Sampling lower bounds: boolean average-case and permutations*

Emanuele Viola†

October 4, 2019

Abstract

We show that for every small AC^0 circuit $C : \{0,1\}^\ell \to \{0,1\}^m$ there exists a multiset S of $2^{m-m^{\Omega(1)}}$ restrictions that preserve the output distribution of C and moreover *polarize min-entropy:* the restriction of C to any $r \in S$ either is constant or has polynomial min-entropy. This structural result is then applied to exhibit an explicit boolean function $h : \{0,1\}^n \to \{0,1\}$ such that for every small AC^0 circuit $C : \{0,1\}^\ell \to \{0,1\}^{n+1}$ the output distribution of C for a uniform input has statistical distance exponentially close to $1/2$ from the distribution $(U, h(U))$ for U uniform in $\{0,1\}^n$. Previous such "sampling lower bounds" either gave exponentially small statistical distance or applied to functions h with large output length.

We also show that the output distribution of a d-local map $f : [n]^\ell \to [n]^n$ for a uniform input has statistical distance at least $1 - 2 \cdot \exp(-n/\log^{\exp(O(d))} n)$ from a uniform permutation of $[n]$. Here d-local means that each output symbol in $[n] = \{1, 2, \ldots, n\}$ depends only on d of the ℓ input symbols in $[n]$. This separates AC^0 sampling from local, because small AC^0 circuits can sample almost uniform permutations. As an application, we prove that any cell-probe data structure for storing permutations π of n elements such that $\pi(i)$ can be retrieved with d non-adaptive probes must use space $\geq \log_2 n! + n/\log^{\exp(O(d))} n$.

*This paper subsumes the unpublished manuscript [Vio17a].

†Supported by NSF CCF award 1813930.

1 Introduction, results, and discussion

A classical objective of complexity theory is to prove lower bounds on the resources required to compute a target function on a given, worst-case input. This goal has been achieved in several restricted models, and it has been extended to other notions such as average-case hardness. More recently, a series of papers [Vio12b, LV12, DW11, Vio14, Vio12c, BIL12, BCS14, Vio16] study computational lower bounds for *sampling tasks*, that is, for sampling a target distribution given uniform random bits. Proving lower bounds for sampling is more challenging than proving standard lower bounds. For example, even though small AC^0 circuits cannot compute the parity function, there exists a poly(n)-size AC^0 circuit (in fact, every output bit depends on just two input bits) which samples the distribution $(U, \mathrm{parity}(U))$, where U is uniform in $\{0,1\}^n$ [Bab87, Kil88]. Perhaps more surprisingly, polynomial-size AC^0 circuits can sample $(U, f(U))$ for every symmetric function f, such as Majority (up to an exponentially small error) [Vio12b]. They can even do this for some important non-symmetric functions f such as inner product [IN96]. These results, and others discussed below, have several applications to algorithms [MV91, Hag91], complexity theory [Bab87], and cryptography [Kil88, IN96, Vio05, BIVW16].

Despite recent progress, the study of "sampling lower bounds" remains largely uncharted. It will require us to come up with new proof techniques, which may be useful even for classical lower bounds. Moreover, the study has connections with other areas. For example, it had an impact on the recent breakthrough for two-source extractors. Specifically, in [Vio14] a new class of sources was introduced (some bits are k-wise independent, the others adversarially chosen), and an extractor was given for them (majority, analyzed using [DGJ$^+$10]), and finally it was asked if better extractors exist. Answering this question affirmatively is a main step in the construction of two-source extractors for polylogarithmic entropy by Chattopadhyay and Zuckerman [CZ16]. Follow-up work [Li16] gives better yet extractors for the same sources. The subsequent papers [CS16, Coh16, BDT16], leading to better and better two-source extractors, use instead the original extractor from [Vio14].

A different connection to data-structure lower bounds is discussed below.

In this paper we prove two new sampling lower bounds. We start with a result about AC^0. The paper [LV12] showed that the output distribution of a small AC^0 circuit is very far from the uniform distribution over an asymptotically good error-correcting code. Specifically the statistical distance is $\geq 1 - \epsilon$ for ϵ polynomially small. The exciting follow-up [BIL12] optimizes ϵ to exponentially small. However, their techniques are tailored to error-correcting codes, and do not apply to other distributions. In particular, they do not apply to distributions of the form $(U, f(U))$ where U is uniform in $\{0,1\}^n$ and f is boolean.

The paper [Vio14] does prove a sampling lower bound for distributions of the form $(U, f(U))$. However, the sampling lower bound only rules out sampling $(U, f(U))$ in AC^0 with exponentially small statistical distance. In this work we strengthen this latter result to an "average-case" lower bound. To illustrate, note that it is easy to sample $(U, f(U))$ with statistical distance $\leq 1/2$ (either $(U, 0)$ or $(U, 1)$ does the job). We give an explicit function h such that the statistical distance must be $1/2 - \epsilon$ for a small ϵ. In other words, a function that cannot be computed much better than random guessing even by a circuit that is allowed to sample the input. Such average-case lower bounds were known for *local samplers*, see [DW12] and cf. [Vio12b].

We first define and fix the notation for statistical distance, and then we state the theorem.

Definition 1. *The* statistical distance *between two distributions A and B on the same sample space is denoted* $\Delta(A, B) = \max_T |\mathbb{P}[A \in T] - \mathbb{P}[B \in T]| = \frac{1}{2}\sum_x |\mathbb{P}[A = x] - \mathbb{P}[B = x]|$.

Theorem 2. *There is $c > 0$ and a polynomial-time computable function $h : \{0,1\}^n \to \{0,1\}$ such that for any circuit $C : \{0,1\}^\ell \to \{0,1\}^{n+1}$ of depth d and size $\exp(n^{c/d})$ we have $\Delta(C(U), (U, h(U))) \geq 1/2 - 2^{-n^{\Omega(1)}}$.*

We write $\exp(x)$ for 2^x. The constants hidden in the $O(.)$ and $\Omega(.)$ notation are absolute. There is no restriction on ℓ, though obviously it is at most the size of the circuit C.

As in [Vio14, DW12] the function h can be taken to be an extractor for bit-block sources with polynomial min-entropy. Bit-block sources are a special case of affine sources whose definition is not needed until much later in this paper (see Definition 15). An explicit extractor for bit-block sources is given by Rao [Rao09]. This extractor is optimized in [Vio14, DW12]. We note that Li [Li16] extracts from general affine sources of polylogarithmic entropy, however his error is polynomial and so insufficient for Theorem 2. The paper [Vio16] also shows the existence of a quadratic polynomial p that extracts from bit-block sources. Hence there exists a quadratic polynomial h for which the above theorem holds, in contrast with the fact mentioned above that polynomial-size AC^0 circuits can sample (exactly) the distribution $(U, p(U))$ for the quadratic polynomial inner product (and in fact for any read-once polynomial).

Sampling permutations. We mentioned earlier that polynomial-size AC^0 circuits can sample $(U, f(U))$ for any symmetric boolean function f. This result relies on surprising algorithms by Matias and Vishkin [MV91] and Hagerup [Hag91] showing that poly(n)-size AC^0 circuits can generate a uniform random permutation of $[n] = \{1, 2, \ldots, n\}$, up to an exponentially small statistical error. (Their context is slightly different, for a streamlined presentation of the said result see [Vio12b].) They give several algorithmic applications of this result, and more applications have been found since then, including constructing efficient secret-sharing schemes [BIVW16].

Another line of works studies generating random permutations using *switching networks*. A recent paper by Czumaj [Czu15] gives an explicit construction of switching networks with depth $O(\log^2 n)$ and $O(n \log n)$ switches that generate a nearly-uniform permutation on n elements, improving on previous work (see [Czu15] for discussion). The paper also conjectures that the depth can be improved to $O(\log n)$, and proves a partial result in this direction.

A switching network of depth d gives rise to a sampler where each output bit is a depth-d decision tree (see Lemma 6.4 in [Vio12b]). Hence [Czu15] shows that decision trees of depth $O(\log^2 n)$ can generate permutations fairly well, and conjectures the same for depth $O(\log n)$.

On the side of lower bounds apparently nothing was known, and the above algorithms and conjectures arguably explain the difficulty of proving negative results. In this paper we prove a lower bound in the cell-probe model, with the restriction that all probes are non-adaptive. Specifically, we divide the memory into ℓ cells of $\log n$ bits (all logarithms in this paper are in base 2 unless otherwise noted), which are initialized uniformly at random. We consider algorithms that output n cells representing a function from $[n]$ to $[n]$ in the natural way. Each output cell only depends on a small number d of input cells. Again, there is no restriction on the number ℓ of input cells the algorithm may use, though $\ell \leq dn$ without loss of generality.

Theorem 3. *Let $f : [n]^\ell \to [n]^n$ be a d-local map, i.e., a map such that each output symbol in $[n]$ depends only on d input symbols in $[n]$. Let $\Pi \in [n]^n$ be a random permutation of n elements. Let $f(U)$ be the output distribution of f for a uniformly chosen U in $[n]^\ell$. Then $\Delta(f(U), \Pi) \geq 1 - 2 \cdot \exp(-n/\log^{c^d} n)$, where c is a universal constant.*

Theorem 3 remains nontrivial for locality up to $d = \epsilon \log\log n$ for a small enough constant ϵ. (The factor 2 in the conclusion makes the bound trivially true if d is larger.) Note that the 1-local identity map $f(x_1, x_2, \ldots, x_n) = (x_1, x_2, \ldots, x_n)$, where $\ell = n$, achieves statistical distance $1 - \exp(-O(n))$, so for small locality the statistical bound in Theorem 3 is not far from optimal.

Previously, lower bounds with statistical distance approaching 1 exponentially fast were only known for the problem mentioned earlier of sampling error-correcting codes [BIL12]. These lower bounds applied to AC^0 samplers. For distributions that can be sampled in AC^0 the previous sampling lower bounds were much weaker [Vio12b]. Thus, this work gives a new separation between the sampling power of AC^0 and small-locality maps.

Another benefit of obtaining such a large statistical-distance lower bound is that it enables an application discussed next.

Data structures. The work [Vio12b] shows that sampling lower bounds with large statistical distance such as in Theorem 3 imply lower bounds for *succinct data structures*. (In fact, as will be evident from the proof of Corollary 4 below, for this implication a sampling lower bound in the special case when ℓ is small would suffice.) Some data-structure lower bounds proved this way appear in [Vio12b, LV12, BIL12], however they are either very weak or concern unnatural data structure problems.

Theorem 3 implies a data-structure lower bound for the problem of storing a permutation $\pi : [n] \to [n]$ so that $\pi(i)$ can be retrieved fast. At one extreme one can use $n \log_2 n$ bits to store the permutation and answer each query $\pi(i)$ by reading just one cell of $\log_2 n$ bits, at the other extreme we can use the information-theoretic minimum amount $\lceil \log_2 n! \rceil = n \log_2 n - \Omega(n)$ of space and answer queries by reading the entire memory. The goal of *succinct data structures* [MRRR12, GM07, GGG+07, GRR08, Pǎt08, Vio12a, Gol09, DPT10, Vio09a, PV10, Pre18] is to understand what is the right tradeoff between the time it takes to answer a query and the *redundancy* of the data structure, the amount of extra space used over the information-theoretic minimum. As a corollary to Theorem 3 we obtain the following tradeoff.

Corollary 4. *Consider any cell-probe data structure for storing permutations π of n elements such that $\pi(i)$ can be retrieved with d non-adaptive probes in cells of $\log_2 n$ bits. The data structure must use space $\geq \log_2 n! + n/\log^{\exp(O(d))} n$ bits.*

We repeat the simple proof from [Vio12b].

Proof. Suppose a data structure exists with redundancy r. Consider filling its $\lceil \log n! \rceil + r$ memory bits uniformly at random. With probability $\geq 2^{-r}$, the memory will be uniform over encodings of permutations. Hence if we run the data structure algorithm on uniform memory we obtain a sampler with statistical distance $< 1 - 2^{-r}$. The result then follows from Theorem 3. □

In particular, for constant time $d = O(1)$ the redundancy is $r \geq n/\text{poly}\log n$. By contrast, for other important problems there are surprising data structures that achieve $d = O(1)$ and $r = O(1)$, and are also non-adaptive [Păt08, DPT10, Pre18] (for an exposition of the relevant result in [DPT10] see [Vio09b], Lectures 23-24). Corollary 4 shows that such amazing data structures do not exist for storing permutations.

Related work and discussion. Previous work has studied the problem of storing π so that $\pi(i)$ *and both* $\pi^{-1}(i)$ can be retrieved fast. [MRRR11] give several data structures for this problem. In particular, they give a data structure that can store a permutation using $\log_2 n! + n/\log^{2-o(1)} n$ bits such that $\pi(i)$ (and both $\pi^{-1}(i)$) can be computed in time $O(\log n)/\log\log n$. This data structure is based on a switching network known as the Benes network. They achieve their saving by "brute-forcing" certain small components.

On the side of lower bounds, Golynski shows in [Gol09, Theorem 4.1] that any cell-probe data structure for representing a permutation $\pi : [n] \to [n]$ so that π can be computed with t cell probes and π^{-1} with t' must use $\log_2 n! + \Omega(n)/(t \cdot t')$ as long as $\max\{t, t'\} \leq c \log^2(n) \log\log n$. This bound essentially matches the data structure in [MRRR11] for $t = \log n$, but tight bounds are not known in other parameter regimes. His technique is unlikely to apply to our simpler problem where we do not have the inverse queries $\pi^{-1}(i)$. In fact, none of the available techniques seems directly applicable for this problem: essentially, the only other technique available is the one in [PV10]. That technique requires that the mutual information between two sets of t queries is $\Omega(t)$, but a calculation shows that in the case of permutations the mutual information is at most $O(t(t/n))$.

However, we thank an anonymous referee for pointing out that, in the case of non-adaptive probes, the technique in [PV10] can be modified to obtain the same result in Corollary 4, without mentioning sampling.

In this paper the data-structure lower bound is obtained as a consequence of a stronger result: a sampling lower bound. We hope that this sampling approach can be useful to attack some of the long-standing open problems on data structures. Two central open problems are improving Siegel's state-of-the-art 1989 lower bound (Theorem 3.1 in [Sie04], rediscovered in [Lar12]; see [Vio17b, Lecture 18] for an exposition), or proving lower bounds for the succinct dictionary problem (Problem 5 in Pătraşcu's obituary [Tho13]). We emphasize that both these problems are also open for non-adaptive probes, and, as also mentioned earlier, some of the best-known data structures are non-adaptive. (The situation is entirely different for *dynamic* data structures, see [BL15].) On the other hand, the succinct data structure for permutations in [MRRR11] does use adaptivity to follow a path in a switching network.

1.1 Techniques for Theorem 2

The starting point for our proof of Theorem 2 is a previous theorem [Vio14], already mentioned above, giving an explicit boolean function $h : \{0,1\}^n \to \{0,1\}$ such that the distribution $(U, h(U))$ cannot be sampled in AC^0. The function h in the proof of this result is an extractor for bit-block sources *with polynomial min-entropy*, where the min-entropy of a distribution D is $\min_a \log_2(1/\mathbb{P}[D = a])$. Note that if a circuit samples exactly $(U, h(U))$ then its output distribution has the same min-entropy of $(U, h(U))$, which is n, and the extractor is useful.

However, this argument fails to give average-case sampling lower bounds: it only rules out exponentially small statistical distance. The problem is that it's possible that distributions X and T over m bits are statistically close yet have antipodal min-entropies. For example, it can be the case that X has min-entropy 1 while T has min-entropy m, and yet $\Delta(X, T) \le 1/2$: just take T to be uniform over m bits and X which is 0^m with probability $1/2$ and uniform otherwise.

Polarizing min-entropy. Computationally, the distribution X above can be sampled by taking the bit-wise And of the uniform distribution with a single input bit b: $X = (U_1 \wedge b, U_2 \wedge b, \ldots, U_n \wedge b)$ for $U = (U_1, U_2, \ldots, U_n)$ uniform in $\{0,1\}^n$ and b uniform in $\{0,1\}$. However, an important observation is that the min-entropy of X can be "polarized" via restrictions. Specifically, if b is fixed to 0 then X is also fixed and so has min-entropy zero, whereas if b is fixed to 1 then X is uniform and so has min-entropy m.

A main theorem of this work shows how to polarize the min-entropy of any AC^0 distribution using restrictions: for every small AC^0 circuit $C : \{0,1\}^\ell \to \{0,1\}^m$ there exists a small multiset S of restrictions that preserves the output distribution of C and yet the restricted circuit always has either zero or polynomial min-entropy. This result would be useless if we allow ourselves 2^m restrictions, and a critical feature of our proof is that it guarantees a much smaller number of restrictions.

We first formally define restrictions and fix notation that is used throughout, then state the theorem. Restrictions have been a main tool in complexity theory since at least [Sub61].

Definition 5. A restriction r over ℓ bits is a string in $\{\star, 0, 1\}^\ell$. We denote by $r(U)$ the distribution over $\{0,1\}^\ell$ obtained by replacing the stars $\star$ in r with uniform bits. For a multi-set S of restrictions we denote by $S(U)$ the distribution obtained by picking a uniform restriction $r \in S$ and outputting $r(U)$. For a function $f : \{0,1\}^\ell \to \{0,1\}^m$ we denote by f_r the function f restricted to r.

Theorem 6. *[Polarizing min-entropy] There is $c > 0$ such that the following holds:*

Let $C : \{0,1\}^\ell \to \{0,1\}^m$ be a depth-d circuit of size $\le \exp(m^{c/d})$. There exists a multiset S of $\le 2^{m - m^{\Omega(1)}}$ restrictions such that:

(1) $\Delta(C(S(U)), C(U)) \le 2^{-m^{\Omega(1)}}$*, and*
(2) for every $r \in S$*, either* C_r *is constant or has min-entropy* $\ge m^{0.9}$*.*
Moreover, for every $r \in S$ *we have that* C_r *is a depth-*t *forest for* $t = O(1)$*.*

The "moreover" statement slightly simplifies some later arguments, but is not essential for now.

It is natural to ask if every distribution satisfies this decomposition. A distribution that does not satisfy it can be sampled as follows: pick m random bits. If their parity is 1, output them. Otherwise output 0^m. It does not satisfy the decomposition because for every restriction of the input bits that leaves some variable unset, the output distribution still has min-entropy 1 (which is neither zero nor polynomial). So one would need restrictions that leave no variables unset. However then one needs $\ge \Omega(2^m)$ restrictions to be close in statistical distance.

Let us sketch how from Theorem 6 one proves the average-case lower bound in Theorem 2. Reasoning similarly to arguments in [Vio12b], the test that witnesses the large statistical distance is the union of two tests. The first contains the set of $\le 2^{m-m^{\Omega(1)}}$ possible outputs corresponding to restrictions under which the circuit is constant. The second contains the set of strings (x, b) where $b \neq h(x)$. The target distribution $(U, h(U))$ never passes the second test, and only rarely passes the first. Instead, the (distribution sampled by the) circuit passes the union of the two tests with probability at least about $1/2$. This is because if after applying the restriction the circuit is constant then it passes the first test. If it is not, then it has large min-entropy. At this point one can further restrict the input to fix the output bit corresponding to h, without significantly changing the min-entropy. Because h is an extractor for such sources, the value of h should be nearly uniform. But since it's fixed, the restricted circuit passes the second test with probability close to $1/2$.

Theorem 6 is proved in two steps. In the first step we give a small set of restrictions that preserves the output distribution of the AC^0 circuit and also collapses it to a shallow decision forest. In the second we further restrict the decision forest to either fix it or argue that its output distribution has large min-entropy.

1.2 Techniques for Theorem 3

Theorem 3 is proved by induction on the locality d. Consider a d-local map f and write $f = (f_1, f_2, \ldots, f_n)$ where f_i is the function outputting the cell i. In the induction step, we start with a relatively standard *covering argument.* That says that either we have (A) a small number of input cells C that intersect the probes made by all the f_i (in other words, every f_i probes a cell in C), or else (B) we have many f_i whose set of probes are disjoint.

In case (A), suppose we fix the contents of the cells C. Because every f_i probes a cell in C, this reduces the locality of f. Thus, we can write our sampler f as a convex combination of samplers with smaller locality, one for each possible fixing of the contents of the cells in C. To analyze this step we show (Corollary 18) that if D is a distribution that is a convex combination of 2^s distributions D_i (the samplers obtained by any possible fixing of the $s = |C| \log n$ bits in the cells C) where each D_i has statistical distance $\ge 1 - \epsilon$ from a target distribution T, then D has distance $\geq 1 - 2^s\epsilon$ from T. By setting the parameters appropriately, we can ensure that $\epsilon \ll 1/2^s$, concluding this case.

In case (B), we have many f_i which are independent. We obtain large statistical distance just considering these independent f_i. The high-level idea is that if the f_i have small entropy, then the result follows because a uniform permutation has large entropy. Otherwise, if the f_i have high entropy we can show by the *birthday paradox* that the outputs of the f_i will collide (i.e., $f_i = f_j$ for some $i \neq j$) with high probability. Since this never happens for permutations, we obtain statistical distance.

Formalizing case (B) requires finding the right notion of "high-entropy." If we have t independent f_i, we define one f_i to be "high-entropy" if for every set S of $t/2$ values, the probability that $f_i \in S$ is $\Omega(|S|/n)$. Now, if there are $t/2$ functions f_i that have high-entropy, then we can run a folklore, simplified proof of the birthday paradox: fix the other $t/2$ functions arbitrarily, and define S to be the set of values they take. By high-entropy and independence, the probability of not having a collision will be

$$(1 - \Omega(|S|/n))^{t/2} \leq e^{-\Omega(t^2/n)}$$

which is small enough when $t = n^{0.5+\Omega(1)}$. Since a uniform permutation by definition never has a collision, we obtain statistical distance $1 - e^{-\Omega(t^2/n)}$.

If, on the other hand, we have $t/2$ functions which are low entropy, we use concentration of measure to show that they will land in their corresponding sets S not often enough. Here again we obtain a statistical distance $1 - e^{-\Omega(t^2/n)}$.

We shall start with $t = \Omega(n)$ for $d = 0$, and then progressively update it via $t \to t^2/n$ from the above abounds. Losing along the way $\log n$ factors that arise from having cells of $\log n$ bits, this gives the bound in Theorem 3.

Organization. Theorems 6 and 2 are proved in Sections 2 and 3. Specifically, in Section 2 we make the first step towards proving Theorem 6 by exhibiting a small set of restrictions that preserves the distribution of and collapses a given AC^0 circuit to a small-depth forest. Then in Section 3 we show that a small-depth forest can be restricted so that it is either constant or has high min-entropy. Combining these results gives Theorem 6. Theorem 2 follows as a corollary.

Theorem 3 is proved in Section 4. We conclude in Section 5 by discussing a number of open problems.

2 Polarizing min-entropy: from AC^0 to decision trees

In this section we take the first step towards proving Theorem 6 by exhibiting for any given AC^0 circuit a small multiset of restrictions that preserves the output distribution and collapses the circuit to a shallow decision forest. We call $f : \{0,1\}^\ell \to \{0,1\}^m$ a *depth-t forest* if every output bit of f is a decision tree of depth t.

Theorem 7. *There is $c > 0$ such that the following holds:*

Let $C : \{0,1\}^\ell \to \{0,1\}^m$ be a circuit of depth d and size s. For any real ϵ, integers t, v, and $p := 1/(m^{1/t} \cdot O(\log s)^{d-1})$ such that $pm \geq c \log m$ there exists a multiset S of restrictions such that:

(1) $\Delta\left(C(U), C(S(U))\right) \leq \epsilon$,

(2) $|S| = (1/\epsilon)^{O(1)} \cdot 2^{m(1-\Omega(p))+v}$,

(3) for all but a $2s2^{-v}$ fraction of the restrictions $r \in S$, *the circuit C_r is a depth-t forest.*

For the proof we shall use the following concentration inequality which is Theorem 7 in [CL06].

Lemma 8. *Let $X_1, X_2, \ldots, X_n$ be non-negative, independent random variables, and let $X = \sum_{i=1}^n X_i$. It holds that*

$$\mathbb{P}[X \leq \mathbb{E}[X] - \lambda] \leq e^{-\lambda^2/2\sum_{i=1}^n \mathbb{E}[X_i^2]}.$$

We note that the corresponding bound for the upper tail (Theorem 6 in [CL06] – also known as Bernstein's inequality) requires an extra term in the exponent which makes it useless for our application. However, we shall be able to get by using only an estimate for the lower tail.

We use the following standard distribution on restrictions.

Definition 9. We denote by R_p^ℓ the distribution on restrictions over ℓ bits where the probabilities of $\star, 0, 1$ are $p, (1-p)/2, (1-p)/2$, and the symbols are independent.

We shall also use a version of boolean hypercontractivity, cf. [O'D14]. To illustrate, consider the "restriction experiment" where we sample a restriction from R_p^ℓ and then we replace the stars with two independent uniform choices, to obtain two strings y and y' in $\{0,1\}^\ell$. We are interested in the probability that both y and y' land in the same set $A \subseteq \{0,1\}^\ell$ of density α. If y and y' were uniform and independent in $\{0,1\}^\ell$, the probability would be α^2, whereas if it was the case that $y = y'$ always then this probability would just be α. The restriction experiment is somewhere in the middle: y and y' have some common and some independent parts. With hypercontractivity we can show that in that case the probability is smaller than α, depending on the parameter p of the restriction:

Lemma 10. *Let $A \subseteq \{0,1\}^\ell$ be a set of density α. Then*

$$\mathbb{P}_{R \in R_p^\ell, U, U'}[R(U) \in A \wedge R(U') \in A] \leq \alpha^{1+\Omega(p)}.$$

Starting with [LV12], where a proof of the lemma can be found, this result has been used several times in the study of the complexity of distributions [Vio14, BIL12].

The key new lemma in this section is the following, stating that for every function $f : \{0,1\}^\ell \to \{0,1\}^m$ we can find a multi-set S of much fewer than 2^m restrictions such that $f(S(U))$ is close to $f(U)$.

Lemma 11. *There is $c > 0$ such that the following holds:*

Let $f : \{0,1\}^\ell \to \{0,1\}^m$ be a function. Suppose $pm \geq c \log m$. Let S be a multiset of $s = (1/\epsilon)^c 2^{m(1-p/c)}$ restrictions sampled independently from R_p^ℓ. Then

$$\mathbb{P}_S[\Delta\left(f(S(U)), f(U)\right) \geq \epsilon] < 1/2.$$

Proof. For $i \in \{0,1\}^m$ let $A'_i \subseteq \{0,1\}^\ell$ be $f^{-1}(i)$. Further partition each A'_i into sets $A_{i,j}$ of measure $|A_{i,j}|/2^\ell = \alpha := \epsilon 2^{-m}$ and set for $A_{i,0}$ of measure $< \alpha$. The total number of sets $A_{i,j}$ is $\leq 1/\alpha + 2^m \leq 2/\alpha$. We shall show:

(#) with probability at least 1/2 over S, for every i and every $j > 0$, $\mathbb{P}[S(U) \in A_{i,j}] \geq (1-\epsilon)\alpha$.

The latter probability is for a fixed S but over the choice of a uniform $r \in S$ and U, cf. Definition 5. Claim (#) guarantees that the total error from those sets is at most ϵ. And the sets $A_{i,0}$ will contribute at most ϵ to the error.

Formally, fix a set $A_{i,j}$ with $j > 0$ and call it A. Let $S = \{R_1, R_2, \ldots, R_s\}$ be the set of random restrictions. For every $k \leq s$ consider the function X_k of the random variable R_k defined as $X_k := \mathbb{P}_U[R_k(U) \in A]$. We have

$$\mathbb{E}[X_k] = \mathbb{E}_{R_k}[\mathbb{P}_U[R_k(U) \in A]] = \mathbb{P}_{R_k,U}[R_k(U) \in A] = \mathbb{P}_U[U \in A] = \alpha. \tag{1}$$

For a fixed choice of S we have

$$\mathbb{P}[S(U) \in A] = \frac{1}{s}\sum_{k=1}^{s} X_k.$$

Our goal is to show that with high probability over S this quantity is very close to its expectation, which by Equation 1 is α. The way in which we use restrictions is that they *reduce the second moment of the* X_k. This allows us to drive the error below 2^{-m} using fewer than 2^m samples.

For any k we have

$$\mathbb{E}[X_k^2] = \mathbb{E}_{R_k}[\mathbb{P}_{U,U'}[R_k(U) \in A \wedge R_k(U') \in A]] \leq \alpha^{1+\Omega(p)}$$

where the inequality is Lemma 10.

Now we can apply Lemma 8 with $\lambda = s\epsilon\alpha$ to get

$$\begin{aligned}
\mathbb{P}_S[\frac{1}{s}\sum_{k \leq s} X_k \leq \alpha - \epsilon\alpha] &\leq \exp(-\frac{s^2\epsilon^2\alpha^2}{2s\alpha^{1+\Omega(p)}}) \\
&= \exp(-s\epsilon^{O(1)} \cdot 2^{-m(1-\Omega(p))}) \\
&= \exp(-(1/\epsilon)^{c-O(1)} \cdot 2^{-mp/c+\Omega(pm)}) \\
&\leq \exp(-(1/\epsilon) \cdot 2^{\Omega(pm)}),
\end{aligned}$$

where we use the definitions of $\alpha = \epsilon/2^m$ and s, and that c is large enough. The number of sets $A_{i,j}$ with $j > 0$ is $\leq 1/\alpha = 2^m/\epsilon$ as noted earlier. Hence by a union bound the probability over S that there exists a set $A = A_{i,j}$ with $j > 0$ for which $\frac{1}{s}\sum_{k \leq s} X_k \leq \alpha - \epsilon\alpha$ is at most

$$(2^m/\epsilon) \cdot \exp(-(1/\epsilon) \cdot 2^{\Omega(pm)}) < 1/2$$

where the inequality relies on the assumption that $pm \geq c \log m$. This proves (#). Finally, we claim that for any fixed S as in (#) the statistical distance $\Delta(f(S(U)), f(U))$ is at most 2ϵ. This can be verified as follows.

$$
\begin{aligned}
&\Delta(f(S(U)), f(U)) \\
=&\frac{1}{2}\sum_{i\in\{0,1\}^m} \left||A_i|2^{-\ell} - \mathbb{P}[S(U)\in A_i]\right| \\
\leq&\frac{1}{2}\sum_{i,j} \left||A_{i,j}|2^{-\ell} - \mathbb{P}[S(U)\in A_{i,j}]\right| \text{ (by the triangle inequality)} \\
=&\sum_{i,j:|A_{i,j}|2^{-\ell}>\mathbb{P}[S(U)\in A_{i,j}]} \left(|A_{i,j}|2^{-\ell} - \mathbb{P}[S(U)\in A_{i,j}]\right) \\
=&\sum_{i,j>0:\alpha>\mathbb{P}[S(U)\in A_{i,j}]} (\alpha - \mathbb{P}[S(U)\in A_{i,j}]) + \sum_{i:|A_{i,0}|2^{-\ell}>\mathbb{P}[S(U)\in A_{i,0}]} (|A_{i,0}|2^{-\ell} - \mathbb{P}[S(U)\in A_{i,0}]) \\
\leq&(1/\alpha)(\epsilon\alpha) + \sum_i |A_{i,0}|2^{-\ell} \\
\leq&(1/\alpha)(\epsilon\alpha) + 2^m\cdot\alpha \\
=&2\epsilon.
\end{aligned}
$$

□

Next we bound the probability that the circuit collapses, using a switching lemma [Ajt83, FSS84, Yao85, Hås87, Hås14, IMP12, PRST16, Ros17, BKST18]. To get exponentially small error we use the version in the latter paper, stated next after a definition.

Definition 12. A function $f : \{0,1\}^\ell \to \{0,1\}^m$ is computable by a *v-common forest of depth t* if there is a decision tree of depth v such that on every input, the function f restricted along a path of this tree is a depth-t forest.

Lemma 13. *[Lemma 14 in [BKST18]] Let $f : \{0,1\}^\ell \to \{0,1\}^m$ be an AC^0 circuit of size s and depth d. Let $p := 1/(m^{1/t}\cdot O(\log s)^{d-1})$ and sample r from R_p^ℓ. Then except with probability $s2^{-v}$ the restricted function f_r is a $O(v)$-common depth-t forest.*

Combining the above two lemmas we can prove Theorem 7.

Proof. (Theorem 7.) First consider the set S' obtained by picking $(1/\epsilon)^{O(1)}2^{m(1-\Omega(p))}$ samples from R_p^ℓ. Each restriction does not collapse the circuit in the sense of Lemma 13 with probability $\leq s2^{-v}$. By Markov inequality, the probability that more than a $2s2^{-v}$ fraction of restrictions does not collapse the circuit is at most 1/2. Combining this with Lemma 11 and a union bound we obtain that there exists a set S' such that

(1') $\Delta\left(C(U), C(S'(U))\right) \leq \epsilon$,

(2') $|S'| = (1/\epsilon)^{O(1)}\cdot 2^{m(1-\Omega(p))}$,

(3') *for all but a $2s2^{-v}$ fraction of the restrictions* $r\in S'$, the circuit C_r is a v-common depth-t forest.

Now construct S as follows. For each $r \in S'$, put in S the 2^v restrictions obtained by further restricting the common part in any possible way. The theorem follows. □

3 Polarizing min-entropy of decision trees, and putting things together

In this section we conclude the proof of Theorem 6, and then use it to prove Theorem 2. First, we prove a result about decision forests. We show that they can be restricted so that they are either fixed or sample a distribution with high min-entropy.

Theorem 14. *Let $f : \{0,1\}^\ell \to \{0,1\}^m$ be a depth-t forest. Let $b := m/2^t$. There exists a set S of $s = 2^{b+bt}$ restrictions such that:*
(1) $f(S(U))$ and $f(U)$ have the same distribution, and
(2) for every $r \in S$, either f_r is constant or f_r has min-entropy $\geq m/2^{O(t)}$.

A similar result in the special case of local sources is implicit in [Vio12b, DW12]. Our dependence on t is exponentially better. Also note that (1) above gives statistical distance zero, while (1) in Theorem 6 for AC^0 only gives that the distance is small. And Theorem 14 gives a *set* (as opposed to multiset) of restrictions. But in the main Theorem 6 we only claimed a multiset because conceivably one can obtain the same restriction in different ways by applying first Theorem 7 and then Theorem 14.

Proof. We say that a decision tree *probes* an input variable if the variable appears in the tree. Note that the number of input variables that are probed by more than 2^{2t} trees is $< m2^t/2^{2t} = b$.

We begin by restricting those b bits arbitrarily. This gives a set of 2^b restrictions. For any such restriction z consider f_z and reason as follows. If f_z has at most b output bits that are not fixed, then we further restrict f_z to make it a constant. The number of such restrictions is $\leq 2^{b\cdot t}$. If we lump together all these restrictions we arrive at a final multiset of restrictions of size 2^{b+bt}, as desired.

There remains to address the functions f_z which have more than b output bits that are not fixed. We claim that in this case the output distribution of f_z has high min-entropy. Here is where we use that f and hence f_z has depth t (as opposed to just being 2^t local). Pick any string $a \in \{0,1\}^m$. We shall show that $\mathbb{P}[f_z(U) = a]$ is small. Consider any output bit that is not constant. Since it corresponds to a non-constant tree of depth t, it takes the corresponding value of a with probability at most $1 - 2^{-t}$. Sample a uniform path in that tree by sampling the corresponding $\leq t$ variables. Now find another non-constant tree that does not use any of the sampled variables, and repeat. Because each input variable is probed by $\leq 2^{2t}$ trees, each repetition changes the output probability of at most $t \cdot 2^{2t}$ trees. We started with $\geq b = m/2^t$ trees. This means we can repeat this process i times as long as

$$i \cdot t \cdot 2^{2t} < m/2^t$$

for which it suffices that $i < m/2^{O(t)}$. Hence

$$\mathbb{P}[f_z(U) = a] \le (1 - 2^{-t})^{m/2^{O(t)}} \le e^{-m/2^{O(t)}}$$

and the min-entropy of f_z is at least $m/2^{O(t)}$. □

Next we put things together and prove our structural result about AC^0 distributions.

Proof. (Theorem 6) Let s be the size of the circuit. We start by applying Theorem 7 with t a sufficiently large constant to be determined later, and $v = m^{c'}$ and $\epsilon = 2^{-v}$ for a constant c' to be set later.

In the obtained set S, replace every restriction that does not collapse the circuit with the all zero string. This only adds $2s2^{-v} \le \sqrt{\epsilon}$ to the statistical distance, where the inequality holds for c small enough so that $2s \le 2^{v/2}$.

The number of restrictions is $|S| = 2^{m(1-\Omega(p))+O(v)} = 2^{m(1-\Omega(p))}$. In the last step we used that $mp = m^{1-1/t}/O(\log s)^{d-1} \ge m^{1-1/t-c}$ and we pick $c' < 1 - 1/t - c$.

Now for every $r \in S$ apply Theorem 14 to C_r. Here comes a change in parameters. We set the value of t in the statement of Theorem 14 to $t' = c'' \log m$ for a constant c'' to be set later, even though C_r is a t-forest for $t = O(1)$. (The proof of that case would not be simpler.) We can do this as long as $c'' \log m \ge t$, which is true for any constants c'' and t for sufficiently large m. That gives, for $b = m/2^{t'}$, a set S_r of $2^{b+bt'} \le 2^{2bt'} = 2^{2\cdot(m^{1-c''})\cdot c'' \log m}$ restrictions such that for every $r' \in S_r$ either $(C_r)_{r'}$ is a constant or has entropy $\ge m/2^{O(t')} = m/m^{O(c'')} \ge m^{0.9}$ where the last inequality holds for small enough c''.

The total number of restrictions, including S and each S_r is

$$\le |S| \cdot 2^{2bt'} \le 2^{m - \Omega(m^{1-1/t-c}) + 2\cdot(m^{1-c''})\cdot c'' \log m}.$$

Picking $1/t + c$ smaller than c'' concludes the proof.

To summarize the constraints on the constants, we need that c and c'' are small enough, that $c' < 1 - 1/t - c$, and that $1/t + c < c''$. □

We can now prove Theorem 2. We rely on a decomposition from [Vio14].

Definition 15. [Bit-block sources] A *bit-block* source over m bits is specified by an m-tuple where each coordinate can be a value in $\{0, 1\}$, a variable X_i, or the complement of a variable X_i. The block-size of the source is the maximum number of occurrences of any single variable. A sample from a bit-block source is obtained by sampling the variables X_i uniformly at random from $\{0, 1\}$.

For example, $(0, X_5, X_5, 1 - X_5, X_3, 1)$ is a bit-block source on 6 bits with entropy 2 and block size 3.

Lemma 16. *[Vio14, Theorem 1.6] Let $f : \{0,1\}^\ell \to \{0,1\}^m$ be a depth-t forest such that $f(U)$ has min-entropy $\ge k$. Then for an $s = \tilde{\Omega}(k^3/(m^2 2^{3t}))$ we have that $f(U)$ is 2^{-s} close to a convex combination of bit-block sources with min-entropy s and block-size $2 \cdot 2^t m/k$.*

The notation $\tilde{\Omega}$ hides polylogarithmic factors in the argument. ([Vio14, Theorem 1.6] mentions *local* sources; we use the simple fact that a depth-t decision tree has locality 2^t.)

Proof. [Theorem 2] Let $h : \{0,1\}^n \to \{0,1\}$ be an extractor for bit-block sources of entropy $n^{0.1}$ and block-size $n^{0.9}$ with error $2^{-n^{\Omega(1)}}$. This is a function h that on any such source Z satisfies $|\Pr[h(Z) = 1] - 1/2| \le 2^{-n^{\Omega(1)}}$. An explicit such extractor is given by Rao [Rao09] and subsequent optimizations [Vio14, DW12].

Let S be the multiset of restrictions given by Theorem 6. By Property (1) in that theorem it suffices to show that the distribution $C(S(U))$ (rather than $C(U)$) is far from $(U, h(U))$ (this switch only incurs an exponentially small error).

Let T be the set of size $\le |S|$ of values C_r for each $r \in S$ such that C_r is constant. The statistical distance is witnessed by the test $V := T \bigcup \{(x, 1 - h(x)) : x \in \{0,1\}^n\}$.

First note that $\mathbb{P}[(U, h(U)) \in V] = \Pr[(U, h(U)) \in T] \le |T| \cdot 2^{-n} = 2^{m - m^{\Omega(1)} - n} = 2^{-m^{\Omega(1)}}$, using that $m = n + 1$.

On the other hand we shall show that $C(S(U)) \in V$ with probability at least about 1/2. Fix any $r \in S$. If C_r is constant then $\mathbb{P}[C_r(U) \in T] = 1$ and we are done. Otherwise, by Property (2) in Theorem 6, C_r has min-entropy $\ge m^{0.9}$, and is a depth-t forest for $t = O(1)$. We now show that $C_r(U)$ lands in V with probability $\ge 1/2 - 2^{-n^{\Omega(1)}}$. In fact, this is going to be true even conditioned on the value of the output bit corresponding to h. Specifically, let f be the depth-t tree in the forest C_r corresponding to the value of h. Sample a uniform path in f. This fixes the bit corresponding to h but only reduces the min-entropy of C_r by $\le O(1)$. Call the resulting forest $C'(U)$. (Note that C' also has the restriction r hardwired into it.)

By Lemma 16, $C'(U)$ is in turn 2^{-s}-close to a convex combination of bit-block sources with entropy s and block-size $\le w$, for $s = \tilde{\Omega}((m^{0.9})^3/(m^2 \cdot (m^{0.1})^3)) \ge \tilde{\Omega}(m^{0.4})$ and $w \le 2m^{0.1}m/m^{0.9} \le O(m^{0.2})$. Hence by the property of the extractor the value of h on the source will be a bit at statistical distance $\le 2^{-n^{\Omega(1)}} + 2^{-s} \le 2^{-n^{\Omega(1)}}$ from uniform. However, the corresponding bit in $C'(U)$ is fixed. Hence $C'(U)$ will land in $\{(x, 1 - h(x)) : x \in \{0,1\}^n\}$ and hence in V with probability $\ge 1/2 - 2^{-n^{\Omega(1)}}$, concluding the proof. □

4 Proof of Theorem 3

In this section we prove Theorem 3. First, in Section 4.1 we show that a convex combination of distributions that are distant from a target distribution remains distant. Then in Section 4.2 we show that any collection of independent random variables is distant from uniform variables conditioned on not colliding. Finally, in Section 4.3 we use these results to prove Theorem 3.

4.1 Combination of far distributions is far

We start with a lemma about two distributions and then we obtain our main result as a corollary.

Lemma 17. *Let p and q and t be distributions over the same domain. Let $r = \frac{1}{2}(p+q)$ be a convex combination of p and q. If $\Delta(p,t) \geq 1-\epsilon$ and $\Delta(q,t) \geq 1-\epsilon$ then $\Delta(r,t) \geq 1-2\epsilon$. Moreover, there exist distributions for which the conclusion is $\Delta(r,t) = 1-2\epsilon$.*

We thank the anonymous referee who suggested the following proof which improves the constant in our original one.

Proof. We have

$$\begin{aligned} 1-\epsilon \leq \Delta(q,t) &= \sum_{x:t(x)>q(x)} (t(x)-q(x)) \\ &= \sum_{x:t(x)>\max\{p(x),q(x)\}} (t(x)-q(x)) + \sum_{x:p(x)\geq t(x)>q(x)} (t(x)-q(x)). \end{aligned}$$

The second sum is $\leq \epsilon$, since otherwise t puts more than ϵ probability mass on points x with $t(x) \leq p(x)$, and the distance of t and p is less than $1-\epsilon$, contradicting our hypothesis.

Hence,

$$\sum_{x:t(x)>\max\{p(x),q(x)\}} (t(x)-q(x)) \geq 1-2\epsilon.$$

Repeating the same argument with p and q swapped we get

$$\sum_{x:t(x)>\max\{p(x),q(x)\}} (t(x)-p(x)) \geq 1-2\epsilon.$$

Taking the average of these two inequalities we get

$$\sum_{x:t(x)>\max\{p(x),q(x)\}} (t(x)-(p(x)+q(x))/2) \geq 1-2\epsilon,$$

as desired.

To prove the last sentence in the lemma statement, consider the domain $\{1,2,3,4\}$ and distributions as follows:

$p(1)=\epsilon, q(1)=0, t(1)=\epsilon/2,$
$p(2)=0, q(2)=\epsilon, t(2)=\epsilon/2,$
$p(3)=0, q(3)=0, t(3)=1-\epsilon,$
$p(4)=1-\epsilon, q(4)=1-\epsilon, t(4)=0.$

Note that $r(i)=p(i)=q(i)$ for $i \in \{3,4\}$. We have $\Delta(p,t)=\Delta(q,t)=\epsilon/2+1-\epsilon = 1-\epsilon/2$, but $\Delta(r,t)=1-\epsilon$. □

Corollary 18. *Let r and t be distributions over the same domain. Suppose that $r = \frac{1}{2^s}\sum_{i=1}^{2^s} p_i$ where each p_i is a distribution with $\Delta(p_i,t) \geq 1-\epsilon$. Then $\Delta(r,t) \geq 1-2^s\epsilon$.*

Proof. We proceed by induction on s. Write $r = \frac{1}{2}(r_1+r_2)$ where the r_i are convex combinations of 2^{s-1} distributions. By hypothesis $\Delta(r_1,t) \geq 1-2^{s-1}\epsilon$, and the same holds for r_2. By Lemma 17, $\Delta(r,t) \geq 1-2^s\epsilon$. □

4.2 Independent vs. permutation

We shall need a lemma about concentration of measure.

Lemma 19. *Let $X_1, X_2, \ldots, X_m$ be boolean random variables such that for every i, conditioned on any outcome of all the variables except X_i, we have $\Pr[X_i = 1] \geq p$. Then we have $\Pr[\sum X_i \leq 0.5pm] \leq \exp(-\Omega(pm))$.*

Similar lemmas have been proved many times. For completeness we give a proof relying on a bound in [PS97]. We use the presentation in [IK10].

Proof. Define $Y_i := 1 - X_i$. We have $\Pr[Y_i = 1] \leq 1 - p$ conditioned on any outcome of all Y the variables except Y_i. We need to bound $\Pr[\sum Y_i \geq m(1 - 0.5p)]$. The variables Y_i satisfy the property that for any set $S \subseteq [m]$, $\Pr[\forall i \in S, Y_i = 1] \leq (1-p)^{|S|}$, because the probability can be written as $\Pr[Y_{i_1} = 1] \cdot \Pr[Y_{i_2} = 1 | Y_{i_1} = 1] \cdots$, where $i_1, i_2, \ldots$ are the elements of S, and each term is at most $1 - p$. So we can apply Theorem 1.1 in [IK10] to obtain

$$\Pr[\sum Y_i \geq m(1 - 0.5p)] \leq e^{-mD(1-0.5p|1-p)}$$

where D is the relative entropy defined as $D(x|y) = x \log_e(x/y) + (1-x) \log_e((1-x)/(1-y))$. From the definition we observe $D(x|y) = D(1-x|1-y)$, hence the above upper bound is $e^{-mD(0.5p|p)}$. Finally, we claim that $D(0.5p|p) \geq \Omega(p)$. This can be verified by calculus or numerically. □

We can now state and prove our main result of this subsection.

Lemma 20. *Let $X_1, X_2, \ldots, X_t$ be t independent random variables over $[n]$, not necessarily uniform. Let Π be a random, uniform permutation over $[n]$. The statistical distance between the x_i and $\Pi(1), \Pi(2), \ldots, \Pi(t)$ is at least $1 - \exp(-\Omega(t^2/n))$.*

Proof. Let $p := 0.5t/n$. Call a variable X_i *low-entropy* if there is a set S_i of size $t/2 = pn$ such that $\Pr[X_i \in S_i] \leq 0.1p$. We consider two cases:

Case 1: There are $\geq t/2$ low-entropy variables X_i:

In this case select any $b := 0.1t$ low-entropy variables. Without loss of generality assume that they are $X_1, X_2, \ldots, X_b$ and let $Y_1, Y_2, \ldots, Y_b$ be the indicator variables corresponding to the events "$X_i \in S_i$". Consider the statistical test "$\sum_{i \leq b} Y_i \geq 0.2p \cdot b$". We have $\mathbb{E}[\sum_{i \leq b} Y_i] \leq 0.1p \cdot b$. The probability that the test passes is at most the probability that $\sum Y_i$ deviates from its expectation by a constant factor. Without loss of generality we can assume that $\mathbb{E}[\sum_{i \leq b} Y_i]$ is exactly $0.1bp$. The variables are independent, and so by a Chernoff bound this probability is at most $\exp(-0.1bp) = \exp(-\Omega(t^2/n))$.

Now let instead $Y_1, Y_2, \ldots, Y_b$ be the indicator variables corresponding to the events "$\Pi(i) \in S_i$". We observe that regardless of the outcome of any other $r \leq b$ variables $\Pi(j)$, $j \neq i$, (note that $\Pi(j)$ determines Y_j)

$$\Pr[Y_i = 1] \geq \frac{|S_i| - r}{n - r} = \frac{0.5t - r}{n - r} \geq \frac{0.4t}{n} = 0.8p.$$

The probability that the test does not pass is at most the probability that $\sum_i Y_i < 0.5(0.8p)b$, and that by Lemma 19 is $\leq \exp(-\Omega(t/n) \cdot b) = \exp(-\Omega(t^2/n))$.

Case 2: There are $< t/2$ low-entropy variables X_i:

In this case there are $\geq t/2$ high-entropy variables, i.e., variables such that for every set S_i of size $t/2$, the probability of landing in S_i is $\geq 0.1p$. Let H be the index set of $t/2$ of these variables, and L be the index set of the other $t/2$ variables (which may or may not be high entropy). The probability that the X_i collide (i.e., two variables take the same value) is at least the probability that the variables collide conditioned on the event that there is no collision among the variables in L. Fix any outcome for the variables in L conditioned on the event that they do not collide. Because they do not collide, they take $t/2$ distinct values. Let S be the set of $t/2$ values they take. Now the probability that the X_i variables collide is at least the probability that some variable in H lands in S. Because the variables are independent, this probability is at least

$$1 - (1 - 0.1p)^{t/2} \geq 1 - e^{-\Omega(pt)} = 1 - e^{-\Omega(t^2/n)}.$$

On the other hand, by definition, the variables $\Pi(i)$ never collide. Hence the statistical test that simply checks if the variables collide gives the desired statistical distance. $\square$

4.3 Proof of Theorem 3

We proceed by induction on d. We can take $d = 0$ as base case. In this case f is constant and the statistical distance is $1 - 1/n!$ which is larger than $1 - 2^{-n/\log n}$.

For the induction step, we do a case analysis depending on whether there are

$$t := n/\log^{c^d/4} n$$

output variables with indexes $T \subseteq [n]$ whose probes intersect the probes of all other variables. In the case in which there are t such variables, by considering any possible fixing for the values of the cells probed by the variables in T the sampled distribution is a convex combination of $2^{t \cdot d \cdot \log n}$ distributions which are $(d-1)$-local. (Note that we can hardwire the values of the fixed cells in the sampler so that the distribution is only a function of the unfixed cells.)

By the induction hypothesis applied to each of these samplers, and Corollary 18 the statistical distance will be

$$1 - 2^{t \cdot d \cdot \log n} \cdot 2^{-n/\log^{c^{d-1}} n}.$$

This quantity equals $1 - 2^{-x}$ where

$$x = \frac{n}{\log^{c^{d-1}} n} - td\log n = n\left(\frac{1}{\log^{c^{d-1}} n} - \frac{d\log n}{\log^{c^d/4} n}\right) \geq 0.5\frac{n}{\log^{c^{d-1}} n} \geq \frac{n}{\log^{c^d} n}.$$

Here the inequalities hold for $d \leq \log n$ say (for else the theorem is trivial – here is where we are using the factor 2 in front of exp in the statement) and a suitable choice of c.

In the case in which there are not t such variables then there are t variables which are independent. (This can be shown by iteratively collecting variables whose probes are disjoint.

We can't stop before we collect t, for else the answer would have been yes.) By Lemma 20 just considering those variables the statistical distance is at least $1-\exp(-\Omega(t^2/n))$. Noting that $t^2/n = n/\log^{c^d/2} n \geq (\log^{c^d/2} n)n/\log^{c^d} n$ concludes the argument for all large enough n.

It may be helpful to add that the fact that we need to square t at each step makes the exponent of the log in the final bound exponential in d.

5 Open problems

The study of the complexity of distributions remains largely uncharted. We discuss next several open problems that seem within reach.

One problem from [Vio14] that is still open is to extract from 2-local sources on n bits with min-entropy $\leq \sqrt{n}$. The problem is that while such sources with min-entropy $n^{1/2+\Omega(1)}$ can be restricted to being affine sources with large entropy, this is not true if we start with entropy $\sqrt{n}$. To see this, consider the 2-local source on n bits sampled from $O(\sqrt{n})$ bits by taking the And of any possible pair. As long as the restriction leaves two input bits unfixed, the output is not affine.

As our techniques for Theorem 2 are based on restrictions, it is natural to ask whether one can prove average-case sampling lower bounds for other models which shrink under restrictions, including various types of formulas, and branching programs. We refer the reader to [IMZ12] for pointers and a recent discussion of these models. For concreteness, the reader can think of de Morgan's formulas. We note that worst-case sampling lower bounds for these models do follow from [Vio14]. However it does not seem immediate to obtain average-case lower bounds via the approach in this paper, because the shrinkage parameter is not large enough.

Another open problem is to efficiently reduce the input length of the samplers. This was studied by Dubrov and Ishai [DI06] and later by Artemenko and Shaltiel [AS17]. The latter paper shows that the output distribution of an AC^0 circuit $C : \{0,1\}^\ell \to \{0,1\}^m$ can be approximately sampled by another AC^0 circuit $C' : \{0,1\}^{\ell'} \to \{0,1\}^m$ using only $\ell' = m^{1+\alpha}$ input bits, for any $\alpha > 0$ and where the depth of C' depends on α. It would be exciting to obtain $\ell' = O(m)$, and it would have an application to specific distributions such as $D = (U, \mathrm{Majority}(U))$: The AC^0 sampler for D in [Vio12b] goes through the generation of a nearly uniform permutation of $[m]$ and thus uses $\geq m \log m$ input bits. It is an open question whether D can be sampled in AC^0 using fewer input bits. A positive answer to this question would follow from a strong derandomization of our Theorem 6. Specifically, if one could construct an AC^0 circuit that given $m - m^\epsilon$ input bits samples a uniform restriction from the multiset S in Theorem 6, then it would be enough to fill the $m^{\epsilon'}$ stars with bounded independence to obtain $\ell' = m + m^{1-\Omega(1)}$.

It would also be interesting to extend Theorem 3 to the case of *adaptive* probes.

Challenge: Let $f : [n]^\ell \to [n]^n$ be a map such that each output symbol depends on $d = O(1)$ adaptively chosen input cells. Show that the output distribution of f has statistical distance

$\Omega(1)$ from a uniform random permutation.

Another question is to separate the power of adaptive and non-adaptive cell probes. Consider the distribution D over $[n]^{R+n}$ where the first R cells D_i are uniform in $[n]^R$ and each other cell is sampled as follows: Pick a uniform, independent index j in $\{1, 2, \ldots, R\}$ and output D_j. By definition D can be sampled with 2 adaptive probes. We conjecture that for $R = n^{1-\Omega(1)}$ sampling D requires a large number of non-adaptive probes.

Finally, recall that the statistical bound in Theorem 3 is not far from optimal for small locality. At the other end of the spectrum, it is an interesting question what is the minimum locality sufficient to reduce the statistical distance to $1 - \Omega(1)$.

Acknowledgments and relationship with [Vio18]. We thank the anonymous referees for the helpful suggestions mentioned in the paper, and others. A preliminary version [Vio18] of this paper proved Theorem 2 with quasi-polynomial error, and raised the problem of improving the error to exponential. The solution in this paper emerged during discussions with Benjamin Rossman and Avishay Tal at the Simons Institute in Fall 2018. The change in the proof amounts to using the switching lemma in [BKST18], which builds on [Hås14] and postdates [Vio18], further restricting the common part, and working out parameters. We thank them for their permission to include the stronger bound in this revision.

References

[Ajt83] Miklós Ajtai. Σ_1^1-formulae on finite structures. *Annals of Pure and Applied Logic*, 24(1):1–48, 1983.

[AS17] Sergei Artemenko and Ronen Shaltiel. Pseudorandom generators with optimal seed length for non-boolean poly-size circuits. *ACM Trans. Computation Theory*, 9(2):6:1–6:26, 2017.

[Bab87] László Babai. Random oracles separate PSPACE from the polynomial-time hierarchy. *Information Processing Letters*, 26(1):51–53, 1987.

[BCS14] Itai Benjamini, Gil Cohen, and Igor Shinkar. Bi-lipschitz bijection between the boolean cube and the hamming ball. In *IEEE Symp. on Foundations of Computer Science (FOCS)*, 2014.

[BDT16] Avraham Ben-Aroya, Dean Doron, and Amnon Ta-Shma. Explicit two-source extractors for near-logarithmic min-entropy. *Electronic Colloquium on Computational Complexity (ECCC)*, 23:88, 2016.

[BIL12] Chris Beck, Russell Impagliazzo, and Shachar Lovett. Large deviation bounds for decision trees and sampling lower bounds for AC0-circuits. In *IEEE Symp. on Foundations of Computer Science (FOCS)*, pages 101–110, 2012.

[BIVW16] Andrej Bogdanov, Yuval Ishai, Emanuele Viola, and Christopher Williamson. Bounded indistinguishability and the complexity of recovering secrets. In *Int. Cryptology Conf. (CRYPTO)*, 2016.

[BKST18] Adam Bene Watts, Robin Kothari, Luke Schaeffer, and Avishay Tal. Exponential separation between shallow quantum circuits and unbounded fan-in shallow classical circuits. 2018.

[BL15] Joshua Brody and Kasper Green Larsen. Adapt or die: Polynomial lower bounds for non-adaptive dynamic data structures. *Theory of Computing*, 11:471–489, 2015.

[CL06] Fan Chung and Linyuan Lu. Concentration inequalities and martingale inequalities: a survey. *Internet Math.*, 3(1):79–127, 2006.

[Coh16] Gil Cohen. Making the most of advice: New correlation breakers and their applications. In *IEEE Symp. on Foundations of Computer Science (FOCS)*, pages 188–196, 2016.

[CS16] Gil Cohen and Leonard J. Schulman. Extractors for near logarithmic min-entropy. *Electronic Colloquium on Computational Complexity (ECCC)*, 23:14, 2016.

[CZ16] Eshan Chattopadhyay and David Zuckerman. Explicit two-source extractors and resilient functions. In *ACM Symp. on the Theory of Computing (STOC)*, pages 670–683, 2016.

[Czu15] Artur Czumaj. Random permutations using switching networks. In *ACM Symp. on the Theory of Computing (STOC)*, pages 703–712, 2015.

[DGJ+10] Ilias Diakonikolas, Parikshit Gopalan, Ragesh Jaiswal, Rocco A. Servedio, and Emanuele Viola. Bounded independence fools halfspaces. *SIAM J. on Computing*, 39(8):3441–3462, 2010.

[DI06] Bella Dubrov and Yuval Ishai. On the randomness complexity of efficient sampling. In *38th ACM Symposium on Theory of Computing (STOC)*, pages 711–720, 2006.

[DPT10] Yevgeniy Dodis, Mihai Pătraşcu, and Mikkel Thorup. Changing base without losing space. In *42nd ACM Symp. on the Theory of Computing (STOC)*, pages 593–602. ACM, 2010.

[DW11] Anindya De and Thomas Watson. Extractors and lower bounds for locally samplable sources. In *Workshop on Randomization and Computation (RANDOM)*, 2011.

[DW12] Anindya De and Thomas Watson. Extractors and lower bounds for locally samplable sources. *ACM Trans. Computation Theory*, 4(1):3, 2012.

[FSS84] Merrick L. Furst, James B. Saxe, and Michael Sipser. Parity, circuits, and the polynomial-time hierarchy. *Mathematical Systems Theory*, 17(1):13–27, 1984.

[GGG+07] Alexander Golynski, Roberto Grossi, Ankur Gupta, Rajeev Raman, and S. Srinivasa Rao. On the size of succinct indices. In *ESA*, pages 371–382, 2007.

[GM07] Anna Gál and Peter Bro Miltersen. The cell probe complexity of succinct data structures. *Theoretical Computer Science*, 379(3):405–417, 2007.

[Gol09] Alexander Golynski. Cell probe lower bounds for succinct data structures. In *20th ACM-SIAM Symp. on Discrete Algorithms (SODA)*, pages 625–634, 2009.

[GRR08] Alexander Golynski, Rajeev Raman, and S. Srinivasa Rao. On the redundancy of succinct data structures. In *SWAT*, pages 148–159, 2008.

[Hag91] Torben Hagerup. Fast parallel generation of random permutations. In *18th Coll. on Automata, Languages and Programming (ICALP)*, pages 405–416. Springer, 1991.

[Hås87] Johan Håstad. *Computational limitations of small-depth circuits*. MIT Press, 1987.

[Hås14] Johan Håstad. On the correlation of parity and small-depth circuits. *SIAM J. on Computing*, 43(5):1699–1708, 2014.

[IK10] Russell Impagliazzo and Valentine Kabanets. Constructive proofs of concentration bounds. In *Workshop on Randomization and Computation (RANDOM)*, pages 617–631. Springer, 2010.

[IMP12] Russell Impagliazzo, William Matthews, and Ramamohan Paturi. A satisfiability algorithm for AC^0. In *ACM-SIAM Symp. on Discrete Algorithms (SODA)*, pages 961–972, 2012.

[IMZ12] Russell Impagliazzo, Raghu Meka, and David Zuckerman. Pseudorandomness from shrinkage. In *IEEE Symp. on Foundations of Computer Science (FOCS)*, pages 111–119, 2012.

[IN96] Russell Impagliazzo and Moni Naor. Efficient cryptographic schemes provably as secure as subset sum. *J. of Cryptology*, 9(4):199–216, 1996.

[Kil88] Joe Kilian. Founding cryptography on oblivious transfer. In *ACM Symp. on the Theory of Computing (STOC)*, pages 20–31, 1988.

[Lar12] Kasper Green Larsen. Higher cell probe lower bounds for evaluating polynomials. In *IEEE Symp. on Foundations of Computer Science (FOCS)*, pages 293–301, 2012.

[Li16] Xin Li. Improved two-source extractors, and affine extractors for polylogarithmic entropy. In *IEEE Symp. on Foundations of Computer Science (FOCS)*, 2016.

[LV12] Shachar Lovett and Emanuele Viola. Bounded-depth circuits cannot sample good codes. *Computational Complexity*, 21(2):245–266, 2012.

[MRRR11] J. Ian Munro, Rajeev Raman, Venkatesh Raman, and S. Srinivasa Rao. Succinct representations of permutations and functions. *CoRR*, abs/1108.1983, 2011.

[MRRR12] J. Ian Munro, Rajeev Raman, Venkatesh Raman, and S. Srinivasa Rao. Succinct representations of permutations and functions. *Theor. Comput. Sci.*, 438:74–88, 2012.

[MV91] Yossi Matias and Uzi Vishkin. Converting high probability into nearly-constant time-with applications to parallel hashing. In *23rd ACM Symp. on the Theory of Computing (STOC)*, pages 307–316, 1991.

[O'D14] Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014.

[Păt08] Mihai Pătraşcu. Succincter. In *49th IEEE Symp. on Foundations of Computer Science (FOCS)*. IEEE, 2008.

[Pre18] Nicola Prezza. Optimal substring-equality queries with applications to sparse text indexing (preliminary version title: In-place sparse suffix sorting). In *ACM-SIAM Symp. on Discrete Algorithms (SODA)*, pages 1496–1508, 2018.

https://arxiv.org/pdf/1803.01723.

[PRST16] Toniann Pitassi, Benjamin Rossman, Rocco A. Servedio, and Li-Yang Tan. Polylogarithmic Frege depth lower bounds via an expander switching lemma. In *ACM Symp. on the Theory of Computing (STOC)*, pages 644–657, 2016.

[PS97] Alessandro Panconesi and Aravind Srinivasan. Randomized distributed edge coloring via an extension of the chernoff-hoeffding bounds. *SIAM J. Comput.*, 26(2):350–368, 1997.

[PV10] Mihai Pătraşcu and Emanuele Viola. Cell-probe lower bounds for succinct partial sums. In *21th ACM-SIAM Symp. on Discrete Algorithms (SODA)*, pages 117–122, 2010.

[Rao09] Anup Rao. Extractors for low-weight affine sources. In *IEEE Conf. on Computational Complexity (CCC)*, pages 95–101, 2009.

[Ros17] Benjamin Rossman. An entropy proof of the switching lemma and tight bounds on the decision-tree size of AC0, 2017. Manuscript. URL: http://www.math.toronto. edu/rossman/logsize.pdf.

[Sie04] Alan Siegel. On universal classes of extremely random constant-time hash functions. *SIAM J. on Computing*, 33(3):505–543, 2004.

[Sub61] B. A. Subbotovskaya. Realizations of linear functions by formulas using +, *, -. *Soviet Mathematics-Doklady*, 2:110–112, 1961.

[Tho13] Mikkel Thorup. Mihai patrascu: Obituary and open problems. *Bulletin of the EATCS*, 109:7–13, 2013.

[Vio05] Emanuele Viola. On constructing parallel pseudorandom generators from one-way functions. In *20th IEEE Conf. on Computational Complexity (CCC)*, pages 183–197, 2005.

[Vio09a] Emanuele Viola. Cell-probe lower bounds for prefix sums, 2009. arXiv:0906.1370v1.

[Vio09b] Emanuele Viola. Gems of theoretical computer science. Lecture notes of the class taught at Northeastern University. Available at http://www.ccs.neu.edu/home/viola/classes/gems-08/index.html, 2009.

[Vio12a] Emanuele Viola. Bit-probe lower bounds for succinct data structures. *SIAM J. on Computing*, 41(6):1593–1604, 2012.

[Vio12b] Emanuele Viola. The complexity of distributions. *SIAM J. on Computing*, 41(1):191–218, 2012.

[Vio12c] Emanuele Viola. Extractors for turing-machine sources. In *Workshop on Randomization and Computation (RANDOM)*, 2012.

[Vio14] Emanuele Viola. Extractors for circuit sources. *SIAM J. on Computing*, 43(2):355–972, 2014.

[Vio16] Emanuele Viola. Quadratic maps are hard to sample. *ACM Trans. Computation Theory*, 8(4), 2016.

[Vio17a] Emanuele Viola. A sampling lower bound for permutations. *Electronic Colloquium on Computational Complexity (ECCC)*, 24:166, 2017.

[Vio17b] Emanuele Viola. Special topics in complexity theory. Lecture notes of the class taught at Northeastern University. Available at http://www.ccs.neu.edu/home/viola/classes/spepf17.html, 2017.

[Vio18] Emanuele Viola. Sampling lower bounds: boolean average-case and permutations. ECCC TR18-060, 2018.

[Yao85] Andrew Yao. Separating the polynomial-time hierarchy by oracles. In *26th IEEE Symp. on Foundations of Computer Science (FOCS)*, pages 1–10, 1985.