Mean-Squared Accuracy of Good-Turing Estimator

Maciej Skorski
University of Luxembourg

Abstract—The brilliant method due to Good and Turing allows for estimating objects not occurring 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 many applications, including language modelling, inference in ecology and estimation of 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 $X_1, \ldots, X_n \sim \text{i.i.d.}$ 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

$$M_0 = \sum_s p_s I(f_s = 0),$$

(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], archaeology [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:

$$\hat{M} = \frac{1}{n} \sum_s I(f_s = 1).$$

(2)

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

$$MSE = \mathbb{E} \left[ (\hat{M} - M_0)^2 \right],$$

(3)

under the constrained alphabet (upper-bounded support)

$$\#S = m.$$

(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, define

$$N_k = \sum_s I(f_s = k),$$

(5)

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

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

(6)

and moreover $\max_{(p_s) \in \mathcal{P}(S)} MSE = \Theta(n^{-1})$ when $\#S = +\infty$, where $\mathcal{P}(S)$ 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.

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]–[33], and others [34]–[36].

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 [37].

II. RESULTS

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

We give the formula for MSE 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:

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

(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 [38]–[40]. We develop such an approximation, and use later for constrained optimization.

\footnote{The proof is omitted in both proceedings and public version.}
Theorem 2. For any distribution \((p_s)\) we have:

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

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 \(n \geq 2\) and \(m = \#S \geq 2\), consider

\[
\max_{\alpha} \alpha(w, c) = w(1 + c)e^{-c} - (we^{-c})^2.
\]

s.t.

\[
0 \leq w \leq 1, \quad w \leq \frac{m}{n} c.
\]

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

\[
\max_{(p_s) \in P(S)} \quad MSE = \frac{\alpha}{n} + O(n^{-2}), \quad \tag{10}
\]

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

Below we show how to numerically compute the optimal value; \(W(\cdot)\) denotes the Lambert-W function [41].

Corollary 1. Under the setup of Theorem 3:

\[
\alpha = \left\{ \begin{array}{ll}
\frac{W(2)^2 + 2W(2)}{4} = 0.608... & m \geq \frac{1}{W(2)} \\
\max_{0 \leq c \leq \frac{m}{n}} \frac{mc(1+c)e^{-c}}{n} & m \leq \frac{1}{W(2)}
\end{array} \right. \quad \tag{11}
\]

For illustration, see Figure 2; note the phase transition in the maximizer, not discussed in prior works. The specific "Uniform-Dirac" shape is also seen in other minim-max and optimization problems [25], [42]–[44]. 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 [45], [46]

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 \(x = \frac{a}{a+b}\).
Our goal is to estimate this expression up to $O$ the program value. The constraint $x$ probability distribution, there are at most two values of $s$.

We can assume that $p_0 = \frac{1}{2}$. Indeed, since $(p_s)$ is a probability distribution, there are at most two values of $s$ such that $p_s > \frac{1}{2}$. The total contribution from all such $s$ to the right-hand side of the equation is at most $O(n^{2-g-n}) \leq O(1)$.

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

$$ (1 - x)^n = 1 - O(nx) $$

(16)

to $x = 1 - \frac{1 - p_0 - p_{s'}}{(1 - p_0)(1 - p_{s'})} = \frac{p_0 p_{s'}}{(1 - p_0)(1 - p_{s'})}$, and obtain:

$$ \left(\frac{1 - p_0 - p_{s'}}{(1 - p_0)(1 - p_{s'})}\right)^{n-2} = 1 - O(np_0 p_{s'}).$$

(17)

This implies:

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

+ $O(n)$, (18)

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

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

$$ \sum_{s \neq s'} p_s p_{s'} (1 - p_s) (1 - p_{s'})^{n-2} =$$

$$ = \left(\sum_{s} p_s (1 - p_s)^{n-2}\right)^2 + O(n^{-1}).$$

(19)

Using the last bound we conclude that:

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

(20)

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

$$ E[N_1^2 - N_1] = n^2 \left(\sum_{s} p_s (1 - p_s)^{n-2}\right)^2 + O(n),$$

(21)

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

$$ \sum_{s \neq s'} p^2_s (1 - p_s)^{n-1} = O(n^{-2})$$

(22)

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

$$ E[N_1^2 - N_1] = E[N_1]^2 + O(n).$$

(23)

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

B. Proof of Theorem 2

The result can be derived from Theorem 1 by relating Poisson and binomial distributions as in [50]; we give a direct argument. Since $E[N_1] = n \sum_s p_s (1 - p_s)^{n-1}$, $E[N_2] = \sum_s p^2_s (1 - p_s)^{n-2}$, and $E[(n - 1) p_s] = O(n^{-1})$ by Lemma 1 with $a = 1, b = n - 2$, it suffices to show that:

$$ \sum_{s} p_s (1 - p_s)^{n-1} = \sum_{s} p_s e^{-np_s} + O(n^{-1})$$

$$ \sum_{s} p^2_s (1 - p_s)^{n-2} = \sum_{s} p^2_s e^{-np_s} + O(n^{-2}).$$

(24)

We can assume $p_s \leq \frac{1}{2}$ (justified as in the proof of Theorem 1).

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

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

(25)

Moreover, we have the series of bounds:

$$ \sum_{s} p_s^k e^{-np_s} = O(n^{-k}), \quad k = 2, 3, 4 $$

(26)

obtained by introducing $q_s = 1 - e^{-p_s}$ so that $p_s^k e^{-np_s} = O(1) q_s^k (1 - q_s)^n$, 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:

$$ \sum_{s} p_s (1 - p_s)^{n-1} = \sum_{s} p_s e^{-(n-1)p_s} + O(n^{-1})$$

$$ \sum_{s} p^2_s (1 - p_s)^{n-2} = \sum_{s} p^2_s e^{-(n-2)p_s} + O(n^{-2}).$$

(27)

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

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

\[
\max_{(p_s)} MISE = \frac{\alpha}{n} + O(n^{-2}),
\]

where \(\alpha\) is the value of the following optimization program:

\[
\max \quad n \sum_s p_s^2 e^{-np_s} + \sum_s p_s e^{-np_s} - \left( \sum_s p_s e^{-np_s} \right)^2
\]

\[\text{s.t.} \quad \forall s \in S : p_s \geq 0, \text{ and } \sum_s p_s = 1, \quad (29)\]

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 [51]); we discuss \(m = +\infty\) at the end of the proof.

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

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

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

\[
a \equiv n, \quad b \equiv 1 - 2 \left( \sum_{s \in S} p_s e^{-np_s} \right). \quad (31)
\]

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:

\[
\frac{\partial}{\partial v} \left[(av^2 + bv)e^{-nv}\right] = \lambda. \quad (32)
\]

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

\[
\frac{\partial}{\partial u} \left[(u^2 + bu)e^{-u}\right] = \lambda. \quad (33)
\]

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:

\[
\{p_s^* : p_s^* > 0\} = \left\{ \frac{c_1}{n}, \frac{c_2}{n}, \frac{c_3}{n} \right\}, \quad (34)
\]

with \(c_s^*\) not necessarily distinct. Furthermore, let

\[
m_i^* = \# \{ s : p_s^* = \frac{c_s}{n} \}. \quad (35)
\]

Then our original program (29) is equivalent to:

\[
\max \quad \frac{\sum_{i=1}^3 (m_i c_i^2 + m_i c_i e^{-c_i})}{n} - \left( \frac{\sum_{i=1}^3 m_i c_i e^{-c_i}}{n} \right)^2
\]

\[\text{s.t.} \quad \begin{cases}
0 \leq c_i \text{ and } 0 \leq m_i \text{ and } m_i \in \mathbb{Z}, & i = 1, 2, 3 \\
\sum_{i=1}^3 c_i m_i = n \\
\sum_{i=1}^3 m_i = m,
\end{cases} \quad (36)\]

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

\[
\max \quad \frac{\sum_{i=1}^3 (m_i c_i^2 + m_i c_i e^{-c_i})}{n} - \left( \frac{\sum_{i=1}^3 m_i c_i e^{-c_i}}{n} \right)^2
\]

\[\text{s.t.} \quad \begin{cases}
0 \leq c_i \text{ and } 0 \leq m_i, & i \in I \\
\sum_{i \in I} c_i m_i \leq n \\
\sum_{i \in I} m_i \leq m,
\end{cases} \quad (37)\]

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 \(m_i c_i^2 e^{-c_i}\) and \(m_i c_i e^{-c_i}\) 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 \((c_1), (m_1)\) is optimal and such that the number \(#\{i : m_i c_i \neq 0\}\) is smallest possible. If \(#\{i : m_i c_i \neq 0\} \leq 1\) there is nothing to prove, thus we assume \(#\{i : m_i c_i \neq 0\} \geq 2\). We can assume \(m_i c_i = 0\) for \(i = 1, 2\) (due to the symmetry). The LICQ holds, as the gradients of possibly active constraints

\[
\frac{\partial}{\partial (c_i)} [(c_i, (m_i))^3] = (0, 0, 1, 0, 0, 0),
\]

\[
\frac{\partial}{\partial (c_i)} [(c_i, (m_i)) m_3] = (0, 0, 0, 0, 0, 1),
\]

\[
\frac{\partial}{\partial (c_i)} [(c_i, (m_i)) \sum_{i=1}^3 c_i m_i] = (m_1, m_2, m_3, c_1, c_2, c_3),
\]

\[
\frac{\partial}{\partial (c_i)} [(c_i, (m_i)) \sum_{i=1}^3 m_i] = (0, 0, 0, 1, 1, 1),
\]

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:

\[
\frac{\partial}{\partial u} [(u^2 + bu)e^{-u}] = n \lambda \mu + n \mu
\]

where \(b = 1 - \frac{2}{n} \sum_{i=1}^3 m_i c_i e^{-c_i}\); 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 \(g(u) \equiv (u^2 + bu)e^{-u}\) 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).

 Indeed \(c_1 \neq c_2\) must be on the different sides of the stationary point (by the mean-value theorem [52], [53]); 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:

\[
\begin{align*}
\max & \quad \frac{(m_1' c_1'^2 + m_1 c_1)e^{-c_1} - (m_1 c_1 e^{-c_1})^2}{n} \\
\text{s.t.} & \quad 0 \leq c_1 \text{ and } 0 \leq m_1 \quad (40) \quad \text{and} \quad c_1 m_1 \leq n \quad m_1 \leq m.
\end{align*}
\]

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

\[
P = \frac{(m_1' c_1'^2 + m_1 c_1)e^{-c_1} - (m_1 c_1 e^{-c_1})^2}{n};
\]

it suffices to construct \((c_1', m_1')\) feasible for the program (36) and such that \(P' = \sum_{i=1}^{m_1} (m_1' c_1'^2 + m_1' c_1) e^{-c_1'} - (\sum_{i=1}^{m_1} m_1' c_1'^2 - (\sum_{i=1}^{m_1} m_1' c_1) e^{-c_1'})^2\)

\[
P - P' \leq O(n^{-1}),
\]

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 \(m_1' = \max\{\{m_1 - 1\}, 1\}, m_2 = 1, m_3 = 0\), also \(c_1' = \frac{m_1 c_1}{m_1'}\), \(c_2' = n - m_1 c_1, c_3' = 0\). Note that \(m_1', c_1' \geq 0\) and \(m_3 = 0\); \(m_4' = n\); \(m_4' = n\); \(c_1' = c_1\); \(c_2' = c_2\); \(c_3' = c_3\); \(c_4' = c_4\).

The bound on \(P - P'\) trivially follows when \(m_3 \leq 2\) because for \(m_1 = O(1)\) and \(m_1' = O(1)\) we have \(P = O(n^{-1}), |P'| = O(n^{-1})\). We further assume that \(m_1 \geq 2\), which implies \(m_1' = \Theta(m_1)\) and \(c_1' = c_1\). To bound \(P - P'\) we observe that:

\[
P - P' = \frac{m_1 c_1}{n} \left((c_1 + 1)e^{-c_1} - (c_1' + 1)e^{-c_1'}\right)
- \frac{c_2'(c_2' + 1)e^{-c_2'}}{n}
- \left(\frac{m_1 c_1 e^{-c_1}}{n}\right)^2 + \left(\frac{m_1 c_1 e^{-c_1} + c_2' e^{-c_2'}}{n}\right)^2
\]

(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 \(h(u) \triangleq (1 + w)e^{-u}\). By the mean-value theorem \(\frac{m_1 c_1}{n} \cdot h(c_1) - h(c_1') = \frac{m_1 c_1}{n} \cdot O(c_1 e^{-\Theta(c_1)}); \text{since } |c_1' - c_1| = O(c_1/m_1) = O(c_1/m_1)\), we can upper-bound as \(O(e^{-\Theta(c_1)})) = O(n^{-1})\).

The second term is negative can be ignored.

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

\[
\frac{|x - y|}{n} = O(e^{-\Theta(c_1)})) = O(n^{-1})\text{ and } x^2 - y^2 \leq O(n^{-1})\).
\]

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

\[
\max \quad w(1 + c)e^{-c} - (we^{-c})^2
\]

\[
\text{s.t.} \quad \begin{cases}
0 \leq w \leq 1 \\
w \leq \frac{c_1}{m_1}.
\end{cases}
\]

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

8) Unbounded Dimension: The supremum \(\alpha\) of (29) can be approached on a sequence of distributions \((p_k(e))\), \(k = 1, 2, \ldots\).

For any fixed \(\epsilon > 0\), let \(p_k\) 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 \((p_k')\) with \(p_k\) which is optimal under the support constraint \(m_k\); the objective can only increase, thus it is still at least \(\alpha - \epsilon\).

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 \(w \leq \frac{n}{m}\) can be ignored. This shows that limiting the support to \(m = O(n)\) we can approximate the supremum up to \(2\epsilon\), for arbitrarily small \(\epsilon\). 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 \(b \triangleq \frac{m_1}{n}\) and consider the program:

\[
\max \quad w(1 + c)e^{-c} - (we^{-c})^2
\]

\[
\text{s.t.} \quad 0 \leq w \leq 1 \quad \text{and} \quad w \leq bc
\]

(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 \(w = \frac{(1+ce^c)}{2}\); the program becomes maximizing \(\frac{1+ce^c}{2} - \frac{(1+ce^c)}{2} \leq \frac{(1+ce^c)}{2} < bc\), 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

\[
\max \quad (1 + c)e^{-c} - e^{-2c},
\]

(45)

and when \(w = bc\) the program becomes

\[
\max \quad b(c + c^2)e^{-c} - b^2 c^2 e^{-2c}.
\]

(46)

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

Suppose that \(b \leq \frac{1}{W(2)}\); then (45) is maximized at \(c = \frac{1}{b}\); this matches the objective of (46) at \(c = \frac{1}{b}\), so the optimal value of (46) is bigger or equal than that of (45).
Suppose now that \( b \geq \frac{1}{W(2)} \), then (45) is maximized at \( c = W(2) \), with the value \( h(W(2)) \). This time we claim that (46) is smaller or equal than (45). If \( \frac{1}{c} \leq \frac{1}{e} \) for optimal \( c \), then under the constraint (45), we have \( b \leq \frac{1}{k} \leq \frac{(k+1)e}{2} \). Since the objective increases in \( h(b) \), this in turn at most \( h(W(2)) \).

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,” 2015.

[10] D. Berend, A. Kontorovich et al., “On the concentration of the missing mass,” Electronic Communications in Probability, vol. 18, 2013.

[11] 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.

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

[13] D. Berend, A. Kontorovich, and G. Zadaganski, “The expected missing mass under an entropy constraint,” Entropy, vol. 19, no. 7, p. 315, 2017.

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

[15] A. Orlitsky and A. T. Suresh, “Competitive distribution estimation: Why is good-turing good?” in NIPS, 2015, pp. 2143–2151.

[16] 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.

[17] Y. Hao and A. Orlitsky, “Doubly-competitive distribution estimation,” in International Conference on Machine Learning, PMLR, 2019, pp. 2614–2623.

[18] ——, “Profile entropy: A fundamental measure for the learnability and compressibility of distributions,” Advances in Neural Information Processing Systems, vol. 33, 2020.

[19] 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.

[20] A. N. 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.

[21] 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.

[22] 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.

[23] Y. Bu, S. Zou, Y. Liang, and V. V. Veeravalli, “Estimation of kl divergence: Optimal minimax rate,” IEEE Transactions on Information Theory, vol. 64, no. 4, pp. 2648–2674, 2018.

[24] G. Obrębski and M. Skorski, “Renyi entropy estimation revisited,” in Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2017). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

[25] M. Mohri, A. Rostamizadeh, and A. Talwalkar, Foundations of machine learning. MIT press, 2018.

[26] 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.
[46] 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

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

[48] 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

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

[50] A. D. Barbour and P. Hall, “On the rate of poisson convergence,” in *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 95, no. 3. Cambridge University Press, 1984, pp. 473–480.

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

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

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