Chebyshev polynomials, moment matching, and optimal estimation of the unseen

Yihong Wu and Pengkun Yang*

April 7, 2015

Abstract

We consider the problem of estimating the support size of a discrete distribution whose minimum non-zero mass is at least $\frac{1}{k}$. Under the independent sampling model, we show that the minimax sample complexity to achieve an additive error of $\epsilon k$ with probability at least 0.5 is within universal constant factors of $\frac{k}{\log k} \log^2 \frac{1}{\epsilon}$, which improves the state-of-the-art result $\frac{k}{\epsilon^2 \log k}$ due to Valiant and Valiant. The optimal procedure is a linear estimator based on the Chebyshev polynomial and its approximation-theoretic properties. We also study the closely related species problem where the goal is to estimate the number of distinct colors in an urn containing $k$ balls from repeated draws. While achieving an additive error proportional to $k$ still requires $\Omega(\frac{k}{\log k})$ samples, we show that with $\Theta(k)$ samples one can strictly outperform a general support size estimator using interpolating polynomials.

1 Introduction

Estimating the support size is a classical problem in statistics with widespread applications. For example, a major task for ecologists is to estimate the number of species [FCW43] from field experiments and linguists are interested in estimating the vocabulary size of Shakespeare based on his complete works [ET76]. Estimating the support size is equivalent to estimating the number of unseen symbols [GS04], which is particularly challenging when the sample size is relatively small compared to the total population size, since a significant portion of the population are never observed in the data.

There is a vast amount of literature devoted to the support size estimation problem. In parametric settings, prior knowledge on the distribution such as uniformity [LP56, HW01] is assumed and traditional estimators such as maximum likelihood estimator (MLE), minimum variance unbiased estimator (MVUE) are frequently used [Har68, MSJ82, Sam68, ET76, LP56, HW01]. When difficult to postulate or justify a suitable prior assumption, various nonparametric approaches are adopted such as the Good-Turing estimator [Goo53], Jackknife estimator [BO79], empirical Bayes approach (e.g., Good-Toulmin estimator [GT56]), and several procedures due to Chao and Lee [Cha84, CL92]. Despite their practical popularity, little is known about the optimality of these estimators or the statistical limits of the problem, until recently Valiant and Valiant [VV11] showed that the sample complexity of estimating the support size within a small fraction of the alphabet size is sub-linear. However, the performance guarantee of linear programming-based estimators in [VV11, VV13], which are very similar to [ET76, Program 2], are still far from being optimal.

*The authors are with the Department of Electrical and Computer Engineering and the Coordinated Science Lab, University of Illinois at Urbana-Champaign, Urbana, IL, {yihongwu,pyang14}@illinois.edu.
We adopt the same problem formulation as in [RRSS09]. Let \( P \) be a discrete distribution over some countable alphabet. Without loss of generality, we assume the alphabet is \( \mathbb{N} \) and denote \( P = (p_1, p_2, \ldots) \). Given \( n \) i.i.d. samples \( X \triangleq (X_1, \ldots, X_n) \) drawn from \( P \), the goal is to estimate the support size

\[
S(P) \triangleq \sum_i 1_{\{p_i > 0\}}.
\]  

(1)

To estimate a functional of an unknown distribution, a sufficient statistic is the histogram of the samples, denoted by \( N = (N_1, N_2, \ldots) \) and

\[
N_i = \sum_{j=1}^n 1_{\{X_j = i\}}.
\]

Therefore \( N \) has a multinomial distribution with parameter \( n \) and \( P \). It is clear that unless we impose certain assumptions on the distribution \( P \), it is impossible to estimate \( S(P) \) within a given accuracy, for otherwise there can be arbitrarily many small masses in the support of \( P \) and the risk for estimating \( S(P) \) is obviously infinite. To this end, we restrict our attention on the parameter space \( D_k \), which consists of all probability distributions on \( \mathbb{N} \) whose minimum non-zero mass is at least \( \frac{1}{k} \); consequently \( S(P) \leq k \) for any \( P \in D_k \). The decision-theoretic fundamental limit of this problem is given by the minimax quadratic risk:

\[
R^*(k,n) \triangleq \inf_{\hat{S}} \sup_{P \in D_k} \mathbb{E}(\hat{S} - S(P))^2,
\]

(2)

where \( \hat{S} \) is measurable with respect to the samples \( X_1, \ldots, X_n \).

Our main result is the following: \(^1\)

\textbf{Theorem 1.} For all \( k, n \geq 2 \), \( R^*(k,n) \asymp k^2 \exp(-r^*(k,n)) \), where

\[
r^*(k,n) \asymp \sqrt{\frac{n \log k}{k}} \lor \frac{n}{k}.
\]

(3)

The lower and upper bound part of Theorem 1 is presented in Section 2 and 3, respectively. To interpret (3), we note that \( r^*(k,n) \asymp \frac{n}{k} \) if and only if \( n \geq k \log k \), which is achieved simply by the plug-in estimator

\[
\hat{S}_\text{seen} \triangleq \sum_i 1_{\{N_i > 0\}},
\]

(4)

that is, the number of observed symbols. Furthermore, \( \hat{S}_\text{seen} \) is consistent if \( \frac{n}{k \log k} \) is sufficiently large since \( \mathbb{P}(\hat{S}_\text{seen} \neq S) \leq \mathbb{E}(\hat{S}_\text{seen} - S)^2 \to 0 \), which can be interpreted as the coupon collector’s problem [MR10]. In contrast, the non-trivial regime is when the samples are relatively scarce, namely \( n \leq k \log k \), in which case \( r^*(k,n) \asymp \sqrt{\frac{n \log k}{k}} \). This rate is achieved by a linear estimator based on the Chebyshev polynomial and its approximation-theoretic properties, which is more sophisticated than the plug-in estimator.

Next we discuss the sample complexity of estimating the support size, denoted by \( n^*(k, \epsilon) \), which is defined as the minimal sample size \( n \) such that there exists an integer-valued estimator \( \hat{S} \)

\(^1\) For any sequences \( \{a_n\} \) and \( \{b_n\} \) of positive numbers, we write \( a_n \gtrsim b_n \) or \( b_n \lesssim a_n \) when \( a_n \geq cb_n \) for some absolute constant \( c \). Finally, we write \( a_n \asymp b_n \) when both \( a_n \gtrsim b_n \) and \( a_n \lesssim b_n \) hold.
measurable to the samples $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$, such that $\mathbb{P}[|\hat{S} - S(P)| > \epsilon k] \leq 0.5$ for any $P \in \mathcal{D}_k$. Clearly, since $\hat{S} - S \in \mathbb{Z}$, the only interesting cases are $\epsilon = 0$ or $\epsilon \geq \frac{1}{k}$, corresponding to exact or approximate recovery of the support size respectively. The next result characterizes the sample complexity within universal constants.

**Theorem 2.** If $\epsilon \geq \frac{1}{k}$, then

$$n^*(k, \epsilon) \asymp \frac{k}{\log k} \log^2 \frac{1}{\epsilon}. \quad (5)$$

If $\epsilon = 0$, then $n^*(k, 0) \asymp k \log k$.

Compared to Theorem 1, the only difference is that here we are dealing with the zero-one loss $1_{\{|S - \hat{S}| > \epsilon k\}}$ instead of the quadratic loss $|S - \hat{S}|^2$. In the proof we shall obtain upper bound for the quadratic risk and lower bound for the zero-one loss, thereby proving both Theorem 1 and 2 simultaneously. Furthermore, the choice of 0.5 as the probability of error in the definition of the sample complexity is entirely arbitrary; replacing it by any constant $\delta \in (0, 1)$ only affects $n^*(k, \epsilon)$ up to constant factors.

Previously, the linear programming estimator in [VV11, Corollary 11] shows that $n^*(k, \epsilon) \lesssim k \epsilon^2 \log k$ samples for any arbitrary $\epsilon > 0$, which has subsequently been improved to $\frac{k}{\epsilon^2 \log k}$ in [VV13, Theorem 2, Fact 9]. The lower bound $n^*(k, \epsilon) \gtrsim \frac{k}{\log k}$ in [VV10, Corollary 9] provides no dependence on $\epsilon$. From Theorem 2 we see that the dependence of sample complexity on the accuracy $\epsilon$ can be improved from polynomial to polylogarithmic, which turns out to be optimal.

The paper is organized as follows: In Section 2 we prove the lower bound part of Theorem 1 and 2. In Section 3 we construct an estimator based on Chebyshev polynomials which achieve the optimal minimax rate. Section 4 considers the closely related species problem which can be viewed as a special case of the general support size problem. In Section 5 we apply our estimators to both synthetic and real data and compare the performance with existing methodologies.

**Notations** Let $\text{Poi}(\lambda)$ denote the Poisson distribution with mean $\lambda$ whose probability mass function is denoted by $\text{poi}(\lambda, j) \triangleq \frac{\lambda^j e^{-\lambda}}{j!}, j \geq 0$. Given a positive random variable $U$, denote the Poisson mixture with respect to the distribution of $U$ by $\mathbb{E}[\text{Poi}(U)]$, whose probability mass function is given by $\frac{1}{k} \mathbb{E}[U^k e^{-U}], k \geq 0$. Let $\text{Bern}(p) = p \delta_1 + (1-p) \delta_0$ denote the Bernoulli distribution. The total variation and Kullback-Leibler divergence between probability measures $P$ and $Q$ are denoted by $\text{TV}(P, Q) \triangleq \frac{1}{2} \int |dP - dQ|$ and $D(P \| Q) \triangleq \int dP \log \frac{dP}{dQ}$ respectively.

## 2 Minimax lower bound

The strategy to prove the lower bound in Theorem 1 is similar to [WY14], which is divided into the case $n \gtrsim k \log k$ and $n \lesssim k \log k$ in Proposition 1 and 2, respectively.

**Proposition 1.** Assume $n \geq \alpha k \log k$ for some constant $\alpha > 0$. For all $k \geq 2$, there exists $C$ depending only on $\alpha$ such that

$$R^*(k, n) \gtrsim k^2 \exp \left(-C \frac{n}{k}\right). \quad (6)$$

**Proposition 2.** There exist universal constants $\alpha, C$ such that for all $n \leq \alpha k \log k$

$$R^*(k, n) \gtrsim k^2 \exp \left(-C \sqrt{\frac{n \log k}{k}}\right). \quad (7)$$
Proposition 1 follows from a simple application of Le Cam’s two-point method [LC86] by considering two possible distributions, namely Bern(0) and Bern($\frac{1}{k}$). Their support sizes differ by one while the generated $n$ samples are statistically indistinguishable. Any estimator then necessarily suffers a risk proportional to the statistical distance.

To prove Proposition 2 we follow the idea in [WY14]: We first construct two positive unit-mean random variables $U,U'$ then draw two random vectors

$$P = \frac{1}{k}(U_1, \ldots, U_k), \quad P' = \frac{1}{k}(U'_1, \ldots, U'_k),$$

(8)

where $U_i, U'_i$ are i.i.d. copies of $U, U'$, respectively. Although $P$ and $P'$ need not be probability distributions, as long as the standard deviation of $U, U'$ are not too big, the law of large numbers ensures that $P, P'$ are supported in a small neighborhood near the probability simplex, which we refer as the set of approximate probability distributions. We then show that the minimax risk over this set is close to that over the original probability simplex. This allows us to use $P, P'$ as priors and apply generalized Le Cam’s method, which involves testing two composite hypothesis as to whether the data are generated from $P$ or $P'$. Note that, with high probability, $S(P)$ and $S(P')$ differ by the separation in their mean values:

$$E[S(P)] - E[S(P')] = k(P[U > 0] - P[U' > 0])$$

$$= k(P[U' = 0] - P[U = 0]).$$

(9)

To simplify the argument we also apply the Poissonization technique where the sample size is a Pois($n$) random variable instead of a fixed number $n$. This does not change the statistical nature of the problem due to the concentration of Pois($n$) around its mean $n$. Given $P$ and under Poisson sampling, the sufficient statistic $N_i \sim$ Pois($nP_i$). Therefore through the i.i.d. construction in (8), $N_i \sim E[Poi(\frac{n}{k}U)]$ and similarly $N'_i \sim E[Poi(\frac{n}{k}U')]$. The indistinguishability follows if the total variation distance between the two product Poisson mixtures is strictly bounded away from one, for which it suffices to show

$$TV(E[Poi(nU/k)], E[Poi(nU'/k)]) \leq \frac{c}{k},$$

(10)

for some constant $c < 1$.

The above construction provides a recipe for the lower bound. To optimize the ingredients boils down to the following one-dimensional optimization problem: Construct two priors $U, U'$ with unit mean which maximize the difference $P[U' = 0] - P[U = 0]$ subject to the total variation distance constraint (10), which, in turn, can be guaranteed by moment matching, i.e., ensuring $U$ and $U'$ have many identical moments., and the $L_\infty$-norm of $U, U'$. To summarize, our lower bound entails solving the following optimization problem:

$$\sup \ P[U' = 0] - P[U = 0]$$

s.t. $E[U] = E[U'] = 1$

$$E[U^j] = E[U'^j], \quad j = 1, \ldots, L + 1,$$

$$U, U' \in \{0\} \cup [1, \lambda]$$

(11)

The final lower bound (7) is obtained from (11) by choosing $L \asymp \log k$ and $\lambda \asymp \frac{k \log k}{n}$.

In order to evaluate the infinite-dimensional linear programming problem (11), by considering the dual program we show (in Appendix F) that (11) coincides exactly with the best uniform
approximation error of function $x \mapsto \frac{1}{x}$ over the interval $[1, \lambda]$ by degree-$L$ polynomials:

$$\inf_{p \in P_L} \sup_{x \in [1, \lambda]} \left| \frac{1}{x} - p(x) \right|,$$

where $P_L$ denotes the set of polynomials of degree $L$. The problem of best polynomial approximation has been well-studied, c.f. [Tim63, DS08]; in particular, the exact formula for the best polynomial that approximates $x \mapsto \frac{1}{x}$ and the optimal approximation error have been obtained in [Tim63, Sec. 2.11.1].

3 Optimal estimator via Chebyshev polynomials

In this section we prove the upper bound part of Theorem 1 and describe the rate-optimal support size estimator. By the same argument in [WY14, Section 2], the risk in (2) is less than a constant factor of that under the Poisson sampling. Therefore in the proof we continue to use Poissonized sample size. Analogous to the lower bound part, we can obtain the upper bound in Theorem 1 by Proposition 3 for the regime of $n \gtrsim k \log k$ and Proposition 4 for $n \lesssim k \log k$ separately. The next proposition provides an upper bound of the risk of plug-in estimator:

**Proposition 3.** For all $n, k \geq 1$,

$$\sup_{P \in \mathcal{P}_k} \mathbb{E}(S(P) - \hat{S}_{\text{seen}}(N))^2 \leq 2k^2 \exp \left( - \frac{n}{k} \right),$$

where $N = (N_1, N_2, \ldots)$ and $N_i \overset{\text{ind}}{\sim} \text{Poi}(np_i)$.

To specify the optimal estimator in the regime of $n \gtrsim k \log k$, we first introduce shifted Chebyshev polynomials. Let $c_0 < c_1$ be constants to be specified. Define $L \triangleq \lfloor c_0 \log k \rfloor$, $r \triangleq \frac{c_1 \log k}{n}$ and $l \triangleq \frac{1}{k}$. Recall that the usual Chebyshev polynomial of degree $L$ is

$$T_L(x) = \cos(L \arccos x) = \frac{1}{2}((x - \sqrt{x^2 - 1})^L + (x + \sqrt{x^2 - 1})^L).$$

The shifted Chebyshev polynomial over the interval $[l, r]$ is given by

$$P_L(x) = -\frac{\cos(L \arccos(\frac{2}{r-l}x - \frac{r+l}{r-l}))}{\cos(L \arccos(-\frac{r+l}{r-l}))} \triangleq \sum_{m=0}^L a_m x^m,$$

which satisfies $P_L(0) = -1$ and hence $a_0 = -1$; the remaining coefficients $a_1, \ldots, a_L$ can be obtained by the formula for Chebyshev polynomial [Tim63, 2.9.12] and the binomial expansion.

Let

$$g_L(j) = \begin{cases} \frac{a_j j!}{j^j} + 1, & j \leq L, \\ 1, & j > L. \end{cases}$$

Obviously $g_L(0) = 0$ since $a_0 = -1$ by design. Define a preliminary estimator by

$$\tilde{S} = \sum_i g_L(N_i).$$
We proceed to explain the reasoning behind the estimator (15). Note that the bias is $\mathbb{E}[\hat{S} - S] = \sum_i \mathbb{E}[g_L(N_i) - 1_{(p_i>0)}]$. Since $g_L(0) = 0$ and $g_L(j) = 1$ for $j > L$, each term in the bias can be written as

$$
\mathbb{E}[g_L(N_i) - 1_{(p_i>0)}] = \mathbb{E}[(g_L(N_i) - 1)1_{(p_i>0)}1_{(N_i \leq L)}] = \sum_{j=0}^{L} e^{-np_i} \frac{(np_i)^j}{j!} a_j \frac{1}{n^j} 1_{(p_i>0)} = e^{-np_i} P_L(p_i) 1_{(p_i>0)}
$$

where $P_L$ is the degree-$L$ polynomial defined in (13). The intuition is that whenever $N_i \leq L = [c_0 \log k]$, then with high probability the corresponding mass must satisfy $p_i \leq \frac{c_1 \log k}{n}$. That is, if $p_i > 0$ and $N_i \leq L$ then $p_i \in \left[\frac{1}{k}, \frac{c_1 \log k}{n}\right]$, and hence $P_L(p_i)$ is bounded by the sup norm of $P_L$ over the interval $\left[\frac{1}{k}, \frac{c_1 \log k}{n}\right]$. Using the properties of Chebyshev polynomials [Tim63, 2.13.14], we can show that (13) is the unique degree-$L$ polynomial that passes through the point $(0, -1)$ and deviates the least from zero over $\left[\frac{1}{k}, \frac{c_1 \log k}{n}\right]$. This explains the coefficients (14) which are chosen to minimize the bias.

Finally, in view of the fact that $S(P) \in [0, k]$, we define our estimator by $\hat{S} \triangleq (\hat{S} \lor 0) \land k$. The next proposition gives an upper bound of its quadratic risk:

**Proposition 4.** If $n \leq \frac{c_0}{2} k \log k$, there exists universal constant $c$ such that

$$
\sup_{P \in \mathcal{D}_k} \mathbb{E}(\hat{S}(N) - S(P))^2 \lesssim k^2 \exp\left(-c \sqrt{\frac{n \log k}{k}}\right),
$$

where $N = (N_1, N_2, \ldots) \overset{ind}{\sim} \text{Poi}(np_i)$.

Our estimator (15) belong to the family of linear estimators:

$$
\hat{S} = \sum_i f(N_i) = \sum_{j \geq 1} h_j f(j),
$$

which a linear combination of the fingerprints $h_j = \sum_i 1_{(N_i = j)}$, i.e., the number of items that appear exactly $j$ times. It should be noted that for estimating the support size (or other permutation-invariant functional of the distribution), the fingerprints form a sufficient statistic which is a further summary of the histogram $N$. Other notable examples of linear estimators include

- **Plug-in estimator (4):** $\hat{S}_{\text{seen}} = h_1 + h_2 + \ldots$

- **Good-Toulmin (empirical Bayes) [GT56]:** for some $t > 0$,

$$
\hat{S}_{\text{GT}} = \hat{S}_{\text{seen}} + th_1 - t^2 h_2 + t^3 h_3 - t^4 h_4 + \ldots
$$

- **Efron-Thistle (Bayesian) [ET76]:** for some $t > 0$ and $J \in \mathbb{N}$,

$$
\hat{S}_{\text{ET}} = \hat{S}_{\text{seen}} + \sum_{j=1}^{J} (-1)^{j+1} t^j b_j h_j,
$$

where $b_j = \mathbb{P}[	ext{Binomial}(J, 1/(t + 1)) \geq j]$. 


By definition, our estimator (15) can be written as
\[ \tilde{S} = \sum_{j=1}^{L} g_L(j)h_j + \sum_{j>L} h_j. \]
By (14), \( g_L \) is also a polynomial of degree \( L \), which is oscillating and results in coefficients with alternating signs (see Fig. 1). Interestingly, this behavior, which is counterintuitive, coincide with many classical estimators, such as (19) and (20).

Remark 1. The technique of polynomial approximation has been previously used for estimating non-smooth functions (\( L_q \)-norms) in Gaussian models [INK87, LNS99, CL11] and more recently for estimating information quantities (entropy and power sums) on large discrete alphabets [WY14, JVW14]. The design principle is to approximate the non-smooth function on a given interval using algebraic or trigonometric polynomials for which unbiased estimators exist and choose the degree to balance the bias (approximation error) and the variance (stochastic error). In many situations, the construction of the estimator is a two-stage process involving sample splitting: First, use half of the sample to test whether the corresponding parameter lies in the interval; Second, use the remaining samples to construct unbiased estimator for the approximating polynomial if the parameter lies in the given interval or apply plug-in estimators otherwise (see, e.g., [WY14, JVW14] and [CL11, Section 5]). While the benefit is to facilitate analysis due to independence of the two subsamples, the downside is obviously sacrificing the statistical accuracy. Here to estimate the support size, we forgo sample splitting and directly design a linear estimator. Instead of using a polynomial as a proxy for the original function and then constructing its unbiased estimator, the best polynomial approximation only arises as a natural step in our analysis of the bias.

4 Species problem

Closely related to estimating the support size is the species problem [Lo92, BF93], which can be described as follows: Given an urn containing \( k \) balls, the goal is to estimate the number of
distinct colors by sampling the balls with or without replacement. This problem is also known as the \textsc{Distinct-Elements} problem in the computer science literature [CCMN00, BYJK+02, RRSS09], where adaptive sampling algorithms are also allowed. Originating in ecology and numismatics, this problem now frequently arises in large-scale databases and data mining [RRSS09, CCMN00]. Previous work mainly focuses on multiplicative error where the optimal sample size to estimate within a factor of $1 \pm \epsilon$ is shown to be $\Theta(k/\epsilon^2)$ [CCMN00, BYKS01]. Instead, our goal here is to investigate the sample complexity for achieving an additive error, say, $\epsilon k$.

Similar to (2), we define the minimax risk for the species problem using samples with and without replacement by

$$R^*_m(k, n) \triangleq \inf_{\hat{S}} \sup_{P \in D'_k} \mathbb{E}(\hat{S}(N) - S(P))^2,$$

$$R^*_h(k, n) \triangleq \inf_{\hat{S}} \sup_{P \in D'_k} \mathbb{E}(\hat{S}(N) - S(P))^2,$$  \hspace{1cm} (21)

respectively. Here the histogram $N$ is distributed according to a multinomial (resp. multivariate hypergeometric) distribution in $R^*_m$ (resp. $R^*_h$). The parameter space $D'_k$ denotes the set of all probability distribution $P = (p_1, p_2, \ldots)$ whose masses are multiples of $1/k$, i.e., $p_i \in \{j/k : j = 0, \ldots, k\}$. From this perspective, the species problem using samples with replacement is a special case of the support size problem formulated in Section 1. Since $D'_k \subset D_k$ by definition, we have $R^*_m(k, n) \leq R^*(k, n)$. On the other hand, since we can simulate samples with replacement from samples without replacement, we have

$$R^*(k, n) \geq R^*_m(k, n) \geq R^*_h(k, n).$$  \hspace{1cm} (22)

Therefore we can always apply the support size estimator in Section 3 in species problem with the same performance guarantee in Proposition 4, which implies that $O\left(\frac{k}{\log k}\right)$ samples are sufficient to achieve an additive error of $\epsilon k$ for any constant $\epsilon$.

Next we give a lower bound $R^*_h$ by relating it to $R^*$. The argument is based on the idea in [RRSS09, Lemma 3.3] (see also [Val12, Lemma 5.14]): Given a distribution $P$ whose minimum non-zero mass is at least $1/k$ for some intermediate parameter $l \in \mathbb{N}$, we draw $k$ i.i.d. samples, whose realizations form an instance of an urn containing $k$ balls. Next we draw $n$ samples from this urn without replacement ($n \leq k$) and apply any estimator for the species problem. Note that these $n$ samples are distributed as if they are independently sampled from the original distribution $P$. This results a support size estimator using $n$ i.i.d samples. Theorem 3 formalizes this intuition.

**Theorem 3.** If $k \geq n$, for any $l$,

$$R^*_h(k, n) \geq \frac{1}{2} R^*(l, n) - R_{\text{plug-in}}(l, k).$$  \hspace{1cm} (23)

Consequently, there exist universal constants $\rho < 1$ and $c$ such that if $n \leq \rho k$ then

$$R^*_h(k, n) \geq k^2 \exp\left(-\frac{cn \log k}{k}\right).$$  \hspace{1cm} (24)

**Proof.** Denote by $\hat{S}(k, \cdot)$ an estimator for the species problem using samples from a $k$-ball urn without replacement. Consider the problem of support size estimation using $n$ i.i.d. samples with minimum non-zero mass $1/k$. We construct a (randomized) estimator using $\hat{S}$ as follows: For any $P \in D_l$, let $X_1, \ldots, X_k \overset{\text{i.i.d.}}{\sim} P$, whose histogram is denoted by the random vector $N \sim \text{Multinomial}(k, P)$. Let $Y_1, \ldots, Y_n$ be sampled from $X_1, \ldots, X_k$ uniformly without replacements and denote their histogram by $N'$. Conditioned on $N$, $N'$ has a multivariate hypergeometric distribution with parameters $n$
and $N/k$; on the other hand, the marginal distribution of $N$ is Multinomial($n, P$). Applying the triangle inequality, we obtain

$$\mathbb{E}[(S(P) - \tilde{S}(k, N'))^2] \leq 2\mathbb{E}[(S(P) - S(N/k))^2] + 2\mathbb{E}[(S(N/k) - \tilde{S}(k, N'))^2]$$

$$\leq 2R_{\text{plug-in}}(l, k) + 2\mathbb{E}[(S(N/k) - \tilde{S}(k, N'))^2 | N]$$

$$\leq 2R_{\text{plug-in}}(l, k) + 2 \sup_{P \in \mathcal{P}_k} \mathbb{E}((\hat{S} - S(P))^2).$$

Therefore for any estimator $\hat{S}$, we have

$$R^*(l, n) \leq \sup_{P \in \mathcal{P}_k} \mathbb{E}[(S(P) - \hat{S}(k, N))^2] \leq 2R_{\text{plug-in}}(l, k) + 2 \sup_{P \in \mathcal{P}_k} \mathbb{E}(\hat{S} - S(P))^2.$$

Taking infimum over all $\hat{S}$ yields the desired (23).

Finally, plugging in the minimax lower bound in (3) under the condition that $n \leq l \log l$, and the upper bound of empirical estimator in (12), then optimizing over the free parameter $l$, we obtain the following corollary:

**Remark 2.** The universal constant $\rho$ in Theorem 3 cannot be $\rho \geq 1$. The RHS in (24) is always positive, while obviously $R_{\text{h}}^*(k, k) = 0$ since we can tell exactly the number of colors when every ball has been accessed. On the other hand, the lower bound in (24) still holds in the special case that $n = k - 1$. By definition $R_{\text{h}}^*(k, k - 1) \leq 1$. Let two urns be $N_1 = (0, 2, 1, \ldots, 1)$ and $N_2 = (1, 1, 1, \ldots, 1)$, then applying Le Cam’s two points method yields that $R_{\text{h}}^*(k, k - 1) \asymp \frac{1}{k}$.

Combining Theorem 1, (22) and Theorem 3, we obtain

$$k^2 \exp\left(-C\sqrt{\frac{n \log k}{k}}\right) \geq R_m^*(k, n) \geq R_{\text{h}}^*(k, n) \asymp k^2 \exp\left(-c\frac{n \log k}{k}\right)$$

for some universal constant $c$ and $C$, whenever $n \leq \rho k$. This implies that for the species problem, the sample complexity for achieving an additive error $\epsilon k$ with constant $\epsilon$ is still $\Theta\left(\frac{k}{\log k}\right)$. However, the dependence on $\epsilon$ is unclear. Nevertheless, next we show that when $n \asymp k$ we can actually achieve the lower bound in (25).

The main observation is that, in the species problem, the probability mass only takes a discrete set of values as opposed to a continuum of values in the general support size problem. Therefore, instead of applying the Chebyshev polynomial to best approximate the step function as done in Section 3, we can use a Lagrange interpolating polynomial to interpolate these points and achieve exactly zero bias. To this end, let $c_0 < c_1$ be constants to be specified later that may differ from those in Section 3. Let $L = \lceil c_0 \log k \rceil$. Define

$$q_L(x) \triangleq -\frac{\prod_{j=1}^L (j - x)}{L!} = \sum_{j=1}^L b_j x^j,$$ \hspace{1cm} (26)

which satisfies $q_L(0) = -1$ and $q_L(j) = 0$ for $j = 1, \ldots, L$. Now we define $g_L$ as in (14) with those new coefficients $b_m$ that

$$g_L(j) \triangleq \begin{cases} 
\frac{b_j k^j j!}{n^j} + 1, & j \leq L, \\
1, & j > L.
\end{cases}$$

Then we define $\tilde{S}$ and $\hat{S}$ the same way as Section 3. $\tilde{S} = \sum_i g_L(N_i), \hat{S} = (\tilde{S} \lor 0) \land k$. The next theorem gives the risk of above estimator $\hat{S}$.
Theorem 4. There exist universal constants $\rho, c$ such that for all $n \geq \rho k$,

$$\sup_{P \in \mathcal{D}'_k} \mathbb{E}(\hat{S}(N) - S(P))^2 \lesssim k^2 \exp(-c \log k),$$

(27)

where $N = (N_1, N_2, \ldots) \overset{\text{ind}}{\sim} \text{Poi}(np_i)$.

Comparing with Theorem 1 where the additive error with $n \approx k$ samples is shown to be $k \exp(-c \sqrt{\log k})$, while in species problem we can achieve $k \exp(-c \log k)$. In other words, in species problem $\Theta(k)$ samples can provide the same performance as $\Theta(k \log k)$ samples in support size estimation. It remains open to determine whether the risk lower bound (25) is optimal for $n$ other than $O\left(\frac{k}{\log k}\right)$ and $\Theta(k)$.

5 Experiments

We evaluate the performance of our estimator numerically on both synthetic and real data in comparison with popular existing procedures. In the experiment we choose the constants $c_0 = 0.8, c_1 = 1$ in (13). The minimum non-zero mass varies in different experimental settings.

For synthetic data we first consider the uniform distribution on one million elements, denoted by Uniform[$10^6$], whose minimum non-zero mass is then $10^{-6}$. The plug-in estimate is simply the number of distinct elements observed in the data. We compare the our results with the Good-Turing estimator \cite{Goo53} and the two estimators proposed by Chao and Lee \cite{CL92}. The result for uniform distribution is shown in Fig. 2 and we can see that Good-Turing estimator is the best while our estimator is converges to truth much faster than plug-in and Chao-Lee estimators. The phenomenon that Good-Turing estimator performs remarkably well in the special case of uniform distributions has been noticed and analyzed in \cite{CL92,DR80}.

\footnote{Here the Good-Turing estimator refers to first estimate the total probability of seen symbols (sample coverage) by $\hat{C} = 1 - \frac{h_1}{n}$ then estimate the support size by $\hat{S} = \frac{S_{\text{seen}}}{\hat{C}}$.}

![Figure 2: Comparison of support size estimators for the uniform distribution over $[10^6]$.](image)

while our estimator is converges to truth much faster than plug-in and Chao-Lee estimators. The phenomenon that Good-Turing estimator performs remarkably well in the special case of uniform distributions has been noticed and analyzed in \cite{CL92,DR80}.
As noted in [CL92] Good-Turing estimator is not always efficient. Next we conduct the experiment for the mixture of a point mass and a uniform distribution: \( \frac{1}{2} \delta_{10^6} + \frac{1}{2} \text{Uniform}[10^6 - 1] \). The minimum non-zero mass is \( \frac{1}{2(10^6-1)} \). In this case the output of Chao-Lee’s estimators is highly unstable and thus the comparison with them is omitted. The result is shown in Fig. 3. As we can see, Good-Turing estimator outperforms the plug-in estimator due to its consideration of sample coverage. However, the convergence rate is much slower than our estimator partly because it only uses the information of how many symbols occurred exactly once \( (h_1) \), instead of the full spectrum of fingerprints. The performance of Chao-Lee estimators turns out to be highly unstable in the presence of mixtures and hence omitted from the plot.

Next we evaluate our estimator by a real data experiment based on the text of *Hamlet*, which contains about 32,000 words in total consisting of about 7,700 distinct words. Here the definition of “distinct word” is any distinguishable arrangement of characters that are delimited by spaces, including letters of lower and upper cases and punctuation. We randomly sample the text with replacement and generate the fingerprints for estimation. The minimum non-zero mass is naturally the reciprocal of the total number of words, \( \frac{1}{32000} \). In this experiment we use the degree-8 Chebyshev polynomial. We also compare our estimator with the one in [VV13]. The results are plotted in Fig. 4, which shows that the estimator in [VV13] has similar convergence rate to ours; however, the computational cost of linear programming is significantly higher than computing linear combinations with pre-determined coefficients. Finally, we feed the fingerprint of entire Shakespearean canon [ET76], which contains 31,534 word types, to our estimator and obtain an estimator of 68,944 for Shakespeare’s vocabulary size, as compared to 66,534 obtained by Efron-Thisted [ET76].

**Acknowledgment**

The authors thank Luca Trevisan for helpful comments pertaining to Theorem 2 and Greg Valiant for discussing [Val12, Lemma 5.14] in his thesis.
A Proof of lower bounds

Proof of Proposition 1. Define two distributions $P = \text{Bern}(0)$ and $Q = \text{Bern}(\frac{1}{k})$. Applying Le Cam’s two-point method [Tsy09] yields that

$$R^*(k, n) \geq \frac{1}{4} (S(P) - S(Q))^2 \exp(-nD(P||Q)) = \frac{1}{4} \exp\left(-\frac{n}{k} (-k \log(1 - \frac{1}{k}))\right).$$

(28)

Note that $-k \log(1 - \frac{1}{k})$ is decreasing when $k \geq 2$, we have

$$R^*(n, k) \geq \frac{1}{4} \exp\left(-\frac{2n}{k}\right) \geq \frac{1}{4} k^2 \exp\left(-\frac{C n}{k}\right),$$

where $C = 2 + \frac{2}{\alpha}$. \hfill \Box

Proof of Proposition 2. First note that it suffices to prove (7) in the regime of $\frac{k}{\log k} \lesssim n \leq \alpha k \log k$, where $\alpha$ is some absolute constant to be specified later. To see this, observe that if $n \lesssim \frac{k}{\log k}$, since by definition $R^*(k, \cdot)$ is monotonic decreasing with respect to the sample size $n$, then

$$R^*(k, n) \gtrsim k^2 \exp\left(-C \sqrt{\frac{n \log k}{k}}\right),$$

where in the last step we used the fact the $\exp\left(-C \sqrt{\frac{n \log k}{k}}\right) \approx 1$ if $0 \leq n \lesssim \frac{k}{\log k}$.

For $0 < \epsilon < 1$, define the set of approximate probability vectors by

$$\mathcal{D}_k(\epsilon) \triangleq \left\{ P = (p_1, p_2, \ldots) : \left| \sum_i p_i - 1 \right| \leq \epsilon, p_i \in \{0\} \cup \left[\frac{1 + \epsilon}{k}, 1\right] \right\}.$$

(29)

which reduces to the original probability distribution space $\mathcal{D}_k$ if $\epsilon = 0$.  

Figure 4: Comparison of different estimates of the total number of distinct words in Hamlet.
Generalizing the minimax quadratic risk to the Poisson sampling model, we define
\[
\tilde{R}^*(k, n, \epsilon) \triangleq \inf \sup_{\tilde{S}', P \in \mathcal{D}_k(\epsilon)} \mathbb{E}(\tilde{S}'(N) - S(P))^2,
\]  
\[ (30) \]
where \( N_i \overset{\text{ind}}{\sim} \text{Poi}(np_i) \) for \( i = 1, \ldots, k \). Nevertheless \( S \) is still a valid functional. The functional \( S \) is still the support size of the normalized \( P \) since \( S(P) = S\left(\frac{P}{\sum P_i}\right) \). The risk defined above is connected to \( R^*(k, n) \) by the following lemma:

**Lemma 1.** For any \( k, n \in \mathbb{N} \) and \( \epsilon < 1/3 \),
\[
R^*(k, n/2) \geq \tilde{R}^*(k, n, \epsilon) - k^2 \exp(-n/50).
\]

To establish a lower bound of \( \tilde{R}^*(k, n, \epsilon) \), we apply generalized Le Cam’s method involving two composite hypothesis. Given two random variables \( U, U' \in [0, k] \) with unit mean we can construct two random vectors by \( P = \frac{1}{k}(U_1, \ldots, U_k) \) and \( P' = \frac{1}{k}(U'_1, \ldots, U'_k) \) with i.i.d. entries. Then \( \mathbb{E}(S(P)) - \mathbb{E}(S(P')) = k(\mathbb{P}[U > 0] - \mathbb{P}[U' > 0]) \). Furthermore, both \( S(P) \) and \( S(P') \) are binomially distributed, which are tightly concentrated at the respective means. We can reduce the problem to the separation on mean values, as shown in the next lemma:

**Lemma 2.** Let \( \lambda > 0 \). Let \( \epsilon = \frac{4\lambda}{\sqrt{k}} \) and \( U, U' \in \{0\} \cup [1 + \epsilon, \lambda] \) be random variables such that \( \mathbb{E}[U] = \mathbb{E}[U'] = 1 \) and \( |\mathbb{P}[U > 0] - \mathbb{P}[U' > 0]| \geq d \). Then
\[
\tilde{R}^*(k, n, \epsilon) \geq \frac{k^2d^2}{32} \left( \frac{7}{8} - k \text{TV}(\mathbb{E}[\text{Poi}(nU/k)], \mathbb{E}[\text{Poi}(nU'/k)]) - \frac{32}{kd^2} \right). \]
\[ (31) \]

Let \( c_0 > 0 \) and \( c_1 > 1 \) be absolute constants to be specified later. Let \( L = \lfloor c_0 \log k \rfloor \) and \( \lambda = \frac{c_1k \log k}{n} > 2 \). Applying Lemma 5 in Appendix F, we obtain two random variables \( U, U' \in \{0\} \cup [2, \lambda] \) such that \( \mathbb{E}[U^j] = \mathbb{E}[U'^j], j = 1, \ldots, L \) and \( \mathbb{P}[U > 0] - \mathbb{P}[U' > 0] = 2E_L(\frac{1}{2}, [2, \lambda]) \). Using \([\text{Tim63}, 2.11.1]\) and change of variables, we get
\[
E_L \left( \frac{1}{x}, [2, \lambda] \right) = \frac{2}{\lambda - 2} E_L \left( \frac{1}{x - \lambda + 2}, [-1, 1] \right) = \frac{1 - 2\lambda^{-1}}{4} \left( \frac{\sqrt{\lambda} - \sqrt{2}}{\sqrt{\lambda} + \sqrt{2}} \right)^L.
\]
Therefore the mean values of support sizes are separated by
\[
\mathbb{P}[U > 0] - \mathbb{P}[U' > 0] = \frac{1 - 2\lambda^{-1}}{2} \left( \frac{\sqrt{\lambda} - \sqrt{2}}{\sqrt{\lambda} + \sqrt{2}} \right)^L \gtrsim \exp \left( -\frac{C}{2} \sqrt{\frac{n \log k}{k}} \right), \]
\[ (32) \]
for some universal constant \( C \). Recall that \( \mathbb{E}[U^j] = \mathbb{E}[U'^j], j = 1, \ldots, L \), where \( L = \lfloor c_0 \log k \rfloor \geq \frac{c_0}{2} \log k \). Let \( c_0 > 4c_1 \) satisfying \( \frac{c_0}{4} \log \frac{c_0}{4c_1} - c_1 > 2 \). Applying \([\text{WY14}, \text{Lemma 3}]\) yields that
\[
\text{TV}(\mathbb{E}[\text{Poi}(nU/k)], \mathbb{E}[\text{Poi}(nU'/k)]) \lesssim k^{-2}.
\]
\[ (33) \]
Since \( n \gtrsim \frac{k}{\log k} \), for sufficiently large \( k \) we have \( \epsilon = \frac{4\lambda}{\sqrt{k}} \leq 1 \). Let \( \alpha \) be such that \( C\sqrt{\alpha} < 1 \). Then the desired lower bound (7) follows from combining Lemma 1 and Lemma 2.
A.1 Proof of lemmas

Proof of Lemma 1. Fix $\delta > 0$. Let $\tilde{S}(\cdot, n)$ be a near-minimax estimator for fixed sample size $n$, i.e.,

$$\sup_{P \in \mathcal{D}_k} \mathbb{E}[(\tilde{S}(N, n) - S(P))^2] \leq \delta + R^*(k, n).$$

(34)

Using these estimators we construct an estimator for the Poisson model. Fix an arbitrary $P = (p_1, p_2, \ldots) \in \mathcal{D}_k(\epsilon)$. Let $N = (N_1, N_2, \ldots)$ with $N_i \overset{\text{ind}}{\sim} \text{Poi}(np_i)$ and let $n' = \sum N_i \sim \text{Poi}(n)$. We construct an estimator for the Poisson sampling model by

$$\tilde{S}(N) = \tilde{S}(N, n').$$

We observe that conditioned on $n' = m$, $N \sim \text{Multinomial}(m, \frac{P}{\sum p_i})$. Recall from the definition of $\mathcal{D}_k(\epsilon)$ that $\sum p_i \leq 1 + \epsilon$ and $p_i \in \{0\} \cup \left[\frac{1+\epsilon}{k}, 1\right]$, then $\frac{P}{\sum p_i} \in \mathcal{D}_k$. Therefore by (34), we have

$$\mathbb{E} \left( \tilde{S}(N) - S \left( \frac{P}{\sum p_i} \right) \right)^2 = \sum_{m=0}^{\infty} \mathbb{E} \left[ \left( \tilde{S}(N,m) - S \left( \frac{P}{\sum p_i} \right) \right)^2 \right] \mathbb{P} \left[ n' = m \right] \mathbb{P} \left[ n' = m \right] + \delta.$$

Now note that for fixed $k$, the minimax risk $n \mapsto R^*(k, n)$ is decreasing and $0 \leq R^*(k, n) \leq k^2$. Since $S(P) = S \left( \frac{P}{\sum p_i} \right)$, $n' = \sum N_i \sim \text{Poi}(n \sum p_i)$ and $|\sum p_i - 1| \leq \epsilon \leq 1/3$, we have

$$\mathbb{E} \left( \tilde{S}(N) - S(P) \right)^2 \leq \sum_{m \geq n/2} R^*(k, m) \mathbb{P} \left[ n' = m \right] + k^2 \mathbb{P} \left[ n' \leq \frac{n}{2} \right] + \delta$$

$$\leq R^*(k, n/2) + k^2 \exp(-n/50) + \delta,$$

where in the last inequality we used the Chernoff bound (see, e.g., [MU05, Theorem 5.4]). By the arbitrariness of $\delta$, the lemma follows. $\square$

Proof of Lemma 2. Define two random vectors

$$P = \left( \frac{U_1}{k}, \ldots, \frac{U_k}{k} \right), \quad P' = \left( \frac{U'_1}{k}, \ldots, \frac{U'_k}{k} \right),$$

(35)

where $U_i$ and $U'_i$ are i.i.d. copies of $U$ and $U'$, respectively. Denote by $N = (N_1, \ldots, N_k)$ and $N' = (N'_1, \ldots, N'_k)$ the corresponding histogram, which satisfy $N_i \overset{\text{ind}}{\sim} \text{Poi}(nU_i/k)$ and $N'_i \overset{\text{ind}}{\sim} \text{Poi}(nU'_i/k)$ conditioned on $P$ and $P'$, respectively. Define the following events:

$$E \triangleq \left\{ \left| \sum_{i} \frac{U_i}{k} - 1 \right| \leq \epsilon, |S(P) - \mathbb{E}[S(P)]| \leq \frac{kd}{4} \right\}, \quad E' \triangleq \left\{ \left| \sum_{i} \frac{U'_i}{k} - 1 \right| \leq \epsilon, |S(P') - \mathbb{E}[S(P')]| \leq \frac{kd}{4} \right\}.$$

Now we define two priors on the set $\mathcal{D}_k(\epsilon)$ by the following conditional distributions:

$$\pi = P_{P|E}, \quad \pi' = P_{P'|E'}.$$
First we consider the separation of the functional value under $\pi$ and $\pi'$. Recall that $\mathbb{E}[S(P)] = k\mathbb{P}[U > 0]$, $\mathbb{E}[S(P')] = k\mathbb{P}[U' > 0]$ and $|\mathbb{P}[U > 0] - \mathbb{P}[U' > 0]| \geq d$ by assumption. By the definition of the events $E, E'$ and the triangle inequality, we obtain that under $\pi$ and $\pi'$,

$$|S(P) - S(P')| \geq \frac{kd}{2}. \quad (36)$$

Now we consider the total variation of the distributions of the histogram under the prior $\pi$ and $\pi'$. The triangle inequality yields

$$\text{TV} (P_N|E; P_{N'|E'}) \leq \text{TV} (P_N|E; P_N) + \text{TV} (P_N, P_{N'}) + \text{TV} (P_{N'}, P_{N'|E'}) = \mathbb{P} [E^c] + \text{TV} (P_N, P_{N'}) + \mathbb{P} [E'^c]. \quad (37)$$

Note that $\epsilon = \frac{\lambda}{\sqrt{k}} \geq 4\sqrt{\frac{\text{var}[U/\text{var}[U']]}{k}}$. Chebyshev’s inequality and the union bound yield that

$$\mathbb{P} [E^c] \leq \mathbb{P} \left[ \left| \sum_i \frac{U_i}{k} - 1 \right| > \epsilon \right] + \mathbb{P} \left[ |S(P) - \mathbb{E}[S(P)]| > \frac{kd}{4} \right] \leq \frac{\text{var}[U]}{k\epsilon^2} + \frac{16}{k^2d^2} \leq \frac{1}{16} + \frac{16}{kd^2}. \quad (38)$$

By the same reasoning,

$$\mathbb{P} [E'^c] \leq \frac{1}{16} + \frac{16}{kd^2}. \quad (39)$$

From the fact that total variation of product distribution can be upper bounded by the summation of individual ones we obtain

$$\text{TV} (P_N, P_{N'}) = \text{TV} (\mathbb{E} [\text{Poi} (nU/k)] \otimes k, \mathbb{E} [\text{Poi} (nU'/k)] \otimes k) \leq k \text{TV}(\mathbb{E} [\text{Poi} (nU/k)], \mathbb{E} [\text{Poi} (nU'/k)]). \quad (40)$$

Plugging (38)–(40) into (37), the desired lower bound follows from (36)–(37) and Le Cam’s lemma [LC86].

### B  Proof of upper bounds

#### Proof of Proposition 3.

Fix a distribution $P$. First we consider the bias:

$$|\mathbb{E}(\hat{S}_{\text{seen}}(P) - S(P))| = |\sum_i \mathbb{P}(N_i \geq 1) - 1 \{p_i > 0\}| = \sum_i (1 - \mathbb{P}(N_i \geq 1)) 1 \{p_i \geq \frac{1}{k}\} = \sum_i \exp(-np_i) 1 \{p_i \geq \frac{1}{k}\} \leq k \exp(-n/k).$$

The variance satisfies

$$\text{var}\hat{S}_{\text{seen}}(P) = \sum_i \text{var} 1 \{N_i > 0\} = \sum_i \text{var} (1 \{N_i > 0\}) 1 \{p_i \geq \frac{1}{k}\}\]

$$= \sum_i \exp(-np_i)(1 - \exp(-np_i)) 1 \{p_i \geq \frac{1}{k}\} \leq \sum_i \exp(-np_i) 1 \{p_i \geq \frac{1}{k}\} \leq k \exp(-n/k).$$

The conclusion follows since $k \geq 1$.  

\[\square\]
Proof of Proposition 4. Define an event by $E \triangleq \cap_{i=1}^{k} \left\{ N_i \leq c_0 \log k \Rightarrow p_i \leq \frac{c_1 \log k}{n} \right\}$, where $c_1 > c_0$. Then union bound and Chernoff bound for Poisson distribution give us
\begin{equation}
P[E^c] \leq \frac{1}{k^{c_1 - c_0 \log \frac{c_1}{c_0} - 1}}.
\end{equation}

Since $0 \leq S(P) \leq k$, we have $|\hat{S} - S(P)| \leq |\hat{S} - S(P)|$ and $|\hat{S} - S(P)| \leq k$. Therefore,
\begin{align*}
E(\hat{S} - S(P))^2 &= E[(\hat{S} - S(P))^21_E] + E[(\hat{S} - S(P))^21_{E^c}] \\
&\leq E[(\hat{S} - S(P))^21_E] + k^2P[E^c].
\end{align*}

Define $\mathcal{E} \triangleq \sum_i (g_L(N_i) - 1)1_{\{N_i \leq L\}}1_{\{1/k \leq p_i \leq c_1 \log k/n\}}$. Note that $N_i > L \Rightarrow p_i > 0$ and $g_L(0) = 0$. The implication in event $E$ yields that $\langle \hat{S} - S(P) \rangle 1_E = \epsilon 1_E$. Combining (42) we obtain that
\begin{equation}
E(\hat{S} - S(P))^2 \leq E[\mathcal{E}^2] + k^2P[E^c].
\end{equation}

By (41) the second term is lower order than the goal in (17) as long as $n \leq k \log k$ and $c_1 \geq 2$ and $c_0$ sufficiently small.

Next we proceed to analyze the mean square error $E[\mathcal{E}^2]$ in (43). The maximal value of the polynomial $P_L$ on $[\frac{1}{k}, \frac{c_1 \log k}{n}]$ can be upper bounded by
\begin{align*}
|P_L(x)| &\leq \frac{1}{\cos L \arccos(-\frac{r+1}{r-1})} = \frac{2}{(\frac{\sqrt{r} - \sqrt{l}}{\sqrt{r} + \sqrt{l}})^L} \\
&\times \left(\frac{\sqrt{r} - \sqrt{l}}{\sqrt{r} + \sqrt{l}}\right)^L \leq \exp \left(-c' \sqrt{\frac{n \log k}{k}}\right),
\end{align*}

for some universal constant $c'$. Therefore by the same steps as (16) we obtain an upper bound of the bias:
\begin{equation}
|E[\mathcal{E}]| \leq \sum_i e^{-np_i}P_L(p_i)1_{\{\frac{1}{k} \leq p_i \leq \frac{c_1 \log k}{n}\}} \leq k \exp \left(-\frac{n}{k} - c' \sqrt{\frac{n \log k}{k}}\right).
\end{equation}

Now we turn to its variance.
\begin{equation}
\text{var} [\mathcal{E}] = \sum_i \text{var} [(g_L(N_i) - 1)1_{\{N_i \leq L\}}] 1_{\{\frac{1}{k} \leq p_i \leq \frac{c_1 \log k}{n}\}}
\leq \sum_i E [(g_L(N_i) - 1)21_{\{N_i \leq L\}}] 1_{\{\frac{1}{k} \leq p_i \leq \frac{c_1 \log k}{n}\}}
= \sum_i e^{-np_i} \left(\sum_{j=0}^{L} \frac{(a_{j}j!}{n^j}) 2 \frac{(np_i)^j}{j!}\right) 1_{\{\frac{1}{k} \leq p_i \leq \frac{c_1 \log k}{n}\}}.
\end{equation}

Recall the polynomial coefficients defined in (13). Also $r \geq 2l$ since $n \leq \frac{c_1}{2} k \log k$. Combining Corollary 1 in Appendix E on the coefficients on shifted Chebyshev polynomials in the numerator and the upper bound of the factor in (44), we can upper bound the coefficients $|\frac{a_{j}j!}{n^{j}}|$ by
\begin{equation}
|\frac{a_{j}j!}{n^{j}}| \leq \frac{j!}{(nr)^{j}} e^{6L} \exp \left(-c' \sqrt{\frac{n \log k}{k}}\right) = \frac{j!}{(c_1 \log k)^{j}} e^{6L} \exp \left(-c' \sqrt{\frac{n \log k}{k}}\right).
\end{equation}
Recall that $L \leq c_0 \log k$ and $c_0 < c_1$. Then

$$\text{var } [\mathcal{E}] \lesssim k^{1+12c_0} \exp \left( -\frac{n}{k} - 2c' \sqrt{\frac{n \log k}{k}} \right) \sum_{j=0}^{L} \frac{j!}{(c_1 \log k)^j}$$

$$\lesssim k^{1+12c_0} \exp \left( -\frac{n}{k} - 2c' \sqrt{\frac{n \log k}{k}} \right) \sum_{j=0}^{L} \sqrt{2\pi L} \left( \frac{c_0}{ec_1} \right)^j$$

$$\lesssim k^{1+12c_0 + c_0 \log \frac{c_0}{ec_1}} (\log k)^{1.5} \exp \left( -\frac{n}{k} - 2c' \sqrt{\frac{n \log k}{k}} \right),$$

which is lower order than the squared bias when $c_0$ is sufficiently small.

Now combining (43)–(50) and note that $n \lesssim k \log k$ then $n k \lesssim \sqrt{n \log k}$ we conclude that

$$\mathbb{E} [\mathcal{E}^2] \leq (\mathbb{E} [\mathcal{E}])^2 + \text{var } [\mathcal{E}] \lesssim k^2 \exp \left( -c \sqrt{\frac{n \log k}{k}} \right),$$

for some universal constant $c$. \hfill \Box

### C Proof of Theorem 2

**Proof.** First we consider the case $\epsilon \geq \frac{1}{k}$. Let $n = \frac{Ck}{\log k} \log^2 \frac{k}{\epsilon}$ for some constant $C$ then by Chebyshev inequality,

$$\mathbb{P} [||\hat{S} - S(P)|| > \epsilon k] \leq \frac{\mathbb{E} |\hat{S} - S(P)|^2}{k^2 \epsilon^2}.$$  

Then Proposition 4 implies that

$$\inf \sup \mathbb{P} [||\hat{S} - S(P)|| > \epsilon k] \leq \frac{R^*(k,n)}{k^2 \epsilon^2} \lesssim \epsilon^{\Delta-2},$$

for some constant $\Delta > 2$. Then $n^*(k,\epsilon) \leq \frac{Ck}{\log k} \log^2 \frac{k}{\epsilon}$.

Now let $n = \frac{ck}{\log k} \log^2 \frac{k}{\epsilon}$ for some constant $c$. By the testing procedure in the proof of Proposition 2, if $\epsilon \geq \frac{1}{k^2}$ for some positive constant $\delta < 1$, we obtain that

$$\inf \sup \mathbb{P} [||\hat{S} - S(P)|| > \epsilon k] \gtrsim 1.$$  

Therefore $n^*(k,\epsilon) \geq \frac{ck}{\log k} \log^2 \frac{k}{\epsilon}$ when $\epsilon \geq \frac{1}{k^2}$. If $\frac{1}{k} \leq \epsilon < \frac{1}{k^2}$ then the monotonicity of $n^*(k,\cdot)$ yields that

$$n^*(k,\epsilon) \geq n^*(k,\epsilon^{-\delta}) \gtrsim \frac{\delta^2 ck}{\log k} \log^2 \frac{k}{\epsilon}.$$  

The upper bound of $n^*(k,0)$ follows immediately from the coupon collector’s problem using plug-in estimator. Again by monotonicity $n^*(k,0) \geq n^*(k,k^{-1}) \gtrsim k \log k$. \hfill \Box
D Proofs for the species problem

Proof of Theorem 4. Entirely analogous to the analysis in (41) and (42) we have

\[ \mathbb{E}(\hat{S}_m - S(P))^2 \leq \mathbb{E}[\mathcal{E}^2] + \frac{k^2}{k^{c_1-c_0}\log \frac{cn}{c_0} - 1}. \]  

(52)

where \( \mathcal{E} \triangleq \sum_i (g_L(N_i) - 1)1_{\{N_i \leq L\}} 1_{\{1/k \leq p_i \leq c_1 \log k / n\}} \). The second term satisfies the goal in (27) when \( c_1 \geq 2 \) and \( c_0 \) is sufficiently small.

Let \( \rho > \frac{c_1 \log k}{n} \leq \frac{L}{k} \), then for \( p_i \leq \frac{c_1 \log k}{n} \) by design the estimator is unbiased \( \mathbb{E}[\mathcal{E}] = e^{-np_i q_L(kp_i)} = 0 \). Now we turn to its variance. Recall the definition of \( q_L \) and \( a_m \) in (26). Similar to (48) we have

\[ \text{var}[\mathcal{E}] \leq \sum_i e^{-np_i} \left( \sum_{j=0}^L \left( \frac{a_j k! j!}{n^j} \right)^2 \frac{(np_i)^j}{j!} \right) 1_{\left\{ \frac{k}{L} \leq p_i \leq \frac{c_1 \log k}{n} \right\}}. \]

For any \( x \in [0, L + 1] \) let \( t \triangleq \lfloor x \rfloor \), then \( t \leq x < t + 1 \) and

\[ \left| \prod_{j=1}^L (j - x) \right| = (t + 1 - x) \cdots ((t + L - t) - x) (x - t) \cdots (x - 1) \leq 1 \cdot 2 \cdots \cdot L. \]

Therefore \( |q_L(x)| \leq 2 \) for all \( x \in [1, L] \). By [Tim63, 2.9.11] and Corollary 1 for \( L \geq 2 \) we can upper bound the coefficients that \( |a_j| \leq \frac{2^6 L}{k!} \). Recall that \( \frac{c_1 \log k}{n} \leq \frac{L}{k} \), then \( \left| \frac{a_j k! j!}{n^j} \right| \leq \frac{2^6 L j!}{(c_1 \log k)^j} \) and

\[ \text{var}[\mathcal{E}] \leq 2k^{1 + 12c_0} e^{-n/k} \sum_{j=0}^L \left( \frac{j!}{(c_1 \log k)^j} \right)^2 \lesssim k^{1 + 12c_0 + c_0 \log \frac{2n}{c_1} (\log k) 1.5 e^{-n/k}, \]

which implies the desired results when \( c_0 \) is sufficiently small.

\[ \square \]

E Bounds on the coefficients of shifted Chebyshev polynomials

Lemma 3. Denote Chebyshev polynomial of degree \( L \) by \( T_L(x) = \cos L \arccos x \triangleq \sum_{i=0}^L a_i x^i \). Then \( a_i \leq e^{2L} \).

Proof. By [Tim63, 2.9.12], if \( L \) is even then \( a_{2l+1} = 0 \) and

\[ |a_{2l}| = 2^{2l} \frac{L}{L + 1} \left( \frac{L + 1}{2l} \right), \quad l = 0, \ldots, \frac{L}{2}, \]

therefore

\[ \frac{1}{2} \left( 1 + \frac{L}{2l} \right)^{2l} \leq |a_{2l}| \leq \left( 1 + \frac{L}{2l} \right)^{2l} e^{2l}, \quad l = 0, \ldots, \frac{L}{2}. \]

The RHS is increasing and at most \( (2e)^L \leq e^{2L} \).

If \( L \) is odd again by [Tim63, 2.9.12] \( a_{2l} = 0 \) and

\[ |a_{2l+1}| = 2^{2l} \frac{L}{2l + 1} \left( \frac{L - 1}{2l + 1} \right), \quad l = 0, \ldots, \frac{L - 1}{2}, \]

18
therefore
\[ \frac{L}{2l+1} \left( 1 + \frac{L-1}{2l} \right)^{2l} \leq |a_{2l+1}| \leq \frac{L}{2l+1} \left( 1 + \frac{L-1}{2l} \right)^{2l} e^{2l}, \quad l = 0, \ldots, \frac{L-1}{2}. \]

The RHS can be upper bounded by \( L(2e)^{L-1} \leq e^{2L} \).

In [Tim63, 2.9.11] an important quantity to bound the coefficients of polynomial on \([l, r]\) is the coefficients of shifted Chebyshev polynomial \( \cos L \arccos \frac{2x-l-r}{r-l} \). If \( l = -r \) and the interval is \([-r, r]\) still symmetric around the origin, then
\[
\cos L \arccos \frac{x}{r} = \sum_{i=0}^{L} \tilde{a}_i x^i = \sum_{i=0}^{L} \frac{a_i}{r^i} x^i.
\]

Then by Lemma 3 \( |\tilde{a}_i| = \tilde{a}_i \leq \frac{e^{2L}}{r^i} \). Note that there is a factor \( \frac{1}{r^i} \), which will appear in more general cases as shown in the next lemma.

First we expand a general composed polynomial. Let
\[
\sum_{i=0}^{L} a_i (mx + b)^i \triangleq \sum_{i=0}^{L} \tilde{a}_i x^i, \tag{53}
\]

then by binomial expansion we obtain
\[
\tilde{a}_i = \sum_{k=i}^{L} \binom{k}{i} m^i b^{k-i} \tag{54}.
\]

**Lemma 4.** Let \( \tilde{a}_i \) be the coefficients of shifted Chebyshev polynomial
\[
\cos L \arccos \frac{2x-l-r}{r-l} = \sum_{i=0}^{L} \tilde{a}_i x^i.
\]

where \( r > l \geq 0 \) and \( L \geq 1 \). Then
\[
|\tilde{a}_i| \leq \frac{1}{r^i} \left( 4e^{3r+l} \right)^L.
\]

**Proof.** Denote by \( a_i \) the coefficients of Chebyshev polynomial of degree \( L \):
\[
\cos L \arccos x = \sum_{i=0}^{L} a_i x^i.
\]

Using (54) and Lemma 3, we obtain that
\[
|\tilde{a}_i| = \sum_{k=i}^{L} a_k \binom{k}{i} \left( \frac{2}{r-l} \right)^i \left( \frac{r+l}{r-l} \right)^{k-i} \leq e^{2L} \left( \frac{2}{r+l} \right)^i \sum_{k=i}^{L} \binom{k}{i} \left( \frac{r+l}{r-l} \right)^k
\leq e^{2L} \left( \frac{2}{r} \right)^i \left( \frac{r+l}{r-l} \right)^L \sum_{k=i}^{L} \binom{k}{i} = e^{2L} \left( \frac{2}{r} \right)^i \left( \frac{r+l}{r-l} \right)^L \left( L+1 \right)
\leq \frac{1}{r^i} \left( 4e^{3r+l} \right)^L.
\]

\[\square\]
Corollary 1. Let \( \tilde{a}_i \) be the same as previous lemma. Assume that \( r > \kappa l \geq 0 \) for some constant \( \kappa > 1 \). Then
\[
|\tilde{a}_i| \leq \frac{1}{r^d} \left( 4e^3 \frac{\kappa+1}{\kappa-1} \right)^L.
\]

F Dual program of the linear programming (11)

Define the following infinite-dimensional linear programming problem:
\[
E^*_1 \triangleq \sup P \left[ U' = 0 \right] - P \left[ U = 0 \right]
\text{s.t. } E \left[ U \right] = E \left[ U' \right] = 1
E \left[ U^j \right] = E \left[ U'^j \right], \quad j = 1, \ldots, L + 1,
U, U' \in \{0\} \cup I,
\]
where \( I = [a, b] \) with \( b > a \geq 1 \). Then (11) is a special case when \( I = [1, \lambda] \).

Lemma 5. \( E^*_1 = E_L(1/x, I) \).

Proof. We first show that (11) coincides with the following optimization problem:
\[
E^*_2 \triangleq \sup E \left[ \frac{1}{X} \right] - E \left[ \frac{1}{X'} \right]
\text{s.t. } E \left[ X^j \right] = E \left[ X'^j \right], \quad j = 1, \ldots, L,
X, X' \in I.
\]
Given any feasible solution \( U, U' \) to (11), construct \( X, X' \) with the following distributions:
\[
P_X(dx) = xP_U(dx),
P_X'(dx) = xP_{U'}(dx),
\]
It is straightforward to verify that \( X, X' \) are feasible for (57) and
\[
E^*_2 \geq E \left[ \frac{1}{X} \right] - E \left[ \frac{1}{X'} \right] = P \left[ U' = 0 \right] - P \left[ U = 0 \right].
\]
Therefore \( E^*_2 \geq E^*_1 \).

On the other hand, given any feasible \( X, X' \) for (57), construct \( U, U' \) with the distributions:
\[
P_U(du) = \left( 1 - E \left[ \frac{1}{X} \right] \right) \delta_0(du) + \frac{1}{u} P_X(du),
P_U'(du) = \left( 1 - E \left[ \frac{1}{X'} \right] \right) \delta_0(du) + \frac{1}{u} P_X'(du),
\]
which are well-defined since \( X, X' \geq 1 \) and hence \( E \left[ \frac{1}{X} \right] \leq 1, E \left[ \frac{1}{X'} \right] \leq 1 \). Then \( U, U' \) are feasible for (11) and hence
\[
E^*_1 \geq P \left[ U' = 0 \right] - P \left[ U = 0 \right] = E \left[ \frac{1}{X} \right] - E \left[ \frac{1}{X'} \right].
\]
Therefore \( E^*_1 \geq E^*_2 \). Finally, the dual of (57) is precisely the best polynomial approximation problem (see, e.g., [WY14, Appendix E]) and hence
\[
E^*_1 = E^*_2 = E_L(1/x, I) = \inf_{p \in \mathcal{P}_L} \sup_{x \in I} \left| \frac{1}{x} - p(x) \right|.
\]
References

[BF93] John Bunge and M Fitzpatrick. Estimating the number of species: a review. *Journal of the American Statistical Association*, 88(421):364–373, 1993.

[BO79] Kenneth P Burnham and W Scott Overton. Robust estimation of population size when capture probabilities vary among animals. *Ecology*, 60(5):927–936, 1979.

[BYJK+02] Ziv Bar-Yossef, TS Jayram, Ravi Kumar, D Sivakumar, and Luca Trevisan. Counting distinct elements in a data stream. In *Proceedings of the 6th Randomization and Approximation Techniques in Computer Science*, pages 1–10. 2002.

[BYKS01] Ziv Bar-Yossef, Ravi Kumar, and D Sivakumar. Sampling algorithms: lower bounds and applications. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 266–275. ACM, 2001.

[CCMN00] Moses Charikar, Surajit Chaudhuri, Rajeev Motwani, and Vivek Narasayya. Towards estimation error guarantees for distinct values. In *Proceedings of the nineteenth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 268–279. ACM, 2000.

[Cha84] Anne Chao. Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of statistics*, pages 265–270, 1984.

[CL92] 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.

[CL11] T.T. Cai and M. G. Low. Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional. *The Annals of Statistics*, 39(2):1012–1041, 2011.

[DR80] JN Darroch and D Ratcliff. A note on capture-recapture estimation. *Biometrics*, pages 149–153, 1980.

[DS08] Vladislav K Dzyadyk and Igor A Shevchuk. *Theory of uniform approximation of functions by polynomials*. Walter de Gruyter, 2008.

[ET76] Bradley Efron and Ronald Thisted. Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, 63(3):435–447, 1976.

[FCW43] Ronald Aylmer 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.

[Goo53] Irving J Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264, 1953.

[GS04] Alberto Gandolfi and C.C.A. Sastri. Nonparametric estimations about species not observed in a random sample. *Milan Journal of Mathematics*, 72(1):81–105, 2004.

[GT56] I.J. Good and G.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.
[Har68] Bernard Harris. Statistical inference in the classical occupancy problem unbiased estimation of the number of classes. *Journal of the American Statistical Association*, pages 837–847, 1968.

[HW01] Shu-Pang Huang and BS Weir. Estimating the total number of alleles using a sample coverage method. *Genetics*, 159(3):1365–1373, 2001.

[INK87] I.A. Ibragimov, A.S. Nemirovskii, and R.Z. Khas’minskii. Some problems on nonparametric estimation in gaussian white noise. *Theory of Probability & Its Applications*, 31(3):391–406, 1987.

[JVW14] Jiantao Jiao, Kartik Venkat, and Tsachy Weissman. Order-optimal estimation of functionals of discrete distributions. *arXiv preprint arXiv:1406.6956*, 2014.

[LC86] L. Le Cam. Asymptotic methods in statistical decision theory. Springer-Verlag, New York, NY, 1986.

[LNS99] Oleg Lepski, Arkady Nemirovski, and Vladimir Spokoiny. On estimation of the $L_r$ norm of a regression function. *Probability theory and related fields*, 113(2):221–253, 1999.

[Lo92] Shaw-Hwa Lo. From the species problem to a general coverage problem via a new interpretation. *The Annals of Statistics*, 20(2):1094–1109, 1992.

[LP56] Richard C Lewontin and Timothy Prout. Estimation of the number of different classes in a population. *Biometrics*, 12(2):211–223, 1956.

[MR10] Rajeev Motwani and Prabhakar Raghavan. *Randomized algorithms*. Chapman & Hall/CRC, 2010.

[MSJ82] JP Marchand and FE Schroeck Jr. On the estimation of the number of equally likely classes in a population. *Communications in Statistics-Theory and Methods*, 11(10):1139–1146, 1982.

[MU05] Michael Mitzenmacher and Eli Upfal. *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge University Press, 2005.

[RRSS09] Sofya Raskhodnikova, Dana Ron, Amir Shpilka, and Adam Smith. Strong lower bounds for approximating distribution support size and the distinct elements problem. *SIAM Journal on Computing*, 39(3):813–842, 2009.

[Sam68] Ester Samuel. Sequential maximum likelihood estimation of the size of a population. *The Annals of Mathematical Statistics*, 39(3):1057–1068, 1968.

[Tim63] Aleksandr Filippovich Timan. *Theory of approximation of functions of a real variable*. Pergamon Press, 1963.

[Tsy09] A.B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Verlag, New York, NY, 2009.

[Val12] Gregory Valiant. *Algorithmic Approaches to Statistical Questions*. PhD thesis, EECS Department, University of California, Berkeley, Sep 2012.
[VV10] Gregory Valiant and Paul Valiant. A CLT and tight lower bounds for estimating entropy. In Electronic Colloquium on Computational Complexity (ECCC), volume 17, page 179, 2010.

[VV11] Gregory Valiant and Paul Valiant. Estimating the unseen: an \(n/\log(n)\)-sample estimator for entropy and support size, shown optimal via new CLTs. In Proceedings of the 43rd annual ACM symposium on Theory of computing, pages 685–694, 2011.

[VV13] 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.

[WY14] Yihong Wu and Pengkun Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. arXiv preprint arXiv:1407.0381, 2014.