Journal of Machine Learning Research 23 (2022) 1-37 Submitted 12/21; Revised 7/22; Published 9/22

Convergence Guarantees for the Good-Turing Estimator

Amichai Painsky AMICHAIP@TAUEX.TAU.AC.IL
Department of Industrial Engineering
Tel Aviv University
Tel Aviv, Israel

Editor: Genevera Allen

Abstract

Consider a finite sample from an unknown distribution over a countable alphabet. The occupancy probability (OP) refers to the total probability of symbols that appear exactly k times in the sample. Estimating the OP is a basic problem in large alphabet modeling, with a variety of applications in machine learning, statistics and information theory. The Good-Turing (GT) framework is perhaps the most popular OP estimation scheme. Classical results show that the GT estimator converges to the OP, for every k independently. In this work we introduce new exact convergence guarantees for the GT estimator, based on worst-case mean squared error analysis. Our scheme improves upon currently known results. Further, we introduce a novel simultaneous convergence rate, for any desired set of occupancy probabilities. This allows us to quantify the unified performance of OP estimators, and introduce a novel estimation framework with favorable convergence guarantees.

Keywords: Good-Turing Estimator, Occupancy Probability, Natural Language Modeling, Missing Mass

1. Introduction

Let p be a probability distribution over an alphabet $\mathcal{X}$ of size m. Let X^n be a sample of n independent observations from p. Large alphabet modeling considers the setup where m is comparable, or even larger, than n. In this regime, more symbols are likely to appear the same number of times. The *occupancy probability* (OP) is defined as the total probability of symbols that appear exactly k times in the sample. Estimating the OP from a given sample is a fundamental problem in statistics, machine learning and related fields. It is mostly relevant in the context of large alphabet modeling. For example, a reasonable estimator shall assign the same probability to symbols that appear the same number of times. Therefore, estimating the total mass of those symbols is of high interest in modeling p. OP estimation is extensively studied in a variety of disciplines, including language modeling (Chen and Goodman, 1999; Drukh and Mansour, 2005), authorship attribution (Efron and Thisted, 1976; Thisted and Efron, 1987; Zhang and Huang, 2007), ecology (Good and Toulmin, 1956; Chao, 1981), genomics (Mao and Lindsay, 2002), information theory (Orlitsky et al., 2004c) and computer science (Zhang, 2005).

A naive scheme for OP estimation is based on the maximum likelihood (ML) principle. That is, the k^{th} OP is assigned a probability proportional to number of symbols with k appearances in the sample. While this approach works well for frequent symbols, it dramatically fails in the presence of rare symbols. For example, a ML estimator would assign zero probability to symbols that do not appear in the samples. A variety of alternatives methods have been suggested over the years. Perhaps the first major contribution is due to Laplace, who addressed the problem by adding a single count to all the symbols in the alphabet (including unobserved symbols). Then, the OP estimator is simply the ML estimator of the modified sample. This scheme is known as the *rule of succession.* The work of Laplace was studied and generalized by many researchers and practitioners (see Section 2 for a detailed discussion). Many years after Laplace, a significant milestone was established in the work of Good and Turing (Good, 1953). The Good-Turing (GT) framework suggests that the k^{th} OP is assigned a probability proportional to the number of symbols with $k+1$ appearance in the sample. This (somewhat unintuitive) approach introduced a major improvement compared to known estimators at the time. Its favorable performance and practical appeal have motivated a large body of research over the years, as discussed in detail in Section 2. To this day, Good-Turing estimators are the most commonly used methods in most practical setups (Orlitsky and Suresh, 2015).

In this work we study the convergence rate of the GT estimator to the occupancy probabilities. We present a new perspective to the problem and introduce improved performance guarantees. Our analysis focuses on the worst-case mean square error (MSE). Specifically, we bound from above the MSE of the GT estimator for any possible probability distribution over $\mathcal{X}$, and apply Markov's inequality to obtain the desired convergence rate. We show that with high probability, the GT estimator convergence to the k^{th} OP in a rate of $O(1/\sqrt{n})$, for a fixed k and unbounded m. Further, we study the case where m is bounded, and introduce improved convergence guarantees for this setup. Next, we introduce convergence guarantees for the case where k grows with n. Here, we show that the convergence rate is $O(\sqrt[4]{k}/\sqrt{n})$. Our bounds improve upon currently known results. Importantly, we also provide the exact convergence coefficients. Further, we study the convergence rate of the ML estimator and compare it to GT for different k and n. Consequently, we derive a hybrid estimator, based on GT (for small k's) and ML (for large k's). Our proposed hybrid estimator attains a convergence rate of $O(n^{-2/\sqrt{5}})$ for every k. This result again improves upon currently known hybrid estimators. Finally, we utilize our MSE framework and introduce a novel convergence rate for all k's, simultaneously. This allows us to characterize the performance of the studied estimators in modeling the complete (unknown) distribution.

This manuscript is an extended version of our preliminary work, presented in (Painsky, 2021). In (Painsky, 2021) we introduced initial results on the refined convergence rates of the GT estimator, for some of the setups mentioned above. We also presented a basic scheme for a simultaneous convergence rate. Here we significantly extend the scope. This includes the following contributions:

- A detailed derivation of the GT convergence rate for fixed and large k.
- An analysis of the GT convergence rate for a bounded alphabet size m.
- A study of a hybrid estimator, and its favorable convergence guarantees.
- Improved simultaneous convergence rate of an OP estimator, based on GT and ML estimators.

- An extensive empirical study which demonstrates the performance of the proposed estimator, compared to currently known schemes.

The rest of the manuscript is organized as follows. In Section 2 we review previous results on OP estimation and the Good-Turing framework. In Section 3 we formally state our problem. Throughout Section 4 we derive convergence rates for the GT estimator in three different setups. First, we focus on the case where k is fixed and m is unbounded (Subsection 4.1). Then, we study the setup where m is bounded and k is still fixed (Subsection 4.2). Next, we consider the case where k varies and grows with n (Subsection 4.3). In Section 5 we study the ML estimator, and compare it to GT in different setups. We introduce our proposed hybrid estimator in Section 6. Next, we present simultaneous convergence guarantees in Section 7. Finally, in Section 8 we compare our suggested framework with currently known estimators in a series of synthetic and real-world experiments. We conclude with a discussion in Section 9.

2. Previous Work

Let $X^n \triangleq \{X_1, \dots, X_n\}$ be n independent samples from an unknown distribution p over a countable alphabet $\mathcal{X}$ of size m. We assume that the alphabet size is unknown and may even be unbounded. Let $N_x(X^n)$ be the number of appearances of the symbol $x \in \mathcal{X}$ in X^n. Let

$$\Phi_k(X^n) = \sum_{u \in \mathcal{X}} \mathbb{1}(N_u(X^n) = k) \tag{1}$$

be the number of symbols that appear k times in X^n, for $0 \leq k \leq n$, where $\mathbb{1}(\cdot)$ is the indicator function. For example, for $\mathcal{X} = \{a, b, c, n\}$ and $X^n = \{b, a, n, a, n, a\}$, we have $\Phi_k = 1$ for $k = 0, , , , 3$. We denote the collection $\{\Phi_k(X^n)\}_{k=0}^n$ as the *frequency of frequencies* (FoF's). Given X^n, the occupancy probability (OP) is defined as the total probability of symbols that appear k times in the sample,

$$M_k(X^n) \triangleq \sum_{u \in \mathcal{X}} p(u) \mathbb{1}(N_u(X^n) = k) \tag{2}$$

for $0 \leq k \leq n$. The OP is also referred to as *urn scheme* (Decrouez et al., 2018), *k-hitting mass* (Drukh and Mansour, 2005) and *k-th combined probability mass* (Chandra et al., 2019). It is important to emphasize that in most OP studies, the alphabet size is assumed to be unknown, while in others it is considered known. We provide examples of two setups later in this section.

OP estimation has been extensively studied over the years. The first contribution to the problem is most likely due to Laplace (1825). In his work, Laplace suggested adding a single count to every symbol in the alphabet. Then, the estimate of the OP's is simply the empirical distribution of the modified sample. The Laplace estimator was later generalized to a family of *add-constant* estimators. An add-c estimator assigns to a symbol that appeared t times a probability proportional to $t + c$, where c is a pre-defined constant. Specifically, the add-c estimator is defined as

$$\hat{M}_k^{AC}(X^n) = \frac{(k + c)\Phi_k(X^n)}{n + cm}, \tag{3}$$

where $c = 1$ corresponds to the Laplace estimator. Add-constant estimators hold many desirable properties, mostly in terms of their simplicity and interpretability (for example, (Krichevsky and Trofimov, 1981)). Unfortunately, when the alphabet size m is large compared to the sample size n, add-constant estimators perform quite poorly (Orlitsky et al., 2003). Additional caveats of add-c estimators were discussed by Gale and Church (1994)

Many years after Laplace, I.J. Good and A.M. Turing achieved a significant milestone in OP estimation while trying to break the Enigma Cipher during World War II (Orlitsky et al., 2003). The idea behind their work is surprisingly simple. Instead of using $\Phi_k(X^n)$ as a statistic for the k^{th} OP, they suggest using $\Phi_{k+1}(X^n)$, the number of symbols with a $k+1$ appearances in the sample. Specifically, the Good-Turing estimator satisfies

$$\hat{M}_k^{GT}(X^n) = \frac{(k+1)\Phi_{k+1}(X^n)}{n}. \tag{4}$$

It is important to emphasize that while add-constant estimators depend on the alphabet size m, the Good-Turing estimator does not assume any knowledge of m, which makes it more robust. Furthermore, although OP estimation is properly defined for both known or unknown alphabet size, most methods focus on the latter.

One of the first analytical contributions to the GT framework is due to McAllester and Schapire (2000), who studied the convergence properties of the GT estimator. In their work, they showed that with high probability,

$$|\hat{M}_k^{GT}(X^n) - M_k(X^n)| = O\left(\frac{\log n}{\sqrt{n}}\right) \tag{5}$$

for small k, while for larger k,

$$|\hat{M}_k^{GT}(X^n) - M_k(X^n)| = O\left(\frac{k}{\sqrt{n}}\right). \tag{6}$$

This result demonstrated, perhaps for the first time, the reason GT performs so well in OP estimation and large alphabet modeling. Notice that (5) and (6) may also be viewed as confidence intervals for the OP's. The work of McAllester and Schapire inspired many studies of the GT framework and the properties of occupancy probabilities. One notable example is due to Drukh and Mansour (2005), who improved (6), showing that with high probability,

$$|\hat{M}_k^{GT}(X^n) - M_k(X^n)| = O\left(\frac{\sqrt[4]{k}}{\sqrt{n}} + \frac{k}{n}\right). \tag{7}$$

In addition, Drukh and Mansour (2005) introduced a lower bound for OP estimation, showing that any estimator based on an independent sample satisfies

$$|\hat{M}_k(X^n) - M_k(X^n)| = \Omega\left(\frac{\sqrt[4]{k}}{\sqrt{n}}\right). \tag{8}$$

The Good-Turing framework was later considered by Gale and Sampson (1995), who introduced the popular *smooth Good-Turing.* This framework suggests a smoothing mechanism to the GT estimator such that

$$\hat{M}_k^{SGT}(X^n) = \frac{(k+1)S\left(\Phi_{k+1}(X^n)\right)}{n}. \tag{9}$$

where $S\left(\Phi_k(X^n)\right)$ is a *smoothed* version of $\Phi_k(X^n)$. Specifically, smoothing is obtained by applying simple linear regression between $\log \Phi_k(X^n)$ and $\log k$. Gale and Sampson (1995) justified the proposed estimator by noticing the erratic behavior of $\Phi_k(X^n)$ as k increases. Further, Gale and Sampson (1995) observed that $\Phi_k(X^n)$ is typically zero for many larger values of k. Therefore, an additional modification is applied to overcome this caveat (Gale and Sampson, 1995; Church and Gale, 1991). The smooth GT estimator is a widely used scheme for OP and large alphabet estimation. Unfortunately, it is difficult to analyze and is mostly explained by heuristic justifications and numerical experimentation (Nadas, 1991).

Additional properties of the occupancy probabilities and Good-Turing related estimators were subject to numerous recent studies. We list the more relevant results for our work. In an early work, Karlin (1967) studied the asymptotic behavior of the frequency of frequencies (1), which he denoted as the *occupancy counts*. This work inspired the study of different attributes of (1) and (2). Decrouez et al. (2018) studied the statistical properties of the OP's, focusing on the expected values of $\Phi_k(X^n)$ and $M_k(X^n)$. Ben-Hamou et al. (2017) derived Bernstein-type concentration inequalities for $\Phi_k(X^n)$ and used them to derive confidence intervals for $M_0(X^n)$. Chandra et al. (2019) studied tail bounds and confidence intervals for $M_0(X^n)$, and derived a convergence rate for the GT estimator under a Poisson sampling regime (that is, where the number of samples n is drawn from a Poisson distribution). Ohannessian and Dahleh (2012) studied the multiplicative consistency of the GT estimator, $M_k^{GT}(X^n)/M_k(X^n)$. They showed that the GT estimator is not universally consistent in this sense, but only enjoys consistency guarantees under regularly varying heavy tails distributions. The bias of the GT estimator was also considered in several key contributions. In an earlier work, Robbins (1955) proposed a variant of the GT estimator and studied its mean and variance. Later, Juang and Lo (1994) showed that the bias of GT is of order of $1/n$, and proposed an alternative estimator that reduces the bias to $O(1/n^2)$. More recently, McAllester and Schapire (2000) showed that the bias of GT is in fact bounded from above by $(k+1)/(n-k)$.

An important special case of OP estimation is the missing mass problem, which refers to the case where $k = 0$. In words, the missing mass is the total probability of symbols that do not appear in the sample. Here too, the Good-Turing estimator is perhaps the most popular estimation scheme. A large body of work focuses on the Good-Turing framework in the context of missing mass estimation. For example, Rajaraman et al. (2017) and Acharya et al. (2018) studied the mean square error of the GT framework for the missing mass problem. Alternatively, Mossel and Ohannessian (2019); Ohannessian and Dahleh (2012); Grabchak and Zhang (2017); Battiston et al. (2020) focused on missing mass estimation under a multiplicative loss. Gao et al. (2013) studied the missing mass asymptotic normality and large deviations. Berend and Kontorovich (2013) studied the concentration of the missing mass. Skorski (2021) derived an upper bound on the variance of the missing mass, that holds for every sample and alphabet size. Budianu and Tong (2004) studied the large deviation of the GT estimator, as they considered an estimate for the number of operating sensors in a wireless sensor network. Cohen et al. (2022) studied the missing mass problem from Lehmann unbiasedness perspective, and derived a Cramér-Rao type lower bound for the MSE of the missing mass. Lijoi et al. (2007), Favaro et al. (2012) and Favaro et al. (2016) focused on the missing mass problem from a Baysian perspective. Recently, Painsky (2022) introduced a novel missing mass estimation scheme which generalizes GT and considers additional FoF's (specifically, both $\Phi_1(X^n)$ and $\Phi_2(X^n)$). The Generalized GT estimator holds a closed form expression, and was shown to attain improved performance guarantees, both in terms of MSE and convergence rates.

It is important to mention additional large alphabet modeling problems, which are closely related to OP and missing mass estimation. Given a sample X^n, Fisher et al. (1943) considered the problem of estimating $U_b(X^n)$, the number of unseen symbols that would be observed, if b additional samples were collected. It can be shown that this problem is equivalent to missing mass estimation for $b = 1$, and alphabet size estimation for $b \to \infty$. (Orlitsky et al., 2016). Inspired by the Good-Turing framework, Good and Toulmin (1956) introduced their approach to the problem, which utilizes a linear combination of the FoF's to estimate $U_b(X^n)$. Their work was studied and generalized by others, including Efron and Thisted (1976); Orlitsky et al. (2016).

Modeling and estimation of large alphabet problems is not limited to the Good-Turing framework. In a recent line of work, Orlitsky et al. (2004b) introduced the concept of *profile maximum likelihood* (PML). The profile of a sample X^n is simply the collection of FoF's, $\Phi(X^n) = \Phi_0(x^n), \Phi_1(X^n), \ldots, \Phi_n(X^n)$, and the PML framework seeks a probability distribution p which maximizes the likelihood of the observed FoF's. Specifically,

$$\hat{p}_{pml} = \arg\max_p \sum_{y^n : \Phi(y^n) = \Phi(X^n)} p(y^n). \tag{10}$$

The profile maximum likelihood principle was extensively studied and showed to posses a number of useful attributes, such as existence over finite discrete domains, majorization by empirical distributions, consistency for distribution estimation under both sorted and unsorted l_1 distances, and competitiveness to other profile-based estimators (see Hao and Orlitsky (2019) and references therein). It was further applied to a variety of large alphabet estimation problems including OP estimation, probability estimation, alphabet size estimation and others (Hao and Orlitsky, 2019). It is important to emphasize that PML is also related to the Good-Turing estimator, as discussed in (Orlitsky et al., 2004c). Unfortunately, the PML estimator is computationally challenging to derive and implement (Orlitsky et al., 2004b,a). Several efficient approximations have been recently suggested (Pavlichin et al., 2019; Anari et al., 2020), mostly focusing on reducing the exponential number of elements considered in the summation above (10).

3. Problem Statement

Given a distribution p, the l_2^2 risk of an OP estimator is defined as

$$R_n(\hat{M}_k, p) \triangleq \mathbb{E}_{X^n \sim p} \left(\hat{M}_k(X^n) - M_k(X^n) \right)^2. \tag{11}$$

Throughout this work we refer to (11) as the risk or the MSE of the estimator $\hat{M}_k(X^n)$, interchangeably. Let $\mathcal{P}$ be a given set of distributions. Then, the worst-case risk over $\mathcal{P}$ is

$$R_n(\hat{M}_k,\mathcal{P}) \triangleq \sup_{p\in\mathcal{P}} R_n(\hat{M}_k,p). \tag{12}$$

Let $\mathcal{K}$ be a collection of OP's indices, $\mathcal{K}\subseteq\{0,\dots,n\}$. Define the additive risk over $\mathcal{K}$ as

$$R_n\big(\hat{M}_k,p,\mathcal{K}\big) \triangleq \sum_{k\in\mathcal{K}} R_n(\hat{M}_k,p), \tag{13}$$

and the worst-case additive risk over $\mathcal{P}$ is defined as

$$R_n(\hat{M}_k,\mathcal{P},\mathcal{K}) \triangleq \sup_{p\in\mathcal{P}} R_n\big(\hat{M}_k,p,\mathcal{K}\big). \tag{14}$$

In this work we focus on two sets of probability distributions $\mathcal{P}$. Let Δ_m be the set of all distributions of an alphabet size m, while Δ be the set of all distributions over any countable alphabet $\mathcal{X}$ (that is, $m\to\infty$). The worst-case framework seeks the most conservative performance guarantees, as it controls the maximal risk in the set. This makes it a robust approach that does not depend on additional modeling assumptions. For this reason, the worst-case framework is a highly popular approach in large alphabet modeling (Krichevsky and Trofimov, 1981; Orlitsky and Suresh, 2015; Rajaraman et al., 2017; Acharya et al., 2018). The worst-case risk may also be used to derive a convergence rate for $\hat{M}_k(X^n)$. For example, by Markov inequality we have that for every $p\in\mathcal{P}$,

$$\mathrm{P}\left(|\hat{M}_k(X^n)-M_k(X^n)|\geq a\right)\leq \frac{R_n\big(\hat{M}_k,p\big)}{a^2}\leq\frac{R_n(\hat{M}_k,\mathcal{P})}{a^2},$$

where the last inequality follows (12). Setting a confidence level of $\delta=R_n(\hat{M}_k,\mathcal{P})/a^2$, we have that with probability of at least $1-\delta$,

$$|\hat{M}_k(X^n)-M_k(X^n)|\leq\sqrt{\frac{R_n(\hat{M}_k,\mathcal{P})}{\delta}}, \tag{15}$$

for every $p\in\mathcal{P}$. Plugging $\mathcal{P}=\Delta_m$ (alternatively, $\mathcal{P}=\Delta$), we obtain a convergence rate that holds for every $p\in\Delta_m$ (alternatively, $p\in\Delta$). Notice that this is also the (marginal) confidence interval for $M_k(X^n)$, in a confidence level of δ over the sample X^n.

Next, consider a collection $\mathcal{K}=\{k_1,\dots,k_\kappa\}$ with a cardinality $|\mathcal{K}|=\kappa$. Define

$$W(\hat{M})=\big[|\hat{M}_{k_1}(X^n)-M_{k_1}(X^n)|,..,|\hat{M}_{k_\kappa}(X^n)-M_{k_\kappa}(X^n)|\big]^T.$$

Then, Markov inequality suggests

$$\mathrm{P}\left(W(\hat{M})^TW(\hat{M})\geq a\right)\leq\frac{1}{a}\sum_{k\in\mathcal{K}}R_n(\hat{M}_k,p)\leq\frac{1}{a}R_n(\hat{M}_k,\mathcal{P},\mathcal{K}).$$

Setting $\delta=R_n(\hat{M}_k,\mathcal{P},\mathcal{K})/a$ we obtain an c-sphere confidence region of a radius

$$r=\sqrt{\frac{R_n(\hat{M}_k,\mathcal{P},\mathcal{K})}{\delta}} \tag{16}$$

for the collection $\mathcal{K}$. As above, the obtained confidence region is also the simultaneous rate of convergence of a set of estimators, $\{\hat{M}_k\}_{k\in\mathcal{K}}$ to the corresponding collection of OP's. Alternatively, we may consider a more conservative approach,

$$\mathrm{P}\Big(\cup_{k\in\mathcal{K}}\Big\{|\hat{M}_k(X^n)-M_k(X^n)|\geq a\Big\}\Big)\leq$$
$$\sum_{k\in\mathcal{K}}\mathrm{P}\left(|\hat{M}_k(X^n)-M_k(X^n)|\geq a\right)\leq\frac{1}{a^2}\sum_{k\in\mathcal{K}}R_n(\hat{M}_k,p)=\frac{1}{a^2}R_n(\hat{M}_k,\mathcal{P},\mathcal{K}),$$

where the first inequality is due to the union bound. As above, setting $\delta=R_n(\hat{M}_k,\mathcal{P},\mathcal{K})/a^2$ we obtain a confidence interval which controls the probability that all OP estimators are simultaneously close to the desired (and unknown) OP's.

In the following sections we study (11) - (14) in a variety of setups and introduce new convergence guarantees for different estimators of the occupancy probabilities. We begin with an analysis of the worst-case risk (12) for the GT estimator, $\hat{M}_k^{GT}(X^n)$.

4. The MSE of the Good-Turing Estimator

The squared error of the GT estimator satisfies

$$\begin{aligned}\left(\hat{M}_k^{GT}(X^n)-M_k(X^n)\right)^2=&\left(\sum_{u\in\mathcal{X}}\frac{k+1}{n}\mathbb{1}(N_u(X^n)=k+1)-p(u)\mathbb{1}(N_u(X^n)=k)\right)\cdot\\&\left(\sum_{v\in\mathcal{X}}\frac{k+1}{n}\mathbb{1}(N_v(X^n)=k+1)-p(v)\mathbb{1}(N_v(X^n)=k)\right)=\\&\left(\frac{k+1}{n}\right)^2\sum_{u,v\in\mathcal{X}}\mathbb{1}(N_u(X^n)=k+1)\mathbb{1}(N_v(X^n)=k+1)-\\&\frac{2(k+1)}{n}\sum_{u,v\in\mathcal{X}}p(u)\mathbb{1}(N_u(X^n)=k)\mathbb{1}(N_v(X^n)=k+1)+\\&\sum_{u,v\in\mathcal{X}}p(u)p(v)\mathbb{1}(N_u(X^n)=k)\mathbb{1}(N_v(X^n)=k).\end{aligned}$$

Therefore, its risk is given by

$$\begin{aligned}R_n\big(\hat{M}_k^{GT},p\big)=\frac{1}{n^2}\sum_{u,v\in\mathcal{X}}&(k+1)^2P_n(k+1,k+1)+\\&2(k+1)np(u)P_n(k,k+1)+n^2p(u)p(v)P_n(k,k).\end{aligned}\tag{17}$$

where $P_n(i,j)=\mathbb{E}_{X^n\sim p}\left(\mathbb{1}(N_u(X^n)=i)\mathbb{1}(N_v(X^n)=j)\right)$ and

$$P_n(i,j)=\begin{cases}\binom{n}{i\;j}p^i(u)p^j(v)(1-p(u)-p(v))^{n-i-j} & u\neq v,\ i+j\leq n\\ \binom{n}{i}p^i(u)(1-p(u))^{n-i} & u=v,\ i=j\\ 0 & o.w.\end{cases}$$

Define $P(u,v) = p^{k+1}(u)p^{k+1}(v)(1-p(u)-p(v))^{n-2k-2}$. Plugging the above to (17) yields

$$R_n\big(\hat{M}_k^{GT}, p\big) = \tag{18}$$
$$\frac{1}{n^2}\binom{n}{k\;k}\sum_{u\neq v} P(u,v)\Big(2k(2k+1) - n - 4nk(p(u)+p(v)) + n^2(p(u)+p(v))^2\Big) +$$
$$\binom{n}{k}\sum_u p^{k+2}(u)(1-p(u))^{n-k} + \left(\frac{k+1}{n}\right)^2\binom{n}{k+1}\sum_u p^{k+1}(u)(1-p(u))^{n-k-1}$$

for $2k < n$. We discuss larger value of k later in Section 4.3. Let us now distinguish between different cases. We begin with the fixed k and unbounded m setup.

4.1 Fixed k Analysis

Rajaraman et al. (2017) studied the MSE of the GT estimator for $k = 0$, under an unbounded alphabet size m. Here, we extend their derivation for any fixed $k \geq 0$ and obtain the following theorem.

Theorem 1 *For a fixed $k \geq 0$ and an unbounded alphabet size m, the MSE of the Good-Turing estimator satisfies*

$$R_n\big(\hat{M}_k^{GT}, p\big) = \frac{-(k+1)^2}{n(n-2k)(n-2k-1)}\mathbb{E}\left(\Phi_{k+1}^2(X^n)\right) +$$
$$\left(\frac{k+1}{n}\right)^2\mathbb{E}\left(\Phi_{k+1}(X^n)\right) + \frac{(k+1)(k+2)}{(n-k)(n-k-1)}\mathbb{E}\left(\Phi_{k+2}(X^n)\right) + o\left(\frac{1}{n}\right). \tag{19}$$

The proof of Theorem 1 closely follows the analysis of Rajaraman et al. (2017), and is provided in Appendix A. We would now like to bound (19) from above, for every possible $p \in \Delta$. For this purpose, we introduce the following propositions.

Proposition 2 *Let p be a probability distribution over a countable alphabet $\mathcal{X}$. Let $\psi : [0,1]^2 \to \mathbb{R}$. Then,*

$$\sum_{u,v\in\mathcal{X},\, u\neq v} p(u)p(v)\psi(p(u),p(v)) \leq \max_{q_1,q_2\in\Delta_2}\psi(q_1,q_2).$$

where $\Delta_2 = \{q_1, q_2 \mid 0 \leq q_1, q_2 \leq 1, q_1 + q_2 \leq 1\}$.

Proposition 3 *Let p be a probability distribution over a countable alphabet $\mathcal{X}$. Let $\phi : [0,1] \to \mathbb{R}$. Then,*

$$\sum_{u\in\mathcal{X}} p(u)\phi(p(u)) \leq \max_{q\in[0,1]}\phi(q).$$

The proofs of Propositions 2 and 3 are provided in Appendix B. Going back to (19), we define

$$f_{n,k}(p) \triangleq \mathbb{E}\left(\Phi_k(X^n)\right) = \binom{n}{k}\sum_u p^k(u)(1-p(u))^{n-k}.$$

Further, notice that

$$-\mathbb{E}_{X^n\sim p}\left(\Phi^2_{k+1}(X^n)\right) \leq -\mathbb{E}^2_{X^n\sim p}\left(\Phi_{k+1}(X^n)\right).$$

Therefore,

$$\begin{aligned} R_n\left(\hat{M}_k^{GT},p\right) \leq & \frac{-(k+1)^2}{n(n-2k)(n-2k-1)} f^2_{n,k+1}(p) + \\ & \frac{(k+1)^2}{n^2} f_{n,k+1}(p) + \frac{(k+1)(k+2)}{(n-k)(n-k-1)} f_{n,k+2}(p) + o\left(\frac{1}{n}\right). \end{aligned} \tag{20}$$

Applying Proposition 3 to $f_{n,k+1}(p)$, we obtain

$$f_{n,k+1}(p) \leq \binom{n}{k+1} \max_{q\in[0,1]} q^k(1-q)^{n-k-1} = \binom{n-1}{k} \text{Bin}\left(k+1; n, \frac{k}{n-1}\right) \triangleq f^{max}_{n,k+1} \tag{21}$$

for $k \geq 1$, where $\text{Bin}(k;n,q)$ is a Binomial distribution with parameters n and q. Let us now study (20). We notice that the first two terms are quadratic (and concave) in $f_{n,k+1}(p)$. Therefore, their maximum is obtained either on the local optimum, $f^*_{n,k+1} = (n-2k)(n-2k-1)/2n$, or on the boundary of the set, $f^{max}_{n,k+1}$. Further, the third term in (20) may also be bounded from above by (21). This leads to the following theorem.

Theorem 4 *For a fixed $k \geq 1$ and an unbounded alphabet size m, the MSE of Good-Turing estimator satisfies the following:*

- *If $f^*_{n,k+1} \leq f^{max}_{n,k+1}$ then,*

$$\begin{aligned} R_n(\hat{M}_k^{GT},\Delta) \leq & \frac{(n-2k)(n-2k-1)(k+1)^2}{4n^3} + \\ & \frac{(k+2)(n-1)}{(n-k-1)^2} Bin\left(k+2; n, \frac{k+1}{n-1}\right) + o\left(\frac{1}{n}\right). \end{aligned} \tag{22}$$

- *otherwise,*

$$\begin{aligned} R_n(\hat{M}_k^{GT},\Delta) \leq & \frac{-(k+1)^2(n-1)^2}{k^2 n(n-2k)^2} Bin^2\left(k+1; n, \frac{k}{n-1}\right) + \\ & \left(\frac{k+1}{n}\right)^2 \binom{n-1}{k} Bin\left(k+1; n, \frac{k}{n-1}\right) + \\ & \frac{(k+2)(n-1)}{(n-k-1)^2} Bin\left(k+2; n, \frac{k+1}{n-1}\right) + o\left(\frac{1}{n}\right). \end{aligned} \tag{23}$$

It is well-known that a Binomial distribution $\text{Bin}(k;n,q)$ converges to a Poisson distribution $\text{Pois}(k;\lambda = nq)$ in cases where n grows and nq is fixed, or at least q tends to zero. Specifically, Prokhorov (1953) showed that $|\text{Bin}(k;n,q) - \text{Pois}(k;nq)| \leq cq$ for a fixed constant c.

We apply Prokhorov result to Theorem 4 and replace the Binomial terms with a Poisson distribution. Further we notice that as n grows, $f^*_{n,k+1} > f^{max}_{n,k+1}$. Therefore, in this setup

$$\begin{aligned} R_n(\hat{M}_k^{GT}, \Delta) \leq & \frac{-(k+1)^2(n-1)^2}{k^2 n(n-2k)^2} \frac{\left(\frac{kn}{n-1}\right)^{2k+2} \exp\left(\frac{-2kn}{n-1}\right)}{((k+1)!)^2} + \\ & \left(\frac{k+1}{n}\right)^2 \left(\frac{n-1}{k}\right) \frac{\left(\frac{kn}{n-1}\right)^{k+1} \exp\left(-\frac{kn}{n-1}\right)}{(k+1)!} + \\ & \frac{(k+2)(n-1)}{(n-k-1)^2} \frac{\left(\frac{(k+1)n}{n-1}\right)^{k+2} \exp\left(-\frac{(k+1)n}{n-1}\right)}{(k+2)!} + o\left(\frac{1}{n}\right). \end{aligned} \tag{24}$$

Finally, we apply Sterling bounds, $\sqrt{2\pi} k^{k+1/2} \exp(-k) \leq k! \leq k^{k+1/2} \exp(-k+1)$ and conclude with Theorem 5.

Theorem 5 *For a fixed $k \geq 1$ and $n >> k$, the MSE of the Good-Turing estimator satisfies*

$$R_n(\hat{M}_k^{GT}, \Delta) \leq \frac{g(k)}{n} + o\left(\frac{1}{n}\right) \tag{25}$$

where

$$g(k) = -\frac{1}{k+1}\left(\frac{k}{k+1}\right)^{2k} + \frac{e}{\sqrt{2\pi}}\left(\sqrt{k+1}\left(\frac{k}{k+1}\right)^{k} + \sqrt{k+2}\left(\frac{k+1}{k+2}\right)^{k+2}\right). \tag{26}$$

For example, $g(k) = 1.198, 1.455, 1.665, 1.849, 2.015, 2.169,\ 2.312$ for $k = 1, \dots, 7$, respectively. Applying (25) to (15), we obtain a convergence rate of order $O\left(1/\sqrt{n}\right)$ for a fixed k, which improves upon (5).

4.2 Bounded Alphabet Size Analysis

The analysis above focuses on the case where the alphabet size is unbounded. Let us now study the bounded alphabet size regime. In (20) we introduce an upper bound for $R_n(\hat{M}_k^{GT}, p)$ which is polynomial in $f_{n,k}(p)$. Then, we bound $f_{n,k}(p)$ from above, for every $p \in \Delta$ (see (21)). Here, we assume that the alphabet size m is bounded and $p \in \Delta_m$. This allows us to derive a tighter upper bound for $f_{n,k}(p)$.

We begin our analysis by noticing that $(1-t)^n \leq e^{-nt}$ for every $t \in \mathbb{R}$ and $n \in \mathbb{R}_+$ (Mitrinovic and Vasic, 1970). Therefore,

$$f_{n,k}(p) = \binom{n}{k} \sum_u p^k(u)(1-p(u))^{n-k} \leq \binom{n}{k} \sum_u p^k(u) \exp(-(n-k)p(u)).$$

Define

$$f^{max}_{n,k,m} = \max_{p \in \Delta_m} \binom{n}{k} \sum_u p^k(u) \exp(-(n-k)p(u)).$$

In (Painsky, 2022), we study the properties of $\sum_u p^r(u) \exp(-np(u))$ for a given alphabet size m. Specifically, Theorem 3 in (Painsky, 2022) shows that $\max_{p \in \Delta_m} \sum_u p^r(u) \exp(-np(u))$

depends on no more then four free parameters, for every n, r and m. This means we may the evaluate $f^{max}_{n,k,m}$ at a relatively low computational cost, even when the dimension of the problem increases. Applying the above to (20) we bound the different terms, similarly to Section 4.1. Specifically, the last term satisfies

$$\frac{(k+1)(k+2)}{(n-k)(n-k-1)} f_{n,k+2}(p) \leq \frac{(k+2)(n-1)}{(n-k)(n-k-1)} f^{max}_{n,k+2,m}, \tag{27}$$

while the first two terms of are quadratic (and concave) in $f_{n,k+1}(p)$. Therefore, their maximum is obtained either on the local optimum, $f^*_{n,k+1} = (n-2k)(n-2k-1)/2n$, or on the boundary of $f_{n,k+1}(p)$ (similarly to Section 4.1). Therefore, as in Theorem 4, we conclude that

- If $f^*_{n,k+1} \leq f^{max}_{n,k+1,m}$ then,

$$R(\hat{M}^{GT}_k, \Delta_m) \leq \frac{(n-2k)(n-2k-1)(k+1)^2}{4n^3} + \frac{(k+2)(k+1)}{(n-k)(n-k-1)} f^{max}_{n,k+2,m} + o\left(\frac{1}{n}\right)$$

- else

$$\begin{aligned} R(\hat{M}^{GT}_k, \Delta_m) \leq & \frac{-(k+1)^2}{n(n-2k)(n-2k-1)} \left(f^{max}_{n,k+1,m}\right)^2 + \\ & \left(\frac{k+1}{n}\right)^2 f^{max}_{n,k+1,m} + \frac{(k+2)(k+1)}{(n-k)(n-k-1)} f^{max}_{n,k+2,m} + o\left(\frac{1}{n}\right) \end{aligned} .$$

Figure 1 illustrates the obtained bounds for different proportions of m/n. Specifically, each curve refers to a different k value, where the lower curve is $k = 1$ and the upper curve is $k = 7$. We normalize the bounds by n to emphasize the difference from the unbounded alphabet size regime (see Theorem 5). First, we notice that for sufficiently large m/n, the curves converge to $g(k)$. However, for smaller values of m/n we observe a significant improvement, compared to the unbounded m setup. This result is not quite surprising. The unbounded alphabet setup is oblivious to the alphabet size, and holds for every m (specifically, for every $p \in \Delta$). Here, we assume that the alphabet size is restricted, and apply this knowledge to narrow the class of distributions we control. Naturally, the difference between the two schemes becomes more evident for smaller m.

A natural question that may follow considers the worst-case distribution, for which the proposed bound is attained. That is, the probability distribution which attains $R(\hat{M}^{GT}_k, \Delta_m)$. Unfortunately, our derivation considers both $f^{max}_{n,k+1,m}$ and $f^{max}_{n,k+2,m}$, where each term is a maximization objective, obtained for a different distribution. However, we do observe that for larger alphabet sizes, both terms obtain their maxima with uniform distributions of different alphabet sizes (both smaller than m). This emphasizes the unique role of the uniform distribution in OP estimation worst-case analysis, as demonstrated in (Rajaraman et al., 2017; Acharya et al., 2018; Painsky, 2022) for the missing mass problem, $k = 0$.

4.3 Large k Analysis

We now focus on the most general setup, and study the case where m, n and k hold no restrictions. Specifically, we assume that m is unbounded and k is not fixed, and may grow

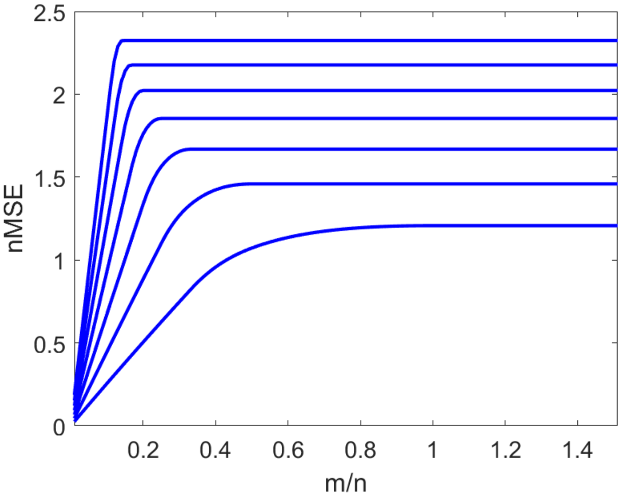

Figure 1: MSE bounds for the GT estimators, for a bounded alphabet size m. The lower curve is $k = 1$ while the upper curve corresponds to $k = 7$

with n. First, we consider the case where $2k < n$. Here, the MSE of the GT estimator satisfies (18). Define

$$\begin{aligned}\rho(q_1, q_2) &= q_1^k q_2^k (1 - q_1 - q_2)^{n-2k-2}\big(2k(2k+1) - n - 4kn(q_1 + q_2) + n^2(q_1 + q_2)^2\big) \qquad (28)\\ \eta(q) &= q^k (1 - q)^{n-k-1}\big((n - k)(k + 1) + n^2 q(1 - q)\big).\end{aligned}$$

Applying Propositions 2 and 3 to (18), we obtain

$$R_n(\hat{M}_k^{GT}, \Delta) \leq \frac{1}{n^2}\binom{n}{k\ k} \max_{q_1,q_2\in\Delta_2} \rho(q_1, q_2) + \frac{1}{n^2}\binom{n}{k} \max_{q\in[0,1]} \eta(q). \tag{29}$$

Simple calculus (see Appendix C) shows that the first term in (29) depends on a single variable. Namely, $\max_{q_1,q_2\in\Delta_2} \rho(q_1, q_2) = \max_{q_1\in[0,1/2]} \rho_1(q_1)$ where $\rho_1(q_1) = \rho(q_1, q_1)$. Let us characterize the maxima of $\rho_1(q_1)$ and $\eta(q)$. We begin with $\rho_1(q_1)$.

$$\rho_1(q_1) = q_1^{2k}(1 - 2q_1)^{n-2k-2}\big((2nq_1 - 2k)^2 + 2k - n\big) \leq q_1^{2k}(1 - 2q_1)^{n-2k-2}(2nq_1 - 2k)^2,$$

where the last inequality is due to $2k < n$. Taking the derivative of this upper bound, we obtain two candidates for a global maximum,

$$q_1^* = \frac{k(n-1)}{n^2} + \frac{1}{2n} \pm \frac{\sqrt{(n-2k)(n-2k+4kn)}}{2n^2}. \tag{30}$$

For the simplicity of notation, we refer to q_1^* as the single maximizer of $\rho_1(q_1)$, provided that $q_1^* \in [0, 1/2]$. In practice, we examine both candidates and choose the maximizer among them. Further, we have

$$\max_{q\in[0,1]} \eta(q) \leq \max_{t_1\in[0,1]} \eta_1(t_1) + \max_{t_2\in[0,1]} \eta_2(t_2) \tag{31}$$

where

$$\begin{aligned}\eta_1(t_1) &= (n-k)(k+1)t_1^k(1-t_1)^{n-k-1} \\ \eta_2(t_2) &= n^2 t_2^{k+1}(1-t_2)^{n-k}.\end{aligned} \tag{32}$$

It is immediate to see that for $k > 0$, we have $t_1^* = k/(n-1)$ and $t_2^* = (k+1)/(n+1)$. Notice that for sufficiently large n, the two maximizers are approximately equivalent. Putting together the above, we obtain

$$\begin{aligned}R_n(\hat{M}_k^{GT}, \Delta) \leq &\frac{1}{n^2}\binom{n}{k\ k}(q_1^*)^{2k}(1-2q_1^*)^{n-2k-2}(2nq_1^*-2k)^2 + \\ &\frac{1}{n^2}\binom{n}{k}(n-k)(k+1)(t_1^*)^k(1-t_1^*)^{n-k-1} + \binom{n}{k}(t_2^*)^{k+1}(1-t_2^*)^{n-k} = \\ &\frac{1}{n^2}\binom{2k}{k}\left(\frac{2nq_1^*-2k}{1-2q_1^*}\right)^2\left(\frac{1}{2}\right)^{2k}\text{Bin}(2k;n,2q_1^*) + \\ &\frac{(n-k)(k+1)}{n^2(1-t_1^*)}\text{Bin}(k;n,t_1^*) + t_2^*\text{Bin}(k;n,t_2^*).\end{aligned} \tag{33}$$

We now bound from above the Binomial terms using Robbin's version of Sterling's bound (Robbins, 1955). We show in Appendix D that

$$\text{Bin}(k;n,q) \leq \frac{1}{\sqrt{2\pi k(1-k/n)}}\exp\left(-nD_{KL}\left(k/n||q\right)\right) \tag{34}$$

where D_{KL} is the Kullback-Leibler divergence. Further, by Sterling's bound, we have that $\binom{2k}{k} \leq \frac{e}{\sqrt{2\pi}}\frac{2^{2k}}{\sqrt{k}}$. Putting together the above, we conclude with the following:

Theorem 6 *For $0 < 2k < n$, the MSE of the Good-Turing Estimator satisfies*

$$\begin{aligned}R_n(\hat{M}_k^{GT}, \Delta) \leq &\frac{(q_1^*-k/n)^2/(q_1^*-1/2)^2}{\sqrt{8\pi^3 e^{-2}k^2(1-2k/n)}}\exp\left(-nD_{KL}(2k/n||2q_1^*)\right) + \\ &\frac{(1-k/n)(k/n+1/n)}{(1-t_1^*)\sqrt{2\pi k(1-k/n)}}\exp\left(-nD_{KL}(k/n||t_1^*)\right) + \\ &\frac{t_2^*}{\sqrt{2\pi k(1-k/n)}}\exp\left(-nD_{KL}(k/n||t_2^*)\right).\end{aligned} \tag{35}$$

Finally, for $2k > n$, the first term in (18) equals zero, and the corresponding first term of the MSE bound (35) eliminates. However, this is a less typical setup for the GT estimator (as we later show). The analysis above introduces an upper bound for the MSE of the GT estimator, for different k and n. Interestingly, the obtained bound (35) strongly depends on the KL divergences between q_1^*, t_1^*, t_2^* and the proportion k/n. It is important to emphasize that a similar derivation also holds for the fixed k regime, discussed in Section 4.1. However, it results in a looser bound.

Let us further study (35). First, we have that $\exp\left(-nD_{KL}(p||q)\right) \leq \exp\left(-n(p-q)^2\right)$, following Painsky and Wornell (2018, 2019). This allows us to quantify the order of the exponential terms and show that they are all $O(1)$. Next, we apply q_1^*,t_1^* and t_2^* to (35) and obtain the following corollary.

Corollary 7 *For $0 < 2k < n$, the MSE of the Good-Turing estimator satisfies*

$$R_n(\hat{M}_k^{GT}, \Delta) \leq \sqrt{\frac{2}{\pi}} \left(\frac{\sqrt{k}}{n\sqrt{1-k/n}} \right) + O\left(\frac{1}{\sqrt{kn}} \right) \tag{36}$$

Importantly, we notice that this bound is $O(\sqrt{k}/n)$, which implies a convergence rate of $O(\sqrt[4]{k}/\sqrt{n})$ for the GT estimator. This result improves upon currently known convergence guarantees (6), (7), and attains Drukh and Mansour (2005) lower bound (8).

5. The MSE of the ML Estimator

To further evaluate our proposed bounds, we study the convergence rate of the maximum likelihood estimator, $\hat{M}_k^{ML} = k\Phi_k(X^n)/n$. Here, we only focus on the case where m holds no restrictions and k is relatively large, as discussed by Drukh and Mansour (2005) and Orlitsky and Suresh (2015). First, for $2k \leq n$, we have

$$\begin{aligned} R_n(\hat{M}_k^{ML}, p) = & \binom{n}{k\ k} \sum_{u \neq v} \left(p(u) - \frac{k}{n} \right) \left(p(v) - \frac{k}{n} \right) p^k(u) p^k(v) (1 - p(u) - p(v))^{n-2k} + \\ & \binom{n}{k} \sum_{u} \left(p(u) - \frac{k}{n} \right)^2 p^k(u) (1 - p(u))^{n-k}. \end{aligned} \tag{37}$$

Applying Propositions 2 and 3 to the above yields

$$R_n^*(\hat{M}_k^{ML}) \leq \binom{n}{k\ k} \max_{q_1, q_2 \in \Delta_2} \psi(q_1, q_2) + \binom{n}{k} \max_{q \in [0,1]} \Phi(q)$$

where

$$\begin{aligned} \psi(q_1, q_2) &= \left(q_1 - \frac{k}{n} \right) \left(q_2 - \frac{k}{n} \right) (q_1 q_2)^{k-1} (1 - q_1 - q_2)^{n-2k} \\ \Phi(q) &= \left(q - \frac{k}{n} \right)^2 q^{k-1} (1-q)^{n-k}. \end{aligned}$$

As in Section 4.3, simple calculus shows that $\max_{q_1, q_2 \in \Delta_2} \psi(q_1, q_2) = \max_{q_2 \in [0,1/2]} \psi_2(q_2)$ where $\psi_2(q_2) = \psi(q_2, q_2)$ (Appendix E). In this case, for $k > 1$,

$$\begin{aligned} q_2^* &= \frac{k}{n} - \frac{k}{n^2} \pm \frac{\sqrt{k\left(1 + \frac{k}{n^2} - \frac{2k}{n}\right)}}{n} \\ q^* &= \frac{k}{n} \left(\frac{2n-1}{2n+2} \right) + \frac{1}{2n+2} \pm \frac{\sqrt{\left(1 + \frac{k}{n}\right)^2 + 8k\left(1 - \frac{k}{n}\right)}}{2n+2}. \end{aligned} \tag{38}$$

For the simplicity of notation, we refer to q_2^* and q^* and as the maximizers of $\psi_2(q_2)$ and $\Phi(q)$, provided that $q_2^* \in [0, 1/2]$ and $q^* \in [0,1]$. Putting together the above, we obtain

$$R_n(\hat{M}_k^{ML}, \Delta) \leq \binom{2k}{k} \left(1 - \frac{k/n}{q_2^*} \right)^2 \left(\frac{1}{2} \right)^{2k} \text{Bin}(2k; n, 2q_2^*) + \left(q^* - \frac{k}{n} \right)^2 \left(\frac{1}{q^*} \right) \text{Bin}(k; n, q^*).$$

Finally, we bound from above the Binomial terms (similarly to (35)) and obtain the following theorem.

Theorem 8 *For $1 < 2k \leq n$, the MSE of the ML estimator satisfies*

$$R_n(\hat{M}_k^{ML}, \Delta) \leq \frac{\left(1-(k/n)/q_2^*\right)^2}{\sqrt{8\pi^3 e^{-2}k^2(1-2k/n)}} \exp\left(-nD_{KL}\left(2k/n||2q_2^*\right)\right) + \frac{\left(q^*-k/n\right)^2/q^*}{\sqrt{2\pi k(1-k/n)}} \exp\left(-nD_{KL}\left(k/n||q^*\right)\right). \tag{39}$$

Here again, for $2k > n$, the first term in (37) equals zero, and the corresponding first term of (39) eliminates. Similarly to the GT analysis (Corollary 7), it can be shown that the exponential terms in (39) are $O(1)$, the quadratic terms are $O(k/n^2)$ and q_1^*, q_2^* are both $O(k/n)$. Then, the ML estimator satisfies

$$R_n(\hat{M}_k^{ML}, \Delta) = O\left(\frac{1}{k^2} + \frac{1}{n\sqrt{k}}\right) \tag{40}$$

for $2k \leq n$ and $R_n(\hat{M}_k^{ML}, \Delta) = O\left(1/n\sqrt{k}\right)$ for $2k > n$.

Figure 2 compares the MSE bounds of the GT and ML estimators, for $k = 0.005n$ and an unbounded alphabet size m. As we can see, for smaller values of n (and k), the GT estimator demonstrates lower worst-case MSE. However, as n grows, the ML estimator outperforms GT. This behavior is not surprising; it is well-known that the GT estimator is superior to the ML only is cases where the number of samples is relatively small. In fact, most practical probability estimators are hybrid implementations of GT (for smaller k) and ML (for larger k) (Drukh and Mansour, 2005; Orlitsky and Suresh, 2015). Finally, we observe that as n increases, our proposed GT bound converges to zero (as shown in (36)). This significantly improves upon currently known convergence guarantees in this setup (7).

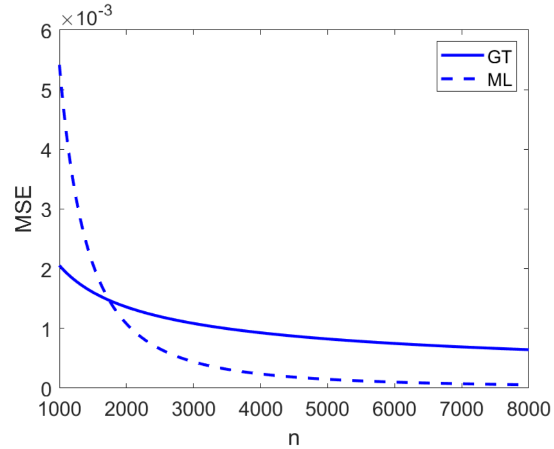

Figure 2: MSE bounds for ML and GT estimators, for $k = 0.005n$

6. The Hybrid Estimator

In the previous sections we show that the GT estimator enjoys favorable performance guarantees for relatively small k. On the other hand, the ML estimator becomes a favorable choice as k increases. We now introduce a hybrid GT-ML estimator, which benefits from both worlds. We focus on the general case where the alphabet size holds no restrictions. We begin our analysis with the case where $2k < n$. Let $0 < r \leq 1$. Define

$$\hat{M}_k^{hyb}(X^n) = \begin{cases} \hat{M}_k^{GT}(X^n) & k \leq n^r \\ \hat{M}_k^{ML}(X^n) & k > n^r \end{cases} \tag{41}$$

We now seek the exact value of r which minimizes the worst-case MSE, for every k.
For $k \leq n^r$ we have

$$R_n(\hat{M}_k^{hyb}, \Delta) = R_n(\hat{M}_k^{GT}, \Delta) = O\left(\frac{\sqrt{k}}{n}\right) = O\left(n^{r/2-1}\right) \tag{42}$$

where the second equality follows from Corollary 7.
For $k > n^r$ we have

$$R_n(\hat{M}_k^{hyb}, \Delta) = R_n(\hat{M}_k^{ML}, \Delta) = O\left(\frac{1}{k^2} + \frac{1}{n\sqrt{k}}\right) = O\left(n^{-2r} + n^{-r/2-1}\right) \tag{43}$$

where the second equality follows from (40). Let us first assume that $2r < r/2 + 1$. Then, the second term of the ML bound is dominated by the first term and

$$R_n(\hat{M}_k^{hyb}, \Delta) = \begin{cases} O(n^{r/2-1}) & k \leq n^r \\ O(n^{-2r}) & k > n^r \end{cases} \tag{44}$$

We seek a value of r which minimizes $R_n(\hat{M}_k^{hyb}, \Delta)$ for every k. Therefore, we require that $r/2 - 1 = -2r$. This leads to $r = 2/5$ and $R_n(\hat{M}_k^{hyb}, \Delta) = O(n^{-4/5})$. Notice that the obtained value of r also satisfies $2r < r/2 + 1$, as desired.

Second, assume that $2r \geq r/2 + 1$, which means that the first term of the ML bound is dominated by the second term. We have that

$$R_n(\hat{M}_k^{hyb}, \Delta) = \begin{cases} O\left(n^{r/2-1}\right) & k \leq n^r \\ O\left(n^{-r/2-1}\right) & k > n^r \end{cases} \tag{45}$$

Here, we cannot attain equality among the different cases of k. Therefore $R_n(\hat{M}_k^{hyb}, \Delta) = O(n^{r/2-1}) = O(n^{-2/3})$, for every $k < n/2$, where the second equality follows from $2r \geq r/2 + 1$. Putting together the above, we conclude that $r = 2/5$ obtains tighter performance guarantees, leading to $R_n(\hat{M}_k^{hyb}, \Delta) = O(n^{-4/5})$, for every $k < n/2$.

Finally, we study the case where $k \geq n/2$, for $r = 2/5$. First, we have that $R_n(\hat{M}_k^{hyb}, \Delta) = R_n(\hat{M}_k^{ML}, \Delta)$ for $n/2 > n^{2/5}$ (or equivalently, $n > 3$). This leads to $R_n(\hat{M}_k^{hyb}, \Delta) = O(n^{-r/2-1}) = O(n^{-6/5}) = O(n^{-4/5})$, which means that our results hold for every k. Theorem 9 summarizes the convergence guarantees of our proposed hybrid estimator.

Theorem 9 *For every $k \geq 0$, the hybrid OP estimator (45), with $r = 2/5$, satisfies*

$$R_n(\hat{M}_k^{hyb}, \Delta) = O\left(n^{-4/5}\right). \tag{46}$$

Applied to (15), we obtain a convergence rate of $O(n^{-2/\sqrt{5}})$. In their work, Drukh and Mansour (2005) introduced a similar hybrid estimator. They showed it attains a convergence rate of $O(n^{-2/5})$ for every k. Our proposed estimator improves upon this convergence rate.

7. Simultaneous Convergence Rates

In the previous sections we derive confidence guarantees for $M_k^{GT}(X^n)$ and $M_k^{ML}(X^n)$, independently for every k. Let us now introduce a simultaneous framework, where we study the convergence rate for multiple k's concurrently.

Let $\mathcal{K}$ be a collection of OP's indices, $\mathcal{K} \subseteq \{0, \dots, n\}$, as defined in Section 3. As shown in the previous sections, the GT estimator is a favorable choice for relatively small k, while the ML estimator performs better for larger k. Therefore, we define $\mathcal{K}^{GT}$ and $\mathcal{K}^{ML}$, such that $\mathcal{K}^{GT} \cup \mathcal{K}^{ML} = \mathcal{K}$ and $\mathcal{K}^{GT} \cap \mathcal{K}^{ML} = \emptyset$. In words, given a collection $\mathcal{K}$ we define a subset of OP's that are estimated by GT, and a subset of OP's that are estimated by ML. In (18) and (37) we derive the risks of the GT and ML estimators, respectively, for every $k < n/2$. Therefore, the additive risk (13) is given by

$$\begin{aligned}
&\sum_{k\in\mathcal{K}} R_n(\hat{M}_k, p) = \sum_{k\in\mathcal{K}^{\mathcal{GT}}} R_n(\hat{M}_k^{GT}, p) + \sum_{k\in\mathcal{K}^{\mathcal{ML}}} R_n(\hat{M}_k^{ML}, p) = \qquad (47)\\
&\quad \sum_{k\in\mathcal{K}^{\mathcal{GT}}} \frac{1}{n^2}\binom{n}{k\ k}\sum_{u\neq v} P(u,v)\left(2k(2k+1) - n - 4k(p(u)+p(v)) + n^2(p(u)+p(v))^2\right) + \\
&\quad \sum_{k\in\mathcal{K}^{\mathcal{ML}}} \binom{n}{k\ k}\sum_{u\neq v}\left(p(u)-\frac{k}{n}\right)\left(p(v)-\frac{k}{n}\right)p^k(u)p^k(v)(1-p(u)-p(v))^{n-2k} + \\
&\quad \sum_{k\in\mathcal{K}^{\mathcal{GT}}}\left(\frac{(k+1)^2}{n^2}\binom{n}{k+1}\sum_u p^{k+1}(u)(1-p(u))^{n-k-1} + \binom{n}{k}\sum_u p^{k+2}(u)(1-p(u))^{n-k}\right) + \\
&\quad \sum_{k\in\mathcal{K}^{\mathcal{ML}}}\binom{n}{k}\sum_u\left(p(u)-\frac{k}{n}\right)^2 p^k(u)(1-p(u))^{n-k},
\end{aligned}$$

where every $k \in \mathcal{K}$ that is not greater than $2n$. Let us study the expression above. We begin with the first two summations in (47). In Appendix F we show that

$$\begin{aligned}
&\sum_{k\in\mathcal{K}^{\mathcal{GT}}} \frac{1}{n^2}\binom{n}{k\ k}\sum_{u\neq v} P(u,v)\left(2k(2k+1) - n - 4k(p(u)+p(v)) + n^2(p(u)+p(v))^2\right) + \qquad (48)\\
&\sum_{k\in\mathcal{K}^{\mathcal{ML}}} \binom{n}{k\ k}\sum_{u\neq v}\left(p(u)-\frac{k}{n}\right)\left(p(v)-\frac{k}{n}\right)p^k(u)p^k(v)(1-p(u)-p(v))^{n-2k} \leq \\
&\max_{q_1,q_2\in\Delta_2}\left(\sum_{k\in\mathcal{K}^{\mathcal{GT}}}\frac{1}{n^2}\binom{n}{k\ k}q_1^k q_2^k(1-q_1-q_2)^{n-2k-2}(n(q_1+q_2)-2k)^2 + \right.\\
&\qquad\left.\sum_{k\in\mathcal{K}^{\mathcal{ML}}}\binom{n}{k\ k}\left(q_1-\frac{k}{n}\right)\left(q_2-\frac{k}{n}\right)q_1^{k-1}q_2^{k-1}(1-q_1-q_2)^{n-2k}\right),
\end{aligned}$$

for every $p \in \Delta$. Similarly to our analysis in the previous sections, Proposition 10, whose proof is provided in Appendix G, shows that the maximization above can be solved over a single parameter.

Proposition 10

$$\max_{q_1,q_2\in\Delta_2}\omega(q_1,q_2) = \max_{q\in[0,1/2]}\omega_1(q_1) \qquad (49)$$

where

$$\omega(q_1,q_2)=\sum_{k\in\mathcal{K}^{\mathcal{GT}}}\frac{1}{n^2}\binom{n}{k\ k}q_1^kq_2^k(1-q_1-q_2)^{n-2k-2}(n(q1+q2)-2k)^2+ \\ \sum_{k\in\mathcal{K}^{\mathcal{ML}}}\binom{n}{k\ k}\left(q_1-\frac{k}{n}\right)\left(q_2-\frac{k}{n}\right)q_1^{k-1}q_2^{k-1}(1-q_1-q_2)^{n-2k} \tag{50}$$

and $\omega_1(q_1)=\omega(q_1,q_1)$.

Plugging Proposition 10 to (48), we show that the first two summations in (47) are bounded from above by

$$\max_{q\in[0,1/2]}\sum_{k\in\mathcal{K}^{\mathcal{GT}}}\frac{1}{n^2}\binom{n}{k\ k}q^{2k}(1-2q)^{n-2k-2}(2nq-2k)^2+ \\ \sum_{k\in\mathcal{K}^{\mathcal{ML}}}\binom{n}{k\ k}\left(q-\frac{k}{n}\right)^2q^{2k-2}(1-2q)^{n-2k}= \\ \max_{q\in[0,1/2]}\sum_{k\in\mathcal{K}^{\mathcal{GT}}}\binom{2k}{k}\left(\frac{1}{2}\right)^{2k}\left(\frac{q-k/n}{q-1/2}\right)^2\mathrm{Bin}(2k;n,2q)+ \\ \sum_{k\in\mathcal{K}^{\mathcal{ML}}}\binom{2k}{k}\left(\frac{1}{2}\right)^{2k}\left(1-\frac{k/n}{q}\right)^2\mathrm{Bin}(2k;n,2q)\le \\ \max_{q\in[0,1/2]}\sum_{k\in\mathcal{K}^{\mathcal{GT}}}\left(\frac{q-k/n}{q-1/2}\right)^2\frac{1}{\sqrt{8\pi^3e^{-2}k^2(1-2k/n)}}\exp(-nD_{KL}(2k/n||2q))+ \\ \sum_{k\in\mathcal{K}^{\mathcal{ML}}}\left(1-\frac{k/n}{q}\right)^2\frac{1}{\sqrt{8\pi^3e^{-2}k^2(1-2k/n)}}\exp(-nD_{KL}(2k/n||2q)) \tag{51}$$

where the last inequality follows from (34) and $\binom{2k}{k}\le\frac{e}{\sqrt{2\pi}}\frac{2^{2k}}{\sqrt{k}}$. We now proceed to the last two summations in (47). Here, we have that

$$\sum_{k\in\mathcal{K}^{\mathcal{GT}}}\left(\frac{(k+1)^2}{n^2}\binom{n}{k+1}\sum_u p^{k+1}(u)(1-p(u))^{n-k-1}+\binom{n}{k}\sum_u p^{k+2}(u)(1-p(u))^{n-k}\right)+ \\ \sum_{k\in\mathcal{K}^{\mathcal{ML}}}\binom{n}{k}\sum_u\left(p(u)-\frac{k}{n}\right)^2p^k(u)(1-p(u))^{n-k}\le \\ \max_{q\in[0,1]}\sum_{k\in\mathcal{K}^{\mathcal{GT}}}\left(\frac{(k+1)^2}{n^2}\left(\frac{n-k}{k+1}\right)\frac{1}{1-q}+q\right)\frac{1}{\sqrt{2\pi k(1-k/n)}}\exp(-nD_{KL}(k/n||q))+ \\ \sum_{k\in\mathcal{K}^{\mathcal{ML}}}\left(q-\frac{k}{n}\right)^2q^{-1}\frac{1}{\sqrt{2\pi k(1-k/n)}}\exp(-nD_{KL}(k/n||q)),$$

for every $p \in \Delta$, where the inequality follows from Proposition 2 and (34). The complete details are provided in Appendix H. To conclude, for $k < n/2$ we have

$$\begin{aligned}
&\sum_{k\in\mathcal{K}} R_n(\hat{M}_k, p) \leq \qquad (52)\\
&\max_{q\in[0,1/2]} \sum_{k\in\mathcal{K}^{GT}} \left(\frac{q-k/n}{q-1/2}\right)^2 \frac{1}{\sqrt{8\pi^3 e^{-2}k^2(1-2k/n)}} \exp(-nD_{KL}(2k/n||2q))+\\
&\qquad \sum_{k\in\mathcal{K}^{ML}} \left(1-\frac{k/n}{q}\right)^2 \frac{1}{\sqrt{8\pi^3 e^{-2}k^2(1-2k/n)}} \exp(-nD_{KL}(2k/n||2q))+\\
&\max_{t\in[0,1]} \sum_{k\in\mathcal{K}^{GT}} \left(\frac{(k+1)^2}{n^2}\left(\frac{n-k}{k+1}\right)\frac{1}{1-t}+t\right) \frac{1}{\sqrt{2\pi k(1-k/n)}} \exp(-nD_{KL}(k/n||t))+\\
&\qquad \sum_{k\in\mathcal{K}^{ML}} \left(t-\frac{k}{n}\right)^2 t^{-1} \frac{1}{\sqrt{2\pi k(1-k/n)}} \exp(-nD_{KL}(k/n||t)),
\end{aligned}$$

for every $p \in \Delta$. Let us now consider the case where $\mathcal{K}$ also consists of indices that are greater than $n/2$. Define $\mathcal{K} = \mathcal{K}^{GT} \cup \mathcal{K}_1^{ML} \cup \mathcal{K}_2^{ML}$, such that $\mathcal{K}^{GT}$ is the collection of all $k \in \mathcal{K}$ for which we apply the GT estimator, $\mathcal{K}_1^{ML}$ is the collection of all $k \in \mathcal{K}$ and $k \leq n/2$, for which we apply the ML estimator, and $\mathcal{K}_2^{ML}$ is the collection of $k \in \mathcal{K}$ and $k > n/2$, for which we apply the ML estimator. As before, the three sets are mutually disjoint. Following (37) for $k > n/2$, the ML satisfies

$$R_n(\hat{M}_k, p) = \binom{n}{k} \sum_u \left(p(u) - \frac{k}{n}\right)^2 p^k(u)(1-p(u))^{n-k}. \qquad (53)$$

Therefore, for every $\mathcal{K} \subseteq \{0, \dots, n\}$, we have that

$$\begin{aligned}
&\sum_{k\in\mathcal{K}} R_n(\hat{M}_k, p) = \qquad (54)\\
&\sum_{k\in\mathcal{K}^{GT}} R_n(\hat{M}_k^{GT}, p) + \sum_{k\in\mathcal{K}_1^{ML}} R_n(\hat{M}_k^{ML}, p) + \sum_{k\in\mathcal{K}_2^{ML}} R_n(\hat{M}_k^{ML}, p) \leq\\
&\max_{q\in[0,1/2]} \sum_{k\in\mathcal{K}^{GT}} \left(\frac{q-k/n}{q-1/2}\right)^2 \frac{1}{\sqrt{8\pi^3 e^{-2}k^2(1-2k/n)}} \exp(-nD_{KL}(2k/n||2q))+\\
&\qquad \sum_{k\in\mathcal{K}_1^{ML}} \left(1-\frac{k/n}{q}\right)^2 \frac{1}{\sqrt{8\pi^3 e^{-2}k^2(1-2k/n)}} \exp(-nD_{KL}(2k/n||2q))+\\
&\max_{t\in[0,1]} \sum_{k\in\mathcal{K}^{GT}} \left(\frac{(k+1)^2}{n^2}\left(\frac{n-k}{k+1}\right)\frac{1}{1-t}+t\right) \frac{1}{\sqrt{2\pi k(1-k/n)}} \exp(-nD_{KL}(k/n||t))+\\
&\qquad \sum_{k\in\mathcal{K}_1^{ML}\cup\mathcal{K}_2^{ML}} \left(t-\frac{k}{n}\right)^2 t^{-1} \frac{1}{\sqrt{2\pi k(1-k/n)}} \exp(-nD_{KL}(k/n||t)),
\end{aligned}$$

for every $p \in \Delta$. Notice that (54) may be numerically evaluated quite efficiently, as it only depends on two parameters, q and t.

The obtained additive risk bound is defined for a given decomposition $\mathcal{K} = \mathcal{K}^{GT} \cup \mathcal{K}_1^{ML} \cup \mathcal{K}_2^{ML}$. We now seek the best decomposition, which minimizes (54). As demonstrated in the previous sections, the GT estimator is a favorable choice for relatively small values of k, while the ML estimator demonstrates improved performance guarantees as k increases. Therefore, we propose a threshold $k^* < n/2$, such that $k \in \mathcal{K}^{GT}$ for all and $k \leq k^*$. In words, for a given collection of indices $\mathcal{K}$, we set a threshold $k^* < n/2$ and split $\mathcal{K}$ into three disjoint sets, $\mathcal{K}^{GT}$ for $k \leq k^*$, $\mathcal{K}_1^{ML}$ for $k^* < k \leq n/2$ and $\mathcal{K}_2^{ML}$ for $k > n/2$. We seek the optimal $k^* \in \{0, \dots, n/2\}$ that minimizes the MSE bound (54).

We now demonstrate our suggested bound, for $\mathcal{K} = \{0, \cdots, n\}$, as n increases. For every value of n we apply our proposed algorithm and find the optimal k_n^*. We compare our result to a marginal bound, which seeks the maximal MSE for every k independently,

$$\sum_{k=0}^{k_n^*} R_n(\hat{M}_k^{GT}, \Delta) + \sum_{k=k_n^*}^{n} R_n(\hat{M}_k^{ML}, \Delta). \tag{55}$$

Figure 3 demonstrates the results we achieve. The right chart compares the MSE of the simultaneous bound (54) with the marginal bound (55). As we can see, the proposed simultaneous bound demonstrates a significant improvement. The left chart in Figure 3 illustrates k_n^* as n grows. Here, we observe that the optimal threshold grows quite slowly with n. Interestingly, for $n = 20,000$ samples, the obtained threshold is only less than 20.

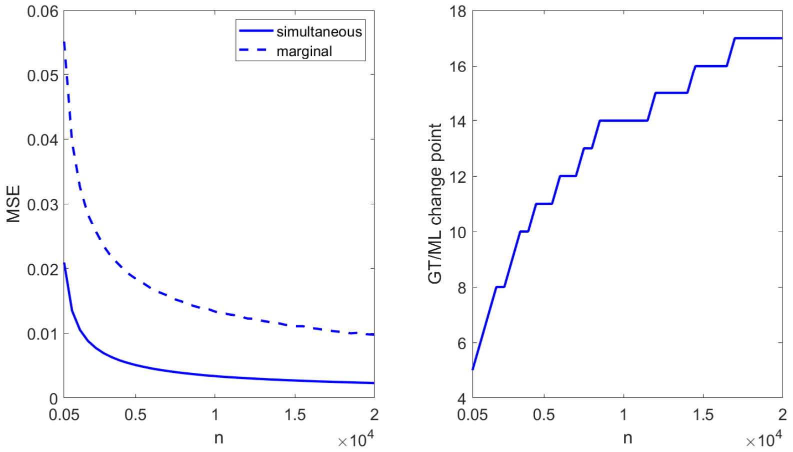

Figure 3: Simultaneous MSE bound. Right: MSE bound for the simultaneous and marginal analysis. Left: the optimal threshold between GT and ML estimators in the simultaneous scheme

8. Experiments

We now illustrate the performance of our proposed estimation schemes. First, we study three example distributions, which are common benchmarks for probability estimation and related problems (Orlitsky and Suresh, 2015). The Zipf's law distribution is a typical benchmark in large alphabet probability estimation; it is a commonly used heavy-tailed distribution, mostly for modeling natural (real-world) quantities in physical and social sciences, linguistics, economics and others fields (Saichev et al., 2009). The Zipf's law distribution follows $p(u; s, m) = u^{-s}/\sum_{v=1}^{m} v^{-s}$ where m is the alphabet size and s is a skewness parameter. Additional example distributions are the uniform, $p(u) = 1/m$, and the step distribution, $p(u) \propto \mathbb{1}(u \leq m/2) + 1/4 \cdot \mathbb{1}(u > m/2)$. In each experiment we draw n samples, and compare the occupancy probabilities $M_k(X^n)$ with their corresponding estimators, for different values of k. Figure 4 demonstrates the results we achieve. The upper row corresponds to the Zipf's Law distribution (with $s = 1.01$), the middle row is the uniform distribution and the lower row is the step distribution. The alphabet size is set to $m = 10,000$ in all the setups. Each plot in figure 4 corresponds to a different k value, as we focus on the MSE. To attain an averaged error, we repeat each experiment 1000 times, and average the squared error. We study a collection of estimation schemes. Specifically, the GT estimator (4), the ML, Laplace (3), the smooth GT estimator (9) and the PML estimator (10). Notice that the PML estimator is implemented by an approximate scheme (Pavlichin et al., 2019), due to the high complexity of the problem (see Section 2 for details). The schemes above are compared to the proposed hybrid GT-ML, as introduced in Section 6. We observe that not all estimators are competitive for every k and n. Hence, non-competitive estimators are omitted from Figure 4 for brevity.

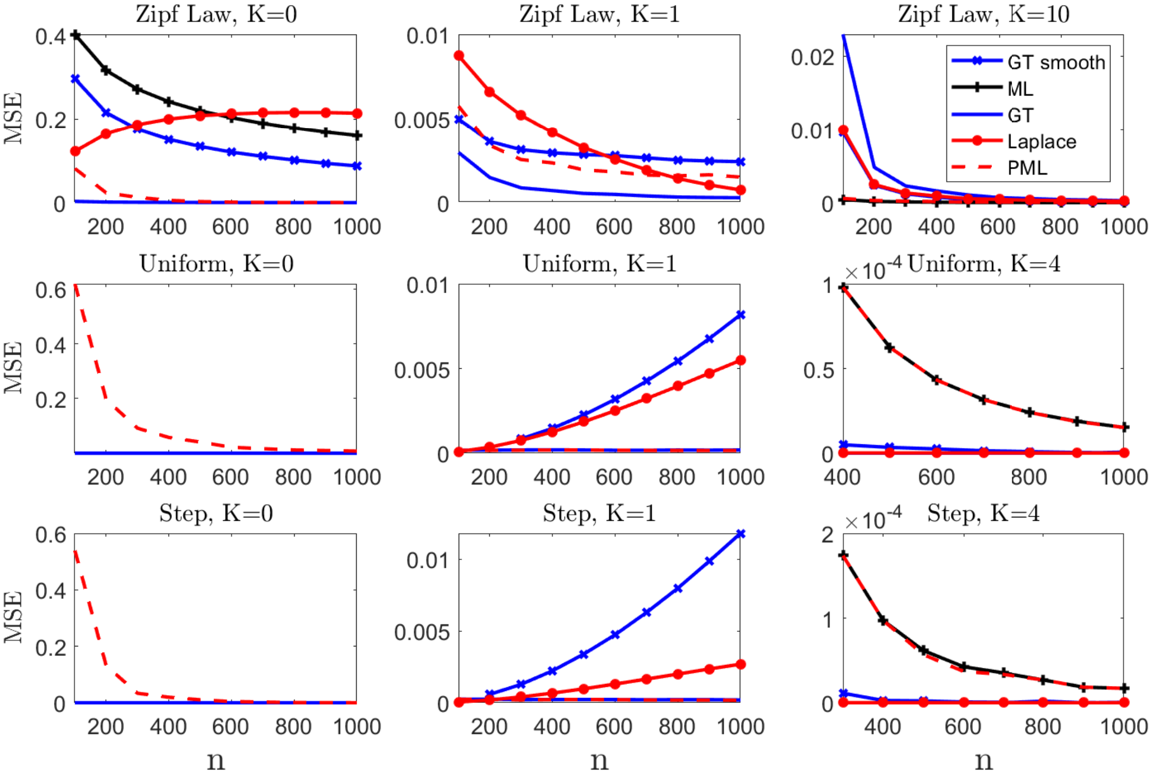

Figure 4: MSE of OP estimation in three synthetic experiments

We begin with the Zipf's Law distribution, as we examine $K = 0$. Here, the GT estimator is a preferable choice, compared to alternative methods. Notice that for $K = 0$, the hybrid estimator is equivalent to GT for every n (and henceforth not presented in the plot). As we proceed to $K = 1$ (the middle plot of the upper row) we observe a similar behavior, where now the ML estimator is omitted from the plot (noncompetitive). Finally, we study a greater value of $k = 10$. Here, the PML and the ML estimators show improved performance for every n. On the other hand, for $k = 10$ the hybrid estimator equals to ML for $n < 317$, and to the GT estimator otherwise. Next, we examine the uniform distribution for $k = 0$. Here, we observe that only the GT and the PML estimators are competitive, where GT outperforms PML quite significantly. Notice that for a uniform distribution, the missing mass, $M_0(X^n)$, is typically quite larger than in the heavy-tailed Zipf's Law distribution. Therefore, estimators which tend to underestimate $M_0(X^n)$ (that is, ML, Laplace and even the smooth GT) under-perform. As we proceed to $k = 1$, we observe a similar behavior where now the smooth GT and Laplace are more competitive, but still outperformed by GT and the PML. Finally, we examine $k = 4$ (larger values of k are very sparse in the uniform distribution setup). Here, the ML and the PML behave quite similarly, and are outperformed by smooth GT, GT and Laplace. Notice that in this setup the hybrid GT-ML estimator is equal to GT for every $n \geq 32$. Last, we study the step distribution. Here we observe a similar behavior to the uniform distribution, where GT (and consequentially the GT-ML estimator) significantly outperform alternative methods.

Next, we turn to real-world experiments. Here, we follow Orlitsky et al. (2016) and study three application domains. Notice that in these real-world settings, the true underlying probability is unknown. Hence, the occupancy probabilities refer to the frequency of symbols in the full data-set. We begin with a corpus linguistic experiment. The popular Broadway play *Hamilton* consists of 20,520 words, of which $m = 3{,}578$ are distinct. We randomly sample n words (with replacement), and estimate the $M_k(X^n)$ from the sample. Once again, we examine different k values for an increasing sample size n. The upper row of Figure 5 demonstrates the results we achieve. First, we observe that for $k = 0$ and $k = 1$, GT is the favorable choice, similarly to the synthetic experiments above. Then, for $k = 8$, PML and ML introduce improved performance. Notice that in this case, the hybrid GT-ML estimator equals to ML for $n < 181$ and to GT estimator otherwise. Next, we focus on a biota analysis. Gao et al. (2007) considered the forearm skin biota of six subjects. They identified a total of 1,221 clones consisting of 182 different species-level operational taxonomic units (SLOTUs). As above, we sample n out of the $m = 1{,}221$ clones with replacement, and estimate the occupancy probabilities. Notice that here, the alphabet size m is relatively small, so we focus on $n \leq 100$. The middle row in Figure 5 demonstrates the results we achieve. For $k = 0$, the GT estimator outperforms alternative methods quite significantly. However, for $k = 1$, the Laplace estimator is superior. Notice that the smooth GT fails to estimate the occupancy probabilities for a small sample size. Finally, we study $k = 8$. Here, the ML and the PML are the preferred schemes. However, notice that the hybrid estimator equals to ML for every examined n, which makes it a preferable choice. Finally, we study census data. The lower row of Figure 5 considers the 2000 United States Census (Bureau, 2014), which lists the frequency of the top $m = 1000$ most common last names in the United States. Here too, we sample n names and estimate $M_k(X^n)$. Similarly to the above, the GT estimator introduces favorable performance for $k = 0$ and $k = 1$, while

ML and PML improve as k increases. On the other hand, the hybrid GT-ML equals GT for every n in $k = 0$ and $k = 1$, and for $n \geq 182$ in the $k = 8$ setup. This again, makes it a preferable choice.

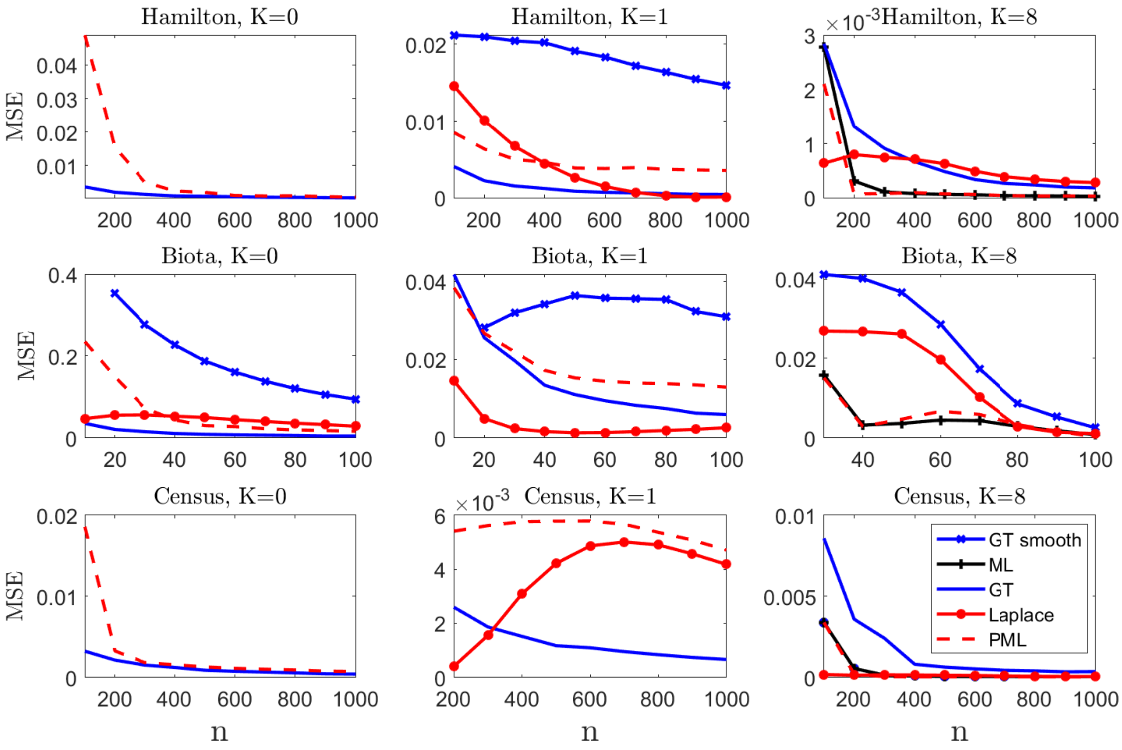

Figure 5: MSE of OP estimation in three real-world experiments

In addition to the MSE, we further examine the mean absolute error (MAE) of the studied estimators. Notice that the MAE is bounded from above (in high probability) by the convergence rates we introduce in the previous sections. The results we attain are similar in spirit to the MSE, and are reported in Appendix I for brevity.

To conclude, we study a total of six distributions and compare the MSE and MAE for different values of k and n. We observe that typically, the GT estimator is a preferable choice for smaller k, while the ML estimator preforms better as k increases. The Laplace and the smooth GT estimators perform quite well in some cases but are noncompetitive in others. As we consider all setups together, the PML and the proposed hybrid GT-ML estimator demonstrate favorable results. Comparing the two, the hybrid estimator outperforms PML mostly for smaller values of k and n (where the errors are typically larger).

9. Discussion and Conclusions

In this work we study the convergence rate of the GT estimator in different setups of interest. We first consider the case where k is fixed and the alphabet size m is either bounded or unbounded. Then, we focus on the case where k is not fixed, and may grow with the sample size n. Next, we study the convergence rate of the ML estimator, for cases where

k is relatively large. We utilize the obtained results and derive a hybrid estimator with a uniform convergence rate that holds for every k. We develop a simultaneous convergence framework, which considers multiple k's simultaneously. We apply our proposed framework and introduce a novel estimator with favorable simultaneous convergence guarantees. The performance of our suggested schemes is demonstrate in extensive numerical experiments.

Our analysis is based on the worst-case MSE. That is, given an OP estimator, we bound from above its MSE over a collection of countable distributions and apply Markov's inequality to obtain the stated convergence rates. Our results provide exact confidence intervals and improve upon currently known convergence guarantees. We further show that the GT estimator asymptotically attains known estimation lower bounds. In addition, our proposed method provides a simple framework to construct a simultaneous confidence region for any desired set of occupancy probabilities. This allows us to quantify the performance of OP estimators, for a collection of k values.

Our suggested framework also provides a simple tool for improving OP estimation accuracy. For example, in Sections 6 and 7 we consider a collection of hybrid estimators, and seek the one which attains the optimal performance guarantees. This methodology may be extended to a broader collection of estimators. For example, we may consider generalized GT estimators (that is, $\hat{M}_k(X^n) = \alpha_k \Phi_k(X^n) + \beta_k \Phi_{k+1}(X^n)$, following (Painsky, 2022) and seek the coefficients which minimize the convergence bound, marginally, or simultaneously for a collection of k's. This way, we introduce novel estimators that improve upon currently known large alphabet frameworks. We consider this scheme for our future work.

Acknowledgments

This research is supported by the Israel Science Foundation grant number 963/21.

Appendix A - A Proof For Theorem 1

We first consider the case where $2k < n$. In this case, the risk of the GT estimator satisfies (18). We focus on the first summation,

$$\frac{1}{n^2}\binom{n}{k\ k}\sum_{u\neq v} P(u,v)\big(2k(2k+1) - n - 4nk(p(u)+p(v)) + n^2(p(u)+p(v))^2\big). \tag{56}$$

Let us study the different terms in (56). First,

$$\binom{n}{k\ k}\sum_{u\neq v} P(u,v)(p(u)+p(v))^2 = \binom{n}{k\ k}\sum_{u\neq v} P(u,v)\left(p^2(u) + 2p(v)p(v) + p^2(v)\right). \tag{57}$$

Lemma 1 of (Rajaraman et al., 2017) states that

$$\sum_{u\neq v} p^i(u)p^j(v)(1-p(u)-p(v))^n \leq \frac{(i-1)!(j-1)!n!}{(n+i+j-2)!} \tag{58}$$

Plugging (58) to (57) yields

$$\binom{n}{k\ k}\sum_{u\neq v}P(u,v)(p(u)+p(v))^2=o\left(\frac{1}{n}\right). \tag{59}$$

Similarly, we have

$$\frac{1}{n^2}\binom{n}{k\ k}\sum_{u\neq v}P(u,v)\big(2k(2k+1)-4nk(p(u)+p(v))\big)=o\left(\frac{1}{n}\right). \tag{60}$$

Therefore, the first term in (18) equals

$$-\frac{1}{n}\binom{n}{k\ k}\sum_{u\neq v}P(u,v)+o\left(\frac{1}{n}\right) \tag{61}$$

and

$$\begin{aligned}R_n(\hat{M}_k^{GT},p)=&-\frac{1}{n}\binom{n}{k\ k}\sum_{u\neq v}P(u,v)+\left(\frac{k+1}{n}\right)^2\binom{n}{k+1}\sum_u p^{k+1}(u)(1-p(u))^{n-k-1}+\\&\binom{n}{k}\sum_u p^{k+2}(u)(1-p(u))^{n-k}+o\left(\frac{1}{n}\right).\end{aligned} \tag{62}$$

We now rewrite (62) in a more compact manner. First, we have that

$$\begin{aligned}\mathbb{E}_{X^n\sim p}\left(\Phi_k\left(X^n\right)\right)=&\mathbb{E}_{X^n\sim p}\left(\sum_u \mathbb{1}(N_u(X^n)=k)\right)=\\&\mathbb{E}_{X^n\sim p}\left(\sum_{u=v}\mathbb{1}(N_u(X^n)=k)\mathbb{1}(N_v(X^n)=k)\right)=\\&\sum_{u=v}P_n(k,k)=\binom{n}{k}\sum_u p^k(u)(1-p(u))^{n-k}.\end{aligned} \tag{63}$$

We begin with the first term in (62),

$$\begin{aligned}&\frac{1}{n}\binom{n}{k\ k}\sum_{u\neq v}P(u,v)=\frac{1}{n}\left(\binom{n}{k\ k}\Big/\binom{n}{k+1\ k+1}\right)\sum_{u\neq v}P_n(k+1,k+1)= \\&\frac{(k+1)^2}{n(n-2k)(n-2k-1)}\mathbb{E}_{X^n\sim p}\sum_{u\neq v}\mathbb{1}(N_u(X^n)=k+1)\mathbb{1}(N_v(X^n)=k+1)=\\&\frac{(k+1)^2}{n(n-2k)(n-2k-1)}\mathbb{E}_{X^n\sim p}\left(\left(\sum_u\mathbb{1}(N_u(X^n)=k+1)\right)^2-\sum_u\mathbb{1}(N_u(X^n)=k+1)\right)=\\&\frac{(k+1)^2}{n(n-2k)(n-2k-1)}\mathbb{E}_{X^n\sim p}\left(\Phi_{k+1}^2(X^n)-\Phi_{k+1}(X^n)\right).\end{aligned} \tag{64}$$

Notice that

$$\mathbb{E}_{X^n\sim p}\left(\Phi_{k+1}(X^n)\right) = \binom{n}{k+1}\sum_u p^{k+1}(u)(1-p(u))^{n-k-1} \leq \binom{n}{k+1}\frac{k!(n-k-1)!}{(n-1)!} = \frac{n}{k+1} \tag{65}$$

where the first equality is due to (63) and the inequality follows from Lemma 2 of (Rajaraman et al., 2017),

$$\sum_{u\in\mathcal{X}} p^i(u)(1-p(u))^n \leq \frac{(i-1)!n!}{(n-1+i)!}. \tag{66}$$

Plugging (65) to (64), we obtain

$$-\frac{1}{n}\binom{n}{k\;k}\sum_{u\neq v} P(u,v) = \frac{-(k+1)^2}{n(n-2k)(n-2k-1)}\mathbb{E}_{X^n\sim p}\left(\Phi^2_{k+1}(X^n)\right) + o\left(\frac{1}{n}\right). \tag{67}$$

We now continue to the second term in (62). Here, we have that

$$\left(\frac{k+1}{n}\right)^2\binom{n}{k+1}\sum_u p^{k+1}(u)(1-p(u))^{n-k-1} = \left(\frac{k+1}{n}\right)^2\mathbb{E}_{X^n\sim p}\left(\Phi_{k+1}(X^n)\right) \tag{68}$$

where the equality follows from (63). Finally, the third term in (62) satisfies

$$\begin{aligned}\binom{n}{k}\sum_u p^{k+2}(u)(1-p(u))^{n-k} = \binom{n}{k}\sum_u p^{k+2}(u)(1-p(u))^{n-k-2}(1-p(u))^2 = \\ \binom{n}{k}\sum_u p^{k+2}(u)(1-p(u))^{n-k-2} - 2\binom{n}{k}\sum_u p^{k+3}(u)(1-p(u))^{n-k-2} + \\ \binom{n}{k}\sum_u p^{k+4}(u)(1-p(u))^{n-k-2} = \binom{n}{k}\sum_u p^{k+2}(u)(1-p(u))^{n-k-2} + o\left(\frac{1}{n}\right) = \\ \frac{(k+1)(k+2)}{(n-k)(n-k-1)}\mathbb{E}_{X^n\sim p}\left(\Phi_{k+2}(X^n)\right) + o\left(\frac{1}{n}\right).\end{aligned} \tag{69}$$

where the third equality follows from (66) and the final equality is due to (63). Putting together (67), (68) and (69), we conclude that

$$\begin{aligned}\mathbb{E}_{X^n\sim p}\left(\hat{M}^{GT}_k - M_k\right)^2 = \frac{-(k+1)^2}{n(n-2k)(n-2k-1)}\mathbb{E}_{X^n\sim p}\left(\Phi^2_{k+1}(X^n)\right) + \\ \left(\frac{k+1}{n}\right)^2\mathbb{E}_{X^n\sim p}\left(\Phi_{k+1}(X^n)\right) + \frac{(k+1)(k+2)}{(n-k)(n-k-1)}\mathbb{E}_{X^n\sim p}\left(\Phi_{k+2}(X^n)\right) + o\left(\frac{1}{n}\right),\end{aligned} \tag{70}$$

as desired.

Appendix B - Proofs for Propositions 2 and 3

A proof for Proposition 2

Let $X \sim p$ and $Y \sim p$ be two independent and identically distributed random variables. Define a random variable $T(X,Y)$, such that $T(u,v)=0$ for $u=v$, and $T(u,v)=\psi(p(u),p(v))$ for $u \neq v$. Then,

$$\mathbb{E}(T(X,Y)) = \sum_{u \neq v} p(u)p(v)\psi(p(u),p(v)) \leq \max_{q_1,q_2 \in \Delta_2} \psi(q_1,q_2)$$

where in the last inequality, the expectation of a random variable is bounded from above by its maximal value.

A proof for Proposition 3

Following Proposition 2, let $X \sim p$ and define a random variable $T(u)$, such that $T(u) = \phi(p(u))$. Then, $\mathbb{E}(T(X)) = \sum_{u,\in\mathcal{X}} p(u)\phi(p(u)) \leq \max_{q\in[0,1]} \phi(q)$.

Appendix C

Proposition 11

$$\max_{q_1,q_2\in\Delta_2} \rho(q_1,q_2) = \max_{q_1\in[0,1/2]} \rho_1(q_1) \tag{71}$$

where $\rho_1(q_1) = \rho(q_1,q_1)$.

Proof We first notice that $\max_{q_1,q_2\in\Delta_2} \rho(q_1,q_2) \geq 0$ since $\rho(0,0)=0$. Next, we show that for every pair $q_1,q_2 \in \Delta_2$ such that $\rho(q_1,q_2) \geq 0$, we have $\rho(q_1,q_2) \leq \rho((q_1+q_2)/2,(q_1+q_2)/2)$. Therefore, we would like to show that

$$\begin{aligned} &q_1^k q_2^k (1-q_1-q_2)^{n-2k-2}\big(2k(2k+1)-n-4kn(q_1+q_2)+n^2(q_1+q_2)^2\big) \leq \\ &\left(\frac{q_1+q_2}{2}\right)^{2k}(1-q_1-q_2)^{n-2k-2}\big(2k(2k+1)-n-4kn(q_1+q_2)+n^2(q_1+q_2)^2\big)\end{aligned} \tag{72}$$

for every q_1,q_2 such that $2k(2k+1)-n-4kn(q_1+q_2)+n^2(q_1+q_2)^2$ is non-negative. This inequality holds since $q_1^k q_2^k \leq \left(\frac{q_1+q_2}{2}\right)^{2k}$, as later shown in the proof of Proposition 12 (Appendix E). ■

Appendix D

We bound from above the Binomial distribution, using Sterling bounds. We have

$$\begin{aligned}\log \mathrm{Bin}(k;n,q) =& \log\binom{n}{k} + k\log(q) + (n-k)\log(1-q) = \\ & \log\binom{n}{rn} + n\left(r\log(q)+(1-r)\log(1-q)\right),\end{aligned} \tag{73}$$

where $r = k/n$. The binomial coefficient satisfies

$$\begin{aligned}\log \binom{n}{rn} =& \log n! - \log(rn)! - \log(n-rn)! \leq \\ & -\log\sqrt{2\pi} + \frac{1}{12n} - \frac{1}{12rn+1} - \frac{1}{12(1-r)n+1} + \\ & \left(n+\frac{1}{2}\right)\log(n) - \left(rn+\frac{1}{2}\right)\log(rn) - \left(n-rn+\frac{1}{2}\right)\log(n-rn) \leq \\ & -\frac{1}{2}\log\left(2\pi nr(1-r)\right) + nH(r)\end{aligned} \tag{74}$$

where the first inequality follows from Robbin's version of Sterling's bound (Robbins, 1955),

$$\sqrt{2\pi}\; n^{n+\frac{1}{2}}e^{-n}e^{\frac{1}{12n+1}} < n! < \sqrt{2\pi}\; n^{n+\frac{1}{2}}e^{-n}e^{\frac{1}{12n}},$$

and the second inequality follows from

$$\frac{1}{12n} - \frac{1}{12rn+1} - \frac{1}{12(1-r)n+1} \leq \frac{1}{12n} - \frac{2}{6n+1} < 0$$

for $0 \leq r \leq 1$, where $H(r)$ is the binary entropy of r, $H(r) = -r\log(r) - (1-r)\log(1-r)$. Therefore,

$$\begin{aligned}\log \mathrm{Bin}(rn; n, q) \leq & -\frac{1}{2}\log\left(2\pi nr(1-r)\right) + nH(r) + n\left(r\log(q) + (1-r)\log(1-q)\right) = \\ & -\frac{1}{2}\log\left(2\pi nr(1-r)\right) - nD_{KL}(r||q)\end{aligned}$$

where $D_{KL}(r||q)$ is the Kullback-Leibler divergence,

$$D_{KL}(r||q) = r\log\frac{r}{q} + (1-r)\log\frac{(1-r)}{(1-q)}.$$

This means that

$$\mathrm{Bin}(k; n, q) \leq \frac{1}{\sqrt{2\pi k(1-k/n)}}\exp\left(-nD_{KL}\left(\frac{k}{n}\middle|\middle| q\right)\right), \tag{75}$$

as desired.

Appendix E

Proposition 12

$$\max_{q_1,q_2\in\Delta_2} \psi(q_1,q_2) = \max_{q_1\in[0,1/2]} \psi_2(q_2) \tag{76}$$

where $\psi_2(q_2) = \psi(q_2, q_2)$

Proof We study $\psi(q_1, q_2)$ for different possible pairs of $q_1, q_2 \in \Delta_2$.

- For $q_1 \leq \frac{k}{n}$ and $q_2 \geq \frac{k}{n}$ (and vice versa):
  We have that $\psi(q_1, q_2) < 0$, while $\psi(\frac{q_1+q_2}{2}, \frac{q_1+q_2}{2}) \geq 0$. Therefore, $\psi(q_1, q_2) \leq \psi(\frac{q_1+q_2}{2}, \frac{q_1+q_2}{2})$.

- For $q_1, q_2 \geq \frac{k}{n}$:
  We would like to show that $\psi(q_1, q_2) \leq \psi(\frac{q_1+q_2}{2}, \frac{q_1+q_2}{2})$. Plugging (38), we require that

  $$\left(q_1 - \frac{k}{n}\right)\left(q_2 - \frac{k}{n}\right) q_1^{k-1} q_2^{k-1} \leq \left(\frac{q_1+q_2}{2} - \frac{k}{n}\right)^2 \left(\frac{q_1+q_2}{2}\right)^{2k-2}$$

  Since both sides of the inequality are positive, this is equivalent to

  $$\begin{aligned} &\log\left(q_1 - \frac{k}{n}\right) + \log\left(q_2 - \frac{k}{n}\right) + (k-1)\left(\log(q_1) + \log(q_2)\right) \leq \\ &2\log\left(\frac{q_1+q_2}{2} - \frac{k}{n}\right) + (2k-2)\log\left(\frac{q_1+q_2}{2}\right). \end{aligned} \tag{77}$$

  Due to the concavity of the log function, we have

  $$\begin{aligned} &\frac{1}{2}\log\left(q_1 - \frac{k}{n}\right) + \frac{1}{2}\log\left(q_2 - \frac{k}{n}\right) \leq \log\left(\frac{q_1+q_2}{2} - \frac{k}{n}\right) \\ &\frac{1}{2}\log(q_1) + \frac{1}{2}\log(q_2) \leq \log\left(\frac{q_1+q_2}{2}\right) \end{aligned} \tag{78}$$

  which justifies (77).

- For $q_1, q_2 \leq \frac{k}{n}$:
  Again, we would like to show that $\psi(q_1, q_2) \leq \psi(\frac{q_1+q_2}{2}, \frac{q_1+q_2}{2})$. First, from the concavity of the log function, $q_1^{k-1} q_2^{k-1} \leq \left(\frac{q_1+q_2}{2}\right)^{2k-2}$. Therefore, we still need to show that $\left(q_1 - \frac{k}{n}\right)\left(q_2 - \frac{k}{n}\right) \leq \left(\frac{q_1+q_2}{2} - \frac{k}{n}\right)^2$. However, we have that

  $$\left(\frac{q_1+q_2}{2} - \frac{k}{n}\right)^2 - \left(q_1 - \frac{k}{n}\right)\left(q_2 - \frac{k}{n}\right) = \left(\frac{q_1-q_2}{2}\right)^2 \geq 0 \tag{79}$$

  which concludes the proof.

■

Appendix F

We have that

$$\begin{aligned}
&\sum_{k\in\mathcal{K}^{\mathcal{GT}}}\frac{1}{n^2}\binom{n}{k\ k}\sum_{u\neq v}P(u,v)\left(2k(2k+1)-n-4k(p(u)+p(v))+n^2(p(u)+p(v))^2\right)+\\
&\sum_{k\in\mathcal{K}^{\mathcal{ML}}}\binom{n}{k\ k}\sum_{u\neq v}\left(p(u)-\frac{k}{n}\right)\left(p(v)-\frac{k}{n}\right)p^k(u)p^k(v)(1-p(u)-p(v))^{n-2k}=\\
&\sum_{u\neq v}p(u)p(v)\Bigg(\sum_{k\in\mathcal{K}^{\mathcal{GT}}}\frac{1}{n^2}\binom{n}{k\ k}p^k(u)p^k(v)(1-p(u)-p(v))^{n-2k-2}\cdot\\
&\qquad\left(2k(2k+1)-n-4k(p(u)+p(v))+n^2(p(u)+p(v))^2\right)+\\
&\qquad\sum_{k\in\mathcal{K}^{\mathcal{ML}}}\binom{n}{k\ k}\left(p(u)-\frac{k}{n}\right)\left(p(v)-\frac{k}{n}\right)p^{k-1}(u)p^{k-1}(v)(1-p(u)-p(v))^{n-2k}\Bigg)\leq\\
&\sum_{u\neq v}p(u)p(v)\Bigg(\sum_{k\in\mathcal{K}^{\mathcal{GT}}}\frac{1}{n^2}\binom{n}{k\ k}p^k(u)p^k(v)(1-p(u)-p(v))^{n-2k-2}(n(p(u)+p(v))-2k)^2+\\
&\qquad\sum_{k\in\mathcal{K}^{\mathcal{ML}}}\binom{n}{k\ k}\left(p(u)-\frac{k}{n}\right)\left(p(v)-\frac{k}{n}\right)p^{k-1}(u)p^{k-1}(v)(1-p(u)-p(v))^{n-2k}\Bigg)\leq\\
&\max_{q_1,q_2\in\Delta_2}\Bigg(\sum_{k\in\mathcal{K}^{\mathcal{GT}}}\frac{1}{n^2}\binom{n}{k\ k}q_1^kq_2^k(1-q_1-q_2)^{n-2k-2}(n(q_1+q_2)-2k)^2+\\
&\qquad\sum_{k\in\mathcal{K}^{\mathcal{ML}}}\binom{n}{k\ k}\left(q_1-\frac{k}{n}\right)\left(q_2-\frac{k}{n}\right)q_1^{k-1}q_2^{k-1}(1-q_1-q_2)^{n-2k}\Bigg),
\end{aligned}\tag{80}$$

where the first inequality follows from

$$\begin{aligned}
&2k(2k+1)-n-4k(p(u)+p(v))+n^2(p(u)+p(v))^2=\\
&(n(p(u)+p(v))-2k)^2+2k-n<(n(p(u)+p(v))-2k)^2
\end{aligned}\tag{81}$$

as $2k<n$, and the last inequality follows from Proposition 2.

Appendix G - Proof of Proposition 10

We show the for every pair $q_1, q_2\in\Delta_2$, we have $\omega(q_1,q_2)\leq\omega\left(\frac{q_1+q_2}{2},\frac{q_1+q_2}{2}\right)$. Specifically,

$$\begin{aligned}
&q_1^kq_2^k(1-q_1-q_2)^{n-2k-2}(n(q1+q2)-2k)^2\leq\\
&\left(\frac{q_1+q_2}{2}\right)^{2k}(1-q_1-q_2)^{n-2k-2}(n(q1+q2)-2k)^2
\end{aligned}\tag{82}$$

and

$$\begin{aligned}
&\left(q_1-\frac{k}{n}\right)\left(q_2-\frac{k}{n}\right)q_1^{k-1}q_2^{k-1}(1-q_1-q_2)^{n-2k}\leq\\
&\left(\frac{q_1+q_2}{2}-\frac{k}{n}\right)^2\left(\frac{q_1+q_2}{2}\right)^{2k-2}(1-q_1-q_2)^{n-2k}
\end{aligned}\tag{83}$$

for every k and every $q_1, q_2 \in \Delta_2$. First, we notice that (82) holds since $q_1^k q_2^k \leq \left(\frac{q_1+q_2}{2}\right)^{2k}$, as shown in the proof of Proposition 12. Second, we observe that (83) holds, following Proposition 12. Therefore, $\omega(q_1, q_2) \leq \omega\left(\frac{q_1+q_2}{2}, \frac{q_1+q_2}{2}\right)$ as desired.

Appendix H

we have that

$$\begin{aligned}
&\sum_{k\in\mathcal{K}^{\mathcal{GT}}}\left(\frac{(k+1)^2}{n^2}\binom{n}{k+1}\sum_u p^{k+1}(u)(1-p(u))^{n-k-1}+\binom{n}{k}\sum_u p^{k+2}(u)(1-p(u))^{n-k}\right)+\\
&\sum_{k\in\mathcal{K}^{\mathcal{ML}}}\binom{n}{k}\sum_u\left(p(u)-\frac{k}{n}\right)^2 p^k(u)(1-p(u))^{n-k}=\\
&\sum_u p(u)\Bigg(\sum_{k\in\mathcal{K}^{\mathcal{GT}}}\frac{(k+1)^2}{n^2}\binom{n}{k+1}p^k(u)(1-p(u))^{n-k-1}+\\
&\qquad\sum_{k\in\mathcal{K}^{\mathcal{GT}}}\binom{n}{k}p^{k+1}(u)(1-p(u))^{n-k}+\\
&\qquad\sum_{k\in\mathcal{K}^{\mathcal{ML}}}\binom{n}{k}\left(p(u)-\frac{k}{n}\right)^2 p^{k-1}(u)(1-p(u))^{n-k}\Bigg)\leq\\
&\max_{q\in[0,1]}\sum_{k\in\mathcal{K}^{\mathcal{GT}}}\left(\frac{(k+1)^2}{n^2}\binom{n}{k+1}q^k(1-q)^{n-k-1}+\binom{n}{k}q^{k+1}(1-q)^{n-k}\right)+ \qquad (84)\\
&\qquad\sum_{k\in\mathcal{K}^{\mathcal{ML}}}\binom{n}{k}\left(q-\frac{k}{n}\right)^2 q^{k-1}(1-q)^{n-k}=\\
&\max_{q\in[0,1]}\sum_{k\in\mathcal{K}^{\mathcal{GT}}}\left(\frac{(k+1)^2}{n^2}\left(\frac{n-k}{k+1}\right)\frac{1}{1-q}+q\right)\mathrm{Bin}(k;n,q)+\sum_{k\in\mathcal{K}^{\mathcal{ML}}}\left(q-\frac{k}{n}\right)^2 q^{-1}\mathrm{Bin}(k;n,q)\leq\\
&\max_{q\in[0,1]}\sum_{k\in\mathcal{K}^{\mathcal{GT}}}\left(\frac{(k+1)^2}{n^2}\left(\frac{n-k}{k+1}\right)\frac{1}{1-q}+q\right)\frac{1}{\sqrt{2\pi k(1-k/n)}}\exp(-nD_{KL}(k/n||q))+\\
&\qquad\sum_{k\in\mathcal{K}^{\mathcal{ML}}}\left(q-\frac{k}{n}\right)^2 q^{-1}\frac{1}{\sqrt{2\pi k(1-k/n)}}\exp(-nD_{KL}(k/n||q)),
\end{aligned}$$

where the first inequality follows from Proposition 3 and the last inequality from (34).

Appendix I

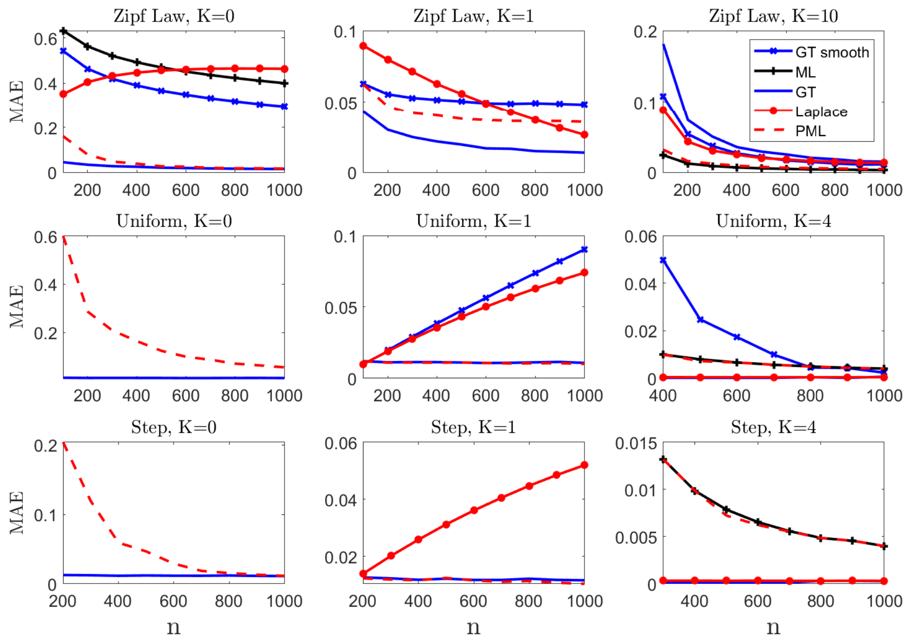

Figure 6: MAE of OP estimation in three synthetic experiments

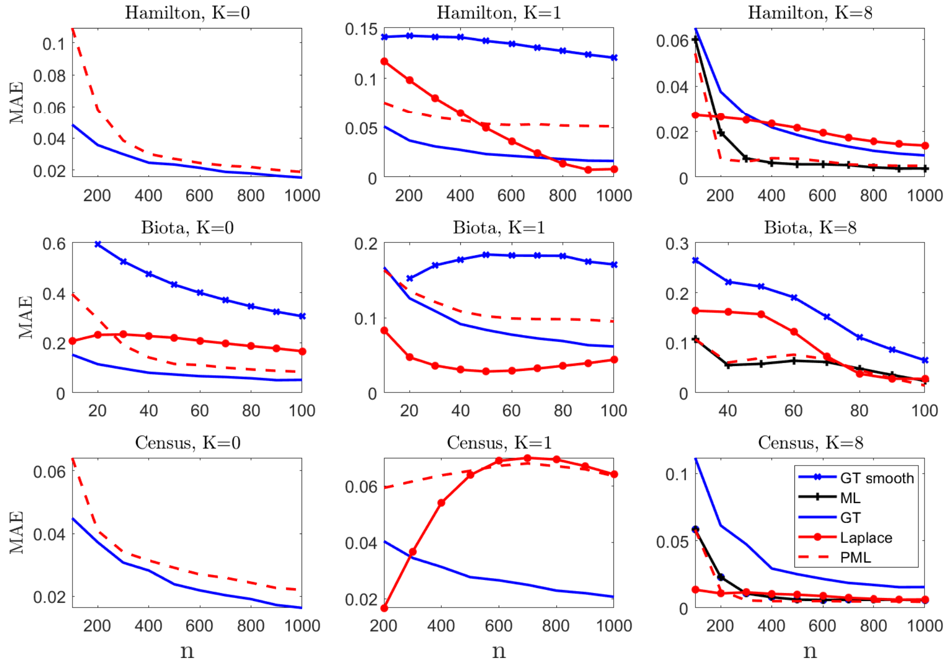

Figure 7: MAE of OP estimation in three real-world experiments

References

Jayadev Acharya, Yelun Bao, Yuheng Kang, and Ziteng Sun. Improved bounds for minimax risk of estimating missing mass. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 326–330. IEEE, 2018.

Nima Anari, Moses Charikar, Kirankumar Shiragur, and Aaron Sidford. Instance based approximations to profile maximum likelihood. *Advances in neural information processing systems*, 33:20272–20285, 2020.

Marco Battiston, Fadhel Ayed, Federico Camerlenghi, and Stefano Favaro. On consistent and rate optimal estimation of the missing mass. In *Annales de l'institut Henri Poincare (B) Probability and Statistics*, 2020.

Anna Ben-Hamou, Stéphane Boucheron, and Mesrob I Ohannessian. Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications. *Bernoulli*, 23:249–287, 2017.

Daniel Berend and Aryeh Kontorovich. On the concentration of the missing mass. *Electronic Communications in Probability*, 18, 2013.

Cristian Budianu and Lang Tong. Good-Turing estimation of the number of operating sensors: a large deviations analysis. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages ii–1029. IEEE, 2004.

US Census Bureau. Frequently occurring surnames from the census 2000. 2014.

Prafulla Chandra, Aditya Pradeep, and Andrew Thangaraj. Improved tail bounds for missing mass and confidence intervals for Good-Turing estimator. In *2019 National Conference on Communications*, pages 1–6. IEEE, 2019.

Anne Chao. On estimating the probability of discovering a new species. *The Annals of Statistics*, pages 1339–1342, 1981.

Stanley F Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–394, 1999.

Kenneth W Church and William A Gale. A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of english bigrams. *Computer Speech & Language*, 5(1):19–54, 1991.

Shir Cohen, Tirza Routtenberg, and Lang Tong. Non-bayesian parametric missing-mass estimation. *IEEE Transactions on Signal Processing*, 2022.

Geoffrey Decrouez, Michael Grabchak, and Quentin Paris. Finite sample properties of the mean occupancy counts and probabilities. *Bernoulli*, 24(3):1910–1941, 2018.

Evgeny Drukh and Yishay Mansour. Concentration bounds for unigram language models. *Journal of Machine Learning Research*, pages 1231–1264, 2005.

Bradley Efron and Ronald Thisted. Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, 63(3):435–447, 1976.

Stefano Favaro, Antonio Lijoi, and Igor Prünster. A new estimator of the discovery probability. *Biometrics*, 68(4):1188–1196, 2012.

Stefano Favaro, Bernardo Nipoti, and Yee Whye Teh. Rediscovery of Good-Turing estimators via bayesian nonparametrics. *Biometrics*, 72(1):136–145, 2016.

Ronald A Fisher, A Steven Corbet, and Carrington B Williams. The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology*, pages 42–58, 1943.

William Gale and Kenneth Church. What's wrong with adding one. *Corpus-Based Research into Language: In honour of Jan Aarts*, pages 189–200, 1994.

William A Gale and Geoffrey Sampson. Good-Turing frequency estimation without tears. *Journal of quantitative linguistics*, 2(3):217–237, 1995.

Fuqing Gao et al. Moderate deviations for a nonparametric estimator of sample coverage. *The Annals of Statistics*, 41(2):641–669, 2013.

Zhan Gao, Chi-hong Tseng, Zhiheng Pei, and Martin J Blaser. Molecular analysis of human forearm superficial skin bacterial biota. *Proceedings of the National Academy of Sciences*, 104(8):2927–2932, 2007.

Irving J Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40:237–264, 1953.

Irving J Good and George H Toulmin. The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika*, 43(1-2):45–63, 1956.

Michael Grabchak and Zhiyi Zhang. Asymptotic properties of turing's formula in relative error. *Machine Learning*, 106(11):1771–1785, 2017.

Yi Hao and Alon Orlitsky. The broad optimality of profile maximum likelihood. *Advances in Neural Information Processing Systems*, 32, 2019.

Biing-Hwang Juang and SH Lo. On the bias of the Turing-Good estimate of probabilities. *IEEE transactions on signal processing*, 42(2):496–498, 1994.

Samuel Karlin. Central limit theorems for certain infinite urn schemes. *Journal of Mathematics and Mechanics*, 17(4):373–401, 1967.

Raphail Krichevsky and Victor Trofimov. The performance of universal encoding. *IEEE Transactions on Information Theory*, 27(2):199–207, 1981.

Pierre-Simon Laplace. *Pierre-Simon Laplace Philosophical Essay on Probabilities: Translated from the fifth French edition of 1825 With Notes by the Translator*, volume 13. Springer Science & Business Media, 1825.

Antonio Lijoi, Ramsés H Mena, and Igor Prünster. Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika*, 94(4):769–786, 2007.

Chang Xuan Mao and Bruce G Lindsay. A poisson model for the coverage problem with a genomic application. *Biometrika*, 89(3):669–682, 2002.

David A McAllester and Robert E Schapire. On the convergence rate of Good-Turing estimators. In *COLT*, pages 1–6, 2000.

Dragoslav S Mitrinovic and Petar M Vasic. *Analytic inequalities*, volume 61. Springer, 1970.

Elchanan Mossel and Mesrob Ohannessian. On the impossibility of learning the missing mass. *Entropy*, 21(1):28, 2019.

Arthur Nadas. Good, jekinek, mercer, and robbins on turing's estimate of probabilities. *American Journal of Mathematical and Management Sciences*, 11(3-4):299–308, 1991.

Mesrob I Ohannessian and Munther A Dahleh. Rare probability estimation under regularly varying heavy tails. In *Conference on Learning Theory*, pages 21–1. JMLR Workshop and Conference Proceedings, 2012.

Alon Orlitsky and Ananda Theertha Suresh. Competitive distribution estimation: Why is Good-Turing good. In *Advances in Neural Information Processing Systems*, pages 2143–2151, 2015.

Alon Orlitsky, Narayana P Santhanam, and Junan Zhang. Always Good Turing: Asymptotically optimal probability estimation. *Science*, 302(5644):427–431, 2003.

Alon Orlitsky, S Sajama, Narayana P Santhanam, Krishnamurthy Viswanathan, and Junan Zhang. Algorithms for modeling distributions over large alphabets. In *International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings.*, pages 304–304. IEEE, 2004a.

Alon Orlitsky, Narayana P Santhanam, Krishnamurthy Viswanathan, and Junan Zhang. On modeling profiles instead of values. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 426–435, 2004b.

Alon Orlitsky, Narayana P Santhanam, and Junan Zhang. Universal compression of memoryless sources over unknown alphabets. *IEEE Transactions on Information Theory*, 50 (7):1469–1481, 2004c.

Alon Orlitsky, Ananda Theertha Suresh, and Yihong Wu. Optimal prediction of the number of unseen species. *Proceedings of the National Academy of Sciences*, 113(47):13283–13288, 2016.

Amichai Painsky. Refined convergence rates of the Good-Turing estimator. In *2021 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE, 2021.

Amichai Painsky. Generalized Good-Turing improves missing mass estimation. *Journal of the American Statistical Association*, pages 1–10, 2022.

Amichai Painsky and Gregory Wornell. On the universality of the logistic loss function. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 936–940. IEEE, 2018.

Amichai Painsky and Gregory W Wornell. Bregman divergence bounds and universality properties of the logarithmic loss. *IEEE Transactions on Information Theory*, 66(3): 1658–1673, 2019.

Dmitri S Pavlichin, Jiantao Jiao, and Tsachy Weissman. Approximate profile maximum likelihood. *J. Mach. Learn. Res.*, 20:122–1, 2019.

Yurii Vasil'evich Prokhorov. Asymptotic behavior of the binomial distribution. *Uspekhi Matematicheskikh Nauk*, 8(3):135–142, 1953.

Nikhilesh Rajaraman, Andrew Thangaraj, and Ananda Theertha Suresh. Minimax risk for missing mass estimation. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 3025–3029. IEEE, 2017.

Herbert Robbins. A remark on Stirling's formula. *The American mathematical monthly*, 62 (1):26–29, 1955.

Alexander I Saichev, Yannick Malevergne, and Didier Sornette. *Theory of Zipf's law and beyond*, volume 632. Springer Science & Business Media, 2009.

Maciej Skorski. On missing mass variance. *arXiv preprint arXiv:2104.07028*, 2021.

Ronald Thisted and Bradley Efron. Did shakespeare write a newly-discovered poem? *Biometrika*, 74(3):445–455, 1987.

Cun-Hui Zhang. Estimation of sums of random variables: examples and information bounds. *The Annals of Statistics*, 33(5):2022–2041, 2005.

Zhiyi Zhang and Hongwei Huang. Turing's formula revisited. *Journal of Quantitative Linguistics*, 14(2-3):222–241, 2007.