The Annals of Statistics
2019, Vol. 47, No. 2, 857–883
https://doi.org/10.1214/17-AOS1665

CHEBYSHEV POLYNOMIALS, MOMENT MATCHING, AND OPTIMAL ESTIMATION OF THE UNSEEN

BY YIHONG WU AND PENGKUN YANG

Yale University and University of Illinois at Urbana-Champaign

We consider the problem of estimating the support size of a discrete distribution whose minimum nonzero mass is at least $\frac{1}{k}$. Under the independent sampling model, we show that the sample complexity, that is, the minimal sample size to achieve an additive error of εk with probability at least 0.1 is within universal constant factors of $\frac{k}{\log k}\log^2\frac{1}{\varepsilon}$, which improves the state-of-the-art result of $\frac{k}{\varepsilon^2\log k}$ in [In *Advances in Neural Information Processing Systems* (2013) 2157–2165]. Similar characterization of the minimax risk is also obtained. Our procedure is a linear estimator based on the Chebyshev polynomial and its approximation-theoretic properties, which can be evaluated in $O(n+\log^2 k)$ time and attains the sample complexity within constant factors. The superiority of the proposed estimator in terms of accuracy, computational efficiency and scalability is demonstrated in a variety of synthetic and real datasets.

1. Introduction.

1.1. *Model.* Estimating the support size of a distribution from data is a classical problem in statistics with widespread applications. For example, a major task for ecologists is to estimate the number of species [11] from field experiments; linguists are interested in estimating the vocabulary size of Shakespeare based on his complete works [10, 26, 34]; in population genetics it is of great interest to estimate the number of different alleles in a population [18]. Estimating the support size is equivalent to estimating the number of unseen symbols, 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.

We adopt the following statistical model [3, 30]. Let P be a discrete distribution over some countable alphabet. Without loss of generality, we assume the alphabet is $\mathbb{N}$ and denote $P=(p_1, p_2, \dots)$. Given n i.i.d. samples $X \triangleq (X_1, \dots, X_n)$ drawn from P, the goal is to estimate the support size

$$S = S(P) \triangleq \sum_i \mathbf{1}_{\{p_i>0\}}. \tag{1}$$

Received June 2016; revised November 2017.
MSC2010 subject classifications. Primary 62G05; secondary 62C20.
Key words and phrases. Support size estimation, large domain, polynomial approximation, high-dimensional statistics, nonparametric inference.

To estimate the distribution or its functionals, a sufficient statistic is the *histogram* of the samples, denoted by $N = (N_1, N_2, \dots)$ and

$$N_i = \sum_{j=1}^{n} \mathbf{1}_{\{X_j = i\}}. \tag{2}$$

Therefore, N has a multinomial distribution with parameter n and P. 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, which are defined as

$$\Phi_j = \sum_{i} \mathbf{1}_{\{N_i = j\}}, \tag{3}$$

that is, the number of symbols that appear exactly j times.

It is clear that unless we impose further assumptions on the distribution P, it is impossible to estimate $S(P)$ within a given accuracy, for otherwise there can be arbitrarily many masses in the support of P that, with high probability, are never sampled and the worst-case risk for estimating $S(P)$ is obviously infinite. To prevent the triviality, a conventional assumption [30] is to impose a lower bound on the nonzero probabilities. Therefore, we restrict our attention to the parameter space $\mathcal{D}_k$, which consists of all probability distributions on $\mathbb{N}$ whose minimum nonzero mass is at least $\frac{1}{k}$; consequently, $S(P) \le k$ for any $P \in \mathcal{D}_k$. The decision-theoretic fundamental limit of this problem is given by the *minimax risk*:

$$R^*(k, n) \triangleq \inf_{\hat{S}} \sup_{P \in \mathcal{D}_k} \mathbb{E}(\hat{S} - S)^2, \tag{4}$$

where the loss function is the mean squared error (MSE) and $\hat{S}$ is an integer-valued estimator measurable with respect to the samples $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} P$.

1.2. *Main results*. Our first main result is the following characterization of the minimax risk.

THEOREM 1. *For all* $k, n \ge 2$,

$$R^*(k, n) = k^2 \exp\left(-\Theta\left(\sqrt{\frac{n}{k}\log\frac{k}{n}} \vee \frac{n}{k} \vee 1\right)\right). \tag{5}$$

Furthermore, if $\frac{k}{\log k} \ll n \ll k \log k$, *as* $k \to \infty$,

$$\begin{aligned} k^2 \exp\left(-(\sqrt{2}e + o(1))\sqrt{\frac{n}{k}\log\frac{k}{n}}\right) &\le R^*(k, n) \\ &\le k^2 \exp\left(-(1.579 + o(1))\sqrt{\frac{n}{k}\log\frac{k}{n}}\right). \end{aligned} \tag{6}$$

To interpret the rate of convergence in (5), we consider three cases:

Simple regime $n \gtrsim k \log k$: we have $R^*(k,n) = k^2 \exp(-\Theta(\frac{n}{k}))$ which can be achieved by the simple plug-in estimator

$$\hat{S}_{\text{seen}} \triangleq \sum_i \mathbf{1}_{\{N_i > 0\}}, \tag{7}$$

that is, the number of observed symbols or the support size of the empirical distribution. Furthermore, if $\frac{n}{k \log k}$ exceeds a sufficiently large constant, all symbols are present in the data and $\hat{S}_{\text{seen}}$ is in fact exact with high probability, namely, $\mathbb{P}[\hat{S}_{\text{seen}} \neq S] \leq \mathbb{E}(\hat{S}_{\text{seen}} - S)^2 \to 0$. This can be understood as the classical coupon collector's problem (cf., e.g., [27]).

Nontrivial regime $\frac{k}{\log k} \ll n \ll k \log k$: In this case, the samples are relatively scarce and the naive plug-in estimator grossly underestimate the true support size as many symbols are simply not observed. Nevertheless, accurate estimation is still possible and the optimal risk is given by $R^*(k,n) = k^2 \exp(-\Theta(\sqrt{\frac{n \log k}{k}}))$. This can be achieved by a linear estimator based on the Chebyshev polynomial and its approximation-theoretic properties. Although more sophisticated than the plug-in estimator, this procedure can be evaluated in $O(n + \log^2 k)$ time.

Impossible regime $n \lesssim \frac{k}{\log k}$: any estimator suffers an error proportional to k in the worst case.

Next we discuss the *sample complexity* of estimating the support size, which is defined as follows:

$$n^*(k,\varepsilon) \triangleq \min\{n \geq 0 \colon \exists \hat{S},\ \text{s.t. } \mathbb{P}[|\hat{S} - S(P)| \geq \varepsilon k] \leq 0.1, \forall P \in \mathcal{D}_k\}, \tag{8}$$

where $\hat{S}$ is an integer-valued estimator measurable with respect to the samples $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P$. Clearly, since $\hat{S} - S$ is an integer, the only interesting case is $\varepsilon \geq \frac{1}{k}$, with $\varepsilon = \frac{1}{k}$ corresponding to the exact estimation of the support size since $|\hat{S} - S| < 1$ is equivalent to $\hat{S} = S$. Furthermore, since $S(P)$ takes values in $[k]$, $n^*(k, \frac{1}{2}) = 0$ by definition. The next result characterizes the sample complexity within universal constant factors that are within a factor of six asymptotically.

THEOREM 2. *Fix a constant $c_0 < \frac{1}{2}$. For all $\frac{1}{k} \leq \varepsilon \leq c_0$,*

$$n^*(k,\varepsilon) \asymp \frac{k}{\log k} \log^2 \frac{1}{\varepsilon}. \tag{9}$$

Furthermore, if $\varepsilon \to 0$ and $\varepsilon = k^{-o(1)}$, as $k \to \infty$,

$$\frac{1+o(1)}{2e^2} \frac{k}{\log k} \log^2 \frac{1}{\varepsilon} \leq n^*(k,\varepsilon) \leq \frac{1+o(1)}{2.494} \frac{k}{\log k} \log^2 \frac{1}{\varepsilon}. \tag{10}$$

Compared to Theorem 1, the only difference is that here we are dealing with the zero-one loss $\mathbf{1}_{\{|S-\hat{S}|\geq\varepsilon k\}}$ 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 Theorems 1 and 2 simultaneously. Furthermore, the choice of 0.1 as the probability of error in the definition of the sample complexity is entirely arbitrary; replacing it by $1-\delta$ for any constant $\delta\in(0,1)$ only affect $n^*(k,\varepsilon)$ up to constant factors.[1]

1.3. *Previous work.* There is a vast amount of literature devoted to the support size estimation problem. In parametric settings, the data generating distribution is assumed to belong to certain parametric family such as uniform or Zipf [8, 22, 26] and traditional estimators, such as maximum likelihood estimator and minimum variance unbiased estimator, are frequently used [10, 17, 18, 22, 25, 32]; see the extensive surveys [2, 12]. When difficult to postulate or justify a suitable parametric assumption, various nonparametric approaches are adopted such as the Good–Turing estimator [14, 31] and variants due to Chao and Lee [5, 6], Jackknife estimator [3], empirical Bayes approach (e.g., Good–Toulmin estimator [15]) and one-sided estimator [24]. Despite their practical popularity, little is known about the performance guarantee of these estimators, let alone their optimality. Next we discuss provable results assuming the independent sampling model in Section 1.1.

For the naive plug-in estimator (7), it is easy to show (see Proposition 2) that to estimate $S(P)$ within $\pm\varepsilon k$ the minimal required number of samples is $\Theta(k\log\frac{1}{\varepsilon})$, which scales logarithmically in $\frac{1}{\varepsilon}$ but linearly in k, the same scaling for estimating the distribution P itself. Recently, Valiant and Valiant [39] showed that the sample complexity is in fact sublinear in k; however, the performance guarantee of the proposed estimators are still far from being optimal. Specifically, an estimator based on a linear program (LP) that is a modification of [10], Program 2, is proposed and shown to achieve $n^*(k,\varepsilon)\lesssim\frac{k}{\varepsilon^{2+\delta}\log k}$ for any arbitrary $\delta>0$ [39], Corollary 11, which has subsequently been improved to $\frac{k}{\varepsilon^2\log k}$ in [41], Theorem 2, Fact 9. The lower bound $n^*(k,\varepsilon)\gtrsim\frac{k}{\log k}$ in [38], Corollary 9, is optimal in k but provides no dependence on ε. These results show that the optimal scaling in terms of k is $\frac{k}{\log k}$ but the dependence on the accuracy ε is $\frac{1}{\varepsilon^2}$, which is even worse than the plug-in estimator. From Theorem 2, we see that the dependence on ε can be improved from polynomial to polylogarithmic $\log^2\frac{1}{\varepsilon}$, which turns out to be optimal. Furthermore, this can be attained by a linear estimator which is far more scalable than linear programming on massive datasets (see the experiment on New

[1]Specifically, upgrading the confidence to $1-\delta$ can be achieved by oversampling by merely a factor of $\log\frac{1}{\delta}$: Let $T=\log\frac{1}{\delta}$. With nT samples, divide them into T batches, apply the n-sample estimator to each batch and aggregate by taking the median. Then Hoeffding's inequality implies the desired confidence.

York Times datasets of one billion words in Section 4). Finally, we mention that a general framework of designing and analyzing linear estimators is given in [40] based on linear programming (as opposed to the approximation-theoretic approach in the current paper).

To close this subsection, we mention two closely related problems whose recent resolution relies on results in the current paper:

- *Species extrapolation*: Given n independent samples drawn from an unknown distribution, the goal is to predict the number of hitherto unseen symbols that would be observed if m additional samples were collected from the same distribution. Originally formulated in [11] and further studied in [5, 10, 15], this problem reduces to support size estimation if $m = \infty$; in contrast, for finite m, this problem remains nontrivial even if no lower bound on the minimum nonzero probability is imposed on the underlying distribution, since very rare species will typically not appear in the new samples. The recent result [28] showed that the furthest range for accurate extrapolation is $m = o(n \log n)$ and obtained the minimax estimation error as a function of m, n for all distributions, where the lower bound is obtained via a reduction to support size estimation studied in this paper.
- *Distinct elements*: In this problem, the goal is to estimate the number of distinct colors based on repeated draws from in an urn consisting of k colored balls. For sampling with replacement, this can be viewed as a restricted case of the model in the present paper, where the distribution $P = (p_i)$ has the special form of $p_i = \frac{k_i}{k}$, with $k_i \in \mathbb{Z}_+$ corresponding to the number of balls of the ith color and $\sum_i k_i = k$. The sample complexity under multiplicative error, that is, estimating $S(P)$ within a factor of α (≥ 2) has been shown to be $\Theta(\frac{k}{\alpha^2})$ in [7]. For additive error, that is, estimating $S(P)$ within $\pm\varepsilon k$, a lower bound has been established in [30], which for constant ε, scales as $k^{1-O(\sqrt{\frac{\log\log k}{\log k}})}$. This, in turn, implies a lower bound for $n^*(k, \varepsilon)$, which is slightly suboptimal compared to the tight bound $\frac{k}{\log k} = k^{1-\frac{\log\log k}{\log k}}$. The sample complexity of the distinct elements problem has been recently shown in [42] to be $\Theta(\frac{k}{\log k} \log \frac{1}{\varepsilon})$ if the desired accuracy satisfies $\varepsilon > k^{-0.5+\delta}$. Compared with that of support size estimation in Theorem 2, we see that the discrete structure of the distribution strictly reduces the sample complexity of the problem.

1.4. *Organization*. The paper is organized as follows: in Section 2, we outline the proof for the lower bound part of Theorems 1 and 2 and the construction of the least favorable priors. In Section 3, we construct an estimator based on Chebyshev polynomials which achieves the minimax risk and the sample complexity within constant factors. In Section 4, we apply our estimators to both synthetic and real data and compare the performance with existing methodologies. Proofs

of the lower and upper bounds are given in Section 5, respectively. Due to space constraints, proofs of technical lemmas are provided in [44], Appendix B in the Supplementary Material.

1.5. *Notation.* For $k \in \mathbb{N}$, let $[k] \triangleq \{1, \ldots, k\}$. The n-fold product of a distribution P is denoted by $P^{\otimes n}$. Let $\mathrm{Poi}(\lambda)$ denote the Poisson distribution with mean λ whose probability mass function is denoted by $\mathrm{poi}(\lambda, j) \triangleq \frac{\lambda^j e^{-\lambda}}{j!}$, $j \geq 0$. Given a positive random variable U, denote the Poisson mixture with respect to the distribution of U by $\mathbb{E}[\mathrm{Poi}(U)]$, whose probability mass function is given by $\frac{1}{j!}\mathbb{E}[U^j e^{-U}]$, $j \geq 0$. Let $\mathrm{Bern}(p) = p\delta_1 + (1-p)\delta_0$ denote the Bernoulli distribution. The total variation and the Kullback–Leibler divergence between probability measures P and Q are denoted by $\mathsf{TV}(P, Q) \triangleq \frac{1}{2}\int |\mathrm{d}P - \mathrm{d}Q|$ and $D(P\|Q) \triangleq \int \mathrm{d}P \log \frac{\mathrm{d}P}{\mathrm{d}Q}$, respectively. We use standard big-O notation, for example, for any positive sequences $\{a_n\}$ and $\{b_n\}$, $a_n = O(b_n)$ or $a_n \lesssim b_n$ if $a_n \leq Cb_n$ for some absolute constant $C > 0$, $a_n = o(b_n)$ or $a_n \ll b_n$ or if $\lim a_n/b_n = 0$. In order to extract non-asymptotic statements from asymptotic ones, we pay extra attention to $o(1)$ terms. Specifically, we write $o_\delta(1)$ as $\delta \to 0$ to indicate convergence that is uniform in all other parameters.

2. Minimax lower bound. The lower bound argument follows the idea in [4, 21, 43] and relies on the generalized Le Cam's method involving two composite hypotheses, also known as the method of fuzzy hypotheses [36]. Specifically, suppose the following (composite) hypothesis testing problem:

$$H_0 : S(P) \leq s,\ P \in \mathcal{D}_k \qquad \text{versus } H_1 : S(P) \geq s + \delta,\ P \in \mathcal{D}_k$$

cannot be tested with vanishing probability of error on the basis of n samples, then the sample complexity of estimating $S(P)$ within δ with high probability must exceed n. In particular, the impossibility to test the above composite hypotheses is shown by constructing two priors (i.e., two random probability vectors) so that the induced distribution of the samples are close in total variation. Next we elaborate the main ingredients of Le Cam's method: (a) construction of the two priors; (b) separation between functional values; (c) bound on the total variation.

Let $\lambda > 1$. Given unit-mean random variables U and U' that take values in $\{0\} \cup [1, \lambda]$, define the following random vectors:

$$\mathsf{P} = \frac{1}{k}(U_1, \ldots, U_k), \qquad \mathsf{P}' = \frac{1}{k}(U_1', \ldots, U_k'), \tag{11}$$

where U_i and U_i' are i.i.d. copies of U and U', respectively. Although P and P' need not be probability distributions, as long as the standard deviations of U and U' are not too big, the law of large numbers ensures that with high probability P and P' lie in a small neighborhood near the probability simplex, which we refer as the set of *approximate* probability distributions. Furthermore, the minimum nonzeros in

P and P' are at least $\frac{1}{k}$. It can be shown that the minimax risk over approximate probability distributions is close to that over the original parameter space $\mathcal{D}_k$ of probability distributions. This allows us to use P and P' as priors and apply Le Cam's method. Note that both $S(\mathsf{P})$ and $S(\mathsf{P}')$ are binomially distributed, which with high probability, differ by the difference in their mean values:

$$\mathbb{E}\big[S(\mathsf{P})\big] - \mathbb{E}\big[S(\mathsf{P}')\big] = k\big(\mathbb{P}[U > 0] - \mathbb{P}\big[U' > 0\big]\big) = k\big(\mathbb{P}\big[U' = 0\big] - \mathbb{P}[U = 0]\big).$$

If we can establish the impossibility of testing whether data are generated from P or P', the resulting lower bound is proportional to $k(\mathbb{P}[U' = 0] - \mathbb{P}[U = 0])$.

To simplify the argument, we apply the Poissonization technique where the sample size is a $\mathrm{Poi}(n)$ random variable instead of a fixed number n. This provably does not change the statistical nature of the problem due to the concentration of $\mathrm{Poi}(n)$ around its mean n. Under Poisson sampling, the histograms (2) still constitute a sufficient statistic, which are distributed as $N_i \overset{\text{ind}}{\sim} \mathrm{Poi}(np_i)$, as opposed to multinomial distribution in the fixed-sample-size model. Therefore, through the i.i.d. construction in (11), $N_i \overset{\text{i.i.d.}}{\sim} \mathbb{E}[\mathrm{Poi}(\frac{n}{k}U)]$ or $\mathbb{E}[\mathrm{Poi}(\frac{n}{k}U')]$. Then Le Cam's lemma is applicable if $\mathsf{TV}(\mathbb{E}[\mathrm{Poi}(\frac{n}{k}U)]^{\otimes k}, \mathbb{E}[\mathrm{Poi}(\frac{n}{k}U')]^{\otimes k})$ is strictly bounded away from one, for which it suffices to show

$$\mathsf{TV}\big(\mathbb{E}\big[\mathrm{Poi}(nU/k)\big], \mathbb{E}\big[\mathrm{Poi}\big(nU'/k\big)\big]\big) \le \frac{c}{k}, \tag{12}$$

for some constant $c < 1$.

The above construction provides a recipe for the lower bound. To optimize the ingredients, it boils down to the following optimization problem (over one-dimensional probability distributions): Construct two priors U, U' with unit mean that maximize the difference $\mathbb{P}[U' = 0] - \mathbb{P}[U = 0]$ subject to the total variation distance constraint (12), which in turn, can be guaranteed by *moment matching*, that is, ensuring U and U' have identical first L moments for some large L, and the L_∞-norms U, U' are not too large. To summarize, our lower bound entails solving the following optimization problem:

$$\begin{aligned} &\sup \mathbb{P}\big[U' = 0\big] - \mathbb{P}[U = 0] \\ &\text{s.t. } \mathbb{E}[U] = \mathbb{E}\big[U'\big] = 1 \\ &\qquad \mathbb{E}\big[U^j\big] = \mathbb{E}\big[U'^j\big], \qquad j = 1, \ldots, L \\ &\qquad U, U' \in \{0\} \cup [1, \lambda]. \end{aligned} \tag{13}$$

The final lower bound is obtained from (13) by choosing $L \asymp \log k$ and $\lambda \asymp \frac{k \log k}{n}$.

In order to evaluate the infinite-dimensional linear programming problem (13), we consider its dual program. It is well known that the problem of best polynomial and moment matching are dual to each other; however, unlike the standard moment matching problem which imposes the equality of moments, the extra constraint in

(13) is that the values of the first moment must equal to one. Therefore, its dual is no longer the best polynomial approximation problem. Nevertheless, for the specific problem (13) which deals with the indicator function $x \mapsto \mathbf{1}_{\{x=0\}}$, via a change of variable we show in the Appendix that (13) coincides exactly with the best uniform approximation error of the function $x \mapsto \frac{1}{x}$ over the interval $[1, \lambda]$ by degree-$(L-1)$ polynomials:

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

where $\mathcal{P}_{L-1}$ denotes the set of polynomials of degree at most $L-1$. This best polynomial approximation problem has been well studied, cf. [9, 35]; 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 [35], Section 2.11.1.

Applying the procedure described above, we obtain the following sample complexity lower bound.

PROPOSITION 1. *Let* $\delta \triangleq \frac{\log \frac{1}{\varepsilon}}{\log k}$ *and* $\tau \triangleq \frac{\sqrt{\log k}/k^{1/4}}{1-2\varepsilon}$. *As* $k \to \infty$, $\delta \to 0$ *and* $\tau \to 0$,

$$n^*(k, \varepsilon) \geq \big(1 - o_\delta(1) - o_k(1) - o_\tau(1)\big) \frac{k}{2e^2 \log k} \log^2 \frac{1}{2\varepsilon}. \tag{14}$$

Consequently, if $\frac{1}{k^c} \leq \varepsilon \leq \frac{1}{2} - c' \frac{\sqrt{\log k}}{k^{1/4}}$ *for some constants* c, c', *then* $n^*(k, \varepsilon) \gtrsim \frac{k}{\log k} \log^2 \frac{1}{2\varepsilon}$.

The lower bounds announced in Theorems 1 and 2 follow from Proposition 1 combined with a simple two-point argument; see Section 5.2.

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. Following the same idea as in the lower bound part, we shall apply the Poissonization technique to simplify the analysis where the sample size is $\mathrm{Poi}(n)$ instead of a fixed number n, and hence the sufficient statistics $N = (N_1, \ldots, N_k) \overset{\text{ind}}{\sim} \mathrm{Poi}(np_i)$. Analogous to (4), the minimax risk under the Poisson sampling is defined by

$$\tilde{R}^*(k, n) \triangleq \inf_{\hat{S}} \sup_{P \in \mathcal{D}_k} \mathbb{E}(\hat{S} - S)^2. \tag{15}$$

Due to the concentration of $\mathrm{Poi}(n)$ near its mean n, the minimax risk with fixed sample size is close to that under the Poisson sampling, as shown in the following lemma, which allows us to focus on the model using Poissonized sample size.

LEMMA 1. *For any* $\beta < 1$,

$$R^*(k,n) \le \frac{\tilde{R}^*(k,(1-\beta)n)}{1-\exp(-n\beta^2/2)}.$$

In the next proposition, we first analyze the risk of the plug-in estimator $\hat{S}_{\text{seen}}$, which yields the optimal upper bound of Theorem 1 in the regime of $n \gtrsim k\log k$. This is consistent with the coupon collection intuition explained in Section 1.2.

PROPOSITION 2. *For all* $n, k \ge 1$,

$$\sup_{P\in\mathcal{D}_k} \mathbb{E}\big(S(P) - \hat{S}_{\text{seen}}(N)\big)^2 \le k^2 e^{-2n/k} + k e^{-n/k}, \tag{16}$$

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

Conversely, for P *that is uniform over* $[k]$, *for any fixed* $\delta \in (0,1)$, *if* $n \le (1-\delta)k\log\frac{1}{\varepsilon}$, *then as* $k\to\infty$,

$$\mathbb{P}\big[\big|S(P) - \hat{S}_{\text{seen}}(N)\big| \le \varepsilon k\big] \le e^{-\Omega(k^\delta)}. \tag{17}$$

In order to remedy the inaccuracy of the plug-in estimate $\hat{S}_{\text{seen}}$ in the regime of $n \lesssim k\log k$, our proposed estimator adds a linear correction term:

$$\hat{S} = \hat{S}_{\text{seen}} + \sum_{j\ge 1} u_j \Phi_j, \tag{18}$$

where the coefficients u_j's are to be specified. Equivalently, the estimator can be expressed in terms of the histogram as

$$\hat{S} = \sum_i g(N_i), \tag{19}$$

where $g : \mathbb{Z}_+ \to \mathbb{R}$ is defined as $g(j) = u_j + 1$ for $j \ge 1$ and $g(0) = 0$. Then the bias of $\hat{S}$ is

$$\begin{aligned}
\mathbb{E}[\hat{S} - S] &= \sum_{i:p_i>0} \mathbb{E}\big[g(N_i) - 1\big] \\
&= \sum_{i:p_i>0} e^{-np_i}\Big(\sum_{j\ge 1} u_j \frac{(np_i)^j}{j!} - 1\Big) \qquad (20)\\
&\triangleq \sum_{i:p_i>0} e^{-np_i} P(p_i),
\end{aligned}$$

where $P(0) = -1$ by design. Therefore, the bias of $\hat{S}$ is at most $S\max_{x\in[p_{\min},1]}|e^{-nx}P(x)|$, and the variance can be upper bounded by $2S\|g\|_\infty^2$ using the Efron–Stein inequality [33]. Next we choose the coefficients in order to

balance the bias and variance. The construction is done using Chebyshev polynomials, which we first introduce.

Recall that the usual Chebyshev polynomial of degree L is

$$T_L(x) = \cos(L \arccos x) = \big(z^L + z^{-L}\big)/2, \tag{21}$$

where z is the solution of the quadratic equation $z + z^{-1} = 2x$. Note that T_L is bounded in magnitude by one over the interval $[-1, 1]$. The shifted and scaled Chebyshev polynomial over the interval $[l, r]$ is given by

$$P_L(x) = -\frac{T_L(\frac{2x-r-l}{r-l})}{T_L(\frac{-r-l}{r-l})} \triangleq \sum_{m=1}^{L} a_m x^m - 1, \tag{22}$$

where the coefficients $a_1, \ldots, a_L$ can be obtained from those of the Chebyshev polynomial [35], 2.9.12, and the binomial expansion, or more directly,

$$a_j = \frac{P_L^{(j)}(0)}{j!} = -\left(\frac{2}{r-l}\right)^j \frac{1}{j!} \frac{T_L^{(j)}(-\frac{r+l}{r-l})}{T_L(-\frac{r+l}{r-l})}. \tag{23}$$

We now let

$$L \triangleq \lfloor c_0 \log k \rfloor, \qquad r \triangleq \frac{c_1 \log k}{n}, \qquad l \triangleq \frac{1}{k}, \tag{24}$$

where $c_0 < c_1$ are constants to be specified, and choose the coefficients of the estimator as

$$u_j = \begin{cases} \dfrac{a_j j!}{n^j} & j = 1, \ldots, L, \\ 0 & \text{otherwise}, \end{cases} \tag{25}$$

The estimator $\hat{S}$ is defined according to (18).

We proceed to explain the reasoning behind the choice (25) and the role of the Chebyshev polynomial. The main intuition is that since $c_0 < c_1$, then with high probability, whenever $N_i \le L = \lfloor c_0 \log k \rfloor$ the corresponding mass must satisfy $p_i \le \frac{c_1 \log k}{n}$. That is, if $p_i > 0$ and $N_i \le L$ then $p_i \in [l, r]$ with high probability, and hence $P_L(p_i)$ is bounded by the sup-norm of P_L over the interval $[l, r]$, which controls the bias in view of (20). In view of the extremal property of Chebyshev polynomials [35], Example 2.13.14, (22) is the unique degree-L polynomial that passes through the point $(0, -1)$ and deviates the least from zero over the interval $[l, r]$. This explains the coefficients (19) which are chosen to minimize the bias. The degree of the polynomial is only logarithmic so that the variance is small.

The next proposition gives an upper bound of the quadratic risk of our estimator (19).

PROPOSITION 3. *Assume the Poissonized sampling model where the histograms are distributed as* $N = (N_1, N_2, \ldots) \overset{\text{ind}}{\sim} \mathrm{Poi}(np_i)$. *Let* $c_0 = 0.558$ *and*

$c_1 = 0.5$. *As* $\delta \triangleq \frac{n}{k \log k} \to 0$ *and* $k \to \infty$, *the bias and variance of* $\hat{S}$ *are upper bounded by*

$$|\mathbb{E}(\hat{S} - S)| \le 2S\big(1 + o_k(1)\big) \exp\Bigg(-\big(1 + o_\delta(1)\big)\sqrt{\kappa \frac{n \log k}{k}}\Bigg),$$

$$\mathsf{var}[\hat{S}] \le O\big(Sk^c\big),$$

for some absolute constant $c < 1$, *and consequently*,

$$\sup_{P \in \mathcal{D}_k} \mathbb{E}\big(\hat{S}(N) - S(P)\big)^2 \le 4k^2\big(1 + o_k(1)\big) \exp\Bigg(-\big(2 + o_\delta(1)\big)\sqrt{\kappa \frac{n \log k}{k}}\Bigg), \tag{26}$$

where $\kappa = 2.494$.

The minimax upper bounds in Theorems 1 and 2 follow from combining Propositions 2 and 3; see Section 5.2.

Note that the optimal estimator (19) relies on the choice of parameters in (24), which in turn, depends on the knowledge of $1/k$, the lower bound on the minimum nonzero probability $p_{\min}$. While k is readily obtainable in certain applications where the samples are uniformly drawn from a database or corpus of known size (see [1, 10] as well as the empirical results in Section 4), it is desirable to construct estimators that are agnostic to $p_{\min}$ and retains the same optimality guarantee. To this end, we provide the following alternative choice of parameters. Let $\tilde{S}$ be the linear estimator defined using the same coefficients in (25), with the approximation interval $[l, r]$ and the degree L in (24) replaced by

$$l = \frac{c_1}{c_0^2} \frac{\log^2(1/\varepsilon)}{n \log n}, \qquad r = \frac{c_1 \log n}{n}, \qquad L = \lfloor c_0 \log n \rfloor. \tag{27}$$

Here, ε is the desired accuracy and the constants c_0, c_1 are the same as in Proposition 3. Following the same analysis as in the proof of Proposition 3, the above choice of parameters leads to the following upper bound of the quadratic risk.

PROPOSITION 4. *Let* c_0, c_1, c *be the same constants as Proposition* 3. *There exist constants* C, C' *such that if* $\varepsilon > n^{-C}$, *then*

$$\mathbb{E}(\tilde{S} - S)^2 \le C'\big(S^2 \varepsilon^{2(1-\sqrt{\alpha})} + Sn^c\big),$$

where $\alpha = \max\big(1 - \frac{c_0^2}{c_1} \frac{n \log n}{k \log^2(1/\varepsilon)}, 0\big)$.

Therefore, whenever the sample size satisfies $n \ge (\frac{c_1}{c_0^2} + o_k(1)) \frac{k}{\log k} \log^2 \frac{1}{\varepsilon}$ and $n \le (\varepsilon^2 k)^{\frac{1}{c}}$, the upper bound is at most $O((\varepsilon k)^2)$, recovering the optimal risk bound in Proposition 3. The new result here is that even when n is not that large the risk degrades gracefully.

We finish this section with a few remarks.

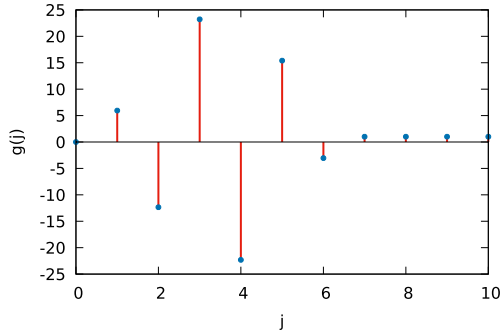

FIG. 1. *Coefficients of estimator $g(j)$ in (19) with $c_0 = 0.45$, $c_1 = 0.5$, $k = 10^6$ and $n = 2 \times 10^5$.*

REMARK 1. Combined with standard concentration inequalities, the mean-square error bound in Proposition 3 can be easily converted to a high-probability bound. In the regime of $n \lesssim k \log k$, for any distribution $P \in \mathcal{D}_k$, the bias of our estimate $\hat{S}$ is at most the uniform approximation error [see (43)]

$$|\mathbb{E}[\hat{S}] - S| \le S \exp\left(-\Theta\left(\sqrt{\frac{n \log k}{k}}\right)\right).$$

The standard deviation is significantly smaller than the bias. Indeed, the coefficients of the linear estimator (19) are uniformly bounded by $\|g\|_\infty^2 \le k^c$ for some absolute constant $c < 1$ (see [44], (70) in the Supplementary Material, as well as Figure 1 for numerical results). Therefore, by Hoeffding's inequality, we have the following concentration bound:

$$\mathbb{P}\big[|\hat{S} - \mathbb{E}[\hat{S}]| \ge tk\big] \le 2 \exp\left(-\frac{t^2 k}{2\|g\|_\infty^2}\right) = \exp\big(-t^2 k^{\Omega(1)}\big).$$

REMARK 2. The estimator (19) belongs to the family of *linear estimators*:

$$\hat{S} = \sum_i f(N_i) = \sum_{j \ge 1} f(j) \Phi_j, \tag{28}$$

which is a linear combination of fingerprints Φ_j's defined in (3).

Other notable examples of linear estimators include:

- Plug-in estimator (7): $\hat{S}_{\text{seen}} = \Phi_1 + \Phi_2 + \cdots$.
- Good–Toulmin estimator [15]: for some $t > 0$,

$$\hat{S}_{\text{GT}} = \hat{S}_{\text{seen}} + t\Phi_1 - t^2\Phi_2 + t^3\Phi_3 - t^4\Phi_4 + \cdots \tag{29}$$

- Efron–Thisted estimator [10]: for some $t > 0$ and $J \in \mathbb{N}$,

$$\hat{S}_{\text{ET}} = \hat{S}_{\text{seen}} + \sum_{j=1}^{J} (-1)^{j+1} t^j b_j \Phi_j, \tag{30}$$

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

By definition, our estimator (19) can be written as

$$\hat{S} = \sum_{j=1}^{L} g(j)\Phi_j + \sum_{j>L} \Phi_j. \tag{31}$$

By (22), P_L is also a polynomial of degree L, which is oscillating and results in coefficients with alternating signs (see Figure 1). Interestingly, this behavior, although counterintuitive, coincides with many classical estimators, such as (29) and (30). The occurrence of negative coefficients can be explained as follows. Note that the rationale of linear estimator is to form a linear prediction the number of unseen Φ_0 using the observed fingerprints $\Phi_1, \Phi_2, \ldots$; this is possible because the fingerprints are correlated. Indeed, since the sum of all fingerprints coincides with the support size, that is, $\sum_{j\geq 0} \Phi_j = S$, for each $j \geq 1$, the random variable Φ_j is negatively correlated with Φ_0, and hence some of the coefficients in the linear estimator are negative.

REMARK 3 (Time complexity). The evaluation of the estimator (28) consists of three parts:

1. Construction of the estimator: $O(L^2) = O(\log^2 k)$, which amounts to computing the coefficients $g(j)$ per (23).
2. Computing the histograms N_i and fingerprints Φ_j: $O(n)$.
3. Evaluating the linear combination: $O(n \wedge k)$, since the number of non-zero terms in the second summation of (28) is at most $n \wedge k$.

Therefore, the total time complexity is $O(n + \log^2 k)$.

REMARK 4. The technique of polynomial approximation has been previously used for estimating nonsmooth functions (L_q-norms) in Gaussian models [4, 19, 21] and more recently for estimating information quantities (entropy and power sums) on large discrete alphabets [20, 43]. The design principle is to approximate the nonsmooth function on a given interval using algebraic or trigonometric polynomials for which unbiased estimators exist; the degree is chosen to balance the bias (approximation error) and the variance (stochastic error). Note that in general uniform approximation by polynomials is only possible on a compact interval. Therefore, in many situations, the construction of the estimator is a two-stage procedure involving *sample splitting*: First, use half of the sample to test whether the corresponding parameter lies in the given interval; Second, use the remaining samples to construct an unbiased estimator for the approximating polynomial if the parameter belongs to the interval or apply plug-in estimators otherwise (see, e.g., [20, 43] and [4], Section 5).

While the benefit of sample splitting is to make the analysis tractable by capitalizing on the independence of the two subsamples, it also sacrifices the statistical accuracy since half of the samples are wasted. In the present paper, to estimate the

support size, we forgo the sample splitting approach 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 of the indicator function arises as a natural step in controlling the bias [see (20)].

4. Experiments. We evaluate the performance of our estimator on both synthetic and real datasets in comparison with popular existing procedures.[2] In the experiments, we choose the constants $c_0 = 0.45, c_1 = 0.5$ in (24), instead of $c_0 = 0.558$ which is optimized to yield the best rate of convergence in Proposition 3 under the i.i.d. sample model. The reason for such a choice is that in the real-data experiments the samples are not necessarily generated independently and dependency leads to a higher variance. By choosing a smaller c_0, the Chebyshev polynomials have a slightly smaller degree, which results in smaller variance and more robustness to model mismatch. Each experiment is averaged over 50 independent trials and the standard deviations are shown as error bars.

Synthetic data. We consider data independently sampled from the following distributions: (a) the uniform distribution with $p_i = \frac{1}{k}$, (b) Zipf distributions with $p_i \propto i^{-\alpha}$ and α being either 1 or 0.5, and (c) an even mixture of geometric distribution and Zipf distribution where for the first half of the alphabet $p_i \propto 1/i$ and for the second half $p_{i+k/2} \propto (1 - \frac{2}{k})^{i-1}$, $1 \le i \le \frac{k}{2}$. The alphabet size k varies in each distribution so that the minimum non-zero mass is roughly 10^{-6}. Accordingly, a degree-6 Chebyshev polynomial is applied. Therefore, according to (31), we apply the polynomial estimator g to symbols appearing at most six times and the plug-in estimator otherwise. In Figure 2 we compare our results with the Good–Turing estimator [14], the Chao 1 estimator [5, 16], the two estimators proposed by Chao and Lee [6] and the linear programming approach proposed by Valiant and Valiant [41]. Here, the Good–Turing estimator refers to first estimate the total probability of seen symbols (sample coverage) by $\hat{C} = 1 - \frac{\Phi_1}{n}$ then estimate the support size by $\hat{S}_{\text{Good-Turing}} = \hat{S}_{\text{seen}}/\hat{C}$; the Chao 1 estimator refers to the bias-corrected form $\hat{S}_{\text{Chao 1}} = \hat{S}_{\text{seen}} + \frac{\Phi_1(\Phi_1-1)}{2(\Phi_2+1)}$. The plug-in estimator simply counts the number of distinct elements observed, which is always outperformed by the Good–Turing estimator in our experiments, and hence omitted in the comparison.

Good–Turing's estimate on sample coverage performs remarkably well in the special case of uniform distributions. This has been noticed and analyzed in [6, 8]. Chao–Lee's estimators are based on Good–Turing's estimate with further correction terms for nonuniform distributions. However, with limited number of samples, if no symbol appears more than once, the sample coverage estimate $\hat{C}$ is zero, and consequently the Good–Turing estimator and Chao–Lee estimators are not even

[2]The implementation of our estimator is available at https://github.com/Albuso0/support.

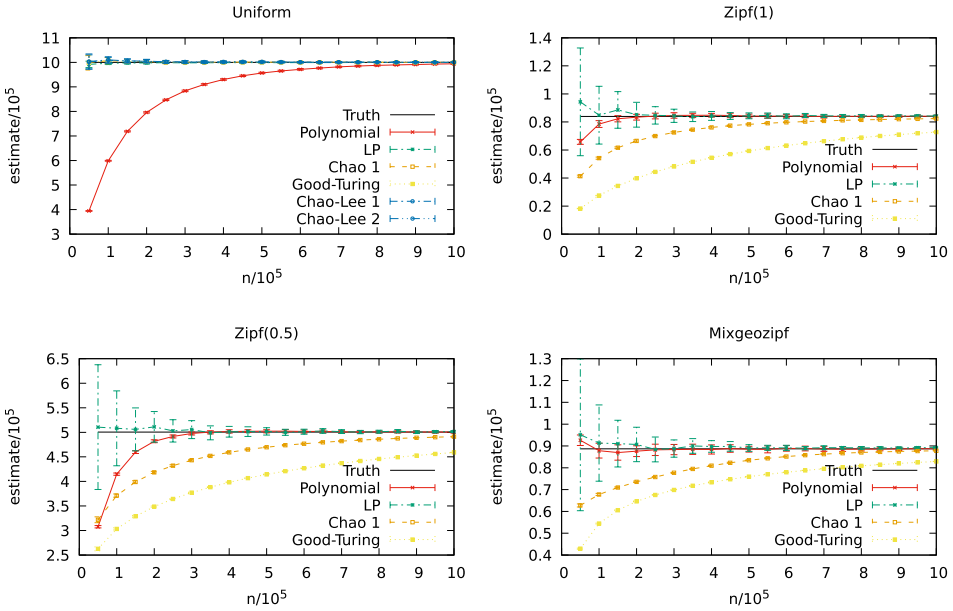

FIG. 2. *Performance comparison under four data-generating distributions.*

well defined. For Zipf and mixture distributions, the output of Chao–Lee's estimators is highly unstable, and thus is omitted from the plots; the convergence rates of the Good–Turing estimator and Chao 1 estimator are much slower than our estimator and the LP estimator, partly because they only use the information of how many symbols occurred exactly once and twice, namely the first two fingerprints Φ_1 and Φ_2, as opposed to the full spectrum of fingerprints $\{\Phi_j\}_{j\geq 1}$, and they suffer provably large bias under nonuniform distributions as simple as mixtures of two uniform distributions (see [44], Appendix C, in the Supplementary Material); the linear programming approach has similar convergence rate to ours but suffers large variance when samples are scarce.

Real data. 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 4800 distinct words. Here and below, the definition of "distinct word" is any distinguishable arrangement of letters that are delimited by spaces, insensitive to cases, with punctuation removed. We randomly sample the text with replacement and generate the fingerprints for estimation. The minimum nonzero mass is naturally the reciprocal of the total number of words, $\frac{1}{32{,}000}$. In this experiment, we use the degree-4 Chebyshev polynomial. We also compare our estimator with the one in [41]. The results are plotted in Figure 3, which shows that the estimator in [41] has similar convergence rate to ours; however, the variance is again much larger and the computational cost of linear programming is significantly higher than linear

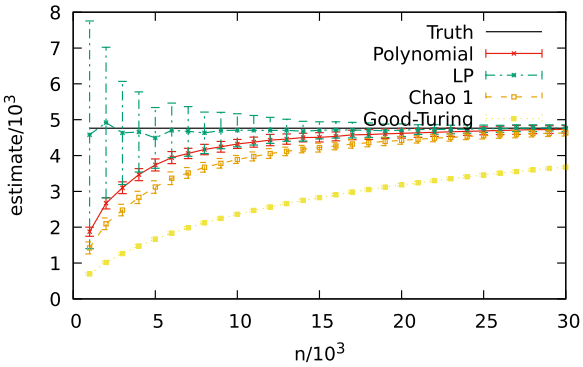

FIG. 3. *Comparison of various estimates of the total number of distinct words in* Hamlet.

estimators, which amounts to computing linear combinations with pre-determined coefficients.

Next we conduct a larger-scale experiment using the *New York Times Corpus* from the years 1987–2007.[3] This corpus has a total of 25,020,626 paragraphs consisting of 996,640,544 words with 2,047,985 distinct words. We randomly sample 1%–50% out of the all paragraphs with replacements and feed the fingerprint to our estimator. The minimum nonzero mass is also the reciprocal of the total number of words, $1/10^9$, and thus the degree-9 Chebyshev polynomial is applied. Using only 20% samples our estimator achieves a relative error of about 10%, which is a systematic error due to the model mismatch: the sampling here is paragraph by paragraph rather than word by word, which induces dependence across samples as opposed to the i.i.d. sampling model for which the estimator is designed; in comparison, the LP estimator[4] suffers a larger bias from this model mismatch. Furthermore, the proposed linear estimator is significantly faster than linear programming based methods: given the sampled data, the curve in Figure 4 corresponding to the LP estimator takes over 5 hours to compute; in contrast, the proposed linear estimator takes only 2 seconds on the same computer, which clearly demonstrate its computational advantage even if one takes into account the fact that our implementation is based on C++ while the LP estimator is in MATLAB.

Finally, we perform the classical experiment of "how many words did Shakespeare know." We feed the fingerprint of the entire Shakespearean canon (see [10], Table 1), which contains 31,534 word types, to our estimator. We choose the minimum nonzero mass to be the reciprocal of the total number of English words, which, according to known estimates, is between 600,000 [29] to 1,000,000 [13],

[3]Dataset available at https://catalog.ldc.upenn.edu/LDC2008T19.

[4]In this large-scale experiment, the original MATLAB code of the linear programming estimator given in [41] is extremely slow; the result in Figure 4 is obtained using an optimized version provided by the author [37].

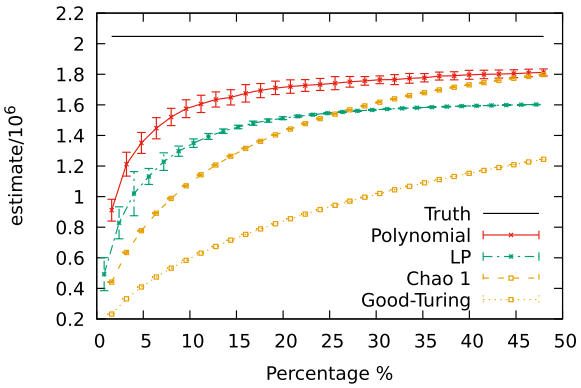

FIG. 4. *Performance comparison using* New York Times Corpus.

and obtain an estimate of 63,148 to 73,460 for Shakespeare's vocabulary size, as compared to 66,534 obtained by Efron–Thisted [10]. Using the alternative choice of parameters that are agnostic to k in Proposition 4, by setting the desired accuracy to be 0.05 and 0.1, we obtain an estimate of 62,355 to 72,454.

5. Proofs.

5.1. *Proofs of Propositions* 1–4.

PROOF OF PROPOSITION 1. For $0 < \nu < 1$, define the set of *approximate* probability vectors by

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

which reduces to the original probability distribution space $\mathcal{D}_k$ if $\nu = 0$. Generalizing the sample complexity $n^*(k, \varepsilon)$ in (8) to the Poisson sampling model over $\mathcal{D}_k(\nu)$, we define

$$\tilde{n}^*(k, \varepsilon, \nu) \triangleq \min\left\{ n \ge 0 : \exists \hat{S},\ \text{s.t.}\ \mathbb{P}\left[\left| \hat{S} - S(P) \right| \ge \varepsilon k \right] \le 0.2, \forall P \in \mathcal{D}_k(\nu) \right\}, \tag{32}$$

where $\hat{S}$ is an integer-valued estimator measurable with respect to $N = (N_1, N_2, \ldots) \overset{\text{ind}}{\sim} \text{Poi}(np_i)$. The sample complexity of the fixed-sample-size and Poissonized model is related by the following lemma.

LEMMA 2. *For any* $\nu \in (0, 1)$ *and any* $\varepsilon \in (0, \frac{1}{2})$,

$$n^*(k, \varepsilon) \ge (1 - \nu)\tilde{n}^*(k, \varepsilon, \nu)\left(1 - O\left(\frac{1}{\sqrt{(1-\nu)\tilde{n}^*(k, \varepsilon, \nu)}}\right)\right). \tag{33}$$

To establish a lower bound of $\tilde{n}^*(k,\varepsilon,\nu)$, 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 $\mathsf{P} = \frac{1}{k}(U_1,\dots,U_k)$ and $\mathsf{P}' = \frac{1}{k}(U'_1,\dots,U'_k)$ with i.i.d. entries. Then $\mathbb{E}[S(\mathsf{P})] - \mathbb{E}[S(\mathsf{P}')] = k(\mathbb{P}[U > 0] - \mathbb{P}[U' > 0])$. Furthermore, both $S(\mathsf{P})$ and $S(\mathsf{P}')$ 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 3. *Let $U, U' \in \{0\} \cup [1+\nu, \lambda]$ be random variables such that $\mathbb{E}[U] = \mathbb{E}[U'] = 1$, $\mathbb{E}[U^j] = \mathbb{E}[U'^j]$ for $j \in [L]$, and $|\mathbb{P}[U > 0] - \mathbb{P}[U' > 0]| = d$, where $\nu \in (0,1)$, $L \in \mathbb{N}$, $d \in (0,1)$ and $\lambda > 1+\nu$. Then, for any $\alpha < 1/2$,*

$$\frac{2\lambda}{k\nu^2} + \frac{2}{k\alpha^2 d^2} + k\left(\frac{en\lambda}{2kL}\right)^L \le 0.6 \quad \Rightarrow \quad \tilde{n}^*\left(k, \frac{(1-2\alpha)d}{2}, \nu\right) \ge n. \tag{34}$$

The proof of Lemma 3 relies on bounds on the total variation distance between two Poisson mixtures with matching moments, as given by the following lemma, which improves upon [43], Lemma 3, in terms of constants. This improvement is crucial for the purpose of obtaining good constants for the sample complexity bounds in (10).

LEMMA 4. *Let V and V' be random variables taking values on $[0,\Lambda]$. If $\mathbb{E}[V^j] = \mathbb{E}[V'^j]$, $j = 1,\dots,L$, then*

$$\mathsf{TV}\big(\mathbb{E}\big[\mathrm{Poi}(V)\big], \mathbb{E}\big[\mathrm{Poi}\big(V'\big)\big]\big) \le \frac{(\Lambda/2)^{L+1}}{(L+1)!}\big(2 + 2^{\Lambda/2-L} + 2^{\Lambda/(2\log 2)-L}\big). \tag{35}$$

In particular, $\mathsf{TV}(\mathbb{E}[\mathrm{Poi}(V)], \mathbb{E}[\mathrm{Poi}(V')]) \le (\frac{e\Lambda}{2L})^L$. Moreover, if $L > \frac{e}{2}\Lambda$, then

$$\mathsf{TV}\big(\mathbb{E}\big[\mathrm{Poi}(V)\big], \mathbb{E}\big[\mathrm{Poi}\big(V'\big)\big]\big) \le \frac{2(\Lambda/2)^{L+1}}{(L+1)!}\big(1 + o(1)\big), \qquad \Lambda \to \infty.$$

Now applying Lemma 7 in the Appendix, we obtain two random variables $U, U' \in \{0\} \cup [1+\nu, \lambda]$ such that $\mathbb{E}[U] = \mathbb{E}[U'] = 1$, $\mathbb{E}[U^j] = \mathbb{E}[U'^j]$, $j = 1,\dots,L$ and

$$\begin{aligned}\mathbb{P}[U > 0] - \mathbb{P}\big[U' > 0\big] &= 2E_{L-1}\left(\frac{1}{x}, [1+\nu, \lambda]\right) \\ &= \frac{(1+\sqrt{\frac{1+\nu}{\lambda}})^2}{1+\nu}\left(1 - \frac{2\sqrt{\frac{1+\nu}{\lambda}}}{1+\sqrt{\frac{1+\nu}{\lambda}}}\right)^L \triangleq d,\end{aligned}$$

where the value of $E_{L-1}(\frac{1}{x}, [1+\nu,\lambda])$ follows from [35], 2.11.1. To apply Lemma 3 and obtain a lower bound of $\tilde{n}^*(k,\varepsilon,\nu)$, we need to pick the parame-

ters depending on the given k and ε to fulfill:

$$\frac{(1-2\alpha)d}{2} \geq \varepsilon, \tag{36}$$

$$\frac{2\lambda}{k\nu^2} + \frac{2}{k\alpha^2 d^2} + k\left(\frac{en\lambda}{2kL}\right)^L \leq 0.6. \tag{37}$$

Let

$$L = \lfloor c_0 \log k \rfloor, \qquad \lambda = \left(\frac{\gamma \log k}{\log(1/2\varepsilon)}\right)^2, \qquad n = C\frac{k}{\log k}\log^2\frac{1}{2\varepsilon},$$
$$\alpha = \frac{1}{k^{1/3}}, \qquad \nu = \sqrt{\sqrt{\lambda/k}(1-2\varepsilon)},$$

for some $c_0, \gamma, C \asymp 1$ to be specified, and by assumption $L, \lambda \to \infty$, $\frac{\alpha}{1-2\varepsilon} = o_k(1)$, $\frac{\nu}{1-2\varepsilon} = o_\tau(1) + o_k(1)$, $1/\lambda = o_\delta(1)$. Since $d \geq \frac{1}{1+\nu}(1 - 2\sqrt{\frac{1+\nu}{\lambda}})^L$, a sufficient condition for (36) is that

$$\left(1 - 2\sqrt{\frac{1+\nu}{\lambda}}\right)^L \geq 2\varepsilon\frac{1+\nu}{1-2\alpha} \quad \Leftrightarrow \quad \frac{\gamma}{c_0} > 2 + o_\tau(1) + o_\delta(1) + o_k(1). \tag{38}$$

Now we consider (37). By the choice of ν and α, we have

$$\nu \gg \sqrt{\lambda/k}, \qquad \alpha \gg 1/\sqrt{k}d,$$

since $1 - 2\varepsilon \gg \frac{\sqrt{\log k}}{k^{1/4}}$, $d \geq \frac{2\varepsilon}{1-2\alpha}$ and $\varepsilon = k^{-o(1)}$. Then the first two terms in (37) vanish. The last term in (37) vanishes as long as the constant $C < \frac{2c_0}{e\gamma^2}e^{-1/c_0}$. By the fact that

$$\sup\left\{\frac{2c_0}{e\gamma^2}e^{-1/c_0} : 0 < 2c_0 < \gamma\right\} = \frac{1}{2e^2},$$

the optimal C satisfying (38) is $\frac{1+o_\delta(1)+o_\tau(1)+o_k(1)}{2e^2}$. Therefore, combining (36)–(37) and applying (34), we obtain a lower bound of $\tilde{n}^*$ that

$$\tilde{n}^*(k, \varepsilon, \nu) \geq \frac{1 + o_\delta(1) + o_\tau(1) + o_k(1)}{2e^2}\frac{k}{\log k}\log^2\frac{1}{2\varepsilon}.$$

Since $1 - 2\varepsilon \gg \frac{\sqrt{\log k}}{k^{1/4}}$, we have $\tilde{n}^*(k, \varepsilon, \nu) \gg \sqrt{k}$. Applying Lemma 2, we conclude the desired lower bound of $n^*(k, \varepsilon)$. □

PROOF OF PROPOSITION 2. First, we consider the bias:

$$\left|\mathbb{E}(\hat{S}_{\text{seen}} - S)\right| = \sum_{i:p_i \geq \frac{1}{k}}\left(1 - \mathbb{P}(N_i \geq 1)\right) = \sum_{i:p_i \geq \frac{1}{k}}\exp(-np_i) \leq S\exp(-n/k).$$

The variance satisfies

$$\mathsf{var}[\hat{S}_{\mathrm{seen}}] = \sum_{i:p_i \geq \frac{1}{k}} \mathsf{var}\mathbf{1}_{\{N_i>0\}} \leq \sum_{i:p_i \geq \frac{1}{k}} \exp(-np_i) \leq S\exp(-n/k).$$

The conclusion follows.

For the negative result, under the Poissonized model and with the samples drawn from the uniform distribution, the plug-in estimator $\hat{S}_{\mathrm{seen}}$ is distributed as $\mathrm{Binomial}(k, 1-e^{-n/k})$. If $n \leq (1-\delta)k\log\frac{1}{\varepsilon} < k\log\frac{1}{\varepsilon}$, then $1-e^{-n/k} < 1-\varepsilon$. By the Chernoff bound,

$$\begin{aligned}\mathbb{P}\big[|\hat{S}_{\mathrm{seen}} - S(P)| \leq \varepsilon k\big] &= \mathbb{P}\big[\mathrm{Binomial}\big(k, 1-e^{-n/k}\big) \geq (1-\varepsilon)k\big] \\ &\leq e^{-kd(1-\varepsilon\|1-e^{-n/k})} = e^{-kd(\varepsilon\|e^{-n/k})},\end{aligned}$$

where $d(p\|q) \triangleq p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$ is the binary divergence function. Since $e^{-n/k} \geq \varepsilon^{1-\delta} > \varepsilon$,

$$d\big(\varepsilon\|e^{-n/k}\big) \geq d\big(\varepsilon\|\varepsilon^{1-\delta}\big) \geq d\big(k^{-1}\|k^{-1+\delta}\big) \asymp k^{-1+\delta},$$

where the middle inequality follows from the fact that $\varepsilon \mapsto d(\varepsilon\|\varepsilon^{1-\delta})$ is increasing near zero. Therefore, $\mathbb{P}[|\hat{S}_{\mathrm{seen}} - S(P)| \leq \varepsilon k] \leq \exp(-\Omega(k^{\delta}))$. □

PROOF OF PROPOSITION 3. First, we consider the bias. By (20), the bias of $\hat{S}$ is

$$\big|\mathbb{E}[\hat{S} - S]\big| \leq \sum_{i:p_i>0} \big|e^{-np_i} P_L(p_i)\big| \leq S \max_{x\in[\frac{1}{k},1]} \big|e^{-nx} P_L(x)\big|, \tag{39}$$

where P_L is the Chebyshev polynomial in (22). Recall that $L = \lfloor c_0 \log k\rfloor$, $l = \frac{1}{k}$, $r = \frac{c_1 \log k}{n}$. Then

$$\max_{x\in[l,r]} \big|P_L(x)\big| = \frac{1}{|T_L(-\frac{r+l}{r-l})|}, \tag{40}$$

$$\max_{x\in(r,1]} \big|e^{-nx} P_L(x)\big| = \frac{\max_{x\in(r,1]} e^{-nx}|T_L(\frac{2x-r-l}{r-l})|}{|T_L(-\frac{r+l}{r-l})|}. \tag{41}$$

We need the following lemma to upper bound (41).

LEMMA 5. *If* $\alpha \triangleq L/\beta = \Omega(1)$, *then*

$$\max_{x\geq 1} e^{-\beta x} T_L(x) = \frac{1}{2}\left(\frac{\alpha + \sqrt{\alpha^2+1}}{e^{\sqrt{1+1/\alpha^2}}}\big(1+o_L(1)\big)\right)^L, \qquad L\to\infty.$$

Applying Lemma 5 to (41) with $L = \lfloor c_0 \log k \rfloor$, $\beta = \frac{nr(1-\delta)}{2} = \frac{c_1 \log k(1-\delta)}{2}$, we obtain that

$$\max_{x \ge r} \left| e^{-nx} T_L\left(\frac{2x - r - l}{r - l}\right)\right| \le \frac{1}{2}\left(\frac{2\rho + \sqrt{(2\rho)^2 + 1}}{e^{\sqrt{1+1/(2\rho)^2}+1/(2\rho)}}\big(1 + o_k(1) + o_\delta(1)\big)\right)^L, \tag{42}$$

where $\rho \triangleq c_0/c_1$. Combining (40) to (42), $\max_{x\in[l,1]} |e^{-nx} P_L(x)| \le \frac{1+o_k(1)+o_\delta(1)}{|T_L(-\frac{1+\delta}{1-\delta})|}$ as long as we pick the constant ρ such that $\frac{2\rho+\sqrt{(2\rho)^2+1}}{e^{\sqrt{1+1/(2\rho)^2}+1/(2\rho)}} < 1 \Leftrightarrow \operatorname{arcsinh}(2\rho) < \frac{1+\sqrt{1+4\rho^2}}{2\rho}$, or equivalently, $\rho < \rho^* \approx 1.1$. Then, by (39), the bias of $\hat{S}$ is at most

$$\begin{aligned} \big|\mathbb{E}[\hat{S} - S]\big| &\le S\frac{1 + o_k(1) + o_\delta(1)}{|T_L(-\frac{1+\delta}{1-\delta})|} \\ &\le 2S\big(1 + o_k(1) + o_\delta(1)\big)\left(1 - \frac{2\sqrt{\delta}}{1+\sqrt{\delta}}\right)^L \qquad (43) \\ &= 2S\big(1 + o_k(1)\big)\exp\left(-\big(1 + o_\delta(1)\big)\sqrt{4c_0\rho\frac{n\log k}{k}}\right). \end{aligned}$$

Now we turn to the variance of $\hat{S}$:

$$\begin{aligned} \mathsf{var}[\hat{S}] &= \sum_{i:p_i>0} \mathsf{var}[u_{N_i}\mathbf{1}_{\{N_i\le L\}}] \\ &\le \sum_{i:p_i>0} \mathbb{E}\big[u_{N_i}^2\mathbf{1}_{\{N_i\le L\}}\big] \qquad (44) \\ &\le \|u\|_\infty^2 \sum_{i:p_i>0} \mathbb{P}[N_i \le L], \end{aligned}$$

where $\Phi_j \triangleq \sum_i \mathbf{1}_{\{N_i=j\}}$ is the fingerprint of samples. The following lemma shows that $|u_j|$ is at most exponential in the degree of the polynomial.

LEMMA 6. *Let a_j be defined as (22) and u_j be defined as (25). Then*

$$\|u\|_\infty \le \frac{e\sqrt{L}}{2}\exp\left(\tau\left(\frac{L}{nr}\right)L\right), \tag{45}$$

where $\tau(x) \triangleq \operatorname{arcsinh}(2x) - \frac{\sqrt{1+4x^2}-1}{2x}$.

From (44) and (45), the variance of $\hat{S}$ is at most

$$\mathsf{var}[\hat{S}] \le S\frac{e^2 L}{4}k^{2c_0\tau(\rho)}. \tag{46}$$

Then, from (43) and (46), we obtain that

$$\sup_{P\in\mathcal{D}_k}\mathbb{E}(\hat{S}-S)^2 \le 4k^2\big(1+o_k(1)\big)\exp\bigg(-2\big(1+o_\delta(1)\big)\sqrt{\frac{2\rho}{\tau(\rho)}\frac{n\log k}{k}}\bigg) + \frac{e^2c_0\log k}{4}k^{1+2c_0\tau(\rho)}.$$

Note that the first term is $4k^{2-o_\delta(1)}$. Therefore, as long as we pick constant c_0 such that $2c_0\tau(\rho)<1$ the second term is lower order than the first term, and thus

$$\sup_{P\in\mathcal{D}_k}\mathbb{E}(\hat{S}-S)^2 \le 4k^2\big(1+o_k(1)\big)\exp\bigg(-2\big(1+o_\delta(1)\big)\sqrt{\frac{2\rho}{\tau(\rho)}\frac{n\log k}{k}}\bigg).$$

The conclusion follows from the fact that $\sup_{\rho<\rho^*}2\rho/\tau(\rho)\approx 2.494$, which corresponds to choosing $c_0\approx 0.558$ and $c_1=0.5$. □

PROOF OF PROPOSITION 4. Let $\delta=l/r$, which is less than some absolute constant C/c_0 when $\varepsilon>n^{-C}$. The upper bound of the mean squared error is essentially the same as the proof of Proposition 3. The bias of $\tilde{S}$ is at most $S\max_{x\in[p_{\min},1]}e^{-nx}|P_L(x)|$ given in (39). For $p_i\in[l,r]$, the bias is upper bounded by the uniform approximation error

$$\max_{x\in[l,r]}\big|P_L(x)\big| \le \frac{1}{|T_L(-\frac{1+\delta}{1-\delta})|} \le 2\bigg(1-\frac{2\sqrt{\delta}}{1+\sqrt{\delta}}\bigg)^L \le 2\varepsilon.$$

For $p_i>r$, following (41)–(42), we have $e^{-np_i}|P_L(p_i)|=o(\varepsilon)$ as long as $c_0/c_1<\rho^*\approx 1.1$. For $p_i\in[p_{\min},l]$, since the shifted Chebyshev polynomial P_L is monotone on $(-\infty,l)$, we have

$$\begin{aligned}\big|P_L(x)\big| &\le \frac{|T_L(\frac{2p_{\min}-r-l}{r-l})|}{|T_L(\frac{-r-l}{r-l})|} = \frac{|T_L(1+\frac{2\alpha\delta}{1-\delta})|}{|T_L(1+\frac{2\delta}{1-\delta})|}\\ &= \exp\big(-\big(1-o_\delta(1)\big)2(1-\sqrt{\alpha})L\sqrt{\delta}\big) \le \varepsilon^{1-\sqrt{\alpha}},\end{aligned}$$

where $\alpha=\frac{l-p_{\min}}{l}\in(0,1)$ denotes the relative deviation of l from $p_{\min}$, and we used the following equation of the Chebyshev polynomial evaluated at $1+x$ for $x>0$:

$$\begin{aligned}T_L(1+x) &= \frac{1}{2}\big(\big(1+x-\sqrt{x^2+2x}\big)^L+\big(1+x+\sqrt{x^2+2x}\big)^L\big)\\ &= \frac{1}{2}\exp\big(\big(1+o_x(1)\big)L\sqrt{2x}\big).\end{aligned}$$

To conclude, the bias of $\tilde{S}$ is at most

$$\max_{x\in[p_{\min},1]}e^{-nx}\big|P_L(x)\big| \le S\varepsilon^{1-\sqrt{(1-p_{\min}/l)\vee 0}}.$$

By similar analysis to (44) and (45), the variance is at most $O(Sn^c)$ for some constant $c < 1$. $\square$

5.2. *Proofs of Theorems 1 and 2.*

PROOF OF THEOREM 1. By the Markov inequality,

$$n^*(k,\varepsilon) > n \quad \Rightarrow \quad R^*(k,n) > 0.1k^2\varepsilon^2.$$

Therefore, our lower bound is

$$R^*(k,n) \geq \sup\{0.1k^2\varepsilon^2 : n^*(k,\varepsilon) > n\} = 0.1k^2\varepsilon_*^2,$$

where $\varepsilon_* \triangleq \{\varepsilon : n^*(k,\varepsilon) > n\}$. By the lower bound of $n^*(k,\varepsilon)$ in (14), we obtain that

$$\varepsilon_* \geq \exp\left(-\left(\sqrt{2}e + o_\delta(1) + o_{\delta'}(1) + o_k(1)\right)\sqrt{\frac{n\log k}{k}}\right),$$

as $\delta \triangleq \frac{n}{k\log k} \to 0$, $\delta' \triangleq \frac{k}{n\log k} \to 0$, and $k \to \infty$. Then we conclude the lower bound part of (6), which implies the lower bound part of (5) when $n \lesssim k\log k$.

For the lower bound part of (5) when $n \gtrsim k\log k$, we apply Le Cam's two-point method [23] by considering two possible distributions, namely $P = \text{Bern}(0)$ and $Q = \text{Bern}(\frac{1}{k})$. Then

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

Since $n \gtrsim k\log k$, we have $n\log(1-\frac{1}{k}) - 2\log k \gtrsim -\frac{n}{k}$.

Combining Lemma 1 and Proposition 3 yields the upper bound part of (6), which also implies the upper bound of (5) when $n \lesssim k\log k$. The upper bound part of (5) when $n \gtrsim k\log k$ follows from Proposition 2. $\square$

PROOF OF THEOREM 2. The lower bound part of (10) follows from Proposition 1. Consequently, we obtain the lower bound part of (9) for $\frac{1}{k^c} \leq \varepsilon \leq c_0$ for the fixed constant $c_0 < 1/2$, where c is some small constant.

The lower bound part of (9) for $\frac{1}{k} \leq \varepsilon \leq \frac{1}{k^c}$ simply follows from the fact that $\varepsilon \mapsto n^*(k,\varepsilon)$ is decreasing:

$$n^*(k,\varepsilon) \geq n^*\big(k, 1/k^c\big) \gtrsim k\log k \asymp \frac{k}{\log k}\log^2\frac{1}{\varepsilon}.$$

By the Markov inequality,

$$R^*(k,n) \leq 0.1k^2\varepsilon^2 \quad \Rightarrow \quad n^*(k,\varepsilon) \leq n. \tag{47}$$

Therefore, our upper bound is

$$n^*(k,\varepsilon) \le \inf\{n : R^*(k,n) \le 0.1k^2\varepsilon^2\}.$$

By the upper bound of $R^*(k,n)$ in (26), we obtain that

$$n^*(k,\varepsilon) \le \frac{1+o_{\delta'}(1)+o_\varepsilon(1)+o_k(1)}{\kappa}\frac{k}{\log k}\log^2\frac{1}{\varepsilon}$$

as $\delta' \triangleq \frac{\log(1/\varepsilon)}{\log k} \triangleq 0$, $\varepsilon \to 0$, and $k \to \infty$. Consequently, we obtain the upper bound part of (9) when $\frac{1}{k^c} \le \varepsilon \le c_0$ for the fixed constant $c_0 < 1/2$, where c is some small constant.

The upper bound part of Theorem 2 when $\frac{1}{k} \le \varepsilon \le \frac{1}{k^c}$ follows from the monotonicity of $\varepsilon \mapsto n^*(k,\varepsilon)$ that

$$n^*(k,\varepsilon) \le n^*(k,1/k) \le 3k\log k \asymp \frac{k}{\log k}\log^2\frac{1}{\varepsilon},$$

where the middle inequality follows from Proposition 2 and (47). □

APPENDIX: DUAL PROGRAM OF (13)

Define the following infinite-dimensional linear program:

$$\begin{aligned}
\mathcal{E}_1^* \triangleq \sup\ & \mathbb{P}[U'=0] - \mathbb{P}[U=0] \\
\text{s.t. } & \mathbb{E}[U] = \mathbb{E}[U'] = 1 \\
& \mathbb{E}[U^j] = \mathbb{E}[U'^j], \qquad j = 1,\dots,L+1, \\
& U, U' \in \{0\} \cup I,
\end{aligned} \tag{48}$$

where $I = [a,b]$ with $b > a \ge 1$ and the variables are probability measures on I (distributions of the random variables U, U'). Then (13) is a special case of (48) with $I = [1,\lambda]$.

LEMMA 7. $\mathcal{E}_1^* = \inf_{p\in\mathcal{P}_L} \sup_{x\in I} |\frac{1}{x} - p(x)|$.

PROOF. We first show that (13) coincides with the following optimization problem:

$$\begin{aligned}
\mathcal{E}_2^* \triangleq \sup\ & \mathbb{E}\left[\frac{1}{X}\right] - \mathbb{E}\left[\frac{1}{X'}\right] \\
\text{s.t. } & \mathbb{E}[X^j] = \mathbb{E}[X'^j], \qquad j = 1,\dots,L, \\
& X, X' \in I.
\end{aligned} \tag{49}$$

Given any feasible solution U, U' to (13), construct X, X' with the following distributions:

$$\begin{aligned}
P_X(\mathrm{d}x) &= x\,P_U(\mathrm{d}x), \\
P_{X'}(\mathrm{d}x) &= x\,P_{U'}(\mathrm{d}x),
\end{aligned} \tag{50}$$

It is straightforward to verify that X, X' are feasible for (49) and

$$\mathcal{E}_2^* \geq \mathbb{E}\left[\frac{1}{X}\right] - \mathbb{E}\left[\frac{1}{X'}\right] = \mathbb{P}\big[U' = 0\big] - \mathbb{P}[U = 0].$$

Therefore, $\mathcal{E}_2^* \geq \mathcal{E}_1^*$.

On the other hand, given any feasible X, X' for (49), construct U, U' with the distributions:

$$\begin{aligned} P_U(\mathrm{d}u) &= \left(1 - \mathbb{E}\left[\frac{1}{X}\right]\right)\delta_0(\mathrm{d}u) + \frac{1}{u}P_X(\mathrm{d}u), \\ P_{U'}(\mathrm{d}u) &= \left(1 - \mathbb{E}\left[\frac{1}{X'}\right]\right)\delta_0(\mathrm{d}u) + \frac{1}{u}P_{X'}(\mathrm{d}u), \end{aligned} \tag{51}$$

which are well defined since $X, X' \geq 1$, and hence $\mathbb{E}[\frac{1}{X}] \leq 1, \mathbb{E}[\frac{1}{X'}] \leq 1$. Then U, U' are feasible for (13), and hence

$$\mathcal{E}_1^* \geq \mathbb{P}\big[U' = 0\big] - \mathbb{P}[U = 0] = \mathbb{E}\left[\frac{1}{X}\right] - \mathbb{E}\left[\frac{1}{X'}\right].$$

Therefore, $\mathcal{E}_1^* \geq \mathcal{E}_2^*$. Finally, the dual of (49) is precisely the best polynomial approximation problem (see, e.g., [43], Appendix E), and hence

$$\mathcal{E}_1^* = \mathcal{E}_2^* = \inf_{p \in \mathcal{P}_L} \sup_{x \in I} \left|\frac{1}{x} - p(x)\right|. \qquad \square$$

Acknowledgments. This work was completed in part when the first author was visiting the Simons Institute for the Theory of Computing, whose generous support is acknowledged. The authors thank Luca Trevisan for helpful comments pertaining to Theorem 2 and Greg Valiant [37] for providing an updated version of the code in [41]. The authors are grateful to Dan Roth and Mark Sammons for help with the datasets used in Figure 4. The authors also thank the anonymous reviewers and editor for constructive comments.

SUPPLEMENTARY MATERIAL

Supplementary material for "Chebyshev polynomials, moment matching and optimal estimation of the unseen" (DOI: 10.1214/17-AOS1665SUPP; .pdf). Due to space constraints, the technical proofs have been given in the supplementary documents [44].

REFERENCES

[1] BAR-YOSSEF, Z., JAYRAM, T., KUMAR, R., SIVAKUMAR, D. and TREVISAN, L. (2002). Counting distinct elements in a data stream. In *Proceedings of the 6th Randomization and Approximation Techniques in Computer Science* 1–10. Springer, Berlin.

[2] BUNGE, J. and FITZPATRICK, M. (1993). Estimating the number of species: A review. *J. Amer. Statist. Assoc.* **88** 364–373.

[3] BURNHAM, K. P. and OVERTON, W. S. (1979). Robust estimation of population size when capture probabilities vary among animals. *Ecology* **60** 927–936.
[4] CAI, T. T. and LOW, M. G. (2011). Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional. *Ann. Statist.* **39** 1012–1041. MR2816346
[5] CHAO, A. (1984). Nonparametric estimation of the number of classes in a population. *Scand. J. Stat.* 265–270. MR0793175
[6] CHAO, A. and LEE, S.-M. (1992). Estimating the number of classes via sample coverage. *J. Amer. Statist. Assoc.* **87** 210–217. MR1158639
[7] CHARIKAR, M., CHAUDHURI, S., MOTWANI, R. and NARASAYYA, V. (2000). Towards estimation error guarantees for distinct values. In *Proceedings of the Nineteenth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems* (*PODS*) 268–279. ACM, New York.
[8] DARROCH, J. and RATCLIFF, D. (1980). A note on capture-recapture estimation. *Biometrics* **36** 149–153.
[9] DZYADYK, V. K. and SHEVCHUK, I. A. (2008). *Theory of Uniform Approximation of Functions by Polynomials*. de Gruyter, Berlin.
[10] EFRON, B. and THISTED, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika* **63** 435–447.
[11] FISHER, R. A., CORBET, A. S. and WILLIAMS, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.* **12** 42–58.
[12] GANDOLFI, A. and SASTRI, C. C. A. (2004). Nonparametric estimations about species not observed in a random sample. *Milan J. Math.* **72** 81–105. MR2099128
[13] Global Language Monitor. Number of words in the English language. https://www.languagemonitor.com/global-english/no-of-words/. Accessed: 2016-02-16.
[14] GOOD, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika* **40** 237–264.
[15] GOOD, I. J. and TOULMIN, G. H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika* **43** 45–63.
[16] GOTELLI, N. J. and COLWELL, R. K. (2011). Estimating species richness. *Biological Diversity*: *Frontiers in Measurement and Assessment* **12** 39–54.
[17] HARRIS, B. (1968). Statistical inference in the classical occupancy problem unbiased estimation of the number of classes. *J. Amer. Statist. Assoc.* **63** 837–847. MR0231480
[18] HUANG, S.-P. and WEIR, B. (2001). Estimating the total number of alleles using a sample coverage method. *Genetics* **159** 1365–1373.
[19] IBRAGIMOV, I. A., NEMIROVSKII, A. S. and KHAS'MINSKII, R. Z. (1987). Some problems on nonparametric estimation in Gaussian white noise. *Theory Probab. Appl.* **31** 391–406.
[20] JIAO, J., VENKAT, K., HAN, Y. and WEISSMAN, T. (2015). Minimax estimation of functionals of discrete distributions. *IEEE Trans. Inform. Theory* **61** 2835–2885.
[21] LEPSKI, O., NEMIROVSKI, A. and SPOKOINY, V. (1999). On estimation of the L_r norm of a regression function. *Probab. Theory Related Fields* **113** 221–253.
[22] LEWONTIN, R. C. and PROUT, T. (1956). Estimation of the number of different classes in a population. *Biometrics* **12** 211–223.
[23] LE CAM, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer, New York. MR0856411
[24] MAO, C. X. and LINDSAY, B. G. (2007). Estimating the number of classes. *Ann. Statist.* **35** 917–930. MR2336874
[25] MARCHAND, J. and SCHROECK JR, F. (1982). On the estimation of the number of equally likely classes in a population. *Comm. Statist. Theory Methods* **11** 1139–1146. MR0652619

[26] MCNEIL, D. R. (1973). Estimating an author's vocabulary. *J. Amer. Statist. Assoc.* **68** 92–96. MR0373124

[27] MITZENMACHER, M. and UPFAL, E. (2005). *Probability and Computing*: *Randomized Algorithms and Probabilistic Analysis*. Cambridge Univ. Press, Cambridge. MR2144605

[28] ORLITSKY, A., SURESH, A. T. and WU, Y. (2016). Optimal prediction of the number of unseen species. *Proc. Natl. Acad. Sci. USA* **113** 13283–13288.

[29] Oxford English Dictinary. http://public.oed.com/about/. Accessed: 2016-02-16.

[30] RASKHODNIKOVA, S., RON, D., SHPILKA, A. and SMITH, A. (2009). Strong lower bounds for approximating distribution support size and the distinct elements problem. *SIAM J. Comput.* **39** 813–842.

[31] ROBBINS, H. E. (1968). Estimating the total probability of the unobserved outcomes of an experiment. *Ann. Math. Stat.* **39** 256–257.

[32] SAMUEL, E. (1968). Sequential maximum likelihood estimation of the size of a population. *Ann. Math. Stat.* **39** 1057–1068.

[33] STEELE, J. M. (1986). An Efron–Stein inequality for nonsymmetric statistics. *Ann. Statist.* **14** 753–758. MR0840528

[34] THISTED, R. and EFRON, B. (1987). Did Shakespeare write a newly-discovered poem? *Biometrika* **74** 445–455.

[35] TIMAN, A. F. (1963). *Theory of Approximation of Functions of a Real Variable*. Pergamon Press, Elmsford, NY.

[36] TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. Revised and extended from the 2004 French original. Translated by Vladimir Zaiats. Springer, New York, NY. MR2724359

[37] VALIANT, G. (2017). Private communication.

[38] VALIANT, G. and VALIANT, P. (2010). A CLT and tight lower bounds for estimating entropy. In *Electronic Colloquium on Computational Complexity* (*ECCC*) **17** 179.

[39] VALIANT, G. and VALIANT, P. (2011). Estimating the unseen: An $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs. In *Proceedings of the* 43*rd Annual ACM Symposium on Theory of Computing* 685–694.

[40] VALIANT, G. and VALIANT, P. (2011). The power of linear estimators. In *Foundations of Computer Science* (*FOCS*), 2011 *IEEE* 52*nd Annual Symposium on* 403–412. IEEE, New York.

[41] VALIANT, P. and VALIANT, G. (2013). Estimating the unseen: Improved estimators for entropy and other properties. In *Advances in Neural Information Processing Systems* 2157–2165.

[42] WU, Y. and YANG, P. (2016). Sample complexity of the distinct elements problem. Available at arXiv:1612.03375.

[43] WU, Y. and YANG, P. (2016). Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Trans. Inform. Theory* **62** 3702–3720.

[44] WU, Y. and YANG, P. (2017). Supplement to "Chebyshev polynomials, moment matching and optimal estimation of the unseen." DOI:10.1214/17-AOS1665SUPP.

DEPARTMENT OF STATISTICS
AND DATA SCIENCE
YALE UNIVERSITY
24 HILLHOUSE AVE
NEW HAVEN, CONNECTICUT 06511
USA
E-MAIL: yihong.wu@yale.edu

DEPARTMENT OF ECE
AND COORDINATED SCIENCE LAB
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
1308 WEST MAIN ST
URBANA, ILLINOIS 61801
USA
E-MAIL: pyang14@illinois.edu