Optimal Prediction of the Number of Unseen Species with Multiplicity

Yi Hao [§] , **Ping Li**
Cognitive Computing Lab
Baidu Research
10900 NE 8th St. Bellevue, WA 98004, USA
yih179@eng.ucsd.edu , pingli98@gmail.com

Abstract

Based on a sample of size n, we consider estimating the number of symbols that appear at least μ times in an independent sample of size $a \cdot n$, where a is a given parameter. This formulation includes, as a special case, the well-known problem of inferring the number of unseen species introduced by [Fisher et al.] in 1943 and considered by many others. Of considerable interest in this line of works is the largest a for which the quantity can be accurately predicted. We completely resolve this problem by determining the limit of estimation to be $a \approx (\log n)/\mu$, with both lower and upper bounds matching up to constant factors. For the particular case of $\mu = 1$, this implies the recent result by [Orlitsky et al.] on the unseen species problem. Experimental evaluations show that the proposed estimator performs exceptionally well in practice. Furthermore, the estimator is a linear combination of symbols' empirical counts, and hence linear-time computable.

1 Introduction

Let Δ denote the collection of all discrete distributions. Given an independent sample X^n from an unknown distribution $p \in \Delta$, we are interested in estimating the number of *unseen* symbols that appear *at least μ times* in an independent sample Y^m from the same distribution p, where m is a given sampling parameter. To be more specific, we study the estimation of

$$U_\mu \triangleq U_\mu(X^n, Y^m) \triangleq \sum_x \mathbb{1}_{N_x=0} \cdot \mathbb{1}_{M_x \geq \mu},$$

where the summation is over all symbols x in the potentially unknown alphabet, and N_x and M_x represent the number of times symbol x appearing in X^n and Y^m, respectively. In addition, let

$$\Phi_i \triangleq \Phi_i(X^n) \triangleq \sum_x \mathbb{1}_{N_x=i}$$

denote the number of symbols appearing i times in the sample X^n. Due to symmetry, the collection of Φ_i's is a *sufficient statistic* for learning U_μ.

For the special case of $\mu = 1$, the quantity U_1 becomes the number of newly observed symbols, and the task reduces to the well-known *unseen species problem*, whose study dates back to more than half a century ago [12]. In the early 1940s, the British chemist and naturalist Corbet had spent two years in Malaya to trap butterflies (X^n). Every time he saw a new species, he recorded how many individuals of that species he had trapped (N_x).

[§]Yi Hao is a Ph.D. candidate in the Department of Electrical and Computer Engineering at the University of California, San Diego. Yi Hao's work was conducted while he was a summer intern at Baidu Research-Bellevue.

After returning to England, Corbet was curious how many new species he would discover if he went back to Malaya for another 2 years. He constructed a table to show the number of species appearing certain number of times (Φ_i), and then presented the problem to Fisher, Father of modern statistics, who provided a remarkable estimator [12] that was later improved on by Good and Toulmin [14],

$$\hat{U}_{\mathrm{GT}} \triangleq - \sum_{i=1}^{n} (-1)^i \Phi_i.$$

Before continuing the discussion, we introduce a convenient notation, $a \triangleq \frac{m}{n}$, and refer to it as the *amplification ratio (or extrapolation ratio)*, a quantity that describes the limit of our estimation power and frequently appears in our subsequent derivations. Henceforth we often write m as na to reflect our intention of inferring future statistics from past observations.

The Good-Toulmin estimator $\hat{U}_{\mathrm{GT}}$ applies to the special case of $a = \mu = 1$ and generalizes to

$$\hat{U}_{\mathrm{GT}} \triangleq - \sum_{i=1}^{n} (-a)^i \Phi_i$$

for general a values, where we kept the same notation. The paper [14] that originally introduced $\hat{U}_{\mathrm{GT}}$ showed that as long as $a \leq 1$, the estimator produces a nearly unbiased estimate with an expected (absolute) deviation of $\mathcal{O}(a\sqrt{n})$, which is negligible as U_1 can potentially be an.

However, the success of this estimator does not extend to the regime where $a > 1$, both in theory and in practice. For example, an experiment in a recent work [29] addressing U_1 estimation shows that even for a simple (shifted) Zipf-law with $p_i \propto 1/(i + 10)$ for $1 \leq i \leq 10^4$, and $a \in [1.3, 1.5]$, estimate $\hat{U}_{\mathrm{GT}}$ can significantly deviate from the actual value and sometimes even becomes negative.

To address this issue, Good and Toulmin applied a smoothing technique to the alternating coefficient sequence $(-a)^i, i = 1, \ldots n$, and replaced it by $(-a)^i \cdot \Pr(\mathrm{bin}(k, 1/(a+1)) \geq i)$ for some properly chosen parameter k. The resulting estimator, which [29] terms as the *smoothed Good-Toulmin (SGT)*, has found numerous applications over the past half-century. However, the statistical properties and optimality of this estimator were not well-understood until the recent work of Orlitsky et al. [29, 31].

Noting that the heuristic multiplies the coefficients by the binomial tail probabilities, Orlitsky et al. first proposed the following general smoothing regime. Let L be an independent nonnegative random variable, and denote $\hat{U}_L \triangleq - \sum_{i=1}^{n} (-a)^i \Pr(L \geq i) \Phi_i$. Subsequently, they showed that for L being either a (proper) binomial or Poisson random variable, the induced estimator performs optimally.

Formally and more generally, U_μ varies from 0 to $m_\mu \triangleq m/\mu = na/\mu$, hence we measure the performance of any n-sample estimator $\hat{U}$ by the *worst-case normalized mean-square error* (NMSE),

$$\mathcal{E}_{n,a}^\mu(\hat{U}) \triangleq \max_{p \in \Delta} \mathbb{E}_p \left(\frac{\hat{U} - U_\mu}{m_\mu} \right)^2.$$

The main result of [29] states that for some Poisson or binomially distributed L, the estimator satisfies $\mathcal{E}_{n,a}^1(\hat{U}_L) \lesssim 1/n^{1/a}$, where $\alpha \lesssim \beta$ abbreviates $\alpha = O(\beta)$. On the other hand, for some absolute constants $c, c' > 0$, and any $a \geq c$ and estimator $\hat{U}$, one has $\mathcal{E}_{n,a}^1(\hat{U}_L) \gtrsim 1/n^{c'/a}$, where $\alpha \gtrsim \beta$ abbreviates $\alpha = \Omega(\beta)$. Combined, the results established the optimality of the smoothing scheme.

1.1 New results

As described previously, the problem that we study here is natural generalization of the unseen species problem, and calls for estimating U_μ, the number of symbols that appear at least μ times in the future sample. One motivation for considering this problem is reproducibility. For example, in the aforementioned butterfly trapping story, a basic task one often wants to perform is checking the existence of *sexual dimorphism* in a newly observed butterfly species, meaning that the two sexes look completely different. This clearly requires inferring U_μ for $\mu \geq 2$.

Replacing butterflies by words, *vocabulary size estimation* [9, 11, 19, 35] aims to determine how many words a writer, say William Shakespeare, knew based on his written works. An intuitive and widely used approach is to simply add up the number of observed (distinct) words and some estimate

of U_1. With the same motivation, we may also want to know how many words fall into a writer's *common vocabulary* (excluding those that appear only once or twice), which calls for estimating U_μ.

For another example, app developers are often interested in knowing how many new users their apps will have in a future time period. In addition, they usually care more about active users who will use the apps for at least a certain number of times. Under appropriate assumptions, this again translates to a U_μ estimation problem. The same rationale applies to many other types of businesses such as advertising, catering, and entertainment, since U_μ is a natural business growth indicator.

There has been a long line of research works on estimating U_1, in which of considerable interest is the largest a for which the quantity can be accurately predicted. The generalization of this "unseen species" problem, on the other hand, is a new problem that we propose and rigorously study. As the subsequent discussion shows, we completely resolved this problem by determining the limit of estimation to be $a \approx (\log n)/\mu$, with both lower and upper bounds matching up to constant factors.

Our estimator is linear-time computable given Φ_i's, and has the form

$$\hat{U}_\mu \triangleq \hat{U}_\mu(X^n, a) \triangleq \sum_{i=1}^{n} s_i \cdot \Phi_i,$$

where for $r \sim \frac{\log n}{a}$, the *smoothing rate*, we denote the *i-th smoothed coefficient* by

$$s_i \triangleq - \sum_{j=0}^{(\mu-1)\wedge i} (-a)^i (-1)^j \binom{i}{j} \Pr(\mathrm{Poi}(r) \geq i+j),$$

where the notation $\alpha \wedge \beta$ abbreviates $\min\{\alpha, \beta\}$. In particular, if we set $\mu = 1$, the estimator reduces to $\hat{U}_L$ in [29] with $L \sim \mathrm{Poi}(r)$. Interestingly, if we remove the smoothing probability, the estimator becomes the vanilla Good-Toulmin estimator $\hat{U}_{\mathrm{GT}}$ when $\mu = 1$, and becomes the following "generalized Good-Toulmin" estimator (denoted by $\hat{U}_{\mathrm{GGT}}$) when $\mu \geq 1$:

$$\hat{U}_{\mathrm{GGT}} \triangleq (-1)^\mu \sum_{i=\mu}^{n} (-a)^i \binom{i-1}{\mu-1} \Phi_i.$$

The deviation of $\hat{U}_{\mathrm{GGT}}$ can be found in Appendix J.

The following two results essentially establish the optimality of $\hat{U}_\mu$ and determine the min-max learning risk of approximating U_μ. In particular, these theorems imply the main result of [29] as a special case. The first theorem bounds the worst-case NMSE of our estimator.

Theorem 1. *There exist absolute constants c and c_0 such that for any parameter $a \in [1, (c \log n)/\mu]$, the estimator $\hat{U}_\mu$ described above satisfies*

$$\mathcal{E}_{n,a}^\mu(\hat{U}_\mu) \lesssim \frac{1}{n^{c_0/a}}.$$

The second theorem lowerly bounds the worst-case NMSE of the best estimator.

Theorem 2. *There exist absolute constants c' and c_0' such that for any $a \geq \max\{1, (c' \log n)/\mu\}$, and any n-sample estimator $\hat{U}$,*

$$\mathcal{E}_{n,a}^\mu(\hat{U}) \gtrsim \frac{1}{n^{c_0'/a}}.$$

1.2 Related Work

Unseen species problem Initiated by the seminal work of [12, 14], species estimation is an important problem in a variety of scientific disciplines. Over the past few decades, the problem has found numerous applications to the estimation of ecological or bacterial diversity [4, 5, 7, 10], bacterial and microbial diversity [13, 28, 30], and writers' vocabulary sizes [11, 35, 9], and the study of database attribute variations [15], immune receptors [32], and genetic variations [24].

A number of estimators were proposed, notable ones include the Good-Toulmin (GT) estimator [14, 29, 31], jackknife estimator [33], abundance coverage estimator (ACE) [6], Chao-Lee estimator [5, 7], and a linear-programming based (LP) estimator [36]. Even just for this U_1 estimation task, the GT estimator and their variants are the only ones known [29, 31] to achieve the optimal worst-case NMSE. The ACE and Chao-Lee estimators are designed for uniform distributions, and LP estimator has exponentially worse NMSE guaranty and runs in polynomial time, inefficient compared with GT or others in [29]. Another way of generalizing this unseen species problem is to consider the setting involving multiple species groups (hence multiple distributions). Paper [17] recently studied this problem as an application of its proposed multi-distribution functional estimation methodology.

Functional estimation   In recent years, a related area that has seen significant advances is distribution functional estimation. For example, one may want to infer the entropy of an unknown distribution from its i.i.d. sample. Other important functionals include support size, L_1 distance to a fixed distribution, general Lipschtiz functionals, and *support coverage*, the expected number of distinct symbols in a new sample. Two types of methods have been proposed: one first finds an approximation of the targeted functional, then find a near-unbiased estimator for the proxy [20, 21, 23, 26, 29, 39, 40]; the other focuses on symmetric functionals, and tackles the problem by computing an estimate of the distribution probability multiset and plugging it into the functional [2, 8, 16, 18, 22, 27, 36, 37].

Our approach essentially falls into the first category, and is closer to [17, 23, 29] utilizing smoothing methods based on Bessel functions, with which we compare carefully in Section 3. Other methods either require assumptions not satisfied in our setting, such as Lipschitzness [21, 18] or the existence of highly concentrated estimators [2, 18]; or not known to work (well) even for U_1 estimation, such as those based on minimax polynomials [26, 27, 39, 40] or linear programming [16, 36, 37] (see above).

Outline   For the rest of the paper, we mainly focus on constructing the estimator and deriving the upper bound. The proofs of the lower bound and technical lemmas are postponed to the supplementary. In Section 2, we present a roadmap for our construction and analysis, and show that Poissonization works well in simplifying the problem. In Section 3, we utilize connections between Bessel functions and Poisson probabilities to approximate the expectation of U_μ, and introduce r as a tunable hyper-parameter controlling the variance-bias tradeoff. Section 4 then finds the explicit form of the estimator and outlines the tools and key lemmas for the bias analysis. Finally, we numerically evaluate the performance of the estimator in Section 5, demonstrating the excellent practicality of our approach.

2   MSE Decomposition and Poisson Sampling

First we address the upper bound in Theorem 1, decompose the problem into several parts, and simplify our reasoning with *Poisson sampling*, a technique to remove symbol-count dependence. Given n, a, and μ, the normalizing factor $M_\mu = na/\mu$ is simply a constant. Hence, it suffices to bound the mean squared error (MSE) of the estimator. On a high level, our analysis follows by

$$\text{MSE} \triangleq \mathbb{E}\left(\hat{U}_\mu - U_\mu\right)^2 \leq 2\text{Var}(\hat{U}_\mu) + 2\text{Var}\left(U_\mu\right) + \left(\mathbb{E}[\hat{U}_\mu] - \mathbb{E}\left[U_\mu\right]\right)^2,$$

based on which we construct $\hat{U}_\mu$ and bound each term in the last line separately.

Variances   As one might expect, a difficulty in analyzing the variance is caused by dependence. Given a sample X^n from some unknown distribution p, our estimator takes the form of $\sum_i s_i \Phi_i$ where Φ_i denotes the number of symbols appearing i times in the sample. While X_i's are independent by our assumption, Φ_i's are dependent random variables, and putting involved coefficients in front of them certainly makes the analysis even harder. A similar difficulty appears when we consider the target random variable U_μ. To get around these obstacles, we apply Steele's inequality [34].

Lemma 1 (Steele's inequality)**.** *If $S(x^n)$ is any real function of n variables and if $X_i, X_i', 1 \leq i \leq n$ are $2n$ i.i.d. random variables then $\text{Var}(S) \leq \frac{1}{2}\mathbb{E}\left[\sum_{i=1}^n (S - S_i)^2\right]$, where $S \triangleq S(X^n)$ and S_i is given by replacing the i-th observation with X_i'.*

The inequality basically states that as long as we have i.i.d. input X^n and care only a functional $S(X^n)$ *that is not too sensitive to input changes*, the variance of $S(X^n)$ will be reasonably bounded. For example, U_μ is a functional of the length-$(n+m)$ i.i.d. sample (X^n, Y^m), where modifying a

single symbol changes U_μ by at most 1. Hence, Lemma 1 bounds its variance as

$$\mathrm{Var}(U_\mu) \leq \frac{1}{2}(n+m).$$

In the supplementary material, we derive a similar bound of $\mathrm{Var}(\hat{U}_\mu) \leq 2n \max_i s_i^2$ together with $|s_i| \leq e^{(2a-1)r}$ under appropriate assumptions, which requires significantly more work.

Absolute bias   Analyzing the bias term $\mathbb{E}[\hat{U}_\mu - U_\mu]$ directly relates to the estimator construction. Intuitively, we first find the exact form of $\mathbb{E}[U_\mu]$, which induces to an unbiased n-sample estimator. Then, we gradually increase the absolute bias, with an intention to reduce the variance (or the coefficient differences by the previous reasoning). The smoothing rate r appearing in the coefficient expressions is a tunable hyper-parameter for this variance-bias tradeoff. In fact, this is the only hyper-parameter for our algorithm, which implies an implementation advantage. See Section 3 for details on how we approximate the target quantity and how r comes into the picture, and Section 4 for how our approximation induces estimator $\hat{U}_\mu$ and how we bound its bias.

To simplify our bias analysis, we also consider a closely related statistical model where the corresponding sample sizes are respectively $N \sim \mathrm{Poi}(n)$ and $N' \sim \mathrm{Poi}(m)$, independent of the samples. This leads to a technique in statistical learning, commonly known as *Poisson sampling*. The reason for employing this *Poissonization* procedure is that it eliminates the dependence between sample counts and turns a binomial probability $\binom{b}{a}q^a(1-q)^{b-a}$ to the simpler Poisson version $e^{-bq}(bq)^a/a!$.

Henceforth, we indicate Poissonization by attaching a letter ᴘ to an expression, for example, U_μ^{P} for U_μ, and $\hat{U}_\mu^{\mathrm{P}}$ for $\hat{U}_\mu$, under the Poisson model. The following lemma partially justifies our reasoning.

Lemma 2. *For any n, a and p, we have* $\left|\mathbb{E}[U_\mu - U_\mu^{\mathrm{P}}]\right| \leq 4$.

A similar result holds for the estimator part.

Lemma 3. *Under the conditions in Theorem 1, we have* $|\mathbb{E}[\hat{U}_\mu - \hat{U}_\mu^{\mathrm{P}}]| \leq 2e^{(2a-1)r}$.

As we show latter in the paper, this bound is also tiny compared with the normalizing factor m_μ.

3   Bessel Functions and Approximation of Unseen

Following the reasoning in the last section, Poissonizing the samples facilitates our derivation without affecting our results. In this section, we will assume that the sample sizes are $\mathrm{Poi}(n)$ and $\mathrm{Poi}(m)$ for the past and future samples. To not further complicate the notation, we still use N_x's and M_x's to denote the counts of symbols in the samples. An important point to note is that now they are all *mutually independent* Poisson random variables.

We introduce a new notation $\lambda_x \triangleq np_x$ for every symbol x. Now, we can write $N_x \sim \mathrm{Poi}(\lambda_x)$ and $M_x \sim \mathrm{Poi}(a\lambda_x)$. For the target quantity U_μ^{P}, taking expectation yields

$$\mathbb{E}\left[U_\mu^{\mathrm{P}}\right] = \sum_x \mathbb{E}\left[\mathbb{1}_{M_x \geq \mu}\right] \cdot \mathbb{E}\left[\mathbb{1}_{N_x=0}\right]$$

$$= \sum_x \left(1 - e^{-a\lambda_x} \sum_{j \leq \mu-1} \frac{(a\lambda_x)^j}{j!}\right) \cdot e^{-\lambda_x}$$

$$\triangleq \sum_x F_\mu(a\lambda_x) \cdot e^{-\lambda_x},$$

where the definition of function F_μ is clear from the context. Leveraging Lemma 1 and the standard exponential tail bounds on Poisson variables, we can show that U_μ^{P} has small variance. In other words, it suffices to approximate the above expectation. Next, express function $F_\mu(y)$ into the summation of μ terms, with each being a polynomial of e^{-y} and y that vanishes at $y=0$,

$$F_\mu(y) = (1 - e^{-y}) + \sum_{1 \leq j \leq \mu-1}\left(-\frac{e^{-y}y^j}{j!}\right).$$

5

Bessel functions We approximate terms in the parentheses by polynomial series induced by the Bessel functions. For completeness, we briefly introduce this function class. The *Bessel functions of the first kind* $J_s(y)$ are defined as the solutions to the Bessel differential equation

$$y^2 \frac{d^2 z}{dy^2} + y \frac{dz}{dy} + (y^2 - s^2)z = 0,$$

for $s \in \mathbb{Z}^+$, which is referred to as the *order* of the function.

For every s, function $J_s(y)$ admits a series expansion [25] at $y = 0$,

$$J_s(y) = \sum_{t=0}^{\infty} \frac{(-1)^t}{t!(t+s)!} \left(\frac{y}{2}\right)^{2t+s}.$$

For our purpose, it is sufficient to use only the even order Bessel functions. For $j \geq 0$, denote

$$g_j(y) \triangleq J_{2j}(2\sqrt{y}) = \sum_{t=0}^{\infty} \frac{(-1)^t y^{t+j}}{t!(t+2j)!}.$$

As we mentioned in Section 1, the problem reduces to the unseen species problem for $\mu = 1$. Hence, we follow the smoothing technique in [29] and approximate the first term $(1 - e^{-y})$ by

$$f_0(y) \triangleq (1 - e^{-y}) - e^{-y} \left[e^{-r} \int_0^y g_0(sr)e^s ds \right] = -\sum_{j=1}^{\infty} \frac{(-y)^j}{j!} \Pr(\mathrm{Poi}(r) \geq j),$$

where r is the aforementioned tunable hyper-parameter to balance the variance and bias. For the j-th term in the expression of F_μ, we follow the technique in [23] and approximate every $-e^{-y}y^j$ by

$$f_j(y) \triangleq -e^{-y}y^j + \int_r^{\infty} e^{-\alpha}\alpha^j g_j(\alpha y)d\alpha = -\int_0^r e^{-\alpha}\alpha^j g_j(\alpha y)d\alpha,$$

where the second equality follows by the lemma below.

Lemma 4. *[23] For any $j \in \mathbb{Z}^+$ and $y \geq 0$,*

$$e^{-y}y^j = \int_0^{\infty} e^{-\alpha}\alpha^j g_j(\alpha y)d\alpha.$$

In other words, function $f_i(y)$ corresponds to the truncated integral form of $e^{-y}y^j$ at level r, where the integral is expressed in terms of the Bessel functions.

It is worth mentioning that the result in [23] (induced by the use of Bessel functions) and that in our paper are orthogonal to a certain extend. The result in [23] addresses the problem of estimating a distribution functional that is roughly Lipschitz, and designs an n-sample estimator that performs as well as the $n\sqrt{\log n}$-sample MLE estimator. The problem we consider here is estimating a random distribution functional that depends on the current sample. Even if we consider the functional's expectation and divide it by our normalization factor m_μ, the transformed functional is not Lipschitz. In addition, our extrapolation factor a can be as large as $(\log n)/\mu$, which, for constant μ values, is of order $\log n$. Paper [17] generalizes the results of [23] to functionals involving multiple distributions, such as the total variation distance, but does not strengthen those results for the one-distribution case.

Another paper that studies the same problem as [23] is [21]. In terms of techniques, [21] leverages integral forms of the best approximation polynomials, and does not rely on smoothing techniques involving Bessel functions. In addition, [21] applies its smoothing technique to only probabilities smaller than $(\log n)/n$, while we apply the aforementioned technique to the entire range $[0, 1]$. In terms of results, [21] requires the functional to be Lipschitz, which is not the case for our problem, and improves the $\sqrt{n}$ factor in [23] to $\log n$, which is not the right factor for our problem. The paper also addresses the problem of estimating $\mathbb{E}[U_1]$, while the estimator is essentially the same as [29]. As the above derivation shows, the corresponding technique only approximates $(1 - e^{-y})$, and by the reasoning in [23], does not address other $e^{-y}y^j$ terms.

Approximation function Consequently, we will approximate $F_\mu(y)$ by

$$\tilde{F}_\mu(y) \triangleq \sum_{j \leq \mu - 1} \frac{f_j(y)}{j!},$$

which implicitly depends on r and approaches $F_\mu(y)$ as $r \to \infty$. Our estimator simply estimates a variant of this function nearly unbiasedly. Intuitively, if r is large, the bias is small, e.g., for $r = \infty$, the bias is zero, but the variance, as shown by the example in Section 1, can be fairly large. On the other hand, if r is small, e.g., for $r = 0$, the variance is brought down to zero but the bias is large.

4 Estimator Construction and Bias Analysis

Approximation and estimator   Following the previous section, we approximate the value of U_μ^{P} by a *near-unbiased* estimator of

$$\tilde{U}_\mu^{\mathrm{P}} \triangleq \sum_x \tilde{F}_\mu(a\lambda_x) \cdot e^{-\lambda_x}.$$

The statistics available to us are essentially indicator functions $\mathbb{1}_{N_x = t}$, or $\frac{\lambda_x^t}{t!} \cdot e^{-\lambda_x}$ in expectation. This suggests we expand each term in the summation as the product of $e^{-\lambda_x}$ and a polynomial series. Equivalently, we find the series expansion of $\tilde{F}_\mu(y)$.

Lemma 5. *For any non-negative number y,*

$$\tilde{F}_\mu(y) = -\sum_{i=1} c_i \cdot \frac{y^i}{i!},$$

where the coefficients are defined as

$$c_i \triangleq (-1)^i \Pr(\mathrm{Poi}(r) \geq i) + \sum_{j=1}^{(\mu-1)\wedge i} (-1)^{i-j} \binom{i}{j} \Pr(\mathrm{Poi}(r) \geq i + j + 1).$$

We postpone the proof of Lemma 5 to the supplementary material, which relies on Fubini's theorem and the series expansions of the incomplete Gamma functions centered at the origin. Next, for the target quantity $\tilde{U}_\mu$, substituting each $\tilde{F}_\mu$ in the expression in Lemma 5 yields

$$\tilde{U}_\mu^{\mathrm{P}} = \sum_x \tilde{F}_\mu(a\lambda_x) \cdot e^{-\lambda_x} = -\sum_x \sum_{i=1}^\infty (c_i a^i) \cdot e^{-\lambda_x} \frac{\lambda_x^i}{i!} = -\sum_{i=1}^\infty (c_i a^i) \cdot \mathbb{E}[\Phi_i].$$

It should be clear that $s_i \approx c_i a^i$ for every i as they are equal if we replace each smoothing probability $\Pr(\mathrm{Poi}(r) \geq i + j + 1)$ by $\Pr(\mathrm{Poi}(r) \geq i + j)$ in the expression of c_i. Hence, given sample statistics Φ_i for $i \geq 1$, a *near-unbiased* estimator of $\tilde{U}_\mu^{\mathrm{P}}$ is

$$\hat{U}_\mu^{\mathrm{P}} = -\sum_{i=1}^\infty s_i \cdot \Phi_i.$$

More concretely, the absolute bias satisfies $|\mathbb{E}[\hat{U}_\mu^{\mathrm{P}}] - \tilde{U}_\mu^{\mathrm{P}}| \leq na \cdot \Pr(\mathrm{Poi}(r) \leq \mu)$, which is proved in the supplementary material leveraging point-wise bounds on Bessel functions.

Bias analysis   Unifying the previous bounds by the triangle inequality, the absolute bias of the final estimator $\hat{U}_\mu$ admits the decomposition

$$\mathrm{Bias} \triangleq \left| \mathbb{E}[\hat{U}_\mu] - \mathbb{E}[U_\mu] \right| \leq \left| \mathbb{E}[\hat{U}_\mu] - \mathbb{E}[\hat{U}_\mu^{\mathrm{P}}] \right| + \left| \mathbb{E}[\hat{U}_\mu^{\mathrm{P}}] - \tilde{U}_\mu^{\mathrm{P}} \right| + \left| \tilde{U}_\mu^{\mathrm{P}} - \mathbb{E}[U_\mu^{\mathrm{P}}] \right| + \left| \mathbb{E}[U_\mu^{\mathrm{P}}] - \mathbb{E}[U_\mu] \right|.$$

Then, the first term on the right-hand side is at most $2e^{(2a-1)r}$ by Lemma 3, the second is at most $na \cdot \Pr(\mathrm{Poi}(r) \leq \mu)$ as stated in the last subsection, and the last is at most 4 by Lemma 2. As for the third term, our construction of function $\tilde{F}_\mu$ and f_j yields

$$\mathrm{Bias}_3 \triangleq \left| \tilde{U}_\mu^{\mathrm{P}} - \mathbb{E}[U_\mu^{\mathrm{P}}] \right|$$

$$\leq \sum_x |\tilde{F}_\mu(a\lambda_x) - F_\mu(a\lambda_x)|$$

$$\leq \sum_x \left(e^{-r} \int_0^{a\lambda_x} |g_0(sr)| e^{s - a\lambda_x} ds + \sum_{j \leq \mu - 1} \frac{1}{j!} \int_r^\infty |g_j(\alpha a\lambda_x)| e^{-\alpha} \alpha^j d\alpha \right).$$

7

Analyzing the last summation calls for point-wise and uniform bounds on $|g_j(y)|$, which further requires bounds on the Bessel functions by $g_j(y) = J_{2j}(2\sqrt{y})$. It turns out that Bessel functions, while being transcendental, behave nicely for nonnegative reals.

Lemma 6. *[1, 38] For any $s, y \geq 0$, we have $|J_s(y)| \leq 1$ and $|J_s(y)| \leq \frac{(y/2)^s}{s!}$.*

Building on this lemma and under appropriate conditions, the supplementary material shows that

$$\text{Bias}_3 \leq an(\mu + 1) \cdot \Pr(\text{Poi}(r) \leq \mu).$$

The proof of Theorem 1 follows from consolidating these bias bounds and the previous variance bounds by the MSE decomposition at the beginning of Section 2.

5 Experiments

Given relevant parameters and sample $X^n \sim p$, our estimator computes a linear combination of Φ_i's, the number of symbols appearing exactly i times in X^n. The estimator is both easy to implement and near-linear-time computable. To demonstrate the effectiveness of our approach, we present in this section experimental results on synthetic data (see the plots on the next page), and in the supplementary material ones on real data.

Hyper-Parameters   Our algorithm has a single hyper-parameter r. In the experiments, we choose

$$r = \frac{\log(n(a+1)^2/(a-1))}{2a}.$$

This equals the r value in Theorem 1, up to absolute constant factors, and is consistent with [29]. In fact, our theorem holds for both choices of r and the value in Theorem 1 is for proof simplification.

Distributions   We consider a support size of $S \triangleq 10,000$ and six distributions: a uniform distribution; a two-step distribution with half the symbols having probability $1/(2S)$, and the other half having probability $3/(2S)$; two Zipf distributions with parameter $\alpha \triangleq 0.5$ and 1, satisfying $p_i \propto i^{-\alpha}$, truncated at $i = S$ and renormalized; a distribution generated by the uniform prior over Δ_S, the S-dimensional simplex; a distribution generated by a Dirichlet-$1/2$ prior.

Experimental settings   We fix the sample size to be $n = S/2$, vary a from 1 to 10, and test three different μ values, $\mu = 1$, 2, and 3. The (expected) true value $\mathbb{E}[U_\mu]$ is shown in the black dashed line, and our estimator is shown in red and blue, with the solid red line representing its mean estimate, and the shaded blue area corresponding to one standard deviation, both based on 100 independent simulations. As these plots demonstrate, our simple estimator performs well in numerous settings. In addition, for $\mu = 1$, our estimator reduces to that in [29]. In the work of [29], this estimator has been compared with several state-of-the-art estimators and outperformed them in nearly every experiment.

In the supplementary, we conduct experiments on larger μ values. We observe that our estimator can provide meaningful results for $\mu = 5$ on the tested distributions. However, as μ further increases to $8 \approx \log n$, the estimator does not produce useful estimates. Both of these are consistent with our theoretical justifications, as the optimal threshold for a is $\Theta((\log n)/\mu)$.

Broader Impact

In this paper, we propose and completely resolve a natural generalization of the well-known unseen species problem. The algorithm we construct is simple in its form and linear-time computable, and performs fairly well on different synthetic and real datasets. We fully generalize the notable work of [29] in terms of the formulation, algorithm, and main results (MSE bounds). As illustrated in Section 1.1 and 1.2, our algorithm has numerous potential applications, such as active app user estimation, vocabulary size estimation, business growth analysis, species diversity estimation, database attribute study, and genetic variation study. A possible downside is that the formulation ignores any prior information, and only assumes that the distribution is discrete. A promising strategy is to incorporate the Bayesian inference framework, which will be our future research direction.
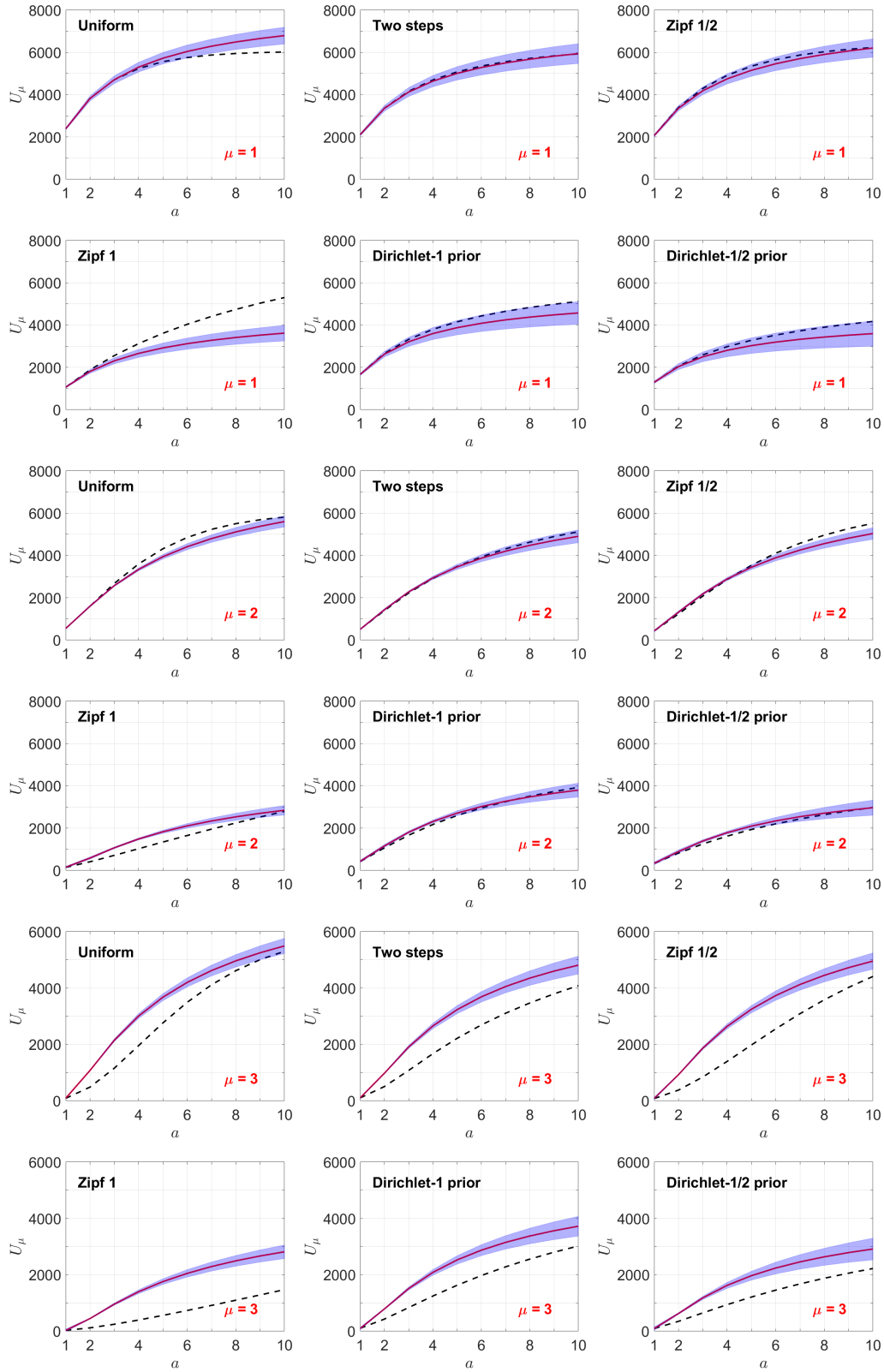
Figure 1: We present three plot sets, for $\mu = 1, 2$, and 3, each containing 6 plots, for 6 different distributions. The dashed (black) curves represent the truth while the solid (red) curves represent our estimate (the mean). The shaded areas illustrate the standard deviation from the mean.

Acknowledgments and Disclosure of Funding

References

[1] Milton Abramowitz and Irene A Stegun. Handbook of mathematical functions with formulas, graphs, and mathematical table. In *US Department of Commerce*. National Bureau of Standards Applied Mathematics series 55, 1965.

[2] Jayadev Acharya, Hirakendu Das, Alon Orlitsky, and Ananda Theertha Suresh. A unified maximum likelihood approach for estimating symmetric properties of discrete distributions. In *International Conference on Machine Learning*, pages 11–21, 2017.

[3] Andrew D Barbour and Peter Hall. On the rate of poisson convergence. *Mathematical Proceedings of the Cambridge Philosophical Society*, 95(3):473–480, 1984.

[4] John Bunge and Michael Fitzpatrick. Estimating the number of species: a review. *Journal of the American Statistical Association*, 88(421):364–373, 1993.

[5] Anne Chao. Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of statistics*, pages 265–270, 1984.

[6] Anne Chao. Species estimation and applications. *Encyclopedia of statistical sciences*, 12, 2004.

[7] Anne Chao and Shen-Ming Lee. Estimating the number of classes via sample coverage. *Journal of the American statistical Association*, 87(417):210–217, 1992.

[8] Moses Charikar, Kirankumar Shiragur, and Aaron Sidford. Efficient profile maximum likelihood for universal symmetric property estimation. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 780–791, 2019.

[9] 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.

[10] Robert K Colwell, Anne Chao, Nicholas J Gotelli, Shang-Yi Lin, Chang Xuan Mao, Robin L Chazdon, and John T Longino. Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of plant ecology*, 5(1): 3–21, 2012.

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

[12] 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.

[13] 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.

[14] IJ Good and GH Toulmin. The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika*, 43(1-2):45–63, 1956.

[15] Peter J Haas, Jeffrey F Naughton, S Seshadri, and Lynne Stokes. Sampling-based estimation of the number of distinct values of an attribute. In *VLDB*, volume 95, pages 311–322, 1995.

[16] Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Local moment matching: A unified methodology for symmetric functional estimation and distribution estimation under wasserstein distance. In *Conference On Learning Theory*, pages 3189–3221, 2018.

[17] Yi Hao and Ping Li. Bessel smoothing and multi-distribution property estimation. In *Conference on Learning Theory*, pages 1817–1876, 2020.

[18] Yi Hao and Alon Orlitsky. The broad optimality of profile maximum likelihood. In *Advances in Neural Information Processing Systems*, pages 10989–11001, 2019.

[19] Yi Hao and Alon Orlitsky. Doubly-competitive distribution estimation. In *International Conference on Machine Learning*, pages 2614–2623, 2019.

[20] Yi Hao and Alon Orlitsky. Unified sample-optimal property estimation in near-linear time. In *Advances in Neural Information Processing Systems*, pages 11106–11116, 2019.

[21] Yi Hao and Alon Orlitsky. Data amplification: Instance-optimal property estimation. In *International Conference on Machine Learning*, 2020.

[22] Yi Hao and Alon Orlitsky. Profile entropy: A fundamental measure for the learnability and compressibility of discrete distributions. In *Advances in Neural Information Processing Systems (To appear)*, 2020.

[23] Yi Hao, Alon Orlitsky, Ananda Theertha Suresh, and Yihong Wu. Data amplification: A unified and competitive approach to property estimation. In *Advances in Neural Information Processing Systems*, pages 8834–8843, 2018.

[24] Iuliana Ionita-Laza, Christoph Lange, and Nan M Laird. Estimating the number of unseen variants in the human genome. *Proceedings of the National Academy of Sciences*, 106(13): 5008–5013, 2009.

[25] GRADSHTEYN IS and IM RYZHIK. Table of integrals, series, and products, corrected and enlarged ed, 1980.

[26] Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. Minimax estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 61(5):2835–2885, 2015.

[27] Jiantao Jiao, Yanjun Han, and Tsachy Weissman. Minimax estimation of the $l\_\{1\}$ distance. *IEEE Transactions on Information Theory*, 64(10):6672–6706, 2018.

[28] Ian Kroes, Paul W Lepp, and David A Relman. Bacterial diversity within the human subgingival crevice. *Proceedings of the National Academy of Sciences*, 96(25):14547–14552, 1999.

[29] 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.

[30] Bruce J Paster, Susan K Boches, Jamie L Galvin, Rebecca E Ericson, Carol N Lau, Valerie A Levanos, Ashish Sahasrabudhe, and Floyd E Dewhirst. Bacterial diversity in human subgingival plaque. *Journal of bacteriology*, 183(12):3770–3783, 2001.

[31] Yury Polyanskiy and Yihong Wu. Dualizing le cam's method, with applications to estimating the unseens. *arXiv preprint arXiv:1902.05616*, 2019.

[32] Harlan S Robins, Paulo V Campregher, Santosh K Srivastava, Abigail Wacher, Cameron J Turtle, Orsalem Kahsai, Stanley R Riddell, Edus H Warren, and Christopher S Carlson. Comprehensive assessment of t-cell receptor β-chain diversity in $\alpha\beta$ t cells. *Blood, The Journal of the American Society of Hematology*, 114(19):4099–4107, 2009.

[33] Eric P Smith and Gerald van Belle. Nonparametric estimation of species richness. *Biometrics*, pages 119–129, 1984.

[34] J Michael Steele et al. An efron-stein inequality for nonsymmetric statistics. *The Annals of Statistics*, 14(2):753–758, 1986.

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

[36] Gregory Valiant and Paul Valiant. Instance optimal learning of discrete distributions. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 142–155, 2016.

[37] Paul Valiant and Gregory Valiant. Estimating the unseen: improved estimators for entropy and other properties. In *Advances in Neural Information Processing Systems*, pages 2157–2165, 2013.

[38] George Neville Watson. *A treatise on the theory of Bessel functions*. Cambridge university press, 1995.

[39] Yihong Wu and Pengkun Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6):3702–3720, 2016.

[40] Yihong Wu, Pengkun Yang, et al. Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *The Annals of Statistics*, 47(2):857–883, 2019.

Appendix

Table of Contents

A   Lemma 1 and Variance Bound

Lemma 1 (Steele's inequality). *If $S(x^n)$ is any real function of n variables and if $X_i, X_i', 1 \leq i \leq n$ are $2n$ i.i.d. random variables then $\mathrm{Var}\,(S) \leq \frac{1}{2} \mathbb{E}\left[\sum_{i=1}^{n}(S - S_i)^2\right]$, where $S \triangleq S(X^n)$ and S_i is given by replacing the i-th observation with X_i'.*

The inequality basically states that as long as we have i.i.d. input X^n and care only a functional $S(X^n)$ *that is not too sensitive to input changes*, the variance of $S(X^n)$ will be reasonably bounded.

As we showed in Section 2, U_μ is a functional of the length-$(n+m)$ i.i.d. sample (X^n, Y^m), and modifying a single symbol changes U_μ by at most 1. Hence, Lemma 1 bounds its variance as

$$\mathrm{Var}(U_\mu) \leq \frac{1}{2}(n + m).$$

As for estimator $\hat{U}_\mu = \sum_{i \geq 1} s_i \Phi_i = \sum_{i \geq 1} s_i (\sum_x \mathbb{1}_{N_x = i})$, where N_x denotes the number of times symbol x appearing in the sample, changing a single symbol changes its value by at most $2 \max_i |s_i|$.

As we will see in Appendix C, the maximum magnitude of the coefficients is bound by $2e^{(2a-1)r}$ from above, implying an upper bound of $8ne^{2(2a-1)r}$ on the variance of our estimator.

B   Proof of Lemma 2

Recall that U_μ is the quantity of interest, and U_μ^{P} is its alternative under Poisson sampling.

Lemma 2. *For any n, a and p, we have $\left|\mathbb{E}[U_\mu - U_\mu^{\mathrm{P}}]\right| \leq 4$.*

Proof. The original symbols counts are defined as N_x and M_x in the main paper. For simplicity, we will use N_x' and M_x' to denote their Poissonized versions. Then the absolute difference becomes

$$\left|\mathbb{E}[U_\mu - U_\mu^{\mathrm{P}}]\right| = \left|\sum_x \mathbb{E}[\mathbb{1}_{N_x=0} \cdot \mathbb{1}_{M_x \geq \mu} - \mathbb{1}_{N_x'=0} \cdot \mathbb{1}_{M_x' \geq \mu}]\right|.$$

By symmetry, it suffices to bound a single term in the summation.

$$
\begin{aligned}
D(p_x) &\overset{(a)}{:=} \left|\mathbb{E}\left[\mathbb{1}_{N_x=0} \cdot \mathbb{1}_{M_x \geq \mu} - \mathbb{1}_{N_x'=0} \cdot \mathbb{1}_{M_x' \geq \mu}\right]\right| \\
&\overset{(b)}{=} \left|\mathbb{E}\left[\mathbb{1}_{M_x \geq \mu} \cdot \left(\mathbb{1}_{N_x=0} - \mathbb{1}_{N_x'=0}\right)\right] + \mathbb{E}\left[\left(\mathbb{1}_{M_x \geq \mu} - \mathbb{1}_{M_x' \geq \mu}\right) \cdot \mathbb{1}_{N_x'=0}\right]\right| \\
&\overset{(c)}{=} \left|\mathbb{E}\left[\mathbb{1}_{M_x \geq \mu}\right] \cdot \mathbb{E}\left[\left(\mathbb{1}_{N_x=0} - \mathbb{1}_{N_x'=0}\right)\right] + \mathbb{E}\left[\left(\mathbb{1}_{M_x \geq \mu} - \mathbb{1}_{M_x' \geq \mu}\right)\right] \cdot \mathbb{E}\left[\mathbb{1}_{N_x'=0}\right]\right| \\
&\overset{(d)}{\leq} \left|\mathbb{E}\left[\mathbb{1}_{N_x=0} - \mathbb{1}_{N_x'=0}\right]\right| + \left|\mathbb{E}\left[\mathbb{1}_{M_x \geq \mu} - \mathbb{1}_{M_x' \geq \mu}\right]\right| \\
&\overset{(e)}{\leq} 4p_x,
\end{aligned}
$$

where (a) defines $D(p_x)$; (b) follows by subtracting and adding the term $\mathbb{1}_{N_x'=0} \cdot \mathbb{1}_{M_x \geq \mu}$; (c) follows by independence between the empirical counts; (d) follows by the fact that the expectation of an indicator function is at most 1; (e) follows by the following inequality:

Lemma 7. *[3] For any $q \in [0,1]$, $Y \sim \mathrm{Poi}(mq)$, $Z \sim \mathrm{bin}(m,q)$, and real function f,*
$$\left|\mathbb{E}\left[f(Y)\right] - \mathbb{E}\left[f(Z)\right]\right| \leq 2q \sup_j |f(j)|.$$

Therefore, the difference we want to bound is at most

$$\left|\mathbb{E}[U_\mu - U_\mu^{\mathrm{P}}]\right| \leq \sum_x D(p_x) \leq \sum_x 4p_x = 4. \qquad \square$$

C   Proof of Lemma 3 and Coefficient Bound

Recall that our estimator takes the form of

$$\hat{U}_\mu \triangleq \hat{U}_\mu(X^n, a) \triangleq \sum_{i=1}^n s_i \cdot \Phi_i.$$

where $\Phi_i = \sum_x \mathbb{1}_{N_x=i}$ denotes the number of symbols appearing exactly i times. Similar to the last section, we will use N_x' and Φ_i' to denote the Poissonized versions of these quantities.

Lemma 3. *Under the conditions in Theorem 1, we have $|\mathbb{E}[\hat{U}_\mu - \hat{U}_\mu^{\mathrm{P}}]| \leq 2e^{(2a-1)r}$.*

Proof. Adopting the Poisson sampling model changes the expectation of the estimator by

$$
\begin{aligned}
|\mathbb{E}[\hat{U}_\mu - \hat{U}_\mu^{\mathrm{P}}]| &= \left|\sum_i (s_i \cdot \Phi_i - s_i \cdot \Phi_i')\right| \\
&\overset{(a)}{\leq} \sum_x \left|\sum_i s_i \cdot \left(\mathbb{1}_{N_x=i} - \mathbb{1}_{N_x'=i}\right)\right| \\
&\overset{(b)}{\leq} \sum_x 2p_x \max_i |s_i| \\
&\overset{(c)}{=} 2 \max_i |s_i|,
\end{aligned}
$$

where (a) follows by the triangle inequality; (b) follows by Lemma 7; (c) follows by $\sum_x p_x = 1$.

To complete our proof of Lemma 3, we establish the following bounds on the estimator's coefficients. Intuitively, this captures how sensitive the estimator is to changes in the input sample.

Lemma 8. *For any positive index $i \in \mathbb{N}$,*

$$|s_i| \le e^{(2a-1)r}.$$

Proof of Lemma 8. For any positive index $i \in \mathbb{N}$,

$$
\begin{aligned}
|s_i| &\overset{(a)}{\le} a^i \sum_{j=0}^{(\mu-1)\wedge i} \frac{i!}{j!(i-j)!} \Pr(\mathrm{Poi}(r) \ge i+j) \\
&\overset{(b)}{\le} a^i \Pr(\mathrm{Poi}(r) \ge i) \left(\sum_{j=0}^{(\mu-1)\wedge i} \frac{i!}{j!(i-j)!} \right) \\
&\overset{(c)}{\le} (2a)^i \Pr(\mathrm{Poi}(r) \ge i) \\
&\overset{(d)}{=} (2a)^i e^{-r} \sum_{j=i}^{\infty} \frac{r^j}{j!} \\
&\overset{(e)}{\le} e^{-r} \sum_{j=i}^{\infty} \frac{(2ar)^j}{j!} \\
&\overset{(f)}{\le} e^{(2a-1)r}.
\end{aligned}
$$

where (a) follows by the definition of s_i and the triangle inequality; (b) follows by the monotonicity of Poisson tail probabilities; (c) follows by the binomial expansion of $(1+1)^i$; (d) follows by the expressions of Poisson probabilities; (e) follows by the monotonicity of Power functions; and (f) follows by the series expansion of the function $\exp(x)$.

Applying this lemma to the right-hand side of our bound on $|\mathbb{E}[\hat{U}_\mu - \hat{U}_\mu^{\mathrm{P}}]|$ completes the proof. $\quad\square$

D   Proof of Lemma 4

Recall that J_j denotes the j-th order Bessel function of the first kind. For $j \ge 0$, we have

$$g_j(y) \triangleq J_{2j}(2\sqrt{y}) = \sum_{t=0}^{\infty} \frac{(-1)^t y^{t+j}}{t!(t+2j)!}.$$

Lemma 4 presents the integral form of $e^{-y}y^j$ expressed in terms of the Bessel functions. We present its proof in [23], which shines light on latter derivations.

Lemma 4. *[23] For any $j \in \mathbb{Z}^+$ and $y \ge 0$,*

$$e^{-y}y^j = \int_0^\infty e^{-\alpha} \alpha^j g_j(\alpha y) d\alpha.$$

Proof. By Fubini's theorem and the series expansion of g_j,

$$
\begin{aligned}
\int_0^\infty e^{-\alpha} \alpha^j g_j(\alpha y) d\alpha &= \int_0^\infty e^{-\alpha} \alpha^j \sum_{i=0}^{\infty} \frac{(-1)^i (\alpha y)^{i+j}}{i!(i+2j)!} d\alpha \\
&= \sum_{i=0}^{\infty} \frac{(-1)^i y^{i+j}}{i!(i+2j)!} \int_0^\infty e^{-\alpha} \alpha^{i+2j} d\alpha.
\end{aligned}
$$

15

Observe that the integral is actually $\Gamma(i + 2j + 1)$ and equals to $(i + 2j)!$,

$$\sum_{i=0}^{\infty} \frac{(-1)^i y^{i+j}}{i!(i + 2j)!} \int_0^{\infty} e^{-\alpha} \alpha^{i+2j} d\alpha = \sum_{i=0}^{\infty} \frac{(-1)^i y^{i+j}}{i!(i + 2j)!}(i + 2j)!$$

$$= \sum_{i=0}^{\infty} \frac{(-1)^i y^{i+j}}{i!}$$

$$= e^{-y} y^j. \qquad \square$$

E Proof of Lemma 5

By the argument in Section 4, instead of estimating U_μ, we will consider its Poissonized version U_μ^{P} and approximate the value of U_μ^{P} by a *near-unbiased* estimator of

$$\tilde{U}_\mu^{\mathrm{P}} \triangleq \sum_x \tilde{F}_\mu(a\lambda_x) \cdot e^{-\lambda_x},$$

where $\lambda_x \triangleq n p_x$ and $\tilde{F}_\mu(y)$ is defined as $\sum_{j \leq \mu-1} \frac{1}{j!} f_j(y)$ for the sequence of real functions

$$f_0(y) = -\sum_{j=1}^{\infty} \frac{(-y)^j}{j!} \Pr(\mathrm{Poi}(r) \geq j) \quad \text{and} \quad f_j(y) = -\int_0^r e^{-\alpha} \alpha^j g_j(\alpha y) d\alpha, \forall j \geq 1.$$

The following lemma finds the series expansion of $\tilde{F}_\mu(y)$.

Lemma 5. *For any non-negative number y,*

$$\tilde{F}_\mu(y) = -\sum_{i=1}^{\infty} c_i \cdot \frac{y^i}{i!},$$

where the coefficients are

$$c_i \triangleq (-1)^i \Pr(\mathrm{Poi}(r) \geq i) + \sum_{j=1}^{(\mu-1)\wedge i} (-1)^{i-j} \binom{i}{j} \Pr(\mathrm{Poi}(r) \geq i + j + 1).$$

Proof. For any $j \geq 1$, the following equations hold.

$$f_j(y) \overset{(a)}{=} -\int_0^r e^{-\alpha} \alpha^j g_j(\alpha y) d\alpha$$

$$\overset{(b)}{=} -\int_0^r e^{-\alpha} \alpha^j \sum_{t=0}^{\infty} \frac{(-1)^j (\alpha y)^{t+j}}{t!(t + 2j)!} d\alpha$$

$$\overset{(c)}{=} -\sum_{t=0}^{\infty} \frac{(-1)^j y^{t+j}}{t!(t + 2j)!} \int_0^r e^{-\alpha} \alpha^{t+2j} d\alpha$$

$$\overset{(d)}{=} -\sum_{t=0}^{\infty} \frac{(-1)^t y^{t+j}}{t!} \Pr(\mathrm{Poi}(r) \geq t + 2j + 1),$$

where (a) follows by the definition of f_j; (b) follows by the series expansions of the Bessel functions; (c) follows by Fubini's theorem; (d) follows by the series expansions the incomplete Gamma functions. Therefore, we can write $\tilde{F}_\mu(y)$ as

$$\sum_{j \leq \mu-1} \frac{f_j(y)}{j!} = -\sum_{t=1}^{\infty} \frac{(-y)^t}{t!} \Pr(\mathrm{Poi}(r) \geq t) - \sum_{j=1}^{\mu-1} \frac{1}{j!} \sum_{t'=0}^{\infty} \frac{(-1)^{t'} y^{t'+j}}{t'!} \Pr(\mathrm{Poi}(r) \geq t' + 2j + 1)$$

$$= -\sum_{i=1}^{\infty} \left((-1)^i \Pr(\mathrm{Poi}(r) \geq i) + \sum_{j=1}^{(\mu-1)\wedge i} \frac{(-1)^{i-j} i!}{j!(i - j)!} \Pr(\mathrm{Poi}(r) \geq i + j + 1) \right) \frac{y^i}{i!},$$

which completes the proof of the lemma. $\qquad \square$

16

F Lemma 6 and Bias Bound

Our construction closely relates to Bessel functions of the first kind. The following lemma provides useful bounds on their magnitude.

Lemma 6. *[1, 38] For any $s, y \geq 0$, we have $|J_s(y)| \leq 1$ and $|J_s(y)| \leq \frac{(y/2)^s}{s!}$.*

As a corollary, we have the following bounds for $g_j(y) = J_{2j}(2\sqrt{y})$, where $j \geq 0$.

Corollary 1 (Extension of a result in [23]). *For $j \geq 0$ and $y \geq 0$, we have $|g_j(y)| \leq 1$, and if in addition, $j \geq 1$, then $|g_j(y)| \leq \frac{y}{j+1}$.*

Proof. The first inequality follows directly from Lemma 6. For the second inequality, by definition of g_j, we can therefore write

$$|g_j(y)| = |J_{2j}(2\sqrt{y})| \leq \frac{y^j}{(2j)!} \leq \frac{y^j}{j+1},$$

where the last step follows from $(j+1) \leq (2j)!$ for any $j \geq 0$. $\qquad\square$

Now, we are ready to establish the last bias bound in Section 4, which states that

$$\text{Bias}_3 \triangleq \left| \tilde{U}_\mu^{\mathrm{P}} - \mathbb{E}[U_\mu^{\mathrm{P}}] \right| \leq \sum_x |\tilde{F}_\mu(a\lambda_x) - F_\mu(a\lambda_x)| \leq an(\mu+1) \cdot \Pr(\text{Poi}(r) \leq \mu).$$

Proof. Recall that λ_x abbreviates np_x. Consolidating the above bounds yields

$$\text{Bias}_3 \overset{(a)}{\leq} \sum_x e^{-r} \left[e^{-a\lambda_x} \int_0^{a\lambda_x} |g_0(sr)| e^s ds \right] + \sum_x \sum_{1 \leq j \leq \mu-1} \frac{1}{j!} \int_r^\infty e^{-\alpha} \alpha^j |g_j(\alpha a\lambda_x)| d\alpha$$

$$\overset{(b)}{\leq} \sum_x e^{-r} \left[e^{-a\lambda_x} \int_0^{a\lambda_x} e^s ds \right] + \sum_x \sum_{1 \leq j \leq \mu-1} \frac{1}{j!} \int_r^\infty e^{-\alpha} \alpha^j \frac{\alpha a\lambda_x}{j+1} d\alpha$$

$$\overset{(c)}{=} e^{-r} \sum_x (1 - e^{-a\lambda_x}) + \sum_x a\lambda_x \sum_{1 \leq j \leq \mu-1} \frac{1}{(j+1)!} \int_r^\infty e^{-\alpha} \alpha^{j+1} d\alpha$$

$$\overset{(d)}{\leq} e^{-r} \sum_x a\lambda_x + e^{-r} \left(\sum_{j=1}^{\mu-1} \sum_{t=0}^{j+1} \frac{r^t}{t!} \right) \sum_x a\lambda_x$$

$$\overset{(e)}{\leq} e^{-r} an \left(1 + \sum_{j=1}^{\mu-1} \sum_{t=0}^{j+1} \frac{r^t}{t!} \right)$$

$$\overset{(f)}{\leq} an(\mu+1) \cdot \Pr(\text{Poi}(r) \leq \mu),$$

where (a) follows by the triangle inequality; (b) follows by Corollary 1; (c) follows by simple algebra; (d) follows by $1 - e^{-x} \leq x, \forall x \geq 0$ and the series expansion of the incomplete gamma function; (e) follows by $\sum_x \lambda_x = n$; (f) follows by adding non-negative terms and by Poisson probabilities. $\quad\square$

G   Estimator Modification

In Section 4, we introduced a quantity $\tilde{U}_\mu^{\mathrm{P}}$ as our new approximation target, which takes the form of

$$\tilde{U}_\mu^{\mathrm{P}} = \sum_x \tilde{F}_\mu(a\lambda_x) \cdot e^{-\lambda_x} = -\sum_x \sum_{i=1}^\infty (c_i a^i) \cdot e^{-\lambda_x} \frac{\lambda_x^i}{i!} = -\sum_{i=1}^\infty (c_i a^i) \cdot \mathbb{E}[\Phi_i'],$$

where Φ_i' denotes the number of symbols appearing exactly i times in the Poissonized sample, and the coefficients c_i's are defined as

$$c_i \triangleq (-1)^i \Pr(\mathrm{Poi}(r) \geq i) + \sum_{j=1}^{(\mu-1)\wedge i} (-1)^{i-j} \binom{i}{j} \Pr(\mathrm{Poi}(r) \geq i + j + 1), \ \forall i.$$

Under Poisson sampling, a simple unbiased estimator is $-\sum_{i=1}^\infty (c_i a^i) \cdot \Phi_i'$. It should also be clear that $s_i \approx c_i a^i$ for every i as they are equal if we replace each smoothing probability $\Pr(\mathrm{Poi}(r) \geq i+j+1)$ by $\Pr(\mathrm{Poi}(r) \geq i+j)$ in the expression of c_i. Hence, given sample statistics Φ_i' for $i \geq 1$, we claimed that a *near-unbiased* estimator of $\tilde{U}_\mu^{\mathrm{P}}$ is

$$\hat{U}_\mu^{\mathrm{P}} = -\sum_{i=1}^\infty s_i \cdot \Phi_i'.$$

Below, we show that the absolute bias satisfies $|\mathbb{E}[\hat{U}_\mu^{\mathrm{P}}] - \tilde{U}_\mu^{\mathrm{P}}| \leq na \cdot \Pr(\mathrm{Poi}(r) \leq \mu)$ by leveraging the point-wise bounds on Bessel functions in the last section.

Proof. The difference between the two estimators is

$$|\mathbb{E}[\hat{U}_\mu^{\mathrm{P}}] - \tilde{U}_\mu^{\mathrm{P}}| = \sum_x d(a\lambda_x),$$

where $\lambda_x = np_x$ and

$$d(y) \triangleq -\sum_{j \leq \mu-1} \frac{1}{j!} \sum_{t=0}^\infty \frac{(-1)^t y^{t+j}}{t!} \Pr(\mathrm{Poi}(r) \geq t + 2j).$$

First we show that $d(y)$ has a small magnitude. Specifically,

$$|d(y)| = \left| \sum_{j \leq \mu-1} \frac{1}{j!} \sum_{t=0}^\infty \frac{(-1)^t y^{t+j}}{t!} \Pr(\mathrm{Poi}(r) = t + 2j) \right|$$

$$= \left| \sum_{j \leq \mu-1} \frac{1}{j!} \sum_{t=0}^\infty \frac{(-1)^t y^{t+j}}{t!} \left(e^{-r} \frac{r^{t+2j}}{(t+2j)!} \right) \right|$$

$$= \left| \sum_{j \leq \mu-1} e^{-r} \frac{r^j}{j!} \sum_{t=0}^\infty \frac{(-1)^t (yr)^{t+j}}{t!(t+2j)!} \right|$$

$$\stackrel{(a)}{=} \left| \sum_{j \leq \mu-1} e^{-r} \frac{r^j}{j!} \cdot g_j(yr) \right|$$

$$\stackrel{(b)}{\leq} \sum_{j \leq \mu-1} e^{-r} \frac{r^j}{j!} \cdot \frac{yr}{j+1}$$

$$= y \cdot \Pr(1 \leq \mathrm{Poi}(r) \leq \mu),$$

where the first three steps follow by algebraic manipulations; (a) follows by the series expansion of g_j (or equivalently the $2j$-th order Bessel function); (b) follows by Corollary 1. Consequently, the absolute bias of our estimate satisfies

$$|\mathbb{E}[\hat{U}_\mu^{\mathrm{P}}] - \tilde{U}_\mu^{\mathrm{P}}| \leq \left| \sum_x d(a\lambda_x) \right| \leq \left(\sum_x a\lambda_x \right) \cdot \Pr(1 \leq \mathrm{Poi}(r) \leq \mu) \leq na \cdot \Pr(\mathrm{Poi}(r) \leq \mu). \quad \square$$

H Lower Bounds

We establish a lower bound on the worst-case MSE of any estimator by connecting the task to the closely related task of support size estimation, which is ill-defined if we simply consider the collection of discrete distributions with bounded support size. Specifically, for an unknown distribution and any finite sample size n, there can be many symbols with tiny probabilities so that it is unlikely to observe any of them in a size-n sample, yet they can have a huge impact on the distribution's support size.

Motivated by applications to database studies, the paper of [40] considered support size estimation for distributions whose positive probabilities are at least $1/k$, where k is the known size of the alphabet (hence also serves as an upper bound on distributions' support sizes).

For any distribution $p \in \Delta$, denote by $S(p) = \sum_x \mathbb{1}_{p_x>0}$ its support size, and by $p_{\min}^+$ its minimum nonzero probability. Paper [40] showed the following lower bound on the minimax MSE (together with a nearly matching upper bound).

Lemma 9. *Assume that $k \log k \gg n \gg \frac{k}{\log k}$ and $k \gg 1$,*

$$\min_{\hat{S}} \max_{p: p_{min}^+ \geq 1/k} \mathbb{E}_{X^n \sim p} \left(\hat{S}(X^n) - S(p) \right)^2 \geq \exp\left(-4\sqrt{\frac{n \log k}{k}} \right) k^2.$$

In the next two subsections, we will connect this estimation problem to ours, and establish Theorem 2.

H.1 Connection between Two Problems

Recall that the quantity of interest is

$$U_\mu = \sum_x \mathbb{1}_{M_x \geq \mu} \cdot \mathbb{1}_{N_x=0}.$$

For some integer k to be determined later, assume that p is a distribution satisfying $p_{\min}^+ \geq 1/k$. For the ease of exposition, define

$$U_\mu^p \triangleq \sum_x \mathbb{E}\left[\mathbb{1}_{M_x \geq \mu}\right] \cdot \mathbb{1}_{N_x=0} = \sum_x \Pr\left(\text{bin}(na, p_x) \geq \mu\right) \cdot \mathbb{1}_{N_x=0},$$

Adding the number of observed symbols in X^n, we obtain

$$\tilde{U}_\mu^p \triangleq U_\mu(X^n, p) + \sum_x \mathbb{1}_{N_x>0} = \sum_x \left(\Pr\left(\text{bin}(na, p_x) \geq \mu\right) \cdot \mathbb{1}_{N_x=0} + \mathbb{1}_{N_x>0}\right).$$

Below, we show that for any $\beta \geq 3$ and $k \leq na/(\beta\mu)$, and distribution p satisfying $p_{\min}^+ \geq 1/k$,

$$\left|\tilde{U}_\mu^p - S(p)\right| \leq S(p) \exp\left(-0.3\frac{na}{k}\right) \leq k \exp\left(-0.3\frac{na}{k}\right).$$

Proof. The absolute difference between $\tilde{U}_\mu^p$ and $S(p)$ satisfies

$$S(p) - \tilde{U}_\mu^p \overset{(a)}{=} \sum_x \left(\mathbb{1}_{p_x>0} \cdot \mathbb{1}_{N_x=0} + \mathbb{1}_{N_x>0}\right) - \sum_x \left(\Pr\left(\text{bin}(na, p_x) \geq \mu\right) \cdot \mathbb{1}_{N_x=0} + \mathbb{1}_{N_x>0}\right)$$

$$\overset{(b)}{=} \sum_x \mathbb{1}_{N_x=0} \cdot \mathbb{1}_{p_x>0} - \sum_x \Pr\left(\text{bin}(na, p_x) \geq \mu\right) \cdot \mathbb{1}_{N_x=0} \cdot \mathbb{1}_{p_x>0}$$

$$\overset{(c)}{=} \sum_x \left(1 - \Pr\left(\text{bin}(na, p_x) \geq \mu\right)\right) \cdot \mathbb{1}_{N_x=0} \cdot \mathbb{1}_{p_x>0}$$

$$\overset{(d)}{\geq} 0,$$

where (a) holds as $\mathbb{1}_{N_x>0} = 0$ surely if $p_x = 0$; (b) follows by term cancelation and the equivalence of $\Pr\left(\text{bin}(na, p_x) \geq \mu\right) = 0$ and $p_x = 0$, for any $1 \leq \mu \leq na$; (c) follows by simply algebra; (d) follows by $\Pr\left(\text{bin}(na, p_x) \geq \mu\right) \leq 1$.

On the other hand, for $k \leq na/(\beta\mu)$ and $\beta \geq 3$,

$$\tilde{U}_\mu^p \overset{(a)}{=} \sum_x \left(\Pr\left(\text{bin}(na, p_x) \geq \mu\right) \cdot \mathbb{1}_{N_x=0} + \mathbb{1}_{N_x>0}\right)$$

$$\overset{(b)}{=} \sum_{x:p_x \geq 1/k} \left((1 - \Pr\left(\text{bin}(na, p_x) < \mu\right)) \cdot \mathbb{1}_{N_x=0} + \mathbb{1}_{N_x>0}\right)$$

$$\overset{(c)}{\geq} \sum_{x:p_x \geq 1/k} \left(1 - \Pr\left(\text{bin}(na, 1/k) < \mu\right) \cdot \mathbb{1}_{N_x=0}\right)$$

$$\overset{(d)}{\geq} \sum_{x:p_x \geq 1/k} \left(1 - \Pr\left(\text{bin}(na, 1/k) < \frac{na}{3k}\right) \cdot \mathbb{1}_{N_x=0}\right)$$

$$\overset{(e)}{\geq} \sum_{x:p_x \geq 1/k} \left(1 - \exp\left(-0.3\frac{na}{k}\right) \cdot \mathbb{1}_{N_x=0}\right)$$

$$\overset{(f)}{\geq} S(p) \left(1 - \exp\left(-0.3\frac{na}{k}\right)\right),$$

where (a) follows by the definition of $\tilde{U}_\mu^p$; (b) follows by the equivalence of $p_x = 0$, $\mathbb{1}_{N_x>0} = 0$ and $\Pr\left(\text{bin}(na, p_x) \geq \mu\right) = 0$; (c) follows by $\mathbb{1}_{N_x=0} + \mathbb{1}_{N_x>0} = 1$; (d) follows by the two conditions $k \leq na/(\beta\mu)$ and $\beta \geq 3$; (e) follows by standard Chernoff bound for binomial random variables; (f) holds as the summation is over x satisfying $p_x \geq 1/k$.

Consequently, we have shown that

$$\left|\tilde{U}_\mu^p - S(p)\right| \leq S(p) \exp\left(-0.3\frac{na}{k}\right) \leq k \exp\left(-0.3\frac{na}{k}\right). \qquad \square$$

H.2   Proof of Theorem 2

Theorem 2. *There exist absolute constants c' and c_0' such that for any $a \geq (c' \log n)/\mu$, and any n-sample estimator $\hat{U}$,*

$$\mathcal{E}_{n,a}^\mu(\hat{U}) \gtrsim \frac{1}{n^{c_0'/a}}.$$

The theorem lowerly bounds the worst-case normalized MSE of the best estimator.

Proof. For any estimator $\hat{U}$ and distribution p, the MSE of estimating U_μ admits

$$\text{MSE} \overset{(a)}{=} \mathbb{E}\left(\hat{U} - U_\mu\right)^2$$

$$\overset{(b)}{\geq} \underset{X^n \sim p}{\mathbb{E}} \left(\hat{U} - U_\mu^p\right)^2$$

$$\overset{(c)}{\geq} \frac{1}{2} \underset{X^n \sim p}{\mathbb{E}} \left(\hat{U} - U_\mu^p + \tilde{U}_\mu^p - S(p)\right)^2 - \underset{X^n \sim p}{\mathbb{E}} \left(\tilde{U}_\mu^p - S(p)\right)^2$$

$$\overset{(d)}{\geq} \frac{1}{2} \underset{X^n \sim p}{\mathbb{E}} \left(\hat{U} + \sum_x \mathbb{1}_{N_x>0} - S(p)\right)^2 - \underset{X^n \sim p}{\mathbb{E}} \left(\tilde{U}_\mu^p - S(p)\right)^2$$

$$\overset{(e)}{\geq} \frac{1}{2} \underset{X^n \sim p}{\mathbb{E}} \left(\hat{U} + \sum_x \mathbb{1}_{N_x>0} - S(p)\right)^2 - \exp\left(-0.6\frac{na}{k}\right) k^2,$$

where (a) follows by definition; (b) follows by the Jensen's inequality; (c) follows by the linearity of expectation and inequality $a^2 \geq \frac{1}{2}(b+a)^2 - b^2$; (d) follows by $\tilde{U}_\mu^p - U_\mu^p = \sum_x \mathbb{1}_{N_x>0}$; (e) follows by the main result of Appendix H.1 and our choice of parameters.

Consider $a\mu = c' \log n$ for some $c' \gg 1$. For $a, \mu \ll \log n$, if we choose $k = \frac{na}{\beta\mu}$, then $\frac{k}{\log k} \ll n \ll k \log k$. Therefore, by Lemma 9 on support size estimation, the minimax MSE of learning U_μ, which we refer to as the MMSE, satisfies

$$
\begin{aligned}
\text{MMSE} &\overset{(a)}{=} \min_{\hat{U}} \max_{p \in \Delta} \mathbb{E}\,(\hat{U} - U_\mu)^2 \\
&\overset{(b)}{\geq} \min_{\hat{U}} \max_{p:\, p_{\min}^+ \geq 1/k} \mathbb{E}\,(\hat{U} - U_\mu)^2 \\
&\overset{(c)}{\geq} \frac{1}{2} \min_{\hat{U}} \max_{p:\, p_{\min}^+ \geq 1/k} \mathbb{E}_{X^n \sim p} \left(\hat{U} + \sum_x \mathbb{1}_{N_x > 0} - S(p) \right)^2 - \exp\left(-0.6\frac{na}{k}\right) k^2 \\
&\overset{(d)}{\geq} \frac{1}{2} \exp\left(-4\sqrt{\frac{n \log k}{k}}\right) k^2 - \exp\left(-0.6\frac{na}{k}\right) k^2,
\end{aligned}
$$

where (a) follows by the definition; (b) follows by maximizing over a subset set of the original; (c) follows by the inequality that we established above; and (d) follows by the lower bound in Lemma 9 and viewing $\hat{U} + \sum_x \mathbb{1}_{N_x > 0}$ as an n-sample support size estimator.

Choose $\beta = (10 \log 4)/\mu + 3$ so that we have both $\beta \geq 3$ for any $\mu \geq 1$ and $0.1 \geq (\log 4)/(\beta\mu)$. To ensure non-triviality, we require

$$
\exp\left(-4\sqrt{\frac{n \log k}{k}}\right) k^2 \geq 4\exp\left(-0.6\frac{na}{k}\right) k^2 \Longleftarrow -4\sqrt{\frac{n \log k}{k}} \geq \log 4 - 0.6\frac{na}{k}
$$

$$
\Longleftrightarrow 4\sqrt{nk \log k} \leq 0.6na - na\frac{\log 4}{\beta\mu}
$$

$$
\Longleftarrow 4\sqrt{nk \log k} \leq 0.5na
$$

$$
\Longleftarrow \log n \ll a\mu,
$$

where the last step follows by $1 \leq a, \mu \ll n$. By definition, we have $k = M_\mu/\beta$. Therefore,

$$
\min_{\hat{U}} \max_{p \in \Delta} \mathbb{E}\,(\hat{U} - U_\mu)^2 \geq \exp\left(-0.6\mu\right) k^2 = \left(\frac{1}{\beta} M_\mu e^{-0.3\mu}\right)^2.
$$

Dividing both sides by M_μ^2 yields

$$
\begin{aligned}
\min_{\hat{U}} \max_{p \in \Delta} \mathbb{E}_{X^n \sim p} \left(\frac{\hat{U} - U_\mu}{M_\mu}\right)^2 &\geq \left(\frac{1}{\beta} e^{-0.3\mu}\right)^2 \\
&= \left(\frac{1}{\beta} \exp\left(-0.3\frac{c' \log n}{a}\right)\right)^2 \\
&\gtrsim \frac{1}{n^{c_0'/a}},
\end{aligned}
$$

where $c_0' \triangleq 0.6c'$ is a positive absolute constant. This completes the proof of Theorem 2. $\qquad\square$

I Experiments on LOTR (Lord of the Rings) Datasets

The task is *vocabulary size estimation*, one that we mentioned at the beginning of Section 1.1. Replacing butterflies by words, *vocabulary size estimation* [9, 11, 19, 35] aims to determine how many words a writer, say William Shakespeare, knew based on his written works. An intuitive and widely used approach is to simply add up the number of observed (distinct) words and some estimate of U_1. With the same motivation, we may also want to know how many words fall into a writer's *common vocabulary* (excluding those that appear only once or twice), which calls for estimating U_μ.

Specifically, the book we used in the experiments is *Lord of the Rings*, written by J. R. R. Tolkien. We found a specific version of this book consisting of 573,479 words, of which 16,335 are distinct. To split the data into two parts, we randomly selected $1/(a+1)$ fraction of the words, and applied our estimator to estimate the number of *new* words that appear at least μ times in the remaining text. Since n is relatively small, we considered $\mu \leq 4$. We simply choose $r = \frac{\log n}{a}$ in the experiments.
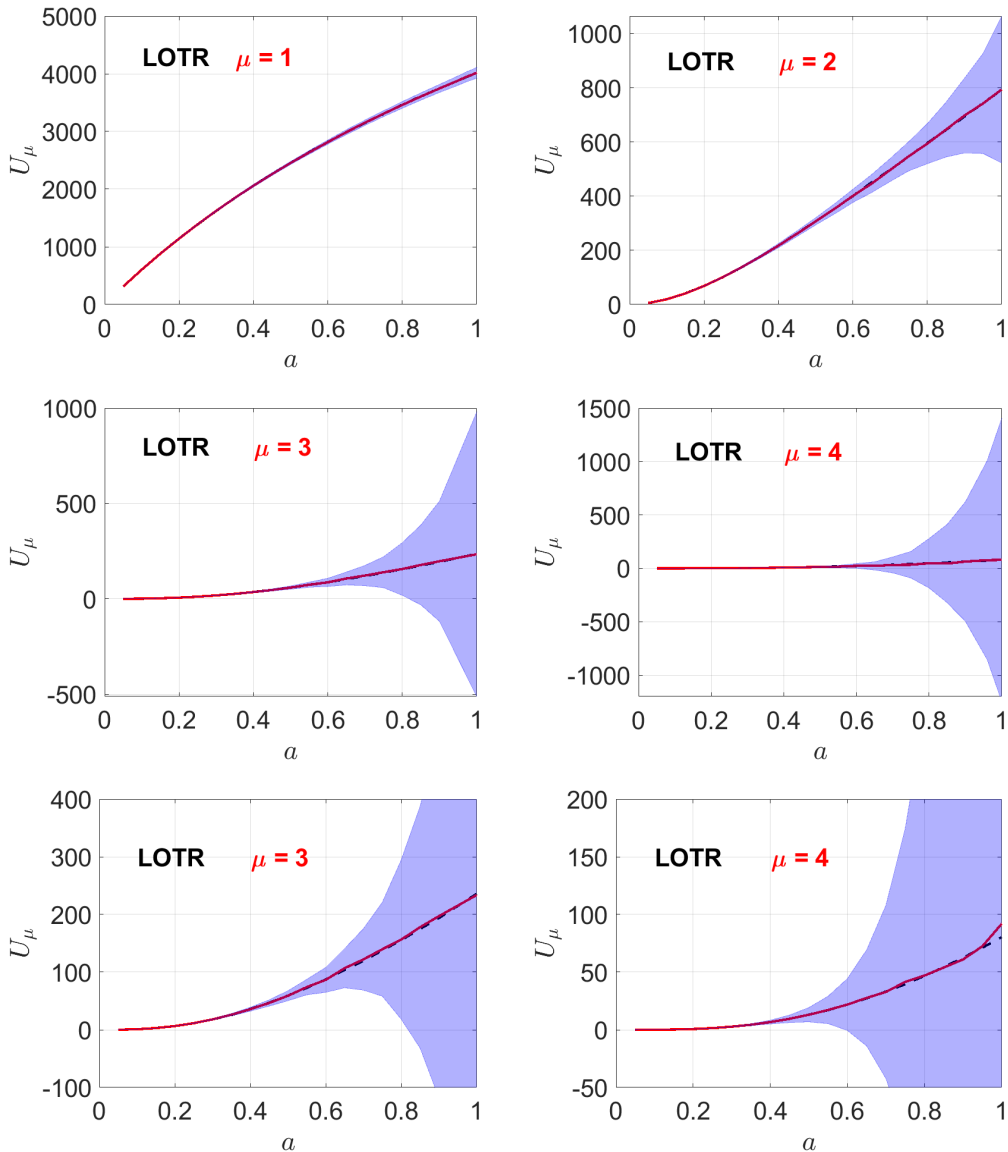

Figure 2: We presents the results for $\mu \in \{1, 2, 3, 4\}$ and the sample amplification ratio $a \leq 1$. The dashed (black) curves represent the truth while the solid (red) curves represent our estimate (the mean). The shaded areas illustrate the standard deviation from the mean. For $\mu = 3$ and $\mu = 4$, since the standard deviations become so large, we repeat them (the bottom two plots) by restricting the y-axes to be smaller, to better visualize the results.

As shown in Figure 2, the proposed estimator provides accurate estimates for $a \leq 1$ and $\mu = 1$. For $\mu = 2$, we can see that the standard deviation becomes much larger when a approaches 1. For $\mu = 3$ and $\mu = 4$, the estimates are still fairly accurate when $a \leq 0.7$. However, for larger a values, the standard deviations become so large that the estimates may not be very reliable.

Figure 3 presents additional plots for $\mu = 1$ and $\mu = 2$, and for a up to 4. For $\mu = 1$, the the estimates are still fairly reliable when $a \leq 2$.
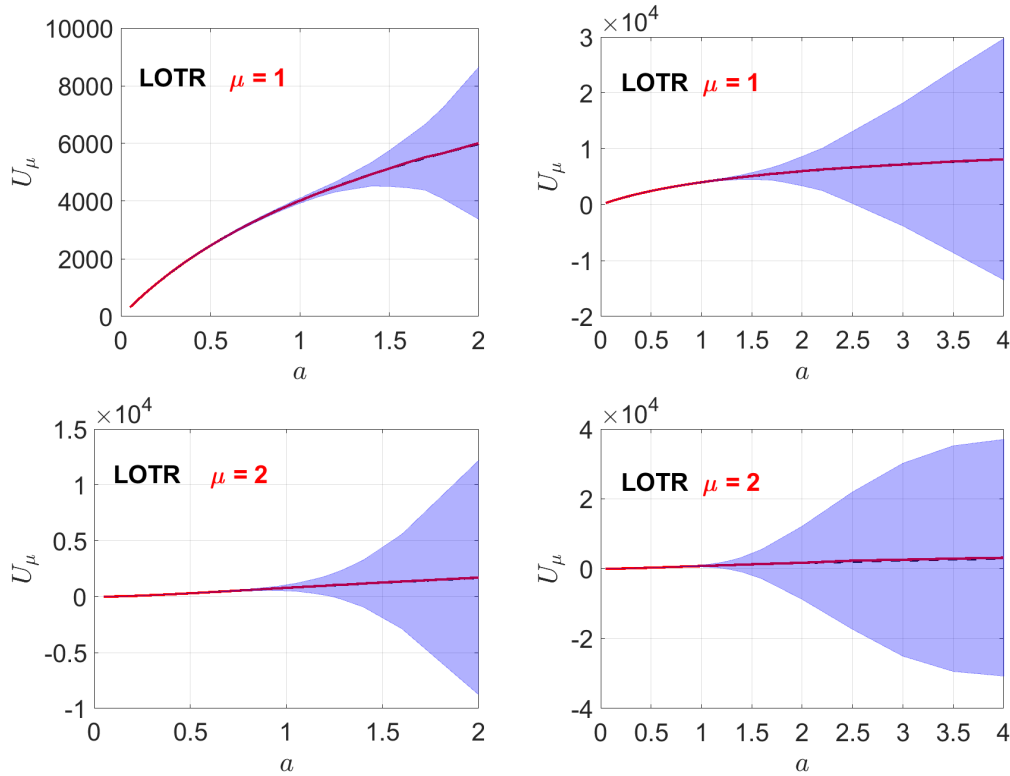

Figure 3: We presents the results for $\mu \in \{1, 2\}$ and the sample amplification ratio a up to 4. The dashed (black) curves represent the truth while the solid (red) curves represent our estimate (the mean). The shaded areas illustrate the standard deviation from the mean.

J   Derivation of the Generalized G-T Estimator for $\mu \geq 1$

In this section, we derive a new estimator which generalizes the G-T estimator for $\mu \geq 1$:

$$\hat{U}_{\text{GGT}} \triangleq (-1)^{\mu} \sum_{i=\mu}^{n} (-a)^{i} \binom{i-1}{\mu-1} \Phi_{i}$$

We obtain this estimator by letting $r \to \infty$ in the proposed estimator using Poisson smoothing:

$$\hat{U}_{\mu} \triangleq \hat{U}_{\mu}(X^{n}, a) \triangleq \sum_{i=1}^{n} s_{i} \cdot \Phi_{i}$$

where

$$s_{i} \triangleq -\sum_{j=0}^{(\mu-1)\wedge i} (-a)^{i} (-1)^{j} \binom{i}{j} \Pr(\text{Poi}(r) \geq i+j)$$

When $r \to \infty$, we have $\Pr(\text{Poi}(r) \geq i+j) = 1$, and the coefficients become

$$s_{i} \triangleq -(-a)^{i} \sum_{j=0}^{(\mu-1)\wedge i} (-1)^{j} \binom{i}{j} = -(-a)^{i} h_{\mu,i} \quad \text{where } h_{\mu,i} = \sum_{j=0}^{(\mu-1)\wedge i} (-1)^{j} \binom{i}{j}$$

The estimator can then be re-written as

$$\hat{U}_{\mu} \triangleq -\sum_{i=1}^{n} (-a)^{i} h_{\mu,i} \cdot \Phi_{i}$$

We can prove that, when $i \leq \mu - 1$, we always have $h_{\mu,i} = 0$, because

$$h_{\mu,i} = \sum_{j=0}^{i} (-1)^{j} \binom{i}{j} = \sum_{j=0}^{i} (-1)^{j} (1)^{i-j} \binom{i}{j} = (1-1)^{i} = 0$$

Therefore, we can write

$$h_{\mu,i} = \begin{cases} 0 & i \leq \mu - 1 \\ \sum_{j=0}^{\mu-1} (-1)^{j} \binom{i}{j} & i > \mu - 1 \end{cases}$$

When $\mu = 3$ and $i > 2$, we have

$$h_{3,i} = 1 - i + \frac{i(i-1)}{2} = \frac{(i-1)(i-2)}{2} = \binom{i-1}{2}$$

When $\mu = 4$ and $i > 3$, we have

$$h_{4,i} = -(i-1) + \frac{i(i-1)}{2} - \frac{i(i-1)(i-2)}{6} = -\frac{(i-1)(i-2)(i-3)}{3 \cdot 2} = -\binom{i-1}{3}$$

Assume that, for $i > \mu - 1$,

$$h_{\mu,i} = (-1)^{\mu-1} \binom{i-1}{\mu-1}$$

Then

$$\begin{aligned} h_{\mu+1,i} &= (-1)^{\mu-1} \binom{i-1}{\mu-1} + (-1)^{\mu} \binom{i}{\mu} = (-1)^{\mu} \left\{ \binom{i}{\mu} - \binom{i-1}{\mu-1} \right\} \\ &= (-1)^{\mu} \left\{ \frac{i!}{\mu!(i-\mu)!} - \frac{(i-1)!}{(\mu-1)!(i-\mu)!} \right\} \\ &= (-1)^{\mu} \frac{(i-1)!}{\mu!(i-\mu-1)!} \left\{ \frac{i}{i-\mu} - \frac{\mu}{i-\mu} \right\} = (-1)^{\mu} \binom{i-1}{\mu} \end{aligned}$$

This proves, by induction, that $h_{\mu,i} = (-1)^{\mu-1} \binom{i-1}{\mu-1}$, and hence we can write the estimator as

$$\hat{U}_{\text{GGT}} = (-1)^{\mu} \sum_{i=\mu}^{n} (-a)^{i} \binom{i-1}{\mu-1} \Phi_{i}$$