

Bias Reduction for Sum Estimation

Talya Eden  

Bar-Ilan University, Ramat Gan, Israel

Jakob Bæk Tejs Houen 


BARC, University of Copenhagen, Denmark

Shyam Narayanan 

MIT, Cambridge, MA, USA

Will Rosenbaum  

Amherst College, MA, USA

Jakub Tětek 

BARC, University of Copenhagen, Denmark

Abstract

In classical statistics and distribution testing, it is often assumed that elements can be sampled exactly from some distribution \mathcal{P} , and that when an element x is sampled, the probability $\mathcal{P}(x)$ of sampling x is also known. In this setting, recent work in distribution testing has shown that many algorithms are robust in the sense that they still produce correct output if the elements are drawn from any distribution \mathcal{Q} that is sufficiently close to \mathcal{P} . This phenomenon raises interesting questions: under what conditions is a “noisy” distribution \mathcal{Q} sufficient, and what is the algorithmic cost of coping with this noise?

In this paper, we investigate these questions for the problem of estimating the sum of a multiset of N real values x_1, \dots, x_N . This problem is well-studied in the statistical literature in the case $\mathcal{P} = \mathcal{Q}$, where the Hansen-Hurwitz estimator [Annals of Mathematical Statistics, 1943] is frequently used. We assume that for some (known) distribution \mathcal{P} , values are sampled from a distribution \mathcal{Q} that is pointwise close to \mathcal{P} . That is, there is a parameter $\gamma < 1$ such that for all x_i , $(1 - \gamma)\mathcal{P}(i) \leq \mathcal{Q}(i) \leq (1 + \gamma)\mathcal{P}(i)$. For every positive integer k we define an estimator ζ_k for $\mu = \sum_i x_i$ whose bias is proportional to γ^k (where our ζ_1 reduces to the classical Hansen-Hurwitz estimator). As a special case, we show that if \mathcal{Q} is pointwise γ -close to uniform and all $x_i \in \{0, 1\}$, for any $\varepsilon > 0$, we can estimate μ to within additive error εN using $m = \Theta(N^{1-\frac{1}{k}}/\varepsilon^{2/k})$ samples, where $k = \lceil (\lg \varepsilon)/(\lg \gamma) \rceil$. We then show that this sample complexity is essentially optimal. Interestingly, our upper and lower bounds show that the sample complexity need not vary uniformly with the desired error parameter ε : for some values of ε , perturbations in its value have no asymptotic effect on the sample complexity, while for other values, any decrease in its value results in an asymptotically larger sample complexity.

2012 ACM Subject Classification Mathematics of computing → Probabilistic algorithms; Theory of computation → Sample complexity and generalization bounds; Theory of computation → Streaming, sublinear and near linear time algorithms; Theory of computation → Lower bounds and information complexity

Keywords and phrases bias reduction, sum estimation, sublinear time algorithms, sample complexity

Digital Object Identifier 10.4230/LIPIcs.APPROX/RANDOM.2023.62

Category RANDOM

Funding *Jakub Tětek*: Supported by the VILLUM Foundation grant 16582.

Acknowledgements We thank the anonymous peer reviewers, whose feedback helped improve our manuscript.



© Talya Eden, Jakob Bæk Tejs Houen, Shyam Narayanan, Will Rosenbaum, and Jakub Tětek; licensed under Creative Commons License CC-BY 4.0

Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2023).

Editors: Nicole Megow and Adam D. Smith; Article No. 62; pp. 62:1–62:21



Leibniz International Proceedings in Informatics

LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

Consider the following simple problem. Let us have values $x_i \in \{0, 1\}$ for $i \in [N]$ and assume we may sample i from a distribution \mathcal{Q} that is pointwise γ -close to uniform (see Definition 2). It is easy to obtain an additive $\pm\gamma N$ approximation of the number of 1's. But is it possible to get a better approximation using a number of samples that is sub-linear in N ? We answer this question positively. Specifically, we solve a more general sum estimation problem, with the above problem being the simplest application. Additionally, we derive lower bounds showing that for a wide range of parameters, the sample complexity of our algorithm is asymptotically tight.

In full generality, estimating the sum of a (multi)set of numbers is a fundamental problem in statistics, and the problem plays an important role in the design of efficient algorithms for large datasets. The basic problem can be formulated as follows: given a multiset of N elements, $S = \{x_1, x_2, \dots, x_N\}$, compute an estimate of the sum $\mu = \sum_{i \in [N]} x_i$.

Assume that the values x_i can be sampled according to some probability distribution \mathcal{Q} over S (equivalently, over $[N]$). The classical work of Hansen and Hurwitz [12] examines the setting in which, when an element $x \in S$ is sampled, the probability $\mathcal{Q}(x)$ can be determined. They introduce the Hansen-Hurwitz estimator defined by

$$\mu_{\text{HH}} = \frac{1}{m} \sum_{j=1}^m \frac{X_j}{\mathcal{Q}(X_j)} \quad (1)$$

where X_1, X_2, \dots, X_m are samples taken from the distribution \mathcal{Q} . This estimator has been used extensively (though often implicitly) in sublinear algorithms [5, 20, 1, 14]. Hansen and Hurwitz prove that (1) is an accurate estimator of μ via the following theorem:

► **Theorem 1** (Hansen & Hurwitz, 1943 [12]). *The value μ_{HH} is an unbiased estimator of the sum μ (i.e., $\mathbb{E}(\mu_{\text{HH}}) = \mu$) and its variance is*

$$\text{Var}[\mu_{\text{HH}}] = \frac{1}{m} \sum_{i=1}^N \mathcal{Q}(i) \left(\frac{x_i}{\mathcal{Q}(i)} - \mu \right)^2. \quad (2)$$

Theorem 1 can be applied to obtain probabilistic guarantees for estimating μ via sampling. For example, if one wishes to compute a $1 \pm \varepsilon$ multiplicative estimate of μ with probability $1 - \delta$, by Chebyshev's inequality, it suffices to take m sufficiently large that $\text{Var}[\mu_{\text{HH}}]/(\varepsilon^2 \mu^2) < \delta$.

In practice, it may however be unreasonable to assume that the probability distribution from which elements are sampled is known precisely. For example, the underlying process generating the samples may be noisy or may induce some underlying bias. We model this situation by assuming that the true sampling distribution \mathcal{Q} is close to some known distribution \mathcal{P} . When an element x is sampled, the probability $\mathcal{P}(x)$ can be determined, but not the true probability $\mathcal{Q}(x)$. We assume that \mathcal{Q} is *pointwise close* to \mathcal{P} in the following sense. The assumption of pointwise closeness turns out to be necessary (as compared to weaker notions of closeness); see the discussion on page 4.

► **Definition 2.** *Let \mathcal{P} and \mathcal{Q} be probability distributions over a (multi)set S . Then for any $\gamma < 1$, we say that \mathcal{Q} is pointwise γ -close to \mathcal{P} if for every $x \in S$, we have*

$$(1 - \gamma)\mathcal{P}(x) \leq \mathcal{Q}(x) \leq (1 + \gamma)\mathcal{P}(x). \quad (3)$$

Given the situation above, one can apply the Hansen-Hurwitz estimator (1) with the known distribution \mathcal{P} in place of the true sample distribution \mathcal{Q} . We define the *positive sum*, μ_+ to be

$$\mu_+ = \sum_{x \in S} |x|. \quad (4)$$

It is straightforward to show that the resulting estimator has bias at most $\gamma\mu_+$, and its variance increases by a factor of $1 + O(\gamma)$. However, the parameter γ may be too large to guarantee the desired error in the estimate of μ . For the above problem of estimating the sum of 0-1 values, this would lead to error of γN , while we want error of εN for $\varepsilon < \gamma$.¹

Our setting is closely related to recent work in distribution testing. For example, it has been noted that many algorithms that rely on a probability oracle are “robust” in the sense that we may do distribution testing to within ε if the oracle’s answers have relative error of, say, $1 \pm \varepsilon/3$ [18, 3]. Our work goes further in the sense that our estimators work also in the setting when the error in the oracle’s answers is greater than the desired error parameter ε . Specifically, our goal is to characterize the (sample) complexity of a task as a function of the oracle error parameter γ and a desired approximation parameter ε . This can also be seen as a generalization of the learning-augmented distribution testing setting where γ is assumed to be constant [6].

1.1 Our Contributions

In this paper, our goal is to estimate the sum $\mu = \sum_{i=1}^N x_i$ with an error that is strictly less than the bias $\gamma\mu_+$ (Equation (4)) guaranteed by μ_{HH} . Specifically, given a desired error parameter ε with $0 < \varepsilon < \gamma$, we wish to estimate μ with bias close to $\varepsilon\mu_+$. In our setting, for each sample we are given a random index $i \in [N]$ drawn from the unknown distribution \mathcal{Q} , along with the value x_i and our estimate $\mathcal{P}(i)$ of the true probability $\mathcal{Q}(i)$. We introduce a family of estimators ζ_1, ζ_2, \dots , where each ζ_k has bias at most $\gamma^k \mu_+$. To motivate the construction of ζ_k , we first re-write the Hansen-Hurwitz estimator in terms of the frequency vector of samples from S . Specifically, if X_1, X_2, \dots, X_m are the sampled elements, define the frequency vector $Y = (Y_1, Y_2, \dots, Y_N)$ by

$$Y_i = |\{j : X_j = i\}|.$$

We define the estimator

$$\xi_1 = \frac{1}{m} \sum_{i=1}^N Y_i \cdot \frac{x_i}{\mathcal{P}(i)}.$$

Note that this estimator can be efficiently implemented, as the items that have not been sampled contribute 0 to the sum. We may thus implement this in time linear in the sample complexity, and do not need to take $O(N)$ time.

In the case where $\mathcal{P} = \mathcal{Q}$, ξ_1 is equivalent to the Hansen-Hurwitz estimator μ_{HH} . More generally, \mathcal{Q} is pointwise γ -close to \mathcal{P} , and ξ_1 has bias at most $\gamma\mu_+$.

The estimator ξ_1 can be generalized as follows. Rather than sampling individual elements, we can examine h -wise collisions between samples, where an h -wise collision consists of h samples resulting in the same outcome.

¹ It is possible to get tighter bounds if parameterizing also by the sum, but for simplicity, we choose to parameterize the error only by N, ε , and γ .

► **Definition 3.** For any positive integer h , we define the h -wise collision estimator ξ_h of $\mu = \sum_i x_i$ to be

$$\xi_h = \frac{1}{\binom{m}{h}} \sum_{i=1}^N \binom{Y_i}{h} \frac{x_i}{(\mathcal{P}(i))^h}. \quad (5)$$

We note that $\binom{Y_i}{h}$ gives the number of h -wise collisions involving the value $X_j = i$. It is straightforward to show that when $\mathcal{Q} = \mathcal{P}$, all ξ_h are unbiased estimators for μ , and that ξ_h has bias $O(h\gamma\mu_+)$ when \mathcal{Q} is pointwise γ -close to \mathcal{P} .

Individually, the estimators ξ_1, ξ_2, \dots are no better than $\xi_1 = \mu_{\text{HH}}$ in terms of bias and variance. As we will show, however, for any positive integer k , a suitable linear combination of the ξ_i can be chosen such that the coefficients of γ^j in the bias cancel out for $j < k$. The resulting estimator then has bias $\leq \gamma^k \mu_+$.

► **Definition 4.** For each positive integer k , we define the bias reducing estimator of order k , ζ_k , to be

$$\zeta_k = \sum_{h=1}^k (-1)^{h+1} \binom{k}{h} \xi_h = \sum_{h=1}^k (-1)^{h+1} \frac{\binom{k}{h}}{\binom{m}{h}} \sum_{i \in [N]} \binom{Y_i}{h} \frac{x_i}{(\mathcal{P}(i))^h}. \quad (6)$$

► **Example 5.** In order to give some intuition about the expression (6), consider the case where \mathcal{P} is the uniform distribution and $k = 2$. Define α_i to be such that $\mathcal{Q}(i) = (1 + \alpha_i)\mathcal{P}(i)$. Note that $|\alpha_i| \leq \gamma$ because \mathcal{Q} is assumed to be pointwise γ -close to uniform. By a simple calculation, we have that $\mathbb{E}[\xi_1] = \sum_{i=1}^N (1 + \alpha_i)x_i$ and $\mathbb{E}[\xi_2] = \sum_{i=1}^N (1 + \alpha_i)^2 x_i = (1 + 2\alpha_i + \alpha_i^2)x_i$. Therefore, it holds that²

$$\mathbb{E}[\zeta_2] = \mathbb{E}[2\xi_1 - \xi_2] = \sum_{i=1}^N 2(1 + \alpha_i)x_i - (1 + 2\alpha_i + \alpha_i^2)x_i = \sum_{i=1}^n (1 + \alpha_i^2)x_i \leq \mu \pm \sum_{i=1}^n \alpha_i^2 |x_i| \leq \mu \pm \gamma^2 \mu_+.$$

This proves that the bias of the estimator is at most $\gamma^2 \mu_+$. Similarly, one can show that the bias of the estimator above is $\leq \gamma^k \mu_+$. We prove this in Section 2.

► **Theorem 6** (Bias portion of Theorem 10). Suppose \mathcal{P} and \mathcal{Q} are probability distributions over $[N]$ with \mathcal{Q} pointwise γ -close to \mathcal{P} . Let ζ_k be defined as in (6). Then

$$|\mathbb{E}[\zeta_k - \mu]| \leq \gamma^k \mu_+.$$

In particular, if $x_i \geq 0$ for all i , then ζ_k has bias at most $\gamma^k \mu$.

This theorem shows that ζ_k reduces the bias to γ^k compared to the bias γ for the Hansen-Hurwitz estimator (equivalent to ζ_1). Theorem 10 additionally bounds the variance of ζ_k , which is required for our applications.

We apply Theorem 6 (or more specifically, Theorem 10) to obtain our main algorithmic results. Our goal is as follows: given sample access to some \mathcal{Q} that is pointwise γ -close to \mathcal{P} and a desired error parameter ε , estimate μ with error ε using as few samples as possible. To this end, we employ a two-stage estimation technique (see Algorithm 2). In the first stage, we use the 1-wise collision estimator ξ_1 (i.e., the Hansen-Hurwitz estimator) to obtain a coarse estimate of μ . Then, the second stage refines this estimate by applying the bias reducing estimator ζ_k with an appropriately chosen k . Specifically, we show the following:

² The first of the two inclusions is not tight in that the absolute value in $|x_i|$ is not necessary, but it becomes necessary for odd values of k .

► **Theorem 7** (Special case of Theorem 10). Define $n = \max_i 1/\mathcal{P}(i)$.³ Suppose \mathcal{Q} is pointwise γ -close to \mathcal{P} , and let Var_{HH} be defined as

$$\text{Var}_{\text{HH}} = \frac{1}{N^2} \sum_{i=1}^N \mathcal{Q}(i) \left(\frac{x_i}{\mathcal{Q}(i)} - \mu \right)^2. \tag{7}$$

That is, Var_{HH} is the variance of the Hansen-Hurwitz estimator for the mean μ_{HH}/N with sample size $m = 1$ (cf. (Equation 2)). For $\varepsilon_1, \varepsilon_2 > 0$, define $k = \lceil (\log \varepsilon_1) / \log \gamma \rceil$. Then using

$$m = O \left(\sqrt[k]{n^{k-1} \varepsilon_2^{-2} \text{Var}_{\text{HH}}} \right)$$

independent samples from \mathcal{Q} , with probability at least $2/3$, Algorithm 2 produces an estimate $\hat{\mu}$ of $\mu = \sum_i x_i$ with absolute error

$$|\mu - \hat{\mu}| \leq \varepsilon_1 \mu + \varepsilon_2.$$

To understand the complexity of this algorithm, we note that when $\mathcal{P} = \mathcal{Q}$, in order to get an error ε_2 the complexity of the Hansen-Hurwitz estimator is $\varepsilon_2^{-2} \text{Var}_{\text{HH}}$. The complexity of our algorithm can thus be seen as a weighted geometric average between the complexity of the Hansen-Hurwitz estimator, and n .

As a corollary of Theorem 7, we obtain a solution to the aforementioned problem of estimating a sum of 0-1 values.

► **Corollary 8.** Suppose \mathcal{Q} is pointwise γ -close to the uniform distribution over $[N]$ and $x_i \in [0, 1]$ for every $i \in [N]$. For any $\varepsilon > 0$ define $k = \lceil (\log \varepsilon) / \log \gamma \rceil$. Then $m = O(N^{1-1/k} \varepsilon^{-2/k})$ samples are sufficient to obtain an estimate of $\mu = \sum_i x_i$ with additive error εN with probability $2/3$.

We note that the asymptotic sample complexities in Theorem 7 and Corollary 8 are non-uniform in γ and ε . In the case of Corollary 8, for any fixed positive integer k and constant $\gamma > 0$, if $\varepsilon = \gamma^k$, then $O(N^{1-1/k})$ samples are sufficient to obtain an εN additive estimate. On the other hand, if $\gamma^{k+1} \leq \varepsilon < \gamma^k$, then our algorithm uses $O(N^{1-1/(k+1)})$ samples. Our next main result shows that this sample complexity is essentially optimal, and perhaps surprisingly, that the non-uniformity of the sample complexity is unavoidable. Specifically, we show the following lower bound.

► **Theorem 9.** Suppose $x_1, x_2, \dots, x_N \in \{0, 1\}$ and N is a parameter. Then for every positive integer k , there exists a positive constant c_k such that for $\varepsilon \leq c_k \gamma^k$, there is a sequence of distributions on $[N]$ that are γ -close to uniform such that the number of samples required to estimate μ within error εN is $\Omega(N^{1-1/(k+1)})$.

This lower bound matches the upper bound of Corollary 8 for a large range of parameters. When $\gamma \leq c_k$ (where c_k is as in the conclusion of the theorem), our algorithm has sample complexity $O(N^{1-1/(k+1)})$ for all $\varepsilon \in [\gamma^{k+1}, \gamma^k]$, while the lower bound shows $\Omega(N^{1-1/(k+1)})$ samples are necessary for all $\varepsilon \in [\gamma^{k+1}, c_k \gamma^k]$. Interestingly, these matching upper and lower bounds show that the asymptotic sample complexity is non-uniform as a let of ε for any fixed (sufficiently small) γ : for every $\varepsilon \in [\gamma^{k+1}, c_k \gamma^k]$, exactly $\Theta(N^{1-1/(k+1)})$ samples are

³ Note that $N \leq n$, where N is the size of the multiset being sampled. In the case where $\mathcal{P}(i) = \Omega(1/N)$ for all i , we have $n = \Theta(N)$. The convention of defining n in this way was previously used in [6].

necessary and sufficient, while for $\varepsilon = \gamma^k$, $\Theta(N^{1-1/k})$ samples are necessary and sufficient. Thus, as a function ε , the sample complexity contains “islands of stability” – intervals in which some perturbations of ε have no effect on the asymptotic sample complexity – while between these intervals, an arbitrarily small (constant) decrease in ε results in a polynomial (in N) increase in the sample complexity.

Discussion and Related Work

Throughout the paper, we assume that the probability distribution \mathcal{Q} from which samples are generated is pointwise close to \mathcal{P} in the sense of Definition 2. Pointwise closeness is a strictly stronger (and less commonly used) distance measure than, for example, total variation distance (i.e., L_1 distance) or other L_p distances. Nonetheless, this relatively strong assumption about the relationship between \mathcal{P} and \mathcal{Q} is necessary in order to obtain any non-trivial guarantee for estimating μ .⁴ Algorithms for generating samples with pointwise approximation guarantees have been studied in the context of sublinear time algorithms [10, 18, 8, 7, 6, 20, 9] as well as Markov chains [16, 13]. In the latter case, *uniform mixing time* gives a bound on complexity of obtaining samples with pointwise guarantees. Interestingly, Hermon [13] shows a result that can be viewed as an analogue of our lower bound for Markov chains (specifically random walks on bounded degree graphs). Namely, small perturbations in the transition probabilities of edges can result in an asymptotic increases in the uniform mixing time.

The problem of estimating the sum is well-studied in statistics. Classical estimators for non-uniform sampling probabilities are described by Hansen and Hurwitz [12] and Horvitz and Thompson [15]. Sum estimation in the related setting where we do not know the sampling probabilities but know that they are proportional to the items’ values has been studied in [17, 2].

Open problem: Sample correctors for uniform sampling

Finally, we state one interesting open problem. The concept of sample correctors from [4] assumes that we may sample from a distribution that is close to some property, and we want to be able to use it to sample from a distribution even closer to satisfying the property. It is natural to ask if one can use $o(n)$ samples from a distribution pointwise γ -close to uniform and simulate a sample with bias $o(\gamma)$.

1.2 Technical Overview

Upper Bound

The goal is to reduce the bias from γ to γ^k . Now the main observation that guides our construction is the following identity:

$$1 + (-1)^{k+1}\gamma^k = \sum_{h=1}^k (-1)^{h+1} \binom{k}{h} (1 + \gamma)^h. \quad (8)$$

This should be compared to our estimator, $\zeta_k = \sum_{h=1}^k (-1)^{h+1} \binom{k}{h} \xi_h$, which clearly mirrors the identity. The reason for this is that the probability of an h -wise collision on position $i \in [N]$ is $\mathcal{Q}(i)^h \approx (1 + \gamma)^h \mathcal{P}(i)^h$ in the worst case when $\frac{\mathcal{Q}(i)}{\mathcal{P}(i)} = 1 + \gamma$. Hence the expectation of the h -wise collision estimator, ξ_h , is approximately bounded by $(1 + \gamma)^h \mu$. This implies that

⁴ As an extreme example, consider the case where only a γ -fraction of values x_i are nonzero. If \mathcal{P} is uniform, then \mathcal{Q} can assign zero mass to the nonzero elements, and still be γ -close to \mathcal{P} with respect to total variation distance. Thus, *any* estimator will return a value that is independent of the actual sum.

when we take the expectation of our estimator then the expression reduces to Equation (8) which shows that the bias is reduced to γ^k . Here, we cheated slightly by assuming that the bias is the same for all the positions $i \in [N]$. This is of course not true, but actual calculation reduces to N instances of Equation (8).

The more delicate part of the analysis of our estimator is the bound on the variance. The main difficulty lies in rewriting $\zeta_k - \mathbb{E}[\zeta_k]$ into something manageable. We note that we can write our estimator ζ_k as

$$\zeta_k = \sum_{i \in [N]} x_i \sum_{h=1}^k (-1)^{h+1} \frac{\binom{k}{h}}{\binom{m}{h}} \frac{\binom{Y_i}{h}}{(\mathcal{P}(i))^h}.$$

We can express the frequency vector Y by random variables as follows: $Y_i = \sum_{j \in [m]} [X_j = i]$. (Here, we use $[X_j = i]$ to denote the indicator random variable of the event $X_j = i$.) This allows us to see that $\binom{Y_i}{h} = \sum_{\substack{I \subseteq [m] \\ |I|=h}} \prod_{j \in I} [X_j = i]$. If we now define the polynomials P_i by

$$P_i(\beta_1, \dots, \beta_m) = \sum_{h=1}^k (-1)^{h+1} \frac{\binom{k}{h}}{\binom{m}{h}} \frac{1}{(\mathcal{P}(i))^h} \sum_{\substack{I \subseteq [m] \\ |I|=h}} \prod_{j \in I} \beta_j,$$

then we can write our estimator as $\zeta_k = \sum_{i \in [N]} x_i P_i([X_1 = i], \dots, [X_m = i])$. We observe that P has degree 1 in each variable so $\mathbb{E}[P_i([X_1 = i], \dots, [X_m = i])] = P_i(\mathcal{Q}(i), \dots, \mathcal{Q}(i))$. Furthermore, it seems reasonable that there should exist polynomials Q_i which have degree 1 in each variable and satisfy $P_i([X_1 = i], \dots, [X_m = i]) - P_i(\mathcal{Q}(i), \dots, \mathcal{Q}(i)) = Q_i([X_1 = i] - \mathcal{Q}(i), \dots, [X_m = i] - \mathcal{Q}(i))$. This will help us in understanding the variance of our final estimator ζ_k by decomposing into variances of the simpler events $[X_1 = i]$. We show that Q_i exist and are defined as follows

$$Q_i(\beta_1, \dots, \beta_m) = (-1)^{k+1} \sum_{h=1}^k \frac{\binom{k}{h}}{\binom{m}{h}} \frac{\gamma_i^{k-h}}{(\mathcal{P}(i))^h} \sum_{\substack{I \subseteq [m] \\ |I|=h}} \prod_{j \in I} \beta_j$$

So we get that $\zeta_k - \mathbb{E}[\zeta_k] = \sum_{i \in [N]} x_i Q_i([X_1 = i] - \mathcal{Q}(i), \dots, [X_m = i] - \mathcal{Q}(i))$. Since $[X_1 = i] - \mathcal{Q}(i)$ is a zero mean variable and Q_i has degree 1 in each variable, it becomes easy to calculate the variance.

A technical detail that we have not touched upon yet is that if $k \geq 2$ then $\text{Var}[\zeta_k]$ becomes very large if $\mu^2 \gg \text{Var}[\mu_{\text{HH}}]$. The reason is that $\text{Var}[\zeta_k]$ depends on $\sum_{i \in [N]} \mathcal{P}(i) \left(\frac{x_i}{\mathcal{P}(i)}\right)^2 = \text{Var}[\mu_{\text{HH}}] + \mu^2$. Thus, if $\mu^2 \gg \text{Var}[\mu_{\text{HH}}]$ then our variance becomes much larger which is a problem. Now imagine that we already have an estimate, W , of μ . We then set $\bar{x}_i = x_i - \mathcal{P}(i)W$ and make a new estimator $\bar{\zeta}_k$ that uses \bar{x}_i instead of x_i . Then $\bar{\zeta}_k$ will estimate $\mu - W$ so $\bar{\zeta}_k + W$ will be an estimator of μ . The trick is that $\text{Var}[\bar{\zeta}_k]$ depends on $\text{Var}[\mu_{\text{HH}}] + (\mu - W)^2$ so if $W = \mu + O(\sqrt{\text{Var}[\mu_{\text{HH}}]})$ then we will have control over the variance. We can get such estimate, W , by using our estimator for $k = 1$ where there are no issues with variance (i.e., the standard deviation of $W - \mu$ only depends on $\sqrt{\text{Var}[\mu_{\text{HH}}]}$ rather than on μ).

Lower Bound

At a high level, our strategy is to define a reduction from the problem of distinguishing two distributions D_1 and D_2 to the problem of estimating $\mu = \sum_i x_i$. More concretely, for each fixed positive integer k , we define distributions D_1 and D_2 with support sizes n_1 and n_2 , respectively, such that:

1. $n_1 = (1 + (\Theta(\gamma))^k)n_2$,
2. D_1 and D_2 are both pointwise γ -close to uniform, and
3. D_1 and D_2 have identical p^{th} frequency moments for $p = 1, 2, \dots, k$.⁵

We describe a reduction showing that for $N = n_1 + n_2$, \mathcal{P} uniform over $[N]$, and $x_i \in \{0, 1\}$ for all $i \in [N]$, any algorithm that distinguishes $\mu = n_1$ from $\mu = n_2$ can also distinguish D_1 from D_2 . We then apply a framework of Raskhodnikova et al. [19], which implies that distinguishing any two distributions whose first k frequency moments are equal requires $\Omega(n^{1-1/(k+1)})$ samples. One difference between our setting and that of [19] is that in our setting, for each sample i we are also given x_i , whereas [19] focuses on support size estimation. However, since the x_i 's are 0 or 1-valued, estimating the sum requires us to estimate either the number of x_i 's which are 0 or the number of x_i 's which are 1. We can apply this observation to reduce the problem to finding two nearly uniform distributions (i.e., both are pointwise γ -close to uniform) that differ in support size by a multiplicative $1 \pm \Theta(\gamma)^k$ factor, yet match in the first k moments.

To do this, we use a combinatorial construction that is inspired by a related lower bound in [6]. For simplicity, we assume that the probability of sampling any fixed item lies in $\{\frac{1}{n}, \frac{1+\gamma}{n}, \dots, \frac{1+k\gamma}{n}\}$ (while these distributions would only be $k \cdot \gamma$ -close to uniform, we can replace γ with γ/k). For the distribution D_1 , we assume the number of elements with probability $\frac{1+i\gamma}{n}$ is a_i , and for the distribution D_2 the number of elements with probability $\frac{1+i\gamma}{n}$ is b_i . Then, our goal is for the support sizes of D_1 and D_2 (which are $\sum a_i, \sum b_i$, respectively) to differ significantly but for the first k moments to match. This means $\sum a_i$ and $\sum b_i$ differ significantly, but $\sum a_i(1+i\gamma)^\ell = \sum b_i(1+i\gamma)^\ell$. If we define $c_i := a_i - b_i$ and think of $(1+i\gamma)^\ell$ as a degree ℓ polynomial $P(i)$, we want $\sum c_i P(i) = 0$ but $\sum c_i$ to be large (roughly $\gamma^k \cdot n$). Finally, we need to make sure that $\sum a_i, \sum b_i$ are both $\Theta(n)$.

To determine the values of each c_i , we utilize the observation that for any polynomial $P(x)$ of degree less than k , the successive differences, i.e., $P(x) - P(x-1)$ is a polynomial of degree less than $k-1$. We can repeatedly take successive differences k times to get a linear combination of $P(0), P(1), \dots, P(k)$ that equals 0 for any polynomial P of degree less than k . We can therefore set $c_i := a_i - b_i$ to be these linear coefficients. Unfortunately, we will have $\sum c_i = 0$ as well. Instead, we replace c_i with $c'_i := c_i/(1+i\gamma)$, so that $\sum c'_i \cdot (1+i\gamma)^\ell = \sum c_i \cdot (1+i\gamma)^{\ell-1} = 0$ for all $1 \leq \ell \leq k$. However, $\sum c'_i = \sum c_i/(1+i\gamma)$ is not expressible as $\sum c_i P(i)$ for a polynomial P of low degree, and will in fact be nonzero, which is exactly what we want. We scale the c'_i terms so that $\sum c'_i = \gamma^k \cdot n$. If we write $a_i = \max(c'_i, 0)$ and $b_i = \max(-c'_i, 0)$, then every a_i and b_i is nonnegative but $a_i - b_i = c'_i$. One can show via some careful combinatorics that after this scaling, $\sum a_i$ and $\sum b_i$ are both $\Theta(n)$, as desired.

2 Sum Estimation

We now give the algorithm for the sum estimation problem. We then state our main result (Theorem 10) as well as the special case for estimating the sum of 0-1 values (Corollary 8) that we discussed in the introduction. We then state and prove Lemma 12 which we then use to prove Theorem 10.

We now state our main theorem. Note that this implies, by a simple substitution, the simpler version mentioned in the introduction as Theorem 7.

⁵ Recall that the p^{th} frequency moment of a distribution D is $\sum_x (D(x))^p$.

Algorithm 1 *EstimateSum*(m, k, W).

- 1 $(X_j)_{j \in [m]} \leftarrow$ take m samples from the distribution \mathcal{Q}
 - 2 For every $i \in [N]$, let Y_i denote the number of times value i was sampled
 $(Y_i = \sum_{j \in [m]} [X_j = i])$
 - 3 For every $h \in [k]$, let ξ_h be the h -wise collision estimator
 $(\xi_h = \frac{1}{\binom{m}{h}} \sum_{i=1}^N \binom{Y_i}{h} \frac{x_i - \mathcal{P}(i)W}{(\mathcal{P}(i))^h})$
 - 4 **return** $W + \sum_{h=1}^k (-1)^{h+1} \binom{k}{h} \xi_h$
-

Algorithm 2 *ImprovedEstimateSum*(m, t, k).

- 1 $W \leftarrow$ *EstimateSum*($t, 1, 0$)
 - 2 **return** *EstimateSum*(m, k, W)
-

► **Theorem 10.** Define $n = \max_i 1/\mathcal{P}(i)$. Suppose \mathcal{Q} is pointwise γ -close to \mathcal{P} , and let Var_{HH} be the variance of μ_{HH}/N defined in Equation (7). For $\varepsilon_1, \varepsilon_2 > 0$, define $k = \lceil (\lg \varepsilon_1) / \lg \gamma \rceil$. Then using

$$m = O\left(\sqrt[k]{n^{k-1} \varepsilon_2^{-2} \text{Var}_{\text{HH}}}\right)$$

independent samples from \mathcal{Q} , with probability at least $2/3$, Algorithm 2 produces an estimate $\hat{\mu}$ of $\mu = \sum_i x_i$ with absolute error

$$|\mu - \hat{\mu}| \leq \varepsilon_1(1 + \gamma) \mathbb{E}_{X \sim \mathcal{P}} [|\mathcal{P}(X)^{-1} x_X - \mu|] + \varepsilon_2$$

This theorem implies in a straightforward way a solution to the problem of estimating the sum of 0-1 values that we discussed in the introduction.

► **Corollary 11.** Suppose \mathcal{Q} is pointwise γ -close to the uniform distribution over $[N]$ and $x_i \in \{0, 1\}$ for every $i \in [N]$. For any $\varepsilon > 0$ define $k = \lceil (\log \varepsilon) / \log \gamma \rceil$. Then $m = O(n^{1-1/k} \varepsilon^{-2/k})$ samples are sufficient to obtain an estimate of $\mu = \sum_i x_i$ with additive error $\varepsilon(\mu + \sqrt{\mu N})$ with probability $2/3$.

Before we can prove our main result, we need the following lemma.

► **Lemma 12** (Analysis of *EstimateSum*($m, k, 0$)). Define $n = \max_i 1/\mathcal{P}(i)$. Let $(x_i)_{i \in [N]}$ be a sequence of numbers and define $\mu = \sum_{i \in [N]} x_i$. Let \mathcal{P} be a probability distribution over $[N]$, and let \mathcal{Q} be another probability distribution over $[N]$ that is pointwise γ -close to \mathcal{P} .

Consider a sequence $(X_j)_{j \in [m]}$ of independent random variables both with distribution \mathcal{Q} . Define $Y_i = \sum_{j \in [m]} [X_j = i]$ for every $i \in [N]$. Define

$$\zeta_k = \sum_{h=1}^k (-1)^{h+1} \frac{\binom{k}{h}}{\binom{m}{h}} \sum_{i \in [N]} \binom{Y_i}{h} \mathcal{P}(i)^{-h} x_i.$$

Then for $k \geq 2$, we get that

$$|\mathbb{E}[\zeta_k] - \mu| \leq \gamma^k \sum_{i \in [N]} |x_i| \tag{9}$$

$$\text{Var}[\zeta_k] \leq \max \left\{ 2(1 + \gamma) \gamma^{2k-2} k^2 \frac{\sum_{i \in [N]} \mathcal{P}(i)^{-1} x_i^2}{m}, 2^k (1 + \gamma)^k k^{3k} \frac{n^{k-1} \sum_{i \in [N]} \mathcal{P}(i)^{-1} x_i^2}{m^k} \right\}, \tag{10}$$

62:10 Bias Reduction for Sum Estimation

and for $k = 1$, we get that

$$|\mathbb{E}[\zeta_1] - \mu| \leq \gamma \sum_{i \in [N]} |x_i - \mathcal{P}(i)\mu| \quad (11)$$

$$\text{Var}[\zeta_1] \leq (1 + \gamma) \frac{\sum_{i \in [N]} \mathcal{P}(i)^{-1} (x_i - \mathcal{P}(i)\mu)^2}{m} \quad (12)$$

We are now ready to prove the main theorem.

Proof of Theorem 10. Let W be an estimate of μ using $\text{EstimateSum}(t, 1, 0)$ where $t = O(1 + \gamma^{2k} \varepsilon_2^{-2} \text{Var}_{\text{HH}})$. Using Lemma 12 we get an estimate of the bias and the variance. Now by Chebyshev's inequality, we easily get that $|W - \mu| \leq \gamma \sum_{i \in [N]} \mathcal{P}(i) |\mathcal{P}(i)^{-1} x_i - \mu| + \max \left\{ \varepsilon_2 / (2\gamma^k), \sqrt{\sum_{i \in [N]} \mathcal{P}(i) (\mathcal{P}(i)^{-1} x_i - \mu)^2} \right\}$ with probability $5/6$. We note that $\sum_{i \in [N]} \mathcal{P}(i) |\mathcal{P}(i)^{-1} x_i - \mu| \leq \sqrt{\sum_{i \in [N]} \mathcal{P}(i) (\mathcal{P}(i)^{-1} x_i - \mu)^2}$ since p -norms are increasing. So we also get that $|W - \mu| \leq O(\sqrt{\sum_{i \in [N]} \mathcal{P}(i) (\mathcal{P}(i)^{-1} x_i - \mu)^2})$

Now we calculate ζ using $\text{EstimateSum}(m, k, W)$ with $m = O\left(\sqrt{n^{k-1} \varepsilon_2^{-2} \text{Var}_{\text{HH}}}\right)$ so ζ corresponds to $\text{ImprovedEstimateSum}(m, t, k)$. We now note that by Lemma 12,

$$\begin{aligned} |(\mathbb{E}[\zeta] - W) - (\mu - W)| &\leq \gamma^k \sum_{i \in [N]} \mathcal{P}(i) |\mathcal{P}(i)^{-1} x_i - W| \\ &\leq \gamma^k |W - \mu| + \gamma^k \sum_{i \in [N]} \mathcal{P}(i) |\mathcal{P}(i)^{-1} x_i - \mu|, \end{aligned}$$

and

$$\begin{aligned} \text{Var}[\zeta] &\leq \max \left\{ 2(1 + \gamma) \gamma^{2k-2} k^2 \frac{\sum_{i \in [N]} \mathcal{P}(i) (\mathcal{P}(i)^{-1} x_i - W)^2}{m}, \right. \\ &\quad \left. 2^k (1 + \gamma)^k k^{3k} \frac{n^{k-1} \sum_{i \in [N]} \mathcal{P}(i) (\mathcal{P}(i)^{-1} x_i - W)^2}{m^k} \right\} \\ &= \max \left\{ 2(1 + \gamma) \gamma^{2k-2} k^2 \frac{\sum_{i \in [N]} \mathcal{P}(i) (\mathcal{P}(i)^{-1} x_i - \mu)^2 + (W - \mu)^2}{m}, \right. \\ &\quad \left. 2^k (1 + \gamma)^k k^{3k} \frac{n^{k-1} \left(\sum_{i \in [N]} \mathcal{P}(i) (\mathcal{P}(i)^{-1} x_i - W)^2 + (W - \mu)^2 \right)}{m^k} \right\} \end{aligned}$$

Assuming that

$$|W - \mu| = \gamma \sum_{i \in [N]} \mathcal{P}(i) |\mathcal{P}(i)^{-1} x_i - \mu| + \max \left\{ \varepsilon_1 / (2\gamma^k), \sqrt{\sum_{i \in [N]} \mathcal{P}(i) (\mathcal{P}(i)^{-1} x_i - \mu)^2} \right\},$$

we get that $|(\mathbb{E}[\zeta] - W) - (\mu - W)| \leq \gamma^k (1 + \gamma) \sum_{i \in [N]} \mathcal{P}(i) |\mathcal{P}(i)^{-1} x_i - \mu| + \varepsilon_2 / 2$. Using that $|W - \mu| \leq O(\sqrt{\sum_{i \in [N]} \mathcal{P}(i) (\mathcal{P}(i)^{-1} x_i - \mu)^2})$ we can then use Chebyshev's inequality to conclude that with probability $5/6$

$$|\zeta - \mathbb{E}[\zeta]| \leq \varepsilon_2 / 2$$

So all in all, with probability at least $2/3$ we that $|\zeta - \mu| \leq \varepsilon_1 (1 + \gamma) \sum_{i \in [N]} \mathcal{P}(i) |\mathcal{P}(i)^{-1} x_i - \mu| + \varepsilon_2$. \blacktriangleleft

3 Lower Bound

In this section, we prove that for a range of parameters – specifically in the setting of Corollary 11 – the sample complexity of our sum estimation algorithm (Algorithm 2) is asymptotically tight.

► **Theorem 13.** *Let k be a positive integer and let $\varepsilon < \gamma < 1$ be positive numbers. Suppose A is an algorithm such that for any $v \in \{0, 1\}^N$ and any distribution \mathcal{P} over $[N]$ that is pointwise γ -close to uniform, A uses samples from \mathcal{P} and returns an estimate of $\|v\|_1 = \sum_i v_i$ within additive error εN with probability $2/3$. Then there exists c_k with $0 < c_k < 1$ such that if $\varepsilon \leq c_k \gamma^k$ then A requires $\Omega(N^{1-1/(k+1)})$ samples.*

While at a first glance this result might seem contradictory to our upper bound (specifically, Corollary 8), it actually reveals the following interesting phenomena. Notice that in our upper bound, the complexity depends on $k = \lceil (\log \varepsilon) / \log \gamma \rceil$, so that, e.g., for $\varepsilon = \gamma^k$, the complexity is $O(N^{1-\frac{1}{k}})$. Once ε becomes slightly smaller, i.e., $\varepsilon = c\gamma^k$ for c satisfying $\gamma \leq c < 1$, the complexity of our algorithm abruptly jumps to $O(N^{1-\frac{1}{k+1}})$. The lower bound implies that this increase in complexity is unavoidable for sufficiently small c . That is, Theorem 13 states that there exists a (sufficiently small) constant c_k , such that indeed once $\varepsilon = c_k \cdot \gamma^k$, the required number of samples is $\Omega(N^{1-\frac{1}{k+1}})$. Interestingly, for all ε satisfying $\gamma^{k+1} \leq \varepsilon \leq c_k \gamma^k$, the asymptotic complexity of sum estimation is the same; the complexity only varies for ε satisfying $c_k \gamma^k \leq \varepsilon \leq \gamma^k$. Our matching upper and lower bounds demonstrate that the sample complexity’s non-uniform dependence on ε is not an artifact, but captures the true complexity of the problem (up to the dependency on $\gamma^{k/2}$ in the numerator of the upper bound). Note that if the conclusion of Theorem 13 held *every* $c < 1$, then this would capture the right dependency on n for *all* possible ranges of γ . Since we only prove the theorem for a sufficiently small c_k , it might be the case that for values ε that are not too much smaller than γ^k , the optimal dependency on N is lower than our stated upper bound. Nonetheless, our upper and lower bounds match (up to constant factors) for all ε satisfying $\gamma^{k+1} \leq \varepsilon \leq c_k \gamma^k$.

As described in Section 1.2, the main technical ingredient in the proof of the theorem is in describing two distributions D_1 and D_2 over ranges $[n_1], [n_2]$, respectively, such that D_1 and D_2 are pointwise γ -close to uniform, $n_1 = (1 + \Theta(\gamma)^k)n_2$, and D_1 and D_2 have matching frequency moments 1 through k . Given these distributions we rely on the framework for proving lower bounds by Raskhodnikova et al. [19], which states that any uniform algorithm that distinguishes two random variables with matching frequency moments 1 through k must perform $\Omega(N^{1-1/(k+1)})$ many samples.

In order to simplify our construction and its analysis, we prove the lower bound for *uniform algorithms*. Here, a uniform algorithm is an algorithm whose output depends only on the “collision statistics” of the samples – i.e., the number of collisions involving each sample, and not the identities of the samples themselves.

► **Definition 14** (Uniform algorithm. Definition 3.2 in [19]). *An algorithm is uniform if it samples indices i_1, \dots, i_m independently with replacement and its output is uniquely determined by (i) the value of the items x_{i_j} and (ii) the set of collisions, where two indices j and j' collide if $i_j = i_{j'}$.*

In particular, a uniform algorithm’s output does *not* depend on the sampled indices themselves. The following lemma asserts that our restriction to uniform algorithms is without loss of generality.

► **Lemma 15** (cf. Theorem 11.12 in [11]). *Suppose there exists an algorithm A such that for any $v \in \{0, 1\}^N$ A uses samples from \mathcal{P} and returns an estimate of $\|v\|_1 = \sum_i v_i$ within additive error εN with probability $2/3$ using s samples in expectation. Then there exists a uniform algorithm A' that achieves the same approximation guarantee using s samples in expectation.*

The proof of Lemma 15 is essentially the same as that of Goldreich’s Theorem 11.12 in [11]. The key idea is that $\sum_i v_i$ is a “label invariant” in the sense that it is unaffected by any permutation of the indices of v . Thus, given an algorithm A as in the hypothesis of Lemma 15, we can obtain uniform algorithm A' with the same approximation guarantee by simply choosing a uniformly random permutation of the indices of v , then using the permuted indices of v as inputs for A . See Theorem 11.12 in [11] for details.

Finally, our lower bound argument requires the following result that is a direct consequence of the work of Raskhodnikova et al. [19].

► **Theorem 16** (Consequence of Lemma 5.3 and Corollary 5.7 from [19]). *Let D_1 and D_2 be distributions over positive integers $b_1 < \dots < b_\ell$, that have matching frequency moments 1 through k . Then for any uniform algorithm A with sample complexity s that distinguishes D_1 and D_2 with high constant probability, $s = \Omega(n^{1-1/(k+1)})$.*

Our main argument for the lower bound applies Theorem 16 in conjunction with a reduction from distinguishing D_1 and D_2 to sum estimation. The main technical ingredient is the following lemma, which asserts the existence of suitable distributions D_1 and D_2 .

► **Lemma 17.** *For every positive integer k and sufficiently large integer n , there exist two distributions D_1, D_2 over $[n_1]$ and $[n_2]$ (respectively) satisfying $n_1 = (1 + \Theta(\gamma)^k)n$, and $n_2 = n$ such that $p_{D_j}(i) \in (1 \pm \gamma)\frac{1}{n}$ for $j \in \{1, 2\}$ and the following holds. For all $\ell \in \{1, 2, \dots, k\}$, it holds that*

$$\sum_{i=1}^{n_1} (p_{D_1}(i))^\ell = \sum_{i=1}^{n_2} (p_{D_2}(i))^\ell.$$

In particular, there exists an absolute constant c_k such that for sufficiently large n , $n_2 \geq (1 + c_k \gamma^k)n_1$ and the above conclusion holds.

Before proving the lemma, we show how the lemma implies Theorem 13.

Proof of Theorem 13. The theorem follows from Theorem 16 together with Lemma 17. Let $N = n_1 + n_2$, where $n_1 = (1 + \Theta(\gamma)^k)n_2$ as in the conclusion of Lemma 17, and consider distinguishing between two possible outcomes \mathcal{O}_1 and \mathcal{O}_2 .

In the first outcome \mathcal{O}_1 , let $S = \{1, 2, \dots, n_1\} \subseteq [N]$ and $T = \{t_1, \dots, t_{n_2}\} := [N] \setminus S$. The distribution \mathcal{Q} will be as follows. With exactly $1/2$ probability, we choose S : if so, we then choose a sample $i \sim D_1$, which will be in $[n_1]$, and output i . Otherwise, we choose T : we then choose a sample $i \sim D_2$, which will be in $[n_2]$, and then output t_i . Here, D_1, D_2 are the distributions from Lemma 17. Finally, we let $v_i = 1$ if $i \in S$ and $v_i = 0$ if $i \in T$.

The second outcome \mathcal{O}_2 is similar but “flipped”. Now, we let $S = \{1, 2, \dots, n_2\} \subseteq [N]$, and $T = \{t_1, \dots, t_{n_1}\} := [N] \setminus S$. With exactly $1/2$ probability, we choose S : if so, we then choose a sample $i \sim D_2$, and output i . Otherwise, we choose T : we then choose a sample $i \sim D_1$, and then output t_i . Finally, we let $v_i = 1$ if $i \in S$ and $v_i = 0$ if $i \in T$.

Under \mathcal{O}_1 , we have that $\sum v_i$ always equals n_1 , whereas under \mathcal{O}_2 , we have that $\sum v_i$ is always n_2 . In addition, since both n_1 and n_2 are $\frac{N}{2} \cdot (1 \pm O(\gamma))$, and since D_1 and D_2 are γ -pointwise close to uniform, the distribution \mathcal{Q} that we sample from in either case is $O(\gamma)$ -pointwise close to uniform. So, we may assume that \mathcal{P} is uniform over $[N]$ in either case.

Now, assume that there exists a uniform algorithm⁶ A that draws samples (i, v_i) either from outcome \mathcal{O}_1 or outcome \mathcal{O}_2 and with probability at least $2/3$, computes an estimate of $\sum_i v_i$ up to additive error $c_k \gamma^k N/5$, where c_k is as in the second conclusion of Lemma 17. Observe that when the error bound on A is satisfied (which occurs with probability at least $2/3$), A 's output distinguishes scenarios \mathcal{O}_1 and \mathcal{O}_2 .

Finally, we observe that distinguishing \mathcal{O}_1 from \mathcal{O}_2 is sufficient to distinguish the distributions D_1 and D_2 . Indeed, under scenario \mathcal{O}_1 , the 1-values and sampled from D_1 , while the 0-values are sampled from D_2 , while the roles are reversed in \mathcal{O}_2 . Thus, the output of A suffices to distinguish D_1 and D_2 . Since A uses s samples in expectation, Theorem 16 and Lemma 17 imply that $s = \Omega(n^{1-1/(k+1)})$, as desired. ◀

We now conclude by proving our main technical lemma.

Proof of Lemma 17. First, we note that

$$\begin{aligned} \sum_{i=0}^k (-1)^i \binom{k}{i} \binom{i}{r} &= \sum_{i=r}^k (-1)^i \binom{k}{i} \binom{i}{r} = \sum_{i=r}^k (-1)^i \cdot \frac{k!}{(k-i)!(i-r)!r!} \\ &= \sum_{i=r}^k (-1)^i \cdot \binom{k}{r} \binom{k-r}{i-r} = (-1)^r \binom{k}{r} \cdot \sum_{j=0}^{k-r} (-1)^j \binom{k-r}{j}. \end{aligned}$$

The last line follows by setting $j = i - r$. Now, note that the summation in the last line equals $(1 - 1)^{k-r} = 0$ if $k > r$, and equals 1 if $k = r$. So, this means that $\sum_{i=0}^k (-1)^i \binom{k}{i} \binom{i}{r} = 0$ for all $0 \leq r < k$.

Next, note that $\binom{i}{r} = \frac{i(i-1)\cdots(i-r+1)}{r!}$ for all integers $i \geq 0$. This is a degree- r polynomial in i . From this observation, it is well-known that every degree at most $k - 1$ polynomial in i can be written as a linear combination of $\binom{i}{0}, \dots, \binom{i}{k-1}$. Therefore, for any polynomial P of degree at most $k - 1$, $\sum_{i=0}^k (-1)^i \binom{k}{i} \cdot P(i) = 0$.

Now, we let the distribution D_1 have exactly a $\frac{\binom{k}{i}}{2^{k-1}}$ fraction of its mass consisting of items each with probability $(1 + \frac{\gamma \cdot i}{k}) \cdot \frac{1}{n_0}$, for each *even* integer $0 \leq i \leq k$. Here, n_0 will be an integer chosen later. Note this means it must have $n_0 \cdot \frac{\binom{k}{i}}{2^{k-1} \cdot (1 + \frac{\gamma \cdot i}{k})}$ points with mass $(1 + \frac{\gamma \cdot i}{k}) \cdot \frac{1}{n_0}$ for each *even* integer $0 \leq i \leq k$. Likewise, we let the distribution D_2 have exactly a $\frac{\binom{k}{i}}{2^{k-1}}$ fraction of its mass consisting of items each with probability $(1 + \frac{\gamma \cdot i}{k}) \cdot \frac{1}{n_0}$, for each *odd* integer $0 \leq i \leq k$. Note that the total fraction of mass for both D_1 and D_2 is clearly 1.

First, we note that for any $1 \leq \ell \leq k$,

$$\sum_{i=1}^{n_1} (p_{D_1}(i))^\ell - \sum_{i=1}^{n_1} (p_{D_2}(i))^\ell \tag{13}$$

$$= \sum_{\substack{i=0 \\ i \text{ even}}}^k n_0 \cdot \frac{\binom{k}{i}}{2^{k-1} \cdot (1 + \frac{\gamma \cdot i}{k})} \cdot \left(\left(1 + \frac{\gamma \cdot i}{k} \right) \cdot \frac{1}{n_0} \right)^\ell \tag{14}$$

$$- \sum_{\substack{i=0 \\ i \text{ odd}}}^k n_0 \cdot \frac{\binom{k}{i}}{2^{k-1} \cdot (1 + \frac{\gamma \cdot i}{k})} \cdot \left(\left(1 + \frac{\gamma \cdot i}{k} \right) \cdot \frac{1}{n_0} \right)^\ell \tag{15}$$

⁶ Again, the assumption that A is uniform is without loss of generality by Lemma 15. For our construction, however, the two scenarios \mathcal{O}_1 and \mathcal{O}_2 can be distinguished by a non-uniform algorithm using $O(\gamma^k)$ samples. Indeed, the two scenarios are distinguished by seeing any value v_i with $n_2 < i \leq n_1$. Following the proof of Lemma 15 (cf. Theorem 11.12 in [11]), the scenarios are indistinguishable to even a non-uniform algorithm we we replace S and T with randomly chosen complementary subsets of $[N]$.

62:14 Bias Reduction for Sum Estimation

$$= \frac{1}{n_0^{\ell-1} \cdot 2^{k-1}} \cdot \sum_{i=0}^k (-1)^i \binom{k}{i} \cdot \left(1 + \frac{\gamma \cdot i}{k}\right)^{\ell-1}. \quad (16)$$

By letting $P(i)$ be the polynomial $\left(1 + \frac{\gamma \cdot i}{k}\right)^{\ell-1}$, we have that $P(i)$ has degree at most $k-1$, so this equals 0, as desired.

Finally, we look at the difference $n_1 - n_2$, i.e., the difference in support size between D_1 and D_2 . This simply equals

$$\sum_{\substack{i=0 \\ i \text{ even}}}^k n_0 \cdot \frac{\binom{k}{i}}{2^{k-1} \cdot \left(1 + \frac{\gamma \cdot i}{k}\right)} - \sum_{\substack{i=0 \\ i \text{ odd}}}^k n_0 \cdot \frac{\binom{k}{i}}{2^{k-1} \cdot \left(1 + \frac{\gamma \cdot i}{k}\right)} = \frac{n_0}{2^{k-1}} \cdot \sum_{i=0}^k (-1)^i \cdot \frac{\binom{k}{i}}{1 + \frac{\gamma}{k} \cdot i}. \quad (17)$$

We now inductively prove (by inducting on $k \geq 1$) that $\sum_{i=0}^k (-1)^i \cdot \frac{\binom{k}{i}}{a + \gamma \cdot i} = \frac{k! \cdot \gamma^k}{a(a+\gamma) \cdots (a+k\gamma)}$ for any real numbers a, γ . For $k = 1$, we have that $\sum_{i=0}^1 (-1)^i \cdot \frac{\binom{1}{i}}{a + \gamma \cdot i} = \frac{1}{a} - \frac{1}{a+\gamma} = \frac{\gamma}{a(a+\gamma)}$. For general k , we can write

$$\sum_{i=0}^k (-1)^i \cdot \frac{\binom{k}{i}}{a + \gamma \cdot i} = \sum_{i=0}^k (-1)^i \cdot \frac{\binom{k-1}{i-1} + \binom{k-1}{i}}{a + \gamma \cdot i} \quad (18)$$

$$= \sum_{i=0}^{k-1} (-1)^i \cdot \frac{\binom{k-1}{i}}{a + \gamma \cdot i} + \sum_{i=1}^k (-1)^i \cdot \frac{\binom{k-1}{i-1}}{a + \gamma \cdot i} \quad (19)$$

$$= \sum_{i=0}^{k-1} (-1)^i \cdot \frac{\binom{k-1}{i}}{a + \gamma \cdot i} - \sum_{j=0}^{k-1} (-1)^j \cdot \frac{\binom{k-1}{j}}{(a + \gamma) + \gamma \cdot j}, \quad (20)$$

where we have set $j = i - 1$. We can now use the inductive hypothesis on $k - 1$ to obtain that this equals

$$\frac{(k-1)! \cdot \gamma^{k-1}}{a(a+\gamma) \cdots (a+(k-1)\gamma)} - \frac{(k-1)! \cdot \gamma^{k-1}}{(a+\gamma) \cdots (a+(k-1)\gamma)(a+k\gamma)} \quad (21)$$

$$= (k-1)! \cdot \gamma^{k-1} \cdot \frac{(a+k\gamma) - a}{a(a+\gamma) \cdots (a+(k-1)\gamma)(a+k\gamma)} \quad (22)$$

$$= k! \cdot \gamma^k \cdot \frac{1}{a(a+\gamma) \cdots (a+(k-1)\gamma)(a+k\gamma)}. \quad (23)$$

Therefore, by setting $a = 1$ and replacing γ with $\gamma' = \gamma/k$, we have that the difference in support size between D_1 and D_2 is

$$\frac{n_0}{2^{k-1}} \cdot \frac{k!}{k^k} \cdot \gamma^k \cdot \frac{1}{(1 + \gamma/k)(1 + 2\gamma/k) \cdots (1 + \gamma)}.$$

Assuming that $\gamma \leq 1/2$, we can apply Stirling's approximation to obtain that this difference is $n_0 \cdot (\gamma/\Theta(1))^k$.

To finish, we will set n_0 appropriately. Note that we wish for D_2 to have support size exactly n . However, all of the points in D_2 has mass between $\frac{1}{n_0}$ and $\frac{1+\gamma}{n_0}$, which means that the support size n_2 must be between $\frac{n_0}{1+\gamma}$ and n_0 . So, we can first set n_0 , and then choose $n = n_2$ to be $n_0 \cdot \sum_{i=0, i \text{ odd}}^k \frac{\binom{k}{i}}{2^{k-1} \cdot (1 + \frac{\gamma \cdot i}{k})}$, which is in the range $\left[\frac{n_0}{1+\gamma}, n_0\right]$, and n_1 to be $n_0 \cdot \sum_{i=0, i \text{ even}}^k \frac{\binom{k}{i}}{2^{k-1} \cdot (1 + \frac{\gamma \cdot i}{k})}$. Both n_1 and $n = n_2$ are in the range $\left[\frac{n_0}{1+\gamma}, n_0\right]$. Indeed, we will have that $\sum_{i=1}^{n_1} (p_{D_1}(i))^\ell = \sum_{i=1}^{n_2} (p_{D_2}(i))^\ell$, and $n_1 - n_2 = \Theta(\gamma)^k \cdot n_0 = \Theta(\gamma)^k \cdot n$. In

addition, because all of the values $p_{D_j}(i)$ are in the range $\left[\frac{1}{n_0}, \frac{1+\gamma}{n_0}\right]$ for both $j = 1$ and $j = 2$, this means that they are also in the range $\left[\frac{1-\gamma}{n}, \frac{1+\gamma}{n}\right]$, as desired. This completes the proof. ◀

References

- 1 Petra Berenbrink, Bruce Krayenhoff, and Frederik Mallmann-Trenn. Estimating the number of connected components in sublinear time. *Information Processing Letters*, 114(11):639–642, 2014.
- 2 Lorenzo Beretta and Jakub Tětek. Better sum estimation via weighted sampling. In *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2303–2338. SIAM, 2022.
- 3 Clément Canonne and Ronitt Rubinfeld. Testing probability distributions underlying aggregated data. In *International Colloquium on Automata, Languages, and Programming*, pages 283–295. Springer, 2014.
- 4 Clément L. Canonne, Themis Gouleakis, and Ronitt Rubinfeld. Sampling correctors. *SIAM J. Comput.*, 47(4):1373–1423, 2018. doi:10.1137/16M1076666.
- 5 Edith Cohen, Nick Duffield, Haim Kaplan, Carstent Lund, and Mikkel Thorup. Algorithms and estimators for summarization of unaggregated data streams. *Journal of Computer and System Sciences*, 80(7):1214–1244, 2014.
- 6 Talya Eden, Piotr Indyk, Shyam Narayanan, Ronitt Rubinfeld, Sandeep Silwal, and Tal Wagner. Learning-based support estimation in sublinear time. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL: <https://openreview.net/forum?id=t1lovEHA3YS>.
- 7 Talya Eden, Saleet Mossel, and Ronitt Rubinfeld. Sampling multiple edges efficiently. In Mary Wootters and Laura Sanità, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2021, August 16-18, 2021, University of Washington, Seattle, Washington, USA (Virtual Conference)*, volume 207 of *LIPICs*, pages 51:1–51:15. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021. doi:10.4230/LIPICs.APPROX/RANDOM.2021.51.
- 8 Talya Eden, Dana Ron, and Will Rosenbaum. The arboricity captures the complexity of sampling edges. In Christel Baier, Ioannis Chatzigiannakis, Paola Flocchini, and Stefano Leonardi, editors, *46th International Colloquium on Automata, Languages, and Programming, ICALP 2019, July 9-12, 2019, Patras, Greece*, volume 132 of *LIPICs*, pages 52:1–52:14. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2019. doi:10.4230/LIPICs.ICALP.2019.52.
- 9 Talya Eden, Dana Ron, and Will Rosenbaum. Almost optimal bounds for sublinear-time sampling of k -cliques in bounded arboricity graphs. In Mikolaj Bojanczyk, Emanuela Merelli, and David P. Woodruff, editors, *49th International Colloquium on Automata, Languages, and Programming, ICALP 2022, July 4-8, 2022, Paris, France*, volume 229 of *LIPICs*, pages 56:1–56:19. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2022. doi:10.4230/LIPICs.ICALP.2022.56.
- 10 Talya Eden and Will Rosenbaum. On sampling edges almost uniformly. In Raimund Seidel, editor, *1st Symposium on Simplicity in Algorithms, SOSA 2018, January 7-10, 2018, New Orleans, LA, USA*, volume 61 of *OASICs*, pages 7:1–7:9. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2018. doi:10.4230/OASICs.SOSA.2018.7.
- 11 Oded Goldreich. *Introduction to Property Testing*. Cambridge University Press, 2017.
- 12 Morris H Hansen and William N Hurwitz. On the theory of sampling from finite populations. *The Annals of Mathematical Statistics*, 14(4):333–362, 1943.
- 13 Jonathan Hermon. On sensitivity of uniform mixing times. In *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques*, volume 54, pages 234–248. Institut Henri Poincaré, 2018.
- 14 Jacob Holm and Jakub Tětek. Massively parallel computation and sublinear-time algorithms for embedded planar graphs. *arXiv preprint*, 2022. arXiv:2204.09035.

- 15 D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. doi:10.1080/01621459.1952.10483446.
- 16 Ben Morris and Yuval Peres. Evolving sets and mixin. In Lawrence L. Larmore and Michel X. Goemans, editors, *Proceedings of the 35th Annual ACM Symposium on Theory of Computing, June 9–11, 2003, San Diego, CA, USA*, pages 279–286. ACM, 2003. doi:10.1145/780542.780585.
- 17 Rajeev Motwani, Rina Panigrahy, and Ying Xu. Estimating sum by weighted sampling. In *International Colloquium on Automata, Languages, and Programming*, pages 53–64. Springer, 2007.
- 18 Krzysztof Onak and Xiaorui Sun. Probability-revealing samples. In Amos J. Storkey and Fernando Pérez-Cruz, editors, *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9–11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, volume 84 of *Proceedings of Machine Learning Research*, pages 2018–2026. PMLR, 2018. URL: <http://proceedings.mlr.press/v84/onak18a.html>.
- 19 Sofya Raskhodnikova, Dana Ron, Amir Shpilka, and Adam Smith. Strong lower bounds for approximating distribution support size and the distinct elements problem. *SIAM Journal on Computing*, 39(3):813–842, 2009.
- 20 Jakub Tětek and Mikkel Thorup. Edge sampling and graph parameter estimation via vertex neighborhood accesses. In Stefano Leonardi and Anupam Gupta, editors, *STOC '22: 54th Annual ACM SIGACT Symposium on Theory of Computing, Rome, Italy, June 20–24, 2022*, pages 1116–1129. ACM, 2022. doi:10.1145/3519935.3520059.

A Proof of Lemma 12

Here, we prove the main technical lemma for our sum estimator.

Proof. For each $i \in [N]$ we define γ_i such that $\mathcal{Q}(i) = (1 + \gamma_i)\mathcal{P}(i)$. Since we know that \mathcal{Q} is pointwise γ -close to \mathcal{P} then $|\gamma_i| \leq \gamma$, and since both \mathcal{Q} and \mathcal{P} are probability distributions then $\sum_{i \in [N]} \gamma_i \mathcal{P}(i) = 0$.

We start by proving the bounds on the expectation, i.e., Equation (9) and Equation (11).

$$\begin{aligned} \mathbb{E}[\zeta_k] &= \mathbb{E} \left[\sum_{h=1}^k (-1)^{h+1} \frac{\binom{k}{h}}{\binom{m}{h}} \sum_{i \in [N]} \binom{Y_i}{h} \mathcal{P}(i)^{-h} x_i \right] \\ &= \sum_{h=1}^k (-1)^{h+1} \frac{\binom{k}{h}}{\binom{m}{h}} \sum_{i \in [N]} \mathbb{E} \left[\binom{Y_i}{h} \right] \mathcal{P}(i)^{-h} x_i. \end{aligned}$$

We use the fact that $Y_i = \sum_{j \in [m]} [X_j = i]$ is a sum of 0-1 variables so

$\binom{Y_i}{h} = \sum_{I \subseteq [m]: |I|=h} \prod_{j \in I} [X_j = i]$. This implies that $\mathbb{E} \left[\binom{Y_i}{h} \right] = \binom{m}{h} \mathcal{Q}(i)^h = \binom{m}{h} (1 + \gamma_i)^h \mathcal{P}(i)^h$. Plugging this in, we get that

$$\begin{aligned} \mathbb{E}[\zeta_k] &= \sum_{h=1}^k (-1)^{h+1} \frac{\binom{k}{h}}{\binom{m}{h}} \sum_{i \in [N]} \binom{m}{h} (1 + \gamma_i)^h \mathcal{P}(i)^h \mathcal{P}(i)^{-h} x_i = \sum_{h=1}^k (-1)^{h+1} \binom{k}{h} \sum_{i \in [N]} (1 + \gamma_i)^h x_i \\ &= \sum_{i \in [N]} x_i \sum_{h=1}^k (-1)^{h+1} \binom{k}{h} (1 + \gamma_i)^h = \sum_{i \in [N]} x_i \left(1 + \sum_{h=0}^k (-1)^{h+1} \binom{k}{h} (1 + \gamma_i)^h \right) \\ &= \sum_{i \in [N]} x_i (1 - (1 - (1 + \gamma_i))^k) = \sum_{i \in [N]} x_i (1 + (-1)^{k+1} \gamma_i^k) \end{aligned}$$

If $k \geq 2$ then we easily see that

$$|\mathbb{E}[\zeta_k] - \mu| = \left| \sum_{i \in [N]} x_i - \sum_{i \in [N]} x_i(1 + (-1)^{k+1} \gamma_i^k) \right| = \left| \sum_{i \in [N]} \gamma_i^k x_i \right| \leq \gamma^k \sum_{i \in [N]} |x_i|$$

For $k = 1$ we will exploit that $\sum_{i \in [N]} \gamma_i \mathcal{P}(i) = 0$.

$$\begin{aligned} |\mathbb{E}[\zeta_1] - \mu| &= \left| \sum_{i \in [N]} \gamma_i x_i \right| = \left| \sum_{i \in [N]} \gamma_i (\mathcal{P}(i)\mu + (x_i - \mathcal{P}(i)\mu)) \right| \\ &= \left| \sum_{i \in [N]} \gamma_i (x_i - \mathcal{P}(i)\mu) \right| \leq \gamma \sum_{i \in [N]} |x_i - \mathcal{P}(i)\mu| \end{aligned}$$

Now we will focus on bounding the variance. First we prove Equation (12).

$$\begin{aligned} \text{Var}[\zeta_1] &= \mathbb{E} \left[\left(\frac{1}{m} \sum_{j \in [m]} \sum_{i \in [N]} x_i ([X_j = i] \mathcal{P}(i)^{-1} - (1 + \gamma_i)) \right)^2 \right] \\ &= \frac{1}{m^2} \sum_{j \in [m]} \mathbb{E} \left[\left(\sum_{i \in [N]} x_i ([X_j = i] \mathcal{P}(i)^{-1} - (1 + \gamma_i)) \right)^2 \right] \end{aligned}$$

We will argue that $\sum_{i \in [N]} x_i ([X_j = i] \mathcal{P}(i)^{-1} - (1 + \gamma_i)) = \sum_{i \in [N]} (x_i - \mathcal{P}(i)\mu) ([X_j = i] \mathcal{P}(i)^{-1} - (1 + \gamma_i))$.

$$\begin{aligned} &\sum_{i \in [N]} x_i ([X_j = i] \mathcal{P}(i)^{-1} - (1 + \gamma_i)) \\ &= \sum_{i \in [N]} (x_i - \mathcal{P}(i)\mu + \mathcal{P}(i)\mu) ([X_j = i] \mathcal{P}(i)^{-1} - (1 + \gamma_i)) \\ &= \sum_{i \in [N]} (x_i - \mathcal{P}(i)\mu) ([X_j = i] \mathcal{P}(i)^{-1} - (1 + \gamma_i)) + \mu \sum_{i \in [N]} \mathcal{P}(i) \gamma_i \end{aligned}$$

Now since $\sum_{i \in [N]} \gamma_i \mathcal{P}(i) = 0$ it follows. We can now bound the variance.

$$\begin{aligned} \text{Var}[\zeta_1] &= \frac{1}{m^2} \sum_{j \in [m]} \mathbb{E} \left[\left(\sum_{i \in [N]} (x_i - \mathcal{P}(i)\mu) ([X_j = i] \mathcal{P}(i)^{-1} - (1 + \gamma_i)) \right)^2 \right] \\ &\leq \frac{1}{m^2} \sum_{j \in [m]} \mathbb{E} \left[\left(\sum_{i \in [N]} \mathcal{P}(i)^{-1} (x_i - \mathcal{P}(i)\mu) [X_j = i] \right)^2 \right] \\ &= \frac{1}{m} \sum_{i \in [N]} \frac{\mathcal{Q}(i)}{\mathcal{P}(i)^2} (x_i - \mathcal{P}(i)\mu)^2 \leq \frac{1 + \gamma}{m} \sum_{i \in [N]} \mathcal{P}(i)^{-1} (x_i - \mathcal{P}(i)\mu)^2 \end{aligned}$$

Now we just need to focus on the case of $k \geq 2$ and prove Equation (10). For this we need the following lemma for which we defer the proof till Appendix A.

► **Lemma 18.** *For all sequences of numbers $(\beta_j)_{j \in [m]}$ and all α the following identity holds:*

$$\sum_{\substack{I \subseteq [m] \\ 0 < |I| \leq k}} (-1)^{|I|+1} \frac{\binom{k}{|I|}}{\binom{m}{|I|}} \left(\prod_{j \in I} \beta_j - (1 + \alpha)^k \right) = (-1)^{k+1} \sum_{\substack{I \subseteq [m] \\ 0 < |I| \leq k}} \frac{\binom{k}{|I|}}{\binom{m}{|I|}} \alpha^{k-|I|} \prod_{j \in I} (\beta_j - (1 + \alpha))$$

62:18 Bias Reduction for Sum Estimation

The idea is to use Lemma 18 to prove that

$$\zeta_k - \mathbb{E}[\zeta_k] = \sum_{\substack{I \subseteq [m] \\ 0 < |I| \leq k}} \frac{\binom{m-|I|}{k-|I|}}{\binom{m}{k}} \alpha^{k-|I|} \sum_{i \in [N]} x_i \prod_{j \in I} (\mathcal{P}(i)^{-1}[X_j = i] - (1 + \gamma_i)). \quad (24)$$

First we use that $\mathbb{E}[\zeta_k] = \sum_{h=1}^k (-1)^h \frac{\binom{k}{h}}{\binom{m}{h}} \sum_{i \in [N]} \mathcal{P}(i)^{-h} x_i \mathbb{E} \left[\binom{Y_i}{h} \right]$ which allow us to rewrite $\zeta_k - \mathbb{E}[\zeta_k]$.

$$\zeta_k - \mathbb{E}[\zeta_k] = \sum_{h=1}^k (-1)^h \frac{\binom{k}{h}}{\binom{m}{h}} \sum_{i \in [N]} \mathcal{P}(i)^{-h} x_i \left(\binom{Y_i}{h} - \mathbb{E} \left[\binom{Y_i}{h} \right] \right)$$

We now again use that $Y_i = \sum_{j \in [m]} [X_j = i]$ is a sum of 0-1 variables so $\binom{Y_i}{h} = \sum_{I \subseteq [m]: |I|=h} \prod_{j \in I} [X_j = i]$ and $\mathbb{E} \left[\binom{Y_i}{h} \right] = \binom{m}{h} (1 + \gamma_i)^h \mathcal{P}(i)^h$.

$$\begin{aligned} & \sum_{h=1}^k (-1)^h \frac{\binom{k}{h}}{\binom{m}{h}} \sum_{i \in [N]} \mathcal{P}(i)^{-h} x_i \left(\binom{Y_i}{h} - \mathbb{E} \left[\binom{Y_i}{h} \right] \right) \\ &= \sum_{h=1}^k (-1)^h \frac{\binom{k}{h}}{\binom{m}{h}} \sum_{i \in [N]} \mathcal{P}(i)^{-h} x_i \sum_{\substack{I \subseteq [m] \\ |I|=h}} \left(\prod_{j \in I} [X_j = i] - (1 + \gamma_i)^h \mathcal{P}(i)^h \right) \\ &= \sum_{i \in [N]} x_i \sum_{h=1}^k \sum_{\substack{I \subseteq [m] \\ |I|=h}} (-1)^h \frac{\binom{k}{h}}{\binom{m}{h}} \mathcal{P}(i)^{-h} \left(\prod_{j \in I} [X_j = i] - (1 + \gamma_i)^h \mathcal{P}(i)^h \right) \\ &= \sum_{i \in [N]} x_i \sum_{\substack{I \subseteq [m] \\ 0 < |I| \leq k}} (-1)^{|I|+1} \frac{\binom{k}{|I|}}{\binom{m}{|I|}} \mathcal{P}(i)^{-|I|} \left(\prod_{j \in I} [X_j = i] - (1 + \gamma_i)^{|I|} \mathcal{P}(i)^{|I|} \right) \\ &= \sum_{i \in [N]} x_i \sum_{\substack{I \subseteq [m] \\ 0 < |I| \leq k}} (-1)^{|I|+1} \frac{\binom{k}{|I|}}{\binom{m}{|I|}} \left(\prod_{j \in I} \mathcal{P}(i)^{-1}[X_j = i] - (1 + \gamma_i)^{|I|} \right) \end{aligned}$$

For each $i \in [N]$, the expression $\sum_{\substack{I \subseteq [m] \\ 0 < |I| \leq k}} (-1)^{|I|+1} \frac{\binom{k}{|I|}}{\binom{m}{|I|}} \left(\prod_{j \in I} \mathcal{P}(i)^{-1}[X_j = i] - (1 + \gamma_i)^{|I|} \right)$ is of the form of Lemma 18. So applying that N times we get that

$$\begin{aligned} & \sum_{i \in [N]} x_i \sum_{\substack{I \subseteq [m] \\ 0 < |I| \leq k}} (-1)^{|I|+1} \frac{\binom{k}{|I|}}{\binom{m}{|I|}} \left(\prod_{j \in I} \mathcal{P}(i)^{-1}[X_j = i] - (1 + \gamma_i)^{|I|} \right) \\ &= \sum_{i \in [N]} x_i (-1)^{k+1} \sum_{\substack{I \subseteq [m] \\ 0 < |I| \leq k}} \frac{\binom{k}{|I|}}{\binom{m}{|I|}} \gamma_i^{k-|I|} \prod_{j \in I} (\mathcal{P}(i)^{-1}[X_j = i] - (1 + \gamma_i)) \\ &= (-1)^{k+1} \sum_{\substack{I \subseteq [m] \\ 0 < |I| \leq k}} \frac{\binom{k}{|I|}}{\binom{m}{|I|}} \sum_{i \in [N]} \gamma_i^{k-|I|} x_i \prod_{j \in I} (\mathcal{P}(i)^{-1}[X_j = i] - (1 + \gamma_i)) \end{aligned}$$

This shows that Equation (24) is true.

Now let $I_1, I_2 \subseteq [m]$ be two different set of indices with $0 < |I_1|, |I_2| \leq k$. Without loss of generality, we can assume that there exists $h \in I_1 \setminus I_2$.

$$T = \left(\sum_{i \in [N]} \gamma_i^{k-|I_1|} x_i \prod_{j \in I_1} (\mathcal{P}(i)^{-1}[X_j = i] - (1 + \gamma_i)) \right) \cdot \left(\sum_{i \in [N]} \gamma_i^{k-|I_2|} x_i \prod_{j \in I_2} (\mathcal{P}(i)^{-1}[X_j = i] - (1 + \gamma_i)) \right)$$

Note that if we multiply this expression out then every term will contain a factor of the form $(\mathcal{P}(i)^{-1}[X_h = s] - (1 + \gamma_i))$ for some $s \in [N]$ and where all the other factors are independent of X_h . Since $\mathbb{E}[(\mathcal{P}(i)^{-1}[X_h = s] - (1 + \gamma_i))] = 0$ we get that $\mathbb{E}[T] = 0$. This implies that

$$\mathbb{E}[(\zeta_k - \mathbb{E}[\zeta_k])^2] = \sum_{\substack{I \subseteq [m] \\ 0 < |I| \leq k}} \left(\frac{\binom{k}{|I|}}{\binom{m}{|I|}} \right)^2 \mathbb{E}[\left(\sum_{i \in [N]} \gamma_i^{k-|I|} x_i \prod_{j \in I} (\mathcal{P}(i)^{-1}[X_j = i] - (1 + \gamma_i)) \right)^2]$$

Let $I \subseteq [N]$ be fixed then

$$\begin{aligned} \mathbb{E}[\left(\sum_{i \in [N]} \gamma_i^{k-|I|} x_i \prod_{j \in I} (\mathcal{P}(i)^{-1}[X_j = i] - (1 + \gamma_i)) \right)^2] &\leq \mathbb{E}[\left(\sum_{i \in [N]} \gamma_i^{k-|I|} x_i \prod_{j \in I} \mathcal{P}(i)^{-1}[X_j = i] \right)^2] \\ &= \sum_{i \in [N]} \gamma_i^{2k-2|I|} x_i^2 \frac{\mathcal{Q}(i)^{|I|}}{\mathcal{P}(i)^{2|I|}} \leq \gamma^{2k-2|I|} \sum_{i \in [N]} x_i^2 \frac{(1 + \gamma_i)^{|I|}}{\mathcal{P}(i)^{|I|}} \end{aligned}$$

Collecting terms we get that

$$\begin{aligned} \mathbb{E}[(\zeta_k - \mathbb{E}[\zeta_k])^2] &\leq \sum_{\substack{I \subseteq [m] \\ 0 < |I| \leq k}} \left(\frac{\binom{k}{|I|}}{\binom{m}{|I|}} \right)^2 \gamma^{2k-2|I|} \sum_{i \in [N]} x_i^2 \frac{(1 + \gamma_i)^{|I|}}{\mathcal{P}(i)^{|I|}} \\ &= \sum_{h=1}^k \frac{\binom{k}{h}^2}{\binom{m}{h}^2} \gamma^{2k-2h} \sum_{i \in [N]} x_i^2 \frac{(1 + \gamma_i)^h}{\mathcal{P}(i)^h} \leq \sum_{h=1}^k \frac{\binom{k}{h}^2}{\binom{m}{h}^2} \gamma^{2k-2h} (1 + \gamma)^h \sum_{i \in [N]} \frac{x_i^2}{\mathcal{P}(i)^h} \\ &\leq \sum_{h=1}^k \frac{k^{2h}}{\binom{m}{h}^2} \gamma^{2k-2h} (1 + \gamma)^h \sum_{i \in [N]} \frac{x_i^2}{\mathcal{P}(i)^h} \leq \max_{h=1}^k \frac{2^h k^{2h}}{\binom{m}{h}^2} \gamma^{2k-2h} (1 + \gamma)^h \sum_{i \in [N]} \frac{x_i^2}{\mathcal{P}(i)^h} \end{aligned}$$

Now we note that for all $i \in [N]$, the map $h \mapsto \frac{2^h k^{2h}}{\binom{m}{h}^2} \gamma^{2k-2h} (1 + \gamma)^h \frac{x_i^2}{\mathcal{P}(i)^h}$ is log-convex since each factor is log-convex. This implies that the map $h \mapsto \frac{2^h k^{2h}}{\binom{m}{h}^2} \gamma^{2k-2h} (1 + \gamma)^h \sum_{i \in [N]} \frac{x_i^2}{\mathcal{P}(i)^h}$ is convex and is thus maximized at the boundary. We then get that

$$\mathbb{E}[(\zeta_k - \mathbb{E}[\zeta_k])^2] \leq \max \left\{ 2(1 + \gamma) \gamma^{2k-2} k^2 \frac{\sum_{i \in [N]} \mathcal{P}(i)^{-1} x_i^2}{m}, 2^k (1 + \gamma)^k k^{3k} \frac{n^{k-1} \sum_{i \in [N]} \mathcal{P}(i)^{-1} x_i^2}{m^k} \right\}$$

which finishes the proof of Equation (10). ◀

Finally, we must prove Lemma 18 which will finish the proof of Lemma 12.

► **Lemma 19.** *For all sequences of numbers $(\beta_j)_{j \in [m]}$ and all α the following identity holds:*

$$\sum_{\substack{I \subseteq [m] \\ 0 < |I| \leq k}} (-1)^{|I|+1} \frac{\binom{k}{|I|}}{\binom{m}{|I|}} \left(\prod_{j \in I} \beta_j - (1 + \alpha)^h \right) = (-1)^{k+1} \sum_{\substack{I \subseteq [m] \\ 0 < |I| \leq k}} \frac{\binom{k}{|I|}}{\binom{m}{|I|}} \alpha^{k-|I|} \prod_{j \in I} (\beta_j - (1 + \alpha))$$

62:20 Bias Reduction for Sum Estimation

First we need a simple lemma.

► **Lemma 19.** For all sequences of numbers $(\beta_j)_{j \in I}$ and all α the following identity holds:

$$\prod_{j \in I} \beta_j - (1 + \alpha)^{|I|} = \sum_{\emptyset \neq J \subseteq I} (1 + \alpha)^{|I| - |J|} \prod_{j \in J} (\beta_j - (1 + \alpha))$$

Proof. Let $\beta'_j := \beta_j - (1 + \alpha)$. Then,

$$\begin{aligned} \prod_{j \in I} \beta'_j - (1 + \alpha)^{|I|} &= \prod_{j \in I} (\beta'_j + (1 + \alpha)) - (1 + \alpha)^{|I|} \\ &= \left[\sum_{J \subseteq I} (1 + \alpha)^{|I| - |J|} \prod_{j \in J} \beta'_j \right] - (1 + \alpha)^{|I|} \\ &= \sum_{\emptyset \neq J \subseteq I} (1 + \alpha)^{|I| - |J|} \prod_{j \in J} (\beta_j - (1 + \alpha)). \end{aligned}$$

Proof of Lemma 18. We start by applying Lemma 19

$$\begin{aligned} &\sum_{\substack{I \subseteq [m] \\ 0 < |I| \leq k}} (-1)^{|I|+1} \frac{\binom{k}{|I|}}{\binom{m}{|I|}} \left(\prod_{j \in I} \beta_j - (1 + \alpha)^{|I|} \right) \\ &= \sum_{\substack{I \subseteq [m] \\ 0 < |I| \leq k}} (-1)^{|I|+1} \frac{\binom{k}{|I|}}{\binom{m}{|I|}} \sum_{\emptyset \neq J \subseteq I} (1 + \alpha)^{|I| - |J|} \prod_{j \in J} (\beta_j - (1 + \alpha)) \end{aligned}$$

We use that $\frac{\binom{k}{|I|}}{\binom{m}{|I|}} = \frac{\binom{m-|I|}{k-|I|}}{\binom{m}{k}}$ which follows from the fact that $\binom{m}{k} \binom{k}{|I|} = \binom{m}{|I|, k-|I|, m-k} = \binom{m}{|I|} \binom{m-|I|}{k-|I|}$.

$$\begin{aligned} &\sum_{\substack{I \subseteq [m] \\ 0 < |I| \leq k}} (-1)^{|I|+1} \frac{\binom{k}{|I|}}{\binom{m}{|I|}} \sum_{\emptyset \neq J \subseteq I} (1 + \alpha)^{|I| - |J|} \prod_{j \in J} (\beta_j - (1 + \alpha)) \\ &= \sum_{\substack{I \subseteq [m] \\ 0 < |I| \leq k}} (-1)^{|I|+1} \frac{\binom{m-|I|}{k-|I|}}{\binom{m}{k}} \sum_{\emptyset \neq J \subseteq I} (1 + \alpha)^{|I| - |J|} \prod_{j \in J} (\beta_j - (1 + \alpha)) \\ &= \frac{1}{\binom{m}{k}} \sum_{\substack{J \subseteq [m] \\ 0 < |J| \leq k}} \left(\prod_{j \in J} (\beta_j - (1 + \alpha)) \right) \sum_{\substack{I \subseteq [m] \\ J \subseteq I, |I| \leq k}} (-1)^{|I|+1} \binom{m-|I|}{k-|I|} (1 + \alpha)^{|I| - |J|} \\ &= \frac{1}{\binom{m}{k}} \sum_{\substack{J \subseteq [m] \\ 0 < |J| \leq k}} \left(\prod_{j \in J} (\beta_j - (1 + \alpha)) \right) \sum_{h=|J|}^k (-1)^{h+1} \binom{m-|J|}{h-|J|} \binom{m-h}{k-h} (1 + \alpha)^{h-|J|} \end{aligned}$$

Now we use that $\binom{m-h}{k-h} \binom{m-|J|}{h-|J|} = \binom{m-|J|}{k-h, m-k, h-|J|} = \binom{m-|J|}{k-|J|} \binom{k-|J|}{h-|J|}$.

$$\begin{aligned} & \frac{1}{\binom{m}{k}} \sum_{\substack{J \subseteq [m] \\ 0 < |J| \leq k}} \left(\prod_{j \in J} (\beta_j - (1 + \alpha)) \right) \sum_{h=|J|}^k (-1)^{h+1} \binom{m-|J|}{h-|J|} \binom{m-h}{k-h} (1 + \alpha)^{h-|J|} \\ &= \frac{1}{\binom{m}{k}} \sum_{\substack{J \subseteq [m] \\ 0 < |J| \leq k}} \left(\prod_{j \in J} (\beta_j - (1 + \alpha)) \right) \sum_{h=|J|}^k (-1)^{h+1} \binom{m-|J|}{k-|J|} \binom{k-|J|}{h-|J|} (1 + \alpha)^{h-|J|} \\ &= \frac{1}{\binom{m}{k}} \sum_{\substack{J \subseteq [m] \\ 0 < |J| \leq k}} (-1)^{|J|+1} \binom{m-|J|}{k-|J|} \left(\prod_{j \in J} (\beta_j - (1 + \alpha)) \right) \sum_{h=0}^{k-|J|} (-1)^h \binom{k-|J|}{h} (1 + \alpha)^h \\ &= \frac{1}{\binom{m}{k}} \sum_{\substack{J \subseteq [m] \\ 0 < |J| \leq k}} (-1)^{|J|+1} \binom{m-|J|}{k-|J|} \left(\prod_{j \in J} (\beta_j - (1 + \alpha)) \right) (1 - (1 + \alpha))^{k-|J|} \\ &= \frac{(-1)^{k+1}}{\binom{m}{k}} \sum_{\substack{J \subseteq [m] \\ 0 < |J| \leq k}} \binom{m-|J|}{k-|J|} \alpha^{k-|J|} \left(\prod_{j \in J} (\beta_j - (1 + \alpha)) \right) \end{aligned}$$

Finally, we again use that $\frac{\binom{k}{|J|}}{\binom{m}{|J|}} = \frac{\binom{m-|J|}{k-|J|}}{\binom{m}{k}}$ to finish the proof.

$$\begin{aligned} & \frac{(-1)^{k+1}}{\binom{m}{k}} \sum_{\substack{J \subseteq [m] \\ 0 < |J| \leq k}} \binom{m-|J|}{k-|J|} \alpha^{k-|J|} \left(\prod_{j \in J} (\beta_j - (1 + \alpha)) \right) \\ &= (-1)^{k+1} \sum_{\substack{J \subseteq [m] \\ 0 < |J| \leq k}} \frac{\binom{k}{|J|}}{\binom{m}{|J|}} \alpha^{k-|J|} \left(\prod_{j \in J} (\beta_j - (1 + \alpha)) \right) \end{aligned}$$