Electronic Colloquium on Computational Complexity, Report No. 196 (2023)

Sampling, Flowers and Communication

Huacheng Yu *
Princeton University
yuhch123@gmail.com

Wei Zhan†
University of Chicago
weizhan@uchicago.edu

Abstract

Given a distribution over $[n]^n$ such that any k coordinates need $k/\log^{O(1)} n$ bits of communication to sample, we prove that any map that samples this distribution from uniform cells requires locality $\Omega(\log(n/k)/\log\log(n/k))$. In particular, we show that for any constant $\delta > 0$, there exists $\varepsilon = 2^{-\Omega(n^{1-\delta})}$ such that $\Omega(\log n/\log\log n)$ non-adaptive cell probes on uniform cells are required to:

- Sample a uniformly random permutation on n elements with error $1-\varepsilon$. This provides an exponential improvement on the $\Omega(\log\log n)$ cell probe lower bound by Viola.
- Sample an n-vector with each element independently drawn from a random $n^{1-\delta}$-vector, with error $1-\varepsilon$. This provides the first adaptive vs non-adaptive cell probe separation for sampling.

The major technical component in our proof is a new combinatorial theorem about flower with small kernel, i.e. a collection of sets where few elements appear more than once. We show that in a family of n sets, each with size $O(\log n/\log\log n)$, there must be $k = \mathrm{poly}(n)$ sets where at most $k/\log^{O(1)} n$ elements appear more than once.

To show the lower bound on sampling permutation, we also prove a new $\Omega(k)$ communication lower bound on sampling uniformly distributed disjoint subsets of $[n]$ of size k, with error $1-2^{-\Omega(k^2/n)}$. This result unifies and subsumes the lower bound for $k = \Theta(\sqrt{n})$ by Ambainis et al., and the lower bound for $k = \Theta(n)$ by Göös and Watson.

1 Introduction

In this paper we examine the complexity of generating a certain distribution, which is a study initially advocated by Viola [Vio12] and then followed up by a long line of works, e.g. [LV12, BIL12, Wat14, Vio16, Vio20, CGZ22]. Specifically, we consider the problem of sampling a distribution over vectors of lengths n with symbols in $[n] = \{1, \ldots, n\}$, given access to ℓ uniformly random cells of $\log n$ bits, via a function $f\colon [n]^\ell \to [n]^n$, where ℓ can be arbitrarily large. We say f is s-local if each output symbol depends on at most s input symbols, or in other words, f can be computed with s non-adaptive cell probes.

We connect this notion of sampling with local maps to the task of sampling with communication protocols, which was studied in e.g. [AST+03, Wat16, GW20, CGZ22]. Following [GW20], for a distribution $\mathcal{D}$ on vectors of length k, we use $\mathsf{Samp}_\varepsilon(\mathcal{D})$ to denote the smallest $\ell \in \mathbb{N}$ such that there exists a private-randomness communication protocol among k parties, with transcript length

*Supported by Simons Junior Faculty Award - AWD1007164.

†Supported by a Simons Investigator Award and by the National Science Foundation grant No. CCF-2007462. This work was partially done when WZ was in Princeton University.

ISSN 1433-8092

always at most ℓ, that the distribution $\mathcal{C}$ on the outputs of the k parties satisfies $\Delta(\mathcal{C}, \mathcal{D}) \le \varepsilon$. Here $\Delta(\cdot,\cdot)$ denotes the statistical (i.e. total-variation) distance, that is,

$$\Delta(\mathcal{C}, \mathcal{D}) = \max_{\mathcal{X} \subseteq [n]^k} \left| \Pr_{X \sim \mathcal{C}}[X \in \mathcal{X}] - \Pr_{X \sim \mathcal{D}}[X \in \mathcal{X}] \right|.$$

Note that merging several parties into one does not increase the communication complexity, and the communication lower bounds in this paper are all proved for two parties, so we do not specify the number of parties when applying these lower bounds.

If the distribution $\mathcal{D}$ is over vectors of length n, we use $\mathcal{D}_I$ to denote the marginal distribution of $\mathcal{D}$ on the coordinates in I for any non-empty set $I \subseteq [n]$. We relate the locality of sampling $\mathcal{D}$ to the communication complexity of sampling $\mathcal{D}_I$ through the our main theorem, which states as follows.

Theorem 1.1. *Let $\mathcal{D}$ be a distribution on $[n]^n$. Suppose $\varepsilon \in (0,1)$ and integers $0 < h \le k < n$ satisfy that $\mathsf{Samp}_{1-\varepsilon}(\mathcal{D}_I) > h \log n$ for every set $I \subset [n]$ of k coordinates. Then for any $s \in \mathbb{N}$ such that $2s^2 \le h$ and*

$$s \le \frac{\log(n/k)}{\log(8k/h) + 4\log\lceil\log(n/k)\rceil}$$

and arbitrary $\ell \in \mathbb{N}$, every s-local map $f\colon [n]^\ell \to [n]^n$ must satisfy $\Delta(f(\mathcal{U}), \mathcal{D}) \ge 1 - \varepsilon$, where $\mathcal{U}$ is the uniform distribution over $[n]^\ell$.

The key component in the proof is a new combinatorial theorem (Theorem 2.1) on the existence of a *flower* with small kernel in a large family of sets, which consists of k sets within which there are at most h elements appearing more than once. The theorem could also be of independent interest since flowers are natural combinatorial objects and may emerge in other applications.

We give two applications of Theorem 1.1 which answers two open problems from [Vio20]. The first application is on sampling permutations. If we take any k coordinates in a random permutation and divide it into two equal parts, then the two parts are uniformly distributed over two disjoint subsets of $[n]$ with size $k/2$. Consequently, we have to prove the following communication lower bound on sampling disjoint sets with fixed size:

Theorem 1.2. *Let $[n]^{(k)}$ denote the collection of all size-k subsets of $[n]$. Let $\mathcal{D}_{n,k}$ be the uniform distribution over*

$$\left\{(X, Y) \mid X, Y \in [n]^{(k)}, X \cap Y = \varnothing\right\}.$$

Then for every $\omega(\log\log n) < k \le n/2$, there exists $\varepsilon = 2^{-\Omega(k^2/n)}$ such that $\mathsf{Samp}_{1-\varepsilon}(\mathcal{D}_{n,k}) = \Omega(k)$.

We note that some special cases of Theorem 1.3 have been known before this work. Ambainis et al. [AST+03] proved the case for $k = \Theta(\sqrt{n})$, while Göös and Watson [GW20] proved the case for $k = \Theta(n)$, or similarly when there is no size restriction (where a uniformly random pair of disjoint subsets almost always both have size $\Theta(n)$). A simpler proof for the result of [GW20] was later found by Chattopadhyay, Goodman and Zuckerman [CGZ22]. However, neither cases are suitable for our application: We need $k = o(n)$ to have a meaningful lower bound on the locality s when applying Theorem 1.1, and we need $k = \omega(\sqrt{n})$ so that ε is small and thus the statistical distance is close to 1. Our Theorem 1.2 subsumes both previous cases and we present a self-contained proof in Section 3.1.

By choosing $k = n^{1-\delta}$ for some $\delta \in (0, 1/2)$ in Theorem 1.2 and $h = \Theta(k/\log^2 n)$ in Theorem 1.1, we immediately obtain the following corollary.

Theorem 1.3. *Let $\mathcal{D}$ be the uniform distribution over all permutations in $[n]^n$. For every constant $\delta > 0$, there exists $s = O(\log n / \log\log n)$ such that every s-local map $f\colon [n]^\ell \to [n]^n$ satisfies $\Delta(f(\mathcal{U}), \mathcal{D}) \geq 1 - 2^{-\Omega(n^{1-\delta})}$.*

Viola in [Vio20] proved a bound $\Delta(f(\mathcal{U}), \mathcal{D}) \geq 1 - \exp(-n/\log^{2^{O(s)}} n)$ for sampling permutation with s-local maps, which is only non-trivial for $s \leq O(\log\log n)$. Therefore, our Theorem 1.3 provides an exponential improvement on the locality lower bound. It also almost answers the long-standing open problem raised at the end of [Vio12], which asked for an $\Omega(\log n)$ locality lower bound. The result also implies a succinct data structure lower bound for storing permutations as follows.

Corollary 1.4. *For every constant $\delta > 0$, there exists $s = O(\log n / \log\log n)$ such that the following holds. Any cell-probe data structure for storing permutations on n elements, such that each element can be retrieved with s non-adaptive probes in cells of $\log n$ bits, must use $\log(n!) + \Omega(n^{1-\delta})$ space.*

The proof of Corollary 1.4 from Theorem 1.3 can be found in [Vio12, Vio20]. Our second application is on sampling tuples in $[n]^n$ with at most k distinct elements:

Theorem 1.5. *Let $\mathcal{E}$ be the following distribution on $[n]^n$ for some $k \in \mathbb{N}_+$: first sample r uniformly from $[n]^k$, and let $x \sim \mathcal{E}$ be that for every $i \in [n]$, independently and uniformly draw $j \in [k]$ and let $x_i = r_j$. For every $k \leq n^{1-\Omega(1)}$, there exists $s = O(\log n / \log\log n)$ such that every s-local map $f\colon [n]^\ell \to [n]^n$ satisfies $\Delta(f(\mathcal{U}), \mathcal{E}) \geq 1 - 2^{-\Omega(k)}$.*

Theorem 1.5 is proved via Theorem 1.1 by showing an $\Omega(k)$ communication lower bound for sampling two equal sets of size k (Lemma 3.3), whose proof is fairly simple and standard. As noted in [Vio20], each output symbol in $\mathcal{E}$ can be sampled with two *adaptive* probes from a uniformly random input in $[k]^n \times [n]^k$. It was conjectured in [Vio20] that non-adaptively a large amount of probes is required. Theorem 1.5 proves the conjecture and thus providing an $O(1)$ vs. $\Omega(\log n / \log\log n)$ separation between adaptive and non-adaptive cell probes for sampling.

Comparisons with [FLRS23] Independently of our work, Filmus, Leigh, Riazanov and Sokolov [FLRS23] recently proved an $\Omega(\log(n/k)/\log\log(n/k))$ lower bound on the decision forest depth for sampling uniformly random n-bit strings of Hamming weight k, which directly implies a lower bound on sampling uniform permutations (as we can assign 1 to the outputs symbols in $\{1, \ldots, k\}$ and 0 to $\{k+1, \ldots, n\}$).

Yet, their method and results are incomparable to ours: Although the lower bound they proved works for adaptive queries, it is only against *bit probes* (as it crucially depends on the fact that $2^s < n$ for the probe number s, which is not true if we replace the left-hand side with n^s); while our lower bound works against *cell probes* on $O(\log n)$-bit cells, albeit being non-adaptive. In addition, their method could only prove a bound on the statistic distance up to $1 - n^{-O(1)}$ while our bounds are exponentially close to 1. However, our method apparently does not work on the Hamming weight sampling problem that they considered.

Nevertheless, our work shares some interesting similarities with [FLRS23]: The asymptotic lower bounds on the number of probes happen to be identical, and their proof also used some variant of the sunflowers (robust sunflowers [Ros14, ALWZ21] to be precise). This indicates that there might be deeper connections between the two works that we are currently not aware of.

2 Flower with Small Kernel

In this section we prove Theorem 1.1. The idea is to characterize every s-local map using a family of n sets each with size at most s, indicating the input symbols that each output symbol depends

on. We want to show that within this family there exists a large flower with small kernel, i.e. k sets such that at most $h < k$ elements appear more than once. In this case broadcasting h random input symbols suffices to sample the corresponding k output symbols, as the rest of the input symbols used for each output are independent.

To put it formally, we first define the flowers.

Definition 1. An *s-family* is a collection of sets $S_1, \ldots, S_n$ such that $|S_i| \le s$ for all i. The family is called a *flower* with *kernel* K, if $S_i \cap S_j \subset K$ for all $i \neq j$, and each $S_i \setminus K$ is called a *petal*.

The readers may find the definition a reminiscence of the well-known notion of *sunflowers* [ER60, ALWZ21, BCW21]. Indeed, for a sunflower it is instead required that $S_i \cap S_j = K$ for all $i \neq j$. In [RR18] flowers are defined similarly to ours, but with the additional requirement that $|K| \le s$. As a result, for a k-petal flower with their definition to exist, the size of the s-family has to be $k^{\Omega(s)}$, similar to the case of sunflowers. Here we show that, if we allow the kernel size $|K|$ to be larger (but still non-trivially small, while trivially the kernel could be the union and have size $\Omega(ks)$), then the dependence on k could be much better.

Theorem 2.1. *For every $s, k > 0$ and $h \ge 2s^2$, if $n \ge (8ks^4/h)^s \cdot k$, then in every s-family $\mathcal{F}$ of n sets, there exists a k-petal flower with kernel K such that $|K| \le h$.*

Proof. We will iteratively construct a sequence of s-families $\mathcal{F}_0, \mathcal{F}_1, \ldots$ until certain requirements are met specified below. Initially, let $\mathcal{F}_0 = \mathcal{F}$.

Suppose that we currently have a set system $\mathcal{F}_t$ with size $|\mathcal{F}_t| = n_t$. We say an element x is heavy (within this round of iteration), if

$$\big|\{S \in \mathcal{F}_t \mid x \in S\}\big| > \frac{n_t}{2ks},$$

and otherwise x is light. Let H be the collection of heavy elements in this round, and we have $|H| \le 2ks^2$ by counting the pairs $(S \in \mathcal{F}_t, x \in S)$. We partition $\mathcal{F}_t$ into subfamilies based on the size of $S \cap H$: For every $0 \le i \le s$, let

$$\mathcal{F}_{t,i} = \big\{S \in \mathcal{F}_t \mid |S \cap H| = i\big\}.$$

If $|\mathcal{F}_{t,0}| < \frac{1}{2} n_t$, then there must be some $i > 0$ such that $|\mathcal{F}_{t,i}| \ge n_t/(2s)$. In this case we continue the iteration by letting

$$\mathcal{F}_{t+1} = \big\{S \setminus H \mid S \in \mathcal{F}_{t,i}, S \cap H \subseteq H_t\big\},$$

where $H_t \subseteq H$ is a set of at most h/s heavy elements chosen to maximize the size of $\mathcal{F}_{t+1}$. In fact, if $|H| \le h/s$ then we can simply take $H_t = H$. Otherwise let H_t be a uniformly random subset of H with size h/s, then for every $S \in \mathcal{F}_{t,i}$ we have

$$\Pr[S \cap H \subseteq H_t] = \binom{|H| - i}{h/s - i} \Big/ \binom{|H|}{h/s} \ge \left(\frac{h/s - i}{|H|}\right)^i \ge \left(\frac{h}{4ks^3}\right)^i.$$

The least inequality is because $h/s - i \ge h/s - s \ge h/(2s)$ and $|H| \le 2ks^2$. Therefore, there must be a choice of H_t such that

$$n_{t+1} = |\mathcal{F}_{t+1}| \ge \left(\frac{h}{4ks^3}\right)^i |\mathcal{F}_{t,i}| \ge \left(\frac{h}{4ks^3}\right)^i \cdot \frac{n_t}{2s}. \tag{1}$$

Notice that in the above scenario the size of sets in $\mathcal{F}_{t+1}$ is reduced by $i > 0$ compared to that in $\mathcal{F}_t$, which means that the iteration goes for at most s rounds until either we have $|\mathcal{F}_{t,0}| \ge \frac{1}{2} n_t$,

or every set in $\mathcal{F}_t$ becomes empty and thus $|\mathcal{F}_{t,0}| = n_t \geq \frac{1}{2}n_t$ also holds. Now we just greedily pick k sets from $\mathcal{F}_{t,0}$ that are disjoint on the light elements. This is always achievable as long as $|\mathcal{F}_{t,0}| \geq k$, because each set touches at most s light elements, which in total, by the definition of light elements, prohibit at most

$$s \cdot \frac{n_t}{2ks} \leq \frac{1}{2}n_t \leq |\mathcal{F}_{t,0}|$$

sets in $\mathcal{F}_{t,0}$ from further choices.

We take the flower to be the original sets in $\mathcal{F}_0$ corresponding to these k picked sets, and take the kernel K to be the union of H_t. Since there are at most s rounds and $|H_t| \leq h/s$ in each round, we have that $|K| \leq h$. Finally, to ensure that $|\mathcal{F}_{t,0}| \geq k$ holds in the final round, notice that the sum of i through the iteration is at most s, and hence by (1) and the assumption on n,

$$|\mathcal{F}_{t,0}| \geq \frac{1}{2}n_t \geq \left(\frac{h}{4ks^3}\right)^s \cdot (2s)^{-s} \cdot n \geq k. \qquad \square$$

Now we can complete the proof of Theorem 1.1 simply by playing around with definitions.

Proof of Theorem 1.1. Given the s-local map $f\colon [n]^\ell \to [n]^n$, we define an s-family $S_1, \ldots, S_n$ over the universe $[\ell]$, where S_i consists of the indices of inputs symbols that the u-th output symbol depends on. By Theorem 2.1 we can find a flower $\{S_i\}_{i\in I}$ with kernel K, where $|I| = k$ and $|K| \leq h$, as long as $h \geq 2s^2$ and $n \geq (8ks^4/h)^s \cdot k$. The later is equivalent to

$$s \cdot [\log(8k/h) + 4\log s] \leq \log(n/k).$$

Our assumption on s implies that $\log s \leq \log\log(n/k)$, and therefore the above inequality holds.

Using f, we design the following communication protocol for k players to sample $\mathcal{D}_I$. The first player samples uniformly $(x_j)_{j\in K}$ from $[n]^{|K|}$, and broadcast it to other players. Then each player, assigned with $i \in I$, uniformly samples inputs symbols in the petal $(x_j)_{j\in S_i\setminus K}$ and output $f(x)_i$ based on $(x_j)_{j\in S_i}$. Note that the output distribution of the protocol is exactly $f(\mathcal{U})_I$, and communication complexity is $|K|\log n \leq h\log n$. Thus if $\mathsf{Samp}_{1-\varepsilon}(\mathcal{D}_I) > h\log n$, then we have $\Delta(f(\mathcal{U}), \mathcal{D}) \geq \Delta(f(\mathcal{U})_I, \mathcal{D}_I) > 1 - \varepsilon$. $\square$

Remark. The dependence on k in Theorem 2.1 is essentially optimal in the regime of interest for the application in Theorem 1.1. In particular, for every $s, h, k > 0$, there exists an s-family $\mathcal{F}$ of

$$n = \left\lfloor \frac{1}{4e}\left(\frac{k}{2h}\right)^s \cdot k \right\rfloor$$

sets where no k sets form a flower with kernel size at most h.

This fact can be shown via a probabilistic construction. Let each set in $\mathcal{F}$ be a random subset of $[k/2]$ by independently and uniformly sample s elements with replacement (so that each set has size at most s). If a k-petal flower exists with kernel $K \subseteq [k/2]$, since all the petals are disjoint outside K, there must be at least $k/2$ sets in the flower that are completely contained in K. When K is fixed with $|K| \leq h$, each set in $\mathcal{F}$ is contained in K with probability at most $(2h/k)^s$. By a union bound over the choices of these $k/2$ sets among the n sets in $\mathcal{F}$ and the choices of K, we have that the probability of $\mathcal{F}$ containing a k-petal flower with kernel size at most h is at most

$$\left(\frac{2h}{k}\right)^{s\cdot k/2} \cdot \binom{n}{k/2} \cdot 2^{k/2} \leq \left[\left(\frac{2h}{k}\right)^s \cdot \frac{4en}{k}\right]^{k/2}$$

which is less than 1 with the choice of n above. Therefore there exists a family $\mathcal{F}$ as required. This means that the $\log(n/k)$ factor in Theorem 1.1 is essential, and justifies our choices of $k = n^{1-\delta}$.

3 Communication Complexity for Sampling

In this section we prove the communication lower bounds for sampling two-part distributions, which by Theorem 1.1 imply lower bounds for sampling with s-local maps. We prove the lower bound for sampling disjoint sets (Theorem 1.2) in Section 3.1, and prove Theorem 1.5 via a lower bound for sampling equal sets (Lemma 3.3) in Section 3.2.

3.1 Sampling Disjoint k-Sets

Recall the sampling lower bound we need to prove, where we use $[n]^{(k)}$ to denote the collection of all size-k subsets of $[n]$.

Theorem 1.2. Let $\mathcal{D}_{n,k}$ be the uniform distribution over

$$\left\{(X,Y) \mid X,Y \in [n]^{(k)}, X \cap Y = \varnothing\right\}.$$

Then for every $\omega(\log\log n) < k \le n/2$, there exists $\varepsilon = 2^{-\Omega(k^2/n)}$ such that $\mathsf{Samp}_{1-\varepsilon}(\mathcal{D}_{n,k}) = \Omega(k)$.

Note that the error bound is tight as $\Delta(\mathcal{U}\times\mathcal{U}, \mathcal{D}_{n,k}) \le 1 - 2^{-O(k^2/n)}$ for uniform distribution $\mathcal{U}$ over $[n]^{(k)}$. This also means that we cannot simply view the protocol as a convex combination of product distributions, and assort to the standard lower bound technique on statistical distance of convex combinations:

$$\Delta\Big(\sum\nolimits_{i\in[m]} p_i\mathcal{D}_i, \mathcal{D}\Big) \ge 1 - \sum\nolimits_{i\in[m]} \big(1 - \Delta(\mathcal{D}_i,\mathcal{D})\big) \tag{2}$$

since the it will only provide a lower bound of $\mathsf{Samp}_{1-\varepsilon}(\mathcal{D}_{n,k}) \ge \Omega(k^2/n)$, which in turn only gives a constant lower bound on the locality s when plugged into Theorem 1.1. Note that (2) was crucially used in the proof by [CGZ22] for $k = \Theta(n)$, and also in the sampling lower bound for permutation of [Vio20], whereas here we need to use some different techniques.

Our actual proof of Theorem 1.2 is similar in framework with the proof by [AST+03] for $k = \Theta(\sqrt{n})$. Their proof depends on the fact that every large rectangle $\mathcal{R} = \mathcal{X}\times\mathcal{Y}$ contains at least a constant fraction of intersecting pairs, which was proved by Babai, Frankl and Simon for their $\Omega(\sqrt{n})$ randomized communication lower bound of computing set disjointness [BFS86]. In the Lemma 3.1 below we prove a version of this fact for larger k, showing that the faction of disjoint pairs in $2^{-O(k)}$ large rectangles in $[n]^{(k)}\times[n]^{(k)}$ is exponentially small in k^2/n. Our proof is based on the work of Alon and Frankl [AF85], in which only $k = \Theta(n)$ was considered but the techniques easily generalize to arbitrary k.

Lemma 3.1. *For every constant $\sigma \in [0,1)$ the following holds. For every rectangle $\mathcal{R} = \mathcal{X}\times\mathcal{Y}$ with $\mathcal{X},\mathcal{Y} \subseteq [n]^{(k)}$ and $|\mathcal{X}|, |\mathcal{Y}| \ge \binom{n}{k}\cdot 2^{-\sigma k}$, we have*

$$\rho(\mathcal{R}) = \frac{|\mathcal{R}\cap \operatorname{supp}\mathcal{D}_{n,k}|}{|\mathcal{R}|} \le 2^{-\Omega(k^2/n)}.$$

Proof. Let $X_1,\dots,X_t$ be independent uniform samples from $\mathcal{X}$ for some t to be chosen later. Let S be the random variable that counts the number of $Y\in\mathcal{Y}$ disjoint from all X_i, that is,

$$S = \left|\left\{Y\in\mathcal{Y} \mid X_i\cap Y = \varnothing, \forall i\in[t]\right\}\right|.$$

By convexity of the function $x\mapsto x^t$, we have

$$\begin{aligned}\mathbb{E}[S] = \sum_{Y\in\mathcal{Y}} \Pr\left[X_i\cap Y = \varnothing, \forall i\in[t]\right] &= \sum_{Y\in\mathcal{Y}} \left[\rho(\mathcal{X}\times\{Y\})\right]^t \\ &\ge |\mathcal{Y}|\cdot\left[\frac{1}{|\mathcal{Y}|}\sum_{Y\in\mathcal{Y}}\rho(\mathcal{X}\times\{Y\})\right]^t = |\mathcal{Y}|\cdot\rho(\mathcal{R})^t. \end{aligned}\tag{3}$$

On the other hand, notice that when $|X_1 \cup \cdots \cup X_t| \geq n/2$, we have

$$S \leq \binom{\lfloor n/2 \rfloor}{k} \leq \binom{n}{k} \cdot 2^{-k} \leq |\mathcal{Y}| \cdot 2^{(\sigma-1)k}.$$

To bound probability of the complement event where $|X_1 \cup \cdots \cup X_t| < n/2$, we use union bound to first choose a set of size $\lfloor n/2 \rfloor$, and then pick every X_i from this set, which means that

$$\Pr\left[|X_1 \cup \cdots \cup X_t| < n/2\right] \leq \binom{n}{\lfloor n/2 \rfloor} \cdot \left[\frac{1}{|\mathcal{X}|}\binom{\lfloor n/2 \rfloor}{k}\right]^t \leq 2^n \cdot 2^{(\sigma-1)kt}.$$

Therefore we conclude that

$$\Pr\left[S > |\mathcal{Y}| \cdot 2^{(\sigma-1)k}\right] \leq 2^n \cdot 2^{(\sigma-1)kt}.$$

And since $S \leq |\mathcal{Y}|$, we can upper bound $\mathbb{E}[S]$ with

$$\begin{aligned}\mathbb{E}[S] &\leq |\mathcal{Y}| \cdot 2^{(\sigma-1)k} + |\mathcal{Y}| \cdot \Pr\left[S > |\mathcal{Y}| \cdot 2^{(\sigma-1)k}\right] \\ &\leq |\mathcal{Y}| \cdot \left(2^{(\sigma-1)k} + 2^n \cdot 2^{(\sigma-1)kt}\right). \end{aligned} \tag{4}$$

Combining (3) and (4) gives us

$$\rho(\mathcal{R})^t \leq 2^{(\sigma-1)k} + 2^n \cdot 2^{(\sigma-1)kt}.$$

Take $t = \frac{2n}{(1-\sigma)k}$, then we have $\rho(\mathcal{R})^t \leq 2^{(\sigma-1)k} + 2^{-n} \leq 2^{-(\sigma-1)k/2}$. Therefore $\rho(\mathcal{R}) \leq 2^{-\Omega(k/t)} \leq 2^{-\Omega(k^2/n)}$. □

Now, to prove Theorem 1.2 from Lemma 3.1, we need to decompose the product distributions from the communication protocol into uniform distributions over rectangles, which is done by decomposing any distribution (on each party) into uniform distributions.

Lemma 3.2. *Given any distribution $\mathcal{D}$ over the universe $[N]$, there are $m \leq 2\log N + 1$ uniform distributions $\mathcal{U}_1, \ldots, \mathcal{U}_m$ over subsets of $[n]$, and a distribution $(p_1, \ldots, p_m)$ over $[m]$ such that*

$$\Delta\Big(\mathcal{D}, \sum\nolimits_{i\in[m]} p_i \mathcal{U}_i\Big) \leq \frac{1}{N}.$$

Proof. For each $x \in [N]$, let d_x be the probability of x under $\mathcal{D}$. Then for $i \in \mathbb{N}_+$, we take $\mathcal{U}_i$ to be the uniform distribution over all x such that the i-th bit after decimal point in the binary representation of d_x is 1, as long as such x exists. It is easy to see that $\mathcal{D} = \sum_{i\in\mathbb{N}_+} 2^{-i}|\operatorname{supp}\mathcal{U}_i| \cdot \mathcal{U}_i$. Let $p_i = 2^{-i}|\operatorname{supp}\mathcal{U}_i|$ for $i \leq \log N$, and let

$$p = 1 - \sum_{i\leq 2\log N} p_i = \sum_{i>2\log N} 2^{-i}|\operatorname{supp}\mathcal{U}_i| \leq \sum_{i>2\log N} 2^{-i}N \leq \frac{1}{N}.$$

Then adding an arbitrary distribution $\mathcal{U}$ gives us

$$\Delta\Big(\mathcal{D}, \sum\nolimits_{i\leq 2\log N} p_i\mathcal{U}_i + p\cdot\mathcal{U}\Big) \leq p \leq \frac{1}{N}.$$ □

Proof of Theorem 1.2. Suppose we have a two-party communication protocol for sampling $\mathcal{D}_{n,k}$ with error $1-\varepsilon$ and transcript length at most $k/3$. Conditioned on any transcript in $\{0,1\}^{k/3}$, the outputs of the two parties are independent and thus generates a product distribution $\mathcal{D}_X \times \mathcal{D}_Y$. By Lemma 3.2, each of $\mathcal{D}_X$ and $\mathcal{D}_Y$ can be approximated by a convex combination of at most $2\log\binom{n}{k} + 1$ uniform distributions with error $1/\binom{n}{k}$, and thus $\mathcal{D}_X \times \mathcal{D}_Y$ can be approximated by a

convex combination of at most $\left(2\log\binom{n}{k}+1\right)^2 = O(k^2\log^2 n)$ uniform distributions over rectangles with error $2/\binom{n}{k} \le 2^{-k}$.

Therefore, altogether we have $M \le O(2^{k/3}k^2\log^2 n)$ rectangles $\mathcal{R}_1,\dots,\mathcal{R}_M$ and a distribution $(p_1,\dots,p_M)$ over $[M]$ such that

$$\Delta\Big(\mathcal{D}_{n,k}, \sum\nolimits_{i\in[M]} p_i\mathcal{R}_i\Big) \le 1-\varepsilon+2^{-k}, \tag{5}$$

where we abuse the notation and use $\mathcal{R}_i$ to also denote the uniform distribution over the rectangle $\mathcal{R}_i = \mathcal{X}_i\times\mathcal{Y}_i$. Let $I\subseteq[M]$ consist of indices i such that $|\mathcal{X}_i|,|\mathcal{Y}_i| \ge \binom{n}{k}\cdot 2^{-k/2}$, and let $\mathcal{R}_I = \sum_{i\in I} p_i\mathcal{R}_i \big/ \sum_{i\in I} p_i$. By Lemma 3.1 we have

$$\begin{aligned}
\Delta(\mathcal{D}_{n,k},\mathcal{R}_I) &\ge 1 - \Pr_{(X,Y)\sim\mathcal{R}_I}[X\cap Y=\varnothing] \\
&= 1 - \sum_{i\in I} p_i\cdot \Pr_{(X,Y)\sim\mathcal{R}_i}[X\cap Y=\varnothing] \Big/ \sum_{i\in I} p_i \\
&= 1 - \sum_{i\in I} p_i\cdot \rho(\mathcal{R}_i) \Big/ \sum_{i\in I} p_i \\
&\ge 1 - 2^{-\Omega(k^2/n)}.
\end{aligned}$$

And when $i\notin I$, if $|\mathcal{X}_i| < \binom{n}{k}\cdot 2^{-k/2}$ then

$$\Delta(\mathcal{D}_{n,k},\mathcal{R}_i) \ge 1 - \Pr_{(X,Y)\sim\mathcal{D}_{n,k}}[X\in\mathcal{X}_i] > 1-2^{-k/2},$$

as the distribution of X is uniform over $[n]^{(k)}$ in $\mathcal{D}_{n,k}$. The same bound holds for the case when $|\mathcal{Y}_i| < \binom{n}{k}\cdot 2^{-k/2}$ for the same reason, and therefore by inequality (2) (whose proof can be found in e.g. [CGZ22, Lemma 2]), we have

$$\Delta\Big(\mathcal{D}_{n,k}, \sum\nolimits_{i\in[M]} p_i\mathcal{R}_i\Big) \ge 1 - 2^{-\Omega(k^2/n)} - M\cdot 2^{-k/2}. \tag{6}$$

Since $k=\omega(\log\log n)$, we can find $\varepsilon = 2^{-\Omega(k^2/n)}$ that makes (5) and (6) contradict, which means that the communication protocol with transcript length $k/3$ does not exist. Therefore $\mathsf{Samp}_{1-\varepsilon}(\mathcal{D}_{n,k}) = \Omega(k)$. □

3.2 Sampling Equal k-Sets

In contrast to Theorem 1.2, the communication lower bound for sampling equal sets can be easily proved via the standard inequality (2).

Lemma 3.3. *Let $\mathcal{E}_{n,k}$ be the uniform distribution over*

$$\Big\{(X,X) \;\Big|\; X\in[n]^{(k)}\Big\}.$$

Then for every $k\le n^{1-\Omega(1)}$, there exists $\varepsilon = n^{-\Omega(k)}$ such that $\mathsf{Samp}_{1-\varepsilon}(\mathcal{E}_{n,k}) = \Omega(k\log n)$.

Proof. Consider any product distribution $\mathcal{X}\times\mathcal{Y}$ over $[n]^{(k)}\times[n]^{(k)}$. We can directly write out the statistical distance as

$$\Delta(\mathcal{E}_{n,k},\mathcal{X}\times\mathcal{Y}) = 1 - \sum_{Z\in[n]^{(k)}} \min\left\{\binom{n}{k}^{-1}, \Pr_{X\sim\mathcal{X}}[X=Z]\cdot\Pr_{Y\sim\mathcal{Y}}[Y=Z]\right\}.$$

We claim that there are at most $2 \cdot \binom{n}{k}^{2/3}$ such Z that the term in the summation is at least $\binom{n}{k}^{-4/3}$, since it would imply that either $\Pr[X = Z] \geq \binom{n}{k}^{-2/3}$ or $\Pr[Y = Z] \geq \binom{n}{k}^{-2/3}$. Therefore we have

$$1 - \Delta(\mathcal{E}_{n,k}, \mathcal{X} \times \mathcal{Y}) \leq 2 \cdot \binom{n}{k}^{2/3} \cdot \binom{n}{k}^{-1} + \binom{n}{k} \cdot \binom{n}{k}^{-4/3} = 3 \cdot \binom{n}{k}^{-1/3} \leq n^{-\Omega(k)},$$

where the last inequality is because $k \leq n^{1-\Omega(1)}$. Since given any communication protocol, conditioned on any transcript the output distribution is a product distribution, with inequality (2) we can conclude the lemma. □

Now recall the distribution $x \sim \mathcal{E}$ in Theorem 1.5: First sample r uniformly from $[n]^k$, and for every $i \in [n]$, independently and uniformly draw $j \in [k]$ and let $x_i = r_j$. In order to use Theorem 1.1, we fix two disjoint sets $I, I' \subseteq [n]$ each containing $3k$ coordinates. We define three sets of symbols in $[n]$ as

$$S = \{r_j \mid j \in [k]\}, \quad X = \{x_i \mid i \in I\}, \quad Y = \{x_i \mid i \in I'\}. \tag{7}$$

It is not hard to see that $S = X = Y$ with constant probability. However, we cannot simply restrict ourselves to this case as it does not provide lower bounds with error exponentially close to 1. Instead, we show that $X \cap Y$ is almost always large, and thus two parties given X and Y separately can independently sample two subsets that are equal with non-negligible probability.

Lemma 3.4. *With probability at least* $1 - 2^{-\Omega(k)}$, $|X \cap Y| \geq k/4$.

Proof. First we have

$$\Pr\left[|S| \leq k/2\right] \leq \binom{n}{k/2} \cdot (k/2)^k \cdot n^{-k} \leq \left(\frac{ek}{2n}\right)^{k/2} \leq 2^{-\Omega(k)}.$$

When $|S| \geq k/2$, we can identify $k/2$ different coordinates and symbols in r. The probability that X hits less than $3k/8$ of them (while missing at least $k/8$) is at most

$$\binom{k/2}{k/8} \cdot \left(1 - \frac{k/8}{k}\right)^{3k} \leq 2^{k/2} \cdot (7/8)^{3k} \leq 2^{-\Omega(k)}.$$

The same inequality holds for Y. And when both X and Y hits at least $3k/8$ different symbols from the set of size $k/2$, they have at least $k/4$ common ones, which happens with probability at least $1 - 2^{-\Omega(k)}$ by union bound. □

Corollary 3.5. *Let* $\mathcal{E}'_{n,k}$ *be the distribution of* (X, Y) *described in* (7). *Then for every* $k \leq n^{1-\Omega(1)}$, *there exists* $\varepsilon = 2^{-\Omega(k)}$ *such that* $\mathsf{Samp}_{1-\varepsilon}(\mathcal{E}'_{n,k}) = \Omega(k \log n)$.

Proof. First, suppose that we have the distribution $\mathcal{E}'_{n,k}$, and we use it to sample $\mathcal{E}_{n,k/4}$ as follows: Give the two sets X and Y separately to the two parties, each of whom samples a uniform subset of size $k/4$ from the set they receives (whenever possible) as the output. By Lemma 3.4, with probability at least $1 - 2^{-\Omega(k)}$, $|X \cap Y| \geq k/4$, conditioned on which the two subsets of size $k/4$ are equal with probability at least $\binom{k}{k/4}^{-1} = 2^{-O(k)}$ as $|X|, |Y| \leq k$. Notice that the sets of size $k/4$ are symmetric under this protocol. This means that with probability $2^{-O(k)}$, which is a probability independent from $\mathcal{E}'_{n,k}$, we output a distribution that is $2^{-\Omega(k)}$-close to $\mathcal{E}_{n,k/4}$.

Now replace $\mathcal{E}'_{n,k}$ with the output distribution of a communication protocol that samples $\mathcal{E}'_{n,k}$ with statistical distance at most $1 - \varepsilon$. Then we get a protocol, without any additional

communication, that with probability $2^{-O(k)}$ outputs a distribution which is $(1-\varepsilon+2^{-\Omega(k)})$-close to $\mathcal{E}_{n,k/4}$. The overall error is thus at most

$$1-2^{-O(k)}(\varepsilon-2^{-\Omega(k)})<1-n^{-o(k)},$$

if we properly choose $\varepsilon=2^{-\Theta(k)}$. Therefore by applying Lemma 3.3 on $\mathcal{E}_{n,k/4}$, we conclude that $\mathsf{Samp}_{1-\varepsilon}(\mathcal{E}'_{n,k})=\Omega(k\log n)$. □

Now Theorem 1.5 follows directly from Theorem 1.1 and Corollary 3.5 by choosing $h=\Theta(k/\log n)$.

Acknowledgments The authors would like to thank Ran Raz, Emanuele Viola and anonymous reviewers for helpful comments and discussions.

References

[AF85] Noga Alon and Peter Frankl. The maximum number of disjoint pairs in a family of subsets. *Graphs and Combinatorics*, 1:13–21, 1985. 6

[ALWZ21] Ryan Alweiss, Shachar Lovett, Kewen Wu, and Jiapeng Zhang. Improved bounds for the sunflower lemma. *Annals of Mathematics*, 194(3):795–815, 2021. 3, 4

[AST+03] Andris Ambainis, Leonard J. Schulman, Amnon Ta-Shma, Umesh V. Vazirani, and Avi Wigderson. The quantum communication complexity of sampling. *SIAM J. Comput.*, 32(6):1570–1585, 2003. 1, 2, 6

[BCW21] Tolson Bell, Suchakree Chueluecha, and Lutz Warnke. Note on sunflowers. *Discrete Mathematics*, 344(7):112367, 2021. 4

[BFS86] László Babai, Peter Frankl, and Janos Simon. Complexity classes in communication complexity theory. In *27th Annual Symposium on Foundations of Computer Science*, pages 337–347. IEEE Computer Society, 1986. 6

[BIL12] Chris Beck, Russell Impagliazzo, and Shachar Lovett. Large deviation bounds for decision trees and sampling lower bounds for AC^0-circuits. In *53rd Annual IEEE Symposium on Foundations of Computer Science, FOCS 2012*, pages 101–110. IEEE Computer Society, 2012. 1

[CGZ22] Eshan Chattopadhyay, Jesse Goodman, and David Zuckerman. The space complexity of sampling. In *13th Innovations in Theoretical Computer Science Conference, ITCS 2022*, volume 215 of *LIPIcs*, pages 40:1–40:23. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2022. 1, 2, 6, 8

[ER60] Paul Erdös and Richard Rado. Intersection theorems for systems of sets. *Journal of the London Mathematical Society*, 1(1):85–90, 1960. 4

[FLRS23] Yuval Filmus, Itai Leigh, Artur Riazanov, and Dmitry Sokolov. Sampling and Certifying Symmetric Functions. In Nicole Megow and Adam Smith, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2023)*, volume 275 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 36:1–36:21, Dagstuhl, Germany, 2023. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. 3

[GW20] Mika Göös and Thomas Watson. A lower bound for sampling disjoint sets. *ACM Trans. Comput. Theory*, 12(3):20:1–20:13, 2020. 1, 2

[LV12] Shachar Lovett and Emanuele Viola. Bounded-depth circuits cannot sample good codes. *Comput. Complex.*, 21(2):245–266, 2012. 1

[Ros14] Benjamin Rossman. The monotone complexity of k-clique on random graphs. *SIAM J. Comput.*, 43(1):256–279, 2014. 3

[RR18] Sivaramakrishnan Natarajan Ramamoorthy and Anup Rao. Lower bounds on non-adaptive data structures maintaining sets of numbers, from sunflowers. In *33rd Computational Complexity Conference, CCC 2018*, volume 102 of *LIPIcs*, pages 27:1–27:16. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2018. 4

[Vio12] Emanuele Viola. The complexity of distributions. *SIAM J. Comput.*, 41(1):191–218, 2012. 1, 3

[Vio16] Emanuele Viola. Quadratic maps are hard to sample. *ACM Trans. Comput. Theory*, 8(4):18:1–18:4, 2016. 1

[Vio20] Emanuele Viola. Sampling lower bounds: Boolean average-case and permutations. *SIAM J. Comput.*, 49(1):119–137, 2020. 1, 2, 3, 6

[Wat14] Thomas Watson. Time hierarchies for sampling distributions. *SIAM J. Comput.*, 43(5):1709–1727, 2014. 1

[Wat16] Thomas Watson. Nonnegative rank vs. binary rank. *Chic. J. Theor. Comput. Sci.*, 2016. 1

ECCC ISSN 1433-8092

https://eccc.weizmann.ac.il