Improved Bounds for Minimax Risk of Estimating Missing Mass

Jayadev Acharya
Cornell University
acharya@cornell.edu

Yelun Bao
Tsinghua University
byl14@mails.tsinghua.edu.cn

Yuheng Kang
Tsinghua University
kyh14@mails.tsinghua.edu.cn

Ziteng Sun
Cornell University
zs335@cornell.edu

***Abstract*—Given n independent draws from a discrete distribution, what is the probability that the next draw will be a symbol that has not appeared before? We study the problem of estimating this missing mass probability under mean squared error. Our results include the following:**

1) **Mean squared error (MSE) of Good-Turing estimator is $\frac{0.608\ldots}{n} + o\left(\frac{1}{n}\right)$.**
2) **Minimax MSE for estimating missing mass of uniform distributions is $\frac{0.570\ldots}{n} + o\left(\frac{1}{n}\right)$.**

We prove that the minimax MSE R^* of missing mass estimation satisfies

$$\frac{0.570\ldots}{n} + o\left(\frac{1}{n}\right) \le R^* \le \frac{0.608\ldots}{n} + o\left(\frac{1}{n}\right).$$

The upper bound characterizes the maximum MSE of the celebrated Good-Turing (GT) estimator, and the lower bound characterizes the minimax MSE for estimating the missing mass for the class of uniform distributions.

I. Introduction

Given independent draws from a discrete distribution, what is the probability that the next draw will be a symbol that has not appeared before? This probability is called as the discovery probability, unseen probability, or as the missing mass probability. This random variable has applications across many scientific disciplines. For example, what is the probability that the next gene variant that we obtain is new, or the probability that the next species that we collect is a new species. .

A. Problem Set Up

Let $\mathcal{X}$ be a countable domain. Let p be an unknown distribution over $\mathcal{X}$. For example, p could be the distribution over all butterfly species ($\mathcal{X}$), some of which are yet to be discovered. Let $X_1^n \stackrel{\text{def}}{=} X_1, \ldots, X_n$ be drawn independently from p. The missing mass of X_1^n with respect to p is the random variable

$$M_0(X_1^n, p) \stackrel{\text{def}}{=} \sum_{x \in \mathcal{X}} p(x) \cdot \mathbb{I}\{x \text{ is not in } X_1^n\},$$

which is the probability that a new draw from p has not appeared before in X_1^n.

A missing mass estimator is a (possibly randomized) function $M : \mathcal{X}^n \to [0,1]$ that maps a length-n sample to a non-negative number in $[0,1]$. Adopting the framework proposed by Rajaraman, Suresh, and Thangaraj [1], we measure the performance of an estimator in the popular *mean squared error* (MSE) metric. More precisely, the MSE of the estimator M for a distribution p is

$$R(M,p) \stackrel{\text{def}}{=} \mathbb{E}\left[\left(M(X_1^n) - M_0(X_1^n, p)\right)^2\right],$$

where the expectation is over the randomness in the input sequences $X_1^n \sim p$ and possibly the randomness of the estimator. The MSE of M for a collection of distributions $\mathcal{P}$ is defined as

$$R(M,\mathcal{P}) \stackrel{\text{def}}{=} \sup_{p \in \mathcal{P}} R(M,p),$$

the worst MSE of M over distributions in $\mathcal{P}$. The *minimax risk*, or minimax MSE of missing mass estimation for $\mathcal{P}$ is

$$R^*(\mathcal{P}) \stackrel{\text{def}}{=} \inf_M R(M,\mathcal{P}) = \inf_M \sup_{p \in \mathcal{P}} R(M,p),$$

the MSE of the *best* estimator for the *worst* distribution in the class.

Let Δ be the class of *all* discrete distributions, namely the class of all distributions over all countable sets $\mathcal{X}$. Without loss of generality we assume that $\mathcal{X} = \mathbb{N}$, and for a distribution p and $i \in \mathbb{N}$, p_i is the probability of i. The focus of this work is to understand the minimax MSE of missing mass estimation for $\mathcal{P} = \Delta$, namely,

$$R^* \stackrel{\text{def}}{=} R^*(\Delta) = \inf_{M:\mathbb{N}^n \to [0,1]} \sup_{p \text{ over } \mathbb{N}} R(M,p). \quad (1)$$

II. Known Results and Our Contributions

Missing mass estimation has a long history, dating back at least half a century [2], [3]. The most

978-1-5386-4780-6/18/$31.00©2018 IEEE

celebrated estimator for missing mass is the Good-Turing estimator [2], denoted $M^{\mathtt{GT}}$ (Definition 1). The first theoretical analysis of the $M^{\mathtt{GT}}$ was given in [4], who among other things showed that $M^{\mathtt{GT}}$ is an *almost* unbiased estimator of the expected missing mass, namely for any p,

$$\mathbb{E}\left[\left|M^{\mathtt{GT}}(X_1^n,p)-M_0(X_1^n,p)\right|\right]\le\frac{1}{n}.$$

They also proved concentration results about the deviation of $M^{\mathtt{GT}}$ from the true M_0, and these arguments can be extended to show that the MSE of the GT estimator satisfies $R(M^{\mathtt{GT}},\Delta)=O\left(\frac{1}{n}\right)$. Arguments for estimating the MSE of Bernoulli random variables (*e.g.,* Chapter 5, Example 1.7 in [5]) can be used to show that $R^*=\Omega(1/n)$. Combining these two,

$$\Omega\left(\frac{1}{n}\right)\le R^*\le R\left(M^{\mathtt{GT}},\Delta\right)=O\left(\frac{1}{n}\right).$$

In a recent work, Rajaraman, Thangaraj, and Suresh [1] initiated the question of determining the precise constant for both the Good-Turing estimator as well as the best possible estimator, namely tight characterization of R^*, and $R(M^{\mathtt{GT}},\Delta)$.

A. Why care for the precise constant?

GT estimators, and its variants have been studied in many problems in probability theory, data compression, language modeling and other fields [6]–[11]. Better estimators for missing mass might provide better performance in practice for these tasks.

Another line of recent work [12]–[17] have studied the concentration and other properties of missing mass. Obtaining better bounds on the minimax MSE of missing mass estimators will inherently require estimators with better variance bounds than those known, and can improve the concentration results for the missing mass. In particular, perhaps finding the precise R^* can shed light on the precise constants in the exponents of missing mass deviation inequalities.

Missing mass estimation is an interesting question in the sense that the quantity we want to estimate is itself a random variable, and it is still simple enough that we might hope to characterize its missing mass precisely, and it might provide tools for establishing tight minimax MSE bounds for other problems.

B. Results

[1] proposed the minimax MSE estimation problem and proved the following bound on the performance of $M^{\mathtt{GT}}$. Up to an additive $\pm o(1/n)$,

$$\frac{0.608\ldots}{n}\le R\left(M^{\mathtt{GT}},\Delta\right)\le\frac{0.6179\ldots}{n}. \qquad (2)$$

They first derive an expression for the MSE of $M^{\mathtt{GT}}$ for any distribution p (Given in (6)). Upper bounding this expression over all distributions gives the upper bound. For the lower bound they consider $\mathcal{U}$, the class of all discrete uniform distributions. By maximizing (6) over $\mathcal{U}$, they obtain the lower bound. In particular, they showed that $R(M^{\mathtt{GT}},\Delta)\ge R(M^{\mathtt{GT}},\mathcal{U})=\frac{\max_{x\in[0,1]}x(1-x-\log x)}{n}+o(1/n)=\frac{0.608\ldots}{n}+o(1/n)$. We denote $\max_{x\in[0,1]}x(1-x-\log x)$ by $\alpha_{\mathtt{GT}}$.

They also study R^*, showed a lower bound that holds for *every estimator.* More precisely, by using the result on minimax MSE of estimating a Bernoulli they showed that $R^*\ge 0.25/n+o(1/n)$.

We prove a number of results for missing mass estimation. Our first result, proved in Section IV establishes the precise minimax MSE of $M^{\mathtt{GT}}$, in particular showing that the class of uniform distributions are the worst case instances.

Theorem 1.

$$R\left(M^{GT},\Delta\right)=R\left(M^{GT},\mathcal{U}\right)=\frac{\alpha_{GT}}{n}+o\left(\frac{1}{n}\right).$$

Then a natural question to ask is: *Does M^{GT} achieve R^*?*, namely is the Good-Turing estimator MSE optimal? While we are not able to answer this question, we will provide some arguments that suggest that $M^{\mathtt{GT}}$ might not achieve R^*. In particular, we study the minimax MSE of missing mass estimation of $\mathcal{U}$, namely $R^*(\mathcal{U})$. If $R^*(\mathcal{U})=R(M^{\mathtt{GT}},\mathcal{U})$, it proves that $M^{\mathtt{GT}}$ is MSE-optimal! If not, then perhaps there is an estimator with smaller minimax MSE than $M^{\mathtt{GT}}$.

Theorem 2. *The Maximum-Likelihood estimator for missing mass for the uniform distribution*

$$R_u\left(M^{ML}\right)=\frac{\alpha_u}{n}+o\left(\frac{1}{n}\right),$$

where α_u is the solution to

$$\alpha_u=\max_{x\in[0,1]}\frac{xe^{-x}(1-e^{-x})^2}{1-e^{-x}-xe^{-x}}\approx 0.570\ldots. \qquad (3)$$

This shows that up to an $o(1/n)$ additive factor, $R^*(\mathcal{U})=\frac{0.570\ldots}{n}<\frac{0.608\ldots}{n}=R(M^{\mathtt{GT}},\mathcal{U})$. This shows that there is a better estimator than $M^{\mathtt{GT}}$ for $\mathcal{U}$. Since $\mathcal{U}$ is the class where $M^{\mathtt{GT}}$ achieves its minimax rate, we believe that $M^{\mathtt{GT}}$ does not achieve R^*.

Since $R^*\ge R^*(\mathcal{U})$, we obtain

$$\frac{0.570\ldots}{n}+o\left(\frac{1}{n}\right)\le R^*(\Delta)\le\frac{0.608\ldots}{n}+o\left(\frac{1}{n}\right).$$

III. Preliminaries and the Estimators

For a sequence X_1^n, and $t \geq 1$, let $\Phi_t(X_1^n)$ be the number of symbols that appear t times in X_1^n. The *profile* of X_1^n, is $\Phi(X_1^n) \stackrel{\text{def}}{=} (\Phi_1(X_1^n), \Phi_2(X_1^n), \ldots)$. When the sequence is clear from the context, we denote $\Phi(X_1^n)$, and $\Phi_t(X_1^n)$ by Φ, and Φ_t respectively. For example, when $X_1^n = \texttt{abracadabra}$, $\Phi_1 = \Phi_2 = 2, \Phi_3 = \Phi_4 = 0, \Phi_5 = 1, \Phi_6 = \Phi_7 = \ldots = 0$, and $\Phi = (2, 2, 0, 0, 1, 0, \ldots)$.

Definition 1. The Good-Turing estimator is

$$M^{\texttt{GT}}(X_1^n) \stackrel{\text{def}}{=} \frac{\Phi_1(X_1^n)}{n}. \tag{4}$$

$M^{\texttt{GT}}$ estimates the missing mass as the fraction of symbols in X_1^n that appear once. $M^{\texttt{GT}}$ is only a function of Φ_1, and gives the same missing mass estimate for any two sequences with the same profile. Such estimators are called symmetric. Optimal symmetric estimators exist for missing mass estimation.

Definition 2. A missing mass estimator M is *symmetric* if for any two sequences X_1^n, and Y_1^n with $\Phi(X_1^n) = \Phi(Y_1^n)$, the distributions of $M(X_1^n)$, and $M(Y_1^n)$ are identical.

Theorem 3. *There is a symmetric estimator that achieves $R^*(\Delta)$.*

The proof is similar in spirit to that in [18], and is omitted due to lack of space.

While profiles form a sufficient statistic for missing mass, and various other problems [19], their distributional form is unwieldy. In particular, it is not known how to compute, or even approximate the profile probabilities (See [20], [21] for some heuristics on this computation). However, in the special case when the underlying distribution is uniform, profile probabilities take a nice form and are easy to compute. Recall that $\mathcal{U}$ is the set of all distributions that are uniform over a subset of $\mathbb{N}$.

Let $m(X_1^n)$ be the number of distinct symbols in X_1^n. The next result shows that for $\mathcal{U}$, $m(X_1^n)$ is a sufficient statistic for the missing mass. This is crucial for proving Theorem 2, since we can restrict our attention to estimators that are just a function of m, for both the upper and the lower bound.

Theorem 4. *$m(X_1^n)$ is a sufficient statistic to achieve $R^*(\mathcal{U})$.*

Proof. Similar to Theorem 3, it can be used to show that for $\mathcal{U}$ profiles are a sufficient statistic. The probability of a profile Φ under the uniform distribution $u^{(k)}$ is

$$\Pr(\Phi) = \frac{k^{\underline{m}}}{k^n} \frac{n!}{\prod_{t\geq 1}((t!)^{\Phi_t} \cdot \Phi_t!)}, \tag{5}$$

where $k^{\underline{m}} = \prod_{i=0}^{m}(k-i)$.
(5) follows from a counting argument similar to Lemma 3 in [22]. Therefore, the distribution over profiles with the same m is independent of the value of k. In other words, conditioned on the value of m, and n the distribution over profiles is independent of k. Hence knowing the exact profile won't give us additional information about k than knowing m. Combining with Theorem 3 gives us the result. □

In Section V we provide a missing mass estimator for $\mathcal{U}$ based on the profile maximum likelihood estimation method. This estimator is a function of m, and achieves the bound in Theorem 2 and strictly outperforms $M^{\texttt{GT}}$ over $\mathcal{U}$.

To provide the lower bound, by Yao's minimax principle, we consider a prior $\mathbb{P}$ over $\mathcal{U}$. $\mathbb{P}$ induces a distribution over $m(X_1^n)$. Given m, and $\mathbb{P}$, MMSE theory states that the MSE-optimal estimator is given by

$$\mathbb{E}[M_0(X_1^n, p)|m].$$

In Section VI, we show that for a properly chosen prior, the MMSE optimal estimator is identical to the ML estimator (ignoring the smaller order terms), proving the lower bound.

IV. Performance of Good-Turing

In this section, we prove Theorem 1. [1] characterized the MSE of GT estimator for any p as:

$$R(M^{\texttt{GT}}, p) = \frac{1}{n}\mathbb{E}\left[\frac{2\Phi_2}{n} + \frac{\Phi_1}{n}\left(1 - \frac{\Phi_1}{n}\right)\right] + o\left(\frac{1}{n}\right). \tag{6}$$

They proved (2) by obtaining upper and lower bounds on the largest possible value of (6) over discrete distributions p. We show that their lower bound is tight, by showing that for any discrete distribution $R(M^{\texttt{GT}}, p) \leq \alpha_{\texttt{GT}}/n + o(1/n)$.

By concavity of $x(1-x)$, $\mathbb{E}\left[\frac{\Phi_1}{n}\left(1 - \frac{\Phi_1}{n}\right)\right] \leq \frac{\mathbb{E}[\Phi_1]}{n}\left(1 - \frac{\mathbb{E}[\Phi_1]}{n}\right)$ and linearity of summations in (6), it will suffice to show that

$$\frac{2\mathbb{E}[\Phi_2]}{n} + \frac{\mathbb{E}[\Phi_1]}{n}\left(1 - \frac{\mathbb{E}[\Phi_1]}{n}\right) \leq \alpha_{\texttt{GT}} + o(1). \tag{7}$$

Note that $\mathbb{E}[\Phi_t] = \sum_{i=1}^{\infty}\binom{n}{t}p_i^t(1-p_i)^{n-t}$ We then show that replacing $(1-p_i)^{n-t}$ with e^{-np_i} for $t = 1, 2$ can affect only the $o(1/n)$ term. Using this replacement, we plug in $\mathbb{E}[\Phi_t]$'s in (7).

From these results, it will suffice to prove that for any $p_1, \ldots,$ such that $p_i \geq 0, \sum_i p_i = 1$

$$n\sum_{i=1}^{\infty} p_i^2 e^{-np_i} + \left(\sum_{i=1}^{\infty} p_i e^{-np_i}\right)\left(1 - \sum_{i=1}^{\infty} p_i e^{-np_i}\right) \leq \alpha_{\mathtt{GT}}.$$

The proof involves invoking Lagrange multipliers, and showing that the maximum value of the expression is achieved for a uniform distribution. For uniform distributions we end up with $\alpha_{\mathtt{GT}}$.

V. Upper Bound on $R^*(\mathcal{U})$

In this section, we provide the upper bound of Theorem 2, the minimax MSE of $\mathcal{U}$. Suppose the support of the uniform distribution is k, then the missing mass is exactly $1 - m/k$. However, since we do not know k, we study the following two step procedure. (a) Obtain an estimate $\hat{k}$ of the support size of p from X_1^n, (b) Output $1 - m/\hat{k}$.

Our estimator of support size is the uniform distribution that assigns the highest probability to the profile, also called as the Profile Maximum Likelihood distribution [23]. For the uniform distributions, this is equivalent to finding the value of k that maximizes (5). Note that this expression depends on k only via $k^{\underline{m}}/k^n$. Therefore, $k_{m,n} \stackrel{\text{def}}{=} \arg\max_k \frac{k^{\underline{m}}}{k^n}$ is the support of the uniform distribution that maximizes the probability of $\Phi(X_1^n)$. Our ML missing mass estimator, which we show achieves the upper bound in Theorem 2, is the following.

$$M^{\mathtt{ML}} = 1 - \frac{m}{k_{m,n}}.$$

Similar to the arguments in [24], it is easy to see that the expression in (V) increases and then decreases in k. Using this, [24] showed that $k_{m,n}$ is one of the two integers adjacent to the solution to

$$\frac{m}{k} = \left(1 - \left(1 - \frac{1}{k}\right)^n\right).$$

For ease of analysis, instead of using $k_{m,n}$, which is always an integer, we use k_e that is a solution to

$$k_e(1 - e^{-n/k_e}) = m. \tag{8}$$

Note that here we have used exponential approximation similar to the previous section.

Suppose k_0 is the true underlying support size. Then, the MSE of our estimator is

$$\mathbb{E}\left[\left(\frac{m}{k_e} - \frac{m}{k_0}\right)^2\right], \tag{9}$$

where the only randomness is in m, since k_e is a deterministic function of m, n. To bound (9), consider two cases:

1) **Case 1.** $5k_0 < n/\log n$. In this case, by a coupon-collector argument, we observe all symbols with high probability.
2) **Case 2.** $5k_0 > \frac{n}{\log n}$. This is the more involved case. Let $m(X_1^n)$ be the random variable that denotes the number of distinct symbols that appear in X_1^n. Note that we actually obtain an instantiation of $m(X_1^n)$ as our m.

Lemma 1. *When* $5k_0 > \frac{n}{\log n}$, *the variance of* $m(X_1^n)$ *is* $k_0\left(e^{-\frac{n}{k_0}} - e^{-\frac{2n}{k_0}} - \frac{n}{k_0}e^{-\frac{2n}{k_0}}\right) + O(1)$.

Let (m_0, k_0) be a solution to (8), where k_0 is the true underlying support size. For a fixed n, we will characterize the solutions around (m_0, k_0).

Lemma 2. *Suppose* (m_0, k_0) *is the solution to* (8), *then for* c *with* $|c| = O(\log n)$, *and* $m' = m_0 + c\sqrt{m_0}$, *then the solution* (m', k') *to* (8) *satisfies*

$$k' = k_0 + c\frac{\sqrt{1 - e^{-\frac{n}{k_0}}}}{1 - e^{-\frac{n}{k_0}} - \frac{n}{k_0}e^{-\frac{n}{k_0}}}\sqrt{k_0} + O(c^2).$$

Applying McDiarmid's inequality, we show that

$$\Pr\left(|m(X_1^n) - m_0| \geq 10\sqrt{k_0}\log n\right) \leq \frac{1}{n^5}. \tag{10}$$

Let $\mathcal{E}$ denote the event that $|m(X_1^n) - m_0| \leq 10\sqrt{k_0}\log n$. Denote the random variable $m(X_1^n)$ by m', and (m', k') be the solution to (8). Therefore, it suffices to bound $\mathbb{E}\left[\left(\frac{m'}{k'} - \frac{m'}{k}\right)^2 \,\middle|\, \mathcal{E}\right]$, since

$$\mathbb{E}\left[\left(\frac{m'}{k'} - \frac{m'}{k}\right)^2\right] \leq \mathbb{E}\left[\left(\frac{m'}{k'} - \frac{m'}{k}\right)^2 \,\middle|\, \mathcal{E}\right] + o(\frac{1}{n}).$$

We prove the following result.

Lemma 3.

$$\mathbb{E}\left[\left(\frac{m'}{k'} - \frac{m'}{k_0}\right)^2 \,\middle|\, \mathcal{E}\right] = \frac{H^6}{D^2}\frac{m_0}{k_0}\mathrm{Var}\left(m(X_1^n)\right) + o\left(\frac{1}{n}\right)$$

where $D = 1 - e^{-n/k_0} - \frac{n}{k_0}e^{-n/k_0}$, *and* $H = \sqrt{1 - e^{-n/k_0}}$.

Assuming Lemma 3 and plugging in the variance expression for $m(X_1^n)$,

$$\frac{H^6}{D^2}\frac{m_0}{k_0}\mathrm{Var}\,(m) = \frac{n}{k_0}\frac{e^{-n/k_0}(1 - e^{-n/k_0})^2}{1 - e^{-n/k_0} - \frac{n}{k_0}e^{-n/k_0}} \cdot \frac{1}{n}. \tag{11}$$

In (11) denote $\frac{n}{k_0}$ by x, then

$$\frac{n}{k_0}\frac{e^{-n/k_0}(1 - e^{-n/k_0})^2}{1 - e^{-n/k_0} - \frac{n}{k_0}e^{-n/k_0}} = \frac{xe^{-x}(1 - e^{-x})^2}{1 - e^{-x} - xe^{-x}}.$$

Let $f(x) = \frac{xe^{-x}(1-e^{-x})^2}{1-e^{-x}-xe^{-x}}$. It has maximum $0.570\ldots$ when $x = 0.801\ldots$.

VI. Lower bound on $R^*(\mathcal{U})$

We invoke key results from the theory of Minimum Mean Squared Error (MMSE) Estimation.

In the first step, we construct a prior, $\mathbb{P}$ over the class $\mathcal{U}$, which will be simply a distribution over the support size k. Since the maximum is always at least the expectation, for any estimator M,

$$\max_p R^*(M,p) \geq \mathbb{E}_{p\sim\mathbb{P}}\left[R(M,p)\right].$$

Suppose that M_u is the optimal missing mass estimator for the class $\mathcal{U}$ that achieves $R^*(\mathcal{U})$. Then plugging in the previous equation,

$$R^*(\mathcal{U}) \geq \mathbb{E}_{p\sim\mathbb{P}}R(M_u,p) \geq \inf_M \mathbb{E}_{p\sim\mathbb{P}}R(M,p).$$

We now invoke MMSE theory to obtain an optimal missing mass estimator given the knowledge of $\mathbb{P}$.

Lemma 4. *Suppose X is a random variable that we want to estimate given that another random variable Y takes on the value y. Then the estimator $\hat{X}$ that minimizes $\mathbb{E}\left[(X-\hat{X})^2\right]$ is $\mathbb{E}\left[X|Y=y\right]$.*

Let D be the random variable denoting the number of distinct symbols that appear in X_1^n. By Theorem 4 we know that D is a sufficient statistic for estimating the missing mass for uniform distributions. In the framework of MMSE estimation, the random variable X is m/k, and the random variable Y is D. Now conditioned on the value of $D=m$, the optimal estimator $M_{\mathbb{P}}(X_1^n)$ is equal to

$$M_{\mathbb{P}}(X_1^n) = \mathbb{E}_{\mathbb{P}}\left[1-\frac{m}{k}|D(X_1^n)=m\right].$$

After constructing the appropriate prior, the bulk of our argument goes into showing that $M_{\mathbb{P}}$ can be replaced by the ML estimator from the previous section while only introducing an error of order $o(\frac{1}{n})$. Once this reduction is done, we can simply use the bound from the previous section.

Acknowledgements

This work is supported by CRII-CIF-1657471, and a Cornell University startup.

References

[1] N. Rajaraman, A. Thangaraj, and A. T. Suresh, "Minimax risk for missing mass estimation," in *ISIT*, 2017, pp. 3025–3029.

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

[3] B. Harris, "Determining bounds on integrals with applications to cataloging problems," *The Annals of Mathematical Statistics*, pp. 521–548, 1959.

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

[5] E. L. Lehmann and G. Casella, *Theory of point estimation*. Springer Science & Business Media, 1998, vol. 31.

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

[7] A. Orlitsky, N. P. Santhanam, and J. Zhang, "Always good turing: Asymptotically optimal probability estimation," in *FOCS*, 2003.

[8] A. B. Wagner, P. Viswanath, and S. R. Kulkarni, "Strong consistency of the good-turing estimator," in *ISIT*, 2006.

[9] E. Drukh and Y. Mansour, "Concentration bounds for unigrams language model," in *COLT*, 2004.

[10] J. Acharya, A. Jafarpour, A. Orlitsky, and A. T. Suresh, "Optimal probability estimation with applications to prediction and classification," in *COLT*, 2013.

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

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

[13] A. Ben-Hamou, S. Boucheron, and M. I. Ohannessian, "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.

[14] B. Y. S. Khanloo and G. Haffari, "Novel bernstein-like concentration inequalities for the missing mass," *arXiv preprint arXiv:1503.02768*, 2015.

[15] C.-H. Zhang and Z. Zhang, "Asymptotic normality of a nonparametric estimator of sample coverage," *The Annals of Statistics*, pp. 2582–2595, 2009.

[16] M. Grabchak and Z. Zhang, "Asymptotic properties of turing's formula in relative error," *Machine Learning*, vol. 106, no. 11, pp. 1771–1785, 2017.

[17] G. Decrouez, M. Grabchak, and Q. Paris, "Finite sample properties of the mean occupancy counts and probabilities," *arXiv preprint arXiv:1601.06537*, 2016.

[18] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White, "Testing that distributions are close," in *FOCS*, 2000, pp. 259–269.

[19] J. Acharya, H. Das, A. Orlitsky, and A. T. Suresh, "A unified maximum likelihood approach for estimating symmetric properties of discrete distributions," in *ICML*, 2017, pp. 11–21.

[20] P. O. Vontobel, "The bethe and sinkhorn approximations of the pattern maximum likelihood estimate and their connections to the valiant-valiant estimate," in *ITA*, 2014, pp. 1–10.

[21] S. Pan, "On the theory and application of pattern maximum likelihood," Ph.D. dissertation, UC San Diego, 2012.

[22] A. Orlitsky, N. P. Santhanam, and J. Zhang, "Universal compression of memoryless sources over unknown alphabets," *IEEE Transactions on Information Theory*, vol. 50, no. 7, pp. 1469– 1481, July 2004.

[23] A. Orlitsky, N. P. Santhanam, K. Viswanathan, and J. Zhang, "On modeling profiles instead of values," in *UAI*, 2004.

[24] A. Orlitsky, N. Santhanam, K. Viswanathan, and J. Zhang, "On estimating the probability multiset," Manuscript, 2011. [Online]. Available: http://www-ee.eng.hawaii.edu/~prasadsn/skelnew59.pdf