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

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

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: $\mathsf{E}[\mathsf{X}_1]/\,\mathsf{E}[\mathsf{X}_2] = \mathsf{E}[\mathsf{X}_1^2]/\,\mathsf{E}[\mathsf{X}_2^2] = \ldots = \mathsf{E}[\mathsf{X}_1^k]/\,\mathsf{E}[\mathsf{X}_2^k]$. Our lower bound method is also applicable to other problems. In particular, it gives new lower bounds for the sample complexity of (1) approximating the entropy of a distribution and (2) approximating how well a given string is compressed by the Lempel-Ziv scheme.

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 extremely 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 approximation 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 hence, time) complex-

*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. Supported by the Israel Science Foundation (grant number 439/06).

ity 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

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.*

We show 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 $\frac{n}{\mathsf{d}}$ colors, for any $\mathsf{d} = n^{o(1)}$. 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 $\Omega\left(n^{1-1/k}\right)$ 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 $\frac{n}{\mathsf{d}_1}$ colors from an input with $\frac{n}{\mathsf{d}_2}$ colors, where d_1 and d_2 are different constants. Surprisingly, 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 $\frac{n}{\mathsf{d}_1}$ and $\frac{n}{\mathsf{d}_2}$ colors, respectively, and $\mathsf{d}_2 \gg \mathsf{d}_1$. 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 $\frac{\mathsf{E}[\mathsf{X}_1]}{\mathsf{E}[\mathsf{X}_2]} = \frac{\mathsf{E}[\mathsf{X}_1^2]}{\mathsf{E}[\mathsf{X}_2^2]} = \cdots = \frac{\mathsf{E}[\mathsf{X}_1^{k-1}]}{\mathsf{E}[\mathsf{X}_2^{k-1}]}$, 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 that have 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 suffices 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 Section 1.3.)

To prove a lower bound, it suffices to consider algorithms that have access only to the *histogram* $(\mathsf{F}_1, \mathsf{F}_2, \mathsf{F}_3, \dots)$ of the selected sample. Namely, 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 $(\mathsf{F}_1, \mathsf{F}_2, \mathsf{F}_3, \dots)$. The main idea is to show that (a) the variables $\mathsf{F}_1, ..., \mathsf{F}_{k-1}$ 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 $0 \le \ell < k$) 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* [15], 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, defined in [17]. In conjunction with 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 shown 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 all $B \le n^{1/4}/\sqrt{\log n}$, 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 $\frac{n}{11}$ colors[1] from inputs with at most $\frac{n}{B}$ colors.*

The next corollary provides an important special case:

Corollary 2.2 *For all $B = n^{o(1)}$, 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). Section 3 shows that to obtain a lower bound on DE it suffices to consider algorithms that take uniform samples with replacement. 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. Section 6 culminates in the proof of Theorem 2.1. Finally, in Section 7, we apply our techniques to the sample complexity of approximating the entropy.

[1] We did not try to optimize the constants (in particular, 11).

3 Algorithms 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 3.1 ([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.

Definition 3.2 (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 3.3 *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 *with multiplicative error $(\sqrt{0.1} \cdot \alpha)$, there is a* uniform *algorithm $\mathcal{A}'$ that takes s samples and provides, with probability at least $\frac{2}{3}$, an approximation for* DE *with multiplicative error α.*

Proof: 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 has at most C colors. However, some of the colors might be missing. One can show [14] 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 has between $0.1 \cdot C$ and C colors. When $\mathcal{A}$ is run on that instance, with probability $\geq \frac{11}{12}$, it outputs 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. ■

Rephrasing Lemma 3.3, using a few details from the reduction in the proof, we obtain Lemma 3.4.

Lemma 3.4 *If every* uniform *algorithm needs at least s queries 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 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. For $\ell = 0, 1, \cdots, s$, let F_ℓ be the number of ℓ-way collisions in the sample. The **histogram** F of the sample is the vector $(F_1, \cdots, F_s)$, 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 $o\left(n^{1-1/k}\right)$ 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, closely related to ℓ-way collisions: A *monochromatic ℓ-tuple* is a set of ℓ samples that have the same color. Notice that the number of k-way collisions can be obtained from the number of monochromatic ℓ-tuples for all $k \geq \ell$ by the Inclusion-Exclusion Principle. Therefore, if for two instances, the expected number of monochromatic ℓ-tuples is the same for all ℓ, then so is the expected number of ℓ-way collisions. In this section, we show how to construct pairs of instances with the same *expectations* on the number of monochromatic ℓ-tuples, for every $\ell < k$. (Section 5 proves that equal expectations imply that the distributions themselves are close.) To express requirements on the instances 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, $\Pr[\mathsf{X} = a_i] = p_i$. Since, on average, each color appears d times, $\mathsf{E}[\mathsf{X}] = \sum_i p_i a_i = d$. Conversely, 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, 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

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

The last equality holds because $\mathsf{E}[\mathsf{X}] = d$ and $\Pr[\mathsf{X} = a_i] = p_i$. We 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 $\frac{\mathsf{E}[\mathsf{X}^\ell]}{\mathsf{E}[\mathsf{X}]}$ 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) *Random variables $\hat{\mathsf{X}}$ and $\tilde{\mathsf{X}}$ have $k-1$ **proportional moments** if*

$$\frac{\mathsf{E}[\tilde{\mathsf{X}}]}{\mathsf{E}[\hat{\mathsf{X}}]} = \frac{\mathsf{E}[\tilde{\mathsf{X}}^2]}{\mathsf{E}[\hat{\mathsf{X}}^2]} = \frac{\mathsf{E}[\tilde{\mathsf{X}}^3]}{\mathsf{E}[\hat{\mathsf{X}}^3]} = \cdots = \frac{\mathsf{E}[\tilde{\mathsf{X}}^{k-1}]}{\mathsf{E}[\hat{\mathsf{X}}^{k-1}]} .$$

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. Additionally, we need that the instances have very different numbers of distinct colors. This corresponds to ensuring that the frequency variables have different expectations.

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

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

To reduce notation, in the rest of the paper, all variables pertaining to the first instance in the pair of instances that are hard to distinguish, are marked by a hat (ˆ) and those pertaining to the second, by a tilde (˜). In statements relevant to both instances, the corresponding variables without hat or tilde are used.

Proof of Theorem 4.5. The rest of the section is devoted to this proof, which is comprised of three parts: an overview, the construction, and its analysis.

Overview. Let $C = \mathsf{E}[\tilde{\mathsf{X}}]/\mathsf{E}[\hat{\mathsf{X}}]$. Then the moments condition (Definition 4.4) can be restated as $C \geq B$ and $(\mathsf{E}[\tilde{\mathsf{X}}], \mathsf{E}[\tilde{\mathsf{X}}^2], \ldots, \mathsf{E}[\tilde{\mathsf{X}}^{k-1}]) = C \cdot (\mathsf{E}[\hat{\mathsf{X}}], \mathsf{E}[\hat{\mathsf{X}}^2], \ldots, \mathsf{E}[\hat{\mathsf{X}}^{k-1}])$. Recall that the supports of $\hat{\mathsf{X}}$ and $\tilde{\mathsf{X}}$ are both contained in $\{a_0, \ldots, a_{k-1}\}$. The main step in our construction is to set $a_j = a^j$ for an appropriate $a > 1$. Let $p_i = \Pr[\mathsf{X} = a_i]$, and $\vec{p} = (p_0, \ldots, p_{k-1})$. Let V denote the $(k-1) \times k$ Vandermonde matrix satisfying $V_{i,j} = (a_j)^i$. Then the vector $(\mathsf{E}[\mathsf{X}], \mathsf{E}[\mathsf{X}^2], \ldots, \mathsf{E}[\mathsf{X}^{k-1}])$ can be represented as the product $V \cdot \vec{p}$. This gives yet another way to formulate the moments condition: $V(C \cdot \vec{\hat{p}} - \vec{\tilde{p}}) = \vec{0}$. 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 $C \cdot \hat{\mathsf{p}}_i$ and negative u_i become $-\tilde{\mathsf{p}}_i$, divided by the corresponding normalization factors. This defines distributions $\hat{\mathsf{X}}$ and $\tilde{\mathsf{X}}$, for each a.

To find an appropriate choice of a and to demonstrate the required properties of our construction, we explicitly compute vector $\vec{u}$ that defines the distributions. The main idea behind this step is to view $\vec{u}$ as coefficients of a polynomial. Let $f(t) = t^{k-1} + u_{k-2}t^{k-2} + \ldots + u_0$ be the unique non-zero polynomial that vanishes on $a, a^2, \ldots, a^{k-1}$. Then $f(t) = \prod_{i=1}^{k-1}(t - a^i)$. Because the set of zeros of f is a geometric sequence, we can show that the coefficients of f also grow rapidly. This enables us to demonstrate that $C > a - 3$, which implies that it is enough to set $a = B + 3$.

Construction. We start by computing the coefficients of the polynomial $f(t)$, described in the overview. For every $0 \leq i \leq k-1$, let $s_i(y_1, \ldots, y_{k-1})$ be the ith symmetric function

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

E.g., $s_2(y_1, \ldots, y_{k-1}) = y_1y_2 + y_1y_3 + y_2y_3$ if $k = 4$ and $i = 2$. In general, $s_0 = 1$ and $s_{k-1}(y_1, ..., y_{k-1}) = y_1 \cdot \ldots \cdot y_{k-1}$. As explained in the overview, the supports of the two distributions we construct are contained in the set $\{1, a, a^2, \ldots, a^{k-1}\}$, where a is a positive integer parameter. Define $s_i(a) \stackrel{\text{def}}{=} s_i(a, a^2, \ldots, a^{k-1})$. Following our previous example, $s_2(a) = a^3 + a^4 + a^5$ and $s_3(a) = a^6$. For our analysis we need the following bounds on $s_i(a)$'s in terms of $s_{k-1}(s)$, proved in [14]:

Claim 4.6 *For all $a > 3$,*

1. $s_{k-2}(a) > s_{k-1}(a)/a$.
2. $s_{k-i}(a) < \frac{s_{k-1}(a)}{a^{\left\lfloor \frac{i-1}{2} \right\rfloor}(a-1)^{i-1}}$ *for all* $2 \leq i \leq k-1$.

Consider the polynomial $f(t) = \prod_{i=1}^{k-1}(t - a^i)$. It is easy to see that $f(t) = (-1)^{k-1} \cdot \sum_{i=0}^{k-1}(-1)^i \cdot s_{k-1-i}(a) \cdot t^i$. The probability of each element in our distributions is determined by the corresponding coefficient of f. We define $\forall i, \ 0 \leq i \leq k-1$:

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

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

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

Analysis. After proving the two auxiliary lemmas, we use them to complete the proof of Theorem 4.5. Lemma 4.7 shows that the distributions $\hat{\mathsf{X}}$ and $\tilde{\mathsf{X}}$ have $k-1$ proportional moments (see Defition 4.3). Lemma 4.8 bounds $\mathsf{E}[\tilde{\mathsf{X}}]$ and $\mathsf{E}[\hat{\mathsf{X}}]$.

Lemma 4.7 *Let $C \stackrel{\text{def}}{=} \hat{\mathsf{N}}(a)/\tilde{\mathsf{N}}(a)$. Then $C \cdot \mathsf{E}[\hat{\mathsf{X}}^\ell] = \mathsf{E}[\tilde{\mathsf{X}}^\ell]$ for $\ell = 1, \ldots, k-1$.*

The proof easily follows from the definition of $\hat{\mathsf{X}}$ and $\tilde{\mathsf{X}}$.

Lemma 4.8 *For all $a > 3$,*
(1) $\mathsf{E}[\hat{\mathsf{X}}] < 1 + \frac{1}{a-3}$; *(2)* $\mathsf{E}[\tilde{\mathsf{X}}] > a - 2$.

Proof: By definition of $\hat{\mathsf{X}}$,

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

By Claim 4.6, Item (2),

$$\begin{aligned} \mathsf{E}[\hat{\mathsf{X}}] &< \frac{s_{k-1}(a)}{\hat{\mathsf{N}}(a)} \cdot \left(1 + \sum_{j=1}^{\lfloor (k-1)/2 \rfloor} \frac{a^{2j}}{a^j(a-1)^{2j}}\right) \\ &< \frac{s_{k-1}(a)}{\hat{\mathsf{N}}(a)} \cdot \left(1 + \sum_{j=1}^{\lfloor (k-1)/2 \rfloor} \frac{1}{(a-2)^j}\right) \\ &< \frac{s_{k-1}(a)}{\hat{\mathsf{N}}(a)} \cdot \left(1 + \frac{1}{a-3}\right) \\ &< 1 + \frac{1}{a-3}. \end{aligned}$$

To bound $\mathsf{E}[\tilde{\mathsf{X}}]$ from below, we first bound $\tilde{\mathsf{N}}(a)$ from above. Recall that $\tilde{\mathsf{N}}(a) = \sum_{j=0}^{\lfloor (k-2)/2 \rfloor} s_{k-2-2j}(a)$. By Claim 4.6, Item (2),

$$\begin{aligned}\tilde{\mathsf{N}}(a) &< s_{k-1}(a) \cdot \sum_{j=0}^{\lfloor (k-2)/2 \rfloor} \frac{1}{a^j (a-1)^{2j+1}} \\ &< s_{k-1}(a) \cdot \left(\frac{1}{a-1} \cdot \left(1 + \frac{1}{a(a-1)^2 - 1} \right) \right) \\ &< s_{k-1}(a)/(a-2).\end{aligned}$$

Since $\tilde{\mathsf{X}}$ takes the value a with probability $s_{k-2}(a)/\tilde{\mathsf{N}}(a)$,

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

The last inequality follows from Claim 4.6, Item (1). The proof of Lemma 4.8 is completed. ■

It remains to find, for every $B > 1$, an a such that $\mathsf{E}[\tilde{\mathsf{X}}]/\mathsf{E}[\hat{\mathsf{X}}] \geq B$. By Lemma 4.8, $\frac{\mathsf{E}[\tilde{\mathsf{X}}]}{\mathsf{E}[\hat{\mathsf{X}}]} > \frac{a-2}{1+1/(a-3)} = a-3$. Thus, if we take $a = B+3$ then $\mathsf{E}[\tilde{\mathsf{X}}]/\mathsf{E}[\hat{\mathsf{X}}] > B$, $\mathsf{E}[\hat{\mathsf{X}}] < 1 + \frac{1}{B}$ and $\mathsf{E}[\tilde{\mathsf{X}}] > B$. This completes the construction and the proof of Theorem 4.5. ■

5 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 $\mathrm{Po}(\lambda)$ takes the value $x \in \mathbb{N}$ with probability $e^{\lambda}\lambda^x/x!$. The expectation and variance of $\mathrm{Po}(\lambda)$ are both λ (for the proof see, e.g., [10]).

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

From this point on we consider Poisson algorithms that get only the histogram of the sample as their input. This is justified by Lemma 5.2, stated next. Batu *et al.* [5] proved a variant of Lemma 5.2 in the context of entropy estimation of distributions. However, the statements and the proofs generalize to estimating symmetric functions over strings and, in particular, to DE.

Recall that two random variables X and Y over a domain S have *statistical difference* δ if $\max_{S' \subset S} |\Pr[X \in S'] - \Pr[Y \in S']| = \delta$.

Lemma 5.2 (generalizes 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 $\mathrm{Po}(\frac{b \cdot s}{n})$.* ***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 (Item (b)) greatly simplifies the analysis of Poisson algorithms.

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 $a_0 < a_1 < \ldots < a_{k-1}$ be integers, and let X be a random variable over these integers with $\Pr[\mathsf{X} = a_i] = p_i$. Then $\mathsf{E}[\mathsf{X}] = \sum_{i=0}^{k-1} p_i \cdot a_i$. Based on X, we define a DE instance D_{X} of length n (that is, a string in $[n]^n$) that contains $\frac{n}{\mathsf{E}[\mathsf{X}]}$ colors. For $i = 0, \ldots, k-1$, D_{X} contains $\frac{np_i}{\mathsf{E}[\mathsf{X}]}$ colors of ***type*** i, where each color of type i appears a_i times. (The full version of this paper [14] contains a general treatment, without the assumption that $\frac{np_i}{\mathsf{E}[\mathsf{X}]}$ is an integer.) By Lemma 5.2, the number of times a given color of type i is seen by a Poisson-s is distributed according to $\mathrm{Po}(\frac{a_i s}{n})$.

Our next main building block in the proof of Theorem 2.1, is the theorem stated below. It shows that if distributions $\hat{\mathsf{X}}$ and $\tilde{\mathsf{X}}$ over integers have $k-1$ proportional moments, then the corresponding instances of DE, $D_{\hat{\mathsf{X}}}$ and $D_{\tilde{\mathsf{X}}}$, 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 $\hat{\mathsf{X}}$ and $\tilde{\mathsf{X}}$ varies as n increases.

Theorem 5.3 (Distinguishability by Poisson Algorithms) *Let $\hat{\mathsf{X}}, \tilde{\mathsf{X}}$ be random variables over positive integers $a_0 < a_1 < \ldots < a_{k-1}$ 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,* $\left|\Pr[\mathcal{A}(D_{\hat{\mathsf{X}}}) = 1] - \Pr[\mathcal{A}(D_{\tilde{\mathsf{X}}}) = 1]\right|$

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

The generality of this bound is required to prove Theorem 2.1. However, the following simpler corollary is sufficient to show that algorithms for DE with additive approximation guarantees require a near-linear number of samples.

Corollary 5.4 *Let $\hat{\mathsf{X}}$ and $\tilde{\mathsf{X}}$ 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 $|\Pr[\mathcal{A}(D_{\hat{\mathsf{X}}}) = 1] - \Pr[\mathcal{A}(D_{\tilde{\mathsf{X}}}) = 1]| = o(1)$.*

We now turn to proving Theorem 5.3. As in Definition 4.1, for $\ell = 0, 1, \ldots, s$, let F_ℓ be a random variable representing the number of ℓ-way collisions a Poisson-s algorithm sees, and let $\mathsf{F} = (\mathsf{F}_1, \mathsf{F}_2, \mathsf{F}_3 \ldots)$ be the corresponding histogram. We can restate Theorem 5.3 in terms of the statistical difference between histogram distributions. Recall that random variables X and Y have statistical difference at most δ if and only if for every algorithm $\mathcal{A}$, $\left|\Pr[\mathcal{A}(X) = 1] - \Pr[\mathcal{A}(Y) = 1]\right| \leq \delta$.

Theorem 5.5 (Distinguishability by Poisson Algorithms, restated) *For $s \leq \frac{n}{2 \cdot a_{k-1}}$, the statistical difference between the histograms $(\hat{\mathsf{F}}_1, \hat{\mathsf{F}}_2, \hat{\mathsf{F}}_3, \ldots)$ and $(\tilde{\mathsf{F}}_1, \tilde{\mathsf{F}}_2, \tilde{\mathsf{F}}_3, \ldots)$ is*

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

For the remainder of this section, assume $s \leq \frac{n}{2 \cdot a_{k-1}}$. The proof of Theorem 5.5 relies on the three following lemmas. Lemma 5.6 states that ℓ-way collisions are very unlikely for $\ell \geq k$, when s is sufficiently small. Lemma 5.7 shows that for both distributions $D_{\hat{\mathsf{X}}}$ and $D_{\tilde{\mathsf{X}}}$, 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 5.10 shows that the number of k-way collisions is distributed almost identically under $D_{\hat{\mathsf{X}}}$ and $D_{\tilde{\mathsf{X}}}$.

Lemma 5.6 *For both distributions $D_{\hat{\mathsf{X}}}$ and $D_{\tilde{\mathsf{X}}}$, the probability of a collision involving $k > 1$ or more balls is at most*

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

Proof: Consider any particular color of type i. The probability that the algorithm sees k or more balls of that color is $\Pr[\mathrm{Po}(a_i s/n) \geq k] \leq \frac{1}{k!}(\frac{a_i s}{n})^k$. Let $C_i = \frac{np_i}{\mathsf{E}[\mathsf{X}]}$ be the number of colors of type i. Taking a union bound over all colors, we can bound the probability that some color appears k or more times. We sum first over types i, and then over colors of a given type:

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

$$< \frac{s^k}{k! \cdot n^{k-1}} \cdot \sum_{i=0}^{k-1} \frac{p_i a_i^k}{\mathsf{E}[\mathsf{X}]} = \frac{s^k}{k! \cdot n^{k-1}} \cdot \frac{\mathsf{E}[\mathsf{X}^k]}{\mathsf{E}[\mathsf{X}]} . \quad (3)$$

Since a_{k-1} is the largest value that X can take, $\mathsf{E}[\mathsf{X}^k] \leq a_{k-1}^{k-1}\,\mathsf{E}[\mathsf{X}]$. Combining this with the bound above, (3), completes the proof. ■

Let $X \approx_\delta Y$ denote that the statistical difference between random variables X and Y is bounded by δ.

Lemma 5.7 *For both distributions $D_{\hat{\mathsf{X}}}$ and $D_{\tilde{\mathsf{X}}}$, $\mathsf{F}_1, \ldots, \mathsf{F}_{k-1}$ are close to independent, that is, $(\mathsf{F}_1, \ldots, \mathsf{F}_{k-1}) \approx_{\delta_2} (\mathsf{F}'_1, \ldots, \mathsf{F}'_{k-1})$, where the variables F'_ℓ are independent, for each ℓ the distributions of F_ℓ and F'_ℓ are identical, and $\delta_2 \leq \frac{k \cdot a_{k-1} \cdot s}{n}$.*

The proof of Lemma 5.7 relies on the following two claims. Claim 5.8 states that the Poisson distribution is a good approximation to the binomial distribution $\mathrm{Bin}(m, p)$ when the parameter p is small. Recall that $\mathrm{Bin}(m, p)$ represents the number of heads in a sequence of m coin flips where the probability of heads is p. Claim 5.9 considers many independent rolls of a biased k-sided die. It shows that 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 independent.

Claim 5.8 ([12, 16]) *The statistical difference between $\mathrm{Bin}(m, p)$ and $\mathrm{Po}(mp)$ is at most p.*

Claim 5.9 *Consider a k-sided die, whose sides are numbered $0, \ldots, k-1$, where side ℓ has probability q_ℓ and $q_0 \geq 1/2$. Let $Z_0, \ldots, Z_{k-1}$ be random variables that count the number of occurrences of each side in a sequence of independent rolls. Let $Z'_1, \ldots, Z'_{k-1}$*

*be **independent** random variables, where for each ℓ, the variable Z'_ℓ is distributed identically to Z_ℓ. Then $(Z_1, ..., Z_{k-1}) \approx_{\delta_4} (Z'_1, \ldots, Z'_{k-1})$ for $\delta_4 = 2(1-q_0)$.*

Proof: Suppose the die is rolled m times. Let $\hat{Z}$ count the number of times that side 0 does not come up, i.e., $\hat{Z} = m - Z_0 = \sum_{\ell=1}^{k-1} Z_\ell$. This count follows a binomial distribution $\hat{Z} \sim \mathrm{Bin}(m, 1-q_0)$. By Claim 5.8, the statistical difference between $\mathrm{Bin}(m, 1-q_0)$ and $\mathrm{Po}\left(m(1-q_0)\right)$ is at most $1-q_0$.

Conditioned on a fixed value of $\hat{Z}$, the variables $Z_1, ..., Z_{k-1}$ follow a multinomial distribution. By Lemma 5.2, if $\hat{Z}$ itself is chosen according to $\mathrm{Po}(m(1-q_0))$, and $Z_1, ..., Z_{k-1}$ are resampled according to this value of $\hat{Z}$, the resulting distribution on $Z_1, ..., Z_{k-1}$ is a vector of independent Poisson random variables distributed according to $\mathrm{Po}(mq_1), ..., \mathrm{Po}(mq_{k-1})$. The statistical difference between the vector of resampled (Poissonized) random variables and the original vector is no greater than the statistical difference between $\mathrm{Po}(m(1-q_0))$ and the original distribution of $\hat{Z}$. For each ℓ, $\mathrm{Po}(mq_\ell)$ approximates the original distribution $\mathrm{Bin}(m, q_\ell)$ within error q_ℓ. The overall statistical distance between $Z_1, ..., Z_{k-1}$ and independent realizations $Z'_1, ..., Z'_{k-1}$ is thus at most $(1-q_0) + \sum_{\ell=1}^{k-1} q_\ell = 2(1-q_0)$. ■

Proof of Lemma 5.7. We can represent F_ℓ as a sum $\mathsf{F}_\ell = \mathsf{F}_\ell^{(1)} + \cdots + \mathsf{F}_\ell^{(k)}$, where $\mathsf{F}_\ell^{(i)}$ is the number of ℓ-way collisions among colors of type i. Since the types are independent, it suffices to show that for each i, the variables $\mathsf{F}_1^{(i)}, \ldots, \mathsf{F}_{k-1}^{(i)}$ are close to being independent. We can then sum the distances over the types to prove the lemma.

Let $\mathsf{F}_0^{(i)}$ denote the number of colors of type i that occur either 0 times, or k or more times, in the sample. The vector $\mathsf{F}_0^{(i)}, \mathsf{F}_1^{(i)}, \ldots, \mathsf{F}_{k-1}^{(i)}$ 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 $e^{-\lambda_i}\lambda_i^\ell/\ell!$, where $\lambda_i = a_i s/n$ for $\ell \in [k-1]$, and outcome 0 with the remaining probability. On each roll, outcome 0 occurs with probability at least $e^{-\lambda_i} \geq 1 - \lambda_i \geq 1/2$.

Claim 5.9 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 $O(\lambda_i)$. Summing over all types, the distance of $\mathsf{F}_1, \ldots, \mathsf{F}_{k-1}$ from independent is $O\left(\sum_i \lambda_i\right) = O\left(\frac{k a_{k-1} s}{n}\right)$. ■

We now give the third lemma needed for the proof of Theorem 5.5.

Lemma 5.10 *For $\ell = 1, ..., k-1$, $\hat{\mathsf{F}}_\ell \approx_{\delta_3} \tilde{\mathsf{F}}_\ell$ where*

$$\delta_3 = O\left(\frac{k \cdot a_{k-1} \cdot s}{n} + \frac{\left(\frac{a_{k-1}}{n}\right)^{k-1} \cdot s^k}{\lfloor \frac{k}{2} \rfloor ! \cdot \lceil \frac{k}{2} \rceil !}\right) .$$

The fact that $\hat{\mathsf{X}}$ and $\tilde{\mathsf{X}}$ have proportional moments is used in the proof of Lemma 5.10 (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 to show that the moment conditions imply that $\hat{\mathsf{F}}_\ell$ and $\tilde{\mathsf{F}}_\ell$ have similar (though not equal) expectations. The proof is quite technical (see the full version of this paper [14]).

Given the three lemmas above, we can easily prove the main result of the section:

Proof of Theorem 5.5. The proof follows by a hybrid argument. Consider a chain of distributions "between" the two histograms of Theorem 5.5. Starting from the "hat" histogram, first replace all counts of collisions greater than k by 0, and then replace each count $\hat{\mathsf{F}}_\ell$ with an independent copy $\hat{\mathsf{F}}'_\ell$ for $\ell \in [k-1]$, as in Lemma 5.7. Next, change each $\hat{\mathsf{F}}'_\ell$ with a corresponding $\tilde{\mathsf{F}}'_\ell$. Finally, replace these independent $\tilde{\mathsf{F}}'_\ell$s with the actual variables $\tilde{\mathsf{F}}_\ell$ and add back the counts of the collisions involving more than k variables to obtain the "tilde" histogram. The resulting chain of distributions has $k+3$ steps, and looks as follows (here δ_1, δ_2 and δ_3 are as defined in Lemmas 5.6, 5.7 and 5.10, respectively):

$$\begin{array}{rcccc}
 & (\hat{\mathsf{F}}_1, & \ldots, & \hat{\mathsf{F}}_{k-1}, & \hat{\mathsf{F}}_k, \hat{\mathsf{F}}_{k+1}, \ldots) \\
\approx_{\delta_1} & (\hat{\mathsf{F}}_1, & \ldots, & \hat{\mathsf{F}}_{k-1}, & 0, 0, \ldots) \\
\approx_{\delta_2} & (\hat{\mathsf{F}}'_1, & \ldots, & \hat{\mathsf{F}}'_{k-1}, & 0, 0, \ldots) \\
\approx_{\delta_3} & (\tilde{\mathsf{F}}'_1, & \ldots, & \hat{\mathsf{F}}'_{k-1}, & 0, 0, \ldots) \\
 & & & \vdots & \\
\approx_{\delta_3} & (\tilde{\mathsf{F}}'_1, & \ldots, & \tilde{\mathsf{F}}'_{k-1}, & 0, 0, \ldots) \\
\approx_{\delta_2} & (\tilde{\mathsf{F}}_1, & \ldots, & \tilde{\mathsf{F}}_{k-1}, & 0, 0, \ldots) \\
\approx_{\delta_1} & (\tilde{\mathsf{F}}_1, & \ldots, & \tilde{\mathsf{F}}_{k-1}, & \tilde{\mathsf{F}}_k, \tilde{\mathsf{F}}_{k+1}, \ldots)
\end{array}$$

By the triangle inequality, the sum of the statistical differences between consecutive distributions in the chain is a bound on the total statistical difference:

$$\begin{aligned}
&2 \cdot \delta_1 + 2 \cdot \delta_2 + (k-1) \cdot \delta_3 \\
&\quad = O\left(\frac{1}{k!} \cdot \left(\frac{a_{k-1}}{n}\right)^{k-1} \cdot s^k + \frac{k \cdot a_{k-1} \cdot s}{n}\right. \\
&\quad \left. + k \cdot \frac{k \cdot a_{k-1} \cdot s}{n} + \frac{k}{\lfloor \frac{k}{2} \rfloor ! \cdot \lceil \frac{k}{2} \rceil !} \cdot \left(\frac{a_{k-1}}{n}\right)^{k-1} \cdot s^k\right).
\end{aligned}$$

The first and second terms are negligible given the others. Removing them yields the claimed bound. ■

6 Proof of Main Lower Bound

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 5.3) and the reductions to uniform algorithms (Lemma 3.4), and to Poisson algorithms (Lemma 5.2).

Recall that our goal is to give a lower bound on the number of queries required for a general algorithm for DE to distinguish inputs with at least $n/11$ colors from inputs with at most n/B colors (for $B > 11$). By combining Lemmas 5.2 and 3.4 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 3.4). Details follow.

Let $\hat{\mathsf{X}}$ and $\tilde{\mathsf{X}}$ obey the moments condition with parameters k and B, and let $D_{\tilde{\mathsf{X}}}$ and $D_{\hat{\mathsf{X}}}$ 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 $\frac{np_i}{\mathsf{E}[\mathsf{X}]}$ 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. (The choice of $\frac{s}{2}$ rather than s samples is made for the convenience of the analysis). Recall that Theorem 4.5 states that there exist $\hat{\mathsf{X}}$ and $\tilde{\mathsf{X}}$ such that $a_{k-1} = (B+3)^{k-1} < (B+3)^k$. We assume that this is in fact the case. By substituting this bound in Theorem 5.3, we get:

$$\left|\Pr[\mathcal{A}(D_{\hat{\mathsf{X}}}) = 1] - \Pr[\mathcal{A}(D_{\tilde{\mathsf{X}}}) = 1]\right| = O\Big(\frac{k\cdot(B+3)^k\cdot s}{n} + \frac{k}{\lfloor\frac{k}{2}\rfloor!\cdot\lceil\frac{k}{2}\rceil!}\cdot\frac{(B+3)^{k(k-1)}\cdot s^k}{n^{k-1}}\Big). \quad (4)$$

Set k and s as functions of B so that the error term in Equation (4) is $o(1)$. 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 $s = \left\lfloor n^{1-\frac{2}{k}}\right\rfloor$. 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 at most $n^{\frac{1}{4}}/\sqrt{\log n}$. To make the calculations easier, assume $n > 16$, so that $k < \sqrt{\log n}$. We handle the two summands in Equation (4) separately. By substituting k, s, and B in the first summand, $\frac{k\cdot(B+3)^k s}{n}$, one can show that it is at most $2^{-\sqrt{\frac{1}{2}\log\log(n)\log(n)}}$. The second summand, $\frac{k}{\lfloor\frac{k}{2}\rfloor!\cdot\lceil\frac{k}{2}\rceil!}\cdot\frac{(B+3)^{k(k-1)}s^k}{n^{k-1}} \leq 2^{-\frac{1}{2}\sqrt{\log\log(n)\log(n)}}$. By Equation (4) and these two bounds, $\left|\Pr[\mathcal{A}(D_{\hat{\mathsf{X}}}) = 1] - \Pr[\mathcal{A}(D_{\tilde{\mathsf{X}}}) = 1]\right| = O\left(2^{-\frac{1}{2}\sqrt{\log\log(n)\log(n)}}\right)$. This completes the proof of Theorem 2.1.

7 Lower Bound for Estimating Entropy

The following problem was introduced by Batu *et al.* [5]. Let $\mathbf{p} = \langle p_1, \ldots, p_n\rangle$ be a discrete distribution over n elements, where p_i is the probability of the ith 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 $\hat{\mathsf{X}}$ and $\tilde{\mathsf{X}}$ over $a_0 = 1, a_1 = 2B, a_2 = 4B-2$, that satisfy the moments condition with parameters 3 and B. Moreover, $\mathsf{E}[\tilde{\mathsf{X}}] = 2$ and $\mathsf{E}[\hat{\mathsf{X}}] = 2B$.*

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

The two distributions and their entropies. As in Section 5, given the two random variables $\hat{\mathsf{X}}$ and $\tilde{\mathsf{X}}$, define two distributions over n elements (or, more precisely, two families of distributions). One distribution, denoted $\mathbf{p}_{\hat{\mathsf{X}}}$, 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 $\mathbf{p}_{\tilde{\mathsf{X}}}$, has support on $\frac{n}{2B}$ elements of weight $\frac{2B}{n}$ each. We define two families of distributions, $F_{\hat{\mathsf{X}}}$ and $F_{\tilde{\mathsf{X}}}$, respectively, where we allow

all permutations over the names (colors) of the elements. Let $D'_{\hat{\mathsf{X}}}$ denote the uniform distribution over $F_{\hat{\mathsf{X}}}$, and let $D'_{\tilde{\mathsf{X}}}$ denote the uniform distribution over $F_{\tilde{\mathsf{X}}}$.

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 $F_{\tilde{\mathsf{X}}}$ is $\beta \log n$, and the entropy of each distribution in $F_{\hat{\mathsf{X}}}$ is $\frac{2B-2}{4B-3} \cdot \log n + \frac{2B-1}{4B-3} \cdot \log \frac{n}{4B-2}$, which is at most $\frac{1+\beta}{2} \log n - 1$ by our choice of B. Thus, the ratio between the entropies is $\frac{1+\beta}{2\beta} - o(1)$.

While Theorem 5.3 is stated for the distributions on strings, $D_{\hat{\mathsf{X}}}$ and $D_{\tilde{\mathsf{X}}}$, and algorithms taking uniform samples from an input string of length n, it is not hard to verify that it also holds for the distributions $D'_{\hat{\mathsf{X}}}$ and $D'_{\tilde{\mathsf{X}}}$ and algorithms that are provided with samples from distributions over n elements. Since $k = 3$ and $a_2 = 2n^{1-\beta}$, to distinguish the two distributions one has 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 discussions. Ronitt's involvement in the initial stages of this project was especially valuable.

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] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *J. Comput. Syst. Sci.*, 58(1):137–147, 1999.

[3] Z. Bar-Yossef. *The complexity of Massive Data Set Computations*. PhD thesis, Computer Science Division, U.C. Berkeley, 2002.

[4] Z. Bar-Yossef, R. Kumar, and D. Sivakumar. Sampling algorithms: lower bounds and applications. In *Proceedings of STOC*, pages 266–275, New York, NY, USA, 2001. ACM Press.

[5] T. Batu, S. Dasgupta, R. Kumar, and R. Rubinfeld. The complexity of approximating the entropy. *SIAM Journal on Computing*, 35(1):132–150, 2005. Conference version appeared in the proceedings of Computational Complexity, 2002.

[6] T. Batu, E. Fischer, L. Fortnow, R. Kumar, R. Rubinfeld, and P. White. Testing random variables for independence and identity. In *Proceedings of FOCS*, pages 442–451, 2001.

[7] T. Batu, L. Fortnow, R. Rubinfeld, W. Smith, and P. White. Testing that distributions are close. In *Proceedings of FOCS*, pages 259–269, 2000.

[8] J. Bunge. Bibliography on estimating the number of classes in a population. Available from `www.stat.cornell.edu/~bunge/`.

[9] M. Charikar, S. Chaudhuri, R. Motwani, and V. R. Narasayya. Towards estimation error guarantees for distinct values. In *PODS*, pages 268–279. ACM, 2000.

[10] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., New York, 1991.

[11] P. Indyk and D. Woodruff. Tight lower bounds for the distinct elements problem. In *Proceedings of FOCS*, pages 283–288, 2003.

[12] Y. V. Prohorov. Asymptotic behavior of the binomial distribution (Russian). *Uspekhii Matematicheskiikh Nauk*, 8(3):135–142, 1953. Moscow.

[13] S. Raskhodnikova, D. Ron, R. Rubinfeld, and A. Smith. Sublinear algorithms for approximating string compressibility. In *Proceedings of the 11th RANDOM*, pages 609–623, 2007.

[14] S. Raskhodnikova, D. Ron, A. Shpilka, and A. Smith. Strong lower bounds for approximating distribution support size and the distinct elements problem. Manuscript, 2007.

[15] W. Szpankowski. *Average Case Analysis of Algorithms on Sequences*. John Wiley & Sons, Inc., New York, 2001.

[16] M. 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.

[17] J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23:337–343, 1977.