

Minimax Risk for Missing Mass Estimation

Nikhilesh Rajaraman, Andrew Thangaraj
 Department of Electrical Engineering
 Indian Institute of Technology Madras
 Chennai 600036, India
 andrew@ee.iitm.ac.in

Ananda Theertha Suresh
 Google Research
 New York, USA
 theertha@google.com

Abstract—The problem of estimating the missing mass or total probability of unseen elements in a sequence of n random samples is considered under the squared error loss function. The worst-case risk of the popular Good-Turing estimator is shown to be between $0.6080/n$ and $0.6179/n$. The minimax risk is shown to be lower bounded by $0.25/n$. This appears to be the first such published result on minimax risk for estimation of missing mass, which has several practical and theoretical applications.

I. INTRODUCTION

Given independent samples from an unknown distribution, missing mass estimation asks for the sum of the probability of the unseen elements. Missing mass estimation is a basic problem in statistics and has wide applications in several fields ranging from language modeling [1], [2] to ecology [3]. Perhaps the most used missing mass estimator is the Good-Turing estimator which was proposed in a seminal paper by I. J. Good and Alan Turing in 1953 [4]. The Good-Turing estimator is used in support estimators [3], entropy estimators [5] and unseen species estimators [6]. To describe the estimator and the results, we need a modicum of nomenclature.

Let p be an underlying unknown distribution over an unknown domain \mathcal{X} . Let $X^n \triangleq (X_1, X_2, \dots, X_n)$ be n independent samples from p . For $x \in \mathcal{X}$, let $N_x(X^n)$ be the number of appearances of x in X^n . Upon observing X^n , our goal is to estimate the missing mass

$$M_0(X^n) \triangleq \sum_{u \in \mathcal{X}} p(u) \mathbb{I}(N_u(X^n) = 0), \quad (1)$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. For example, if $\mathcal{X} = \{a, b, c, d\}$ and $X^3 = bcb$, then $M_0(X^3) = p(a) + p(d)$. The above sampling model for estimation is termed the multinomial model. We note that $1 - M_0(X^n)$ is often referred as sample coverage in the literature [7].

An estimator for missing mass $\hat{M}_0(X^n)$ is a mapping from $\mathcal{X}^n \rightarrow [0, 1]$. For a distribution p , the ℓ_2^2 risk of the estimator $\hat{M}_0(X^n)$ is

$$R_n(\hat{M}_0, p) \triangleq E_{X^n \sim p} [(\hat{M}_0(X^n) - M_0(X^n))^2],$$

and the worst-case risk over all distributions is

$$R_n(\hat{M}_0) \triangleq \max_p R_n(\hat{M}_0, p),$$

and minimax mean squared loss or minimax risk is

$$R_n^* = \min_{\hat{M}_0} R_n(\hat{M}_0).$$

The goal of this paper is to characterize R_n^* .

A. Good-Turing estimator and previous results

Let

$$\Phi_i(X^n) \triangleq \sum_{u \in \mathcal{X}} \mathbb{I}(N_u(X^n) = i)$$

denote the number of symbols that have appeared i times in X^n , $1 \leq i \leq n$. For example, if $X^3 = abc$, then $\Phi_1 = 3$ and $\Phi_i = 0$ for all $i > 1$. The Good-Turing estimator [4] for the missing mass is

$$M^{\text{GT}}(X^n) \triangleq \frac{\Phi_1(X^n)}{n}.$$

One of the first theoretical analysis of the Good-Turing estimator was in [8], where it was shown that

$$|E[M^{\text{GT}}(X^n) - M_0(X^n)]| \leq \frac{1}{n}. \quad (2)$$

This shows that the bias of the Good-Turing estimator falls as $1/n$. They further showed that with probability $\geq 1 - \delta$,

$$|M^{\text{GT}}(X^n) - M_0(X^n)| \leq \frac{2}{n} + \sqrt{\frac{2 \ln(3/\delta)}{n}} (1 + 2 \ln(3n/\delta)).$$

Various properties of the Good-Turing estimator and several variations of it have been analyzed for distribution estimation and compression [9], [10], [11], [12], [13], [14], [15]. Several concentration results on missing mass estimation are also known [16], [17]. Despite all this work, the risk of the Good-Turing estimator and the minimax risk of missing mass estimation have still not been conclusively established.

B. New results

Unlike parameters of a distribution, missing mass itself is a function of the observed sample and that makes finding the exact minimax risk difficult.

We first analyze the risk of the Good-Turing estimator and show that for any distribution p ,

$$R_n(M^{\text{GT}}, p) = \frac{1}{n} E \left[\frac{2\Phi_2}{n} + \frac{\Phi_1}{n} \left(1 - \frac{\Phi_1}{n} \right) \right] + o\left(\frac{1}{n}\right),$$

where Φ_i is abbreviated notation for $\Phi_i(X^n)$. By maximizing the RHS in the first equation above over all distributions, in Theorem 4, we show that

$$\frac{0.6080}{n} + o\left(\frac{1}{n}\right) \leq R_n(M^{\text{GT}}) \leq \frac{0.6179}{n} + o\left(\frac{1}{n}\right).$$

We note that under the multinomial model, the numbers of occurrences of symbols are correlated, and this makes finding the worst case distribution for the Good-Turing estimator difficult.

We then prove estimator-independent information-theoretic lower bounds on R_n^* using two approaches. We first compute the lower bound via Dirichlet prior approach [18]. In Lemma 7, we show that

$$R_n^* \geq \frac{4}{27n}.$$

We then improve the constant by reducing the problem of missing mass estimation to that of distribution estimation. In particular, in Theorem 11, we show that

$$R_n^* \geq \frac{1}{4n} + o\left(\frac{1}{n}\right)$$

Combining the lower and upper bounds, we get

$$\frac{0.25}{n} + o\left(\frac{1}{n}\right) \leq R_n^* \leq \frac{0.6179}{n} + o\left(\frac{1}{n}\right),$$

Finding the exact minimax risk for the missing mass estimation problem remains an open question.

The rest of the paper is organized as follows. In Section II, we analyze the Good-Turing estimator. In Section III-A, we use Dirichlet prior approach to obtain lower bounds and in Section III-B we obtain lower bounds via reduction.

II. RISK OF GOOD-TURING ESTIMATOR

The analysis of [8] can be extended to characterize the risk of the Good-Turing estimator for missing mass. The squared error of the Good-Turing estimator $M^{\text{GT}}(X^n)$ can be written down as follows:

$$\begin{aligned} & (M^{\text{GT}}(X^n) - M_0(X^n))^2 \\ &= \left(\sum_{u \in \mathcal{X}} \frac{1}{n} \mathbb{I}(N_u = 1) - p(u) \mathbb{I}(N_u = 0) \right) \\ & \quad \left(\sum_{v \in \mathcal{X}} \frac{1}{n} \mathbb{I}(N_v = 1) - p(v) \mathbb{I}(N_v = 0) \right) \\ &= \frac{1}{n^2} \sum_{u, v \in \mathcal{X}} \left(\mathbb{I}(N_u = 1) \mathbb{I}(N_v = 1) \right. \\ & \quad \left. - 2np(u) \mathbb{I}(N_u = 0) \mathbb{I}(N_v = 1) \right. \\ & \quad \left. + n^2 p(u)p(v) \mathbb{I}(N_u = 0) \mathbb{I}(N_v = 0) \right) \quad (3) \end{aligned}$$

For $u, v \in \mathcal{X}$, $E[\mathbb{I}(N_u = i) \mathbb{I}(N_v = j)] = \mathbb{P}(N_u = i, N_v = j)$. Using the notation $P_n(i, j) = \mathbb{P}(N_u(X^n) = i, N_v(X^n) = j)$, we get

$$\begin{aligned} R_n(M^{\text{GT}}, p) &= \frac{1}{n^2} \sum_{u, v \in \mathcal{X}} \left(P_n(1, 1) - 2np(u)P_n(0, 1) \right. \\ & \quad \left. + n^2 p(u)p(v)P_n(0, 0) \right). \quad (4) \end{aligned}$$

The probability $P_n(i, j)$ can be written down as

$$P_n(i, j) = \begin{cases} \binom{n}{i, j} p(u)^i p(v)^j (1 - p(u) - p(v))^{n-i-j}, & u \neq v, \\ \binom{n}{i} p(u)^i (1 - p(u))^{n-i}, & u = v, i = j, \end{cases} \quad (5)$$

where $\binom{n}{i, j} = \frac{n!}{i!j!(n-i-j)!}$ and $\binom{n}{i} = \frac{n!}{i!(n-i)!}$. The summation in (4) is first split into two cases: $u \neq v$ and $u = v$. Denoting $P(u, v) = p(u)p(v)(1 - p(u) - p(v))^{n-2}$, we have, for $u \neq v$,

$$\begin{aligned} p(u)p(v)P_n(0, 0) &= (1 - p(u) - p(v))^2 P(u, v), \\ p(u)P_n(0, 1) &= n(1 - p(u) - p(v))P(u, v), \\ P_n(1, 1) &= n(n-1)P(u, v). \end{aligned}$$

For $u = v$, observe that $P_n(0, 1) = 0$. Using the above observations, the summation in (4) simplifies to

$$\begin{aligned} R_n(M^{\text{GT}}, p) &= \frac{1}{n} \sum_{\substack{u, v \in \mathcal{X} \\ v \neq u}} P(u, v) \left[n(p(u) + p(v))^2 - 1 \right] \\ & \quad + \frac{1}{n} \sum_{u \in \mathcal{X}} \left[p(u)(1 - p(u))^{n-1} + np(u)^2(1 - p(u))^n \right]. \quad (6) \end{aligned}$$

The following lemma is useful in bounding certain terms in the first summation above as a function of n , independent of the unknowns \mathcal{X} and p .

Lemma 1. For $i \geq 1, j \geq 1$,

$$\sum_{u, v \in \mathcal{X}, u \neq v} p(u)^i p(v)^j (1 - p(u) - p(v))^n \leq \frac{(i-1)! (j-1)! n!}{(n+i+j-2)!}.$$

Proof: Let X and Y be a pair of independent and identical random variables with marginal distribution p . Define a random variable $T(X, Y)$, whose value $T(u, v) = 0$ for $u = v$ and, for $u \neq v$,

$$T(u, v) = \binom{n+i+j-2}{i-1, j-1} p(u)^{i-1} p(v)^{j-1} (1 - p(u) - p(v))^n.$$

We see that $T(X, Y)$ is a probability for $X \neq Y$, and that it takes values in $[0, 1]$ in all cases. Therefore, its expectation

$$\begin{aligned} E[T(X, Y)] &= \sum_{\substack{u, v \in \mathcal{X} \\ u \neq v}} p(u)p(v)T(u, v) \\ &= \sum_{\substack{u, v \in \mathcal{X} \\ u \neq v}} \binom{n+i+j-2}{i-1, j-1} p(u)^i p(v)^j (1 - p(u) - p(v))^n \leq 1, \end{aligned}$$

which concludes the proof. \blacksquare

A useful univariate version of Lemma 1 is the following.

Lemma 2. For $i \geq 1$,

$$\sum_{u \in \mathcal{X}} p(u)^i (1 - p(u))^n \leq \frac{(i-1)! n!}{(n+i-1)!}.$$

Proof: For $X \sim p$, define $T(X) = \binom{n+i-1}{i-1} p(X)^{i-1} (1 - p(X))^n$ and follow the proof of Lemma 1. \blacksquare

Using Lemma 1, observe that

$$\sum_{u, v \in \mathcal{X}, u \neq v} P(u, v) (p(u) + p(v))^2 = o(1/n). \quad (7)$$

Therefore, the risk can be written as

$$R_n(M^{GT}, p) = \frac{1}{n} \left[\sum_{u \in \mathcal{X}} p(u)(1-p(u))^{n-1} - \sum_{\substack{u, v \in \mathcal{X} \\ v \neq u}} P(u, v) + \sum_{u \in \mathcal{X}} np(u)^2(1-p(u))^n \right] + o(1/n). \quad (8)$$

The summation terms above can be rewritten as follows:

$$\begin{aligned} \sum_{u \in \mathcal{X}} p(u)(1-p(u))^{n-1} &= E \left[\frac{\Phi_1(X^n)}{n} \right], \quad (9) \\ \sum_{u \in \mathcal{X}} np(u)^2(1-p(u))^n &= \frac{2}{n-1} \sum_{u \in \mathcal{X}} P_n(2, 0)(1-p(u))^2 \\ &\stackrel{(a)}{=} \frac{2}{n-1} \sum_{u \in \mathcal{X}} P_n(2, 0) \pm o\left(\frac{1}{n}\right) \\ &= E \left[\frac{2\Phi_2(X^n)}{n} \right] \pm o\left(\frac{1}{n}\right), \quad (10) \end{aligned}$$

where (a) follows using Lemma 2.

$$\begin{aligned} \sum_{\substack{u, v \in \mathcal{X} \\ v \neq u}} P(u, v) &= \frac{1}{n(n-1)} \sum_{\substack{u, v \in \mathcal{X} \\ v \neq u}} P_n(1, 1) \\ &= \frac{1}{n(n-1)} E \left[\sum_{\substack{u, v \in \mathcal{X} \\ v \neq u}} \mathbb{I}(N_u(X^n) = 1) \mathbb{I}(N_v(X^n) = 1) \right] \\ &= E \left[\frac{1}{n(n-1)} \Phi_1(X^n) (\Phi_1(X^n) - 1) \right] \\ &= E \left[\frac{\Phi_1^2(X^n)}{n} \right] \pm o(1). \quad (11) \end{aligned}$$

Using the above expressions in (8), we get the following characterization of the risk.

Theorem 3. *The risk of the Good-Turing estimator under squared error loss satisfies*

$$R_n(M^{GT}, p) = \frac{1}{n} E \left[\frac{2\Phi_2}{n} + \frac{\Phi_1}{n} \left(1 - \frac{\Phi_1}{n} \right) \right] + o\left(\frac{1}{n}\right). \quad (12)$$

A. Upper bound on risk

To obtain a tight upper bound on the risk, we start with the following upper bound on one of the terms in (8):

$$\sum_{u \in \mathcal{X}} np(u)^2(1-p(u))^n \leq \sum_{u \in \mathcal{X}} p(u) \left(np(u)e^{-np(u)} \right) \leq e^{-1}, \quad (13)$$

where the first step follows because $1-x \leq e^{-x}$ for a fraction x , and the second step follows because $te^{-t} \leq e^{-1}$ for $t \geq 0$. Using (9), (10) and (13) in (8), an upper bound for the risk of the Good-Turing estimator is

$$R_n(M^{GT}, p) \leq \frac{1}{n} E \left[\frac{\Phi_1}{n} \left(1 - \frac{\Phi_1}{n} \right) \right] + \frac{e^{-1}}{n} \pm o\left(\frac{1}{n}\right)$$

$$\leq \frac{0.25 + e^{-1}}{n} \pm o\left(\frac{1}{n}\right), \quad (14)$$

where the last step follows because $x(1-x) \leq 0.25$ for a fraction x . The above constant $e^{-1} + 0.25 \approx 0.6179$ is not best possible, and could be marginally improved by more careful analysis. However, we show that the improvement is not significant through a lower bound on $R_n(M^{GT}) = \max_p R_n(M^{GT}, p)$ by picking p to be a suitable uniform distribution.

B. Lower bound on the Good-Turing worst-case risk

A lower bound can be obtained for the worst case risk of the Good-Turing estimator by evaluating the risk for the uniform distribution p_U on \mathcal{X} . Let $|\mathcal{X}| = cn$ and $p_U(x) = \frac{1}{cn}$ for all $x \in \mathcal{X}$, where c is a positive constant. Using (8), we get

$$\begin{aligned} R_n(M^{GT}, p_U) &= \frac{1}{n} \left[\frac{cn \cdot n}{(cn)^2} \left(1 - \frac{1}{cn} \right)^n + \frac{cn}{cn} \cdot \left(1 - \frac{1}{cn} \right)^{n-1} \right. \\ &\quad \left. - \left(\frac{cn}{cn} \cdot \left(1 - \frac{1}{cn} \right)^{n-1} \right)^2 \right] + o\left(\frac{1}{n}\right) \\ &\stackrel{(a)}{=} \frac{1}{n} \left(\left(\frac{1}{c} + 1 \right) \left(1 - \frac{1}{cn} \right)^n - \left(1 - \frac{1}{cn} \right)^{2n} \right) + o\left(\frac{1}{n}\right) \\ &\stackrel{(b)}{=} \frac{1}{n} \left(\left(\frac{1}{c} + 1 \right) e^{-\frac{1}{c}} - e^{-\frac{2}{c}} \right) + o\left(\frac{1}{n}\right) \quad (15) \end{aligned}$$

where the reasoning for the steps is as follows:

- replacing $\left(1 - \frac{1}{cn} \right)^{n-1}$ with $\left(1 - \frac{1}{cn} \right)^n (1 + o(1))$.
- using the fact that $\left(1 - \frac{1}{cn} \right)^n = e^{-1/c} (1 + o(1))$.

The coefficient of $\frac{1}{n}$ in (15) can be maximized numerically to obtain a maximum value of 0.6080 at $c \approx 1.1729$. Hence, from (14) and (15), we have:

Theorem 4. *The worst-case risk of the Good-Turing estimator satisfies the following bounds:*

$$\frac{0.6080}{n} + o\left(\frac{1}{n}\right) \leq R_n(M^{GT}) \leq \frac{0.6179}{n} + o\left(\frac{1}{n}\right). \quad (16)$$

Therefore, the constant in (14) is fairly tight.

III. LOWER BOUNDS ON THE MINIMAX RISK

In this section, we consider lower bounds on the squared error risk of an arbitrary estimator of missing mass. The main result is that the minimax risk is lower-bounded by c/n for a constant c . Two methods are described for finding lower bounds - the first one is a Dirichlet prior approach, and the second one is reduction of the missing mass problem to a distribution estimation problem.

Both approaches provide the same order of $1/n$ for the lower bound, but the second reduction approach provides a better constant. However, the Dirichlet prior approach has significant potential for further optimization for better constants, and is an interesting extension of the standard prior method to the case of estimation of random variables such as missing mass, which depend on both the distribution p and the sample X^n .

A. Lower Bounds via Prior Distributions

The first approach is to bound the minimax risk by the average risk obtained by averaging over a family of distributions with a prior. Let P be a random variable over a family of distributions \mathcal{P} , having an alphabet $\mathcal{X} = \{0, 1, 2, \dots, k-1\}$. In the following section, the missing mass will be denoted as $M_0(X^n, p)$ to explicitly show the dependence on the distribution p .

Lemma 5. For any missing mass estimator $\hat{M}_0(X^n)$ and a random variable P over a family of distributions \mathcal{P} ,

$$\begin{aligned} \min_{\hat{M}_0} \max_{p \in \mathcal{P}} \mathbb{E}_{X^n \sim p} \left(M_0(X^n, p) - \hat{M}_0(X^n) \right)^2 \\ \geq \mathbb{E}_{X^n \sim P} \left[\text{var}_{P|X^n} [M_0(X^n, P) | X^n] \right] \end{aligned}$$

Proof:

$$\begin{aligned} \min_{\hat{M}_0} \max_{p \in \mathcal{P}} \mathbb{E} \left(M_0(X^n, p) - \hat{M}_0(X^n) \right)^2 \\ \geq \min_{\hat{M}_0} \mathbb{E}_P \left(\mathbb{E}_{X^n|P} \left(M_0(X^n, P) - \hat{M}_0(X^n) \middle| P \right)^2 \right) \\ \stackrel{(a)}{=} \min_{\hat{M}_0} \mathbb{E}_{X^n} \left(\mathbb{E}_{P|X^n} \left(M_0(X^n, P) - \hat{M}_0(X^n) \middle| X^n \right)^2 \right) \\ \stackrel{(b)}{=} \mathbb{E}_{X^n \sim P} \left[\text{var}_{P|X^n} [M_0(X^n, P) | X^n] \right] \end{aligned}$$

where (a) follows from the law of total expectation and (b) follows from the fact that (a) is minimized when $\hat{M}_0(X^n) = \mathbb{E}_{P|X^n} (M_0(X^n, P) | X^n)$. ■

Lemma 5 gives us a family of bounds depending on the distribution of the prior P . The RHS in Lemma 5 can be computed exactly for a Dirichlet prior with some analysis.

Lemma 6. Suppose P has a Dirichlet distribution $\text{Dir}(k, \alpha)$, where $\alpha = (\alpha_0, \alpha_1, \dots, \alpha_{k-1})$. Then, we have

$$\begin{aligned} \mathbb{E}_{X^n} \left[\text{var}_{P|X^n} [M_0(X^n, P) | X^n] \right] \\ = \frac{B(a, n)}{(a+n)^2 (a+n+1)} \left(\sum_{u \in \mathcal{X}} \frac{\alpha_u (a+n) - \alpha_u^2}{B(a - \alpha_u, n)} \right. \\ \left. - \sum_{u \in \mathcal{X}} \sum_{v \in \mathcal{X}, v \neq u} \frac{\alpha_u \alpha_v}{B(a - \alpha_u - \alpha_v, n)} \right), \end{aligned}$$

where $B(\cdot, \cdot)$ is the Beta function and $a = \sum_{u \in \mathcal{X}} \alpha_u$.

We skip the details for want of space.

Let $\alpha = (\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n})$ and $k = cn^2$. For this choice of parameters, the expression in Lemma 6 can be bounded as

$$\mathbb{E}_{X^n} \left[\text{var}_{P|X^n} [M_0(X^n, P) | X^n] \right] \geq \frac{1}{n} \cdot \frac{c}{(c+1)^3} + o\left(\frac{1}{n}\right),$$

where, once again, we skip the details. The coefficient of $\frac{1}{n}$ attains a maximum value of $\frac{4}{27}$ when $c = \frac{1}{2}$, which results in the following bound on the minimax risk:

Lemma 7.

$$\min_{\hat{M}_0} \max_{p \in \mathcal{P}} \mathbb{E} \left(M_0(X^n, p) - \hat{M}_0(X^n) \right)^2 \geq \frac{4}{27n} + o\left(\frac{1}{n}\right)$$

The bound is worse than the $\frac{1}{4n}$ bound obtained from distribution estimation in the next section, but it can possibly be improved by better selection of the prior.

B. Lower bounds via Distribution Estimation

To bound the minimax risk for missing mass estimation, one approach is to reduce the problem to that of estimating a distribution. Let \mathcal{P} be the set of distributions over the set $\mathcal{X} = \{0, 1\}$ such that for all $p \in \mathcal{P}$, $p(0) \geq \frac{1}{2}$. A known result (refer [19], [20] for instance) states that the minimax ℓ^2 loss in estimating $p(0)$ is $\frac{1}{4n}$. More precisely, let $\hat{p}(X^n)$ be an estimator for $p(0)$ from a random sample X^n distributed according to p . Then, we have

Lemma 8.

$$\min_{\hat{p}(0)} \max_{p \in \mathcal{P}} \mathbb{E}_{X^n \sim p} (p(0) - \hat{p}(X^n))^2 = \frac{1}{4n} + o\left(\frac{1}{n}\right)$$

For an arbitrary positive integer k , let \mathcal{P}_c be the set of distributions over the set $\mathcal{X} = \{0, 1, 2, \dots, k-1\}$, such that for any $p_c \in \mathcal{P}_c$, we have $p_c(0) \geq \frac{1}{2}$ and $p_c(i) = \frac{1-p_c(0)}{k}$ for all $i \geq 1$. We can use Lemma 8 to obtain minimax bounds in estimating $p_c(0)$ for this family of distributions as well. Let $\hat{p}_c(X^n)$ be an estimator for p_c from a random sample X^n distributed according to p_c . Let $\hat{p}_c(X^n, i)$ be the probability \hat{p}_c assigns to the symbol i .

Lemma 9.

$$\min_{\hat{p}_c(0)} \max_{p_c \in \mathcal{P}_c} \mathbb{E} (p_c(0) - \hat{p}_c(X^n, 0))^2 \geq \frac{1}{4n} + o\left(\frac{1}{n}\right)$$

Proof: Suppose we want to estimate an unknown distribution $p \in P$ and we have an estimator \hat{p}_c for distributions in \mathcal{P}_c . Then we can use \hat{p}_c to estimate p as follows. Take the observed sample distributed according to p , and if it is 0, keep it as it is. If it is 1, then replace it with a uniformly sampled random variable over $\{1, 2, \dots, k\}$. The result of this sampling process is a distribution p_c in \mathcal{P}_c with $p_c(0) = p(0)$. Thus, any estimator for distributions in \mathcal{P}_c can be reduced to an estimator for distributions in \mathcal{P} and

$$\begin{aligned} \min_{\hat{p}_c(0)} \max_{p_c \in \mathcal{P}_c} \mathbb{E}_{X^n \sim p_c} (p_c(0) - \hat{p}_c(X^n, 0))^2 \\ \geq \min_{\hat{p}(0)} \max_{p \in \mathcal{P}} \mathbb{E}_{X^n \sim p} (p(0) - \hat{p}(X^n))^2 \end{aligned}$$

and the proof follows from Lemma 8. ■

Lemma 10. Let $k = e^n$. With probability at least $1 - 1/2^n$, the missing mass $M_0(X^n)$ satisfies

$$M_0(X^n) = 1 - p(0) + O(ne^{-n}).$$

Proof: Probability of symbol 0 appearing at least once in X^n is $1 - (1 - p(0))^n \geq 1 - 1/2^n$. Furthermore, at most n distinct symbols from $1, 2, \dots, k-1$ can appear in X^n . Hence, with probability $1 - 1/2^n$, the observed mass $1 - M_0(X^n)$ satisfies

$$p(0) \leq 1 - M_0(X^n) \leq p(0) + (1 - p(0))ne^{-n},$$

and hence follows the lemma. ■

From Lemmas 9 and 10, we can obtain a lower bound of $1/4n$ on the minimax risk of missing mass estimation. Combining the lower bound with the upper bound on the risk of the Good-Turing estimator from Theorem 4, we have the following:

Theorem 11. *The minimax risk of missing mass estimation, denoted R_n^* , satisfies the following bounds:*

$$\frac{0.25}{n} + o\left(\frac{1}{n}\right) \leq R_n^* \leq \frac{0.6179}{n} + o\left(\frac{1}{n}\right).$$

IV. SUMMARY AND FUTURE DIRECTIONS

We studied the problem of missing mass estimation and showed that the minimax risk lies between $0.617/n$ and $1/4n$. We further showed that the risk of the Good-Turing estimator lies between $0.608/n$ and $0.617/n$.

Our results pose several interesting questions for future work. Two natural questions are: (1) are there priors which yield better lower bounds on the minimax risk of missing mass? and (2) are there estimators that have better risk than the Good-Turing estimator?

We finally remark that it might be interesting to see if the minimax risk results imply better concentration results for the missing mass and the Good-Turing estimator.

V. ACKNOWLEDGEMENTS

Authors thank Alon Orlitsky for helpful discussions. Ananda Theertha Suresh thanks Jayadev Acharya for helpful comments.

REFERENCES

- [1] W. A. Gale and G. Sampson, "Good-Turing frequency estimation without tears," *Journal of Quantitative Linguistics*, vol. 2, no. 3, pp. 217–237, 1995.
- [2] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," in *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, ser. ACL '96, 1996, pp. 310–318.
- [3] 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.
- [4] I. J. Good, "The population frequencies of species and the estimation of population parameters," *Biometrika*, vol. 40, no. 3-4, pp. 237–264, 1953.
- [5] V. Q. Vu, B. Yu, and R. E. Kass, "Coverage-adjusted entropy estimation," *Statistics in medicine*, vol. 26, no. 21, pp. 4039–4060, 2007.
- [6] 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.
- [7] R. K. Colwell, A. Chao, N. J. Gotelli, S.-Y. Lin, C. X. Mao, R. L. Chazdon, and J. T. Longino, "Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages," *Journal of Plant Ecology*, vol. 5, no. 1, pp. 3–21, 2012.
- [8] D. A. McAllester and R. E. Schapire, "On the convergence rate of Good-Turing estimators," in *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*, ser. COLT '00. Morgan Kaufmann Publishers Inc., 2000, pp. 1–6.
- [9] A. Orlitsky, N. P. Santhanam, and J. Zhang, "Always Good Turing: Asymptotically optimal probability estimation," in *Foundations of Computer Science*, 2003, pp. 179–188.
- [10] E. Druk and Y. Mansour, "Concentration bounds for unigrams language model," in *Conference on Learning Theory*, 2004, pp. 170–185.
- [11] A. B. Wagner, P. Viswanath, and S. R. Kulkarni, "Strong consistency of the Good-Turing estimator," in *International Symposium on Information Theory*, 2006, pp. 2526–2530.
- [12] —, "A better Good-Turing estimator for sequence probabilities," in *Information Theory, 2007. ISIT 2007. IEEE International Symposium on*. IEEE, 2007, pp. 2356–2360.
- [13] M. I. Ohannessian and M. A. Dahleh, "Rare probability estimation under regularly varying heavy tails," in *Conference on Learning Theory*, 2012, pp. 21.1–21.24.
- [14] J. Acharya, A. Jafarpour, A. Orlitsky, and A. T. Suresh, "Optimal probability estimation with applications to prediction and classification," in *Conference on Learning Theory*, 2013, pp. 764–796.
- [15] A. Orlitsky and A. T. Suresh, "Competitive distribution estimation: Why is Good-Turing good," in *Advances in Neural Information Processing Systems*, 2015, pp. 2143–2151.
- [16] D. Berend, A. Kontorovich *et al.*, "On the concentration of the missing mass," *Electronic Communications in Probability*, vol. 18, no. 3, pp. 1–7, 2013.
- [17] 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.
- [18] R. E. Krichevskiy, "Laplace's law of succession and universal encoding," *IEEE Transactions on information theory*, vol. 44, no. 1, pp. 296–303, 1998.
- [19] E. Lehmann and G. Casella, *Theory of Point Estimation*. Springer, 1998, ch. 5, pp. 311–312.
- [20] S. Kamath, A. Orlitsky, D. Pichapati, and A. T. Suresh, "On learning distributions from their samples," in *COLT*, 2015, pp. 1066–1100.