Math. Stat. Learn. 1 (2018), 37–72
DOI 10.4171/MSL/1-1-2

Mathematical Statistics and Learning

Sample complexity of the distinct elements problem

Yihong Wu and Pengkun Yang

Abstract. We consider the distinct elements problem, where the goal is to estimate the number of distinct colors in an urn containing k balls based on n samples drawn with replacements. Based on discrete polynomial approximation and interpolation, we propose an estimator with additive error guarantee that achieves the optimal sample complexity within $O(\log\log k)$ factors, and in fact within constant factors for most cases. The estimator can be computed in $O(n)$ time for an accurate estimation. The result also applies to sampling without replacement provided the sample size is a vanishing fraction of the urn size.

One of the key auxiliary results is a sharp bound on the minimum singular values of a real rectangular Vandermonde matrix, which might be of independent interest.

Mathematics Subject Classification (2010). 62G05; 41A05, 41A10, 62C20, 62D05.

Keywords. Sampling large population, nonparametric statistics, discrete polynomial approximation, orthogonal polynomials, Vandermonde matrix, minimaxity.

1. The Distinct Elements problem

The Distinct Elements problem [5] refers to the following question:

> *Given n balls randomly drawn from an urn containing k colored balls, how to estimate the total number of distinct colors in the urn?*

Originating from ecology, numismatics, and linguistics, this problem is also known as the *species problem* in the statistics literature [2, 29]. Apart from the theoretical interests, it has a wide array of applications in various fields, such as estimating the number of species in a population of animals [14, 20], the number of dies used to mint an ancient coinage [12], and the vocabulary size of an author [13]. In computer science, this problem frequently arises in large-scale databases, network monitoring, and data mining [3, 5, 37], where the objective is to estimate the types of database entries or IP addresses from limited observations, since it is typically impossible to have full access to the entire database or keep track of all the network traffic. The key challenge in the Distinct Elements problem is the following: given a small set of samples where most of the colors are not observed, how to accurately extrapolate the number of unseens?

1.1. Main results. The fundamental limit of the Distinct Elements problem is characterized by the sample complexity, i.e. the smallest sample size needed to estimate the number of distinct colors with a prescribed accuracy and confidence level. A formal definition is the following:

Definition 1. The sample complexity $n^*(k, \Delta)$ is the minimal sample size n such that there exists an integer-valued estimator $\widehat{C}$ based on n balls drawn independently with replacements from the urn, such that $\mathbb{P}[|\widehat{C} - C| \geq \Delta] \leq 0.1$ for any urn containing k balls with C different colors.[1]

The main results of this paper provide bounds and constant-factor approximations of the sample complexity in various regimes summarized in Table 1, as well as computationally efficient algorithms. Below we highlight a few important conclusions drawn from Table 1:

From linear to sublinear. From the result for $k^{0.5+\delta} \leq \Delta \leq ck$ in Table 1, we conclude that the sample complexity is sublinear in k if and only if $\Delta = k^{1-o(1)}$, which also holds for sampling without replacement. To estimate within a constant fraction of balls $\Delta = ck$ for any small constant c, the sample complexity is $\Theta(k/\log k)$, which coincides with the general support size estimation problem [45, 47] (see Section 1.2 for a detailed comparison). However, in other regimes we can achieve better performance by exploiting the discrete nature of the Distinct Elements problem.

From linear to superlinear. The transition from linear to superlinear sample complexity occurs near $\Delta = \sqrt{k}$. Although the exact sample complexity near $\Delta = \sqrt{k}$ is not completely resolved in the current paper, the lower bound and upper bound in Table 1 differ by a factor of at most $\log\log k$. In particular, the estimator via interpolation can achieve $\Delta = \sqrt{k}$ with $n = O(k \log\log k)$ samples, and achieving a precision of $\Delta \leq k^{0.5-o(1)}$ requires strictly superlinear sample size.

To establish the sample complexity, our lower bounds are obtained under zero-one loss and our upper bounds are under the (stronger) quadratic loss. Hence we also obtain the following characterization of the minimax mean squared error (MSE) of the Distinct Elements problem:

$$\min_{\widehat{C}} \max_{k\text{-ball urn}} \mathbb{E}\left(\frac{\widehat{C} - C}{k}\right)^2 = \exp\left\{-\Theta\left(\left(1 \vee \frac{n \log k}{k}\right) \wedge \left(\log k \vee \frac{n}{k}\right)\right)\right\}$$
$$= \begin{cases} \Theta(1), & n \leq k/\log k, \\ \exp\left(-\Theta(n \log k/k)\right), & k/\log k \leq n \leq k, \\ \exp\left(-\Theta(\log k)\right), & k \leq n \leq k \log k, \\ \exp\left(-\Theta(n/k)\right), & n \geq k \log k, \end{cases}$$

where $\widehat{C}$ denotes an estimator using n samples with replacements and C is the number of distinct colors in a k-ball urn.

[1] Clearly, since $\widehat{C} - C \in \mathbb{Z}$, we shall assume without loss of generality that $\Delta \in \mathbb{N}$, with $\Delta = 1$ corresponding to the exact estimation of the number of distinct elements.

Δ	Lower bound	Upper bound	Estimator
≤ 1	$\Theta(k \log k)$		Naïve
$[1, \sqrt{k}(\log k)^{-\delta}]$	$\Theta(k \log k/\Delta^2)$		Interpolation (Section 2.4)
$[\sqrt{k}(\log k)^{-\delta}, k^{0.5+\delta}]$	$\Omega(k(1 \vee \log k/\Delta^2))$	$O(k \log \frac{\log k}{1 \vee \log \Delta^2/k})$	
$[k^{0.5+\delta}, ck]$	$\Theta(\frac{k}{\log k} \log k/\Delta)$		ℓ_2-approximation (Section 2.2)
$[ck, (0.5-\delta)k]$	$k \exp(-\sqrt{O(\log k \log\log k)})$ [37][2]	$O(\frac{k}{\log k})$	

Table 1. Summary of the sample complexity $n^*(k, \Delta)$, where δ is any sufficiently small constant, c is an absolute positive constant less than 0.5 (same over the table), and the notations $a \wedge b$ and $a \vee b$ stand for $\min\{a, b\}$ and $\max\{a, b\}$, respectively. The estimators are linear with coefficients obtained from either interpolation or ℓ_2-approximation.

1.2. Related work.

Statistics literature. The Distinct Elements problem is equivalent to estimating the number of species (or classes) in a finite population, which has been extensively studied in the statistics (see surveys [2, 21]) and the numismatics literature (see survey [12]). Motivated by various practical applications, a number of statistical models have been introduced for this problem, the most popular four being (cf. [2, Figure 1]):

The multinomial model. n samples are drawn uniformly at random with replacement;

The hypergeometric model. n samples are drawn uniformly at random without replacement;

The Bernoulli model. each individual is observed independently with some fixed probability, and thus the total number of samples is a binomial random variable;

The Poisson model. the number of observed samples in each class is independent and Poisson distributed, and thus the total sample size is also a Poisson random variable.

These models are closely related: conditioned on the sample size, the Bernoulli model coincides with the hypergeometric one, and Poisson model coincides with the multinomial one; furthermore, hypergeometric model can simulate multinomial

[2] A more precise result from [37] is the following: for $\Delta \in [ck, 0.5k - 2k^{3/4}\sqrt{\log k}]$, $n^*(k, \Delta) \geq k \exp\big(-\sqrt{O\big(\log k\big(\log\log k + \log(k/(k/2 - \Delta))\big)\big)}\big)$.

one and is hence more informative. The multinomial model is adopted as the main focus of this paper and the sample complexity in Definition 1 refers to the number of samples with replacement. In the undersampling regime where the sample size is significantly smaller than the population size, all four models are approximately equivalent. See Appendix A for a rigorous justification and detailed comparisons.

Under these models various estimators have been proposed such as unbiased estimators [19], Bayesian estimators [22], variants of Good–Turing estimators [7], etc. None of these methodologies, however, have a provable worst-case guarantee. Finally, we mention a closely related problem of estimating the number of connected components in a graph based on sampled induced subgraphs. In the special case where the underlying graph consists of disjoint cliques, the problem is exactly equivalent to the Distinct Elements problem [17].

Computer science literature. The interests in the Distinct Elements problem also arise in the database literature, where various intuitive estimators [24, 33] have been proposed under simplifying assumptions such as uniformity, and few performance guarantees are available. More recent work in [4, 5] obtained the optimal sample complexity under the *multiplicative* error criterion, where the minimum sample size to estimate the number of distinct elements within a factor of α is shown to be $\Theta(k/\alpha^2)$. For this task, it turns out the least favorable scenario is to distinguish an urn with unitary color from one with *almost* unitary color, the impossibility of which implies large multiplicative error. However, the optimal estimator performs poorly compared with others on an urn with many distinct colors [5], the case where most estimators enjoy small multiplicative error. In view of the limitation of multiplicative error, additive error is later considered by [37, 43]. To achieve an additive error of ck for a constant $c \in (0, \frac{1}{2})$, the result in [5] only implies an $\Omega(1/c)$ sample complexity lower bound, whereas a much stronger lower bound scales like $k^{1-O(\sqrt{\log\log k/\log k})}$ obtained in [37], which is almost linear. Determining the optimal sample complexity under additive error is the focus of the present paper.

The Distinct Elements problem can be viewed as a special case of the Support Size problem, where the goal is to estimate the cardinality of the support of an unknown discrete distribution, whose nonzero probabilities are at least $\frac{1}{k}$, based on independent samples. Improving previous results in [45], the optimal sample complexity has been recently determined in [47] to be

$$\Theta\left(\frac{k}{\log k}\log^2\frac{k}{\Delta}\right). \tag{1.1}$$

Samples drawn from a k-ball urn with replacement can be viewed as i.i.d. samples from a distribution supported on the set $\{\frac{1}{k}, \frac{2}{k}, \ldots, \frac{k}{k}\}$. From this perspective, any support size estimator, as well as its performance guarantee, is applicable to the Distinct Elements problem.

We briefly describe and compare the strategy to construct estimators in [47] and the current paper. Both are based on the idea of *polynomial approximation*, a powerful tool to circumvent the nonexistence of unbiased estimators [28]. The key is to approximate the function to be estimated by a polynomial, whose degree is chosen to balance the approximation error (bias) and the estimation error (variance). The worst-case performance guarantee for the Support Size problem in [47] is governed by the uniform approximation error over an interval where the probabilities may reside. In contrast, in the Distinct Elements problem, samples are generated from a distribution supported on a *discrete* set of values. Uniform approximation over a discrete subset leads to smaller approximation error and, in turn, improved sample complexity. It turns out that $O(\frac{k}{\log k}\log k/\Delta)$ samples are sufficient to achieve an additive error of Δ that satisfies $k^{0.5+O(1)} \le \Delta \le O(k)$, which strictly improves the sample complexity (1.1) for the Support Size problem, thanks to the discrete structure of the Distinct Elements problem.

The Distinct Elements problem considered here is not to be confused with the formulation in the streaming literature, where the goal is to approximate the number of distinct elements in the observations with low space complexity, see, e.g. [16, 27]. There, the proposed algorithms aim to optimize the memory consumption, but still require a full pass of every ball in the urn. This is different from the setting in the current paper, where only random samples drawn from the urn are available.

1.3. Organization. The paper is organized as follows: In Section 2 we describe a unified approach to construct estimators via discrete polynomial approximation, whose bias is analyzed in Section 2.2 and variance is upper bounded in Sections 2.3 and 2.4 separately. In Section 3 we obtain lower bounds on the sample complexity in Table 1 which establish the optimality of the proposed estimators. Section 4 explains how sample complexity bounds summarized in Table 1 follow from various results in Sections 2 and 3. Connections between the four sampling model mentioned in Section 1.2 are detailed in Appendix A. Proofs of auxiliary results are deferred to Appendix B and Appendix C.

1.4. Notations. All logarithms are with respect to the natural base. The transpose of a matrix A is denoted by $A^\top$. Let $\mathbf{1}$ denote the all-one column vector. Let $\|\cdot\|_p$ denote the vector ℓ_p-norm, for $1 \le p \le \infty$. Let $\mathrm{Poi}(\lambda)$ be the Poisson distribution with mean λ, $\mathrm{Bern}(p)$ be the Bernoulli distribution with mean p, $\mathrm{Binomial}(n, p)$ be the binomial distribution with n trials and success probability p, and $\mathrm{Hypergeometric}(N, K, n)$ be the hypergeometric distribution with probability mass function $\binom{K}{k}\binom{N-K}{n-k}/\binom{N}{n}$, for $0 \vee (n + K - N) \le k \le n \wedge K$. The n-fold product of a distribution P is denoted by $P^{\otimes n}$. We use standard big-O notations: for any positive sequence $\{a_n\}$ and $\{b_n\}$, $a_n = O(b_n)$ or $a_n \lesssim b_n$ if $a_n \le cb_n$ for some absolute constant $c > 0$, or equivalently, $\sup_n a_n/b_n < \infty$; $a_n = \Omega(b_n)$ or $a_n \gtrsim b_n$

if $b_n = O(a_n)$; $a_n = \Theta(b_n)$ or $a_n \asymp b_n$ if both $a_n = O(b_n)$ and $b_n = O(a_n)$; $a_n = o(b_n)$ if $\lim a_n/b_n = 0$; $a_n = \omega(b_n)$ if $b_n = o(a_n)$. Furthermore, the subscript in $o_n(1)$ indicates convergence in n that is uniform in all other parameters. We use the notations $a \wedge b$ and $a \vee b$ for $\min\{a, b\}$ and $\max\{a, b\}$, respectively. For $M \in \mathbb{N}$, let $[M] \triangleq \{1, \ldots, M\}$. For $\alpha \in \mathbb{R}$ and $S \subset \mathbb{R}$, let $\alpha S \triangleq \{\alpha x : x \in S\}$.

2. Linear estimators via discrete polynomial approximation

In this section we develop a unified framework to construct linear estimators and analyze its performance. Note that linear estimators (i.e. linear combinations of fingerprints) have been previously used for estimating distribution functionals [36, 45–47]. As commonly done in the literature, we assume the *Poisson sampling model*, where the sample size is a random variable $\mathrm{Poi}(n)$ instead of being exactly n. Under this model, the histograms of the samples, which count the number of balls in each color, are independent which simplifies the analysis. Any estimator under the Poisson sampling model can be easily modified for fixed sample size, and vice versa, thanks to the concentration of the Poisson random variable near its mean. Consequently, the sample complexities of these two models are close to each other, as shown in Corollary 1 in Appendix A.

2.1. Performance guarantees for general linear estimators. Recall that C denotes the number of distinct colors in a urn containing k colored balls. Let k_i denote the number of balls of the ith color in the urn. Then $\sum_i k_i = k$ and $C = \sum_i \mathbf{1}_{\{k_i > 0\}}$. Let $X_1, X_2, \ldots$ be independently drawn with replacement from the urn. Equivalently, the X_i's are i.i.d. according to a distribution $P = (p_i)_{i \geq 1}$, where $p_i = k_i/k$ is the fraction of balls of the ith color. The observed data are $X_1, \ldots, X_N$, where the sample size N is independent from $(X_i)_{i \geq 1}$ and is distributed as $\mathrm{Poi}(n)$. Under the Poisson model (or any of the sampling models described in Section 1.2), the *histograms* $\{N_i\}$ are sufficient statistics for inferring any aspect of the urn configuration; here N_i is the number of balls of the ith color observed in the sample, which is independently distributed as $\mathrm{Poi}(np_i)$. Furthermore, the *fingerprints* $\{\Phi_j\}_{j \geq 1}$, which are the histogram of the histograms, are also sufficient for estimating any permutation-invariant distributional property [35,43], in particular, the number of colors. Specifically, the jth fingerprint Φ_j denotes the number of colors that appear exactly j times. Note that $U \triangleq \Phi_0$, the number of unseen colors, is not observed.

The naïve estimator, "what you see is what you get," is simply the number of observed distinct colors, which can be expressed in terms of fingerprints as

$$\hat{C}_{\mathrm{seen}} = \sum_{j \geq 1} \Phi_j .$$

This is typically an underestimator because $C = \hat{C}_{\text{seen}} + U$. In turn, our estimator is

$$\tilde{C} = \hat{C}_{\text{seen}} + \hat{U}, \tag{2.1}$$

which adds a linear correction term

$$\hat{U} = \sum_{j \geq 1} u_j \Phi_j, \tag{2.2}$$

where the coefficients u_j's are to be specified. Since the fingerprints $\Phi_0, \Phi_1, \ldots$ are dependent (for example, they sum up to C), (2.2) serves as a linear predictor of $U = \Phi_0$ in terms of the observed fingerprints. Equivalently, in terms of histograms, the estimator has the following decomposable form:

$$\tilde{C} = \sum_{i=1}^{\infty} g(N_i), \tag{2.3}$$

where $g: \mathbb{Z}_+ \to \mathbb{R}$ satisfies $g(0) = 0$ and $g(j) = 1 + u_j$ for $j \geq 1$. In fact, any estimator that is linear in the fingerprints can be expressed of the decomposable form (2.3).

The main idea to choose the coefficients u_j is to achieve a good trade-off between the variance and the bias. In fact, it is instructive to point out that linear estimators can easily achieve exactly zero bias, which, however, comes at the price of high variance. To see this, note that the bias of the estimator (2.3) is

$$\mathbb{E}[\tilde{C}] - C = \sum_{i \geq 1} \big(\mathbb{E}[g(N_i)] - 1\big),$$

where

$$|\mathbb{E}[g(N_i) - 1]| = e^{-np_i} \left| -1 + \sum_{j=1}^{\infty} k_i^j \frac{u_j (n/k)^j}{j!} \right| \leq e^{-n/k} \max_{a \in [k]} |\phi(a) - 1|, \tag{2.4}$$

and

$$\phi(a) \triangleq \sum_{j \geq 1} a^j \frac{u_j (n/k)^j}{j!}$$

is a (formal) power series with $\phi(0) = 0$. The right-hand side of (2.4) can be made zero by choosing ϕ to be, e.g. the Lagrange interpolating polynomial that satisfies $\phi(0) = -1$ and $\phi(i) = 0$ for $i \in [k]$, namely,

$$\phi(a) = \frac{(-1)^{k+1}}{k!} \prod_{i=1}^{k} (a - i);$$

however, this strategy results in a high-degree polynomial ϕ with large coefficients, which, in turn, leads to a large variance of the estimator.

To reduce the variance of our estimator, we only use the first L fingerprints in (2.2) by setting $u_j = 0$ for all $j > L$, where L is chosen to be proportional to $\log k$. This restricts the polynomial degree in (2.4) to at most L and, while possibly incurring bias, reduces the variance. A further reason for only using the first few fingerprints is that higher-order fingerprints are *almost uncorrelated* with the number of unseens Φ_0. For instance, if red balls are observed for $n/2$ times, the only information this reveals is that approximately half of the urn are red. In fact, the correlation between Φ_0 and Φ_j decays exponentially (see Appendix B for a proof). Therefore for $L = \Theta(\log k)$, $\{\Phi_j\}_{j>L}$ offer little predictive power about Φ_0. Moreover, if a color is observed at most L times, say, $N_i \le L$, this implies that, with high probability, $k_i \le M$, where $M = O(kL/n)$, thanks to the concentration of Poisson random variables. Therefore, effectively we only need to consider those colors that appear in the urn for at most M times, i.e. $k_i \in [M]$, for which the bias is at most

$$\begin{aligned} |\mathbb{E}[g(N_i) - 1]| &\le e^{-n/k} \max_{a \in [M]} |\phi(a) - 1| \\ &= e^{-n/k} \max_{x \in [M]/M} |p(x) - 1| = e^{-n/k} \, \|Bw - \mathbf{1}\|_\infty , \end{aligned} \tag{2.5}$$

where $p(x) \triangleq \phi(Mx) = \sum_{j=1}^{L} w_j x^j$, $w = (w_1, \ldots, w_L)^\top$, and

$$w_j \triangleq \frac{u_j (Mn/k)^j}{j!}, \quad B \triangleq \begin{pmatrix} 1/M & (1/M)^2 & \cdots & (1/M)^L \\ 2/M & (2/M)^2 & \cdots & (2/M)^L \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix} \tag{2.6}$$

is a (partial) Vandermonde matrix. Lastly, since $\hat{C}_{\text{seen}} \le C \le k$, we define the final estimator to be $\widetilde{C}$ projected to the interval $[\hat{C}_{\text{seen}}, k]$. We have the following error bound:

Proposition 1. *Assume the Poisson sampling model. Let*

$$L = \alpha \log k, \quad M = \frac{\beta k \log k}{n}, \tag{2.7}$$

for any $\beta > \alpha$ such that L and M are integers. Let $w \in \mathbb{R}^L$. Let $\widetilde{C}$ be defined in (2.1) with $u_j = w_j \, j! (\frac{k}{nM})^j$ for $j \in [L]$ and $u_j = 0$ otherwise. Define $\hat{C} \triangleq (\widetilde{C} \vee \hat{C}_{\text{seen}}) \wedge k$. Then

$$\begin{aligned} \mathbb{E}(\hat{C} - C)^2 \le k^2 e^{-2n/k} \, \|Bw - \mathbf{1}\|_\infty^2 + k e^{-n/k} \\ + k \max_{m \in [M]} \mathbb{E}_{N \sim \mathrm{Poi}(nm/k)}[u_N^2] + k^{-(\beta - \alpha \log \frac{e\beta}{\alpha} - 3)}. \end{aligned} \tag{2.8}$$

Proof. Since $\widehat{C}_{\text{seen}} \le C \le k$, $\widehat{C}$ is always an improvement of $\widetilde{C}$. Define the event

$$E \triangleq \cap_{i=1}^{k} \{N_i \le L \Rightarrow kp_i \le M\},$$

which means that whenever $N_i \le L$ we have $p_i \le M/k$. Since $\beta > \alpha$, applying the Chernoff bound and the union bound yields $\mathbb{P}[E^c] \le k^{1-\beta+\alpha \log e\beta/\alpha}$, and thus

$$\begin{aligned} \mathbb{E}(\widehat{C} - C)^2 &\le \mathbb{E}\big((\widehat{C} - C)\mathbf{1}_E\big)^2 + k^2 \mathbb{P}[E^c] \\ &\le \mathbb{E}\big((\widetilde{C} - C)\mathbf{1}_E\big)^2 + k^{3-\beta+\alpha \log e\beta/\alpha}. \end{aligned} \tag{2.9}$$

The decomposable form of $\widetilde{C}$ in (2.3) leads to

$$(\widetilde{C} - C)\mathbf{1}_E = \sum_{i:k_i \in [M]} (g(N_i) - 1)\mathbf{1}_{\{N_i \le L\}} \triangleq \mathcal{E}.$$

In view of the bias analysis in (2.5), we have

$$|\mathbb{E}[E]| \le \sum_{i:k_i \in [M]} e^{-nk_i/k} \left\| Bw - \mathbf{1} \right\|_\infty \le k e^{-n/k} \left\| Bw - \mathbf{1} \right\|_\infty . \tag{2.10}$$

Recall that $g(0) = 0$ and $g(j) = u_j + 1$ for $j \in [L]$. Since N_i is independently distributed as $\text{Poi}(nk_i/k)$, we have

$$\begin{aligned} \mathsf{var}[E] &= \sum_{i:k_i \in [M]} \mathsf{var}\left[(g(N_i) - 1)\mathbf{1}_{\{N_i \le L\}}\right] \le \sum_{i:k_i \in [M]} \mathbb{E}\left[(g(N_i) - 1)^2 \mathbf{1}_{\{N_i \le L\}}\right] \\ &= \sum_{i:k_i \in [M]} \left(e^{-nk_i/k} + \mathbb{E}[u_{N_i}^2]\right) \le k e^{-n/k} + k \max_{m \in [M]} \mathbb{E}_{N \sim \text{Poi}(nm/k)}[u_N^2]. \end{aligned} \tag{2.11}$$

Combining the upper bound on the bias in (2.10) and the variance in (2.11) yields an upper bound on $\mathbb{E}[\mathcal{E}^2]$. Then the MSE in (2.8) follows from (2.9). □

Proposition 1 suggests that the coefficients of the linear estimator can be chosen by solving the following linear programming (LP):

$$\min_{w \in \mathbb{R}^L} \left\| Bw - \mathbf{1} \right\|_\infty \tag{2.12}$$

and showing that the solution does not have large entries. Instead of the ℓ_∞-approximation problem (2.12), whose optimal value is difficult to analyze, we solve the ℓ_2-approximation problem as a relaxation:

$$\min_{w \in \mathbb{R}^L} \left\| Bw - \mathbf{1} \right\|_2, \tag{2.13}$$

which is an upper bound of (2.12), and is in fact within an $O(\log k)$ factor since $M = O(k \log k/n)$ and $n = \Omega(k/\log k)$. In the remainder of this section, we consider two separate cases:

(i) $M > L$ ($n \lesssim k$): In this case, the linear system in (2.13) is overdetermined and the minimum is non-zero. Surprisingly, as shown in Section 2.2, the exact optimal value can be found in closed form using discrete orthogonal polynomials. The coefficients of the solution can be bounded using the minimum singular value of the matrix B, which is analyzed in Section 2.3.

(ii) $M \leq L$ ($n \gtrsim k$): In this case, the linear system is underdetermined and the minimum in (2.13) is zero. To bound the variance, it turns out that the coefficients bound obtained from the minimum singular value is not precise enough in this regime. Instead, we express the coefficients in terms of Lagrange interpolating polynomials and use Stirling numbers to obtain sharp variance bounds. This analysis in carried out in Section 2.4.

We finish this subsection with two remarks:

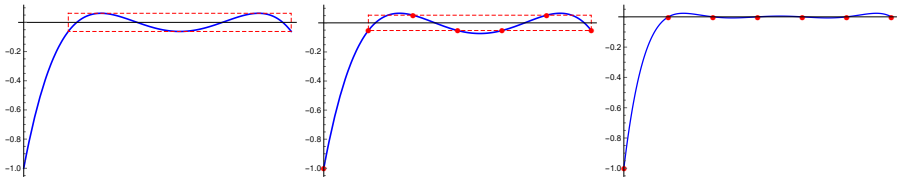
Remark 1 (Discrete versus continuous approximation). The optimal estimator for the Support Size problem in [47] has the same linear form as (2.1); however, since the probabilities can take any values in an interval, the coefficients are found to be the solution of the continuous polynomial approximation problem

$$\inf_{p} \max_{x \in [\frac{1}{M}, 1]} |p(x) - 1| = \exp\left(-\Theta\left(\frac{L}{\sqrt{M}}\right)\right). \tag{2.14}$$

where the infimum is taken over all degree-L polynomials such that $p(0) = 0$, achieved by the (appropriately shifted and scaled) Chebyshev polynomial [40]. In contrast, in Section 2.2 we show that the discrete version of (2.14), which is equivalent to the LP (2.12), satisfies

$$\inf_{p} \max_{x \in \{\frac{1}{M}, \frac{2}{M}, \ldots, 1\}} |p(x) - 1| = \mathsf{poly}(M) \exp\left(-\Theta\left(\frac{L^2}{M}\right)\right), \tag{2.15}$$

provided $L < M$. The difference between (2.14) and (2.15) explains why the sample complexity (1.1) for the Support Size problem has an extra log factor compared to that of the Distinct Elements problem in Table 1. When the sample size n is large enough, interpolation is used in lieu of approximation. See Fig. 1 for an illustration.



(a) Continuous approximation (b) Discrete approximation (c) Interpolation

Figure 1. Continuous and discrete polynomial approximations for $M = 6$ and degree $L = 4$, where (a) and (b) plot the optimal solution to (2.14) and (2.15) respectively. The interpolating polynomial in (c) requires a higher degree $L = 6$.

Remark 2 (Time complexity). The time complexity of the estimator (2.1) consists of:

(a) Computing histograms N_i and fingerprints Φ_j of n samples: $O(n)$;

(b) Computing the coefficients w by solving the least square problem in (2.5): $O(L^2(M+L))$;

(c) Evaluating the linear combination (2.1): $O(n \wedge k)$.

As shown in Table 1, for an accurate estimation the sample complexity is $n = \Omega(k/\log k)$, which implies $L = O(\log k)$ and $M = O(\log^2 k)$. Therefore, the overall time complexity is $O(n + \log^4 k) = O(n)$.

2.2. Exact solution to the ℓ_2-approximation. Next we give an explicit solution to the ℓ_2-approximation problem (2.13). In general, the optimal solution is given by $w^* = (B^\top B)^{-1} B^\top \mathbf{1}$ and the minimum value is the Euclidean distance between the all-one vector $\mathbf{1}$ and the column span of B, which, in the case of $M > L$, is non-zero (since B has linearly independent columns). Taking advantage of the Vandermonde structure of the matrix B in (2.6), we note that (2.13) can be interpreted as finding the orthogonal projection of the constant function onto the linear space of polynomials of degree between 1 and L defined on the discrete set $[M]/M$. Using the orthogonal polynomials with respect to the counting measure, known as *discrete Chebyshev (or Gram) polynomials* (see [38, Section 2.8] or [34, Section 2.4.2]), we show that, surprisingly, the optimal value of the ℓ_2-approximation can be found in closed form:

Lemma 1. *For all $L \geq 1$ and $M \geq L+1$,*

$$\min_{w \in \mathbb{R}^L} \|Bw - \mathbf{1}\|_2 = \left[\frac{\binom{M+L+1}{L+1}}{\binom{M}{L+1}} - 1 \right]^{-1/2} = \left[\exp\left(\Theta\left(\frac{L^2}{M} \right) \right) - 1 \right]^{-1/2}. \quad (2.16)$$

Proof. Define the following inner product between functions f and g:

$$\langle f, g \rangle \triangleq \sum_{i=1}^{M} f\left(\frac{i}{M}\right) g\left(\frac{i}{M}\right) \quad (2.17)$$

and the induced norm $\|f\| \triangleq \sqrt{\langle f, f \rangle}$. The least square problem (2.16) can be equivalently formulated as

$$\min_{w \in \mathbb{R}^L} \|-1 + w_1 x + w_2 x^2 + \cdots + w_L x^L\|. \quad (2.18)$$

This can be analyzed using the orthogonal polynomials under the inner product (2.17), which we describe next.

Recall the discrete Chebyshev polynomial [38, Sec. 2.8]: for $x = 0, 1, \ldots, M-1$,

$$t_m(x) \triangleq \frac{1}{m!} \Delta^m p_m(x) = \frac{1}{m!} \sum_{j=0}^{m} (-1)^j \binom{m}{j} p_m(x+m-j), \quad 0 \leq m \leq M-1, \quad (2.19)$$

where

$$p_m(x) \triangleq x(x-1)\cdots(x-m+1)(x-M)(x-M-1)\cdots(x-M-m+1), \quad (2.20)$$

and Δ^m denotes the mth order forward difference. The polynomials $\{t_0, \ldots, t_{M-1}\}$ are orthogonal with respect to the counting measure over the discrete set $\{0, 1, \ldots, M-1\}$; in particular, we have (cf. [38, Sec. 2.8.2, 2.8.3]):

$$\sum_{x=0}^{M-1} t_m(x)t_\ell(x) = 0, \quad m \neq \ell,$$
$$\sum_{x=0}^{M-1} t_m^2(x) = \frac{M(M^2-1^2)(M^2-2^2)\cdots(M^2-m^2)}{2m+1} \triangleq c(M,m).$$

By appropriately shifting and scaling the set of polynomials t_m, we define an orthonormal basis for the set of polynomials of degree at most $L \leq M-1$ under the inner product (2.17) by

$$\phi_m(x) = \frac{t_m(Mx-1)}{\sqrt{c(M,m)}}, \quad m = 0, \ldots, L. \quad (2.21)$$

Since $\{\phi_m\}_{m=0}^L$ constitute a basis for polynomials of degree at most L, the least square problem (2.18) can be equivalently formulated as

$$\min_{a:\sum_{i=1}^L a_i\phi_i(0)=-1} \left\| \sum_{i=0}^{L} a_i\phi_i \right\| = \min_{a:\langle a,\phi(0)\rangle=-1} \|a\|_2,$$

where $\phi(0) \triangleq (\phi_0(0), \ldots, \phi_L(0))$, $a = (a_0, \ldots, a_L)$, and $\langle\cdot,\cdot\rangle$ denotes vector inner product. Thus, the optimal value is clearly $1/\|\phi(0)\|_2$, achieved by $a^* = -\phi(0)/\|\phi(0)\|_2^2$.

From (2.20) we have

$$p_m(0) = p_m(1) = \cdots = p_m(m-1) = 0.$$

By the formula of t_m in (2.19), we obtain

$$t_m(-1) = \frac{1}{m!}(-1)^m p_m(-1) = (-1)^m \prod_{j=1}^{m}(M+j).$$

In view of the definition of ϕ_m in (2.21), we have

$$\phi_m(0) = \frac{t_m(-1)}{\sqrt{c(M,m)}} = \frac{(-1)^m \prod_{j=1}^m (M+j)}{\sqrt{\frac{M\prod_{j=1}^m(M^2-j^2)}{2m+1}}} = (-1)^m \sqrt{\frac{2m+1}{M}\prod_{j=1}^{m}\frac{M+j}{M-j}}.$$

Therefore

$$\|\phi(0)\|_2^2 = \sum_{m=0}^{L} \frac{2m+1}{M} \prod_{j=1}^{m} \frac{M+j}{M-j} = \frac{\binom{M+L+1}{L+1}}{\binom{M}{L+1}} - 1,$$

where the last equality follows from induction since

$$\frac{\binom{M+L+1}{L+1}}{\binom{M}{L+1}} - \frac{\binom{M+L}{L}}{\binom{M}{L}} = \frac{2L+1}{M} \prod_{j=1}^{L} \frac{M+j}{M-j}.$$

This proves the first equality in (2.16).

The second equality in (2.16) is a direct consequence of Stirling's approximation. If $M = L+1$, then

$$\frac{\binom{M+L+1}{L+1}}{\binom{M}{L+1}} = \binom{2(L+1)}{L+1} = \exp\big(\Theta(L)\big). \tag{2.22}$$

If $M \geq L+2$, denoting $x = (L+1)/M$ and applying $n! = \sqrt{2\pi n}(n/e)^n(1+\Theta(1/n))$ when $n \geq 1$, we have

$$\begin{aligned}
\frac{\binom{M+L+1}{L+1}}{\binom{M}{L+1}} &= \frac{(M+L+1)!(M-L-1)!}{(M!)^2} = \frac{(M(1+x))!(M(1-x))!}{(M!)^2} \\
&= \frac{\left(\begin{array}{r}\sqrt{2\pi M(1+x)}\big(\frac{M(1+x)}{e}\big)^{M(1+x)}\sqrt{2\pi M(1-x)}\big(\frac{M(1-x)}{e}\big)^{M(1-x)} \\ \big(1+\Theta\big(\frac{1}{M(1+x)}+\frac{1}{M(1-x)}\big)\big)\end{array}\right)}{2\pi M\big(\frac{M}{e}\big)^{2M}\big(1+\Theta\big(\frac{1}{M}\big)\big)} \\
&= \sqrt{1-x^2}\exp\big(M\big((1+x)\log(1+x)+(1-x)\log(1-x)\big)\big)\frac{1+\Theta\big(\frac{1}{M(1-x^2)}\big)}{1+\Theta\big(\frac{1}{M}\big)} \\
&= \exp\left(\Theta(Mx^2)+\frac{1}{2}\log(1-x^2)+\log\frac{1+\Theta\big(\frac{1}{M(1-x^2)}\big)}{1+\Theta\big(\frac{1}{M}\big)}\right),
\end{aligned} \tag{2.23}$$

where the last step follows from $(1+x)\log(1+x)+(1-x)\log(1-x) = \Theta(x^2)$ when $0 \leq x \leq 1$. In the exponent of (2.23), the term $\Theta(Mx^2)$ dominates when $M \geq L+2$. Applying (2.22) and (2.23) to the exact solution (2.16) yields the desired approximation. □

2.3. Minimum singular values of real rectangle Vandermonde matrices. In Proposition 1 the variance of our estimator is bounded by the magnitude of coefficients u, which is related to the polynomial coefficients w by (2.6). A classical result from approximation theory is that if a polynomial is bounded over a compact interval,

its coefficients are at most exponential in the degree [40, Theorem 2.9.11]: for any degree-L polynomial $p(x) = \sum_{i=0}^{L} w_i x^i$,

$$\max_{0\le i\le L} |w_i| \le \max_{x\in[0,1]} |p(x)| \exp\big(O(L)\big), \tag{2.24}$$

which is tight when p is the Chebyshev polynomial. This fact has been applied in statistical contexts to control the variance of estimators obtained from best polynomial approximation [9, 26, 47, 48]. In contrast, for the Distinct Elements problem, the polynomial is only known to be bounded over the discretized interval. Nevertheless, we show that the bound (2.24) continues to hold as long as the discretization level exceeds the degree:

$$\max_{0\le i\le L} |w_i| \le \max_{x\in\{\frac{1}{M},\frac{2}{M},\ldots,1\}} |p(x)| \exp\big(O(L)\big), \tag{2.25}$$

provided that $M \ge L+1$ (see Remark 3 after Lemma 2). Clearly, (2.25) implies (2.24) by sending $M \to \infty$. If $M \le L$, a coefficient bound like (2.25) is impossible, because one can add to p an arbitrary degree-L interpolating polynomial that evaluates to zero at all M points.

To bound the coefficients, note that the optimal solution of ℓ_2-approximation is $w^* = (B^\top B)^{-1} B^\top \mathbf{1}$, and consequently

$$\|w^*\|_2 \le \frac{\|\mathbf{1}\|_2}{\sigma_{\min}(B)}, \tag{2.26}$$

where $\sigma_{\min}(B)$ denotes the smallest singular value of B. Let

$$\bar{B} \triangleq [\mathbf{1}, B] = \begin{pmatrix} 1 & 1/M & (1/M)^2 & \cdots & (1/M)^L \\ 1 & 2/M & (2/M)^2 & \cdots & (2/M)^L \\ 1 & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & 1 \end{pmatrix}$$

which is an $M \times (L+1)$ Vandermonde matrix and satisfies $\sigma_{\min}(\bar{B}) \le \sigma_{\min}(B)$ since $\bar{B}$ has one extra column. The Gram matrix of $\bar{B}$ is an instance of *moment matrices*. A moment matrix associated with a probability measure μ is a Hankel matrix M given by $M_{i,j} = m_{i+j-2}$, where $m_\ell = \int x^\ell \mathrm{d}\mu$ denotes the ℓth moment of μ. Then $\frac{1}{M}\bar{B}^\top\bar{B}$ is the moment matrix associated with the uniform distribution over the discrete set $\{\frac{1}{M}, \frac{2}{M}, \ldots, 1\}$, which converges to the uniform distribution over the interval $(0, 1)$. The moment matrix of the uniform distribution is the famous *Hilbert matrix* H, with

$$H_{ij} = \frac{1}{i+j-1},$$

which is a well-studied example of ill-conditioned matrices in the numerical analysis literature. In particular, it is known that the condition number of the $L \times L$ Hilbert

matrix is $O((1+\sqrt{2})^{4L}/\sqrt{L})$ [41] and the operator norm is $\Theta(1)$, and thus the minimum singular value is exponentially small in the degree. Therefore we expect the discrete moment matrix $\frac{1}{M}\bar{B}^\top\bar{B}$ to behave similarly to the Hilbert matrix when M is large enough. Interestingly, we show that this is indeed the case as soon as M exceeds L (otherwise the minimum singular value is zero).

Lemma 2. *For all $M \geq L+1$,*

$$\sigma_{\min}\left(\frac{\bar{B}}{\sqrt{M}}\right) \geq \frac{1}{L^2 2^{7L}(2L+1)}\left(\frac{M+L}{eM}\right)^{L+0.5}. \tag{2.27}$$

Remark 3. The inequality (2.25) follows from Lemma 2 since the coefficient vector $w = (w_0, \ldots, w_L)$ satisfies

$$\|w\|_\infty \leq \|w\|_2 \leq \frac{1}{\sigma_{\min}(\bar{B})}\|\bar{B}w\|_2 \leq \frac{\sqrt{M}}{\sigma_{\min}(\bar{B})}\|\bar{B}w\|_\infty.$$

Remark 4. The extreme singular values of square Vandermonde matrices have been extensively studied (c.f. [1, 18] and the references therein). For rectangular Vandermonde matrices, the focus was mainly with nodes on the unit circle in the complex domain [6, 15, 30] with applications in signal processing. In contrast, Lemma 2 is on rectangular Vandermonde matrices with real nodes. The result on integers nodes in [11] turns out to be too crude for the purpose of this paper.

Proof. Note that $\bar{B}^\top\bar{B}$ is the Gramian of monomials $\mathbf{x} = (1, x, x^2, \ldots, x^L)^\top$ under the inner product defined in (2.17). When $M \geq L+1$, the orthonormal basis $\phi = (\phi_0, \ldots, \phi_L)^\top$ under the inner product (2.17) are given in (2.21). Let $\phi = \mathbf{L}\mathbf{x}$ where $\mathbf{L} \in \mathbb{R}^{(L+1)\times(L+1)}$ is a lower triangular matrix and $\mathbf{L}$ consists of the coefficients of ϕ. Taking the Gramian of ϕ yields that $I = \mathbf{L}(\bar{B}^\top\bar{B})\mathbf{L}^\top$, i.e. $\mathbf{L}^{-1}$ can be obtained from the Cholesky decomposition: $\bar{B}^\top\bar{B} = (\mathbf{L}^{-1})(\mathbf{L}^{-1})^\top$. Then[3]

$$\sigma^2_{\min}(\bar{B}) = \frac{1}{\|\mathbf{L}\|^2_{\mathrm{op}}} \geq \frac{1}{\|\mathbf{L}\|^2_F}, \tag{2.28}$$

where $\|\cdot\|_{\mathrm{op}}$ denotes the ℓ_2 operator norm, which is the largest singular value of L, and $\|\cdot\|_F$ denotes the Frobenius norm. By definition, $\|\mathbf{L}\|^2_F$ is the sum of all squared coefficients of $\phi_0, \ldots, \phi_L$. A useful method to bound the sum-of-squares of the coefficients of a polynomial is by its maximal modulus over the unit circle on the complex plane. Specifically, for any polynomial $p(z) = \sum_{i=0}^n a_i z^i$, we have

$$\sum_{i=0}^n |a_i|^2 = \frac{1}{2\pi}\oint_{|z|=1} |p(z)|^2 \mathrm{d}z \leq \sup_{|z|=1} |p(z)|^2. \tag{2.29}$$

[3]The lower bound (2.28), which was also obtained in [8, (1.13)] using Cauchy–Schwarz inequality, is tight up to polynomial terms in view of the fact that $\|\mathbf{L}\|_F \leq (L+1)\|\mathbf{L}\|_{\mathrm{op}}$.

Therefore

$$\begin{aligned}\sigma_{\min}(\bar{B}) \geq \frac{1}{\|\mathbf{L}\|_F} &\geq \frac{1}{\sqrt{\sum_{m=0}^{L} \sup_{|z|=1} |\phi_m(z)|^2}} \\ &\geq \frac{1}{\sqrt{L+1}} \frac{1}{\sup_{0\leq m\leq L, |z|=1} |\phi_m(z)|}.\end{aligned} \tag{2.30}$$

For a given M, the orthonormal basis $\phi_m(x)$ in (2.21) is proportional to the discrete Chebyshev polynomials $t_m(Mx-1)$. The classical asymptotic result for the discrete Chebyshev polynomials shows that [38, (2.8.6)]

$$\lim_{M\to\infty} M^{-m} t_m(Mx) = P_m(2x-1),$$

where P_m is the Legendre polynomial of degree m. This gives the intuition that $t_m(x) \approx M^m$ for real-valued $x \in [0, M]$. We have the following non-asymptotic upper bound (proved in Appendix C) for t_m over the complex plane:

Lemma 3. *For all* $0 \leq m \leq M-1$,

$$|t_m(z)| \leq m^2 2^{6m} \sup_{0\leq\xi\leq m} \big(|z+\xi| \vee M\big)^m. \tag{2.31}$$

Applying (2.31) on the definition of ϕ_m in (2.21), for any $|z| = 1$ and any $M \geq L+1$, we have

$$|\phi_m(z)| = \frac{|t_m(Mz-1)|}{\sqrt{c(M,m)}} \leq \frac{m^2 2^{7m} M^m}{\sqrt{\frac{M(M^2-1^2)(M^2-2^2)\cdots(M^2-m^2)}{2m+1}}}.$$

The right-hand side is increasing with m. Therefore,

$$\begin{aligned}\sup_{0\leq m\leq L, |z|=1} |\phi_m(z)| &\leq \frac{L^2 2^{7L} M^L}{\sqrt{\frac{M(M^2-1^2)(M^2-2^2)\cdots(M^2-L^2)}{2L+1}}} \\ &= \frac{1}{\sqrt{M}} L^2 2^{7L} \sqrt{2L+1} \sqrt{\frac{M^{2L+1}}{\binom{M+L}{2L+1}(2L+1)!}}.\end{aligned}$$

Combining (2.30), we obtain

$$\begin{aligned}\sigma_{\min}\left(\frac{\bar{B}}{\sqrt{M}}\right) &\geq \frac{1}{L^2 2^{7L}\sqrt{(L+1)(2L+1)}} \sqrt{\frac{\binom{M+L}{2L+1}(2L+1)!}{M^{2L+1}}} \\ &\geq \frac{1}{L^2 2^{7L}(2L+1)} \left(\frac{M+L}{eM}\right)^{L+0.5},\end{aligned}$$

where in the last inequality we used $\binom{n}{k} \geq (n/k)^k$ and $n! \geq (n/e)^n$. □

Using the optimal solution w^* to the ℓ_2-approximation problem (2.13) as the coefficient of the linear estimator $\widehat{C}$, the following performance guarantee is obtained by applying Lemma 1 and Lemma 2 to bound the bias and variance, respectively:

Theorem 1. *Assume the Poisson sampling model. Then,*

$$\mathbb{E}(\widehat{C} - C)^2 \le k^2 \exp\left(-\Theta\left(1 \vee \frac{n \log k}{k} \wedge \log k\right)\right). \tag{2.32}$$

Proof. If $n \le \frac{k}{\log k}$, then the upper bound in (2.32) is $\Theta(k^2)$, which is trivial thanks to the thresholds that $\widehat{C} = (\widetilde{C} \vee \widehat{C}_{\text{seen}}) \wedge k$. It is hereinafter assumed that $n \ge \frac{k}{\log k}$, or equivalently $M \le \frac{\beta}{\alpha^2} L^2$; here M, L are defined in (2.7) and the constants α, β are to be determined later. Then, from Lemma 1,

$$\|Bw^* - \mathbf{1}\|_\infty \le \|Bw^* - \mathbf{1}\|_2 \le \exp\left(-\Theta\left(\frac{L^2}{M}\right)\right). \tag{2.33}$$

In view of (2.26) and Lemma 2, we have

$$\|w^*\|_\infty \le \|w^*\|_2 \le \frac{\|\mathbf{1}\|_2}{\sigma_{\min}(B)} \le \exp\big(O(L)\big).$$

Recall the connection between u_j and w_j in (2.6). For $1 \le j \le L < \beta \log k$, we have

$$u_j = w_j \frac{j!}{(\beta \log k)^j} \le \frac{w_j}{\beta \log k}.$$

Therefore,

$$\|u^*\|_\infty \le \frac{\|w^*\|_\infty}{\beta \log k} \le \frac{\exp\big(O(L)\big)}{\beta \log k}. \tag{2.34}$$

Applying (2.33) and (2.34) to Proposition 1, we obtain

$$\mathbb{E}(\widehat{C} - C)^2 \le k^2 \exp\left(-\frac{2n}{k} - \Theta\left(\frac{n \log k}{k}\right)\right) + k e^{-n/k} + k \frac{\exp\big(O(\log k)\big)}{(\beta \log k)^2} + k^{-(\beta - \alpha \log \frac{e\beta}{\alpha} - 3)}.$$

Then the desired (2.32) holds as long as β is sufficiently large and α is sufficiently small. □

2.4. Lagrange interpolating polynomials and Stirling numbers. When we sample at least a constant faction of the urn, i.e. $n = \Omega(k)$, we can afford to choose α and β in (2.7) so that $L = M$ and B is an invertible matrix. We choose the coefficient $w = B^{-1}\mathbf{1}$ which is equivalent to applying *Lagrange interpolating polynomial* and achieves exact zero bias. To control the variance, we can follow the approach in

Section 2.3 by using the bound on minimum singular value of the matrix B, which implies that the coefficients are $\exp(O(L))$ and yields a coarse upper bound

$$O\left(k \frac{\log k}{1 \vee \log \Delta^2/k}\right)$$

on the sample complexity. As previously announced in Table 1, this bound can be improved to

$$O\left(k \log \frac{\log k}{1 \vee \log \Delta^2/k}\right)$$

by a more careful analysis of the Lagrange interpolating polynomial coefficients expressed in terms of the Stirling numbers, which we introduce next.

The Stirling numbers of the first kind are defined as the coefficients of the falling factorial $(x)_n$ where

$$(x)_n = x(x-1)\dots(x-n+1) = \sum_{j=1}^{n} s(n,j)x^j.$$

Compared to the coefficients w expressed by the Lagrange interpolating polynomial:

$$\sum_{j=1}^{M} w_j x^j - 1 = -\frac{(1-xM)(2-xM)\dots(M-xM)}{M!},$$

we obtain a formula for the coefficients w in terms of the Stirling numbers:

$$w_j = \frac{(-1)^{M+1} M^j}{M!} s(M+1, j+1), \quad 1 \le j \le M.$$

Consequently, the coefficients of our estimator u_j are given by

$$u_j = (-1)^{M+1} \frac{j!}{M!} \left(\frac{k}{n}\right)^j s(M+1, j+1). \tag{2.35}$$

The precise asymptotics the Stirling number is rather complicated. In particular, the asymptotic formula of $s(n,m)$ as $n \to \infty$ for fixed m is given by [25] and the uniform asymptotics over all m is obtained in [32] and [39]. The following lemma (proved in Appendix C) is a coarse non-asymptotic version, which suffices for the purpose of constant-factor approximations of the sample complexity.

Lemma 4.

$$|s(n+1, m+1)| = n! \left(\Theta\left(\frac{1}{m}\left(1 \vee \log \frac{n}{m}\right)\right)\right)^m \tag{2.36}$$

We construct $\widehat{C}$ as in Proposition 1 using the coefficients u_j in (2.35) to achieve zero bias. The variance upper bound by the coefficients u is a direct consequence of the upper bound of Stirling numbers in Lemma 4. Then we obtain the following mean squared error (MSE):

Theorem 2 (Interpolation). *Assume the Poisson sampling model. If $n > \eta k$ for some sufficiently large constant η, then*

$$\mathbb{E}(\widehat{C} - C)^2 \le k e^{-\Theta(n/k)} + k^{-0.5-3.5\frac{k}{n}\log(k/en)} + \begin{cases} k\exp\left(\frac{k^2\log k}{n^2}e^{-\Theta(n/k)}\right), & n \lesssim k\log\log k, \\ k\left(\Theta(k/n)\log\frac{k^2\log k}{n^2}\right)^{2n/k}, & k\log\log k \lesssim n \lesssim k\sqrt{\log k}, \\ 0, & n \gtrsim k\sqrt{\log k}. \end{cases}$$

Proof. In Proposition 1, fix $\beta = 3.5$ and $\alpha = \beta k/n$ so that $L = M$. Our goal is to show an upper bound of

$$\max_{\lambda \in n/k[M]} \mathbb{E}_{N\sim\mathrm{Poi}(\lambda)}[u_N^2] = \max_{\lambda\in n/k[M]} \sum_{j=1}^{M} u_j^2 e^{-\lambda}\frac{\lambda^j}{j!}. \tag{2.37}$$

Here the coefficients u_j are obtained from (2.35) and, in view of (2.36), satisfy:

$$|u_j| \le \left(\frac{\eta k}{n}\left(1 \vee \log\frac{M}{j}\right)\right)^j, \quad 1 \le j \le M, \tag{2.38}$$

for some universal constant η. We consider three cases separately:

Case I: $n \ge \sqrt{\beta}k\sqrt{\log k}$. In this case we have $n/k \ge M$. The maximum of each summand in (2.37) as a function of $\lambda \in \mathbb{R}$ occurs at $\lambda = j$. Since $j \le n/k$, the maximum over $\lambda \in n/k[M]$ is attained at $\lambda = n/k$. Then,

$$\max_{\lambda\in n/k[M]} \mathbb{E}_{N\sim\mathrm{Poi}(\lambda)}[u_N^2] = \mathbb{E}_{N\sim\mathrm{Poi}(n/k)}[u_N^2]. \tag{2.39}$$

In view of (2.38) and $j \ge 1$, we have $|u_j| \le (\Theta(k/n)\log M)^j$. Then,

$$\begin{aligned}\mathbb{E}_{N\sim\mathrm{Poi}(n/k)}[u_N^2] &\le \mathbb{E}_{N\sim\mathrm{Poi}(n/k)}\left(\Theta\left(\frac{k\log M}{n}\right)^2\right)^N \\ &= \exp\left(\frac{n}{k}\left(\Theta\left(\frac{k\log M}{n}\right)^2 - 1\right)\right) = e^{-\Theta(n/k)},\end{aligned}$$

as long as $n \gtrsim k\log\log k$ and thus $(k\log M)/n \lesssim 1$. Therefore,

$$\max_{\lambda\in n/k[M]} \mathbb{E}_{N\sim\mathrm{Poi}(\lambda)}[u_N^2] \le e^{-\Theta(n/k)}, \quad n \gtrsim k\sqrt{\log k}. \tag{2.40}$$

Case II: $\eta k \log\log k \le n \le \sqrt{\beta} k \sqrt{\log k}$. We apply the following upper bound:

$$\begin{aligned}
&\max_{\lambda \in n/k[M]} \mathbb{E}_{N\sim\mathrm{Poi}(\lambda)}[u_N^2] \\
&\quad = \max_{\lambda \in n/k[M]} \mathbb{E}_{N\sim\mathrm{Poi}(\lambda)}[u_N^2 \mathbf{1}_{\{N \ge n/k\}}] + \max_{\lambda \in n/k[M]} \mathbb{E}_{N\sim\mathrm{Poi}(\lambda)}[u_N^2 \mathbf{1}_{\{N < n/k\}}] \\
&\quad \le \max_{n/k \le j \le M} |u_j|^2 + e^{-\Theta(n/k)}.
\end{aligned} \tag{2.41}$$

where the upper bound of the second addend is analogous to (2.39) and (2.40). Since $\eta k/n \le 1$, the right-hand side of (2.38) is decreasing with j when $j \ge M/e$. It suffices to consider $j \le M/e$, when the maximum as a function of $j \in \mathbb{R}$ occurs at $j^* \le Me^{-n/\eta k}$. Since $Me^{-n/\eta k} \le n/k$ when $n \ge \eta k \log\log k$, the maximum over $n/k \le j \le M$ is attained at $j = n/k$. Applying (2.38) with $j = n/k$ to (2.41) yields

$$\max_{\lambda \in n/k[M]} \mathbb{E}_{N\sim\mathrm{Poi}(\lambda)}[u_N^2] \le \left(\Theta\left(\frac{k}{n}\right) \log \frac{k^2 \log k}{n^2}\right)^{2n/k} + e^{-\Theta(n/k)}. \tag{2.42}$$

Case III: $\eta k \le n \le \eta k \log\log k$. We apply the upper bound of expectation by the maximum:

$$\max_{\lambda \in n/k[M]} \mathbb{E}_{N\sim\mathrm{Poi}(\lambda)}[u_N^2] \le \max_{j \in [M]} u_j^2.$$

Since $\eta k/n \le 1$, the right-hand side of (2.38) is decreasing with j when $j \ge M/e$, so it suffices to consider $j \le M/e$. Denoting $x = \log M/j$ and $\tau = \Theta(k/n)$, in view of (2.38), we have $|u_j| \le \exp(Me^{-x} \log(\tau x))$, which attains maximum at x^* satisfying $e^{1/x^*}/x^* = \tau$. Then,

$$|u_j| \le \exp\left(Me^{-x^*} \log(\tau x^*)\right) = \exp\left(Me^{-x^*}/x^*\right) < \exp\left(M\tau e^{-1/\tau}\right).$$

where the last inequality is because of $\tau > 1/x^*$. Therefore,

$$\max_{\lambda \in n/k[M]} \mathbb{E}_{N\sim\mathrm{Poi}(\lambda)}[u_N^2] \le \exp\left(\frac{k^2 \log k}{n^2} e^{-\Theta(n/k)}\right), \quad k \lesssim n \lesssim k \log\log k. \tag{2.43}$$

Applying the upper bounds in (2.40), (2.42) and (2.43) to Proposition 1 concludes the proof. □

Remark 5. It is impossible to bridge the gap near $\Delta = \sqrt{k}$ in Table 1 using the technology of interpolating polynomials that aims at zero bias, since its worst-case variance is at least $k^{1+\Omega(1)}$ when $n = O(k)$. To see this, note that the variance term given by (2.11) is

$$\sum_{p_i} \mathbb{E}_{N\sim\mathrm{Poi}(np_i)}[u_N^2] = \sum_{p_i} \sum_{j=1}^{L} u_j^2 e^{-np_i} \frac{(np_i)^j}{j!}. \tag{2.44}$$

Consider the distribution Uniform$[n/j_0]$ with $j_0 = Le^{-2n/k} = \Omega(\log k)$, which corresponds to an urn where each of the n/j_0 colors appears equal number of times. By the formula of coefficient u_j in (2.35) and the characterization from Lemma 4, the $j = j_0$ term in the summation of (2.44) is of order $\frac{n}{j_0}(\frac{k}{n}\log M/j_0)^{2j_0} = \frac{n}{j_0}2^{2j_0}$, which is already $k^{1+\Omega(1)}$.

3. Optimality of the sample complexity

In this section we develop lower bounds of the sample complexity which certify the optimality of estimators constructed in Section 2. We first give a brief overview of the lower bound in [5, Theorem 1], which gives the optimal sample complexity under the multiplicative error criterion. The lower bound argument boils down to considering two hypothesis: in the null hypothesis, the urn consists of only one color; in the alternative, the urn contains $2\Delta + 1$ distinct colors, where $k - 2\Delta$ balls share the same color as in the null hypothesis, and all other balls have distinct colors. These two scenarios are distinguished if and only if a second color appears in the samples, which typically requires $\Omega(k/\Delta)$ samples. This lower bound is optimal for estimating within a multiplicative factor of $\sqrt{\Delta}$, which, however, is too loose for additive error Δ.

In contrast, instead of testing whether the urn is monochromatic, our first lower bound is given by testing whether the urn is maximally colorful, that is, containing k distinct colors. The alternative contains $k - 2\Delta$ colors, and the numbers of balls of two different colors differ by at most one. In other words, the null hypothesis is the uniform distribution on $[k]$ and the alternative is close to uniform distribution with smaller support size. The sample complexity, which is shown in Theorem 3, gives the lower bound in Table 1 for $\Delta \le \sqrt{k}$.

Theorem 3. *If* $1 \le \Delta \le \frac{k}{2}$, *then*

$$n^*(k, \Delta) \ge \Omega\left(\frac{k - 2\Delta}{\sqrt{k}}\right). \tag{3.1}$$

If $1 \le \Delta < \frac{k}{4}$, *then*

$$n^*(k, \Delta) \ge \Omega\left(k \operatorname{arccosh}\left(1 + \frac{k}{4\Delta^2}\right)\right) \asymp \begin{cases} k\log(1 + \frac{k}{\Delta^2}), & \Delta \le \sqrt{k}, \\ \frac{k^{3/2}}{\Delta}, & \Delta \ge \sqrt{k}. \end{cases} \tag{3.2}$$

Proof. Consider the following two hypotheses: The null hypothesis H_0 is an urn consisting of k distinct colors; The alternative H_1 consists of $k - 2\Delta$ distinct colors, and each color appears either $b_1 \triangleq \lfloor \frac{k}{k-2\Delta} \rfloor$ or $b_2 \triangleq \lceil \frac{k}{k-2\Delta} \rceil$ times. In terms of distributions, H_0 is the uniform distribution $Q = (\frac{1}{k}, \ldots, \frac{1}{k})$; H_1 is the closest

perturbation from the uniform distribution: randomly pick disjoint sets of indices $I, J \subseteq [k]$ with cardinality $|I| = c_1$ and $|J| = c_2$, where c_1 and c_2 satisfy

$$\begin{aligned} &\text{(number of colors)} \quad c_1 + c_2 = k - 2\Delta, \\ &\text{(number of balls)} \quad c_1 b_1 + c_2 b_2 = k. \end{aligned}$$

Conditional on $\theta \triangleq (I, J)$, the distribution $P_\theta = (p_{\theta,1}, \ldots, p_{\theta,k})$ is given by

$$p_\theta = \begin{cases} b_1/k, & i \in I, \\ b_2/k, & i \in J. \end{cases}$$

Put the uniform prior on the alternative. Denote the marginal distributions of the n samples $X = (X_1, \ldots, X_n)$ under H_0 and H_1 by Q_X and P_X, respectively. Since the distinct colors in H_0 and H_1 are separated by 2Δ, to show that the sample complexity $n^*(k, \Delta) \geq n$, it suffices to show that no test can distinguish H_0 and H_1 reliably using n samples. A further sufficient condition is a bounded χ^2 divergence [42]

$$\chi^2(P_X \| Q_X) \triangleq \int \frac{P_X^2}{Q_X} - 1 \leq O(1).$$

The remainder of this proof is devoted to upper bounds of the χ^2 divergence.

Since $P_{X|\theta} = P_\theta^{\otimes n}$ and $Q_X = Q^{\otimes n}$, we have

$$\begin{aligned} \chi^2(P_X \| Q_X) + 1 = \int \frac{P_X^2}{Q_X} &= \int \frac{(\mathbb{E}_\theta P_{X|\theta})(\mathbb{E}_{\theta'} P_{X|\theta'})}{Q_X} \\ &= \mathbb{E}_{\theta,\theta'} \int \frac{P_{X|\theta} P_{X|\theta'}}{Q_X} = \mathbb{E}_{\theta,\theta'} \left(\int \frac{P_\theta P_{\theta'}}{Q} \right)^n, \end{aligned}$$

where θ' is an independent copy of θ. By the definition of P_θ and Q,

$$\begin{aligned} \int \frac{P_\theta P_{\theta'}}{Q} &= \frac{b_1^2}{k} |I \cap I'| + \frac{b_2^2}{k} |J \cap J'| + \frac{b_1 b_2}{k} \big(|I \cap J'| + |J \cap I'| \big) \\ &= 1 + \sum_{i=1}^4 A_i, \end{aligned} \tag{3.3}$$

where

$$A_1 \triangleq \frac{b_1^2}{k} \Big(|I \cap I'| - \frac{c_1^2}{k} \Big), \quad A_2 \triangleq \frac{b_2^2}{k} \Big(|J \cap J'| - \frac{c_2^2}{k} \Big),$$
$$A_3 = \frac{b_1 b_2}{k} \Big(|I \cap J'| - \frac{c_1 c_2}{k} \Big), \quad \text{and} \quad A_4 = \frac{b_1 b_2}{k} \Big(|J \cap I'| - \frac{c_1 c_2}{k} \Big)$$

are centered random variables. Applying $1 + x \leq e^x$ and Cauchy–Schwarz inequality, we obtain

$$\chi^2(P_X \| Q_X) + 1 \leq \mathbb{E}[e^{n \sum_{i=1}^4 A_i}] \leq \prod_{i=1}^4 \big(\mathbb{E}[e^{4n A_i}] \big)^{1/4}. \tag{3.4}$$

Consider the first term $\mathbb{E}[e^{4nA_1}]$. Note that $|I \cap I'| \sim \text{Hypergeometric}(k, c_1, c_1)$, which is the distribution of the sum of c_1 samples drawn without replacement from a population of size k which consists of c_1 ones and $k - c_1$ zeros. By the convex stochastic dominance of the binomial over the hypergeometric distribution [23, Theorem 4], for $Y \sim \text{Binomial}(c_1, c_1/k)$, we have

$$\begin{aligned}(\mathbb{E}[e^{4nA_1}])^{1/4} &\le \left(\mathbb{E}\left[\exp\left(\frac{4nb_1^2}{k}\left(Y - c_1^2/k\right)\right)\right]\right)^{1/4} \\ &\le \exp\left(\frac{c_1^2}{4k}\left(\exp\left(\frac{4nb_1^2}{k}\right) - 1 - \frac{4nb_1^2}{k}\right)\right) \\ &\le \exp\left(\frac{c_1^2}{4k}\left(\exp\left(\frac{4nb_2^2}{k}\right) - 1 - \frac{4nb_2^2}{k}\right)\right), \end{aligned} \tag{3.5}$$

where the last inequality follows from the fact that $x \mapsto e^x - 1 - x$ is increasing when $x > 0$. Other terms in (3.4) are bounded analogously and we have

$$\begin{aligned}&\chi^2(P_X \| Q_X) + 1 \\ &\quad\le \exp\left(\frac{c_1^2 + c_2^2 + 2c_1c_2}{4k}\left(\exp\left(\frac{4nb_2^2}{k}\right) - 1 - \frac{4nb_2^2}{k}\right)\right) \\ &\quad= \exp\left(\frac{(k-2\Delta)^2}{4k}\left(\exp\left(\frac{4n}{k}\left\lceil\frac{k}{k-2\Delta}\right\rceil^2\right) - 1 - \frac{4n}{k}\left\lceil\frac{k}{k-2\Delta}\right\rceil^2\right)\right).\end{aligned} \tag{3.6}$$

If $k - 2\Delta \ge \sqrt{k}$, the upper bound (3.6) implies that $n^*(k, \Delta) \ge \Omega(k-2\Delta/\sqrt{k})$ since the χ^2-divergence is finite with $O(k-2\Delta/\sqrt{k})$ samples, using the inequality that $e^x - 1 - x \le x^2/2$ for $x \ge 0$; if $k - 2\Delta \le \sqrt{k}$, the lower bound is trivial since $k-2\Delta/\sqrt{k} \le 1$.

Now we prove the refined estimate (3.2) for $1 \le \Delta < k/4$, in which case $|I| = c_1 = k - 4\Delta$, $|J| = c_2 = 2\Delta$, and $b_1 = 1$, $b_2 = 2$. When c_1 is close to k, $\text{Hypergeometric}(k, c_1, c_1)$ is no longer well approximated by $\text{Binomial}(c_1, c_1/k)$, and the upper bound in (3.5) yields a loose lower bound for the sample complexity. To fix this, note that in this case the set $K \triangleq (I \cup J)^c$ has small cardinality $|K| = 2\Delta$. The equality in (3.3) can be equivalently represented in terms of J, J' and K, K' by

$$\int \frac{P_\theta P_{\theta'}}{Q} = 1 + \frac{|J \cap J'| + |K \cap K'| - |J \cap K'| - |K \cap J'|}{k}.$$

By upper bounds analogous to (3.4)–(3.6),

$$\chi^2(P_X \| Q_X) + 1 \le \prod_{i=1}^{4} \left(\mathbb{E}[e^{4nB_i}]\right)^{1/4},$$

where

$$B_1 \triangleq \frac{1}{k}\Big(|J \cap J'| - \frac{(2\Delta)^2}{k}\Big), \quad B_2 \triangleq \frac{1}{k}\Big(|K \cap K'| - \frac{(2\Delta)^2}{k}\Big),$$
$$B_3 \triangleq -\frac{1}{k}\Big(|J \cap K'| - \frac{(2\Delta)^2}{k}\Big), \quad \text{and} \quad B_4 \triangleq -\frac{1}{k}\Big(|K \cap J'| - \frac{(2\Delta)^2}{k}\Big).$$

Note that $|J \cap J'|$, $|K \cap K'|$, $|J \cap K'|$, $|K \cap J'|$ are all distributed as Hypergeometric$(k, 2\Delta, 2\Delta)$, which is dominated by Binomial$(2\Delta, {}^{2\Delta}/_{k})$. For $Y \sim$ Binomial$(2\Delta, {}^{2\Delta}/_{k})$, we have

$$\big(\mathbb{E}[e^{4nB_i}]\big)^{1/4} \le \left(\mathbb{E}\left[\exp\left(t\left(Y - \frac{(2\Delta)^2}{k}\right)\right)\right]\right)^{1/4} \le \exp\left(\frac{(2\Delta)^2}{4k}(e^t - 1 - t)\right).$$

with $t = {}^{4n}/_{k}$ for $i = 1, 2$ and $t = -{}^{4n}/_{k}$ for $i = 3, 4$. Therefore,

$$\begin{aligned}\chi^2(P_X \| Q_X) + 1 &\le \exp\left(\frac{\Delta^2}{k}\big(2e^{4n/k} + 2e^{-4n/k} - 4\big)\right) \\ &= \exp\left(\frac{4\Delta^2}{k}\big(\cosh(4n/k) - 1\big)\right). \end{aligned} \tag{3.7}$$

The upper bound (3.7) yields the sample complexity

$$n^*(k, \Delta) \ge \Omega\Big(k \operatorname{arccosh}\Big(1 + \frac{k}{4\Delta^2}\Big)\Big). \qquad \square$$

Now we establish another lower bound for the sample complexity of the Distinct Elements problem for sampling without replacement. Since we can simulate sampling with replacement from samples obtained without replacement (see (A.1) for details), it is also a valid lower bound for $n^*(k, \Delta)$ defined in Definition 1. On the other hand, as observed in [37, Lemma 3.3] (see also [44, Lemma 5.14]), any estimator $\widehat{C}$ for the Distinct Elements problem with sampling without replacement leads to an estimator for the Support Size problem with slightly worse performance: Suppose we have n i.i.d. samples drawn from a distribution P whose minimum non-zero probability is at least $1/\ell$. Let $\widehat{C}_{\text{seen}}$ denote the number of distinct elements in these samples. Equivalently, these samples can be viewed as being generated in two steps: first, we draw k i.i.d. samples from P, whose realizations form an instance of a k-ball urn with $\widehat{C}_{\text{seen}}$ distinct colors; next, we draw n samples from this urn without replacement ($n \le k$), which clearly are distributed according to $P^{\otimes n}$. Suppose $\widehat{C}_{\text{seen}}$ is close to the actual support size of P. Then applying any algorithm for the Distinct Elements problem to these n i.i.d. samples constitutes a good support size estimator. Lemma 5 formalizes this intuition.

Lemma 5. *Suppose an estimator $\widehat{C}$ takes n samples from a k-ball urn ($n \le k$) without replacement and provides an estimation error of less than Δ with probability at least $1-\delta$. Applying $\widehat{C}$ with n i.i.d. samples from any distribution P with minimum non-zero mass $1/\ell$ and support size $S(P)$, we have*

$$|\widehat{C} - S(P)| \le 2\Delta$$

with probability at least $1 - \delta - \binom{\ell}{\Delta}\left(1 - \frac{\Delta}{\ell}\right)^k$.

Proof. Suppose that we take k i.i.d. samples from $P = (p_1, p_2, \dots)$, which form a k-ball urn consisting of C distinct colors. By the union bound,

$$\mathbb{P}[|C - S(P)| \ge \Delta] \le \sum_{\substack{I:|I|=\Delta,\\ p_i \ge 1/\ell,\, i\in I}} \left(1 - \sum_{i\in I} p_i\right)^k \le \binom{\ell}{\Delta}\left(1 - \frac{\Delta}{\ell}\right)^k.$$

Next we take n samples without replacement from this urn and apply the given estimator $\widehat{C}$. By assumption, conditioned on any realization of the k-ball urn, $|\widehat{C} - C| \le \Delta$ with probability at least $1-\delta$. Then $|\widehat{C} - S(P)| \le 2\Delta$ with probability at least $1 - \delta - \binom{\ell}{\Delta}\left(1 - \frac{\Delta}{\ell}\right)^k$. Marginally, these n samples are identically distributed as n i.i.d. samples from P. □

Combining with the sample complexity of the Support Size problem in (1.1), Lemma 5 leads to the following lower bound for the Distinct Elements problem:

Theorem 4. *Fix a sufficiently small constant c. For any $1 \le \Delta \le ck$,*

$$n^*(k, \Delta) \ge \Omega\left(\frac{k}{\log k} \log \frac{k}{\Delta}\right).$$

The same lower bound holds for sampling without replacement.

Proof. By the lower bound of the support size estimation problem obtained in [47, Theorem 2], if $n \le \frac{\alpha\ell}{\log \ell} \log^2 \ell/2\Delta$ and $2\Delta \le c_0\ell$ for some fixed constants $c_0 < 1/2$ and α, then for any $\widehat{C}$, there exists a distribution P with minimum non-zero mass $1/\ell$ such that $|\widehat{C} - S(P)| \le 2\Delta$ with probability at most 0.8. Applying Lemma 5 yields that, using n samples without replacement, no estimator can provide an estimation error of Δ with probability 0.9 for an arbitrary k-ball urn, provided $\binom{\ell}{\Delta}\left(1 - \frac{\Delta}{\ell}\right)^k \le 0.1$. Consequently, as long as $2\Delta \le c_0\ell$ and $\binom{\ell}{\Delta}\left(1 - \frac{\Delta}{\ell}\right)^k \le 0.1$, we have

$$n^*(k, \Delta) \ge \frac{\alpha\ell}{\log \ell} \log^2 \frac{\ell}{2\Delta}.$$

The desired lower bound follows from choosing $\ell \asymp \frac{k}{\log k/\Delta}$. □

4. Proof of results in Table 1

Below we explain how the sample complexity bounds summarized in Table 1 are obtained from various results in Section 2 and Section 3:

– The upper bounds are obtained from the worst-case MSE in Section 2 and the Markov inequality. In particular, the case of $\Delta \le \sqrt{k}(\log k)^{-\delta}$ follows from the second and the third upper bounds of Theorem 2; the case of $\sqrt{k} \le \Delta \le k^{0.5+\delta}$ follows from the first upper bound of Theorem 2; the case of $k^{1-\delta} \le \Delta \le ck$ follows from Theorem 1. By monotonicity, we have the $O(k \log\log k)$ upper bound when $\sqrt{k}(\log k)^{-\delta} \le \Delta \le \sqrt{k}$, the $O(k/\log k)$ upper bound when $\Delta \ge ck$, and the $O(k)$ upper bound when $k^{0.5+\delta} \le \Delta \le k^{1-\delta}$.

– The lower bound for $\Delta \le \sqrt{k}$ follows from Theorem 3; the lower bound for $k^{0.5+\delta} \le \Delta \le ck$ follows from Theorem 4. These further implies the $\Omega(k)$ lower bound for $\sqrt{k} \le \Delta \le k^{0.5+\delta}$ by monotonicity.

A. Connections between various sampling models

As mentioned in Section 1.2, four popular sampling models have been introduced in the statistics literature: the multinomial model, the hypergeometric model, the Bernoulli model, and the Poisson model. The connections between those models are explained in details in this section, as well as relations between the respective sample complexities.

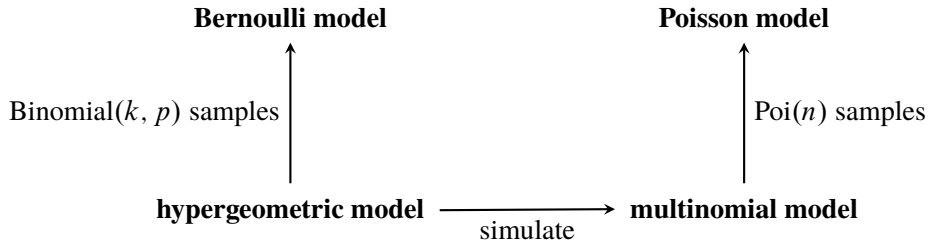


Figure 2. Relations between the four sampling models. In particular, hypergeometric (resp., multinomial) model reduces to the Bernoulli (resp., Poisson) model when the sample size is binomial (resp., Poisson) distributed.

The connections between different models are illustrated in Fig. 2. Under the Poisson model, the sample size is a Poisson random variable; conditioned on the sample size, the samples are i.i.d. which is identical to the multinomial model. The same relation holds as the Bernoulli model to the hypergeometric model. Given samples $(Y_1, \ldots, Y_n)$ uniformly drawn from a k-ball urn without replacement

(hypergeometric model), we can simulate $(X_1, \dots, X_n)$ drawn with replacement (multinomial model) as follows: for each $i = 1, \dots, n$, let

$$X_i = \begin{cases} Y_i, & \text{with probability } 1 - \frac{i-1}{k}, \\ Y_m, & \text{with probability } \frac{i-1}{k}, \quad m \sim \text{Uniform}([i-1]). \end{cases} \tag{A.1}$$

In view of the connections in Fig. 2, any estimator constructed for one specific model can be adapted to another. The adaptation from multinomial to hypergeometric model is provided by the simulation in (A.1), and the other direction is given by Lemma 5 (without modifying the estimator). The following result provides a recipe for going between fixed and randomized sample size:

Lemma 6. *Let N be an $\mathbb{N}$-valued random variable.*

(a) *Given any*[4] *$\widehat{C}$ that uses n samples and succeeds with probability at least $1 - \delta$, there exists $\widehat{C}'$ using N samples that succeeds with probability at least $1 - \delta - \mathbb{P}[N < n]$.*

(b) *Given any $\widetilde{C}$ using N samples that succeeds with probability at least $1 - \delta$, there exists $\widetilde{C}'$ that uses n samples and succeeds with probability at least $1 - \delta - \mathbb{P}[N > n]$.*

Proof. (a) Denote the samples by $X_1, \dots, X_N$. Following [37, Lemma 5.3(a)], define $\widehat{C}'$ as

$$\widehat{C}' = \begin{cases} \widehat{C}(X_1, \dots, X_n), & N \geq n, \\ 0, & N < n. \end{cases}$$

Then $\widehat{C}'$ succeeds as long as $N \geq n$ and $\widehat{C}$ succeeds, which has probability at least $1 - \delta - \mathbb{P}[N < n]$.

(b) Denote the samples by $X_1, \dots, X_n$. Draw a random variable m from the distribution of N and define $\widetilde{C}'$ as

$$\widetilde{C}' = \begin{cases} \widetilde{C}(X_1, \dots, X_m), & m \leq n, \\ 0, & m > n. \end{cases}$$

The given estimator $\widetilde{C}$ fails with probability

$$\sum_{j \geq 0} \mathbb{P}[\widetilde{C} \text{ fails} \mid N = j] \mathbb{P}[N = j] \leq \delta.$$

Consequently,

$$\sum_{j=0}^{n} \mathbb{P}[\widetilde{C} \text{ fails} \mid N = j] \mathbb{P}[N = j] \leq \delta.$$

[4] More precisely, here and below $\widehat{C}$ is understood as a sequence of estimators indexed by the sample size $(X_1, \dots, X_n) \mapsto \widehat{C}(X_1, \dots, X_n)$.

The estimator $\widetilde{C}'$ fails with probability at most

$$\sum_{j=0}^{n} \mathbb{P}[\widetilde{C} \text{ fails} \mid m = j]\mathbb{P}[m = j] + \mathbb{P}[m > n] \le \delta + \mathbb{P}[m > n],$$

which completes the proof. □

The adaptations of estimators between different sampling models imply the relations of the fundamental limits on the corresponding sample complexities. Extending Definition 1, let $n_M^*(k, \Delta, \delta)$, $n_H^*(k, \Delta, \delta)$, $n_B^*(k, \Delta, \delta)$, and $n_P^*(k, \Delta, \delta)$ be the minimum expected sample size under the multinomial, hypergeometric, Bernoulli, and Poisson sampling model, respectively, such that there exists an estimator $\widehat{C}$ satisfying $\mathbb{P}[|\widehat{C} - C| \ge \Delta] \le \delta$. Combining Chernoff bounds (see, e.g. [31, Theorem 4.4, 4.5, and 5.4]), we obtain Corollary 1, in which the connection between multinomial and Poisson models gives a rigorous justification of the assumption on the Poisson sampling model in Section 2.

Corollary 1. *The following relations hold:*

- n_H^* *versus* n_M^**:*
  - (a) $n_H^*(k, \Delta, \delta) \le n_M^*(k, \Delta, \delta)$*;*
  - (b) $n_H^*(k, \Delta, \delta) \le n \Rightarrow n_M^*(k', 2\Delta, \delta + \binom{k'}{\Delta}(1 - \frac{\Delta}{k'})^k) \le n$*, for any* $k' \in \mathbb{N}$*. In particular, if* δ *is a constant, then we can choose* $k' = \Theta(k/\log k/\Delta)$.
- n_P^* *versus* n_M^**:*
  - (c) $n_P^*(k, \Delta, \delta) \le n \Rightarrow n_M^*(k, \Delta, \delta + (e/4)^n) \le 2n$*;*
  - (d) $n_M^*(k, \Delta, \delta) \le n \Rightarrow n_P^*(k, \Delta, \delta + (2/e)^n) \le 2n$.
- n_B^* *versus* n_H^**:*
  - (e) $n_B^*(k, \Delta, \delta) \le n \Rightarrow n_H^*(k, \Delta, \delta + (e/4)^n) \le 2n$*;*
  - (f) $n_H^*(k, \Delta, \delta) \le n \Rightarrow n_B^*(k, \Delta, \delta + (2/e)^n) \le 2n$.

B. Correlation decay between fingerprints

Recall that the fingerprints are defined by $\Phi_j = \sum_i \mathbf{1}_{\{N_i = j\}}$, where N_i denotes the histogram of samples. Under the Poisson model, $N_i \overset{\text{ind}}{\sim} \text{Poi}(np_i)$. Then

$$\text{cov}(\Phi_j, \Phi_{j'}) = -\sum_i \mathbb{P}[N_i = j]\mathbb{P}[N_i = j'], \quad j \ne j',$$

$$\text{var}[\Phi_j] = \sum_i \mathbb{P}[N_i = j]\big(1 - \mathbb{P}[N_i = j]\big).$$

The correlation coefficient between Φ_0 and Φ_j follows immediately:

$$\begin{aligned}
&|\rho(\Phi_0, \Phi_j)| \\
&= \sum_i \frac{\mathbb{P}[N_i = 0]\mathbb{P}[N_i = j]}{\sqrt{\sum_l \mathbb{P}[N_l = 0]\big(1 - \mathbb{P}[N_l = 0]\big) \sum_l \mathbb{P}[N_l = j]\big(1 - \mathbb{P}[N_l = j]\big)}} \\
&\le \sum_i \frac{\mathbb{P}[N_i = 0]\mathbb{P}[N_i = j]}{\sqrt{\mathbb{P}[N_i = 0]\big(1 - \mathbb{P}[N_i = 0]\big)\mathbb{P}[N_i = j]\big(1 - \mathbb{P}[N_i = j]\big)}} \qquad \text{(B.1)} \\
&= \sum_i \sqrt{\frac{\mathbb{P}[N_i = 0]}{1 - \mathbb{P}[N_i = 0]} \frac{\mathbb{P}[N_i = j]}{1 - \mathbb{P}[N_i = j]}} = \sum_i \sqrt{\frac{e^{-\lambda_i}}{1 - e^{-\lambda_i}} \frac{\frac{e^{-\lambda_i}\lambda_i^j}{j!}}{1 - \frac{e^{-\lambda_i}\lambda_i^j}{j!}}},
\end{aligned}$$

where $\lambda_i = np_i$. Note that

$$\max_{x>0} \frac{e^{-x}x^j}{j!} = \frac{e^{-j}j^j}{j!} \to 0 \quad \text{as } j \to \infty.$$

Therefore, for any $x > 0$,

$$\frac{e^{-x}}{1 - e^{-x}} \frac{\frac{e^{-x}x^j}{j!}}{1 - \frac{e^{-x}x^j}{j!}} = \frac{1}{j!} \frac{e^{-2x}x^j}{1 - e^{-x}}\big(1 + o_j(1)\big), \qquad \text{(B.2)}$$

where $o_j(1)$ is uniform as $j \to \infty$. Taking derivative, the function $x \mapsto \frac{e^{-2x}x^j}{1-e^{-x}}$ on $x > 0$ is increasing if and only if $x + e^x(j - 2x) - j > 0$, and the maximum is attained at $x = j/2 + o_j(1)$. Therefore, applying $j! > (j/e)^j$,

$$\frac{1}{j!} \frac{e^{-2x}x^j}{1 - e^{-x}} \le \big(1 + o_j(1)\big)2^{-j}. \qquad \text{(B.3)}$$

Combining (B.1)–(B.3), we conclude that

$$|\rho(\Phi_0, \Phi_j)| \le k2^{-j/2}\big(1 + o_j(1)\big).$$

C. Proof of auxiliary lemmas

Proof of Lemma 3. For any $z \in \mathbb{C}$, we can represent the forward difference in (2.19) as an integral:

$$\begin{aligned}
\Delta^m f(z) &= f(z + m) - \binom{m}{1} f(z + m - 1) + \cdots + (-1)^m f(z) \\
&= \int_{[0,1]^m} f^{(m)}(z + x_1 + \cdots + x_m)\mathrm{d}x_1 \cdots \mathrm{d}x_m.
\end{aligned}$$

Therefore,

$$|t_m(z)| = \left|\frac{1}{m!}\Delta^m p_m(z)\right| \le \frac{1}{m!}\sup_{0\le\xi\le m}|p_m^{(m)}(z+\xi)|. \tag{C.1}$$

Recall the definition of p_m in (2.20). Let $p_m(z) = \sum_{l=0}^{2m} a_\ell z^\ell$. Let

$$z(z-1)\cdots(z-m+1) = \sum_{i=0}^{m} b_i z^i$$

and

$$(z-M)(z-M-1)\cdots(z-M-m+1) = \sum_{i=0}^{m} c_i z^i.$$

Expanding the product and collecting the coefficients yields a simple upper bound:

$$|b_i| \le 2^m (m-1)^{m-i},$$
$$|c_i| \le 2^m (M+m-1)^{m-i} \le 2^m (2M)^{m-i} \le 2^{2m} M^{m-i}.$$

Since

$$\sum_{\ell=0}^{2m} a_\ell z^\ell = \Big(\sum_{i=0}^{m} b_i z^i\Big)\Big(\sum_{j=0}^{m} c_j z^j\Big),$$

for $\ell \ge m$,

$$\begin{aligned}|a_\ell| &= \Big|\sum_{i=\ell-m}^{m} b_i c_{\ell-i}\Big| \le \sum_{i=\ell-m}^{m} 2^{3m}(m-1)^{m-i} M^{m-\ell+i} \\ &= 2^{3m} M^{2m-\ell} \sum_{i=\ell-m}^{m} \left(\frac{m-1}{M}\right)^{m-i} \le m 2^{3m} M^{2m-\ell}.\end{aligned}$$

Taking mth derivative of p_m, we obtain

$$\begin{aligned}|p_m^{(m)}(z)| &= \left|\sum_{j=0}^{m} a_{j+m}\frac{(j+m)!}{j!} z^j\right| \\ &\le \sum_{j=0}^{m} |a_{j+m} M^j| \binom{m+j}{m} m! \left|\frac{z}{M}\right|^j \le m 2^{3m} M^m m! (2e)^m \sum_{j=0}^{m} \left|\frac{z}{M}\right|^j \\ &\le m^2 2^{6m} M^m m! \left(\frac{|z|}{M} \vee 1\right)^m = m^2 2^{6m} m! \big(|z| \vee M\big)^m.\end{aligned}$$

Then the desired (2.31) follows from (C.1). □

Proof of Lemma 4. The following uniform asymptotic expansions of the Stirling numbers of the first kind was obtained in [10, Theorem 2]:

$$|s(n+1,m+1)| = \begin{cases} \frac{n!}{m!}(\log n+\gamma)^m\big(1+o(1)\big), & 1 \le m \le \sqrt{\log n}, \\ \frac{\Gamma(n+1+R)}{\Gamma(R)R^{m+1}\sqrt{2\pi H}}\big(1+o(1)\big), & \sqrt{\log n} \le m \le n-n^{1/3}, \\ \binom{n+1}{m+1}\big(\frac{m+1}{2}\big)^{n-m}\big(1+o(1)\big), & n-n^{1/3} \le m \le n, \end{cases}$$

where γ is Euler's constant, R is the unique positive solution to $h'(x) = 0$ with

$$h(x) \triangleq \log \frac{\Gamma(x+n+1)}{\Gamma(x+1)x^m}, \quad H = R^2h''(R),$$

and all $o(1)$ terms are uniform in m. In the following we consider each range separately and prove the non-asymptotic approximation in (2.36).

Case I. For $1 \le m \le \sqrt{\log n}$, Stirling's approximation gives

$$\frac{n!}{m!}(\log n+\gamma)^m = n!\left(\Theta\left(\frac{\log n}{m}\right)\right)^m.$$

Case II. For $n-n^{1/3} \le m \le n$,

$$\begin{aligned}\binom{n+1}{m+1}\left(\frac{m+1}{2}\right)^{n-m} &= \frac{n!}{m!}\left(\Theta\left(\frac{m}{n-m}\right)\right)^{n-m} \\ &= n!\exp\left(m\left(\frac{n-m}{m}\log\left(\Theta\left(\frac{m}{n-m}\right)\right)-\log\Theta(m)\right)\right) \\ &= n!\left(\Theta\left(\frac{1}{m}\right)\right)^m.\end{aligned}$$

Case III. For $\sqrt{\log n} \le m \le n-n^{1/3}$, note that

$$h(x) = \sum_{i=1}^{n}\log(x+i) - m\log x,$$

and thus

$$H = R^2h''(R) = m - \sum_{i=1}^{n}\frac{R^2}{(R+i)^2} \le m.$$

By [32, Lemma 4.1], $H = \omega(1)$ in this range. Hence,

$$|s(n+1,m+1)| = \frac{\Gamma(n+1+R)}{\Gamma(R)R^{m+1}}\big(\Theta(1)\big)^m = \frac{n!}{R^m}\frac{\Gamma(n+1+R)}{n!\Gamma(R+1)}\big(\Theta(1)\big)^m, \quad \text{(C.2)}$$

where R is the solution to $x(\frac{1}{x+1}+\cdots+\frac{1}{x+n}) = m$. Bounding the sum by integrals, we have

$$R\log\left(1+\frac{n}{R+1}\right) \le m \le R\log\left(1+\frac{n}{R}\right).$$

If $\sqrt{\log n} \le m \le n/e$, then $R \asymp m/\log n/m$, and hence

$$1 \le \frac{\Gamma(n+1+R)}{n!\Gamma(R+1)} \le \left(O\left(\frac{n+R}{R}\right)\right)^R = \exp\big(O(m)\big).$$

In view of (C.2), we have

$$|s(n+1,m+1)| = \frac{n!}{(\Theta(R))^m},$$

which is exactly (2.36) when $m \le n/e$. If $n/e \le m \le n-n^{1/3}$, then $R \asymp n^2/(n-m)$, and

$$\begin{aligned}\frac{1}{R^m}\frac{\Gamma(n+1+R)}{n!\Gamma(R+1)} &= R^{-m}\left(\Theta\left(\frac{n+R}{n}\right)\right)^n \\ &= \exp\left(-m\log\Theta\left(\frac{n^2}{n-m}\right)+n\log\Theta\left(\frac{n}{n-m}\right)\right) \\ &= \exp\left(-m\log\Theta(n)+(n-m)\log\Theta\left(\frac{n}{n-m}\right)\right) \\ &= \exp\big(-m\log\Theta(n)\big).\end{aligned}$$

Combining (C.2) yields that

$$|s(n+1,m+1)| = n!\big(\Theta\big(1/n\big)\big)^m,$$

which coincides with (2.36) since $n \asymp m$ is this range. □

Acknowledgements. This research has been supported in part by the National Science Foundation under the grant agreement IIS-14-47879 and CCF-15-27105 and an NSF CAREER award CCF-1651588. The authors thank Greg Valiant for helpful discussion on Lemma 5.14 in his thesis [44]. We thank the anonymous referees for constructive comments which have helped to improve the presentation of the paper.

References

[1] B. Beckermann, The condition number of real Vandermonde, Krylov and positive definite Hankel matrices, *Numer. Math.*, **85** (2000), no. 4, 553–577. Zbl 0965.15003 MR 1771780

[2] J. Bunge and M Fitzpatrick, Estimating the number of species: a review, *J. Am. Stat. Assoc.*, **88** (1993), no. 421, 364–373.

[3] Z. Bar-Yossef, T. S. Jayram, R. Kumar, D Sivakumar, and L. Trevisan, Counting distinct elements in a data stream, in *Proceedings of the 6th Randomization and Approximation Techniques in Computer Science*, 1–10, Lecture Notes in Comput. Sci., 2483, Springer, Berlin, 2002. Zbl 1028.68949 MR 2047015

[4] Z. Bar-Yossef, R. Kumar, and D Sivakumar, Sampling algorithms: lower bounds and applications, in *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, 266–275, ACM, New York, 2001. Zbl 1323.68292 MR 2120325

[5] M. Charikar, S. Chaudhuri, R. Motwani, and V. Narasayya, 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, 2000.

[6] A. Córdova, W. Gautschi, and S. Ruscheweyh, Vandermonde matrices on the circle: spectral properties and conditioning, *Numer. Math.*, **57** (1990), no. 1, 577–591. Zbl 0728.15003 MR 1062367

[7] A. Chao and S.-M. Lee, Estimating the number of classes via sample coverage, *J. Am. Stat. Assoc.*, **87** (1992), no. 417, 210–217. Zbl 0850.62145 MR 1158639

[8] Y. Chen and N. Lawrence, Small eigenvalues of large hankel matrices, *J. Phys. A*, **32** (1999), no. 42, 7305–7315. Zbl 0999.15007 MR 1747170

[9] T. T. Cai and M. G. Low, Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional, *Ann. Statist.*, **39** (2011), no. 2, 1012–1041. Zbl 1277.62101 MR 2816346

[10] R Chelluri, L. B Richmond, and N. M Temme, Asymptotic estimates for generalized Stirling number, *Analysis (Munich)*, **20** (2000), no. 1, 1–14. Zbl 1093.11504 MR 1757065

[11] A. Eisinberg, P. Pugliese, and N. Salerno, Vandermonde matrices on integer nodes: the rectangular case, *Numer. Math.*, **87** (2001), no. 4, 663–674. Zbl 0974.65029 MR 1815730

[12] W. W. Esty, Estimation of the size of a coinage: A survey and comparison of methods, *Numismatic Chronicle*, **146** (1986), 185–215. Zbl 0603.62051 MR 925622

[13] B. Efron and R. Thisted, Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, **63** (1976), no. 3, 435–447. Zbl 0344.62088

[14] R. A. Fisher, A. S. Corbet, and C. B. Williams, The relation between the number of species and the number of individuals in a random sample of an animal population, *J. Animal Ecol.*, **12** (1943), no. 1, 42–58.

[15] P. J. S. G. Ferreira, Super-resolution, the recovery of missing samples and Vandermonde matrices on the unit circle, in *Proceedings of the Workshop on Sampling Theory and Applications (Loen, Norway, August 1999)*, 1999.

[16] P. Flajolet, É. Fusy, O. Gandouet, and F. Meunier, Hyperloglog: The analysis of a near-optimal cardinality estimation algorithm, in *2007 Conference on Analysis of Algorithms, AofA 07*, 127–145, Discrete Math. Theor. Comput. Sci. Proc., AH, Assoc. Discrete Math. Theor. Comput. Sci., Nancy, 2007. Zbl 1192.68959 MR 2509517

[17] O. Frank, Estimation of the number of connected components in a graph by using a sampled subgraph, *Scand. J. Statist.*, **5** (1978), no. 4, 177–188. Zbl 0391.05049 MR 515656

[18] W. Gautschi, How (un)stable are Vandermonde systems?, in *Asymptotic and computational analysis (Winnipeg, MB, 1989)*, 193–210, Lecture Notes in Pure and Appl. Math., 124, Dekker, New York, 1990. Zbl 0707.15003 MR 1052434

[19] L. A. Goodman, On the estimation of the number of classes in a population, *Ann. Math. Statistics*, **20** (1949), 572–579. Zbl 0035.09102 MR 32165

[20] I. J. Good, The population frequencies of species and the estimation of population parameters, *Biometrika*, **40** (1953), no. 3-4, 237–264. Zbl 0051.37103 MR 61330

[21] A. Goldenshluger and V. Spokoiny, On the shape–from–moments problem and recovering edges from noisy Radon data, *Probab. Theory Related Fields*, **128** (2004), no. 1, 123–140. Zbl 1035.62001 MR 2027297

[22] B. M. Hill, Posterior moments of the number of species in a finite population and the posterior probability of finding a new species, *J. Am. Stat. Assoc.*, **74** (1979), no. 367, 668–673. Zbl 0417.62079 MR 548263

[23] W. Hoeffding, Probability inequalities for sums of bounded random variables, *J. Am. Stat. Assoc.*, **58** (1963), no. 301, 13–30. Zbl 0127.10602 MR 144363

[24] W.-C. Hou, G. Ozsoyoglu, and B. K. Taneja, Statistical estimators for relational algebra expressions, in *Proceedings of the seventh ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, 276–287, ACM, 1988.

[25] C. Jordan, *Calculus of finite differences*, Carver Chelsea Publishing Co., New York, 1947. Zbl 0154.33901 MR 183987

[26] J. Jiao, K. Venkat, Y. Han, and T. Weissman, Minimax estimation of functionals of discrete distributions, *IEEE Trans. Inform. Theory*, **61** (2015), no. 5, 2835–2885. Zbl 1359.62104 MR 3342309

[27] D. M. Kane, J. Nelson, and D. P. Woodruff, An optimal algorithm for the distinct elements problem, in *Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 41–52, ACM, 2010.

[28] O. Lepski, A. Nemirovski, and V. Spokoiny, On estimation of the L_r norm of a regression function, *Probab. Theory Related Fields*, **113** (1999), no. 2, 221–253. Zbl 0921.62103 MR 1670867

[29] S.-H. Lo, From the species problem to a general coverage problem via a new interpretation, *Ann. Statist.*, **20** (1992), no. 2, 1094–1109. Zbl 0778.62028 MR 1165608

[30] A. Moitra, Super-resolution, extremal functions and the condition number of Vandermonde matrices, in *STOC'15. Proceedings of the 2015 ACM Symposium on Theory of Computing*, 821–830, ACM, New York, 2015. Zbl 1321.68421 MR 3388262

[31] M. Mitzenmacher and E. Upfal, *Probability and computing: Randomized algorithms and probabilistic analysis*, Cambridge University Press, 2005. Zbl 1092.60001 MR 2144605

[32] L Moser and M Wyman, Asymptotic development of the Stirling numbers of the first kind, *Journal of the London Mathematical Society*, **1** (1958), no. 2, 133–146. Zbl 0081.28202 MR 98202

[33] J. F. Naughton and S Seshadri, On estimating the size of projections, in *ICDT '90 (Paris, 1990)*, 499–513, Lecture Notes in Comput. Sci., 470, Springer, Berlin, 1990. Zbl 0789.68041 MR 1085744

[34] A. F. Nikiforov, V. B. Uvarov, and S. K. Suslov, *Classical orthogonal polynomials of a discrete variable*, Translated from the Russian, Springer Series in Computational Physics, Springer-Verlag, Berlin, 1991. Zbl 0743.33001 MR 1149380

[35] L. Paninski, Estimation of entropy and mutual information, *Neural Comput.*, **15** (2003), no. 6, 1191–1253. Zbl 1052.62003

[36] L. Paninski, Estimating entropy on m bins given fewer than m samples, *IEEE Trans. Inform. Theory*, **50** (2004), no. 9, 2200–2203. Zbl 1315.94025 MR 2097210

[37] S. Raskhodnikova, D. Ron, A. Shpilka, and A. Smith, Strong lower bounds for approximating distribution support size and the distinct elements problem, *SIAM J. Comput.*, **39** (2009), no. 3, 813–842. Zbl 1194.68124 MR 2538840

[38] G. Szegö, *Orthogonal polynomials*, 4th edition, American Mathematical Society, Colloquium Publications, XXIII, American Mathematical Society, Providence, R.I., 1975. Zbl 0305.42011 MR 372517

[39] N. M. Temme, Asymptotic estimates of Stirling numbers, *Stud. Appl. Math.*, **89** (1993), no. 3, 233–243. Zbl 0784.11007 MR 1223774

[40] A. F. Timan, *Theory of approximation of functions of a real variable*, Translated from the Russian by J. Berry, English translation edited and editorial preface by J. Cossar, International Series of Monographs in Pure and Applied Mathematics, 34, A Pergamon Press Book, The Macmillan Co., New York 1963. Zbl 0117.29001 MR 192238

[41] J. Todd, The condition of the finite segments of the Hilbert matrix, in *Contributions to the solution of systems of linear equations and the determination of eigenvalues*, 109?116, National Bureau of Standards Applied Mathematics Series, 39, U.S. Government Printing Office, Washington, D.C., 1954. Zbl 0058.01003 MR 68304

[42] A. B. Tsybakov, *Introduction to Nonparametric Estimation*, Introduction to nonparametric estimation, Revised and extended from the 2004 French original, Translated by Vladimir Zaiats, Springer Series in Statistics, Springer, New York, 2009. Zbl 1176.62032 MR 2724359

[43] P. Valiant, Testing symmetric properties of distributions, *SIAM J. Comput.*, **40** (2011), no. 6, 1927–1968. Zbl 1233.62112 MR 2863200

[44] G. Valiant, *Algorithmic Approaches to Statistical Questions*, Ph.D. thesis, EECS Department, University of California, Berkeley, Sep. 2012. MR 3152616

[45] G. Valiant and P. Valiant, Estimating the unseen: an $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs, in *STOC'11. Proceedings of the 43rd ACM Symposium on Theory of Computing*, 685–694, ACM, New York, 2011. Zbl 1288.68186 MR 2932019

[46] G. Valiant and P. Valiant, The power of linear estimators, in *2011 IEEE 52nd Annual Symposium on Foundations of Computer Science. FOCS 2011*, 403–412, IEEE Computer Soc., Los Alamitos, CA, 2011. Zbl 1292.68159 MR 2932716

[47] Y. Wu and P. Yang, Chebyshev polynomials, moment matching, and optimal estimation of the unseen, 2015. arXiv:1504.01227

[48] Y. Wu and P. Yang, Minimax rates of entropy estimation on large alphabets via best polynomial approximation, *IEEE Trans. Inform. Theory*, **62** (2016), no. 6, 3702–3720. Zbl 1359.94375 MR 3506758

Received 20 May, 2017; revised 14 January, 2018

Y. Wu, Department of Statistics and Data Science, Yale University, New Haven, CT 06511, USA
E-mail: yihong.wu@yale.edu
P. Yang, Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
E-mail: pyang14@illinois.edu