max_{(p_s)} MSE = \frac{\alpha}{n} + O(n^{-2}), \tag{28}$$

where α is the value of the following optimization program:

$$\begin{aligned} \max \quad & n\sum_s p_s^2 \mathrm{e}^{-np_s} + \sum_s p_s \mathrm{e}^{-np_s} - (\sum_s p_s \mathrm{e}^{-np_s})^2 \\ \text{s.t.} \quad & \forall s \in S: p_s \geqslant 0, \text{ and } \sum_s p_s = 1, \end{aligned} \tag{29}$$

and the optimal point gives the probability distribution realizing the maximum. We assume that $\#S = m < +\infty$, then the optimal solution p^* exists (by the extreme value theorem [45]); we discuss $m = +\infty$ at the end of the proof.

2) First-Order Conditions: The KKT condition gives:

$$\forall s \in S: \quad a \cdot \frac{\partial}{\partial p_s}[p_s^2 \mathrm{e}^{-np_s}] + b \cdot \frac{\partial}{\partial p_s}[p_s \mathrm{e}^{-np_s}] = \lambda, \tag{30}$$

with coefficients a, b that do not change with $s \in S$:

$$a \triangleq n, \quad b \triangleq 1 - 2\left(\sum_{s \in S} p_s \mathrm{e}^{-np_s}\right). \tag{31}$$

We conclude that for the optimal solution the components of $(p_s)_{s \in S}$ take values in the set of solutions v to the equation:

$$\frac{\partial}{\partial v}\left[(av^2 + bu)\mathrm{e}^{-nv}\right] = \lambda. \tag{32}$$

3) Optimum is 3-Mixture: We now argue that the equation has at most 3 positive solutions in u. To this end, let us introduce $u = vn$ and use $a = n$ to simplify the equation:

$$\frac{\partial}{\partial u}\left[(u^2 + bv)\mathrm{e}^{-u}\right] = \lambda. \tag{33}$$

By Lemma 2, the left-hand side changes its monotonicity at most twice. Thus, the equation has at most three solutions, and the optimal p_S takes at most 3 distinct non-zero values.

4) 6-D Program for 3-Mixture: By the previous step:

$$\{p_s^* : p_s^* > 0\} = \left\{\frac{c_1^*}{n}, \frac{c_2^*}{n}, \frac{c_3^*}{n}\right\}, \tag{34}$$

with c_i^* not necessarily distinct. Furthermore, let

$$m_i^* = \#\left\{s : p_s^* = \frac{c_i^*}{n}\right\}. \tag{35}$$

Then our original program (29) is equivalent to:

$$\begin{aligned} &\max \ \frac{\sum_{i=1}^3 (m_i c_i^2 + m_i c_i)\mathrm{e}^{-c_i}}{n} - \left(\frac{\sum_{i=1}^3 m_i c_i \mathrm{e}^{-c_i}}{n}\right)^2 \\ &\text{s.t.} \begin{cases} 0 \leqslant c_i \text{ and } 0 \leqslant m_i \text{ and } m_i \in \mathbb{Z}, \quad i = 1,2,3 \\ \sum_{i=1}^3 c_i m_i = n \\ \sum_{i=1}^3 m_i \leqslant m, \end{cases} \end{aligned} \tag{36}$$

with the optimal solution $(c_i^*), (m_i^*)$.

5) Step 6: 2-D Program for Continuous Relaxation: In the previous program m_i are integers; we consider the relaxation

$$\begin{aligned} &\max \ \frac{\sum_{i \in I} (m_i c_i^2 + m_i c_i)\mathrm{e}^{-c_i}}{n} - \left(\frac{\sum_{i \in I} m_i c_i \mathrm{e}^{-c_i}}{n}\right)^2 \\ &\text{s.t.} \begin{cases} 0 \leqslant c_i \text{ and } 0 \leqslant m_i, \quad i \in I \\ \sum_{i \in I} c_i m_i \leqslant n \\ \sum_{i \in I} m_i \leqslant m, \end{cases} \end{aligned} \tag{37}$$

where $I = \{1, 2, 3\}$. We first prove that the maximum is achieved (not obvious, as c_i are not bounded). Indeed, if the maximum is achieved as a limit with $c_i \to +\infty$, then the contributions to the objective $m_i c_i^2 \mathrm{e}^{-c_i}$ and $m_i c_i \mathrm{e}^{-c_i}$ tend to zero regardless of the values of m_i, because m_i is bounded; in the limit we obtain the same value as when setting $c_i = 0$ and an arbitrary fixed value for m_i (this preserves the constraints).

We next argue that the maximum occurs at a point such that $m_i c_i = 0$ for at least two indices i; this means that we can assume $|I| = 1$ and simplify the program to two variables. Suppose that $(c_i), (m_i)$ is optimal and such that the number $\#\{i : m_i c_i \neq 0\}$ is smallest possible. If $\#\{i : m_i c_i \neq 0\} \leqslant 1$ there is nothing to prove, thus we assume $\#\{i : m_i c_i \neq 0\} \geqslant 2$. We can assume $m_i c_i \neq 0$ for $i = 1, 2$ (due to the symmetry). The LICQ holds, as the gradients of *possibly* active constraints

$$\begin{aligned} \frac{\partial}{\partial[(c_i),(m_i)]}[c_3] &= (0,0,1,0,0,0) \\ \frac{\partial}{\partial[(c_i),(m_i)]}[m_3] &= (0,0,0,0,0,1) \\ \frac{\partial}{\partial[(c_i),(m_i)]}\left[\sum_{i=1}^3 c_i m_i\right] &= (m_1, m_2, m_3, c_1, c_2, c_3) \\ \frac{\partial}{\partial[(c_i),(m_i)]}\left[\sum_{i=1}^3 m_i\right] &= (0,0,0,1,1,1), \end{aligned} \tag{38}$$

are linearly independent (here we use $m_1, m_2 \neq 0$). The KKT condition shows that $u = c_1, c_2$ satisfy the system:

$$\begin{aligned} (u^2 + bu)\mathrm{e}^{-u} &= n\lambda u + n\mu \\ \frac{\partial}{\partial u}\left[(u^2 + bu)\mathrm{e}^{-u}\right] &= n\lambda, \end{aligned} \tag{39}$$

where $b = 1 - \frac{2}{n}\sum_{i=1}^3 m_i c_i \mathrm{e}^{-c_i}$; the first equation is the condition for m_i multiplied by n, and the second equation is the condition for c_i multiplied by n and divided by m_i (here we use again $m_1, m_2 \neq 0$). Equivalently $g(u) \triangleq (u^2 + bu)\mathrm{e}^{-u}$ is tangent to the straight line $n\lambda u + n\mu$ at points $u = c_1, c_2$. We claim this is not possible, unless $c_1 = c_2$ (see Figure 3). Otherwise $c_1 \neq c_2$ must be on the different sides of the stationary point (by the mean-value theorem [46], [47]); to match the slopes they need to be in the two intervals where g decreases (Lemma 2), so necessarily $b < 0$; the intercept is below zero for the first interval (with the start-point at 0), and above zero for the second interval (with the end-point at $+\infty$) In turn, $c_1 = c_2$ reduces to $c_1 = 0$; to see that we define $c_1', c_2', c_3' = 0, c_1 + c_2, c_3$ and $m_1', m_2', m_3' = 0, m_1 + m_2, m_3$, then replace c_i, m_i by c_i', m_i', preserving the objective value and constraints. But $\#\{i : m_i c_i' \neq 0\} = \#\{i : m_i c_i \neq 0\} - 1$. This proves our claim that $m_i c_i = 0$ for two indices i.

Our relaxed program is equivalent to:

$$\begin{aligned} &\max \quad \frac{(m_1 c_1^2 + m_1 c_1)\mathrm{e}^{-c_1}}{n} - \left(\frac{m_1 c_1 \mathrm{e}^{-c_1}}{n}\right)^2 \\ &\text{s.t.} \quad \begin{cases} 0 \leqslant c_1 \text{ and } 0 \leqslant m_1 \\ c_1 m_1 \leqslant n \\ m_1 \leqslant m. \end{cases} \end{aligned} \tag{40}$$

6) Relaxation Gap is Small: We argue that the last step (relaxation) changes the optimal value by at most $O(n^{-1})$.

To this end, suppose that c_1, m_1 is optimal to (40) and let $P = \frac{(m_1 c_1^2 + m_1 c_1)\mathrm{e}^{-c_1}}{n} - \left(\frac{m_1 c_1 \mathrm{e}^{-c_1}}{n}\right)^2$; it suffices to construct $(c_i'), (m_i')$ feasible for the program (36) and such that $P' = \frac{\sum_{i=1}^3 (m_i' {c_i'}^2 + m_i' c_i')\mathrm{e}^{-c_i'}}{n} - \left(\frac{\sum_{i=1}^3 m_i' c_i' \mathrm{e}^{-c_i'}}{n}\right)^2$ satisfies

$$P - P' \leqslant O(n^{-1}), \tag{41}$$

because the optimal value of (36) is upper-bounded by P (by relaxation) and lower-bounded by P' (by feasibility).

The optimal value is clearly positive and thus $m_1, c_1 > 0$. Define $m_1' = \max\{\lfloor m_1 - 1 \rfloor, 1\}$, $m_2' = 1, m_3' = 0$, also $c_1' = m_1 c_1 / m_1'$, $c_2' = n - m_1 c_1, c_3' = 0$. Note that $m_i', c_i' \geqslant 0$ and m_i' are integers; moreover $\sum_i m_i' \leqslant \max\{m_1, 2\} \leqslant m$ and $\sum_i m_i' c_i' = n$. Thus, $(c_i'), (m_i')$ is feasible for (36).

The bound on $P - P'$ trivially follows when $m_1 \leqslant 2$ because for $m_i = O(1)$ and $m_i' = O(1)$ we have $P = O(n^{-1})$, $|P'| = O(n^{-1})$. We further assume that $m_1 \geqslant 2$, which implies $m_1' = \Theta(m_1)$ and $c_1' = \Theta(c_1)$. To bound $P - P'$ we observe that:

$$\begin{aligned} P - P' = \frac{m_1 c_1}{n}&((c_1+1)\mathrm{e}^{-c_1} - (c_1'+1)\mathrm{e}^{-c_1'}) \\ &- \frac{c_2'(c_2'+1)\mathrm{e}^{-c_2'}}{n} \\ - \left(\frac{m_1 c_1 \mathrm{e}^{-c_1}}{n}\right)^2 &+ \left(\frac{m_1 c_1 \mathrm{e}^{-c_1'} + c_2' \mathrm{e}^{-c_2'}}{n}\right)^2, \end{aligned} \tag{42}$$

(we used $m_1 c_1 = m_1' c_1'$ and $m_2' = 1, m_3' = 0$) so it remains to show that these three terms are at most $O(n^{-1})$.

The first term is $\frac{m_1 c_1}{n} \cdot (h(c_1) - h(c_1'))$ with $h(u) \triangleq (1+u)\mathrm{e}^{-u}$. By the mean-value theorem $\frac{m_1 c_1}{n} \cdot |h(c_1) - h(c_1')| = \frac{m_1 c_1}{n} \cdot O(c_1 \mathrm{e}^{-\Theta(c_1)}) \cdot |c_1' - c_1|$; since $|c_1' - c_1| = O(c_1/m_1') = O(c_1/m_1)$, we can upper-bound as $\frac{O(c_1^2 \mathrm{e}^{-\Theta(c_1)})}{n} = O(n^{-1})$.

The second term is negative can be ignored.

The third term equals $x^2 - y^2 = (x-y)(x+y)$ with $x = \frac{m_1 c_1 \mathrm{e}^{-c_1'} + c_2' \mathrm{e}^{-c_2'}}{n}$ and $y = \frac{m_1 c_1 \mathrm{e}^{-c_1}}{n}$. We have $0 \leqslant x + y \leqslant O(1)$ by the constraint $m_1' c_1' \leqslant n$. In turn, $x - y = \frac{m_1 c_1 (\mathrm{e}^{-c_1} - \mathrm{e}^{-c_1'})}{n} + O(n^{-1})$. By the mean-value theorem applied to $h(u) = \mathrm{e}^{-u}$ we get $x - y = \frac{m_1 c_1}{n} \cdot |c_1' - c_1| \cdot O(1)\mathrm{e}^{-\Theta(c_1)}$; since $|c_1' - c_1| = O(c_1/m_1)$ we get $|x - y| = \frac{O(c_1 \mathrm{e}^{-\Theta(c_1)})}{n} = O(n^{-1})$ and $x^2 - y^2 \leqslant O(n^{-1})$.

7) Summing Up: With $w = \frac{c_1 m_1}{n}, c_1 = c$ we write (40) as:

$$\begin{aligned} \max \quad & w(1+c)\mathrm{e}^{-c} - (w\mathrm{e}^{-c})^2 \\ \text{s.t.} \quad & \begin{cases} 0 \leqslant w \leqslant 1 \\ w \leqslant \frac{m}{n} c. \end{cases} \end{aligned} \tag{43}$$

We proved that the gap w.r.t (29) is $O(n^{-1})$, and thus for the optimal value α the worst-case MSE is $\alpha/n + O(n^{-2})$. The probability distribution achieving this value can be constructed from optimal w, c as explained in the previous step; namely on $m_1' = \max\{\lfloor m_1 - 1 \rfloor, 1\}$ elements the probability mass equals $\frac{c_1}{n} = \frac{w}{m_1}$, and on other $m_2' = 1$ elements the probability mass is $\frac{n - m_1 c}{n} = 1 - w$. This completes the proof when $m < +\infty$.

8) Unbounded Dimension: The supremum α of (29) can be approached on a sequence of distributions $(p_s^{(k)})$, $k = 1, 2, \ldots$.

For any fixed $\epsilon > 0$, let (p_s) be any distribution such that the the objective value is at least $\alpha - \epsilon$. Since the objective of (29) is continuous with respect to the total variation distance, by a mass-shifting argument we can find (p_s') with finite support m' such that the objective is at least $\alpha - 2\epsilon$. Replace (p_s') with (p_s'') which is optimal under the support constraint m'; the objective can only increase, thus it is still at least $\alpha - 2\epsilon$. When $m > \frac{n}{W(2)}$ the objective in Theorem 3 has its maximum at $w = 1$ and $c = W(2)$, as we will see in Corollary 1, and then the constraint $w \leqslant \frac{m}{n}$ can be ignored. This shows that limiting the support to $m = O(n)$ we can approximate the supremum up to 2ϵ, for arbitrarily small ϵ. Thus, for $m = +\infty$ the theorem also holds (the constraint with $m = +\infty$ is automatically satisfied, hence to be ignored).

D. *Proof of Corollary 1*

We denote $b \triangleq \frac{m}{n}$ and consider the program:

$$\begin{aligned} \max \quad & w(1+c)\mathrm{e}^{-c} - (w\mathrm{e}^{-c})^2 \\ \text{s.t.} \quad & 0 \leqslant w \leqslant 1 \quad \text{and } w \leqslant bc \end{aligned} \tag{44}$$

with respect to the parameter $b > 0$.

The program has no local maximum with $0 < w < 1$ and $w < bc$ (also seen in Figure 1). Indeed, otherwise the first order condition for w would give $w = \frac{(1+c)\mathrm{e}^c}{2}$; the program becomes maximizing $\frac{1+c}{2} - \left(\frac{1+c}{2}\right)^2$ subject to $\frac{(1+c)\mathrm{e}^c}{2} < bc$, and the first order condition for c gives $c = 0$, a contradiction.

Since at the optimal point $w \neq 0$ (the objective would be zero), we are left with two cases. For $w = 1$ the program is

$$\max_{\frac{1}{b} \leqslant c} (1+c)\mathrm{e}^{-c} - \mathrm{e}^{-2c}, \tag{45}$$

and when $w = bc$ the program becomes

$$\max_{c \leqslant \frac{1}{b}} b(c + c^2)\mathrm{e}^{-c} - b^2 c^2 \mathrm{e}^{-2c}. \tag{46}$$

In what follows we use the fact that $h(c) \triangleq (1+c)\mathrm{e}^{-c} - \mathrm{e}^{-2c}$ increases for $0 < c < W(2)$ and decreases when $W(2) < c$.

Suppose that $b \geqslant \frac{1}{W(2)}$, then (45) is maximized at $c = W(2)$; for $bc \leqslant 1$ we have $b(c+c^2)\mathrm{e}^{-c} - b^2c^2\mathrm{e}^{-2c} \leqslant (bc)^2(c+1)\mathrm{e}^{-c} - (bc)^2\mathrm{e}^{-2c} \leqslant (c+1)\mathrm{e}^{-c} - \mathrm{e}^{-2c}$, so we conclude that the optimal value of (46) is smaller or equal than that of (45).

Suppose now that $b \leqslant \frac{1}{W(2)}$; then (45) is maximized at $c = \frac{1}{b}$; this value matches the objective of (46) at $c = \frac{1}{b}$, thus the optimal value of (46) is bigger or equal than that of (45).

V. Conclusion

This work determines the worst mean-squared error of the Good-Turing estimator given the sample and alphabet size, completing upon prior results for unrestricted distributions.

References

[1] H. E. Robbins *et al.*, "Estimating the total probability of the unobserved outcomes of an experiment," *The Annals of Mathematical Statistics*, vol. 39, no. 1, pp. 256–257, 1968.

[2] T.-J. Shen, A. Chao, and C.-F. Lin, "Predicting the number of new species in further taxonomic sampling," *Ecology*, vol. 84, no. 3, pp. 798–804, 2003.

[3] A. Chao and T.-J. Shen, "Nonparametric estimation of shannon's index of diversity when there are unseen species in sample," *Environmental and ecological statistics*, vol. 10, no. 4, pp. 429–443, 2003.

[4] A. Chao, Y. Wang, and L. Jost, "Entropy and the species accumulation curve: a novel entropy estimator via discovery rates of new species," *Methods in Ecology and Evolution*, vol. 4, no. 11, pp. 1091–1100, 2013.

[5] A. Chao, R. K. Colwell, C.-H. Chiu, and D. Townsend, "Seen once or more than once: Applying good–turing theory to estimate species richness using only unique observations and a species list," *Methods in Ecology and Evolution*, vol. 8, no. 10, pp. 1221–1232, 2017.

[6] B. Efron and R. Thisted, "Estimating the number of unseen species: How many words did shakespeare know?" *Biometrika*, vol. 63, no. 3, pp. 435–447, 1976.

[7] D. R. McNeil, "Estimating an author's vocabulary," *Journal of the American Statistical Association*, vol. 68, no. 341, pp. 92–96, 1973.

[8] W. A. Gale and G. Sampson, "Good-turing frequency estimation without tears," *Journal of quantitative linguistics*, vol. 2, no. 3, pp. 217–237, 1995.

[9] N. Myrberg Burström, "A tale of buried treasure, some good estimations, and golden unicorns: The numismatic connections of alan turing." 2015.

[10] C. Budianu and L. Tong, "Estimation of the number of operating sensors in sensor network," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2. IEEE, 2003, pp. 1728–1732.

[11] ——, "Good-turing estimation of the number of operating sensors: a large deviations analysis," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2. IEEE, 2004, pp. ii–1029.

[12] V. Q. Vu, B. Yu, and R. E. Kass, "Coverage-adjusted entropy estimation," *Statistics in medicine*, vol. 26, no. 21, pp. 4039–4060, 2007.

[13] Z. Zhang, "Entropy estimation in turing's perspective," *Neural computation*, vol. 24, no. 5, pp. 1368–1389, 2012.

[14] C. X. Mao and B. G. Lindsay, "A poisson model for the coverage problem with a genomic application," *Biometrika*, vol. 89, no. 3, pp. 669–682, 2002.

[15] P. I. Koukos and N. M. Glykos, "On the application of good-turing statistics to quantify convergence of biomolecular simulations," *Journal of chemical information and modeling*, vol. 54, no. 1, pp. 209–217, 2014.

[16] I. J. Good, "The population frequencies of species and the estimation of population parameters," *Biometrika*, vol. 40, no. 3-4, pp. 237–264, 1953.

[17] N. Rajaraman, A. Thangaraj, and A. T. Suresh, "Minimax risk for missing mass estimation," in *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2017, pp. 3025–3029.

[18] J. Acharya, Y. Bao, Y. Kang, and Z. Sun, "Improved bounds for minimax risk of estimating missing mass," in *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2018, pp. 326–330.

[19] M. I. Ohannessian and M. A. Dahleh, "Rare probability estimation under regularly varying heavy tails," in *Conference on Learning Theory*. JMLR Workshop and Conference Proceedings, 2012, pp. 21–1.

[20] E. Mossel and M. I. Ohannessian, "On the impossibility of learning the missing mass," *Entropy*, vol. 21, no. 1, p. 28, 2019.

[21] W. W. Esty, "Confidence intervals for the coverage of low coverage samples," *The Annals of Statistics*, pp. 190–196, 1982.

[22] W. W. Esty *et al.*, "A normal limit law for a nonparametric estimator of the coverage of a random sample," *The Annals of Statistics*, vol. 11, no. 3, pp. 905–912, 1983.

[23] D. A. McAllester and R. E. Schapire, "On the convergence rate of good-turing estimators." in *COLT*, 2000, pp. 1–6.

[24] D. McAllester and L. Ortiz, "Concentration inequalities for the missing mass and for histogram rule error," *Journal of Machine Learning Research*, vol. 4, no. Oct, pp. 895–911, 2003.

[25] D. Berend, A. Kontorovich *et al.*, "On the concentration of the missing mass," *Electronic Communications in Probability*, vol. 18, 2013.

[26] A. Ben-Hamou, S. Boucheron, M. I. Ohannessian *et al.*, "Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications," *Bernoulli*, vol. 23, no. 1, pp. 249–287, 2017.

[27] D. Berend and A. Kontorovich, "The missing mass problem," *Statistics & Probability Letters*, vol. 82, no. 6, pp. 1102–1110, 2012.

[28] D. Berend, A. Kontorovich, and G. Zagdanski, "The expected missing mass under an entropy constraint," *Entropy*, vol. 19, no. 7, p. 315, 2017.

[29] A. Orlitsky, N. P. Santhanam, and J. Zhang, "Always good turing: Asymptotically optimal probability estimation," *Science*, vol. 302, no. 5644, pp. 427–431, 2003.

[30] A. Orlitsky and A. T. Suresh, "Competitive distribution estimation: Why is good-turing good." in *NIPS*, 2015, pp. 2143–2151.

[31] M. Falahatgar, M. Ohannessian, A. Orlitsky, and V. Pichapati, "The power of absolute discounting: all-dimensional distribution estimation," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 6663–6672.

[32] F. Ayed, M. Battiston, F. Camerlenghi, S. Favaro *et al.*, "A good-turing estimator for feature allocation models," *Electronic Journal of Statistics*, vol. 13, no. 2, pp. 3775–3804, 2019.

[33] S. Cohen, T. Routtenberg, and L. Tong, "Non-bayesian parametric missing-mass estimation," *arXiv preprint arXiv:2101.04329*, 2021.

[34] R. Cohen, L. Katzir, and A. Yehezkel, "Cardinality estimation meets good-turing," *Big data research*, vol. 9, pp. 1–8, 2017.

[35] M. Skorski, "Good-turing-mse, github," https://github.com/maciejskorski/Good-Turing-MSE, 2021.

[36] A. Chao and S.-M. Lee, "Estimating the number of classes via sample coverage," *Journal of the American statistical Association*, vol. 87, no. 417, pp. 210–217, 1992.

[37] A. Gnedin, B. Hansen, J. Pitman *et al.*, "Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws," *Probability surveys*, vol. 4, pp. 146–171, 2007.

[38] C.-H. Zhang, Z. Zhang *et al.*, "Asymptotic normality of a nonparametric estimator of sample coverage," *The Annals of Statistics*, vol. 37, no. 5A, pp. 2582–2595, 2009.

[39] R. M. Corless, G. H. Gonnet, D. E. Hare, D. J. Jeffrey, and D. E. Knuth, "On the lambertw function," *Advances in Computational mathematics*, vol. 5, no. 1, pp. 329–359, 1996.

[40] N. Johnson, S. Kotz, and N. Balakrishnan, "Beta distributions," *Continuous univariate distributions. 2nd ed. New York, NY: John Wiley and Sons*, pp. 221–235, 1994.

[41] A. Gupta and S. Nadarajah, *Handbook of Beta Distribution and Its Applications*, ser. Statistics: A Series of Textbooks and Monographs. Taylor & Francis, 2004. [Online]. Available: https://books.google.at/books?id=cVmnsxa-VzwC

[42] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[43] L. Biegler, *Nonlinear Programming: Concepts, Algorithms, and Applications to Chemical Processes*, ser. MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics, 2010. [Online]. Available: https://books.google.at/books?id=ZmIC7w9QnPEC

[44] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear programming: theory and algorithms*. John Wiley & Sons, 2013.

[45] M. Lovric, *Vector Calculus*. Wiley, 2007. [Online]. Available: https://books.google.at/books?id=hDdyDwAAQBAJ

[46] J. Matkowski, "A mean-value theorem and its applications," *Journal of mathematical analysis and applications*, vol. 373, no. 1, pp. 227–234, 2011.

[47] C. B. Boyer, *The history of the calculus and its conceptual development:(The concepts of the calculus)*. Courier Corporation, 1959.