

Strong Lower Bounds for Approximating Distribution Support Size and the Distinct Elements Problem

Sofya Raskhodnikova* Dana Ron† Amir Shpilka‡ Adam Smith*

June 4, 2007

Abstract

We consider the problem of approximating the support size of a distribution from a small number of samples, when each element in the distribution appears with probability at least $\frac{1}{n}$. This problem is closely related to the problem of approximating the number of distinct elements in a sequence of length n . For both problems, we prove a nearly linear in n lower bound on the query complexity, applicable even for approximation with *additive* error.

At the heart of the lower bound is a construction of two positive integer random variables, X_1 and X_2 , with very different expectations and the following condition on the first k moments: $E[X_1]/E[X_2] = E[X_1^2]/E[X_2^2] = \dots = E[X_1^k]/E[X_2^k]$. Our lower bound method is also applicable to other problems. In particular, it gives a new lower bound for the number of samples needed to approximate the entropy of a distribution. Another application of our lower bound is to the problem of approximating the compressibility of a string according to the Lempel-Ziv compression scheme.

*Pennsylvania State University, USA. Email: {sofya,asmith}@cse.psu.edu. Research done while at the Weizmann Institute of Science, Israel. A.S. was supported at Weizmann by the Louis L. and Anita M. Perlman Postdoctoral Fellowship.

†Tel Aviv University, Ramat Aviv, Israel. Email: danar@eng.tau.ac.il. Supported by the Israel Science Foundation (grant number 89/05).

‡Technion, Haifa, Israel. Email: shpilka@cs.technion.ac.il.

1 Introduction

In this work we consider the following problem, which we call DISTRIBUTION-SUPPORT-SIZE (DSS): *Given access to independent samples from a distribution where each element appears with probability at least $\frac{1}{n}$, approximate the distribution support size.* This problem is closely related to another natural problem, known as DISTINCT-ELEMENTS (DE): *Given access to a sequence of length n , approximate the number of distinct elements in the sequence.* Both of these fundamental problems arise in many contexts and have been extensively studied. In statistics, DSS is known as estimating the number of species in a population (see the list of hundreds of references in [8]). Typically, the input distribution is assumed to come from a specific family. DE arises in databases and data mining, for example in the design of query optimizers and the detection of denial-of-service attacks (see [9, 1] and references therein). Because of the overwhelming size of modern databases, a significant effort has focused on solving DE with particular, efficient classes of algorithms: streaming algorithms [2, 4, 11], which make a single pass through the data and use very little memory, and sampling-based algorithms [9, 4], which query only a small number of positions in the input.

This paper looks at the complexity of sampling-based approximation algorithms for DSS and DE. To the best of our knowledge, previous works consider only multiplicative approximations for these problems. Charikar *et al.* [9] and Bar-Yossef *et al.* [4] prove that approximating DE within multiplicative error α requires $\Omega\left(\frac{n}{\alpha^2}\right)$ queries into the input sequence. This lower bound is tight [9]. Its proof boils down to the observation that every algorithm requires $\Omega\left(\frac{n}{\alpha^2}\right)$ queries to distinguish a sequence of n identical elements from the same sequence with α^2 unique elements inserted in random positions. Stated in terms of the DSS problem, the difficulty is in distinguishing a distribution with a single element in its support from a distribution with support size α^2 , where all but one of the elements have weight $1/n$. A good metaphor for the distinguishing task in this argument is finding a needle in a haystack.

This needle-in-a-haystack lower bound leaves open the question of the complexity of DSS when the support size is a non-negligible fraction of n . In other words, is it possible to obtain efficient *additive* approximation algorithms for DSS and DE? This work gives a strong lower bound for the sample (and thus time) complexity of such algorithms. Our techniques also lead to lower bounds on the sample complexity of approximating the compressibility of a string and the entropy of a distribution. We describe our results in more detail in the rest of this section.

1.1 An Almost Linear Lower Bound for Approximation with an Additive Error

First we discuss how DSS and DE are related. An instance of DSS where all probabilities are multiples of $\frac{1}{n}$ is equivalent to a DE instance that can be accessed only by taking independent uniform samples with replacement. Thus, the following problem is a special case of DSS and a restriction of DE: *Given n balls, each of a single color, approximate the number of distinct colors by taking independent uniform samples of the balls with replacement.*

It turns out that this restriction of DE can be made without loss of generality. In principle, an algorithm for DE is allowed to make arbitrary adaptive queries to the input. However, Bar-Yossef [3] shows that algorithms that (a) take uniform random samples with replacement and (b) see the input positions corresponding to the samples, are essentially as good for solving DE as general algorithms. We strengthen his result to algorithms that sample uniformly with replacement but are *oblivious* to the input positions corresponding to the samples. Hence, to obtain lower bounds for both DSS and general DE, it suffices to prove bounds for the restriction of DE above. From this point on we refer interchangeably to the two variants of DE and use the terms “balls” for the positions/samples and “colors” for distinct elements.

Main lower bound. We prove that even if we allow an additive error, so that the multiplicative lower bound [9, 4] does not apply, approximating DE (and hence DSS) requires an almost linear number of queries. Specifically, $n^{1-o(1)}$ queries are necessary to distinguish an input with $\frac{n}{11}$ colors from an input with $\frac{n}{d}$ colors, for any $d = n^{o(1)}$. In particular, obtaining additive error $\frac{n}{12}$ requires $n^{1-o(1)}$ samples. In the above statements and in all that follows, distinguishing means *distinguishing with success probability at least 2/3*.

Such a strong lower bound for an additive approximation may seem surprising. It is easy to prove an $\Omega(\sqrt{n})$ bound on the query complexity of approximating DE with an additive error (recall that we may assume without loss of generality that the algorithm samples uniformly with replacement): with fewer queries it is hard to distinguish an instance with n colors where each color appears once from an instance with $\frac{n}{2}$ colors where each color appears twice. In both cases an algorithm taking $o(\sqrt{n})$ samples is likely to see only unique colors (no collisions). With $\Omega(\sqrt{n})$ samples, 2-way collisions become likely even if all colors appear only a constant number of times in the input. In general, with $\Omega(n^{1-1/k})$ samples, k -way collisions become likely. Intuitively, it seems that one should be able to use statistics on the number of collisions to efficiently distinguish an input with $\frac{n}{d_1}$ colors from an input with $\frac{n}{d_2}$ colors, where d_1 and d_2 are different constants. Surprisingly, it turns out that in our case looking at k -way collisions, for constant k (and even k that is a slowly growing function of n), does not help.

1.2 Techniques

Moment conditions and frequency variables. To prove our lower bound, we construct two input instances that are hard to distinguish, where the inputs have $\frac{n}{d_1}$ and $\frac{n}{d_2}$ colors, respectively, and $d_2 \gg d_1$. The requirements on the number of colors imply that, unlike in the “needle in a haystack” lower bound of [9, 4], the instances being distinguished must have linear Hamming distance. Previous techniques do not apply here, and we need a more subtle argument to show that they are indistinguishable. At the heart of the construction are two positive integer random variables, X_1 and X_2 , that correspond to the two input instances. These random variables have very different expectations (which translate to different numbers of colors) and many *proportional moments*, that is $\frac{\mathbb{E}[X_1]}{\mathbb{E}[X_2]} = \frac{\mathbb{E}[X_1^2]}{\mathbb{E}[X_2^2]} = \dots = \frac{\mathbb{E}[X_1^{k-1}]}{\mathbb{E}[X_2^{k-1}]}$, for some $k = \omega(1)$. The construction of these random variables proceeds by formulating the problem in terms of polynomials and bounding their coefficients, and it is the most technically delicate step of our lower bound (see Section 4).

Let F_ℓ be the number of ℓ -way collisions, that is the number of colors that appear exactly ℓ times in the sample. As explained in the discussion of the main lower bound, computing F_ℓ for small ℓ gives a possible strategy for distinguishing two DE instances. Intuitively, we will ensure that this strategy fails for the instances we construct, by requiring that the expected value of F_ℓ is the same for both instances. To this end, for each instance of DE we define its *frequency variable* to be the outcome of a mental experiment where we choose a *color* uniformly at random and count how many times it occurs in the instance. We prove that the expectation of F_ℓ is the same for two instances if their frequency variables X_1 and X_2 have at least ℓ proportional moments. Thus, the construction mentioned above leads to a pair of instances where F_ℓ has the same expectation for small values of ℓ .

Instances with frequency variables with proportional moments are indistinguishable. Our second technical contribution is to show that constructing frequency variables with proportional moments is sufficient for proving lower bounds on sample complexity: namely, the corresponding instances are indistinguishable given few samples. (This gives a general technique for proving lower bounds on sample complexity: if the quantity to be approximated can be expressed in terms of the distribution of an input’s frequency variable, then it is sufficient to construct two integer variables with proportional moments for which the

quantity differs significantly. We illustrate this generality by also deriving bounds for entropy estimation, discussed in the next subsection.)

To prove a lower bound, it suffices to consider algorithms that have access only to the *histogram* (F_1, F_2, F_3, \dots) of the selected sample. That is, the algorithm is only given the number of colors in the sample that appear once, twice, thrice, etc. The restriction to histograms was also applied in [7, 6]. The difficulty of proving indistinguishability based on proportional moments lies in translating guarantees of *equal expectations* of the variables F_ℓ , to a guarantee of *close distributions* on the vectors (F_1, F_2, F_3, \dots) . The main idea is to show that (a) the variables F_1, \dots, F_{k-1} can each be faithfully approximated by a Poisson random variable with the same expectation, and (b) they are close to being independent. The explanation for the latter, counter-intuitive statement comes from the following experiment: consider many independent rolls of a biased k -sided die. If one side of the die appears with probability close to 1, then the variables counting the number of times each of the other sides appears are close to being independent. In our scenario, side ℓ of the die (for $0 \leq \ell < k$) occurs when a particular color appears ℓ times in the sample. Any given color is most likely not to appear at all, so side 0 of the die is overwhelmingly likely and the counts of the remaining outcomes are nearly independent.

The proofs use a technique called *Poissonization* [14], in which one modifies a probability experiment to replace a fixed quantity (e.g. the number of samples) with a variable one which follows a Poisson distribution. This breaks up dependencies between variables, and makes the analysis tractable.

1.3 Results for Other Problems

As shown in [13], DE is closely related to the problem of approximating the compressibility of a string according to the Lempel-Ziv compression scheme (the version in [16]). By applying the reduction in [13], the lower bound we give for DE implies a lower bound on the complexity of approximating compressibility according to this scheme. The resulting lower bound for compressibility shows that the algorithm given in [13] cannot be significantly improved.

Furthermore, our lower bound method can be extended to other problems where one needs to compute quantities invariant under the permutation of the balls and the colors. In particular, as we show in Section 7, our method gives a lower bound of $\Omega\left(n^{\frac{2}{6\alpha^2-3+o(1)}}\right)$ on approximating the entropy of a distribution over n elements to within a multiplicative factor of α . In particular, when α is close to 1, this bound is close to $\Omega(n^{2/3})$. It can be combined with the $\Omega\left(n^{\frac{1}{2\alpha^2}}\right)$ bound in [5] to give $\Omega\left(n^{\max\left\{\frac{1}{2\alpha^2}, \frac{2}{6\alpha^2-3+o(1)}\right\}}\right)$.

2 Main Result

As noted in the introduction, DE with algorithms that sample uniformly with replacement is a special case of DSS where all probabilities are integer multiples of $\frac{1}{n}$. Hence, Theorem 2.1, stated next, directly implies a lower bound for DSS as well.

Theorem 2.1 *For every sufficiently large n and for every $B \leq n^{1/4}/\sqrt{\log n}$, the following holds for $k = k(n, B) = \left\lfloor \sqrt{\frac{\log n}{\log B + \frac{1}{2} \log \log n}} \right\rfloor$. Every algorithm for DE needs to perform $\Omega\left(n^{1-\frac{2}{k}}\right)$ queries to distinguish inputs with at least $\frac{n}{11}$ colors¹ from inputs with at most $\frac{n}{B}$ colors.*

¹It is quite plausible that 11 is not the smallest constant that can be obtained.

The next corollary provides an important special case:

Corollary 2.2 *For any $B = n^{o(1)}$, distinguishing inputs of DE with at least $n/11$ colors from inputs with at most n/B colors requires $n^{1-o(1)}$ queries.*

To prove Theorem 2.1 we construct a pair of DE instances that are hard to distinguish (though they contain a very different number of colors). As mentioned in the introduction, in order to obtain a lower bound on DE it suffices to consider algorithms that take uniform samples with replacement. The proof is given in Appendix A. In Section 4 we construct integer random variables that satisfy the moments condition, as described in the introduction. Section 5 shows that frequency variables with proportional moments lead to indistinguishable instances of DE. In Section 6 we prove Theorem 2.1 based on the results from the preceding sections. Finally, in Section 7 we discuss the application of our techniques to the sample complexity of approximating the entropy.

3 Algorithms for DE with Uniform Samples

In general, an algorithm for DE is allowed to make arbitrary (adaptive) queries to the input. We first establish that it is enough to consider algorithms for DE that sample uniformly at random with replacement. The proof of Lemma 3.2 below appears in Appendix A.

Definition 3.1 (Uniform algorithm) *An algorithm is uniform if it takes independent samples with replacement and only gets to see the colors of the samples, but not the input positions corresponding to them.*

Lemma 3.2 *If every uniform algorithm needs at least s queries in order to distinguish DE instances with at least C_1 colors from DE instances with at most C_2 colors, then every algorithm needs $\Omega(s)$ queries in order to distinguish DE instances with at least $0.1 \cdot C_1$ colors from DE instances with at most C_2 colors.*

4 Frequency Variables and the Moments Condition

This section defines and constructs the *frequency variables* needed for the main lower bound, as described in the introduction. To begin, note that permuting color names in the input (e.g., painting all pink balls orange and vice versa) clearly does not change the number of colors. Intuitively, all colors play the same role, and the only useful information in the sample is the number of colors that appear exactly once, exactly twice, etc. This motivates the following definition.

Definition 4.1 (Collisions and Histograms) *Consider s samples taken by an algorithm. An ℓ -way collision occurs if a color appears exactly ℓ times in the sample. We denote by F_ℓ , for $\ell = 0, 1, \dots, s$, the number of ℓ -way collisions in the sample. The histogram F of the sample is the vector (F_1, \dots, F_s) , indicating for each non-zero ℓ how many colors appear exactly ℓ times in the sample.*

One can prove that any uniform algorithm for DE can be simulated by a uniform algorithm that only sees a histogram of the sample. (We omit the proof since it follows from the formal argument further below).

To prove our lower bound, we will define a pair of DE instances that contain a significantly different number of colors, but for which the corresponding distributions on histograms are indistinguishable. First, observe that if the algorithm takes $o(n^{1-1/k})$ samples, and each color appears at most a constant number

of times, then with high probability no k -way collisions occur. Hence, it suffices to restrict our attention to ℓ -way collisions for $\ell < k$. Next we consider the following notion, which is closely related to ℓ -way collisions: A *monochromatic ℓ -tuple* is a set of ℓ samples that have the same color.² If we are able to get similar distributions on the number of ℓ -tuples (for the two different instances), then we get similar distributions on ℓ -way collisions. In this section, we show how to construct pairs of instances with the same *expectations* on the number of ℓ -tuples, for every $\ell < k$. (Section 5 proves that making the expectations equal implies that the distributions themselves will be close.) In order to express these expectations concisely, we define, for each instance of DE, a corresponding *frequency variable*.

Definition 4.2 (Frequency Variable) Consider an instance of DE with $\frac{n}{d}$ colors. Group colors into types according to how many times they appear in the input: say, p_i fraction of the colors are of type i and each of them appears a_i times. Consider a mental experiment where we choose a color uniformly at random and count how many times it occurs in the instance. The frequency variable X is a random variable representing the number of balls of a color chosen uniformly at random, as described in the experiment.

By definition, $\Pr[X = a_i] = p_i$. Since, on average, each color appears d times, $E[X] = \sum_i p_i a_i = d$. Note that for any integer random variable X which takes value a_i with probability p_i , if the numbers $p_i \frac{n}{d}$ are integers, then we can easily construct a DE instance with frequency variable X .

Suppose an algorithm takes s uniform samples with replacement from an instance with $\frac{n}{d}$ colors as described in Definition 4.2. The probability that a particular ℓ -tuple is monochromatic is $\sum_i p_i \frac{n}{d} \left(\frac{a_i}{n}\right)^\ell$, since there are $p_i \frac{n}{d}$ colors of type i and each gets sampled with probability $\frac{a_i}{n}$. The expected number of monochromatic ℓ -tuples in s samples is thus

$$\binom{s}{\ell} \sum_i p_i \frac{n}{d} \left(\frac{a_i}{n}\right)^\ell = \binom{s}{\ell} \frac{1}{n^{\ell-1}} \frac{1}{d} \sum_i p_i a_i^\ell = \binom{s}{\ell} \frac{1}{n^{\ell-1}} \frac{E[X^\ell]}{E[X]}.$$

The last equality holds because $E[X] = d$ and $\Pr[X = a_i] = p_i$. We will consider s for which this expression goes to 0 when ℓ is at least some fixed k . We want to construct a pair of instances such that for the remaining ℓ (which are smaller than k), the *expected* number of monochromatic ℓ -tuples is the same. This corresponds to making $\frac{E[X^\ell]}{E[X]}$ the same for both instances. This, in turn, leads to the following condition on the corresponding frequency variables, which is the core of our lower bound.

Definition 4.3 (Proportional Moments) We say that two random variables \hat{X} and \tilde{X} have $k - 1$ proportional moments if $\frac{E[\tilde{X}]}{E[\hat{X}]} = \frac{E[\tilde{X}^2]}{E[\hat{X}^2]} = \frac{E[\tilde{X}^3]}{E[\hat{X}^3]} = \dots = \frac{E[\tilde{X}^{k-1}]}{E[\hat{X}^{k-1}]}$.

We will see in Section 5 that when two frequency variables have $k - 1$ proportional moments, the corresponding instances are indistinguishable by algorithms that take (roughly) fewer than $n^{1-\frac{1}{k}}$ samples. However, we need, additionally, that the instances have very different numbers of distinct colors. This corresponds to insisting that the frequency variables have different expectations.

Definition 4.4 (Moments Condition) We say that two random variables \hat{X} and \tilde{X} satisfy the moments condition with parameters k and B if $\frac{E[\tilde{X}]}{E[\hat{X}]} \geq B$ and \hat{X} and \tilde{X} have $k - 1$ proportional moments.

²The relationship between monochromatic tuples and collisions is straightforward: a monochromatic ℓ -tuple implies an ℓ' way collision of the same color for some $\ell' \geq \ell$, and conversely an ℓ' -way collision implies that there are $\binom{\ell'}{\ell}$ monochromatic ℓ -tuples of the same color.

Theorem 4.5 (R.V.'s Satisfying the Moments Condition) *For all integers $k > 1$ and $B > 1$, there exist random variables \hat{X} and \tilde{X} over positive integers $a_0 < a_1 < \dots < a_{k-1}$ that satisfy the moments condition with parameters k and B . Moreover, for these variables $a_i = (B + 3)^i$, $E[\tilde{X}] > B$ and $E[\hat{X}] < 1 + \frac{1}{B}$.*

The remainder of this section is devoted to sketching the proof of this theorem. To reduce notation, all variables pertaining to the first instance in the pair of instances that are hard to distinguish, are marked by a hat ($\hat{\cdot}$) and those pertaining to the second, by a tilde ($\tilde{\cdot}$). In statements relevant to both instances, the corresponding variables without hat or tilde are used.

We begin by giving an overview of the construction and its analysis. Let $C = E[\tilde{X}] / E[\hat{X}]$. Then the moments condition (Definition 4.4) can be restated as $(E[\tilde{X}], E[\tilde{X}^2], \dots, E[\tilde{X}^{k-1}]) = C \cdot (E[\hat{X}], E[\hat{X}^2], \dots, E[\hat{X}^{k-1}])$ and $C \geq B$. Recall that the supports of \hat{X} and \tilde{X} are both contained in $\{a_0, \dots, a_{k-1}\}$. The main step in our construction is to set $a_j = a^j$ for an appropriate choice of a . Let $p_i = \Pr[X = a_i]$, and $\vec{p} = (p_0, \dots, p_{k-1})$. Let V denote the $(k-1) \times k$ Vandermonde matrix satisfying $V_{i,j} = (a_j)^i$. Then the vector $(E[X], E[X^2], \dots, E[X^{k-1}])$ can be represented as the product $V \cdot \vec{p}$. This gives yet another way to formulate the moments condition: $V(C \cdot \vec{p} - \vec{p}) = \vec{0}$. For a fixed a , there is a unique (up to a factor) non-zero vector \vec{u} satisfying $V \cdot \vec{u} = \vec{0}$. To obtain probability vectors $\vec{\hat{p}}$ and $\vec{\tilde{p}}$ from \vec{u} , we let positive coordinates u_i become $C \cdot \hat{p}_i$ and negative u_i become $-\tilde{p}_i$, divided by the corresponding normalization factors. This defines distributions \hat{X} and \tilde{X} , for each a .

In the proof we give a lower bound on C in terms of k and a . It implies that it is enough to set $a = B + 3$. The main idea behind the analysis is to view \vec{u} as coefficients of a polynomial. Let $f(t) = t^{k-1} + u_{k-2}t^{k-2} + \dots + u_0$ be the unique non-zero polynomial that vanishes on a, a^2, \dots, a^{k-1} . Then $f(t) = \prod_{i=1}^{k-1} (t - a^i)$. Because the set of zeros of f is a geometric sequence, it turns out that the coefficients of f also grow rapidly, and this enables us to give a lower bound on C .

Proof of Theorem 4.5. Following the overview above, the distributions will be based on the evaluations of certain multivariate polynomials at a particular point. Specifically, for every $0 \leq i \leq k-1$ let $s_i(y_1, \dots, y_{k-1})$ be the i th symmetric function

$$s_i(y_1, \dots, y_{k-1}) = \sum_{\substack{T \subseteq [k-1] \\ |T|=i}} \prod_{j \in T} y_j. \quad (1)$$

For example, if $k = 4$ and $i = 2$ then $s_2(y_1, \dots, y_{k-1}) = y_1y_2 + y_1y_3 + y_2y_3$. In general, $s_0 = 1$ and $s_{k-1}(y_1, \dots, y_{k-1}) = y_1 \cdot \dots \cdot y_{k-1}$.

Let a be some integer that we fix later, and define $s_i(a) \stackrel{\text{def}}{=} s_i(a, a^2, \dots, a^{k-1})$. Following our previous example, $s_2(a) = a^3 + a^4 + a^5$, while $s_3(a) = a^6$. In general, as we shall prove in detail, $s_i(a)$ is larger than $s_{i-1}(a)$ for sufficiently large a .

Consider the polynomial $f(t) = \prod_{i=1}^{k-1} (t - a^i)$. It is easy to see that $\sum_{i=0}^{k-1} (-1)^i \cdot s_{k-1-i}(a) \cdot t^i$. We construct two distributions from the coefficients of f . The supports of the distributions are contained in the set $\{1, a, a^2, \dots, a^{k-1}\}$. We define

$$\forall i, 0 \leq i \leq k-1 \quad \Pr[\hat{X} = a^i] = \begin{cases} s_{k-1-i}(a) / \hat{N}(a) & \text{for even } i \\ 0 & \text{for odd } i \end{cases} \quad (2)$$

$$\forall i, 0 \leq i \leq k-1 \quad \Pr[\tilde{X} = a^i] = \begin{cases} 0 & \text{for even } i \\ s_{k-1-i}(a) / \hat{N}(a) & \text{for odd } i \end{cases} \quad (3)$$

where $\hat{N}(a) \stackrel{\text{def}}{=} \sum_{j=0}^{\lfloor (k-1)/2 \rfloor} s_{k-1-2j}(a)$ and $\tilde{N}(a) \stackrel{\text{def}}{=} \sum_{j=0}^{\lfloor (k-2)/2 \rfloor} s_{k-2-2j}(a)$ are normalization factors.

We now show that for an appropriate choice of the parameter a , the distributions \hat{X} and \tilde{X} satisfy the second part of the moments condition. Let $C \stackrel{\text{def}}{=} \hat{N}(a)/\tilde{N}(a)$.

Lemma 4.6 \hat{X} and \tilde{X} satisfy: $C \cdot \mathbb{E}[\hat{X}^\ell] = \mathbb{E}[\tilde{X}^\ell]$ for $\ell = 1, \dots, k-1$.

Proof: By definition of \hat{X} and \tilde{X} ,

$$\begin{aligned} C \cdot \mathbb{E}[\hat{X}^\ell] - \mathbb{E}[\tilde{X}^\ell] &= C \cdot \sum_{\substack{0 \leq i \leq k-1 \\ i \text{ even}}} (a^i)^\ell \cdot \frac{s_{k-1-i}(a)}{\hat{N}(a)} - \sum_{\substack{0 \leq i \leq k-1 \\ i \text{ odd}}} (a^i)^\ell \cdot \frac{s_{k-1-i}(a)}{\tilde{N}(a)} \\ &= \frac{1}{\tilde{N}(a)} \cdot \sum_{i=0}^{k-1} (-1)^i \cdot s_{k-1-i}(a) \cdot (a^\ell)^i = \frac{(-1)^{k-1}}{\tilde{N}(a)} \cdot f(a^\ell) = 0. \quad \blacksquare \end{aligned} \tag{4}$$

The following lemma, proved in Appendix B, bounds $\mathbb{E}[\tilde{X}]$, $\mathbb{E}[\hat{X}]$ and the ratio between them.

Lemma 4.7 (1) $\mathbb{E}[\hat{X}] < 1 + \frac{1}{a-3}$. (2) $\mathbb{E}[\tilde{X}] > a-2$.

It remains to find, for every $B > 1$, an a such that $\mathbb{E}[\tilde{X}]/\mathbb{E}[\hat{X}] \geq B$. By Lemma 4.7, $\frac{\mathbb{E}[\tilde{X}]}{\mathbb{E}[\hat{X}]} > \frac{a-2}{1+1/(a-3)} = a-3$. Thus, if we take $a = B+3$ then $\mathbb{E}[\tilde{X}]/\mathbb{E}[\hat{X}] > B$, $\mathbb{E}[\hat{X}] < 1 + \frac{1}{B}$ and $\mathbb{E}[\tilde{X}] > B$. This completes the construction and the proof of Theorem 4.5. \blacksquare

5 Indistinguishability by Poisson Algorithms

Even though uniform algorithms are much simpler than general algorithms, they still might be tricky to analyze because of dependencies between the numbers of balls of various colors that appear in the sample. Batu *et al.* [5, conference version] noted that such dependencies are avoided when an algorithm takes a random number of samples according to a *Poisson* distribution. The Poisson distribution $\text{Po}(\lambda)$ takes on the value $x \in \mathbb{N}$ with probability $e^{-\lambda} \lambda^x / x!$.

Definition 5.1 We call a uniform algorithm *Poisson- s* if the number of samples it takes is a random variable, distributed as $\text{Po}(s)$.

To prove a lower bound for DE, it is sufficient to consider Poisson algorithms that get only the histogram of the sample as their input. This is justified by Lemma D.1 (see Appendix D). Intuitively, a Poisson algorithm can simulate a general algorithm with only a small loss in parameters.

As we explained, we prove Theorem 2.1 by constructing a pair of instances that are hard to distinguish. They correspond to the pair of frequency variables satisfying the moments condition that we constructed in the proof of Theorem 4.5. Defining DE instances based on frequency variables is straightforward if we make an integrality assumption described below. Specifically, for $k > 1$ let $a_0 < a_1 < \dots < a_{k-1}$ be integers, and let X be a random variable over these integers with $\Pr[X = a_i] = p_i$. Then $\mathbb{E}[X] = \sum_{i=0}^{k-1} p_i \cdot a_i$.

Based on X , we define a DE instance D_X of length n (that is a string in $[n]^n$) that contains $\frac{n}{\mathbb{E}[X]}$ colors. For $i = 0, \dots, k-1$, D_X contains $\frac{np_i}{\mathbb{E}[X]}$ colors of type i , where each color of type i appears a_i times. See Appendix C for a general treatment, without the assumption that $\frac{np_i}{\mathbb{E}[X]}$ is an integer.

Our next main building block in the proof of Theorem 2.1, is the theorem stated below. It shows that if two distributions \hat{X}, \tilde{X} over integers have $k-1$ proportional moments, then the corresponding instances of DE, $D_{\hat{X}}$ and $D_{\tilde{X}}$ cannot be distinguished by a Poisson algorithm that looks only at histograms and uses fewer than about $n^{1-\frac{1}{k}}$ samples. In fact, the bound is more complicated, since it depends on how the maximum value, a_{k-1} , in the support of \hat{X} and \tilde{X} varies as n increases.

Theorem 5.2 (Distinguishability by Poisson Algorithms) *Let \hat{X}, \tilde{X} be random variables over positive integers $a_0 < a_1 < \dots < a_{k-1}$ which have $k-1$ proportional moments. For any Poisson algorithm \mathcal{A} that looks only at histograms and takes $s \leq \frac{n}{2 \cdot a_{k-1}}$ samples in expectation,*

$$\left| \Pr[\mathcal{A}(D_{\hat{X}}) = 1] - \Pr[\mathcal{A}(D_{\tilde{X}}) = 1] \right| = O \left(k^2 a_{k-1}^{2/3} \left(\frac{s}{n} \right)^{2/3} + \frac{k}{\lfloor \frac{k}{2} \rfloor! \cdot \lceil \frac{k}{2} \rceil!} \cdot a_{k-1}^{k-1} \cdot \frac{s^k}{n^{k-1}} \right).$$

The generality of this bound is required to prove Theorem 2.1. However, the following corollary is sufficient to show that additive approximations for DE require a near-linear number of samples, and it is considerably simpler to read.

Corollary 5.3 *Let \hat{X} and \tilde{X} be fixed (w.r.t. n) random variables which have $k-1$ proportional moments. If $s = o(n^{1-\frac{1}{k}})$, then for any Poisson- s algorithm \mathcal{A} , we have $|\Pr[\mathcal{A}(D_{\hat{X}}) = 1] - \Pr[\mathcal{A}(D_{\tilde{X}}) = 1]| = o(1)$.*

We now turn to proving Theorem 5.2. Recall that two random variables X and Y over a domain S have *statistical difference* δ if $\max_{S' \subset S} |\Pr[X \in S'] - \Pr[Y \in S']| = \delta$. Equivalently, for every algorithm \mathcal{A} , $|\Pr[\mathcal{A}(X) = 1] - \Pr[\mathcal{A}(Y) = 1]| \leq \delta$. We write $X \approx_\delta Y$ to indicate that X and Y have statistical difference at most δ .

For $\ell = 0, 1, \dots, s$, let F_ℓ be a random variable representing the number of ℓ -way collisions a Poisson- s algorithm sees (recall Definition 4.1), and let $F = (F_1, F_2, F_3 \dots)$ be the corresponding histogram. We can restate Theorem 5.2 in terms of histograms:

Theorem 5.4 (Distinguishability by Poisson Algorithms, Restated) *For $s \leq \frac{n}{2 \cdot a_{k-1}}$, the statistical difference between the histograms $(\hat{F}_1, \hat{F}_2, \hat{F}_3, \dots)$ and $(\tilde{F}_1, \tilde{F}_2, \tilde{F}_3, \dots)$ is at most*

$$O \left(k^2 a_{k-1}^{2/3} \left(\frac{s}{n} \right)^{2/3} + \frac{k}{\lfloor \frac{k}{2} \rfloor! \cdot \lceil \frac{k}{2} \rceil!} \cdot a_{k-1}^{k-1} \cdot \frac{s^k}{n^{k-1}} \right).$$

For the remainder of this section, assume $s \leq \frac{n}{2 \cdot a_{k-1}}$. The proof of Theorem 5.4 relies on the three following lemmas, proved in Appendix D. Lemma 5.5 states that ℓ -way collisions are very unlikely for $\ell \geq k$, when s is sufficiently small. Lemma 5.6 shows that for both instances, the distribution on histograms is close to the product of its marginal distributions, that is, the components of the histogram are close to being independent. Finally, Lemma 5.7 shows that k -way collisions have close distributions for DE instances $D_{\hat{X}}$ and $D_{\tilde{X}}$.

Lemma 5.5 *For both distributions, the probability of a collision involving $k > 1$ or more balls is at most*

$$\delta_1 = O \left(\frac{a_{k-1}^{k-1}}{k!} \cdot \frac{s^k}{n^{k-1}} \right).$$

Lemma 5.6 For both distributions, F_1, \dots, F_{k-1} are close to independent, that is, $(F_1, \dots, F_{k-1}) \approx_{\delta_2} (F'_1, \dots, F'_{k-1})$, where the variables F'_ℓ are independent, for each ℓ the distributions of F_ℓ and F'_ℓ are identical, and $\delta_2 = O(k^2 \cdot (\frac{a_{k-1} \cdot s}{n})^{2/3})$.

Lemma 5.7 Let $\delta_3 = O\left(\frac{k \cdot a_{k-1} \cdot s}{n} + \frac{1}{\lfloor \frac{k}{2} \rfloor! \cdot \lceil \frac{k}{2} \rceil!} \cdot \left(\frac{a_{k-1}}{n}\right)^{k-1} \cdot s^k\right)$. Then $\hat{F}_\ell \approx_{\delta_3} \tilde{F}_\ell$ for $\ell \in [k-1]$.

The fact that \hat{X} and \tilde{X} have proportional moments is used in the proof of Lemma 5.7 (the other two lemmas hold as long as the a_i 's are bounded). The main idea of the proof is to approximate F_ℓ by a Poisson random variable with the same expectation, and show that the moment conditions imply that \hat{F}_ℓ and \tilde{F}_ℓ have similar (though not equal) expectations. The proof is quite technical (see Appendix D). However, given the three lemmas above, we can easily prove the main result of the section:

Proof of Theorem 5.4. The proof follows by a hybrid argument. Consider a chain of distributions “between” the two histograms of Theorem 5.4. Starting from the “hat” histogram, we first replace all counts of collisions greater than k by 0, and then replace each count \hat{F}_ℓ with an independent copy \hat{F}'_ℓ for $\ell \in [k-1]$, as in Lemma 5.6. Next, change each \hat{F}'_ℓ with a corresponding \tilde{F}'_ℓ . Finally, replace these independent \tilde{F}'_ℓ s with the real, dependent variables \tilde{F}_ℓ and add back in the counts of the collisions involving more than k variables to obtain the “tilde” histogram. The resulting chain has $k+3$ steps. By the triangle inequality, we can sum these differences to obtain a bound on the difference between the two histograms. In symbols, the chain of distribution looks as follows (where δ_1, δ_3 and δ_2 are as defined in Lemmas 5.5, 5.7 and 5.6, respectively):

$$\begin{aligned} (\hat{F}_1, \dots, \hat{F}_{k-1}, \hat{F}_k, \dots) &\approx_{\delta_1} (\hat{F}_1, \dots, \hat{F}_{k-1}, 0, \dots) \approx_{\delta_2} (\hat{F}'_1, \dots, \hat{F}'_{k-1}, 0, \dots) \approx_{\delta_3} (\tilde{F}'_1, \dots, \tilde{F}'_{k-1}, 0, \dots) \\ &\approx_{\delta_3} \dots \approx_{\delta_3} (\tilde{F}'_1, \dots, \tilde{F}'_{k-1}, 0, \dots) \approx_{\delta_2} (\tilde{F}_1, \dots, \tilde{F}_{k-1}, 0, \dots) \approx_{\delta_1} (\tilde{F}_1, \dots, \tilde{F}_{k-1}, \tilde{F}_k, \dots) \end{aligned}$$

The total statistical difference is at most $2 \cdot \delta_1 + 2 \cdot \delta_2 + (k-1) \cdot \delta_3$

$$= O\left(\frac{1}{k!} \cdot \left(\frac{a_{k-1}}{n}\right)^{k-1} \cdot s^k + k^2 \cdot \left(\frac{a_{k-1} \cdot s}{n}\right)^{2/3} + k \cdot \frac{k \cdot a_{k-1} \cdot s}{n} + \frac{k}{\lfloor \frac{k}{2} \rfloor! \cdot \lceil \frac{k}{2} \rceil!} \cdot \left(\frac{a_{k-1}}{n}\right)^{k-1} \cdot s^k\right).$$

The first and third terms are negligible given the others. Removing them yields the claimed bound. \blacksquare

6 Proof of Main Lower Bound (Theorem 2.1)

We now prove the main lower bound (Theorem 2.1) by combining the construction of distributions satisfying the moments condition (Theorem 4.5) with the bound on distinguishability by Poisson algorithms (Theorem 5.2) and the reductions to Uniform algorithms (Lemma 3.2), and to Poisson algorithms (Lemma D.1).

Recall that our goal is to give a lower bound on the number of queries required for a general algorithm to distinguish inputs with at least $n/11$ colors from inputs with at most n/B colors (for $B > 11$). By combining Lemmas D.1 and 3.2 it suffices to give a lower bound on s for a Poisson- s algorithm that uses only the histogram of the samples and distinguishes inputs with at least $\frac{10}{11}n$ colors from inputs with at most n/B colors (the main source of loss is Lemma 3.2). Details follow.

Let \hat{X} and \tilde{X} obey the moments condition with parameters k and B , and let $D_{\tilde{X}}$ and $D_{\hat{X}}$ be the corresponding DE instances. By Theorem 4.5 these instances have at least $n(1 - \frac{1}{B}) > \frac{10}{11}n$ and at most n/B colors, respectively. (Here we continue to assume for simplicity that $\frac{np_i}{E[\tilde{X}]}$ is an integer for all i and both distributions.) We now turn to bounding the statistical difference of the corresponding histogram distributions.

Consider any Poisson algorithm \mathcal{A} that looks only at histograms and takes $\frac{s}{2}$ samples (where the choice of $\frac{s}{2}$ rather than s samples is made for the convenience of the analysis). Recall that Theorem 4.5 stated that there exist \hat{X} and \tilde{X} such that $a_{k-1} = (B+3)^{k-1} < (B+3)^k$, so we may assume that this is in fact the case. By substituting this bound in Theorem 5.2, we get:

$$\left| \Pr[\mathcal{A}(D_{\hat{X}}) = 1] - \Pr[\mathcal{A}(D_{\tilde{X}}) = 1] \right| = O \left(k^2 \left(\frac{(B+3)^k \cdot s}{n} \right)^{2/3} + \frac{k}{\lfloor \frac{k}{2} \rfloor! \cdot \lfloor \frac{k}{2} \rfloor!} \cdot \frac{(B+3)^{k(k-1)} \cdot s^k}{n^{k-1}} \right). \quad (5)$$

We set k and s as functions of B so that the error term in Equation (5) is bounded from above by $o(1)$. Given B , define q by the equality $B = \log(n)^q$. Set $k = \left\lfloor \sqrt{\frac{\log(n)}{(q+\frac{1}{2}) \log \log(n)}} \right\rfloor$, and $s = \lfloor n^{1-\frac{2}{k}} \rfloor$. To ensure $s \geq 1$, we need $k > 2$, so we restrict q to be $0 < q < \frac{\log n}{4 \log \log n} - \frac{1}{2}$. In particular, B is bounded from above by $n^{\frac{1}{4}} / \sqrt{\log n}$. To make the calculations easier assume that $n > 16$, so that $k < \sqrt{\log n}$. We handle the two summands in Equation (5) separately. By substituting k , s , and B as set above in the first summand, $k^2 \left(\frac{(B+3)^k s}{n} \right)^{2/3}$, it can be shown that it is upper-bounded by $2^{-\frac{1}{3}} \sqrt{\log \log(n) \log(n)}$. The second summand, $\frac{k}{\lfloor \frac{k}{2} \rfloor! \cdot \lfloor \frac{k}{2} \rfloor!} \cdot \frac{(B+3)^{k(k-1)} s^k}{n^{k-1}}$ is upper-bounded by $2^{-\frac{1}{2}} \sqrt{\log \log(n) \log(n)}$. By Equation (5) and these two bounds, $\left| \Pr[\mathcal{A}(D_{\hat{X}}) = 1] - \Pr[\mathcal{A}(D_{\tilde{X}}) = 1] \right| = O \left(2^{-\frac{1}{3}} \sqrt{\log \log(n) \log(n)} \right)$. This completes the proof of Theorem 2.1.

7 A Lower Bound for Approximating the Entropy

The following problem was introduced by Batu *et al.* [5]. Let $\mathbf{p} = \langle p_1, \dots, p_n \rangle$ be a discrete distribution over n elements, where p_i is the probability of the i th element. Given access to independent samples generated according to the distribution \mathbf{p} , we would like to approximate its entropy: $H(\mathbf{p}) = -\sum_{i=1}^n p_i \log p_i$. Batu *et al.* showed how to obtain an α -factor approximation in time $\tilde{O} \left(n^{\frac{1+\eta}{\alpha^2}} \right)$, provided that $H(\mathbf{p}) = \Omega \left(\frac{\alpha}{\eta} \right)$. They also proved a lower bound of $\Omega \left(n^{\frac{1}{2\alpha^2}} \right)$ that holds even when $H(\mathbf{p}) = \Omega \left(\frac{\log n}{\alpha^2} \right)$. (Without a lower bound on $H(\mathbf{p})$, the time complexity is unbounded.)

Here we use our technique to obtain a lower bound of $\Omega \left(n^{\frac{2}{6\alpha^2-3+o(1)}} \right)$, improving on the $\Omega \left(n^{\frac{1}{2\alpha^2}} \right)$ lower bound for relatively small α . When α is close to 1, the bound is close to $n^{2/3}$ (rather than $n^{1/2}$).

We first provide a different construction of random variables that satisfy the moments condition (Definition 4.4) for the special case of $k = 3$. This much simpler construction gives random variables with support on smaller integers than the more general construction in Theorem 4.5, leading to better bounds.

Lemma 7.1 (R.V.'s Satisfying the Moments Condition with $k = 3$) *For all integers $B > 1$, there exist random variables \hat{X} and \tilde{X} over $a_0 = 1, a_1 = 2B, a_2 = 4B - 2$ that satisfy the moments condition with parameters 3 and B . Moreover, for these variables, $\mathbb{E}[\tilde{X}] = 2$ and $\mathbb{E}[\hat{X}] = 2B$.*

Proof: Set $\Pr[\hat{X} = a_0] = 1 - \frac{1}{4B-3}$, $\Pr[\hat{X} = a_1] = 0$, $\Pr[\hat{X} = a_2] = \frac{1}{4B-3}$, and $\Pr[\tilde{X} = a_0] = 0$, $\Pr[\tilde{X} = a_1] = 1$, $\Pr[\tilde{X} = a_2] = 0$. By definition of \hat{X} and \tilde{X} , $\mathbb{E}[\hat{X}] = 2$, $\mathbb{E}[\hat{X}^2] = 4B$, while $\mathbb{E}[\tilde{X}] = 2B$ and $\mathbb{E}[\tilde{X}^2] = 4B^2$, so that $\frac{\mathbb{E}[\tilde{X}]}{\mathbb{E}[\hat{X}]} = B$, and $\frac{\mathbb{E}[\tilde{X}^2]}{\mathbb{E}[\hat{X}^2]} = \frac{\mathbb{E}[\tilde{X}^2]}{\mathbb{E}[\hat{X}^2]}$, as required. ■

The two distributions and their entropies. Similarly to what was shown in Section 5, given the two random variables \hat{X} and \tilde{X} , define two distributions over n elements (or, more precisely, two families of distributions). One distribution, denoted $\mathbf{p}_{\hat{X}}$, has support on $\frac{n}{2} \cdot \frac{4B-4}{4B-3}$ elements of weight $\frac{1}{n}$ each and $\frac{n}{2} \cdot \frac{1}{4B-3}$ elements of weight $\frac{4B-2}{n}$ each. The second distribution, denoted $\mathbf{p}_{\tilde{X}}$, has support on $\frac{n}{2B}$ elements of weight $\frac{2B}{n}$ each. As stated above, we can define two families of distributions, $F_{\hat{X}}$ and $F_{\tilde{X}}$, respectively, where we allow all permutations over the names (colors) of the elements. Let $D'_{\hat{X}}$ denote the uniform distribution over $F_{\hat{X}}$, and let $D'_{\tilde{X}}$ denote the uniform distribution over $F_{\tilde{X}}$.

Let $B = B(n)$ be of the form $B = \frac{1}{2}n^{1-\beta}$ for $\beta < 1$. Then the entropy of each distribution in $F_{\tilde{X}}$ is $\beta \log n$, and the entropy of each distribution in $F_{\hat{X}}$ is $\frac{2B-2}{4B-3} \cdot \log n + \frac{2B-1}{4B-3} \cdot \log \frac{n}{4B-2}$, which is lower bounded by $\frac{1+\beta}{2} \log n - 1$ by our choice of B . Thus, the ratio between the entropies is $\frac{1+\beta}{2\beta} - o(1)$.

While Theorem 5.2 was stated for the distributions on strings, $D_{\hat{X}}$ and $D_{\tilde{X}}$, and any algorithm that takes uniform samples from an input string of length n , it is not hard to verify that it also holds for the distributions $D'_{\hat{X}}$ and $D'_{\tilde{X}}$ and any algorithm that is provided with samples from distributions over n elements. Since $k = 3$ and $a_2 = 2n^{1-\beta}$, in order to distinguish the two distributions it is necessary to observe $\Omega\left(\left(\frac{n}{a_2}\right)^{2/3}\right) = \Omega\left(n^{2\beta/3}\right)$ samples. In other words, $\Omega\left(n^{2\beta/3}\right) = \Omega\left(n^{\frac{2}{6\alpha^2-3+o(1)}}\right)$ samples are required for $\alpha = \left(\sqrt{\frac{1+\beta}{2\beta}} - o(1)\right)$ -estimating the entropy.

Acknowledgments. We would like to thank Oded Goldreich, Omer Reingold and Ronitt Rubinfeld for helpful comments.

References

- [1] A. Akella, A. R. Bharambe, M. Reiter, and S. Seshan. Detecting DDoS attacks on ISP networks. In *Proceedings of the Workshop on Management and Processing of Data Streams*, 2003.
- [2] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *J. Comput. Syst. Sci.*, 58(1):137–147, 1999.
- [3] Ziv Bar-Yossef. *The complexity of Massive Data Set Computations*. PhD thesis, Computer Science Division, U.C. Berkeley, 2002.
- [4] Ziv Bar-Yossef, Ravi Kumar, and D. Sivakumar. Sampling algorithms: lower bounds and applications. In *STOC '01: Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 266–275, New York, NY, USA, 2001. ACM Press.
- [5] Tugkan Batu, Sanjoy Dasgupta, Ravi Kumar, and Ronitt Rubinfeld. The complexity of approximating the entropy. *SIAM Journal on Computing*, 35(1):132–150, 2005.
- [6] Tugkan Batu, Eldar Fischer, Lance Fortnow, Ravi Kumar, Ronitt Rubinfeld, and Patrick White. Testing random variables for independence and identity. In *Proceedings of the Forty-Second Annual Symposium on Foundations of Computer Science*, pages 442–451, 2001.
- [7] Tugkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren Smith, and Patrick White. Testing that distributions are close. In *Proceedings of the Forty-First Annual Symposium on Foundations of Computer Science*, pages 259–269, 2000.

- [8] John Bunge. Bibliography on estimating the number of classes in a population. www.stat.cornell.edu/~bunge/bibliography.htm.
- [9] Moses Charikar, Surajit Chaudhuri, Rajeev Motwani, and Vivek R. Narasayya. Towards estimation error guarantees for distinct values. In *PODS*, pages 268–279. ACM, 2000.
- [10] Thomas Cover and Joy Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., New York, 1991.
- [11] Piotr Indyk and David Woodruff. Tight lower bounds for the distinct elements problem. In *Proceedings of the Forty-Forth Annual Symposium on Foundations of Computer Science*, pages 283–288, 2003.
- [12] Yu. V. Prohorov. Asymptotic behavior of the binomial distribution (Russian). *Uspekhi Matematicheskikh Nauk*, 8(3):135–142, 1953. Moscow.
- [13] Sofya Raskhodnikova, Dana Ron, Ronitt Rubinfeld, Amir Shpilka, and Adam Smith. Sublinear algorithms for approximating string compressibility. Manuscript, 2007.
- [14] Wojciech Szpankowski. *Average Case Analysis of Algorithms on Sequences*. John Wiley & Sons, Inc., New York, 2001.
- [15] Michael Weba. Bounds for the total variation distance between the binomial and the poisson distribution in case of medium-sized success probabilities. *J. Appl. Probab.*, 36(1):97–104, 1999.
- [16] Jacob Ziv and Abraham Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23:337–343, 1977.

A Algorithms for DE with Uniform Samples

In this section we show that restricted algorithms that take samples uniformly at random with replacement are essentially as good for DE as general algorithms.

First, consider algorithms that take their samples uniformly at random *without replacement* from $[n]$. The following lemma, appearing in Bar-Yossef’s thesis [3, Page 88], shows that such algorithms are essentially as good for solving DE as general algorithms.

Lemma A.1 ([3]) *For any function invariant under permutations of input elements (ball positions), any algorithm that makes s queries can be simulated by an algorithm that takes s samples uniformly at random without replacement and has the same guarantees on the output as the original algorithm.*

The main idea in the proof of the lemma is that the new algorithm, given input w , can simulate the old algorithm on $\pi(w)$, where π is a random permutation of the input, dictated by the random samples chosen by the new algorithm. Since the value of the function (in our case, the number of colors) is the same for w and $\pi(w)$, the guarantees on the old algorithm hold for the new one.

Next, we would like to go from the algorithms that sample uniformly *without replacement* to the ones that sample uniformly *with replacement* and find out the corresponding color, but not the input position that was queried. Bar-Yossef proved that for all functions invariant under permutations, algorithms that take $O(\sqrt{n})$ uniform samples *without replacement* can be simulated by algorithms that take the same number of samples *with replacement*. The idea is that with so few samples, an algorithm sampling *with replacement*

is likely to never look at the same input position twice. To prove a statement along the same lines for algorithms that take more samples, Bar-Yossef allows them to see not only the color of each sample, but also which input position was queried (this allows the algorithm to ignore replaced samples). One can avoid giving this extra information to an algorithm for DE, with a slight loss in the approximation factor. Recall that we call an algorithm **uniform** if it takes independent samples with replacement and only gets to see the colors of the samples, but not the input positions corresponding to them.

Lemma A.2 *Let $\alpha = \alpha(n)$, such that $\sqrt{0.1} \cdot \alpha \geq 1$. For every algorithm \mathcal{A} that makes s queries, and provides, with probability at least $\frac{11}{12}$, an approximation for DE that is within a multiplicative factor of at most $(\sqrt{0.1} \cdot \alpha)$ from the correct value, there is a uniform algorithm \mathcal{A}' that takes s samples and provides, with probability at least $\frac{2}{3}$, an approximation for DE that is within a multiplicative factor of at most α from the correct value.*

A rephrasing of Lemma A.2 (using a few details concerning the reduction in the proof of the lemma) gives us Lemma 3.2.

Proof of Lemma A.2. Conduct the following mental experiment: let algorithm \mathcal{A}' generate an instance of DE by taking n uniform samples from its input and recording their colors. If there are $C = C(n)$ colors in the input of \mathcal{A}' , the generated instance will have at most C colors. However, some of the colors might be missing. We will show later (see Claim A.3 with s set to n) that with probability $\geq \frac{3}{4}$ at least $0.1 \cdot C$ colors appear in the instance. That is, with probability $\geq \frac{3}{4}$, the instance generated in our mental experiment will have between $0.1 \cdot C$ and C colors. When \mathcal{A} is run on that instance, with probability $\geq \frac{11}{12}$, it will output an answer between $\frac{0.1 \cdot C}{\sqrt{0.1} \cdot \alpha} = \sqrt{0.1} \cdot \frac{C}{\alpha}$ and $\sqrt{0.1} \cdot \alpha \cdot C$. Thus, if \mathcal{A}' runs \mathcal{A} on this instance and multiplies its answer by $\sqrt{10}$, it will get an α -multiplicative approximation to C with probability $\geq 1 - \frac{1}{4} - \frac{1}{12} \geq \frac{2}{3}$, as promised. The final observation is that since each color in the instance is generated independently, \mathcal{A}' can run \mathcal{A} on that instance, generating colors on demand, resulting in s samples instead of n . ■

Claim A.3 *Let $s = s(n) \leq n$. Then s independent samples from a distribution with $C = C(n)$ elements, where each element has probability $\geq \frac{1}{n}$, yield at least $\frac{Cs}{10n}$ distinct elements, with probability $\geq \frac{3}{4}$.*

Proof: For $i \in [C]$, let X_i be the indicator variable for the event that color i is selected in s samples. Then $X = \sum_{i=1}^C X_i$ is a random variable for the number of distinct colors. Since each color is selected with probability at least $\frac{1}{n}$ for each sample,

$$\mathbb{E}[X] = \sum_{i=1}^C \mathbb{E}[X_i] \geq C \left(1 - \left(1 - \frac{1}{n} \right)^s \right) \geq C \left(1 - e^{-(s/n)} \right) \geq (1 - e^{-1}) \frac{Cs}{n}. \quad (6)$$

The last inequality holds because $1 - e^{-x} \geq (1 - e^{-1}) \cdot x$ for all $x \in [0, 1]$.

We now use Chebyshev's inequality to bound the probability that X is far from its expectation. For any distinct pair of colors i, j , the covariance $\mathbb{E}(X_i X_j) - \mathbb{E}(X_i) \mathbb{E}(X_j)$ is negative (knowing that one color was not selected makes it more likely for any other color to be selected). Since X is a sum of Bernoulli variables, $\text{Var}[X] \leq \mathbb{E}[X]$. For any fixed δ ,

$$\Pr[X \leq \delta \mathbb{E}[X]] \leq \Pr[|X - \mathbb{E}[X]| \geq (1 - \delta) \mathbb{E}[X]] \leq \frac{\text{Var}[X]}{((1 - \delta) \mathbb{E}[X])^2} \leq \frac{1}{(1 - \delta)^2 \mathbb{E}[X]}. \quad (7)$$

Set $\delta = 3 - \sqrt{8}$. If $\mathbb{E}[X] \geq \frac{4}{(1 - \delta)^2}$, then by Equations (7,6), with probability $\geq \frac{3}{4}$, variable $X \geq \delta \mathbb{E}[X] \geq \delta(1 - e^{-1}) \frac{Cs}{n} > \frac{Cs}{10n}$, as stated in the claim. Otherwise, that is, if $\mathbb{E}[X] < \frac{4}{(1 - \delta)^2}$, Equation (6)

implies that $\frac{4\delta}{(1-\delta)^2} > \delta(1-e^{-1})\frac{Cs}{n}$. Substituting $3 - \sqrt{8}$ for δ gives $1 > \frac{Cs}{10n}$. In other words, the claim for this case is that at least one color appears among the samples, which, clearly, always holds. ■

B Proof of Lemma 4.7

We first restate Lemma 4.7.

Lemma 4.7 (1) $E[\hat{X}] < 1 + \frac{1}{a-3}$. (2) $E[\tilde{X}] > a - 2$.

In order to prove Lemma 4.7 we shall need the following upper bound on the value of $s_i(a)$.

Lemma B.1 *We have the following bounds on the $s_i(a)$'s:*

1. $s_{k-1}(a)/a < s_{k-2}(a) < s_{k-1}(a)/(a-1)$.
2. For every $3 \leq i \leq k-1$, we have that $s_{k-i}(a) < s_{k-1}(a) \cdot \frac{1}{a^{\lfloor \frac{i-1}{2} \rfloor} (a-1)^{i-1}}$.

Proof: By the definition of $s_i(a)$ we get that

$$s_i(a) = s_i(a, a^2, \dots, a^{k-1}) = \sum_{\substack{T \subset [k-1] \\ |T|=i}} \prod_{j \in T} a^j \quad (8)$$

$$= \left(\prod_{j=1}^{k-1} a^j \right) \cdot \sum_{\substack{T \subset [k-1] \\ |T|=i}} \prod_{j \notin T} a^{-j} \quad (9)$$

$$= s_{k-1}(a) \cdot \sum_{\substack{T \subset [k-1] \\ |T|=i}} \prod_{j \notin T} a^{-j} \quad (10)$$

$$= s_{k-1}(a) \cdot \sum_{\substack{R \subset [k-1] \\ |R|=k-1-i}} \prod_{j \in R} a^{-j} \quad (11)$$

In particular we get that

$$s_{k-2}(a) = s_{k-1}(a) \cdot \left(\sum_{j=1}^{k-1} a^{-j} \right)$$

Since

$$\frac{1}{a} < \left(\sum_{j=1}^{k-1} a^{-j} \right) < \frac{1}{a-1}$$

the first part of the lemma follows.

We now prove the second part of the lemma. By Equation (8), it suffices to prove that

$$\sum_{\substack{R \subset [k-1] \\ |R|=i-1}} \prod_{j \in R} a^{-j} < \frac{1}{a^{\lfloor \frac{i-1}{2} \rfloor} (a-1)^{i-1}}$$

$$\sum_{\substack{R \subset [k-1] \\ |R|=i-1}} \prod_{j \in R} a^{-j} = \left(\sum_{1 \leq j_1 < j_2 < \dots < j_{i-1} \leq k-1} \prod_{\ell=1}^{i-1} a^{-j_\ell} \right) \quad (12)$$

$$< \left(\sum_{1 \leq j_1 < j_2 \leq k-1} a^{-j_1 - j_2} \right)^{\lfloor \frac{i-1}{2} \rfloor} \cdot \left(\sum_{\ell=1}^{k-1} a^{-\ell} \right)^{i-1-2\lfloor \frac{i-1}{2} \rfloor} \quad (13)$$

$$< \left(\sum_{j=3}^{k-1} \left\lfloor \frac{j-1}{2} \right\rfloor \cdot a^{-j} \right)^{\lfloor \frac{i-1}{2} \rfloor} \cdot \frac{1}{(a-1)^{i-1-2\lfloor \frac{i-1}{2} \rfloor}} \quad (14)$$

$$< \frac{1}{(a-1)^{i-1-2\lfloor \frac{i-1}{2} \rfloor}} \cdot \left(\frac{1}{2} \cdot \sum_{j=3}^{\infty} (j-1) \cdot a^{-j} \right)^{\lfloor \frac{i-1}{2} \rfloor} \quad (15)$$

$$= \frac{1}{(a-1)^{i-1-2\lfloor \frac{i-1}{2} \rfloor}} \cdot \left[\frac{1}{2} \cdot \frac{d}{da} \left(- \sum_{j=2}^{\infty} a^{-j} \right) \right]^{\lfloor \frac{i-1}{2} \rfloor} \quad (16)$$

$$= \frac{1}{(a-1)^{i-1-2\lfloor \frac{i-1}{2} \rfloor}} \cdot \left[-\frac{1}{2} \cdot \frac{d}{da} \left(\frac{1}{a-1} - \frac{1}{a} \right) \right]^{\lfloor \frac{i-1}{2} \rfloor} \quad (17)$$

$$= \frac{1}{(a-1)^{i-1-2\lfloor \frac{i-1}{2} \rfloor}} \cdot \left[\frac{1}{2} \cdot \left(\frac{1}{(a-1)^2} - \frac{1}{a^2} \right) \right]^{\lfloor \frac{i-1}{2} \rfloor} \quad (18)$$

$$< \frac{1}{(a-1)^{i-1-2\lfloor \frac{i-1}{2} \rfloor}} \cdot \left(\frac{1}{a(a-1)^2} \right)^{\lfloor \frac{i-1}{2} \rfloor} \quad (19)$$

$$< \frac{1}{a^{\lfloor \frac{i-1}{2} \rfloor} (a-1)^{i-1}}. \quad (20)$$

Equation (12), which is the main step in the above sequence of equations, is obtained by pairing up all $i-1$ indices (but at most 1), and ignoring the constraint that the pairs should be distinct. ■

Proof of Lemma 4.7. By definition of \tilde{X} ,

$$E[\hat{X}] = \frac{1}{\hat{N}(a)} \cdot \sum_{j=0}^{\lfloor (k-1)/2 \rfloor} s_{k-1-2j}(a) a^{2j} \quad (21)$$

$$< \frac{1}{\hat{N}(a)} \cdot s_{k-1}(a) \cdot \left(1 + \sum_{j=1}^{\lfloor (k-1)/2 \rfloor} \frac{a^{2j}}{a^j (a-1)^{2j}} \right) \quad (22)$$

$$< \frac{1}{\hat{N}(a)} \cdot s_{k-1}(a) \cdot \left(1 + \sum_{j=1}^{\lfloor (k-1)/2 \rfloor} \frac{1}{(a-2)^j} \right) \quad (23)$$

$$< \frac{1}{\hat{N}(a)} \cdot s_{k-1}(a) \cdot \left(1 + \frac{1}{a-3} \right) \quad (24)$$

$$< 1 + \frac{1}{a-3}. \quad (25)$$

In order to lower-bound $\mathbb{E}[\tilde{X}]$, we upperbound $\tilde{N}(a)$. Recall that

$$\tilde{N}(a) = \sum_{j=0}^{\lfloor (k-2)/2 \rfloor} s_{k-2-2j}(a).$$

By the second item of Lemma B.1 we have that

$$\tilde{N}(a) < \frac{s_{k-1}(a)}{a-1} + s_{k-1}(a) \cdot \sum_{j=1}^{\lfloor (k-2)/2 \rfloor} \frac{1}{a^j (a-1)^{2j+1}} \quad (26)$$

$$< s_{k-1}(a) \cdot \left(\frac{1}{a-1} + \frac{1}{(a-1) \cdot (a(a-1)^2 - 1)} \right) < s_{k-1}(a)/(a-2). \quad (27)$$

Therefore,

$$\mathbb{E}[\tilde{X}] > \frac{s_{k-2}(a) \cdot a}{\tilde{N}(a)} > \frac{s_{k-2}(a) \cdot a}{s_{k-1}(a)/(a-2)} > a-2. \quad (28)$$

■

C Defining Instances of DE Based on Integer Random Variables

In order to prove Theorem 2.1 we will construct a pair of DE instances based on \hat{X} and \tilde{X} (as defined in Theorem 4.5). For any given length n , we would like to find instances whose frequency variables are exactly \hat{X} and \tilde{X} . This may not always be possible since the probabilities may not be multiples of $\frac{1}{n}$. Instead, we choose instances that approximate the desired frequency variables; for large n this is sufficient. For every variable X and length n , we define a corresponding instance D_X .

Definition C.1 (The Instance D_X) Let $a_0 < a_1 < \dots < a_{k-1}$ be $k > 1$ integers, and let X be a random variable defined over these integers where $\Pr[X = a_i] = p_i$, so that in particular $\mathbb{E}[X] = \sum_{i=0}^{k-1} p_i \cdot a_i$. Based on X , we define a DE instance of length n (that is a string in $[n]^n$) that contains M_X colors, where $M_X = \sum_{i=0}^{k-1} \left\lfloor \frac{np_i}{\mathbb{E}[X]} \right\rfloor + n - \sum_{i=0}^{k-1} \left\lfloor \frac{np_i}{\mathbb{E}[X]} \right\rfloor \cdot a_i$. (Note that if $\frac{np_i}{\mathbb{E}[X]}$ is an integer for every i then $M_X = \frac{n}{\mathbb{E}[X]}$.) For $i = 0, \dots, k-1$, D_X contains $\left\lfloor \frac{np_i}{\mathbb{E}[X]} \right\rfloor$ colors of type i , where each color of type i appears a_i times. In addition, there are $n - \sum_{i=0}^{k-1} \left\lfloor \frac{np_i}{\mathbb{E}[X]} \right\rfloor \cdot a_i$ colors that appear once each.

The names and order of the colors in D_X are unimportant. For concreteness, assign labels from 1 to M_X in increasing order of the number of times each color appears, and arrange the symbols in order of their color names in the string.

Note that because it is sufficient to prove lower bounds for algorithms that sample uniformly at random, and look only at the *histogram* of the colors appearing in their sample, we do not care about the labels or order of colors in the instances we construct – the algorithms we care about are oblivious to them.

D Indistinguishability by Poisson Algorithms

Batu *et al.* [5, conference version] proved a variant of the following lemma in the context of entropy estimation of distributions. However, the statements and the proofs also apply to estimating symmetric functions

over strings and, in particular, to DE. Recall that we call a uniform algorithm Poisson- s if the number of samples it takes is a random variable, distributed as $\text{Po}(s)$.

Lemma D.1 (following conference version of [5])

- (a) *Poisson algorithms can simulate uniform algorithms. Specifically, for every uniform algorithm \mathcal{A} that uses at most $\frac{s}{2}$ samples, there is a Poisson- s algorithm \mathcal{A}' such that for every input w , the statistical difference between the distributions $\mathcal{A}(w)$ and $\mathcal{A}'(w)$ is $o(1/s)$.*
- (b) *If the input to DE contains b balls of a particular color, then the number of balls of that color seen by a Poisson- s algorithm is distributed as $\text{Po}(\frac{b \cdot s}{n})$. Moreover, it is independent of the number of balls of all other colors in the sample.*
- (c) *For any function invariant under permutations of the alphabet symbols (color names), any Poisson algorithm can be simulated by an algorithm that gets only the histogram of the sample as its input. The simulation has the same approximation guarantees as the original algorithm.*

The independence of the number of occurrences of different colors in the sample (Part (b) above) will be very useful in analyzing the distributions seen by the algorithm.

We now prove the three intermediate lemmas used in the proof of Theorem 5.4. In Appendix E we state some properties of the Poisson distribution, which we use in these proofs. Recall that we consider inputs with $C_i = \left\lfloor \frac{p_i n}{\mathbb{E}[\mathbf{X}]} \right\rfloor$ colors of type i , for $i = 0, \dots, k-1$, where each color of type i appears a_i times, and with $C_k = n - \sum_{i=0}^{k-1} \left\lfloor \frac{p_i n}{\mathbb{E}[\mathbf{X}]} \right\rfloor \cdot a_i$ additional colors that appear once each. We say that the C_k “left-over” colors are of type k , and define $a_k = 1$. Note that $C_k < \sum_{i=0}^{k-1} a_i$ (since $\mathbb{E}[\mathbf{X}] = \sum_{i=0}^{k-1} p_i \cdot a_i$ and so $n = \sum_{i=0}^{k-1} \frac{p_i n}{\mathbb{E}[\mathbf{X}]} \cdot a_i$).

The independence of the number of different colors (see Lemma D.1(b)) makes it easy to understand the distribution on the number of ℓ -way collisions. For a color that appears a_i times in the input, the number of balls of that color in the sample is distributed according to $\text{Po}(\lambda_i)$, where $\lambda_i = \frac{a_i s}{n}$. The probability that this color appears exactly ℓ times in the sample is $\frac{\lambda_i^\ell}{\ell!} e^{-\lambda_i}$.

Proof of Lemma 5.5. Consider any particular color of type i . The probability that the algorithm sees k or more balls of that color is $\Pr[\text{Po}(\lambda_i) \geq k] \leq \frac{\lambda_i^k}{k!}$. Summing over all colors (according to types), we can bound the probability that some color appears k or more times by:

$$\sum_{i=0}^k C_i \cdot \frac{\lambda_i^k}{k!} = \sum_{i=0}^k C_i \cdot \frac{1}{k!} \left(\frac{a_i s}{n} \right)^k \tag{29}$$

$$= \frac{s^k}{k! \cdot n^{k-1}} \cdot \left(\sum_{i=0}^k \frac{C_i \cdot a_i^k}{n} + \frac{C_k \cdot a_k^k}{n} \right) \tag{30}$$

$$< \frac{s^k}{k! \cdot n^{k-1}} \cdot \left(\sum_{i=0}^{k-1} \frac{p_i a_i^k}{\mathbb{E}[\mathbf{X}]} + 1 \right) \tag{31}$$

$$= O \left(\frac{s^k \cdot a_{k-1}^{k-1}}{k! \cdot n^{k-1}} \right) \tag{32}$$

where we have used the fact that $a_k = 1$, $C_k < n$, and $\sum_{i=0}^{k-1} p_i a_i^k < a_{k-1}^{k-1} \cdot \mathbb{E}[X]$. ■

Proof of Lemma 5.7. Observe that F_ℓ , the number of ℓ -way collisions, is a sum of independent Bernoulli random variables, one for each color, with probability $\frac{1}{\ell!} \cdot e^{-\frac{as}{n}} \cdot \left(\frac{as}{n}\right)^\ell$ of being 1 if the color appeared a times in the input. Hence, the number of ℓ -way collisions is a sum of independent binomial random variables, one for each type. That is,

$$F_\ell \sim \sum_{i=0}^k \text{Bin} \left(C_i, \frac{e^{-\lambda_i} \lambda_i^\ell}{\ell!} \right), \quad (33)$$

where $\text{Bin}(m, p)$ is the number of heads in a sequence of m independent coin flips, each of which has probability p of heads.

When p is small, the Poisson distribution $\text{Po}(\lambda = pm)$ is a good approximation to $\text{Bin}(m, p)$; the statistical difference between the two is at most p (see Lemma E.1, Item (3)). Since the sum of independent Poisson variables is also a Poisson variable (see Lemma E.1, Item (2)),

$$F_\ell \approx_{\gamma_\ell} \text{Po} \left(\lambda^{(\ell)} = \sum_{i=0}^k C_i \cdot \frac{\lambda_i^\ell}{\ell!} e^{-\lambda_i} \right) \quad (34)$$

where

$$\gamma_\ell \leq \sum_{i=0}^k \frac{e^{-\lambda_i} \lambda_i^\ell}{\ell!} \leq \sum_{i=0}^k \lambda_i \leq \frac{k \cdot a_{k-1} \cdot s}{n}. \quad (35)$$

In the last equality we have used the fact that $a_k = 1$ and $a_i < a_{k-1}$ for every $i < k - 1$.

To bound the statistical difference between \hat{F}_ℓ and \tilde{F}_ℓ from above, it is enough to bound the difference between $\hat{\lambda}^{(\ell)}$ and $\tilde{\lambda}^{(\ell)}$, since the statistical difference between $\text{Po}(\hat{\lambda}^{(\ell)})$ and $\text{Po}(\tilde{\lambda}^{(\ell)})$ is at most $|\hat{\lambda}^{(\ell)} - \tilde{\lambda}^{(\ell)}|$ (see Lemma E.1, Item (5)).

Substituting $\frac{a_i}{n} \cdot s$ for λ_i and using the fact that $e^{-\lambda_i} = \sum_{j=0}^{k-\ell-1} (-1)^j \cdot \frac{\lambda_i^j}{j!} + (-1)^{k-\ell} \cdot O\left(\frac{\lambda_i^{k-\ell}}{(k-\ell)!}\right)$, (where we define $0! = 1$) we get that

$$\lambda^{(\ell)} = \frac{1}{\ell!} \cdot \sum_{j=0}^{k-\ell} T_j^{(\ell)} \quad (36)$$

where

$$T_j^{(\ell)} = (-1)^j \cdot \frac{1}{j!} \cdot \frac{s^{\ell+j}}{n^{\ell+j}} \cdot \sum_{i=0}^k C_i \cdot a_i^{\ell+j} \quad (37)$$

for $0 \leq j \leq k - \ell - 1$, and

$$T_{k-\ell}^{(\ell)} = (-1)^{k-\ell} \cdot O\left(\frac{1}{(k-\ell)!} \cdot \frac{s^k}{n^k} \cdot \sum_{i=0}^k C_i \cdot a_i^k\right) \quad (38)$$

For each j , $0 \leq j \leq k - \ell$, we have that

$$\frac{s^{\ell+j}}{n^{\ell+j}} \cdot \sum_{i=0}^k C_i \cdot a_i^{\ell+j} = \frac{s^{\ell+j}}{n^{\ell+j}} \cdot \left(\sum_{i=0}^{k-1} \left\lfloor \frac{p_i n}{\mathbb{E}[X]} \right\rfloor \cdot a_i^{\ell+j} + C_k \right) \quad (39)$$

$$\leq \frac{s^{\ell+j}}{n^{\ell+j-1}} \cdot \frac{1}{\mathbb{E}[X]} \cdot \sum_{i=0}^{k-1} p_i \cdot a_i^{\ell+j} + \frac{s^{\ell+j}}{n^{\ell+j}} \cdot \sum_{i=0}^{k-1} a_i \quad (40)$$

$$= \frac{s^{\ell+j}}{n^{\ell+j-1}} \cdot \frac{\mathbb{E}[X^{\ell+j}]}{\mathbb{E}[X]} + \frac{s^{\ell+j}}{n^{\ell+j}} \cdot \sum_{i=0}^{k-1} a_i \quad (41)$$

and similarly

$$\frac{s^{\ell+j}}{n^{\ell+j}} \cdot \sum_{i=0}^k C_i \cdot a_i^{\ell+j} = \frac{s^{\ell+j}}{n^{\ell+j}} \cdot \left(\sum_{i=0}^{k-1} \left\lfloor \frac{p_i n}{\mathbb{E}[X]} \right\rfloor \cdot a_i^{\ell+j} + C_k \right) \quad (42)$$

$$\geq \frac{s^{\ell+j}}{n^{\ell+j-1}} \cdot \frac{1}{\mathbb{E}[X]} \cdot \sum_{i=0}^{k-1} p_i \cdot a_i^{\ell+j} - \frac{s^{\ell+j}}{n^{\ell+j}} \cdot \sum_{i=0}^{k-1} a_i^{\ell+j} \quad (43)$$

$$= \frac{s^{\ell+j}}{n^{\ell+j-1}} \cdot \frac{\mathbb{E}[X^{\ell+j}]}{\mathbb{E}[X]} - \frac{s^{\ell+j}}{n^{\ell+j}} \cdot \sum_{i=0}^{k-1} a_i^{\ell+j} \quad (44)$$

The moment condition on \hat{X} and \tilde{X} states that $\frac{\mathbb{E}[\hat{X}^{\ell+j}]}{\mathbb{E}[\hat{X}]} = \frac{\mathbb{E}[\tilde{X}^{\ell+j}]}{\mathbb{E}[\tilde{X}]}$ for $j = 0, \dots, k - \ell - 1$. Thus,

$$\left| \hat{\lambda}^{(\ell)} - \tilde{\lambda}^{(\ell)} \right| = O \left(\frac{1}{\ell!} \cdot \sum_{j=0}^{k-\ell} \frac{s^{\ell+j}}{n^{\ell+j}} \cdot 2 \sum_{i=0}^{k-1} a_i^{\ell+j} + \frac{1}{\ell!(k-\ell)!} \cdot \frac{s^k}{n^{k-1}} \cdot \max \left\{ \frac{\mathbb{E}[\hat{X}^k]}{\mathbb{E}[\hat{X}]}, \frac{\mathbb{E}[\tilde{X}^k]}{\mathbb{E}[\tilde{X}]} \right\} \right) \quad (45)$$

The ratio $\frac{\mathbb{E}[X^k]}{\mathbb{E}[X]}$ is at most $(a_{k-1})^{k-1}$, the expression $\frac{1}{\ell!(k-\ell)!}$ is maximized for $\ell = \lfloor \frac{k}{2} \rfloor$, and

$$\frac{1}{\ell!} \cdot \sum_{j=0}^{k-\ell} \frac{s^{\ell+j}}{n^{\ell+j}} \cdot 2 \sum_{i=0}^{k-1} a_i^{\ell+j} \leq \frac{1}{\ell!} \cdot \left(\frac{s \cdot a_{k-1}}{n} \right)^\ell \cdot 2k \cdot \sum_{j=0}^{k-\ell} \left(\frac{s \cdot a_{k-1}}{n} \right)^j = O \left(\frac{k \cdot a_{k-1} \cdot s}{n} \right) \quad (46)$$

where the last equality uses the fact that $\frac{s \cdot a_{k-1}}{n} \leq \frac{1}{2}$. Therefore,

$$\left| \hat{\lambda}^{(\ell)} - \tilde{\lambda}^{(\ell)} \right| = O \left(\frac{1}{\lfloor \frac{k}{2} \rfloor! \cdot \lceil \frac{k}{2} \rceil!} \cdot \left(\frac{a_{k-1}}{n} \right)^{k-1} \cdot s^k + \frac{k \cdot a_{k-1} \cdot s}{n} \right) \quad (47)$$

Summing this together with the error, denoted γ_ℓ , introduced by approximating a sum of binomials with a Poisson variable, proves the lemma. \blacksquare

In order to prove Lemma 5.6 we need the following lemma concerning multinomial variables, whose proof is provided in Appendix E.

Lemma D.2 Consider a k -sided die, whose sides are numbered $0, \dots, k - 1$, where side ℓ has probability q_ℓ and $q_0 \geq 1/2$. Let Z_0, \dots, Z_{k-1} be random variables that count the number of occurrences of each side in a sequence of independent rolls. Let Z'_1, \dots, Z'_{k-1} be independent random variables, where for each ℓ , the variable Z'_ℓ is distributed identically to Z_ℓ . Then $(Z_1, \dots, Z_{k-1}) \approx_{\delta_4} (Z'_1, \dots, Z'_{k-1})$ for $\delta_4 = O(k(1 - q_0)^{2/3})$.

Proof of Lemma 5.6. We can write F_ℓ as a sum $F_\ell = F_\ell^{(1)} + \dots + F_\ell^{(k)}$, where $F_\ell^{(i)}$ is the number of ℓ -way collisions among colors of type i . Since the types are independent, it is sufficient to show that for each i , the variables $F_1^{(i)}, \dots, F_{k-1}^{(i)}$ are close to being independent. We can then sum the distances over the types to prove the lemma.

Let $F_0^{(i)}$ denote the number of colors of type i that occur either 0 times, or k or more times, in the sample. The vector $F_0^{(i)}, F_1^{(i)}, \dots, F_{k-1}^{(i)}$ follows a multinomial distribution. It counts the outcomes of an experiment in which C_i independent, identical dice are rolled, and each one produces outcome ℓ with probability $e^{-\lambda_i} \lambda_i^\ell / \ell!$, for $\ell \in [k-1]$, and outcome 0 with the remaining probability. On each roll, outcome 0 occurs with probability at least $e^{-\lambda_i} \geq 1 - \lambda_i \geq 1/2$ (recall that $\lambda_i = \frac{a_i \cdot s}{n} \leq 1/2$).

Lemma D.2 shows that when one outcome occupies almost all the mass in such an experiment, the counts of the remaining outcomes are close to independent — within distance $O(k \cdot \lambda_i^{2/3})$. Summing over all types, the distance of F_1, \dots, F_{k-1} from independent is $O\left(k \cdot \sum_i \lambda_i^{2/3}\right) = O\left(k^2 \cdot \left(\frac{a_{k-1}s}{n}\right)^{2/3}\right)$. ■

E Properties of the Poisson Distribution and Proof of Lemma D.2

We start with some useful properties of the Poisson distribution:

- Lemma E.1**
1. If $X \sim \text{Po}(\lambda)$, then $\mathbb{E}[X] = \text{Var}[X] = \lambda$.
 2. If $X \sim \text{Po}(\lambda)$, $Y \sim \text{Po}(\lambda')$ and X, Y are independent, then $X + Y \sim \text{Po}(\lambda + \lambda')$.
 3. The statistical difference between $\text{Bin}(m, p)$ and $\text{Po}(mp)$ is at most p .
 4. For $\lambda > 0$, the statistical difference between $\text{Po}(\lambda)$ and $\text{Po}(\lambda + \epsilon\sqrt{\lambda})$ is $O(\epsilon)$.
 5. The statistical difference between $\text{Po}(\lambda)$ and $\text{Po}(\lambda')$ is at most $|\lambda - \lambda'|$.

Note: Item (5) provides a good bound when λ is near or equal to 0. In most settings, Item (4) is more useful.

Proof: Items (1) and (2) can be found in any standard probability text. For Item (3) (and other bounds on the Poisson approximation to the binomial), see [12] or [15, Bound b_1]. To prove item (4), first compute the *relative entropy* (also called *Kullback-Liebler divergence*) between $\text{Po}(\lambda')$ and $\text{Po}(\lambda)$. For probability distributions p, q , the relative entropy is $D(p||q) = \sum_x p(x) \ln \frac{p(x)}{q(x)}$. The statistical difference between p and q is at most $\sqrt{2 \ln(2) D(p||q)}$ (see, e.g., [10, Lemma 12.6.1]). If $X \sim \text{Po}(\lambda + \Delta)$, then the relative entropy in our case is

$$D(\text{Po}(\lambda + \epsilon) || \text{Po}(\lambda)) = \mathbb{E}_X \left[\ln \left(\frac{e^{-\lambda - \Delta} (\lambda + \Delta)^X / X!}{e^{-\lambda} \lambda^X / X!} \right) \right] = -\Delta + (\lambda + \Delta) \ln \left(\frac{\lambda + \Delta}{\lambda} \right).$$

Since $\ln(1+x) \leq x$, the relative entropy is at most Δ^2/λ , and the statistical difference is at most $\Delta \sqrt{\frac{2 \ln(2)}{\lambda}}$. Setting $\Delta = \epsilon\sqrt{\lambda}$, we obtain the desired bound.

Finally, to prove Item (5) write $\text{Po}(\lambda + \Delta)$ as a sum of two independent Poisson variables X_λ, X_Δ with parameters λ and Δ respectively. Conditioned on the event $X_\Delta = 0$, the sum is distributed as $\text{Po}(\lambda)$. This event occurs with probability $e^{-\Delta} \geq 1 - \Delta$. The statistical difference between $\text{Po}(\lambda)$ and $\text{Po}(\lambda + \Delta)$ is thus at most Δ , as desired. ■

We next repeat and prove Lemma D.2.

Lemma D.2 Consider a k -sided die, whose sides are numbered $0, \dots, k-1$, where side ℓ has probability q_ℓ and $q_0 \geq 1/2$. Let Z_0, \dots, Z_{k-1} be random variables that count the number of occurrences of each side in a sequence of independent rolls. Let Z'_1, \dots, Z'_{k-1} be independent random variables, where for each ℓ , the variable Z'_ℓ is distributed identically to Z_ℓ . Then $(Z_1, \dots, Z_{k-1}) \approx_{\delta_4} (Z'_1, \dots, Z'_{k-1})$ for $\delta_4 = O(k(1-q_0)^{2/3})$.

Proof: Suppose the die is rolled m times. For each $\ell > 1$, the number of occurrences of side ℓ is a binomial: $Z_\ell \sim \text{Bin}(m, q_\ell)$. By Item (3) in Lemma E.1, the difference between $\text{Bin}(m, q_\ell)$ and $\text{Po}(\lambda = mq_\ell)$ is at most q_ℓ .

Consider Z_ℓ , conditioned on the values of $Z_1, \dots, Z_{\ell-1}$. The distribution of Z_ℓ is still binomial but has different parameters: $Z_\ell \sim \text{Bin}\left(m - S_\ell, \frac{q_\ell}{1-Q_\ell}\right)$, where $S_\ell = \sum_{i=1}^{\ell-1} Z_i$ and $Q_\ell = \sum_{i=1}^{\ell-1} q_i$. We can approximate this by a Poisson variable with parameter $\lambda' = (m - S_\ell) \frac{q_\ell}{1-Q_\ell}$. This approximation introduces an error (statistical difference) of at most q_ℓ .

Now the sum S_ℓ is also binomial, with parameters m, Q_ℓ . It has expectation mQ_ℓ and variance $mQ_\ell(1-Q_\ell)$. By Chebyshev's inequality, $S_\ell \in mQ_\ell \pm \sqrt{\frac{mQ_\ell(1-Q_\ell)}{\gamma}}$ with probability at least $1-\gamma$. When this occurs,

$$\lambda' = (m - S_\ell) \frac{q_\ell}{1-Q_\ell} = \frac{mq_\ell}{1-Q_\ell} \left(1 - Q_\ell \pm \sqrt{\frac{Q_\ell(1-Q_\ell)}{m\gamma}}\right) = mq_\ell \pm \sqrt{mq_\ell} \cdot \sqrt{\frac{q_\ell Q_\ell}{\gamma(1-Q_\ell)}}.$$

Thus, $\lambda' = \lambda + \sqrt{\frac{q_\ell Q_\ell}{\gamma(1-Q_\ell)}} \sqrt{\lambda}$ with probability at least $1-\gamma$. The statistical difference between $\text{Po}(\lambda)$ and $\text{Po}(\lambda')$ is $O\left(\sqrt{\frac{q_\ell Q_\ell}{\gamma(1-Q_\ell)}}\right)$, by Lemma E.1, Item (4).

Putting these approximations together: we can replace Z_ℓ by an independent copy of itself, Z'_ℓ , and change the distribution on the vector (Z_1, \dots, Z_ℓ) by at most $\gamma + 2q_\ell + O\left(\sqrt{\frac{q_\ell Q_\ell}{\gamma(1-Q_\ell)}}\right)$. This is minimized when $\gamma = \sqrt[3]{\frac{q_\ell Q_\ell}{1-Q_\ell}}$. In symbols:

$$(Z_1, \dots, Z_{\ell-1}, Z_\ell) \approx_{O\left(\sqrt[3]{\frac{q_\ell Q_\ell}{1-Q_\ell}}\right)} (Z_1, \dots, Z_{\ell-1}, Z'_\ell).$$

Now replace the Z_ℓ s with Z'_ℓ s one at a time. By the triangle inequality, the distance between (Z_1, \dots, Z_{k-1}) and (Z'_1, \dots, Z'_{k-1}) is at most the sum of the errors introduced at each step. This sum is $O(k \sqrt[3]{\frac{q_\ell Q_\ell}{1-Q_\ell}})$. Since $q_0 \geq \frac{1}{2}$, the total distance is $O(k \cdot (1-q_0)^{2/3})$. ■