SIAM J. COMPUT.
Vol. 39, No. 3, pp. 813–842

STRONG LOWER BOUNDS FOR APPROXIMATING DISTRIBUTION SUPPORT SIZE AND THE DISTINCT ELEMENTS PROBLEM*

SOFYA RASKHODNIKOVA[†], DANA RON[‡], AMIR SHPILKA[§], AND 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. Charikar, Chaudhuri, Motwani, and Narasayya [in *Proceedings of the Nineteenth ACM SIGMOD–SIGACT–SIGART Symposium on Principles of Database Systems*, 2000, pp. 268–279] and Bar-Yossef, Kumar, and Sivakumar [in *Proceedings of the Thirty-Third Annual ACM Symposium on Theory of Computing*, ACM Press, New York, 2001, pp. 266–275] proved that multiplicative approximation for these problems within a factor $\alpha > 1$ requires $\Theta(\frac{n}{\alpha^2})$ queries to the input sequence. Their lower bound applies only when the number of distinct elements (or the support size of a distribution) is very small. For both problems, we prove a nearly linear in n lower bound on the query complexity, applicable even when the number of distinct elements is large (up to linear in n) and 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: $\mathsf{E}[\mathsf{X}_1]/\mathsf{E}[\mathsf{X}_2] = \mathsf{E}[\mathsf{X}_1^2]/\mathsf{E}[\mathsf{X}_2^2] = \cdots = \mathsf{E}[\mathsf{X}_1^k]/\,\mathsf{E}[\mathsf{X}_2^k]$. It is related to a well-studied mathematical question, the *truncated Hamburger problem*, but differs in the requirement that our random variables have to be supported on integers. Our lower bound method is also applicable to other problems and, in particular, gives a new lower bound for the sample complexity of approximating the entropy of a distribution.

Key words. approximation algorithms, distinct elements problem, distribution support size, lower bounds, Poissonization

AMS subject classifications. 68Q17, 68Q25

DOI. 10.1137/070701649

1. Introduction.

In this work we consider the following problem, which we call DISTRIBUTION-SUPPORT-SIZE:

> *Given a parameter n and 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:

> *Given access to a sequence of length n, approximate the number of* distinct *elements in the sequence.*

*Received by the editors August 30, 2007; accepted for publication (in revised form) May 15, 2009; published electronically August 6, 2009. A preliminary version of this paper appeared in the proceedings of FOCS 2007 [RRSS07].

http://www.siam.org/journals/sicomp/39-3/70164.html

[†]Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA 16802 (sofya@cse.psu.edu, asmith@cse.psu.edu). The research of these authors was partly supported by the National Science Foundation (NSF/CCF grant 0729171) and was partly done while at the Weizmann Institute of Science, Israel. The fourth author's research was supported at Weizmann by the Louis L. and Anita M. Perlman Postdoctoral Fellowship.

[‡]Department of Electrical Engineering Systems, Tel Aviv University, Ramat Aviv, Israel (danar@eng.tau.ac.il). This author's research was supported by the Israel Science Foundation (grant 89/05).

[§]Department of Computer Science, Technion, Haifa, 32000, Israel (shpilka@cs.technion.ac.il). This author's research was supported by the Israel Science Foundation (grant 439/06).

Both of these fundamental problems arise in many contexts and have been extensively studied. In statistics, Distribution-Support-Size is known as estimating the number of species in a population (see the list of hundreds of references in [Bun]). Typically, the input distribution is assumed to come from a specific family. Distinct-Elements arises in databases and data mining, for example, in the design of query optimizers, and the detection of denial-of-service attacks (see [CCMN00, ABRS03] and the references therein). Because of the overwhelming size of modern databases, a significant effort has focused on solving Distinct-Elements with extremely efficient classes of algorithms: streaming algorithms [FM85, AMS99, GT02, BKS02, BJK$^+$02, IW03, BHR$^+$07], which make a single pass through the data and use very little memory, and sampling-based algorithms [CCMN00, BKS01], which query only a small number of positions in the input.

This paper looks at the complexity of sampling-based approximation algorithms for Distribution-Support-Size and Distinct-Elements. Previous works consider multiplicative approximation for these problems. Charikar et al. [CCMN00] give an algorithm for approximating Distinct-Elements within a factor α with $O\left(\frac{n}{\alpha^2}\right)$ queries into the input sequence.[1] Charikar et al. and Bar-Yossef, Kumar, and Sivakumar [BKS01] prove a matching lower bound of $\Omega\left(\frac{n}{\alpha^2}\right)$ queries. 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 Distribution-Support-Size 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 Distribution-Support-Size when the support size is a nonnegligible fraction of n. In other words, is it possible to obtain efficient algorithms for Distribution-Support-Size and Distinct-Elements that have *additive* error at most βn, where $\beta \in (0, \frac{1}{2})$? Since additive approximation for these problems has not been explicitly studied before this work, the only known algorithm with additive error[2] follows directly from the multiplicative approximation algorithm of Charikar et al. [CCMN00] and runs in time $\Theta\left((1-2\beta)n\right)$. In contrast, the "needle-in-a-haystack" multiplicative lower bound arguments [CCMN00, BKS01] imply only that algorithms with additive error βn require $\Omega(1/\beta)$ samples (by setting $\alpha = \sqrt{\beta n}$). We give a strong lower bound for the sample (and hence, 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 Distribution-Support-Size and Distinct-Elements are related. An instance of Distribution-Support-Size where all probabilities are multiples of $\frac{1}{n}$ is equivalent to a Distinct-Elements instance that can be accessed only by taking independent uniform samples with replacement. Thus, the follow-

[1] We say an algorithm approximates a function f within a factor $\alpha > 1$ if for every input x, with probability at least 2/3, the algorithm's output lies between $\frac{f(x)}{\alpha}$ and $\alpha f(x)$.

[2] This algorithm works by setting $\alpha^2 = 1/(1-2\beta)$, running a multiplicative approximation algorithm to get an output $\hat{c}$, and outputting $\frac{\min(\alpha\cdot\hat{c}, n)+\hat{c}/\alpha}{2}$.

ing problem is a special case of DISTRIBUTION-SUPPORT-SIZE and a restriction of DISTINCT-ELEMENTS: *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 DISTINCT-ELEMENTS can be made without loss of generality. In principle, an algorithm for DISTINCT-ELEMENTS is allowed to make arbitrary adaptive queries to the input. However, Bar-Yossef, Kumar, and Sivakumar [BKS01] and Bar-Yossef [Bar02] show 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 DISTINCT-ELEMENTS as general algorithms. We strengthen their 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 DISTRIBUTION-SUPPORT-SIZE and general DISTINCT-ELEMENTS, it suffices to prove bounds for the restriction of DISTINCT-ELEMENTS above.

Main lower bound. We prove that even if we allow additive error which is a constant fraction of n, so that the multiplicative lower bound [CCMN00, BKS01] implies only that a constant number of queries are necessary, approximating DISTINCT-ELEMENTS (and hence DISTRIBUTION-SUPPORT-SIZE) 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 $\frac{n}{23}$ requires $n^{1-o(1)}$ samples. If we restrict our attention to algorithms that sample balls uniformly with replacement (or if we consider DISTRIBUTION-SUPPORT-SIZE), then the bound can be strengthened: $n^{1-o(1)}$ samples are necessary to distinguish an input with $n - \frac{n}{\mathsf{d}}$ colors from an input with $\frac{n}{\mathsf{d}}$ colors, for any $\mathsf{d} = n^{o(1)}$. In particular, obtaining additive error $(\frac{1}{2} - \delta)n$ requires $n^{1-o(1)}$ samples for any constant $\delta > 0$. (Note that one can obtain additive error $\frac{n}{2}$ without taking any samples at all, since the number of distinct colors is always between 1 and n.) In the above statements and in all that follow, distinguishing means *distinguishing with success probability at least* 2/3. It is an open question to close the gap between our lower bounds and the trivial $O(n)$ upper bound.

To contrast our result with previous bounds, consider a scenario where we receive a petition with n signatures and at least $n/15$ distinct people need to have signed in order for the petition to be valid. Our results imply that even if only $n/100$ people actually signed the petition, convincing ourselves that the petition is invalid requires reading a nearly linear number of signatures. Previous results are based on the difficulty of distinguishing petitions with a single distinct signer from petitions with many signers; in our scenario, they yield only a constant lower bound on the query complexity. More generally, needle-in-a-haystack lower bound techniques are very weak in the scenario of distinguishing a valid petition from a petition with $n/100$ distinct signers: These techniques rely on constructing positive and negative instances for the problem that differ in very few entries, while in our scenario, the instances being distinguished must have linear Hamming distance.

We note that it is easy to prove an $\Omega\left(\sqrt{n}\right)$ bound on the query complexity of approximating DISTINCT-ELEMENTS 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 $o\left(\sqrt{n}\right)$ samples is likely to see only unique colors (no collisions). With $\Omega\left(\sqrt{n}\right)$ 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. One might hope 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. However, we show that 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.

Moments condition 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 [CCMN00, BKS01], 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,

$$\frac{\mathsf{E}[\mathsf{X}_1]}{\mathsf{E}[\mathsf{X}_2]} = \frac{\mathsf{E}[\mathsf{X}_1^2]}{\mathsf{E}[\mathsf{X}_2^2]} = \cdots = \frac{\mathsf{E}[\mathsf{X}_1^{k-1}]}{\mathsf{E}[\mathsf{X}_2^{k-1}]} \tag{1}$$

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. The problem of constructing two such random variables is related to the "truncated moments problem" (see, e.g., [And70, CF91]), and the similarities and differences between the two problems are further discussed in section 4 (after the statement of Theorem 4.5).

Let F_ℓ be the number of ℓ-way collisions, that is, the number of colors that appear exactly ℓ times in the random sample obtained by an algorithm that samples the input positions uniformly with replacement (recall that such algorithms are essentially as good as general algorithms for solving DISTINCT-ELEMENTS). As explained in the discussion of the main lower bound, computing F_ℓ for small ℓ gives a possible strategy for distinguishing two DISTINCT-ELEMENTS 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 DISTINCT-ELEMENTS 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 actually gives a general technique for proving lower bounds on sample complexity; 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 [BFR$^+$00, BFF$^+$01]. 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 $\mathsf{F}_1, \dots, \mathsf{F}_{k-1}$ 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, possibly counterintuitive, 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* [Szp01], in which one modifies a probability experiment to replace a fixed quantity (e.g., the number of samples) with a variable quantity which follows a Poisson distribution. This breaks up dependencies between variables and makes the analysis tractable.

1.3. Results for other problems.

Compressibility. As shown in [RRRS07], Distinct-Elements is closely related to the problem of approximating how well the Lempel–Ziv compression scheme, defined in [ZL77], compresses an input string x. Let $C_{\mathrm{LZ}}(x)$ denote the length of the compressed version of x. An approximation algorithm for C_{LZ} with multiplicative factor $\alpha \ge 1$ and additive error βn, where $\beta \in [0,1]$, has to produce, given input x of length n, an estimate $\widehat{C}$ that with probability $\frac{2}{3}$ satisfies $\frac{C_{\mathrm{LZ}}(x)}{\alpha} - \beta n \le \widehat{C} \le \alpha \cdot C_{\mathrm{LZ}}(x) + \beta n$. While [RRRS07] shows that analogous problems for other compression schemes (such as run-length encoding) have time complexity independent of n, their approximation algorithm for C_{LZ} runs in time $\tilde{O}\left(\frac{n}{\alpha^3 \beta}\right)$. This implies that for all $\epsilon > 0$, one can distinguish, in sublinear time $\tilde{O}(n^{1-\epsilon})$, the case that $C_{\mathrm{LZ}}(x) = O(n^{1-\epsilon})$ from the case that $C_{\mathrm{LZ}}(x) = \Omega(n)$ (by setting $\alpha = c_1 n^{\epsilon/2}$ and $\beta = c_2 n^{-\epsilon/2}$ for an appropriate choice of c_1 and c_2). In conjunction with the reduction from Distinct-Elements to approximating C_{LZ}, presented in [RRRS07], the lower bound we give for Distinct-Elements implies that the algorithm for C_{LZ} cannot be improved significantly. In particular, distinguishing the case that $C_{\mathrm{LZ}}(x) = O(n^{1-\epsilon})$ from the case that $C_{\mathrm{LZ}}(x) = \tilde{\Omega}(n)$ requires reading $\Omega(n^{1-c\sqrt{\epsilon}})$ symbols of x (where c is a constant). For more precise details (on the reduction and the exact form of the resulting lower bound), see [RRRS07].

Entropy estimation. Our methodology yields a general technique for proving lower bounds on the sample complexity for other problems where one needs to compute quantities invariant under permutation of the balls and the colors. Namely, 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 apply the technique to estimating the entropy of an unknown distribution. In section 7, we give a lower bound of $\Omega(n^{\frac{2}{6\alpha^2-3+o(1)}})$ for approximating the entropy of a distribution over n elements to within a multiplicative factor of α. When α is close to 1, this bound is close to $\Omega(n^{2/3})$. It can be combined with the $\Omega(n^{\frac{1}{2\alpha^2}})$ bound of Batu et al. [BDKR05] to give $\Omega(n^{\max\{\frac{1}{2\alpha^2}, \frac{2}{6\alpha^2-3+o(1)}\}})$.

1.4. Subsequent research. Subsequent to our work, Valiant [Val08] provided a novel, generic lower bound for testing symmetric properties of distributions, using a generalization of the techniques in this paper. In particular, for estimating entropy, [Val08] provides a lower bound of $n^{(\frac{1}{\alpha^2}-o(1))}$ samples, a significant strengthening of the bound provided here. For distribution support, [Val08] provides a slightly weaker statement than ours, but with the same qualitative interpretation: Namely, for any constants d_1 and d_2, distinguishing distributions with at least $\frac{n}{\mathsf{d}_1}$ from those with at most $\frac{n}{\mathsf{d}_2}$ colors requires $n^{1-o(1)}$ samples.[3]

2. Main results. As noted in the introduction, Distinct-Elements with algorithms that sample uniformly with replacement is a special case of Distribution-Support-Size where all probabilities are integer multiples of $\frac{1}{n}$. Theorem 2.1, stated next, gives a lower bound on Distinct-Elements algorithms, first for the restricted case of uniform sampling (and hence also for Distribution-Support-Size), and then for the general case.

Theorem 2.1. *For all $T \geq 2n^{3/4}\sqrt{\log n}$, if we set*

$$k = k(n,T) = \left\lfloor \sqrt{\frac{\log n}{\log n - \log T + \frac{1}{2}\log\log n + 1}} \right\rfloor,$$

then the following hold.

1. *Every algorithm for* Distinct-Elements that takes uniform samples without replacement *needs to perform $\Omega\big(n^{1-\frac{2}{k}}\big)$ queries to distinguish inputs with at least $\boldsymbol{n-T}$ colors from inputs with at most $\boldsymbol{T}$ colors.*
   The same bound holds for algorithms for Distribution-Support-Size*, even under the promise that probabilities are integer multiples of $1/n$.*
2. *Every algorithm for* Distinct-Elements*,* regardless of how it accesses the input*, needs to perform $\Omega\big(n^{1-\frac{2}{k}}\big)$ queries to distinguish inputs with at least $\boldsymbol{\frac{n}{11}}$ colors from inputs with at most $\boldsymbol{T}$ colors.*

The next corollary provides a simpler form of the lower bound for T that is not too large and not too small, and another simple form for sufficiently large T. The corollary is obtained by setting $T = n^{1-\epsilon}$ in the main theorem.

Corollary 2.2. *The following hold:*

- *If $\frac{5\log\log n+10}{\log n} \leq \epsilon \leq 1/16$, then distinguishing inputs of* Distinct-Elements *with at least $\frac{n}{11}$ colors from inputs with at most $n^{1-\epsilon}$ colors requires $\Omega(n^{1-3\sqrt{\epsilon}})$ queries.*
- *If $\epsilon < \frac{5\log\log n+10}{\log n}$, then distinguishing inputs of* Distinct-Elements *with at least $\frac{n}{11}$ colors from inputs with at most $n^{1-\epsilon}$ colors requires $n^{1-O(\sqrt{\log\log n/\log n})}$ queries.*

In both statements, the input with $n/11$ colors may be taken to have $n - n^{1-\epsilon}$ colors when the algorithm is restricted to uniform samples without replacement or when the problem to be solved is Distribution-Support-Size.

To prove Theorem 2.1 we construct a pair of Distinct-Elements 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 Distinct-Elements it suffices to consider algorithms that take uniform samples with replacement. We use this to deduce part 2 of Theorem 2.1 from part 1. In section 4, we construct integer random variables that satisfy the moments condition, as described in the introduction (equation (1)). Section 5 shows that frequency variables with proportional moments

[3] The result for distribution support requires a slight extension of the techniques in the conference paper [Val08]. The extension will appear in the forthcoming full version.

lead to indistinguishable instances of Distinct-Elements. 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.

3. Algorithms for Distinct-Elements with uniform samples.

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

First, consider algorithms that take their samples uniformly at random *without replacement* from $[n]$. The following lemma by Bar-Yossef, Kumar, and Sivakumar [BKS01, Lemma 9] shows that such algorithms are essentially as good for solving Distinct-Elements as general algorithms.

Lemma 3.1 (see [BKS01]). *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 algorithms that sample uniformly *without replacement* to ones that sample uniformly *with replacement* and find out the corresponding color, but not the input position that was queried. Bar-Yossef, Kumar, and Sivakumar [BKS01, full version, Lemma 4.17] (also in [Bar02, Lemma 4.19]) 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 Distinct-Elements, 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 gets to see only the colors of the samples, but not the input positions corresponding to them.*

As noted in the introduction, a uniform algorithm for Distinct-Elements is equivalent to an algorithm for the special case of Distribution-Support-Size where all probabilities are integer multiples of $\frac{1}{n}$.

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* Distinct-Elements *within a factor of $(\sqrt{0.1} \cdot \alpha)$, there is a* uniform *algorithm $\mathcal{A}'$ that takes s samples and provides, with probability at least $\frac{2}{3}$, an approximation for* Distinct-Elements *within a factor of α.*

Proof. We define algorithm $\mathcal{A}'$ as follows. It simulates algorithm $\mathcal{A}$. Whenever $\mathcal{A}$ makes a query to a new position of the input, $\mathcal{A}'$ takes a uniform sample with replacement from its input and records its color. This color is used in the simulation as an answer to $\mathcal{A}$'s current query and all subsequent queries to the same position. $\mathcal{A}'$ returns the output of $\mathcal{A}$, multiplied by $\sqrt{10}$. Note that the randomness that $\mathcal{A}'$ uses for sampling is *independent* of the coins of $\mathcal{A}$.

Clearly, if $\mathcal{A}$ makes s queries, then $\mathcal{A}'$ takes at most s samples. To analyze its accuracy, consider algorithm $\mathcal{A}''$ that first runs $\mathcal{A}'$ and then completes its record to a full Distinct-Elements instance with n colors by taking (at least $n-s$) uniform samples with replacement from its input and recording their colors for the positions that $\mathcal{A}$ did not query. $\mathcal{A}''$ outputs the same answer as $\mathcal{A}'$. Thus, it is enough to analyze the accuracy of $\mathcal{A}''$.

If there are $C = C(n)$ colors in the input of $\mathcal{A}''$, the recorded instance has at most C colors. However, some of the colors might be missing. We will show later (see Claim 3.5 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 recorded instance has between $0.1 \cdot C$ and C colors. Because the coins of $\mathcal{A}$ are independent of those used by $\mathcal{A}'$ for sampling, we can apply $\mathcal{A}$'s accuracy guarantee: When $\mathcal{A}$ is run on the recorded 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, since $\mathcal{A}''$ 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. □

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

Lemma 3.4. *Let $C_1 = C_1(n)$ and $C_2 = C_2(n)$, where $0.1 \cdot C_1 > C_2$. If every* uniform *algorithm needs at least s queries to distinguish* Distinct-Elements *instances with at least C_1 colors from* Distinct-Elements *instances with at most C_2 colors, then* every *algorithm needs $\Omega(s)$ queries to distinguish* Distinct-Elements *instances with at least $0.1 \cdot C_1$ colors from* Distinct-Elements *instances with at most C_2 colors.*

The following claim was used in the proof of Lemma 3.3.

Claim 3.5. *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,

$$\mathsf{E}[X] = \sum_{i=1}^{C} \mathsf{E}[X_i] \geq C\left(1 - \left(1 - \frac{1}{n}\right)^s\right) \geq C\left(1 - e^{-(s/n)}\right) \geq \left(1 - e^{-1}\right)\frac{Cs}{n}. \tag{2}$$

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 $\mathsf{Cov}[X_i, X_j] = \mathsf{E}[X_i X_j] - \mathsf{E}[X_i]\,\mathsf{E}[X_j]$ is negative (if color i was not selected, it is more likely that color j was selected). Therefore,

$$\begin{aligned}\mathsf{Var}[X] &= \mathsf{Var}\left[\sum_{i=1}^{C} X_i\right] \\ &= \sum_{i=1}^{C} \mathsf{Var}[X_i] + 2 \sum_{i,j:1 \leq i < j \leq C} \mathsf{Cov}[X_i, X_j] \\ &\leq \sum_{i=1}^{C} \mathsf{Var}[X_i]\end{aligned}$$

$$\leq \sum_{i=1}^{C} \mathsf{E}[X_i] = \mathsf{E}[X].$$

The last inequality holds because X_i is a Bernoulli variable for each color i, and, consequently, $\mathsf{Var}[X_i] = \mathsf{E}[X_i](1 - \mathsf{E}[X_i]) \leq \mathsf{E}[X_i]$. By Chebyshev's inequality and since $\mathsf{Var}[X] \leq \mathsf{E}[X]$, for any fixed $\delta < 1$,

$$\begin{aligned} \Pr\left[X \leq \delta\, \mathsf{E}[X]\right] &\leq \Pr\left[|X - \mathsf{E}[X]| \geq (1-\delta)\, \mathsf{E}[X]\right] \\ &\leq \frac{\mathsf{Var}[X]}{((1-\delta)\, \mathsf{E}[X])^2} \\ &\leq \frac{1}{(1-\delta)^2\, \mathsf{E}[X]}. \end{aligned} \tag{3}$$

Set $\delta = 3 - \sqrt{8}$. If $\mathsf{E}[X] \geq \frac{4}{(1-\delta)^2}$, then by (2) and (3), with probability $\geq \frac{3}{4}$,

$$X \geq \delta\, \mathsf{E}[X] \geq \delta(1 - e^{-1})\frac{Cs}{n} > \frac{Cs}{10n},$$

as stated in the claim. Otherwise, that is, if $\mathsf{E}[X] < \frac{4}{(1-\delta)^2}$, equation (2) 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. □

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 $\ell = 0, 1, \dots, s$, let F_ℓ be the number of ℓ-way collisions in the sample. The* histogram *F of the sample is the vector $(F_1, \dots, F_s)$, indicating for each nonzero ℓ how many colors appear exactly ℓ times in the sample.*

One can prove that any uniform algorithm for Distinct-Elements can be simulated by a uniform algorithm that sees only a histogram of the sample. (We omit the proof since it follows from the formal argument provided subsequently.)

To prove our lower bound, we will define a pair of Distinct-Elements instances that contain a significantly different number of colors, but for which the corresponding distributions on histograms are indistinguishable. In what follows we provide an intuitive discussion that will lead us to our main formal definitions and claims.

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 ℓ-way collisions can be obtained from the number of monochromatic ℓ'-tuples for all $\ell' \geq \ell$ by applying a simple recursive formula.[4] 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 DISTINCT-ELEMENTS, a corresponding *frequency variable.*

DEFINITION 4.2 (frequency variable). *Suppose we are given an instance of* DISTINCT-ELEMENTS *with $\frac{n}{d}$ colors. Group colors into* types *according to how many times they appear in the input: say, a 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,

$$\mathsf{E}[\mathsf{X}] \;=\; \sum_i p_i a_i \;=\; d. \tag{4}$$

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 DISTINCT-ELEMENTS 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

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

The last equality follows from (4) and from the fact that $\Pr[\mathsf{X} = a_i] = p_i$. 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 $\widetilde{\mathsf{X}}$ have $k-1$* proportional moments *if*

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

[4] That is, let us denote by $M_{\ell'}$ the number of monochromatic ℓ'-tuples, and recall that F_ℓ denotes the number of ℓ-way collisions. Then, assuming that $F_k = M_k = 0$, we have that $F_{k-1} = M_{k-1}$, $F_{k-2} = M_{k-2} - (k-1)F_{k-1}$, and, in general, $F_\ell = M_\ell - \sum_{\ell'=\ell+1}^{k-1} \binom{\ell'}{\ell} F_{\ell'}$.

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* $\widetilde{\mathsf{X}}$ satisfy the moments condition *with parameters* k *and* B *if* $\hat{\mathsf{X}}$ *and* $\widetilde{\mathsf{X}}$ *have* $k-1$ *proportional moments and* $\frac{\mathsf{E}[\widetilde{\mathsf{X}}]}{\mathsf{E}[\hat{\mathsf{X}}]} \geq B$.

Theorem 4.5 (random variables satisfying the moments condition). *For all integers* $k > 1$ *and* $B > 1$*, there exist random variables* $\hat{\mathsf{X}}$ *and* $\widetilde{\mathsf{X}}$ *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* $a_i = (B+3)^i$*,* $\mathsf{E}[\widetilde{\mathsf{X}}] > B$*, and* $\mathsf{E}[\hat{\mathsf{X}}] < 1 + \frac{1}{B}$.

The relation to the "truncated moments problem." Our technique for constructing random variables that satisfy the moments condition and the moments condition problem itself are related to a well-studied mathematical problem known as the "truncated moments problem," or the "truncated Hamburger problem" (see, e.g., [And70, CF91]). In this problem we are given some domain of interest K (e.g., the interval $K = [a, b]$, the real line $K = \mathbb{R}$, etc.) and a set of values $\gamma_0, \gamma_1, \ldots, \gamma_m$. The goal is to construct a measure (random variable) μ on K such that $\int_K t^j d\mu(t) = \gamma_j$ for every $0 \leq j \leq m$. This is called the *truncated* problem, as opposed to the *full* moments problem (see, e.g., [Akh65]), since we are interested only in a finite number of moments. Note that γ_j is the jth moment of the random variable constructed. In [CF91] the authors give necessary and sufficient conditions on $\{\gamma_j\}_{j=0}^m$ for the solvability of the problem. Moreover, they show that a solution exists in which the support of μ is finite and is in fact of size $(m+1)/2$ (when m is odd). Next we briefly discuss the similarities and differences between the Hamburger problem and our moments condition.

At first glance it may seem that there is an easy reduction from the problem of constructing random variables that satisfy the moments condition to the Hamburger problem. Indeed, pick any set $\{\gamma_j\}_{j=0}^m$ that satisfies the sufficient conditions provided in [Akh65] and construct the random variable $\hat{\mathsf{X}}$, which, as stated above, has a finite support. Then construct the (finitely supported) random variable $\widetilde{\mathsf{X}}$ for $\{B \cdot \gamma_j\}_{j=0}^m$ (that also satisfies the sufficient conditions). It is clear that $\hat{\mathsf{X}}$ and $\widetilde{\mathsf{X}}$ satisfy the moments condition. The problem with this approach is that we cannot guarantee that $\hat{\mathsf{X}}$ and $\widetilde{\mathsf{X}}$ are supported on integers, which is what we need. Moreover, it is not clear what the largest element of the support is. Our techniques require a bound on this quantity (see Theorem 5.6).

In addition to the similarity between the definitions of the problems, another similarity is the techniques used in the proofs: In both cases the Vandermonde matrix and its properties play an important role. This is not surprising since the moments of a random variable with a finite support are vectors in the image of the appropriate Vandermonde matrix.

To conclude, the Hamburger problem and the problem of constructing random variables that satisfy the moments condition are related. However, it is not clear how to use a solution for the former problem in order to solve the latter problem when the support of the distributions should be on integers and the largest element in the support should be bounded. Our result (Theorem 4.5) can be seen as a complementary result that adds to the literature on the Hamburger problem.

Proof of Theorem 4.5. The rest of the section is devoted to this proof, which comprises three parts: an overview, the construction, and its analysis. To reduce notation, in the proof, and more generally 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 ($\hat{}$), and those pertaining to the second by a tilde ($\tilde{}$). In statements relevant to both instances, the corresponding variables without hat or tilde are used.

Overview. Let $C = \mathsf{E}[\widetilde{\mathsf{X}}]/\mathsf{E}[\hat{\mathsf{X}}]$. Then the moments condition (Definition 4.4) can be restated as $(\mathsf{E}[\widetilde{\mathsf{X}}], \mathsf{E}[\widetilde{\mathsf{X}}^2], \ldots, \mathsf{E}[\widetilde{\mathsf{X}}^{k-1}]) = C \cdot (\mathsf{E}[\hat{\mathsf{X}}], \mathsf{E}[\hat{\mathsf{X}}^2], \ldots, \mathsf{E}[\hat{\mathsf{X}}^{k-1}])$ and $C \geq B$. Recall that the supports of $\hat{\mathsf{X}}$ and $\widetilde{\mathsf{X}}$ 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: $V(C \cdot \vec{\hat{p}} - \vec{\tilde{\mathsf{p}}}) = \vec{0}$. For a fixed a, there is a unique (up to a factor) nonzero vector $\vec{u}$ satisfying $V \cdot \vec{u} = \vec{0}$. To obtain probability vectors $\vec{\hat{p}}$ and $\vec{\tilde{\mathsf{p}}}$ 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 $\widetilde{\mathsf{X}}$ for each a.

To find an appropriate choice of a and to demonstrate the required properties of our construction, we explicitly compute a vector $\vec{u}$ that defines the distributions. The main idea behind this step is to view $\vec{u}$ as coefficients of a polynomial. Let $f(t) = t^{k-1} + u_{k-2}t^{k-2} + \cdots + u_0$ be the unique nonzero 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

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

For example, $s_2(y_1, y_2, y_3) = y_1y_2 + y_1y_3 + y_2y_3$. In general, $s_0 = 1$ and $s_{k-1}(y_1, \ldots, y_{k-1}) = y_1y_2 \cdots y_{k-1}$. 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 $s_3(a) = a^6$. In general, one can show that $s_{i-1}(a) < s_i(a)$ for sufficiently large a. For our analysis we need only some bounds on the $s_i(a)$'s in terms of $s_{k-1}(a)$, proved in Claim 4.8.

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, divided by a normalization factor. We define

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

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

where

$$\hat{\mathsf{N}}(a) \stackrel{\text{def}}{=} \sum_{j=0}^{\lfloor (k-1)/2 \rfloor} s_{k-1-2j}(a) \quad \text{and} \quad \widetilde{\mathsf{N}}(a) \stackrel{\text{def}}{=} \sum_{j=0}^{\lfloor (k-2)/2 \rfloor} s_{k-2-2j}(a)$$

are normalization factors.

Analysis. After proving auxiliary lemmas, Lemmas 4.6 and 4.7, we use them to complete the proof of Theorem 4.5. Lemma 4.6 shows that the distributions $\hat{\mathsf{X}}$ and $\widetilde{\mathsf{X}}$ have $k-1$ proportional moments (see Definition 4.3). Lemma 4.7 bounds $\mathsf{E}[\widetilde{\mathsf{X}}]$ and $\mathsf{E}[\hat{\mathsf{X}}]$.

LEMMA 4.6. *Let* $C \stackrel{\text{def}}{=} \hat{\mathsf{N}}(a)/\widetilde{\mathsf{N}}(a)$. *Then* $C \cdot \mathsf{E}[\hat{\mathsf{X}}^\ell] = \mathsf{E}[\widetilde{\mathsf{X}}^\ell]$ *for* $\ell = 1, \dots, k-1$.

Proof. By definition of $\hat{\mathsf{X}}$ and $\widetilde{\mathsf{X}}$,

$$\begin{aligned} C \cdot \mathsf{E}[\hat{\mathsf{X}}^\ell] - \mathsf{E}[\widetilde{\mathsf{X}}^\ell] &= C \cdot \sum_{\substack{0 \le i \le k-1 \\ i \text{ even}}} (a^i)^\ell \cdot \frac{s_{k-1-i}(a)}{\hat{\mathsf{N}}(a)} - \sum_{\substack{0 \le i \le k-1 \\ i \text{ odd}}} (a^i)^\ell \cdot \frac{s_{k-1-i}(a)}{\widetilde{\mathsf{N}}(a)} \\ &= \frac{1}{\widetilde{\mathsf{N}}(a)} \cdot \sum_{i=0}^{k-1} (-1)^i \cdot s_{k-1-i}(a) \cdot (a^\ell)^i = \frac{(-1)^{k-1}}{\widetilde{\mathsf{N}}(a)} \cdot f(a^\ell) = 0. \end{aligned}$$ □

LEMMA 4.7. *For all* $a > 3$,

1. $\mathsf{E}[\hat{\mathsf{X}}] < 1+\frac{1}{a-3}$;
2. $\mathsf{E}[\widetilde{\mathsf{X}}] > a-2$.

To prove Lemma 4.7, we bound the $s_i(a)$'s in terms of $s_{k-1}(a)$. Later, we express the expectations of $\hat{\mathsf{X}}$ and $\widetilde{\mathsf{X}}$ in terms of the $s_i(a)$'s. We then use the upper bound on the $s_i(a)$'s to get the upper bound on $\mathsf{E}[\hat{\mathsf{X}}]$, and both the lower bound on $s_{k-2}(a)$ and the upper bound on the $s_i(a)$'s to get the lower bound on $\mathsf{E}[\widetilde{\mathsf{X}}]$.

CLAIM 4.8. *For all* $a > 3$,

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

Proof. By the definition,

$$\begin{aligned} s_{k-1-i}(a) = s_{k-1-i}(a, a^2, \dots, a^{k-1}) &= \sum_{\substack{T \subseteq [k-1] \\ |T| = k-1-i}} \prod_{j \in T} a^j \\ &= \left(\prod_{j=1}^{k-1} a^j \right) \cdot \sum_{\substack{T \subseteq [k-1] \\ |T| = k-1-i}} \prod_{j \notin T} a^{-j} \\ &= s_{k-1}(a) \cdot \sum_{\substack{T \subseteq [k-1] \\ |T| = k-1-i}} \prod_{j \notin T} a^{-j} \\ &= s_{k-1}(a) \cdot \sum_{\substack{R \subseteq [k-1] \\ |R| = i}} \prod_{j \in R} a^{-j}. \end{aligned} \tag{5}$$

In particular, $s_{k-2}(a) = s_{k-1}(a) \cdot \left(\sum_{j=1}^{k-1} a^{-j}\right) > \frac{s_{k-1}(a)}{a}$, proving the first part of the claim.

We now prove the second part of the claim. For fixed integers k and a, and for all $i \in [k-1]$, let

$$\rho(i) \stackrel{\text{def}}{=} \sum_{\substack{R\subseteq[k-1]\\|R|=i}} \prod_{j\in R} a^{-j} = \sum_{\substack{R\subseteq[k-1]\\|R|=i}} a^{-\sum_{j\in R} j}.$$

By (5), it suffices to show that

$$\rho(i) < a^{-i(i-1)/2}(a-1)^{-i}. \tag{6}$$

We prove this by induction on i. By definition, $\rho(1) = \sum_{j=1}^{k-1} a^{-j} < \frac{1}{a-1}$, proving (6) for $i = 1$. For $i > 1$, observe that every subset $R \subseteq [k-1]$ of cardinality i can be represented as $R = R' \cup \{j\}$, where $|R'| = i-1$ and $j = \max_{j'\in R} j'$ (so that $j \in \{i, \ldots, k-1\}$). Moreover, each pair (R', j) corresponds to at most one R. Therefore,

$$\rho(i) \le \rho(i-1) \cdot \left(\sum_{j=i}^{k-1} a^{-j}\right) < \rho(i-1) \cdot \frac{a^{-i+1}}{a-1}. \tag{7}$$

By the induction hypothesis, $\rho(i-1) < a^{-(i-1)(i-2)/2}(a-1)^{-i+1}$. Substituting this into the previous equation, we get

$$\rho(i) < \frac{a^{-(i-1)(i-2)/2} \cdot a^{-i+1}}{(a-1)^{-i+1} \cdot (a-1)} = \frac{a^{-i(i-1)/2}}{(a-1)},$$

as claimed in (6). □

Proof of Lemma 4.7. 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 4.8(2),

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

To bound $\mathsf{E}[\widetilde{\mathsf{X}}]$ from below, we first bound $\widetilde{\mathsf{N}}(a)$ from above. Recall that $\widetilde{\mathsf{N}}(a) = \sum_{j=0}^{\lfloor (k-2)/2 \rfloor} s_{k-2-2j}(a)$. By Claim 4.8(2),

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

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

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

The last inequality follows from Claim 4.8(1). Thus, the proof of Lemma 4.7 is completed. □

In order to complete the proof of Theorem 4.5 it remains to find, for every $B > 1$, an a such that $\mathsf{E}[\widetilde{\mathsf{X}}]/\,\mathsf{E}[\hat{\mathsf{X}}] \geq B$. By Lemma 4.7,

$$\frac{\mathsf{E}[\widetilde{\mathsf{X}}]}{\mathsf{E}[\hat{\mathsf{X}}]} > \frac{a-2}{1 + \frac{1}{a-3}} = a - 3.$$

Thus, if we take $a = B + 3$, then $\mathsf{E}[\widetilde{\mathsf{X}}]/\,\mathsf{E}[\hat{\mathsf{X}}] > B$, $\mathsf{E}[\hat{\mathsf{X}}] < 1 + \frac{1}{B}$, and $\mathsf{E}[\widetilde{\mathsf{X}}] > B$. This completes the construction and the proof of Theorem 4.5.

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. [BDKR02] (conference version of [BDKR05]) 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 a random variable distributed according to $\mathrm{Po}(\lambda)$ are both λ. Several properties of the Poisson distribution are collected in Claim A.2 (see Appendix A).

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 5.3, stated next. Batu et al. [BDKR02] (conference version of [BDKR05]) proved a variant of Lemma 5.3 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 DISTINCT-ELEMENTS. In Lemma 5.3 below, and throughout this section, we use statistical difference to bound the distinguishability of distributions.

DEFINITION 5.2 (statistical difference). *Distributions P and Q over a domain S have* statistical difference *(also called* total variation distance*) δ if* $\max_{S' \subseteq S} |P(S') - Q(S')| = \delta$. *We write $P \approx_\delta Q$ to denote that P and Q have statistical difference* at most δ.

For two random variables $X \sim P$ and $Y \sim Q$, we say that X and Y have statistical difference δ (and write $X \approx_\delta Y$) when $P \approx_\delta Q$.

Statistical difference provides a convenient measure of distinguishability between random variables: If X and Y have small statistical difference δ, then any algorithm $\mathcal{A}$ will behave almost identically on the two variables; that is, $\Pr[\mathcal{A}(X) = 1] = \Pr[\mathcal{A}(Y) = 1] \pm \delta$. We have collected several standard properties of statistical difference in Claim A.1 (see Appendix A).

Lemma 5.3 (Poissonization lemma, generalizes [BDKR02], conference version of [BDKR05]).

(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 at most $4/s$.*

(b) *If the input to* Distinct-Elements *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.*

Item (a) implies that it suffices to show lower bounds for Poisson algorithms in order to prove similar bounds for uniform algorithms. The independence of the number of occurrences of different colors in the sample (item (b)) greatly simplifies the analysis of Poisson algorithms. We prove Lemma 5.3 in Appendix A.

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 Distinct-Elements instances based on frequency variables is straightforward if we make an integrality assumption described below. Specifically, for $k > 1$ let $a_0 < a_1 < \cdots < a_{k-1}$ 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 Distinct-Elements instance D_{X} of length n (that is, a string in $[n]^n$) that contains $\frac{n}{\mathsf{E}[\mathsf{X}]}$ colors. For $i = 0, \ldots, k-1$, instance D_{X} contains $\frac{np_i}{\mathsf{E}[\mathsf{X}]}$ colors of *type* i, where each color of type i appears a_i times. We now give a precise definition that does not rely on the assumption that $\frac{np_i}{\mathsf{E}[\mathsf{X}]}$ is an integer for every i.

Definition 5.4 (the instance D_{X}). *For $k > 1$, let $a_0 < a_1 < \cdots < a_{k-1}$ be integers, and let X be a random variable over these integers defined by* $\Pr[\mathsf{X} = a_i] = p_i$. *Observe that $\mathsf{E}[\mathsf{X}] = \sum_{i=0}^{k-1} p_i \cdot a_i$. Based on X, we form a* Distinct-Elements *instance D_{X} of length n (that is, a string in $[n]^n$) that contains M_{X} colors, where $M_{\mathsf{X}} = \sum_{i=0}^{k-1} \left\lfloor \frac{np_i}{\mathsf{E}[\mathsf{X}]} \right\rfloor + n - \sum_{i=0}^{k-1} \left\lfloor \frac{np_i}{\mathsf{E}[\mathsf{X}]} \right\rfloor \cdot a_i$. (Note that if $\frac{np_i}{\mathsf{E}[\mathsf{X}]}$ is an integer for every i, then $M_{\mathsf{X}} = \frac{n}{\mathsf{E}[\mathsf{X}]}$.) For $i = 0, \ldots, k-1$, instance D_{X} contains $\left\lfloor \frac{np_i}{\mathsf{E}[\mathsf{X}]} \right\rfloor$ colors of* type i, *each appearing a_i times. In addition, there are $n - \sum_{i=0}^{k-1} \left\lfloor \frac{np_i}{\mathsf{E}[\mathsf{X}]} \right\rfloor \cdot a_i$ colors that appear once each. We refer to these singleton colors as being of type k and set $a_k = 1$.*

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.

The following claim bounds the distortion introduced to handle nonintegral values of $\frac{np_i}{\mathsf{E}[\mathsf{X}]}$.

Claim 5.5. *In the instance D_{X} of Definition* 5.4, *the following hold.*

1. *The number of colors of type k, called C_k, is at most $\sum_{i=1}^{k-1} a_i \leq k \cdot a_{k-1}$.*
2. *The number of distinct colors M_{X} is at least $\frac{n}{\mathsf{E[X]}}$ and at most $\frac{n}{\mathsf{E[X]}} + C_k$.*

Proof. We can write n as $\sum_{i=0}^{k-1} \frac{np_i}{\mathsf{E[X]}}$, and hence rewrite C_k as

$$C_k = \sum_{i=0}^{k-1} \left(\frac{np_i}{\mathsf{E[X]}} - \left\lfloor \frac{np_i}{\mathsf{E[X]}} \right\rfloor \right) \cdot a_i \, .$$

This is at most $\sum_{i=0}^{k-1} a_i$. Since a_{k-1} is larger than all other a_i, the sum is at most $k \cdot a_{k-1}$. Finally, to bound the number of distinct colors, we write $M_{\mathsf{X}} - \frac{n}{\mathsf{E[X]}} = \sum_{i=0}^{k-1} \left(\left\lfloor \frac{np_i}{\mathsf{E[X]}} \right\rfloor - \frac{np_i}{\mathsf{E[X]}} \right) + C_k$. Using the expression above for C_k, we get

$$M_{\mathsf{X}} - \tfrac{n}{\mathsf{E[X]}} = \sum_{i=0}^{k-1} \left(\tfrac{np_i}{\mathsf{E[X]}} - \left\lfloor \tfrac{np_i}{\mathsf{E[X]}} \right\rfloor \right) \cdot (a_i - 1).$$

This is nonnegative since the a_i's are all positive integers and are bounded above by our previous expression for C_k. □

Note that because it suffices 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 study are oblivious to them.

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 Distinct-Elements, $D_{\hat{\mathsf{X}}}$ 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 5.6 (distinguishability by Poisson algorithms). *Let $\hat{\mathsf{X}}, \tilde{\mathsf{X}}$ be random variables over positive integers $a_0 < a_1 < \cdots < a_{k-1}$ that 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,*

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

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

Corollary 5.7. *Let $\hat{\mathsf{X}}$ and $\tilde{\mathsf{X}}$ be* fixed (w.r.t. n) *random variables that 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 5.6. As in Definition 4.1, for $\ell = 0, 1, 2, \ldots$, 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 (the sequence is infinite since a Poisson algorithm can, in principle, see collisions involving an arbitrary number of elements of the same color). We can restate Theorem 5.6 in terms of the statistical difference between histogram distributions.

THEOREM 5.8 (distinguishability by Poisson algorithms, restated). *For $s \le \frac{n}{2\cdot a_{k-1}}$, the statistical difference between histogram random variables* $(\hat{\mathsf{F}}_1, \hat{\mathsf{F}}_2, \hat{\mathsf{F}}_3, \dots)$ *and* $(\widetilde{\mathsf{F}}_1, \widetilde{\mathsf{F}}_2, \widetilde{\mathsf{F}}_3, \dots)$ *is*

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

For the remainder of this section, assume $s \le \frac{n}{2\cdot a_{k-1}}$. The proof of Theorem 5.8 relies on the following three lemmas. Lemma 5.9 states that ℓ-way collisions are very unlikely for $\ell \ge k$, when s is sufficiently small. Lemma 5.10 shows that for both instances $D_{\hat{\mathsf{X}}}$ and $D_{\widetilde{\mathsf{X}}}$, 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 5.12 shows that the number of ℓ-way collisions is distributed almost identically when sampling from $D_{\hat{\mathsf{X}}}$ and from $D_{\widetilde{\mathsf{X}}}$, for every $\ell < k$. The fact that $\hat{\mathsf{X}}$ and $\widetilde{\mathsf{X}}$ have proportional moments is used only for this last result; the first two lemmas hold as long as the a_i's are bounded.

LEMMA 5.9. *For both instances $D_{\hat{\mathsf{X}}}$ and $D_{\widetilde{\mathsf{X}}}$, 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. By Lemma 5.3(b), the probability that the algorithm sees k or more balls of that color is

$$\Pr_{X\sim\mathrm{Po}(a_i s/n)}[X \ge k] = e^{-a_i s/n}\sum_{t\ge k}\frac{1}{t!}\left(\frac{a_i s}{n}\right)^t.$$

Since $t! > k!(t-k)!$, each of the terms in the sum is bounded above by the product $\frac{1}{k!}(\frac{a_i s}{n})^k\cdot\frac{1}{(t-k)!}(\frac{a_i s}{n})^{t-k}$. We can factor out the term $\frac{1}{k!}(\frac{a_i s}{n})^k$ from the sum, and obtain the following bound:

$$\begin{aligned}\Pr_{X\sim\mathrm{Po}(a_i s/n)}[X \ge k] &= e^{-a_i s/n}\sum_{t\ge k}\tfrac{1}{t!}(\tfrac{a_i s}{n})^t\\ &\le \tfrac{1}{k!}(\tfrac{a_i s}{n})^k e^{-a_i s/n}\sum_{t\ge k}\tfrac{1}{(t-k)!}(\tfrac{a_i s}{n})^{t-k}\\ &= \tfrac{1}{k!}(\tfrac{a_i s}{n})^k.\end{aligned}$$

Let $C_i = \lfloor\frac{np_i}{\mathsf{E}[\mathsf{X}]}\rfloor$ denote the number of colors of type i for $0 \le i \le k-1$, and let $C_k = n - \sum_{i=0}^{k-1}\lfloor\frac{np_i}{\mathsf{E}[\mathsf{X}]}\rfloor\cdot a_i$ denote the number of colors of type k (which appear once each). 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:

$$\begin{aligned}\sum_{i=0}^{k} C_i\cdot\frac{1}{k!}\left(\frac{a_i s}{n}\right)^k &= \frac{s^k}{k!\cdot n^{k-1}}\cdot\sum_{i=0}^{k}\frac{C_i\cdot a_i^k}{n}\\ &\le \frac{s^k}{k!\cdot n^{k-1}}\cdot\left(\sum_{i=0}^{k-1}\left\lfloor\frac{p_i a_i^k}{\mathsf{E}[\mathsf{X}]}\right\rfloor + \frac{1}{n}\cdot C_k\right)\\ &\le \frac{s^k}{k!\cdot n^{k-1}}\cdot\left(\frac{\mathsf{E}[\mathsf{X}^k]}{\mathsf{E}[\mathsf{X}]}+1\right).\end{aligned} \tag{8}$$

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 in (8) completes the proof. $\square$

Recall that we write $X \approx_\delta Y$ to denote that the statistical difference between the random variables X and Y is at most δ.

LEMMA 5.10. *For both instances $D_{\hat{\mathsf{X}}}$ 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 $\delta_2 \leq \frac{2k \cdot a_{k-1} \cdot s}{n}$.*

The proof of Lemma 5.10 relies on the following claim, which 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.

As mentioned above, Claims A.1 and A.2 state several standard properties of statistical difference and the Poisson distribution, respectively, that are used below. One of these properties is a bound on the statistical difference between the Poisson distribution and the binomial distribution (in Claim A.2(3)). We use $\mathrm{Bin}(m,p)$ to denote the binomial distribution with parameters m and p, that is, the distribution on the number of heads in a sequence of m independent coin flips, each of which comes up heads with probability p.

CLAIM 5.11. *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 m 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 $(Z_1, \ldots, Z_{k-1}) \approx_{\delta_4} (Z'_1, \ldots, Z'_{k-1})$ for $\delta_4 \leq 2(1-q_0)$, regardless of the number of times m that the die is rolled.*

Proof. Let $Z_{\neq 0}$ count the number of times that side 0 does not come up; i.e., $Z_{\neq 0} = m - Z_0 = \sum_{\ell=1}^{k-1} Z_\ell$. This count follows a binomial distribution $Z_{\neq 0} \sim \mathrm{Bin}(m, 1-q_0)$. By Claim A.2(3), 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 $Z_{\neq 0}$, the variables $Z_1, \ldots, Z_{k-1}$ follow a multinomial distribution. As in the proof of Lemma 5.3, if $Z_{\neq 0}$ itself is chosen according to $\mathrm{Po}(m(1-q_0))$, and $Z_1, \ldots, Z_{k-1}$ are resampled according to this value of $Z_{\neq 0}$, the resulting distribution on $Z_1, \ldots, Z_{k-1}$ is a vector of independent Poisson random variables distributed according to $\mathrm{Po}(mq_1), \ldots, \mathrm{Po}(mq_{k-1})$.

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 $Z_{\neq 0}$, by the data processing inequality of Claim A.1(3) (note that in both cases, the vector of counts is the same randomized function of the value of $Z_{\neq 0}$).

For each ℓ, $\mathrm{Po}(mq_\ell)$ approximates the original distribution $\mathrm{Bin}(m, q_\ell)$ within error q_ℓ (Claim A.2(3)). By the triangle inequality for statistical difference, the overall statistical distance between $Z_1, \ldots, Z_{k-1}$ and independent realizations $Z'_1, \ldots, Z'_{k-1}$ is thus at most $(1-q_0) + \sum_{\ell=1}^{k-1} q_\ell = 2(1-q_0)$. $\square$

Proof of Lemma 5.10. Observe that the number of ℓ-way collisions F_ℓ 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, $\mathsf{F}_\ell = \mathsf{F}_\ell^{(1)} + \cdots + \mathsf{F}_\ell^{(k)}$, 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 $\mathsf{F}_1^{(i)}, \dots, \mathsf{F}_{k-1}^{(i)}$ 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 $\mathsf{F}_0^{(i)}, \mathsf{F}_1^{(i)}, \dots, \mathsf{F}_{k-1}^{(i)}$ 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 5.11 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 $2\lambda_i$. Summing over all types, the distance of $\mathsf{F}_1, \dots, \mathsf{F}_{k-1}$ from being independent is at most $2\sum_i \lambda_i \leq \frac{2ka_{k-1}s}{n}$.

(N.B.: Subsequent to this work, Valiant [Val08] applied a more general form of Claim 5.11 to obtain a bound of $\frac{2a_{k-1}s}{n}$ on the distance from being independent.) □

We now give the third lemma needed for the proof of Theorem 5.8.

LEMMA 5.12. *For $\ell = 1, \dots, k-1$, $\hat{\mathsf{F}}_\ell \approx_{\delta_3} \widetilde{\mathsf{F}}_\ell$, where*

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

As noted previously, the fact that $\hat{\mathsf{X}}$ and $\widetilde{\mathsf{X}}$ have proportional moments is used in the proof of Lemma 5.12 (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 moments condition implies that $\hat{\mathsf{F}}_\ell$ and $\widetilde{\mathsf{F}}_\ell$ have similar (though not equal) expectations.

Proof. Recall, as in the proof of Lemma 5.10, that the number of ℓ-way collisions F_ℓ is a sum of independent binomial random variables, one for each type. More specifically,

$$\mathsf{F}_\ell \sim \sum_{i=0}^{k} \mathrm{Bin}\left(C_i\ , \frac{e^{-\lambda_i}\lambda_i^{\ell}}{\ell!}\right). \tag{9}$$

When p is small, the Poisson distribution $\mathrm{Po}(\lambda = pm)$ is a good approximation to the binomial distribution $\mathrm{Bin}(m, p)$; the statistical difference between the two is at most p (Claim A.2(3)). To approximate F_ℓ with a Poisson variable, we replace terms in the sum (9) with Poisson random variables, one at a time. Recall that if a random variable Y is independent of random variables X and X' such that $X \approx_\delta X'$, then $X + Y \approx_\delta X' + Y$ (this follows from items 3 (data processing) and 5 (independent pairs) of the statistical difference, stated in Claim A.1). Hence each replacement of a binomial by a Poisson random variable induces a change of at most p in statistical difference in the distribution of the sum. By the triangle inequality, we can sum the differences to obtain a bound on the total change induced by these replacements. Since the sum of independent Poisson variables is also a Poisson variable (see Claim A.2(2)), if we let $\lambda^{(\ell)} = \sum_{i=0}^{k} C_i \cdot \frac{\lambda_i^{\ell}}{\ell!} e^{-\lambda_i}$ and let $X_{\lambda^{(\ell)}}$ denote a random variable distributed as $\mathrm{Po}(\lambda^{(\ell)})$, then

$$\mathsf{F}_\ell \approx_{\gamma_\ell} X_{\lambda^{(\ell)}} \sim \mathrm{Po}\left(\lambda^{(\ell)} = \sum_{i=0}^{k} C_i \cdot \frac{\lambda_i^{\ell}}{\ell!} e^{-\lambda_i}\right),$$

where

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

In the second inequality we used $\frac{e^{-\lambda_i}\lambda_i^{\ell-1}}{\ell!} \leq 1$ for $0 \leq \lambda_i \leq 1$, and in the last inequality we used the fact that $a_k = 1$ and $a_i < a_{k-1}$ for every $i < k-1$ (so that $a_k + a_0 \leq a_{k-1}$).

Next, to bound the statistical difference between $\hat{\mathsf{F}}_\ell$ and $\widetilde{\mathsf{F}}_\ell$ from above, it is enough to bound the difference between $\hat{\lambda}^{(\ell)}$ and $\widetilde{\lambda}^{(\ell)}$, since the statistical difference between $\mathrm{Po}(\hat{\lambda}^{(\ell)})$ and $\mathrm{Po}(\widetilde{\lambda}^{(\ell)})$ is at most $|\hat{\lambda}^{(\ell)} - \widetilde{\lambda}^{(\ell)}|$ (see Claim A.2(4)).

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

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

where

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

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

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

Recall from Claim 5.5 that $C_k \leq \sum_{i=0}^{k-1} a_i$. Thus, for each j, $0 \leq j \leq k-\ell$, we have that

$$\begin{aligned}
\frac{s^{\ell+j}}{n^{\ell+j}} \cdot \sum_{i=0}^{k} C_i \cdot a_i^{\ell+j} &= \frac{s^{\ell+j}}{n^{\ell+j}} \cdot \left(\sum_{i=0}^{k-1} \left\lfloor \frac{p_i n}{\mathsf{E[X]}} \right\rfloor \cdot a_i^{\ell+j} + C_k\right) \\
&\leq \frac{s^{\ell+j}}{n^{\ell+j-1}} \cdot \frac{1}{\mathsf{E[X]}} \cdot \sum_{i=0}^{k-1} p_i \cdot a_i^{\ell+j} + \frac{s^{\ell+j}}{n^{\ell+j}} \cdot \sum_{i=0}^{k-1} a_i \\
&= \frac{s^{\ell+j}}{n^{\ell+j-1}} \cdot \frac{\mathsf{E[X^{\ell+j}]}}{\mathsf{E[X]}} + \frac{s^{\ell+j}}{n^{\ell+j}} \cdot \sum_{i=0}^{k-1} a_i.
\end{aligned}$$

Similarly (using $C_k \geq 0$),

$$\begin{aligned}
\frac{s^{\ell+j}}{n^{\ell+j}} \cdot \sum_{i=0}^{k} C_i \cdot a_i^{\ell+j} &= \frac{s^{\ell+j}}{n^{\ell+j}} \cdot \left(\sum_{i=0}^{k-1} \left\lfloor \frac{p_i n}{\mathsf{E[X]}} \right\rfloor \cdot a_i^{\ell+j} + C_k\right) \\
&\geq \frac{s^{\ell+j}}{n^{\ell+j}} \cdot \left(\sum_{i=0}^{k-1} \left(\frac{p_i n}{\mathsf{E[X]}} - 1\right) \cdot a_i^{\ell+j} + C_k\right) \qquad (10) \\
&\geq \frac{s^{\ell+j}}{n^{\ell+j-1}} \cdot \frac{1}{\mathsf{E[X]}} \cdot \sum_{i=0}^{k-1} p_i \cdot a_i^{\ell+j} - \frac{s^{\ell+j}}{n^{\ell+j}} \cdot \sum_{i=0}^{k-1} a_i^{\ell+j} \\
&= \frac{s^{\ell+j}}{n^{\ell+j-1}} \cdot \frac{\mathsf{E[X^{\ell+j}]}}{\mathsf{E[X]}} - \frac{s^{\ell+j}}{n^{\ell+j}} \cdot \sum_{i=0}^{k-1} a_i^{\ell+j}.
\end{aligned}$$

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

$$\begin{aligned}&\left|\hat{\lambda}^{(\ell)} - \widetilde{\lambda}^{(\ell)}\right| \\ &= O\left(\frac{1}{\ell!} \cdot \sum_{j=0}^{k-\ell} \frac{s^{\ell+j}}{n^{\ell+j}} \cdot 2\sum_{i=0}^{k-1} a_i^{\ell+j} + \frac{1}{\ell!(k-\ell)!} \cdot \frac{s^k}{n^{k-1}} \cdot \max\left\{\frac{\mathsf{E}[\hat{\mathsf{X}}^k]}{\mathsf{E}[\hat{\mathsf{X}}]}, \frac{\mathsf{E}[\widetilde{\mathsf{X}}^k]}{\mathsf{E}[\widetilde{\mathsf{X}}]}\right\}\right).\end{aligned}$$

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

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

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

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

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

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

Proof of Theorem 5.8. The proof follows by a hybrid argument. Consider a chain of distributions "between" the two histograms of Theorem 5.8.

- Starting from the "hat" histogram, first replace all counts of collisions greater than k by 0 (that is, by a random variable that takes the value 0 with probability 1). By Claim A.1(4), the resulting distribution has statistical difference at most δ_1 from the original, where δ_1 is the bound from Lemma 5.9 on the probability that any ℓ-way collisions occur for $\ell \ge k$.
- Next, replace each count $\hat{\mathsf{F}}_\ell$ with an independent copy $\hat{\mathsf{F}}'_\ell$ for $\ell \in [k-1]$, as in Lemma 5.10, introducing a change of δ_2.
- For the following $k-1$ steps, replace each $\hat{\mathsf{F}}'_\ell$ with a corresponding $\widetilde{\mathsf{F}}'_\ell$. By Lemma 5.12, each of these steps introduces a change of at most δ_3.
- Finally, replace these independent $\widetilde{\mathsf{F}}'_\ell$'s with the actual variables $\widetilde{\mathsf{F}}_\ell$ 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 has the following form (here δ_1, δ_2, and δ_3 are as defined in Lemmas 5.9, 5.10, and 5.12, respectively):

$$\begin{aligned}(\hat{\mathsf{F}}_1, \dots, \hat{\mathsf{F}}_{k-1}, \hat{\mathsf{F}}_k, \dots) &\approx_{\delta_1} (\hat{\mathsf{F}}_1, \dots, \hat{\mathsf{F}}_{k-1}, 0, \dots) \approx_{\delta_2} (\hat{\mathsf{F}}'_1, \dots, \hat{\mathsf{F}}'_{k-1}, 0, \dots) \\ &\approx_{\delta_3} (\widetilde{\mathsf{F}}'_1, \hat{\mathsf{F}}'_2, \dots, \hat{\mathsf{F}}'_{k-1}, 0, \dots) \\ &\approx_{\delta_3} \cdots \approx_{\delta_3} (\widetilde{\mathsf{F}}'_1, \dots, \widetilde{\mathsf{F}}'_{k-2}, \hat{\mathsf{F}}'_{k-1} 0, \dots) \\ &\approx_{\delta_3} (\widetilde{\mathsf{F}}'_1, \dots, \widetilde{\mathsf{F}}'_{k-1}, 0, \dots) \approx_{\delta_2} (\widetilde{\mathsf{F}}_1, \dots, \widetilde{\mathsf{F}}_{k-1}, 0, \dots) \\ &\approx_{\delta_1} (\widetilde{\mathsf{F}}_1, \dots, \widetilde{\mathsf{F}}_{k-1}, \widetilde{\mathsf{F}}_k, \dots).\end{aligned}$$

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

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

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

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

Proof of Theorem 2.1(1). Recall that in part 1 of the theorem, our goal is to give a lower bound on the number of queries required for a *uniform* algorithm for Distinct-Elements to distinguish inputs with at least $n - T$ colors from inputs with at most T colors. By Lemma 5.3, it suffices to give a lower bound on s for a Poisson-s algorithm that uses only the histogram of the samples.

Set $B = \frac{2n}{T}$ and

$$k = \left\lfloor \sqrt{\frac{\log n}{\log B + \frac{1}{2}\log\log n}} \right\rfloor = \left\lfloor \sqrt{\frac{\log n}{\log n - \log T + \frac{1}{2}\log\log n + 1}} \right\rfloor,$$

as in the statement of Theorem 2.1. Next, construct integer random variables $\hat{\mathsf{X}}$ and $\widetilde{\mathsf{X}}$ that obey the moments condition with parameters k and B, and let $D_{\widetilde{\mathsf{X}}}$ and $D_{\hat{\mathsf{X}}}$ be the corresponding Distinct-Elements instances. By Theorem 4.5, these random variables have expectation $\mathsf{E}[\widetilde{\mathsf{X}}] > B$ and $\mathsf{E}[\hat{\mathsf{X}}] < 1+\frac{1}{B}$, respectively, and are supported on integers less than $a_{k-1} = (B+3)^{k-1}$.

The corresponding Distinct-Elements instances have at least $\frac{n}{1+\frac{1}{B}} > n - T$ and at most $\frac{n}{B} + a_{k-1}k$ colors, respectively (Claim 5.5). To obtain a slightly cleaner bound, note that $a_{k-1} \leq (B+3)^k \leq 2^{\log(B)\sqrt{\log n/\log(B\log^{1/2} n)}} \leq 2^{\sqrt{\log B \log n}}$. Since $B < n^{1/4}$, we get $ka_{k-1} \leq k\sqrt{n} \leq \sqrt{n\log n} < \frac{n}{B}$, so the instance $D_{\widetilde{\mathsf{X}}}$ has at most $\frac{2n}{B} = T$ colors, as desired for Theorem 2.1.

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.) According to Theorem 5.6,

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

Recall that $a_{k-1} = (B+3)^{k-1} < (B+3)^k$. We assume that this is in fact the case. Therefore,

$$\left|\Pr[\mathcal{A}(D_{\hat{\mathsf{X}}}) = 1] - \Pr[\mathcal{A}(D_{\widetilde{\mathsf{X}}}) = 1]\right|$$
$$= O\left(\frac{k^2 \cdot (B+3)^k \cdot s}{n} + \frac{k}{\lfloor \frac{k}{2} \rfloor ! \cdot \lceil \frac{k}{2} \rceil !} \cdot \frac{(B+3)^{k(k-1)} \cdot s^k}{n^{k-1}} \right). \tag{11}$$

We can set k and s as functions of B so that the error term in (11) is $o(1)$: Given B, define q by the equality $B = \log(n)^q$; it follows that $k = \left\lfloor\sqrt{\frac{\log(n)}{(q+\frac{1}{2})\log\log(n)}}\right\rfloor$. Finally, set $s = \left\lfloor n^{1-\frac{2}{k}}\right\rfloor$. 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 (11) separately. We begin with the first summand:

$$\begin{aligned}
k^2 \cdot \frac{(B+3)^k s}{n} &< \log(n) \cdot \frac{\left(\log(n)^{q+\frac{1}{2}}\right)^k n^{1-\frac{2}{k}}}{n} \\
&= \log(n)\frac{\left(\log(n)^{q+\frac{1}{2}}\right)^k}{n^{\frac{2}{k}}} \\
&\leq \log(n)\frac{\left(2^{\log\log(n)\cdot(q+\frac{1}{2})}\right)^{\sqrt{\log(n)/((q+\frac{1}{2})\log\log(n))}}}{2^{\log(n)\cdot 2\sqrt{((q+\frac{1}{2})\log\log(n))/\log(n)}}} \\
&= \log(n)\frac{2^{\sqrt{(q+\frac{1}{2})\log\log(n)\log(n)}}}{2^{2\sqrt{(q+\frac{1}{2})\log\log(n)\log(n)}}} \\
&= 2^{\log\log(n)-\sqrt{(q+\frac{1}{2})\log\log(n)\log(n)}} \\
&< 2^{-\sqrt{\log\log(n)}(\sqrt{\frac{1}{2}\log(n)}-\sqrt{\log\log(n)})} \\
&< 2^{-\sqrt{\frac{1}{4}\log\log(n)\log(n)}}\,. \qquad (12)
\end{aligned}$$

The inequality in (12) holds for sufficiently large n.

We estimate the second summand in a similar fashion.

$$\begin{aligned}
\frac{k}{\left\lfloor\frac{k}{2}\right\rfloor! \cdot \left\lceil\frac{k}{2}\right\rceil!} \cdot \frac{(B+3)^{k(k-1)} s^k}{n^{k-1}} &= \frac{k}{(B+3)^k \left\lfloor\frac{k}{2}\right\rfloor! \cdot \left\lceil\frac{k}{2}\right\rceil!} \cdot \frac{(B+3)^{k^2} s^k}{n^{k-1}} \\
&< \frac{2}{(B+3)^k} \cdot \frac{(\log(n)^{q+\frac{1}{2}})^{k^2} n^{k-2}}{n^{k-1}} \\
&= \frac{2}{(B+3)^k} \cdot \frac{\left(\log(n)^{q+\frac{1}{2}}\right)^{\frac{\log n}{(q+\frac{1}{2})\log\log n}}}{n} \\
&= \frac{2}{(B+3)^k} \\
&< 2^{-\frac{1}{2}\sqrt{\log\log(n)\log(n)}}\,. \qquad (13)
\end{aligned}$$

Combining (11), (12), and (13) we get that

$$\left|\Pr[\mathcal{A}(D_{\hat{\mathsf{X}}}) = 1] - \Pr[\mathcal{A}(D_{\tilde{\mathsf{X}}}) = 1]\right| = O\left(2^{-\frac{1}{2}\sqrt{\log\log(n)\log(n)}}\right).$$

By applying Lemma 5.3(a) and recalling that $o(1/s) = o(n^{-(1-2/k)}) = o(2^{-\frac{1}{2}\log(n)})$, the proof of Theorem 2.1(1) is completed. □

Proof of Theorem 2.1(2). The reduction to uniform algorithms allows us to deduce part 2 of the theorem from part 1. Applying Lemma 3.4, with $C_1 = n-T$ and $C_2 = T$, we conclude that no *general* algorithm for Distinct-Elements can distinguish inputs with at least $\frac{n-T}{10}$ colors from those with at most T colors. For n large enough, $\frac{n-T}{10} > \frac{n}{11}$, and we obtain the desired statement. □

7. A lower bound for approximating the entropy.

The following problem was introduced by Batu et al. [BDKR05]. 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 $\tilde{O}\big(n^{\frac{1+\eta}{\alpha^2}}\big)$, provided that $H(\mathbf{p}) = \Omega\big(\frac{\alpha}{\eta}\big)$. They also proved a lower bound of $\Omega\big(n^{\frac{1}{2\alpha^2}}\big)$ that holds even when $H(\mathbf{p}) = \Omega\big(\frac{\log n}{\alpha^2}\big)$. (Without a lower bound on $H(\mathbf{p})$, the time complexity is unbounded.)

Here we use our technique to obtain a lower bound of $\Omega\big(n^{\frac{2}{6\alpha^2 - 3 + o(1)}}\big)$, improving on the $\Omega\big(n^{\frac{1}{2\alpha^2}}\big)$ 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 in the more general construction in Theorem 4.5, leading to better bounds.

LEMMA 7.1 (random variables satisfying the moments condition with $k = 3$). *For all integers $B > 1$, there exist random variables $\hat{\mathsf{X}}$ and $\widetilde{\mathsf{X}}$ over $a_0 = 1$, $a_1 = 2B$, $a_2 = 4B - 2$ that satisfy the moments condition with parameters 3 and B. Moreover, $\mathsf{E}[\widetilde{\mathsf{X}}] = 2$ 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 $\Pr[\widetilde{\mathsf{X}} = a_0] = \Pr[\widetilde{\mathsf{X}} = a_2] = 0$, $\Pr[\widetilde{\mathsf{X}} = a_1] = 1$. By definition of $\hat{\mathsf{X}}$ and $\widetilde{\mathsf{X}}$,

$$\mathsf{E}[\hat{\mathsf{X}}] = \left(1 - \frac{1}{4B-3}\right) \cdot 1 + \frac{1}{4B-3} \cdot (4B-2) = \frac{4B - 4 + 4B - 2}{4B-3} \quad = \quad 2$$

and

$$\mathsf{E}[\hat{\mathsf{X}}^2] = \left(1 - \frac{1}{4B-3}\right) \cdot 1^2 + \frac{1}{4B-3} \cdot (4B-2)^2 = \frac{4B - 4 + 16B^2 - 16B + 4}{4B-3} \quad = \quad 4B$$

while

$$\mathsf{E}[\widetilde{\mathsf{X}}] = 1 \cdot 2B = 2B \quad \text{and} \quad \mathsf{E}[\widetilde{\mathsf{X}}^2] = 4B^2.$$

As required, $\frac{\mathsf{E}[\widetilde{\mathsf{X}}]}{\mathsf{E}[\hat{\mathsf{X}}]} = B$, and $\frac{\mathsf{E}[\hat{\mathsf{X}}^2]}{\mathsf{E}[\hat{\mathsf{X}}]} = \frac{\mathsf{E}[\widetilde{\mathsf{X}}^2]}{\mathsf{E}[\widetilde{\mathsf{X}}]}$. □

The two distributions and their entropies. As in section 5, given the two random variables $\hat{\mathsf{X}}$ and $\widetilde{\mathsf{X}}$, 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 (see below on why we may assume that these quantities are all integers). The second distribution, denoted $\mathbf{p}_{\widetilde{\mathsf{X}}}$, has support on $\frac{n}{2B}$ elements of weight $\frac{2B}{n}$ each. We define two families of distributions, $F_{\hat{\mathsf{X}}}$ and $F_{\widetilde{\mathsf{X}}}$, 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 $D'_{\widetilde{\mathsf{X}}}$ denote the uniform distribution over $F_{\widetilde{\mathsf{X}}}$.

Let $B = B(n)$ be of the form $B = \frac{1}{2}n^{1-\beta}$ for rational $\beta < 1$. (If β is rational, there is an infinite family of integers n for which the numbers of colors of each type are all integers.) Then the entropy of each distribution in $F_{\widetilde{\mathsf{X}}}$ is $\beta \log n$, and the entropy of each distribution in $F_{\hat{\mathsf{X}}}$ is

$$\begin{aligned}
&\frac{2B-2}{4B-3}\cdot\log n+\frac{2B-1}{4B-3}\cdot\log\frac{n}{4B-2}\\
&=\frac{1}{2}\cdot\left(\log n+\log\frac{n}{4B-2}\right)-\frac{1}{8B-6}\cdot\left(\log n-\log\frac{n}{4B-2}\right)\\
&\geq\frac{1}{2}\cdot\big(\log n+\log n^{\beta}-1\big)-\frac{\log(2n^{1-\beta})}{4n^{1-\beta}-6}\\
&\geq\frac{1+\beta}{2}\log n-1,
\end{aligned}$$

where the last inequality holds for sufficiently large n. Thus, the ratio between the entropies is $\frac{1+\beta}{2\beta}-o(1)$.

While Theorem 5.6 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 $\Omega\big(\big(\frac{n}{a_2}\big)^{2/3}\big)=\Omega\big(n^{2\beta/3}\big)$ samples. In other words, $\Omega\big(n^{2\beta/3}\big)=\Omega\big(n^{\frac{2}{6\alpha^2-3+o(1)}}\big)$ samples are required for $\alpha=\big(\sqrt{\frac{1+\beta}{2\beta}}-o(1)\big)$-estimating the entropy.

Appendix A. Statistical difference and Poisson sampling. We state here some useful properties of statistical difference and use them to prove the Poissonization lemma (Lemma 5.3).

CLAIM A.1 (properties of statistical difference). *For all random variables X, Y, Z, X', Y' over a discrete domain S, the following hold.*

1. *The statistical difference between X and Y equals*

$$\frac{1}{2}\sum_{a\in S}\big|\Pr[X=a]-\Pr[Y=a]\big|.$$

2. *(Triangle inequality) If $X\approx_{\delta_1}Y$ and $Y\approx_{\delta_2}Z$, then $X\approx_{\delta_1+\delta_2}Z$.*
3. *(Data processing inequality) If $X\approx_{\delta}Y$, then $f(X)\approx_{\delta}f(Y)$ for all functions f.*
4. *If $\Pr[X=a]\geq 1-\delta$ for some constant a, then $(X,Y)\approx_{\delta}(a,Y)$.*
5. *If $X\approx_{\delta}X'$ and Y is independent of both X and X', then $(X,Y)\approx_{\delta}(X',Y)$.*
6. *$X\approx_{\delta}Y$ if and only if there exist (possibly dependent) random variables X^*, Y^* such that marginal distributions of X^* and Y^* are the same as those of X and Y, respectively, and $\Pr[X^*=Y^*]\geq 1-\delta$.*

Proof. Items 1 through 5 are standard; see, for example, Vadhan's thesis [Vad99, Fact 2.2.2].

To prove item 6, suppose first that X', Y' exist that are distributed individually as X, Y and are equal with probability $1-\delta$. Let E denote the event that $X'=Y'$. For any event S', we can write $\Pr[X\in S']=\Pr[X'\in S']=\Pr[X'\in S'\mid E]\cdot\Pr[E]+\Pr[X'\in S'\mid\bar{E}]\cdot\Pr[\bar{E}]$, and similarly for $\Pr[Y\in S']$. The first term in the two developments is the same. Thus $\Pr[X\in S']-\Pr[Y\in S']=\Pr[\bar{E}]\cdot\big(\Pr[X'\in S'\mid\bar{E}]-\Pr[Y'\in S'\mid\bar{E}]\big)$. This is at most $\Pr[\bar{E}]\leq\delta$ in absolute value, as desired.

To prove the other direction of the "if and only if" in item 6, suppose that $X\approx_{\delta}Y$. For each element a in the universe S, let $u_a=\min\{\Pr[X=a],\Pr[Y=a]\}$.

The sum $\sum_a u_a$ is the common area under the probability mass functions of X and Y. Consequently, $\sum_a u_a = 1-\delta^*$, where $\delta^* \leq \delta$ is the actual statistical difference between X and Y (since statistical difference equals half of the area under the probability mass function of X that is not also under the mass function of Y, and vice versa). Let U be drawn according to the probabilities

$$\Pr[U = a] = \frac{u_a}{1-\delta^*}.$$

(These probabilities sum up to 1, and so the distribution is well defined). Define independent random variables X_{rest} and Y_{rest} so that

$$\Pr[X_{rest} = a] = \frac{\max\{0, \Pr[X = a] - u_a\}}{\delta^*},$$
$$\Pr[Y_{rest} = a] = \frac{\max\{0, \Pr[Y = a] - u_a\}}{\delta^*}.$$

X_{rest} and Y_{rest} are also well-defined random variables.

Finally, with probability $1-\delta^*$ set $X^* = Y^* = U$, and with probability δ^* set $X^* = X_{rest}$ and $Y^* = Y_{rest}$. It is easy to verify that X^* and Y^* have the same marginal distributions as X and Y and are equal with probability $1-\delta^* \geq 1-\delta$. □

Claim A.2 (properties of the Poisson distribution).

1. *If* $X \sim \mathrm{Po}(\lambda)$, *then* $\mathsf{E}[X] = \mathsf{Var}[X] = \lambda$.
2. *If* $X \sim \mathrm{Po}(\lambda)$, $Y \sim \mathrm{Po}(\lambda')$, *and* X, Y *are independent, then* $X + Y \sim \mathrm{Po}(\lambda + \lambda')$.
3. *The statistical difference between* $\mathrm{Bin}(m, p)$ *and* $\mathrm{Po}(mp)$ *is at most* p.
4. *The statistical difference between* $\mathrm{Po}(\lambda)$ *and* $\mathrm{Po}(\lambda')$ *is at most* $|\lambda - \lambda'|$.

Note that item 4 provides a good bound when λ is near or equal to 0. One can strengthen the bound to $O\big(\frac{|\lambda-\lambda'|}{\sqrt{1+\min(\lambda,\lambda')}}\big)$. We do not need the latter bound here.

Proof. Items 1 and 2 can be found in any standard probability text (see, e.g., [Was04]). For item 3 (and other bounds on the Poisson approximation to the binomial), see [Pro53] or [Web99, Bound b_1].

Finally, we prove item 4: A Poisson random variable $Y \sim \mathrm{Po}(\lambda + \Delta)$ can be written as a sum of two independent Poisson variables $X_\lambda \sim \mathrm{Po}(\lambda)$ and $X_\Delta \sim \mathrm{Po}(\Delta)$ (this is possible because of item 2). Conditioned on the event $X_\Delta = 0$, the sum is distributed as $\mathrm{Po}(\lambda)$. This event occurs with probability $e^{-\Delta} \geq 1 - \Delta$. The statistical difference between $\mathrm{Po}(\lambda)$ and $\mathrm{Po}(\lambda + \Delta)$ is thus at most Δ, as desired. (Recall that the statistical difference between distributions p and q is at most Δ if and only if there is a pair of (dependent) random variables (A, B) such that the marginal distributions of A and B are p and q, respectively, and $\Pr[A = B] \geq 1 - \Delta$.) □

We now turn to the proof of Lemma 5.3.

Proof of Lemma 5.3. Let X_λ denote a random variable distributed as $\mathrm{Po}(\lambda)$.

(a) Consider the Poisson-s algorithm $\mathcal{A}'$ that outputs “fail” if it receives fewer than $s/2$ samples, and runs $\mathcal{A}$ on its first $s/2$ samples otherwise. The statistical difference between $\mathcal{A}$ and $\mathcal{A}'$ is at most $\Pr[X_s < \frac{s}{2}]$, since the variables have equal distribution conditioned on an event of mass $1 - \Pr[X_s < \frac{s}{2}]$ (see Claim A.1(6)). The Chebyshev inequality provides a good enough bound here. Since X_s has expectation and variance s, we obtain

$$\Pr[X_s < \tfrac{s}{2}] \leq \Pr[|X_s - \mathsf{E}[X_s]| \geq \tfrac{s}{2}] \leq \frac{\mathsf{Var}[X_s]}{(s/2)^2} = \frac{4}{s}.$$

(b) The conversion from multinomial to Poisson sampling is common in, for example, statistical analysis of categorical data. We prove the equivalence here for completeness. Consider an algorithm that *independently* samples a number of balls distributed as $\mathrm{Po}(\frac{b_c \cdot s}{n})$ for each color, where b_c is the number of times that color c appears. First, note that the total number of samples it takes is distributed as a Poisson random variable (by Claim A.2) with parameter $\sum_{\text{colors } c} \frac{b_c \cdot s}{n} = s$.

We now show that, conditioned on this sampler taking a particular number of samples s_0, the list of colors it sees is distributed identically to the case where s_0 balls are sampled uniformly at random. This is sufficient to show that the independent sampler produces the same distribution as the sampler which first selects the number of balls according to a Poisson random variable and then samples uniformly with replacement.

Consider a particular run in which e_c balls of color c are seen. In the case of s_0 uniform samples without replacement, this arises with multinomial probability: $\frac{s_0!}{\prod_c e_c!} \prod_c (\frac{b_c}{n})^{e_c}$. In the independent Poisson sampling case, we obtain conditional probability

$$\frac{\prod_c \Pr[X_{\frac{b_c \cdot s}{n}} = e_c]}{\Pr[X_s = s_0]} = \frac{\prod_c \exp(-\frac{b_c \cdot s}{n})}{\exp(-s)} \cdot \frac{\prod_c (\frac{b_c \cdot s}{n})^{e_c}}{s^{s_0}} \cdot \frac{s_0!}{\prod_c e_c!} .$$

Since the sum of the ball numbers b_c is exactly n, the first term of this product is one. Since the counts e_c sum to s_0, the occurrences of s in the second term cancel out, leaving exactly the desired multinomial probability.

(c) If a function f is invariant under permutations of the input colors, then applying a random permutation to the input colors before running a Poisson algorithm $\mathcal{A}$ will not change $\mathcal{A}$'s approximation guarantee. One can simulate the distribution of $\mathcal{A}$'s (permuted) input from the histogram of the sample, which counts the number of occurrences of different elements, by assigning a random color to each of the distinct elements, duplicating it a number of times according to the histogram, and randomly ordering the resulting set of colors. This produces exactly the same distribution that $\mathcal{A}$ would get if the names of the colors in the original input were permuted. □

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. We are also grateful to Salil Vadhan and anonymous referees for detailed comments that tremendously helped us with the writing of this work.

REFERENCES

[ABRS03] A. Akella, A. R. Bharambe, M. Reiter, and S. Seshan, *Detecting DDoS attacks on ISP networks*, in Proceedings of the ACM SIGMOD/PODS Workshop on Management and Processing of Data Streams, 2003.

[Akh65] N. I. Akhiezer, *The Classical Moment Problem and Some Related Questions in Analysis*, Hafner, New York, 1965.

[AMS99] N. Alon, Y. Matias, and M. Szegedy, *The space complexity of approximating the frequency moments*, J. Comput. System Sci., 58 (1999), pp. 137–147.

[And70] T. Ando, *Truncated moment problems for operators*, Acta Sci. Math. (Szeged), 31 (1970), pp. 319–334.

[Bar02] Z. Bar-Yossef, *The Complexity of Massive Data Set Computations*, Ph.D. thesis, Computer Science Division, University of California at Berkeley, Berkeley, CA, 2002.

[BJK$^+$02] Z. Bar-Yossef, T. S. Jayram, R. Kumar, D. Sivakumar, and L. Trevisan, *Counting distinct elements in a data stream*, in Proceedings of the 6th International Workshop on Randomization and Approximation Techniques, Springer-Verlag, London, 2002, pp. 1–10.

[BKS01] Z. Bar-Yossef, R. Kumar, and D. Sivakumar, *Sampling algorithms: Lower bounds and applications*, in Proceedings of the Thirty-Third Annual ACM Symposium on Theory of Computing, ACM Press, New York, 2001, pp. 266–275; full version available online at http://www.ee.technion.ac.il/people/zivby/papers/sampling/sampling_full.ps.

[BKS02] Z. Bar-Yossef, R. Kumar, and D. Sivakumar, *Reductions in streaming algorithms, with an application to counting triangles in graphs*, in Proceedings of the Thirteenth Annual ACM-SIAM Symposium on Discrete Algorithms, ACM, New York, SIAM, Philadelphia, 2002, pp. 623–632.

[BDKR02] T. Batu, S. Dasgupta, R. Kumar, and R. Rubinfeld, *The complexity of approximating entropy*, in Proceedings of the Thirty-Fourth Annual ACM Symposium on Theory of Computing, 2002, pp. 678–687.

[BDKR05] T. Batu, S. Dasgupta, R. Kumar, and R. Rubinfeld, *The complexity of approximating the entropy*, SIAM J. Comput., 35 (2005), pp. 132–150.

[BFF$^+$01] T. Batu, E. Fischer, L. Fortnow, R. Kumar, R. Rubinfeld, and P. White, *Testing random variables for independence and identity*, in Proceedings of the 42nd Annual IEEE Symposium on Foundations of Computer Science, 2001, pp. 442–451.

[BFR$^+$00] T. Batu, L. Fortnow, R. Rubinfeld, W. Smith, and P. White, *Testing that distributions are close*, in Proceedings of the 41st Annual IEEE Symposium on Foundations of Computer Science, 2000, pp. 259–269.

[BHR$^+$07] K. Beyer, P. J. Haas, B. Reinwald, Y. Sismanis, and R. Gemulla, *On synopses for distinct-value estimation under multiset operations*, in Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data, ACM, New York, 2007, pp. 199–210.

[Bun] J. Bunge, *Bibliography on Estimating the Number of Classes in a Population*, http://www.stat.cornell.edu/~bunge/bibliography.htm.

[CCMN00] M. Charikar, S. Chaudhuri, R. Motwani, and V. R. Narasayya, *Towards estimation error guarantees for distinct values*, in Proceedings of the Nineteenth ACM SIGMOD–SIGACT–SIGART Symposium on Principles of Database Systems, 2000, pp. 268–279.

[CF91] R. E. Curto and L. A. Fialkow, *Recursiveness, positivity, and truncated moment problems*, Houston J. Math., 17 (1991), pp. 603–635.

[FM85] P. Flajolet and G. N. Martin, *Probabilistic counting algorithms for data base applications*, J. Comput. System Sci., 31 (1985), pp. 182–209.

[GT02] P. B. Gibbons and S. Tirthapura, *Distributed streams algorithms for sliding windows*, in Proceedings of the Fourteenth Annual ACM Symposium on Parallel Algorithms and Architectures, 2002, pp. 63–72.

[IW03] P. Indyk and D. Woodruff, *Tight lower bounds for the distinct elements problem*, in Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science, 2003, pp. 283–288.

[Pro53] Yu. V. Prohorov, *Asymptotic behavior of the binomial distribution*, Uspekhi Mat. Nauk, 8 (1953), pp. 135–142 (in Russian).

[RRRS07] S. Raskhodnikova, D. Ron, R. Rubinfeld, and A. Smith, *Sublinear algorithms for approximating string compressibility*, in Proceedings of the 11th RANDOM, Springer-Verlag, Berlin, Heidelberg, 2007, pp. 609–623.

[RRSS07] S. Raskhodnikova, D. Ron, A. Shpilka, and A. Smith, *Strong lower bounds for approximating distribution support size and the distinct elements problem*, in Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science, IEEE Computer Society, Washington, DC, 2007, pp. 559–569.

[Szp01] W. Szpankowski, *Average Case Analysis of Algorithms on Sequences*, Wiley-Interscience, New York, 2001.

[Vad99] S. P. Vadhan, *A Study of Statistical Zero-Knowledge Proofs*, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, 1999.

[Val08] P. Valiant, *Testing symmetric properties of distributions*, in Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing, R. E. Ladner and C. Dwork, eds., ACM, New York, 2008, pp. 383–392.

[Was04] L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference*, Springer-Verlag, New York, 2004.

[Web99] 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 (1999), pp. 97–104.

[ZL77] J. Ziv and A. Lempel, *A universal algorithm for sequential data compression*, IEEE Trans. Inform. Theory, 23 (1977), pp. 337–343.