Optimal streaming and tracking distinct elements with high probability

Jarosław Błasiok*

April 4, 2018

Abstract

The distinct elements problem is one of the fundamental problems in streaming algorithms — given a stream of integers in the range $\{1, \ldots, n\}$, we wish to provide a $(1 + \varepsilon)$ approximation to the number of distinct elements in the input. After a long line of research optimal solution for this problem with constant probability of success, using $\mathcal{O}(\frac{1}{\varepsilon^2} + \lg n)$ bits of space, was given by Kane, Nelson and Woodruff in 2010.

The standard approach used in order to achieve low failure probability δ, is to take a median of $\lg \delta^{-1}$ parallel repetitions of the original algorithm and report the median of computed answers. We show that such a multiplicative space blow-up is unnecessary: we provide an optimal algorithm using $\mathcal{O}(\frac{\lg \delta^{-1}}{\varepsilon^2} + \lg n)$ bits of space — matching known lower bounds for this problem. That is, the $\lg \delta^{-1}$ factor does not multiply the $\lg n$ term. This settles completely the space complexity of the distinct elements problem with respect to all standard parameters.

We consider also *strong tracking* (or *continuous monitoring*) variant of the distinct elements problem, where we want an algorithm which provides an approximation of the number of distinct elements seen so far, at all times of the stream. We show that this variant can be solved using $\mathcal{O}(\frac{\lg \lg n + \lg \delta^{-1}}{\varepsilon^2} + \lg n)$ bits of space, which we show to be optimal.

1 Introduction

Estimating the number of distinct elements in the data stream is one of the first, and one of the most fundamental problems in streaming algorithms. In this problem, we observe a *data stream*, i.e. a sequence of elements $x^{(1)}, x^{(2)}, \ldots, x^{(T)} \in \{1, \ldots n\}$, and we wish to provide a $(1+\varepsilon)$-approximation for the number of distinct elements in this sequence, using small space S. This can be trivially achieved with $\mathcal{O}(\min(n, T \lg n))$ bits of memory by either storing all elements encountered in the stream, or by storing a bitmask, keeping a single bit for every possible element of the universe. We wish to provide a probabilistic algorithm using significantly smaller space (allowing for small failure probability).

This problem was first studied by Flajolet and Martin in their seminal paper [FM83] in FOCS 1983, which started a long line of research on subsequently improved algorithms [AMS96, BJK+02, GT01, DF03, EVF06, FFGM07, Gib01].

Kane, Nelson and Woodruff in 2010 [KNW10] proposed an optimal algorithm for counting the number of distinct elements in the stream with failure probability $\frac{1}{3}$ — their algorithm provided an $(1+\varepsilon)$ approximation to the number of distinct elements using $\mathcal{O}(\frac{1}{\varepsilon^2} + \lg n)$ bits — the matching lower bound has been shown prior to this [Woo04, AMS96, BC09]. The standard black-box method of reducing the failure probability of estimation algorithm of this kind is to repeat it independently $\mathcal{O}(\lg \delta^{-1})$ times in parallel, and use the median of reported answers as the final estimation. This method, applied to the algorithm mentioned above, uses $\mathcal{O}(\lg \delta^{-1}(\frac{1}{\varepsilon^2} + \lg n))$ bits of space.

*John A. Paulson School of Engineering and Applied Sciences, Harvard University, 33 Oxford Street, Cambridge, MA 02138, USA. Email: `jblasiok@g.harvard.edu`. Supported by ONR grant N00014-15-1-2388.

On the other hand, Jayram and Woodruff in [JW13] developed a technique for proving lower bounds for streaming problems in the high success probability regime. Their technique allowed them to show that for number of natural streaming problems the naive repetition method is optimal — for example this is the case for estimation of the ℓ_p pseudonorm (with $p \in [0, 2]$) of frequency vector in so called *strict turnstile* streaming model. In the same paper they proved a lower bound for the distinct elements problem of form $\Omega(\frac{\lg \delta^{-1}}{\varepsilon^2})$. For constant ε, this left a gap between an upper bound $\mathcal{O}(\lg \delta^{-1} \lg n)$ and lower bound $\Omega(\lg n + \lg \delta^{-1})$.

It was known that one should *not* expect a lower bound $\Omega(\lg \delta^{-1} \lg n)$ for this problem. Already [KNW10] showed that for some constant ε, one can achieve failure probability $\delta = \frac{1}{\operatorname{poly} \lg n}$ using only $\mathcal{O}(\lg n)$ bits, and in [JW13] it was observed that for every constant ε there is an algorithm using $\mathcal{O}(\lg n)$ bits with failure probability $\delta = \frac{\lg \lg n}{\lg n}$. In this paper we completely resolve the question about space complexity of the distinct elements problem in the high success probability regime, showing that the Jayram, Woodruff lower bound was optimal.

Continuous monitoring Recently, the space complexity of *tracking* problems in data streams has been considered — namely we say that streaming algorithm provides *strong tracking* of a statistic f of the input stream, if after every update it reports quantity $\tilde{f}$ such that

$$\mathbb{P}(\forall t, (1-\varepsilon) f^{(t)} \leq \tilde{f}^{(t)} \leq (1+\varepsilon) f^{(t)}) > 1 - \delta.$$

The first result of this form that we are aware of, was proven in [KNW10] as a subroutine for non-tracking estimation of distinct elements. They showed that one can achieve tracking of F_0 with some constant approximation factor, using $\mathcal{O}(\lg n)$ bits of space. The question whether one can achieve strong tracking without the naive union bound over all positions of the stream was explicitly asked later in [HTY14], where they also proposed an algorithm for estimation of the ℓ_p-pseudonorm of the frequency vector, for $p \in (0, 2]$. Their algorithm yield improvement over the baseline approach for very long input streams $\lg T = \omega(\lg n)$. The strong tracking of ℓ_2 has been later improved in [BCWY16, BCI$^+$17], where interesting results are achieved even in the more standard regime of parameters, with n and T that are polynomially related. They showed that one can solve strong tracking of ℓ_2 using $\mathcal{O}(\frac{\lg n \lg \lg T}{\varepsilon^2})$ bits, as compared to naive bound of form $\mathcal{O}(\frac{\lg n \lg T}{\varepsilon^2})$. The improved algorithm for strong tracking of ℓ_p with $0 < p < 2$ was provided in [BDN17].

Our contribution. We provide an optimal streaming algorithm for distinct elements problem in the high probability regime, using $\mathcal{O}(\lg n + \frac{\lg \delta^{-1}}{\varepsilon^2})$ bits of space. This result completely settles the space complexity of this problem with respect to all standard parameters.

We also show a strong tracking algorithm for distinct elements with space $\mathcal{O}(\lg n + \frac{\lg \delta^{-1} + \lg \lg n}{\varepsilon^2})$, together with a matching lower bound — we prove that $\Omega(\frac{\lg \lg n}{\varepsilon^2})$ term is necessary. The $\Omega(\frac{\lg \delta^{-1}}{\varepsilon^2} + \lg n)$ lower bound was already known even for easier non-tracking version of Distinct Elements problem.

This is a first matching lower bound for any strong tracking problem, where the non-trivial algorithm is achievable. This shows a separation between the traditional estimation problem and strong tracking variation when $\varepsilon = o(\sqrt{\frac{\lg \lg n}{\lg n}})$. On the other hand, in the regime $\varepsilon = \Omega(\sqrt{\frac{\lg \lg n}{\lg n}})$ the strong tracking problem is not harder than one-shot estimation (up to constant factors).

The update time of our algorithm is $\operatorname{poly}(\lg n, \lg \delta^{-1})$. The only bottleneck is the pseudorandom construction described in Section 8. In particular, by substituting this construction with a random walk over an expander graph of super-constant degree, it is possible to achieve update time $\mathcal{O}(\lg \delta^{-1} + \lg \lg n)$, with slightly worse space complexity $\mathcal{O}(\frac{\lg \delta^{-1}}{\varepsilon^2} + \lg \delta^{-1} \lg \lg n + \lg n)$.

2 Notation

For a natural number m, by $[m]$ we denote set $\{1, \ldots m\}$. For a finite set A, by $\#A \in \mathbb{Z}$ or $|A|$ we denote the cardinality of A. For $X \in \Sigma$, where $|\Sigma| = 2^n$, we will write $\bar{X} \in \{0, 1\}^n$ to be a bit representation of X.

For a bitvector $y \in \{0,1\}^n$ we denote $|y| = \#\{i : y_i = 1\}$. For two bitvectors x, y, we take $x \vee y \in \{0,1\}^n$ to be the bitvector with $(x \vee y)_j = 1$ if and only if $x_i = 1$ or $y_i = 1$.

In the paper n will be used to denote the size of the universe from which the elements in the input stream are chosen, T — the length of the stream, and $x^{(1)}, x^{(2)}, \ldots x^{(t)} \in [n]$ are those elements. Set $S^{(t)} := \{x^{(1)}, \ldots x^{(t)}\}$ is the set of all distinct elements seen up to a time step t, and $F_0^{(t)} := \#S^{(t)}$.

Throughout the paper we use notation $A \lesssim B$, to denote the existence of an absolute constant C such that $A \leq CB$, where A and B themselves may depend on a number of parameters.

3 Overview of our approach

3.1 Constant factor approximation with high probability

The main goal of Section 4 is to show a streaming algorithm that provides an $\mathcal{O}(1)$-approximation to the number of distinct elements at all times in the stream (i.e. $\mathcal{O}(1)$-strong tracking), with probability $1 - \delta$ using optimal space $\mathcal{O}(\lg n + \lg \delta^{-1})$ bits. That is, we want to provide estimate $\tilde{F}_0(t)$, such that

$$\mathbb{P}(\forall t,\ \frac{1}{C} F_0^{(t)} \leq \tilde{F}_0^{(t)} \leq C F_0^{(t)}) > 1 - \delta$$

where $F_0^{(t)}$ is a number of distinct elements on the input among $x_1, \ldots x_t$.

Note that in this regime of parameters, if one has an algorithm estimating number of distinct elements using space complexity $\mathcal{O}(\lg n + \lg \delta^{-1})$, one can set $\delta = \delta'/n$, and apply a union bound over all insertions to the stream, to get a strong tracking algorithm for the same problem with failure probability δ' and space complexity $\mathcal{O}(\lg n + \lg \delta^{-1})$. As such, we can without loss of generality focus on the strong tracking version, and this stronger guarantee is going to be useful in order to ensure that the algorithm can be implemented using small space.

To discuss the main idea behind our approach, for the sake of presentation we will first consider a random oracle model — here we assume that the algorithm is augmented with the access to a uniformly random function $R : \{0,1\}^* \to \{0,1\}$ (all the values of R are uniform and independent); in particular the space to store such a function does not count towards the space complexity of the algorithm, and the failure probability is understood over a selection of the oracle. For space complexity of such an algorithm, we will count only the amount of information passed between observations of elements from the input; we can use larger space to process an element from the input. This allows us to talk in a meaningful way about space complexity $o(\lg n)$, even though any single element in the stream already take $\Theta(\lg n)$ bits to store.

Let us start with discussion on how to design an algorithm using $\mathcal{O}(\lg \lg n + \lg \delta^{-1})$ bits of space in the random oracle model. It is well-known that given access to a random hash function $h : [n] \to \{0,1\}^{\lg n}$, if we fix some set $S \subset [n]$, then $\hat{X} := \max_{s \in S} \mathrm{lsb}(h(s))$ is such that $2^{\hat{X}}$ is a constant factor approximation to $|S|$ with probability $2/3$, where lsb is the least-significant-bit function [FM83]. Indeed, to argue that this is true, we can consider subsets $S_k \subset S$ given by $S_k := \{s \in S : \mathrm{lsb}(h(s)) \geq k\}$ — every such subset corresponds to sub-sampling S by a factor 2^k, and we should expect that the last non-empty set S_k is the one corresponding to sub-sampling by a factor roughly $\frac{1}{|S|} = 2^{-\lg |S|}$. By repeating this estimator $\mathcal{O}(\lg \delta^{-1})$ times in parallel and taking median, one can achieve $\mathcal{O}(\lg \delta^{-1} \lg \lg n)$ bit complexity for failure probability δ under the random oracle model. To improve this construction to $\mathcal{O}(\lg \lg n + \lg \delta^{-1})$ bits, let us take $w := \mathcal{O}(\lg \lg n + \lg \delta^{-1})$ independent estimators as above. Instead of storing all those estimators $\hat{X}_k$ independently, we can store the median (which takes $\mathcal{O}(\lg \lg n)$ bits), and deviations $\hat{X}_i - \mathrm{median}(\{\hat{X}_1, \ldots \hat{X}_w\})$. One can show that with high probability at all times throughout the stream the median is a good estimator of the number of distinct elements seen so far, and moreover — because the deviations $\hat{X}_i - \lg |S|$ are random variables that are extremely well concentrated around zero — on average over all the counters we will use constant number of bits per counter to store all deviations from median, at all time steps.

Getting rid of the random oracle assumption is much more technical — without access to the random oracle, it is known ([AMS96]) that one can use a pairwise independent hash function $h : [n] \to \{0,1\}^{\lg n}$

to get a constant success probability — and a seed to such a hash function can be stored using $\mathcal{O}(\lg n)$ bits. This, together with median over parallel repetitions of the estimator, yields simple $\mathcal{O}(\lg n \lg \delta^{-1})$ space algorithm with failure probability δ.

To improve upon that, we can observe that in this setting it is not necessary for all the $w = \mathcal{O}(\lg \lg n + \lg \delta^{-1})$ estimators to use independent seeds for the underlying pairwise-independent hash functions h_i. Instead, we can consider a fully explicit constant degree expander graph, with set of vertices $[N]$ corresponding to the set of seeds for pairwise independent hash functions. We would choose the first seed for h_1 uniformly at random, but subsequent seeds are chosen by a random walk over this expander graph. In such a way, we can succinctly store all the seeds using $\mathcal{O}(\lg \delta^{-1} + \lg n)$-bits of space, and the standard Chernoff-bounds for expander walks [Vad12, Theorem 4.22] imply that median of estimators generated in such a way is still constant factor approximation for the number of distinct elements, except with small failure probability δ. This yields an algorithm with space complexity $\mathcal{O}(\lg n + \lg \lg n \lg \delta^{-1})$, if we store $\hat{X}_i$ explcitly — still falling short of our goal of $\mathcal{O}(\lg n + \lg \delta^{-1})$ bits of space.

Unfortunately, we *cannot* argue, as before in the random oracle model, that we can succinctly store all counters $\hat{X}_i$ generated via such an expander walk by considering the median and deviations from the median separately — sufficiently strong concentration bounds are *not true* for a constant degree expander walk.

Instead, inspired by the construction of a sampler in [Mek17], we show that by composing a number of pseudorandom objects (i.e. pairwise independent hash functions, short walks over super-constant expander graphs, averaging samplers obtained from celebrated construction of strong extractors [Zuc97], and standard sub-sampling methods), we can generate about $\frac{w}{\lg w}$ groups of estimators, such that each group has about $\lg w$ estimators, and with probability $1 - \exp(-\Omega(w)) = 1 - \delta$ for at least half of the groups the median yields a good estimation of F_0 at all times, while simultaneously the "good" groups take at all times take $\mathcal{O}(1)$-bits on average per estimator to store (if we store estimators within a group by storing just the deviations from the median).

It is essential for this argument that size of each group $w_1 = \mathcal{O}(\lg w)$ is greater than $C \lg \lg n$ — intuitively, if we consider a random group of such a size, probability that we need too many bits to store compactly such a group at any fixed step t is bounded by $\exp(-\Omega(w_1)) = \frac{1}{\mathrm{polylog}(n)}$, and therefore we can union bound over all $\mathcal{O}(\lg n)$ positions where F_0 grows by factor of two, without affecting the failure probability too much.

The details of this pseudorandom construction are presented in Section 8. This is the main technical difficulty in proving the following theorem.

Theorem 1. *There is a streaming algorithm with space complexity $\mathcal{O}(\lg n + \lg \delta^{-1})$ bits, that with probability $1 - \delta$ reports a constant factor approximation to number of distinct elements in the stream after every update.*

This space complexity is optimal [AMS96, JW13].

The algorithm could be significantly simplified, and would mimic exactly the algorithm in the random oracle scenario, if we had an explicit sampler $\Gamma : \{0,1\}^s \times [w] \to [n]$ satisfying that for any $f : [n] \to \mathbb{R}_+$ with $\mathbb{P}_{X \sim \mathrm{Unif}([n])}(f(X) > \lambda) \leq \exp(-e^{\lambda})$, we have $\mathbb{P}_{S \sim \mathrm{Unif}_s}(\sum_{i \leq w} f(\Gamma(S, i)) > Cw) < \exp(-\Omega(w))$, with seed length $s = \mathcal{O}(\lg N + w)$. An object satisfying similar guarantees for a wider class of all $f(X)$ with subgaussian tails could potentially be useful in other applications as well. [1]

Remark 2. *The update time of this algorithm is $\mathcal{O}(\mathrm{poly}(\lg \delta^{-1}, \lg n))$. The only bottleneck is the pseudorandom construction we are using. If we give up on succinctly storing estimates $\hat{X}_i^{(t)}$, and store them explicitly, we can replace this pseudorandom construction with a single random walk over constant degree expander graph. There are expander graphs that allow evaluation of the neighbour function in constant time [GG81]. Such a modification would give an algorithm using slightly worse space $\mathcal{O}(\lg n + \lg \delta^{-1} \lg \lg n)$, but $\mathcal{O}(\lg \delta^{-1} + \lg \lg n)$ update time for strong tracking F_0 with constant factor approximation.*

It is possible to carry this construction over to our subsequent result, achieving $\mathcal{O}(\lg \delta^{-1}(\frac{1}{\varepsilon^2} + \lg \lg n) + \lg n)$ bits of space for high accuracy regime, and $\mathcal{O}(\lg \delta^{-1} \frac{\lg \lg n}{\varepsilon^2} + \lg n)$ bits of space for tracking, all with update time $\mathcal{O}(\lg \lg n + \lg \delta^{-1})$.

[1] Non-constructive existence of samplers like that can be proven using probabilistic method, and reduction to γ-Strong samplers similar to Lemma 35.

3.2 High accuracy regime

In Section 5 we discuss how to use the previous construction to achieve a high accuracy estimation of number of distinct elements, with probability $1-\delta$. We prove the following theorem.

Theorem 3. *For every ε, δ there is an algorithm using $\mathcal{O}(\frac{\lg \delta^{-1}}{\varepsilon^2} + \lg n)$ bits, which, at the end of the stream reports $1+\varepsilon$ approximation to the number of distinct elements, with probability $1-\delta$.*

Remark 4. *[AMS96, JW13] This space complexity is optimal — every algorithm that estimates number of distinct elements up to $(1+\varepsilon)$ factor with probability $1-\delta$ needs to use space at least $\Omega(\frac{\lg \delta^{-1}}{\varepsilon^2} + \lg n)$ bits.*

In fact, given ideas in the work of Kane *et al.* [KNW10] and results obtained in the previous section, getting correct dependence on the error parameter ε is routine, although somewhat tedious.

We consider separately two ranges of parameters: if $\varepsilon < \left(\frac{1}{\lg n}\right)^{1/4}$, the KNW algorithm (given access as a black-box to the strong tracking with some constant approximation), using space $\mathcal{O}(\frac{1}{\varepsilon^2})$ and $\mathcal{O}(\lg n)$ random bits has probability $\frac{2}{3}$ of providing $(1+\varepsilon)$ approximation to the number of distinct elements. We can instantiate $\mathcal{O}(\lg \delta^{-1})$ parallel copies of this algorithm, providing them access to the same strong-tracker with failure probability $\delta/2$. To reduce the amount of randomness necessary for all those runs (naively, we would have to store $\mathcal{O}(\lg n \lg \delta^{-1})$ random bits), we pick them using a walk over a constant-degree expander graph. That is, random bits for first instance of a KNW algorithm are completely uniform, but bits for subsequent runs of KNW are chosen by following a random edge in an expander graph. We can use standard Chernoff-bounds for expander walks, as in [Gil98], to show that failure probability of such an algorithm is still at most δ.

On the other hand, if $\varepsilon > \left(\frac{1}{\lg n}\right)^{1/4}$, we can assume without loss of generality that $\lg \delta^{-1} > \Omega(\sqrt{\lg n})$, because in this case $\frac{\sqrt{\lg n}}{\varepsilon^2} = o(\lg n)$ anyway, and our target is space complexity of form $\mathcal{O}(\frac{\lg \delta^{-1}}{\varepsilon^2} + \lg n)$. In this case, we can instantiate $\mathcal{O}(\lg\lg n + \lg \delta^{-1})$ parallel copies of a KNW algorithm, using the pseudorandom construction as described in Section 8. Identical analysis as for a constant approximation factor in Section 4 can be used to deduce correctness of such an approach.

In fact the space guarantees of the KNW algorithm, as it was originally analyzed, applied only when $\varepsilon \le \sqrt{\lg n}$ — as this could be assumed without loss of generality in the original setting. We provide a more delicate analysis of the space consumption of this algorithm in Appendix A (specifically Theorem 12), that is sufficient for our purposes.

3.3 Strong tracking of distinct elements

In Section 6 we discuss how to achieve the $(1+\varepsilon)$-strong tracking guarantee for F_0 estimation.

First, let us observe that an algorithm estimating F_0 with small failure probability already translates into some upper bound on the space complexity for the tracking problem. Given that number of distinct elements in the stream is increasing, and our estimators proposed in Section 5 are monotone as well, it is enough to look at a sequence of positions $t_1, t_2, \ldots t_s$ such that $F_0^{(t_i)} = (1+\varepsilon)^i$. If the estimate is within $(1+\varepsilon)$ from the actual number of distinct elements at all points t_i, we can deduce a strong tracking with accuracy $1+\mathcal{O}(\varepsilon)$: for $t_i \le t \le t_{i+1}$ we have $\hat{F}_0^{(t)} \le \hat{F}_0^{(t_2)} \le (1+\varepsilon)F_0^{(t_{i+1})} \le (1+\varepsilon)^2 F_0^{(t)}$, and similarly. There are at most $\lg_{1+\varepsilon} n = \mathcal{O}(\frac{\lg n}{\varepsilon})$ such positions t_i, so by setting failure probability $\delta := \delta' \left(\frac{\lg n}{\varepsilon}\right)^{-1}$ in Theorem 3, we can deduce that there is an algorithm satisfying $1+\varepsilon$ strong tracking of F_0 with probability $1-\delta$, using $\mathcal{O}(\frac{\lg\lg n + \lg \delta^{-1}}{\varepsilon^2} + \lg n + \frac{\lg \varepsilon^{-1}}{\varepsilon^2})$ bits of space.

We show that, by opening up the [KNW10] construction and more detailed analysis, it is possible to remove the additive $\frac{\lg \varepsilon^{-1}}{\varepsilon^2}$ term, and obtain an optimal algorithm for F_0 tracking.

Theorem 5. *There is an algorithm for $1+\varepsilon$ strong tracking of the number of distinct elements in the stream, using $\mathcal{O}(\frac{\lg\lg n + \lg \delta^{-1}}{\varepsilon^2} + \lg n)$ bits of space.*

To describe the overview of our contribution, let us first discuss the high-level idea behind the [KNW10] algorithm. Let us focus, for the sake of this discussion, on the random oracle model. Consider a fixed set $S \subset [n]$ (the set of distinct elements seen at the end of the stream), a random hash function $h : [n] \to \{0,1\}^{\lg n}$, and sets $S_k := \{s \in S : \mathrm{lsb}(s) \geq k\}$ — those sets correspond roughly to subsampling S by a factor of 2^k. If we already have access to a constant factor approximation of $|S|$, we can zoom in onto set S_k for which we expect $|S_k| = \Theta(\frac{1}{\varepsilon^2})$. Clearly $2^k|S_k|$ is an unbiased estimator of $|S|$, and moreover, we can relatively easily see that the standard deviation of $2^k|S_k|$ is of order $\mathcal{O}(\varepsilon|S|)$. This implies that if we had a way to estimate size of $|S_k|$ up multiplicative factor $(1+\varepsilon)$ that would be enough to get an $(1+\mathcal{O}(\varepsilon))$ approximation for $|S|$.

In order to do this, we can check a hash function $h_2 : [n] \to [P]$ for $P \approx \frac{100}{\varepsilon^2}$. We wish to recover $|S_k|$ from $|h_2(S_k)|$. This is reminiscent of a famous balls-and-bins thought experiment: we are throwing $|S_k|$ balls randomly into P bins, and we try to estimate number of balls, given number of non-empty bins. Let us define $\Phi_P(t)$ to be the expected number of non-empty bins, after throwing t balls at random into P bins, we have $\mathbb{E}_{h_2} |h_2(S_k)| = \Phi_P(|S_k|)$. We claim that, as long as $|S_k| \leq \frac{P}{20}$, we should expect that $|S_k| \approx_{1+\varepsilon} \Phi^{-1}(|h_2(S_k)|)$. This is because $\sqrt{\mathrm{Var}(|h_2(S_k)|)} = \Theta(\sqrt{|S_k|}) = \Theta(\varepsilon|S_k|)$, so $|h_2(S_k)| = \Phi(|S_k|) \pm \mathcal{O}(\varepsilon|S_k|)$, but Φ_P is bi-Lipschitz in the regime $t \leq P/20$ — i.e. for any $a \leq b \leq P/20$ we have $0.9(b-a) \leq \Phi(b-a) \leq b-a$. Those facts put together justify the claim that $|S| \approx_{1+\varepsilon} 2^k \Phi^{-1}(|h_2(S_k)|)$.

Using about $\frac{\lg \lg n}{\varepsilon^2}$ bits, we can keep track of $|h_2(S_k)|$ for all k — it is enough for each $p \in [P]$ to store $\max\{\mathrm{lsb}(s) : s \in S, h_2(s) = t\}$. In [KNW10] it is discussed, among other things, how to reduce the space complexity to $\mathcal{O}(1/\varepsilon^2)$-bits, and how to remove the random oracle assumption, by using compositions of bounded-wise independent hash functions. We describe this algorithm in Appendix A, together with more detailed analysis of the distribution of space complexity of this algorithm.

In order to achieve smaller space of the tracking algorithm, let us focus on a specific k and consider evolution of $|S_k^{(t)}|$ over the updates to the stream, where $S^{(t)} := \{x_1, \dots x_t\}$, and $S_k^{(t)} := \{x \in S^{(t)} : \mathrm{lsb}(s) \geq k\}$. More specifically, let us take $K := 2^k \varepsilon^{-2}$, and let us look at the stream, given the promise that $|S^{(T)}| < K/100$. We wish to say that with probability $2/3$ simultaneously all times $|S_k^{(t)}|2^k$ gives us an approximation of $|S^{(t)}|$ up to additive term $\pm\varepsilon K$. Moreover, we want to say that $\Phi^{-1}(|h_2(S_k^{(t)})|)$ yields at all time approximation to $|S_k^{(t)}|$, again with additive error $\pm\varepsilon K 2^{-k}$. We could amplify this success probability to $1 - \Theta(\lg n)$ using $\mathcal{O}(\lg \lg n)$ repetitions of the whole algorithm, and union bound over all possible settings of k. Note that there are only $\mathcal{O}(\lg n)$ values of k to union bound over, as opposed to $\mathcal{O}(\varepsilon^{-1} \lg n)$ distinct positions in the stream where $F_0^{(t)}$ grows by a multiplicative factor of $(1+\varepsilon)$.

In the random oracle model both facts — the fact that for all t we have $S_k^{(t)} 2^k \simeq |S^{(t)}| \pm \varepsilon K$, as well as the fact that for all t, we have $\Phi^{-1}(h_2(S_k^{(t)})) = S_k^{(t)} \pm 2^{-k}\varepsilon K$ can be proven by the Doob's martingale inequality. In particular, the fact that $2^k|S_k^{(t)}|$ is an approximation to $|S^{(t)}|$ at all times t, follows directly (after shifting and rescaling) from the fact that for a random walk $Y_t := \sum_{i \leq t} X_i$, where X_i are arbitrary random variables satisfying $\mathbb{E} X_i = 0$ and $\mathbb{E} X_i^2 = 1$, we have $\sup_{t \leq T} |Y_t| = \mathcal{O}(\sqrt{T})$ with good probability. The main technical difficulty in the strong tracking part of this paper lies in dropping the random oracle assumption, and showing some variation on Doob's martingale inequality under bounded independent hash functions. In particular, we show the following lemma on the deviations of random walk that might be of independent interest

Lemma 6. *Let $X_1, \dots X_T$ be collection of 4-wise independent random variables, with $\mathbb{E} X_i = 0$, and $\mathbb{E} X_i^2 = 1$, and let $Y_t := \sum_{i \leq t} X_i$, then*

$$\mathbb{P}(\sup_{t \leq T} |Y_t| > \lambda) \lesssim \frac{T}{\lambda^2}.$$

A result of the same spirit can be deduced from [BCI+17, Theorem 10] when X_i are uniform ± 1 random variables — in our case, however, the steps X_i are significantly less well-behaved, i.e. $\mathbb{E} X_i^4$ is already extremely large, even compared to T.

Lemma 6 already implies the first part of the argument: that for all t we have $2^k|S_k^{(t)}| = |S^{(t)}| \pm \mathcal{O}(\varepsilon K)$. To control deviations of $h_2(|S_k^{(t)}|)$ from its expectation, we use h_2 to be a composition of pairwise independent hash function $h_3 : [n] \to [K^2]$, and poly($\lg \frac{1}{\varepsilon}$)-wise independent function $h_4 : [P^2] \to [P]$. We should expect that $|h_3(S_k^{(T)})| = |S_k^{(T)}|$, i.e. function h_3 has no collisions with probability $9/10$, and we care about deviations of $|h_4(\tilde{S}^{(t)})|$, where $\tilde{S}^{(t)} \subset [P^2]$ is such that $|\tilde{S}^{(t)}| = |S^{(t)}|$.

Consider $\phi : [P]^{P/20} \to \mathbb{N}$, such that $\phi(r_1, r_2, \dots r_{P/20}) := \#\{j : \exists i,\ r_i = j\}$ — in the random oracle model, bounding the deviations $|h_4(\tilde{S}^{(t)})| - \mathbb{E}\,|h_4(\tilde{S}^{(t)})|$ can be reduced to bounding the deviations of the Doob's martingale $Y_t := \mathbb{E}_{\hat{X}}\, \phi(r_1, \dots r_t, \hat{X}_{t+1}, \dots \hat{X}_T)$, where $\hat{X} \sim \mathrm{Unif}([P])$. In this setting the Doob's martingale inequality yields

$$\mathop{\mathbb{P}}_{r_1, \dots r_{P/20} \sim \mathrm{Unif}([P])} \left(\sup_{t \leq P/20} |Y_t - \mathbb{E}\, Y_t| > \lambda \right) \leq \frac{\mathrm{Var}\, \phi}{\lambda^2} \tag{1}$$

where x_i and $\hat{X}$ are independent and uniform. Finally, this together with bi-Lipschitz property of function Φ_K in the range of interest, implies that indeed we have $\forall t, \Phi^{-1}(|h_4(h_3(S_k^{(t)}))|) = |S_k^{(t)}| \pm \mathcal{O}(\varepsilon K)$.

In our case, variables r_i have only bounded-wise independence, and the process Y_t above is no longer a martingale. We deal with this, by showing that ϕ can be approximated (in some sense, under the distributions of interest) by a poly($\lg P$)-degree polynomial, and we show that under some additional restrictions, processes as above induced by degree d polynomials, satisfy the same Eq. (1), even if variables r_i are only 4-wise independent.

3.4 Strong tracking lower bound

In Section 7 we show the optimality of the strong tracking algorithm proposed in the previous section. We prove the following theorem.

Theorem 7. *Every algorithm solving $1+\varepsilon$ strong tracking for F_0 estimation with probability $\frac{2}{3}$ needs to use $\Omega(\frac{\lg \lg n}{\varepsilon^2})$ bits of space.*

Together with previously known lower bound $\Omega(\frac{\lg \delta^{-1}}{\varepsilon^2} + \lg n)$ for F_0 estimation, this shows a lower bound that exactly matches our upper bound discussed earlier.

In order to show this, we introduce a k-round communication game, where at round k, Alice observes input $x_k \in \{0,1\}^n$, Bob observes input $y_k \in \{0,1\}^n$, and they all observe all the previous inputs $(x_i, y_i)_{i \leq k-1}$. In the k-th round, Alice sends a message to Bob, and Bob is supposed to report $(1+\varepsilon)$ approximation to number of ones in a string $x_k \vee y_k$ (with element-wise or). The protocol is successful if and only if simultaneously at all rounds Bob reports correct (approximate) answer. We show that existence of a strong tracking algorithm implies low-communication protocol for this kind of game with $k = \Theta(\lg n)$ rounds — which in turn, implies a one-round one-way communication protocol for estimation of $x \vee y$ with small failure probability $\delta = \Theta(1/k)$. This would contradict known communication complexity lower bound for small failure probability of distinct element counting [JW13].

3.5 Pseudorandom construction

In Section 8, we prove the main derandomization lemma used in the algorithm described in Section 4 for constant factor approximation of the number of distinct elements. Take $w = \Theta(\lg \delta^{-1} + \lg \lg n)$ — we wish to use w instantiations of the basic estimator, each instantiation is uniquely determined by a seed to pairwise independent hash function used for the estimator (such a seed is of length $\mathcal{O}(\lg n)$, let us call the number of different seeds N).

We pick $w_2 := \Theta(\lg w)$, a random walk of length w_2 over an expander graph with vertices $[N]$ and degree quasipoly(w_2) will be *bad* with probability $\exp(-\Omega(w_2))$ — by which we mean that either the median of all the estimators produced by this walk is at some point far from actual F_0, or that we need more than Cw_2 bits to store all the values of the estimators by storing median and deviations from median). A single random walk like this is going to need $\mathcal{O}(\lg n + w_2 \lg^2 w_2)$ random bits. If we consider now space of all those random walks, we can use known construction of averaging samplers to get a sample of size $W = \mathrm{poly}(w)$, such that with probability $1 - \exp(-\Omega(w))$ the fraction of failed random walks is the same as in the entire space. If we condition on the event that the sampler succeeded, by taking $\frac{w}{\lg w}$ independent elements from the sample $[W]$, we can see that at more than half of them is bad with probability $\exp(-cw_1)^{w/\lg w} = \exp(-cw)$. As we are taking $\frac{w}{\lg w}$ independent elements from the universe of size $\mathcal{O}(\mathrm{poly}(w))$ we need only $\frac{w}{\lg w}\mathcal{O}(\lg w) = \mathcal{O}(w)$ random bits to achieve this.

4 Constant factor approximation with high probability

In this section we prove Theorem 1, assuming existence of specific pseudo-random objects described in Lemma 10. The proof of Lemma 10 itself is postponed later in Section 8. We will first state few necessary definitions, followed by a statement of this lemma, then we proceed with the proof of Theorem 1.

Definition 8 (doubly-exponential tail)**.** *We say that a function $f : [M] \to \mathbb{R}_+$ satisfies doubly-exponential tail bounds if $\mathbb{P}_{s\sim \mathrm{Unif}([M])}(f(s) > \lambda) < \exp(-e^{\lambda})$.*

Definition 9 (C-small set)**.** *Consider a finite universe $[M]$, and equipped with R functions $g_1, \dots g_R : [M] \to \mathbb{R}_+$. We will say that a sequence $S \in [M]^*$ is C-small with respect to $g_1, \dots g_R$, if $\forall t \leq R$, $\sum_{X\in S} g_t(X) \leq C|S|$.*

Equipped with those definitions, we are ready to state the pseudorandom lemma.

Lemma 10. *For any $w \geq K_0 R^{1/2}$, there exist w_1, w_2 with $w_1 w_2 = \Theta(w)$, and an explicit function $\Xi : \{0,1\}^s \times [w_1] \to [M]^{w_2}$, such that for any $g_1, \dots g_R : [M] \to \mathbb{R}_+$ with doubly-exponential tail bounds, we have with probability at least $1 - \exp(\Omega(w))$ over a random selection of the seed $U \in \{0,1\}^s$, that majority of sequences $\Xi(U, k)$ is C-small for some universal constant C.*

That is

$$\exists \Xi, \forall g_1, \dots g_R, \underset{U\sim \mathrm{Unif}(\{0,1\}^s)}{\mathbb{P}}(\#\{k : \Xi(U,k) \text{ is } C\text{-small}\} > w_1/2) \geq 1 - \exp(-\Omega(w)),$$

where $g_1, \dots g_R$ above must satisfy doubly-exponential tail bounds.

The seed length in this construction is $s = \mathcal{O}(\lg M + w)$, and Ξ can be evaluated in time and space $\mathrm{poly}(s)$.

Let us now proceed with the proof of Theorem 1.

Fix a stream of updates $x^{(1)}, x^{(t)}, \dots x^{(T)} \in [n]$, and corresponding sets $S^{(t)} := \{x^{(i)} : i \leq t\}$.

Consider $[N]$ as a set, with implicit bijection to a family of pairwise independent hash functions from $[n]$ to $[n]$, where $\lg N = \Theta(\lg n)$. For each $i \in N$, we have corresponding hash function $h_i : [n] \to [n]$, and estimates $Y_i^{(t)} := \max\{\mathrm{lsb}(h_i(s)) : s \in S^{(t)}\}$ — the estimate for $\lg S^{(t)}$ given by hash function h_i. We will focus on the error of those estimators $\hat{Y}_i^{(t)} := Y_i^{(t)} - \lg S^{(t)}$.

Fact 11. *The error terms satisfy following subexponential tail bound*

$$\underset{i\sim \mathrm{Unif}([M])}{\mathbb{P}}(|\hat{Y}_i^{(t)}| > \lambda) \lesssim 2^{-\lambda}$$

Proof. Consider random set $S_k^{(t)} := \{s \in S^{(t)} : \mathrm{lsb}(h_i(s)) \leq k\}$.

For $k = \lg |S^{(t)}| - \lambda$ we have $\mathbb{E}\,|S_k| \simeq 2^{\lambda}$, and $\mathrm{Var}\,|S_k| \leq \mathbb{E}\,|S_k|$, hence $\mathbb{P}(|S_k| = 0) \leq 2^{-\lambda}$ by Chebyshev inequality, and therefore $\mathbb{P}(\hat{Y}_i^{(t)} > \lambda) < 2^{-\lambda}$.

For the lower tail bound it is enough to consider Markov inequality: if $k = \lg |S^{(t)}| + \lambda$, we have $\mathbb{E}\,|S_k| \leq 2^{-k}$, and $\mathbb{P}(|S_k| \geq 1) \leq 2^{-k}$. □

We will be interested in $Z_i^{(t)} := \lg(2 + |\hat{Y}_i^{(t)}|)$ which is proportional to the number of bits necessary to write down deviation $Y_i^{(t)}$ from $\lg S^{(t)}$. The Fact 11 implies that $Z_i^{(t)}$ have doubly-exponential tail bounds, up to some rescaling: $\mathbb{P}_i(Z_i^{(t)} > c\lambda) \leq \exp(-e^{\lambda})$ for some constant c.

Let us take a sequence $t_1, t_2, \dots t_R$, such that $|S^{(t_k)}| = 2^k$, where $R = \Theta(\lg n)$. We can now apply Lemma 10, with $M := N$ and functions $g_k : [M] \to \mathbb{R}_+$ given by $g_k(i) = Z_i^{(t_k)}/c$ and $w = \Theta(\lg \delta^{-1} + \lg n)$.

The final algorithm will be following: in the initialization phase, we choose a uniformly random string $S \in \{0,1\}^s$ and store it. Consider now $\Xi_S : [w_1] \to [N]^{[w_2]}$, as in the statement of Lemma 10, which for each value $t \in [w_1]$ yields a group of w_2 seeds for pairwise independent hash function from $[n]$ to $[n]$. For every such group, for example a group $\Xi_S(k)$, we store all the values $\{Y_i^{(t)} : i \in \Xi_s(k)\}$ in the compressed form: we store separately a median of all estimates within group, and the differences between $Y_i^{(t)}$ and aforementioned median. If at any point in time a size of the whole description of a given group in bits exceeds some $C_2 w_2$ we mark this group as broken and we stop updating it (where $C_2 = 2Cc$). Clearly, the total space complexity is bounded by $\mathcal{O}(s + C_2 w_2 w_1) = \mathcal{O}(s + w) = \mathcal{O}(\lg n + \lg \delta^{-1})$.

We claim that from the C-smallness condition, we can deduce that for a majority of $t \in [w_1]$, at all times both the median of all the estimates within group is close to the actual $|S^{(t)}|$, and the total space to store the whole group is bounded by $C_2 w_2$. If this is the case, then as an estimate for $|S^{(t)}|$ we can just report the median over groups $t \in [w_1]$ that are not marked as broken, of all the medians within a group of estimates $Y_i^{(t)}$, and the correctness of the algorithm follows.

To finish the argument, we need to show that every C-small group of estimators indeed yields a good approximation for F_0, and is stored succinctly at all times (i.e. never becomes marked as broken). Consider a C-small group $H \in [N]^{[w_2]}$. First, we will argue that at all times t_k, we have $\lg(\text{median}(i \in H : Y_i^{(t_k)})) = \lg |S^{(t)}| \pm C_3$. Indeed, we know that on average over all $i \in H$, we have $\mathbb{E}_{i \in H} \lg |2 + Y_i^{(t_k)}| < C$, therefore for at least $2/3$ fraction of $i \in H$, we have $\lg |2 + \hat{Y}_i^{(t_k)}| \leq 3C$, which means that for those i we have $\hat{Y}_i^{(t_k)} \leq 2^{3C}$. This, together with the definition of $\hat{Y}$ implies the claim with $C_3 = 2^{3C}$. To argue that we are storing group H using $\mathcal{O}(w_2)$ bits, let $M^{(t_k)}$ be the median of all $Y_i^{(t_k)}$ over i in H. The space to store the group is given by $\mathcal{O}(\lg \lg n)$ bits to store the median within a group, and $\mathcal{O}(\sum_{i \in H} \lg(2 + |Y_i - M|))$ to store the rest. We have bound

$$\sum_{i \in H} \lg\left(2 + |Y_i^{(t_k)} - M^{(t_k)}|\right) \leq \sum_{i \in H} \lg\left(2 + |Y_i^{(t_k)} - \lg F_0^{(t_k)}|\right) + \sum_{i \in H} \lg(1 + |M^{(t_k)} - \lg F_0^{(t_k)}|) \leq \mathcal{O}(w)$$

where the first sum is bounded because of the C-smallness condition.

Finally, we have to say that if the group satisfies those two properties at all times t_k, then those properties are satisfied (with larger constants) at all time steps $t \leq T$. To see that, fix some t between t_k and t_{k+1}. Note that $Y_i^{(t)}$ is non-decreasing with respect to t, and we have

$$|Y_i^{(t)} - M^{(t)}| \leq |Y_i^{(t_{k+1})} - M^{(t_k)}| \leq |Y_i^{(t_{k+1})} - M^{(t_{k+1})}| + |M^{(t_{k+1})} - M^{(t_k)})|.$$

Moreover, by triangle inequality

$$|M^{(t_{k+1})} - M^{(t_k)}| \leq |M^{(t_{k+1})} - \lg F_0^{(t_{k+1})}| + |M^{(t_k)} - \lg F_0^{(t_k)}| + |\lg F_0^{(t_{k+1})} - \lg F_0^{(t_k)}|$$

and each of those terms is bounded by constant.

This implies

$$\begin{aligned}\sum_{i \in H} \lg\left(2 + |Y_i^{(t)} - M^{(t)}|\right) &\leq \sum_{i \in H} \left[\lg\left(1 + |Y_i^{(t_{k+1})} - M^{(t_{k+1})}|\right) + \lg\left(1 + |M^{(t_k)} - M^{(t_{k+1})}|\right)\right] \\ &= \mathcal{O}(|H|) = \mathcal{O}(w_2).\end{aligned}$$

This completes the proof of the correctness of the algorithm — at any step t, all C-small groups are not marked as broken, and all of them report a $(1 + \varepsilon)$ approximation. Strictly more than half of all the groups is C-small, hence the median of all groups that are still active, has to be a $(1 + \varepsilon)$ approximation to the quantity of interest as well.

5 High accuracy regime

In this section we prove Theorem 3. As a building block we will use algorithm discussed in [KNW10, Section 3.2]. In the Appendix we prove the following, qualitatively stronger bounds on the space complexity of their algorithm. The construction of the algorithm, and correctness analysis was already present in [KNW10] — correctness can be also deduced from the discussion in Section 6, where we discuss this algorithm in detail, and show stronger guarantees for a slight variation of it. Note that in the original paper the guarantees on the space complexity of this algorithm were proven when $\frac{1}{\varepsilon^2} > \lg n$, as this could be assumed without loss of generality in their setting. For us, the scenario when $\frac{1}{\varepsilon^2} < \lg n$ is relevant.

Theorem 12. *There is an algorithm $\mathcal{F}_\varepsilon$ which gives a $(1+\varepsilon)$-approximation to $F_0^{(t)}$ with probability at least $\frac{5}{6}$, assuming access to an oracle providing strong tracking of $F_0^{(t)}$ with constant factor approximation C, and oracle access to $\mathcal{O}(\lg n + \operatorname{poly}\lg(1/\varepsilon))$ additional random bits. The space usage of this algorithm at any given time t (excluding random bits mentioned above), denoted by $W^{(t)}$, satisfies*

$$\mathbb{P}(W^{(t)} > \frac{C_2}{\varepsilon^2}) \leq \varepsilon^4 \tag{2}$$

and

$$\mathbb{P}(W^{(t)} > \frac{\lambda}{\varepsilon^2}) \leq \exp(-e^{\Omega(\lambda)}). \tag{3}$$

Moreover for $t_1 < t_2$ such that $|S^{(t_1)}| \geq |S^{(t_2)}|/2$ we have

$$W^{(t_1)} \leq W^{(t_2)} + \mathcal{O}(\frac{1}{\varepsilon^2}) \tag{4}$$

for some universal constant C.

We will show, how assuming this theorem we can prove Theorem 3, leveraging tools described in Section 4.

First of all, note that on a way to prove Theorem 3, we can assume without loss of generality that $\frac{1}{\varepsilon^2} < \lg n$, for if it is not the case, we can just use $\lg \delta^{-1}$ parallel repetitions of the KNW algorithm, to achieve δ failure probability with space $\mathcal{O}(\lg n \lg \delta^{-1} + \frac{\lg \delta^{-1}}{\varepsilon^2}) = \mathcal{O}(\frac{\lg \delta^{-1}}{\varepsilon^2})$. In particular, this implies that the number of random bits used in Theorem 12 is $\mathcal{O}(\lg n + \operatorname{poly}\lg \frac{1}{\varepsilon^2}) = \mathcal{O}(\lg n)$.

We consider two separate cases, depending on relation between ε and $\lg n$. First, let us discuss case when $\varepsilon < \left(\frac{1}{\lg n}\right)^{1/4}$. In this scenario, Eq. (2) implies that for any specific position with probability $1 - \frac{C}{\lg n}$ the total space consumption of a single instance of KNW algorithm use space $\mathcal{O}(\frac{1}{\varepsilon^2})$. Because of Eq. (4), we can union bound only over positions for which $F_0^{(t)}$ grows by a factor of two (there are $\mathcal{O}(\lg n)$ such positions), to ensure that with probability $\frac{5}{6}$ single instantiation of the algorithm at all times use space $\mathcal{O}(\frac{1}{\varepsilon^2})$.

We will use $\mathcal{O}(\lg \delta^{-1})$ parallel instantiations of this algorithm. We use an algorithm which existence is guaranteed by Theorem 1 instantiated with failure probability δ to provide a strong-tracking oracle for all those implementations simultaneously. Instead of using $\lg \delta^{-1}$ independent seeds across different instantiation of the algorithm $\tilde{F}_\varepsilon$, we consider the following standard pseudorandom object raising from random walks over explicit low degree expander graphs.

Definition 13. *[Vad12, Chapter 3] A function $\Gamma : \{0,1\}^s \times [w] \to [M]$ is (ε, δ)-*averaging sampler*, if for any function $f : [M] \to [0,1]$ and random variables $Y_i := \Gamma(U, i)$ for uniformly random U, we have*

$$\mathbb{P}(|\sum_{i \leq w} f(Y_i) - \mu w| > \varepsilon w) < \delta,$$

where $\mu := \mathbb{E}_{Y \sim \mathrm{Unif}([M])} f(Y)$.

A (ε, δ) sampler is called explicit *if Γ can be computed in polynomial time in s and w.*

Theorem 14. *[Vad12, Corollary 4.41] For every ε, δ there exist an explicit (ε, δ)-averaging sampler, with number of samples $w = \mathcal{O}_\varepsilon(\lg \delta^{-1})$ and seed length $s = \lg M + \mathcal{O}_\varepsilon(w)$, where $\mathcal{O}_\varepsilon$ notation hides constant depending on ε.*

Moreover Γ can be computed using space $\mathcal{O}(s + w)$.

Consider $[M]$ to be the space of all possible random strings that were to be supplied to the algorithm $\mathcal{F}_\varepsilon$, and note that $\lg M = \mathcal{O}(\lg n)$. Let us fix an input stream, and condition on specific realization of the constant approximation tracking oracle (assuming that it succeeded — we can bound the failure probability by δ). For $k \in [M]$ let $\tilde{F}_0(k)$ be the approximation to $F_0^{(T)}$ reported by algorithm $\mathcal{F}_\varepsilon$ while supplied random string corresponding to k. We can define $f : [M] \to \{0, 1\}$, to be $f(k) := 1$ if and only if $|\tilde{F}_0(k) - F_0^{(T)}| < \varepsilon F_0^{(T)}$. Clearly, we have $\mathbb{E}_{k \sim \mathrm{Unif}([M])} f(k) > \frac{5}{6}$ — this follows from the correctness guarantee for algorithm $\mathcal{F}_\varepsilon$.

Consider now Γ, a $(\frac{1}{6}, \delta)$ averaging sampler as in Theorem 14. Except with failure probability δ over uniform random seed S it will yield us a sequence $\Gamma_S(1), \ldots, \Gamma_S(w)$ of seeds, such that at least $\frac{2}{3}w$ amongst $\tilde{F}_0(\Gamma_S(1)), \ldots \tilde{F}_0(\Gamma_S(w))$ yields a good approximation to F_0. In this case, if we report median of all $\tilde{F}_0(k)$, it will be a valid answer.

The space of this algorithm is $\mathcal{O}(\lg n + \lg \delta^{-1})$ for the constant approximation oracle, $\mathcal{O}(\lg n + \lg \delta^{-1})$ for storing the seed to the averaging sampler, and $\mathcal{O}(\frac{\lg \delta^{-1}}{\varepsilon^2})$ for storing all w instantiations of $\mathcal{F}_\varepsilon$ algorithm. This yields total space complexity $\mathcal{O}(\lg n + \lg \delta^{-1})$. The failure probability of each one of the three phases is bounded by δ, hence the total failure probability is bounded by 3δ, and the result follows by rescaling δ by a constant factor.

Let us now turn our attention to the analysis of the second case, where $\varepsilon > \left(\frac{1}{\lg n}\right)^{1/4}$. In this case, the proof will make use of Eq. (3), and mimic the proof of Theorem 1. Note that in this regime of parameters, we can assume without loss of generality, that $\lg \delta^{-1} \geq C\sqrt{\lg n}$, as otherwise we could take δ' such that $\lg \delta'^{-1} = \lg \delta^{-1} + C\sqrt{\lg n}$, and the additional $\frac{C\sqrt{\lg n}}{\varepsilon^2}$ term in the space complexity will be dominated by $\mathcal{O}(\lg n)$ term anyway.

First of all, by naive failure probability amplification, after adjusting other constant in Theorem 12, we can actually assume that failure probability of this algorithm is small constant c_0. We will apply Lemma 10 where the universe $[M]$ is given by $M = 2^{\hat{r}}$, with r being the number of random bits accessed by this new adjusted algorithm (in particular $\lg M = \Theta(\lg n)$).

Let us take a sequence $t_1, \ldots t_{R-1}$ where $t_1 = 0$, and each t_j for $j > 1$ is smallest such that $F_0^{(t_j)} \geq 2F_0^{(t_{j-1})}$. Clearly, $R = \mathcal{O}(\lg n)$.

We will use $g_1, \ldots g_R$ to be given by $g_i(m) := \frac{\varepsilon^2}{C} W_m^{(t_i)}$ for $i \leq R - 1$, where by $W_m^{(t_i)}$ we denote the space consumption of the instantiation of the algorithm described in Theorem 12 with random bits given by $m \in [M]$. Finally, we pick $g_R(m)$ to be 0 if the instance of algorithm corresponding to random bits m succeeds to provide $1 + \varepsilon$ approximation, and some large C_0 if it succeeds. Given that failure probability c_0 is small enough depending on C_0, we can ensure that this function g_R indeed satisfy doubly-exponential tail bounds. For all previous functions $g_1, \ldots g_{R-1}$, doubly exponential tail bounds are guaranteed by Eq. (3). Finally, we can apply Lemma 10 with $w = \Theta(\lg \delta^{-1})$ — we assumed that $\lg \delta^{-1} = \Theta(\sqrt{\lg n})$, so the assumptions of this lemma are satisfied.

The sampler Ξ guaranteed by Lemma 10 returns a sequence of groups of estimators, such that most of those groups are C-small (except with small failure probability δ over the choice of the seed). We wish to argue that if the sampler succeeds (i.e. most of the reported groups is C-small), then the algorithm will use small space, and will correctly return $(1 + \varepsilon)$ approximation for number of distinct elements. As in Section 4, we can discard any group that becomes too large over the course of algorithm, hence the total space consumption is $\mathcal{O}(\frac{w}{\varepsilon^2} + \lg n) = \mathcal{O}(\frac{\lg \delta^{-1}}{\varepsilon^2} + \lg n)$. By C-smallness condition restricted to functions $g_1, \ldots g_{R-1}$ and Eq. (4) majority of the groups (all C-small groups) are never discarded in this way — the argument for this is identical as in the proof of Theorem 1. We need to argue, that reporting median of all the medians within surviving groups indeed yields $(1+\varepsilon)$-approximation to the number of distinct elements. This is guaranteed by C-smallness condition applied to function g_R — indeed, for large enough C_0 we can ensure that any C-small group of estimators have at least $\frac{2}{3}$ fraction of estimators reporting value that is within $(1+\varepsilon)$ to the actual answer.

6 Strong tracking of distinct elements

In this section we prove Theorem 5. Let us first state a technical lemma essential in the argument. After stating this lemma we will show how, together with Lemma 6, those two imply Theorem 5. The rest of this section will be devoted to proving those lemmas.

Definition 15. *For a finite universe $[K]$, let $\phi_K : [K]^* \to \mathbb{N}$ be given by*

$$\phi_K(r_1, \ldots r_s) := \#\{j \in K : \exists i, r_i = j\}.$$

Moreover, let $\Phi_K : \mathbb{N} \to \mathbb{N}$ be given by

$$\Phi_K(t) := \mathbb{E}\, \phi(X_1, \ldots X_t)$$

over $X_1, \ldots X_t$ uniformly random in $[K]$ and independent.

We will skip the index P, when the underlying universe is clear from context.

Lemma 16. *Consider a sequence $X_1, \ldots X_R \in [K]$ of uniform random variables, q-wise independent for some $q = \Theta(\operatorname{polylog} R)$, where $R \leq K/20$. Then*

$$\sup_{t \leq R} |\phi(X_1, \ldots X_t) - \Phi(t)| = \mathcal{O}(\sqrt{R})$$

with probability $3/4$.

The following fact will also be useful

Fact 17. *We can calculate Φ exactly as $\Phi_K(t) = t\left(1 - \left(1 - \frac{1}{t}\right)^n\right)$ for $t > 0$, and $\Phi_K(0) = 0$.*

Moreover for all $0 \leq \alpha_1, \alpha_2 \leq K/20$, we have $0.9|\alpha_2 - \alpha_1| \leq |\Phi(\alpha_2) - \Phi(\alpha_1)| \leq |x_2 - x_1|$.

Finally, for $t < P/20$, we have $\operatorname{Var}(\phi(X_1, \ldots X_t)) \leq \Phi(t) \leq t$.

Proof of Theorem 5. First we will discuss how, given an upper bound P on the number of distinct elements, we can analyze a variant of the algorithm in [KNW10] to argue, that in fact at all times t it provides a $\pm\varepsilon K$ additive approximation to $|S^{(t)}|$, without any additional space blowup. This can be used to say that after amplifying the failure probability to $\delta/\lg n$, by union bound over all positions where $|S^{(t)}|$ grows by a factor of two, we can obtain strong tracking guarantee with failure probability δ.

Take $P = \frac{100}{\varepsilon^2}$, and let us consider a 8-wise independent hash-function $h_1 : [n] \to [n]$, and random sets $S_k^{(t)} := \{s \in S^{(T)} : \operatorname{lsb}(h_1(s)) \geq 2^k\}$ as previously. Let us consider in addition a pairwise independent hash function $h_3 : [n] \to [P^2]$, and finally a polylog(P)-wise independent hash function $h_4 : [P^2] \to [P]$. Define $h_2 : [n] \to [P]$ to be the composition $h_2 := h_4 \circ h_3$.

In the Appendix A it is discussed how, given oracle access to constant factor strong tracking, we can maintain a sketches of size $\mathcal{O}(\frac{1}{\varepsilon^2})$ on average (with some small constant probability of failure), such that we can recover $|h_2(S_k)|$ for any k at any point of the stream.

Let us fix any $K < n$, and let k be such that $2^{-k}K \approx \frac{1}{10\varepsilon^2}$. We wish to show that with probability $\frac{9}{10}$ we have

$$\forall t \leq T_0, |\Phi^{-1}(|h_2(S_k^{(t)})|)2^k| = |S^{(t)}| \pm \mathcal{O}(\varepsilon K) \tag{5}$$

where T_0 is such that $|S^{(T_0)}| = K$.

If this were true, we could repeat the construction $\mathcal{O}(\lg \lg n + \lg \delta^{-1})$ times, to amplify success probability for the median estimator to $1 - \frac{\delta}{\lg n}$, and use a union bound to ensure that Eq. (5) is satisfied for all k simultaneously. This can kind of amplification can be implemented exactly as described in Section 5.

Given access to strong tracking oracle with constant failure probability, we know which set $|S_k|$ to use, at any given time, to estimate $|S^{(t)}|$, as above.

We only need to show that Eq. (5) indeed holds with large constant probability. We can assume without generality that $|S^{(t)}| = t$, i.e. all the elements in the input stream are distinct.

First of all, we will show that $\forall t \leq T_0$, $|S_k^{(t)}|2^k = |S^{(t)}| \pm \varepsilon K$. Indeed, note that if we take $X_k := |S_k^{(t)}|2^k - |S^{(t-1)}|2^k - 1$, we can see that X_k are 4-wise independent (because hash function h_1 was assumed to be 4-wise independent), and satisfy $\mathbb{E}\, X_k = 0$, $\mathbb{E}\, X_k^2 \approx 2^k$. By applying Lemma 6 we see that $\sup_{t \leq T_0} \left| 2^k |S_k^{(t)}| - t \right| = \mathcal{O}(\sqrt{2^k T_0}) = \mathcal{O}(\varepsilon K)$.

By birthday paradox, with probability 9/10 we have $|h_3(S_k^{(t)})| = |S_k^{(t)}|$, i.e. function h_3 has no collisions in the part of the stream of interest. Moreover, by Lemma 16, conditioned on $S_k^{(t)}$ we have that with high probability at all times t in the range of interest that $|h_2(S_k^{(t)})| = \Phi(|S_k^{(t)}|) \pm \mathcal{O}(\frac{1}{\varepsilon})$. This together with Fact 17, implies that $\Phi^{-1}(|h_2(S_k^{(t)})|)$ yields at all times a good approximation of $S_k^{(t)}$, and by composing with the previous argument, this shows Eq. (5) $\square$

The rest of this section will be devoted to proving Lemma 6 and Lemma 16.

Lemma 18. *Let $X_1, X_2, \ldots X_N$ be a sequence of random variables, satisfying $\mathbb{E}\, X_i^2 \leq 1$, and for $i \neq j$, both $\mathbb{E}\, X_i X_j = 0$ and $\mathbb{E}\, X_i^2 X_j^2 \leq 1$. If we take $S_t := \sum_{k<t} X_t$, then*

$$\mathbb{P}(\sup_{t \leq N} |S_t| > \gamma \sqrt{N}) \lesssim \frac{1}{\gamma^2}$$

Proof. Let us assume without loss of generality that $N = 2^n$. Take $A_0 := \{0, N\}$, and for $k \leq n$, take $A_k := \{j 2^{n-k} : j \in \{0, 1, \ldots 2^k\}\}$.

For every $t \in A_k \setminus A_{k-1}$, it has two neighbours in A_{k-1}: those are $t + \delta_k, t - \delta_k \in A_k$, where $\delta_k := 2^{n-k+1}$.

Observe that $\mathbb{E}_{X \sim \mathcal{D}} (S_t - S_{t-\delta_k})^2 \leq \delta_k$, and similarly $\mathbb{E}_{X \sim \mathcal{D}} (S_t - S_{t+\delta_k})^2 \leq \delta_k$, and moreover $\mathbb{E}_{X \sim \mathcal{D}} (S_t - S_{t-\delta_k})^2 (S_t - S_{t+\delta_k})^2 \leq \delta_k^2$ — this can be shown by expanding both sums $(S_t - S_{t-\delta_k})$ and $(S_{t+\delta_k} - S_t)$ — because of 4-wise independence and $\mathbb{E}\, X_i X_j = 0$ for $i \neq j$, we have $\mathbb{E}(S_t - S_{t-\delta_k})^2 (S_{t+\delta_k} - S_t)^2 = \mathbb{E} \sum_{1 \leq i \leq \delta_k} \sum_{1 \leq j \leq \delta_k} X_{t-\delta_k+i}^2 X_{t+j}^2 \leq \delta_k^2$.

Those bounds on second moments, together with Markov inequality, yield a bound

$$\mathbb{P}(\min\{|S_t - S_{t-\delta_k}|, |S_t - S_{t+\delta_k}|\} > \lambda \sqrt{\delta_k}) \leq \mathbb{P}((S_t - S_{t-\delta_k})^2 (S_t + S_{t+\delta_k})^2 > \lambda^4 \delta_k^2) \leq \frac{1}{\lambda^4}$$

For $k \geq 1, t \in A_k$ define a bad event $E_{k,t}$ to be $\min\{|S_t - S_{t-\delta_k}|, |S_t - S_{t+\delta_k}|\} > \gamma \sqrt{\delta_k} 2^{k/3}$. We have $\mathbb{P}(E_{k,t}) \leq \frac{2^{-4k/3}}{\gamma^4}$. Since $|A_k \setminus A_{k-1}| = 2^{k-1}$, for every k we have $\mathbb{P}(\exists t \in A_k \setminus A_{k-1} : E_{k,t}) \leq \frac{2^{-k/3}}{\gamma^4}$, and finally by taking union bound over all k, we have $\mathbb{P}(\exists k, t \in A_k \setminus A_{k-1} : E_{k,t}) \lesssim \frac{1}{\gamma^4}$.

We now claim, that if no of the events $E_{k,t}$ happened, we have

$$\sup_{t \leq N} |S_t| \leq |S_N| + \gamma \sum_{k \leq n} \sqrt{\delta_k} 2^{k/3}. \tag{6}$$

Before we prove that, let us observe that $\sum_{k \leq n} \sqrt{\delta_k} 2^{k/3} \leq \sqrt{N} \sum_{k \leq n} 2^{-k/6} = \mathcal{O}(\sqrt{N})$.

We will show the following fact by induction over k_0

$$\sup_{t \in A_{k_0}} S_t \leq S_N + \gamma \sum_{k \leq k_0} \sqrt{\delta_k} 2^{k/3}.$$

Clearly for $k_0 = 0$ this is satisfied. Moreover, for $k_0 > 1$, if $t \in A_{k_0} - A_{k_0 - 1}$, we have some $\tilde{t} \in A_{k_0 - 1}$, with $|S_t - S_{\tilde{t}}| \leq \gamma \sqrt{\delta_{k_0}} 2^{k_0/3}$ — this follows from the fact that event $E_{k_0,t}$ did not happen. This yields

$$|S_t| \leq |S_{\tilde{t}}| + |S_t - S_{\tilde{t}}| \leq \sup_{t^* \in A_{k_0 - 1}} |S_{t^*}| + \sqrt{\delta_{k_0}} 2^{k_0/3}$$

On the other hand, $\mathbb{E}\, S_N^2 \leq N$, hence by Chebyshev inequality $\mathbb{P}(|S_N| > \sqrt{N} \gamma) \leq \frac{1}{\gamma^2}$. This, together with inequality Eq. (6) yields a tail bound $\mathbb{P}(\sup_{t \leq N} S_t > K \sqrt{N} \gamma) = \frac{1}{\gamma^2} + \mathcal{O}(\frac{1}{\gamma^4})$ for some universal constant K; after changing γ by this constant factor, we conclude the statement of the lemma. $\square$

Observe that Lemma 6 is a direct corollary from Lemma 18 — the 4-wise independent random variables as in the statement of this lemma satisfy all the assumptions of Lemma 18.

For the proof of the Lemma 16 we will need following statement of the Chernoff inequality

Theorem 19 (Chernoff bound [BR94, Lemma 2.3])**.** *If $X_1, \dots X_n$ are r-wise independent random variables satisfying $0 < X_i < 1$ almost surely, with $\mu := \mathbb{E} \sum X_i$ then for $\lambda > 2$ we have*

$$\mathbb{P}(\sum X_i > \lambda\mu) < \exp(-\Omega(\max\{\mu\lambda, r\} \lg \lambda))$$

Lemma 20. *For any P and $R \le P/20$, there exist a polynomial $\hat{\phi} : \{0,1\}^{\lg P \times R} \to \mathbb{R}$ of degree $\mathcal{O}(\lg^2 P)$ with integer coefficients, such that for every distribution $X_1, \dots X_R$ which is at least $r = \mathrm{poly}(\lg P)$-wise independent, we have*

$$\|\hat{\phi}(\bar{X_1}, \dots \bar{X_R}) - \phi(X_1, \dots X_R)\|_p = o(1)$$

for any $p \le C \lg\lg P$, where C is a constant. In particular

$$\mathbb{P}(\hat{\phi}(\bar{X_1}, \dots \bar{X_R}) \ne \phi(X_1, \dots X_R)) = o(1). \tag{7}$$

Above, $\bar{X}_i$ denotes a binary representation of X_i.

Proof. Consider polynomials $\mathrm{EQ}_k : \mathbb{R}^{\lg P} \to \mathbb{R}$, such that EQ restricted to the hypercube $\{0,1\}^{\lg P}$ has values $\{0,1\}$ and it takes value 1 only for $\bar{k}$ (for a number $k \in [P]$, we write $\bar{k} \in \{0,1\}^{\lg P}$ to be the binary representation of k). There is such a multilinear polynomial of degree $\lg P$.

Consider moreover polynomial $I_0 : \mathbb{R} \to \mathbb{R}$ of degree $d = \mathcal{O}(\lg P)$ defined as

$$I_0(x) := 1 - \sum_{0 \le i \le d} (-1)^i \binom{x}{i}$$

Let us observe that for $x \in \mathbb{N}$, with $x \le d$ we have

$$I_0(x) = \begin{cases} 0 & \text{for } x = 0 \\ 1 & \text{otherwise} \end{cases}$$

and moreover $|I_0(x)| \lesssim \binom{x}{d+1}$ for any $x > 0$.

Let us define now

$$\hat{\phi}(\bar{X}_1, \dots \bar{X}_R) := \sum_{k \in [P]} I_0(\sum_{j \le R} \mathrm{EQ}(\bar{X}_j, k))$$

and note that this is a polynomial of degree $d \lg P = \mathcal{O}(\lg^2 P)$.

Given an instantiation of random variables $X_1, \dots X_R$, we will take $B_i := \#\{j : X_j = i\}$ defined for every $i \in [P]$, and moreover we will define $M(X_1, \dots X_R) := \max_{i \in [P]} B_i$.

We claim that $\hat{\phi}(\bar{X}_1, \dots \bar{X}_R) = \phi(X_1, \dots X_R)$ as long as $M(X) \le d$, and for $M(X) > d$ we have $|\hat{\phi}(X_1, \dots X_R) - \phi(X_1, \dots X_R)| < N\binom{M(X)}{d} < N \exp(c_0 d \ln \frac{M(X)}{d})$ for some constant c_0. This yields

$$\begin{aligned} \mathbb{E}\, |\hat{\phi} - \phi|^q &\le \sum_{k > \lg d} \mathbb{P}(2^k < M(X) < 2^{k+1})\, \mathbb{E}\left[N^q \exp(c_0 q d \lg \frac{M(X)}{d}) \middle| 2^k < M(X) < 2^{k+1} \right] \\ &\le \sum_{k > \lg d} \mathbb{P}(M(X) > 2^k) \exp(c_0 q d \lg \frac{2^k}{d} + q \ln N). \end{aligned} \tag{8}$$

We can bound the tail probabilities of $M(X)$ as follows

$$\begin{aligned} \mathbb{P}(M(X) > \lambda) &\le \sum_{j \le R} \mathbb{P}(B_j > \lambda) \\ &\le R\, \mathbb{P}(B_1 > \lambda) \\ &\le R \exp(-C \min(\lambda, r) \lg \lambda) \end{aligned} \tag{9}$$

where the last inequality follows from the Chernoff bound Theorem 19, because $\mathbb{E}\, B_i < 1/20$.

Eq. (9) together with Eq. (8) yields

$$\mathbb{E}\,|\hat{\phi} - \phi|^q \leq \sum_{k > \lg d} \exp(-C(k \min(2^k, r)) + dq \lg \frac{2^k}{d} + q \lg R)$$

The exponents in this sum are quickly decaying, so the whole sum is of the same order as the first term, namely

$$\exp(-\Theta(d \lg d) + \Theta(dq) + q \lg R)$$

if we pick $r > d^{\mathcal{O}(1)}$.

Hence, for $q \ll \lg d$ and $d \gg \lg R$ we have

$$\mathbb{E}\,|\hat{\phi} - \phi|^q \leq \exp(-\Omega(d \lg d)).$$

□

We will now show that for any distribution with enough independence, specific types of random walks associated with functions ϕ and $\hat{\phi}$ stay close to each other with good probability. If variables $X_1, \ldots X_R$ are uniform and independent, processes S_t described below are just Doobs martignales associated with function ϕ, that were used to show correctness of the algorithm in the random-oracle model.

Lemma 21. *Consider $\hat{\phi}$ to be the polynomial from Lemma 20, and let $X_1, \ldots X_R \in \Sigma$ be a sequence of k-wise independent random variables with marginal uniform distribution, where $k = \mathrm{poly}(d)$.*

Consider $S_t := \mathbb{E}_{X'}\, \phi(X_1, \ldots X_t, X'_{t+1}, \ldots X'_R)$, and $\hat{S}_t := \mathbb{E}_{X'}\, \hat{\phi}(X_1, \ldots X_t, X'_{t+1}, \ldots X'_R)$, where X' are independent random variables, distributed uniformly over Σ. Then

$$\mathbb{P}(\exists t,\ |\hat{S}_t - S_t| > \lambda) \lesssim \frac{R}{\lambda^2}. \tag{10}$$

Proof. For single t we have

$$\begin{aligned} \|\hat{S}_t - S_t\|_2^2 &\leq \mathbb{E}_X |\mathbb{E}_{X'} \phi(X_1, \ldots X_t, X'_{t+1}, \ldots, X'_R) \\ &\quad - \hat{\phi}(X_1, \ldots X_t, X'_{t+1}, \ldots X'_R)|^2 \\ &\leq \mathbb{E}_{X,X'} |\phi(X_1, \ldots X_t, X'_{t+1}, \ldots X'_R) \\ &\quad - \hat{\phi}(X_1, \ldots X_t, X'_{t+1}, \ldots X'_R)|^2 \\ &\lesssim 1, \end{aligned}$$

where the last inequality follows from Lemma 20.

Hence, $\mathbb{P}(|\hat{S}_t - S_t| > \lambda) \lesssim \frac{1}{\lambda^2}$, and by union bound

$$\mathbb{P}(\exists t,\ |\hat{S}_t - S_t| > \lambda) \lesssim \frac{R}{\lambda^2}$$

□

Lemma 22. *For $X_1, \ldots X_R$ and $\hat{S}_i$ defined as in the Lemma 21, if $\Delta_i := \hat{S}_i - \hat{S}_{i-1}$, then*

$$\mathbb{E}\,\Delta_i^2 \lesssim 1 \tag{11}$$

$$\mathbb{E}\,\Delta_i \Delta_j = 0 \qquad \text{for } i \neq j \tag{12}$$

$$\mathbb{E}\,\Delta_i^2 \Delta_j^2 \lesssim 1 \qquad \text{for } i \neq j \tag{13}$$

Proof. Note that all the expressions in the statement of the lemma are expectations of polynomials of degree at most $4d$, and variables $X_1, \dots X_R$ are r-wise independent for $r > 4d$. We can without loss of generality prove this theorem, assuming that $X_1, \dots X_R$ are instead independent.

In that case Δ_i is a sequence of increments of a Doob's martingale, and therefore $\mathbb{E}\,\Delta_i = 0$, and $\mathbb{E}\,\Delta_i \Delta_j = 0$ follows immediately.

We need to show the last condition, i.e. $\mathbb{E}\,\Delta_i^2 \Delta_j^2 \lesssim 1$ for $i \neq j$. Let us first bound this expression as follows

$$\begin{aligned}
\mathbb{E}\,\Delta_i^2 \Delta_j^2 &= \int_t \mathbb{P}(\Delta_i^2 \Delta_j^2 > t)\,\mathrm{d}t \\
&\leq \int_t \mathbb{P}(\Delta_i > t^{1/4})\,\mathrm{d}t + \mathbb{P}(\Delta_j > t^{1/4})\,\mathrm{d}t \\
&\leq \int_t \frac{\mathbb{E}\,\Delta_i^8 + \mathbb{E}\,\Delta_j^8}{t^2}\,\mathrm{d}t \\
&\lesssim \mathbb{E}\,\Delta_i^8 + \mathbb{E}\,\Delta_j^8
\end{aligned}$$

It is enough now to show that $\mathbb{E}\,\Delta_i^8 \lesssim 1$. Indeed, by definition of Δ_i and Jensens inequality, $\mathbb{E}\,\Delta_i^8$ is bounded by

$$\mathop{\mathbb{E}}_{X,X'} (\hat{\phi}(X_1, \dots X_N) - \hat{\phi}(X_1, \dots X_i', X_{i+1}, \dots X_N))^8$$

If we use $X^{(i)}$ to denote a sequence $(X_1, \dots X_i', \dots X_N)$, we have $\|\hat{\phi}(X) - \hat{\phi}(X^{(i)})\|_8 \leq \|\hat{\phi}(X) - \phi(X)\|_8 + \|\phi(X) - \phi(X^{(i)})\|_8 + \|\hat{\phi}(X^{(i)}) - \phi(X^{(i)})\|_8 \lesssim 1$ where the outer terms are bounded by Lemma 20, and the bound $|\phi(X) - \phi(X^{(i)})| \leq 1$ follows from the fact that value of function ϕ changes by at most one if we change only a single argument. □

Corollary 23. *For $\hat{S}_k$ defined as above, we have $\mathbb{P}(\sup_{k \leq R} |\hat{S}_k - \hat{S}_0| \geq \lambda) \lesssim \frac{R}{\lambda^2}$.*

Proof. Folows from Lemma 22 and Lemma 18. □

Corollary 24. *For S_t defined as in Lemma 21*

$$\mathbb{P}(\sup_t |S_t - S_0| \geq \lambda) \lesssim \frac{R}{\lambda^2}$$

Proof. Clearly

$$\mathbb{P}(\sup_t |S_t - S_0| \geq \lambda) \leq \mathbb{P}(\exists t,\ |S_t - \hat{S}_t| \geq \lambda/3) + \mathbb{P}(\sup_t |\hat{S}_t - \hat{S}_0| \geq \lambda/3)$$

and the claim follows from Corollary 24 and Lemma 21. □

Remark 25. *If $\hat{X}_{t+1}, \dots \hat{X}_R$ are all uniform and independent, then for any setting of variables $X_1, \dots X_t$ we have*

$$\mathop{\mathbb{E}}_{\hat{X}} \phi(X_1, \dots X_t, \hat{X}_{t+1}, \dots \hat{X}_R) = \Phi\left[\Phi^{-1}(\phi(X_1, \dots X_t)) + R - t\right].$$

We are finally ready to prove the last technical lemma, stating that for bounded-wise independence balls-and-bins experiment, the number of non-empty bins stays close to its expectation at all times.

Proof of Lemma 16. By Corollary 24 we conclude that with probability $\frac{9}{10}$ we have

$$\forall t, \mathop{\mathbb{E}}_{X'} \phi(X_1, \dots X_t, X'_{t+1}, \dots, X'_R) - \Phi(R) \leq \mathcal{O}(\sqrt{R}).$$

Using bi-Lipschitz properties of Φ_R (Fact 17), we deduce that equation above imply

$$\forall t, \Phi^{-1}(\mathop{\mathbb{E}}_{X'} \phi(X_1, \dots X_t, X'_{t+1}, \dots . X'_R)) = R \pm \mathcal{O}(\sqrt{R}).$$

Applying Remark 25, we deduce

$$\forall t, \Phi^{-1}(\phi(X_1, \dots X_t)) = t \pm \mathcal{O}(\sqrt{R})$$

and finally, again using bi-Lipschitz continuity of Φ, we deduce

$$\forall t, \phi(X_1, \dots, X_t) = \Phi(t) \pm \mathcal{O}(\sqrt{R})$$

□

7 Strong tracking lower bound

In this chapter we prove Theorem 32 — $\Omega(\frac{\lg \lg n}{\varepsilon^2})$ lower bound for strong tracking of distinct elements. To this end, we introduce concept of T-game — model of communication-complexity game tailored to the lower bound in question.

Definition 26 (T-game)**.** *For any relation $\mathcal{R} \subset \{0,1\}^n \times \{0,1\}^n \times \Sigma$, we consider T-game $T(\mathcal{R}, k)$ with k-rounds, to be communication problem with two parties, Alice and Bob defined as follows. In each round of the game*

- *Alice receives her input $x_k \in \{0,1\}^n$, and Bob receives his input $y_k \in \{0,1\}^n$.*
- *Alice receives Bobs input y_{k-1} from the previous round, and Bob observes Alice input x_{k-1} from the previous round.*
- *Alice and Bob can observe private random coins $r_k^1, r_k^2 \in \{0,1\}^*$.*
- *Alice can send a message a_k to Bob that depends on all her observations.*
- *Bob reports to the judge his output $z_k \in \Sigma$.*
- *Bob can send a message b_k to Alice.*

We say that protocol P succeeds *on input $(x_1, y_1), \dots (x_k, y_k)$ and random coins $\left((r_i^1, r_i^2)\right)_{i \in [k]}$ if $\forall_k (x_k, y_k, z_k) \in \mathcal{R}$. For any protocol P by Alice and Bob, we define complexity $C(P)$ of the protocol to be the largest length of a_k, or b_k send by any party.*

For a distribution μ over pair of strings $\{0,1\}^n \times \{0,1\}^n$, let $\mathcal{P}_{\mu,\delta}$ be the set of all protocols that succeed with probability $1-\delta$, given as input sequence of independent samples $(x_1, x_2), \dots (x_k, y_k) \sim \mu$. We define

$$D_{\mu,\delta}(T(\mathcal{R}, k)) := \inf_{P \in \mathcal{P}_{\mu,\delta}} C(P)$$

Definition 27. *For relation $\mathcal{R} \subset \{0,1\}^n \times \{0,1\}^n \times \Sigma$, we will denote by $D^{\rightarrow}_{\mu,\delta}(\mathcal{R})$ the one-way deterministic communication complexity of $\mathcal{R}$ under distribution μ of inputs for Alice and Bob.*

The following lemma connects complexity of T-game based on relation $\mathcal{R}$, with one-way communication complexity of the relation $\mathcal{R}$ itself.

Lemma 28. *For every protocol for a k-round T-game with failure probability δ, over independent samples distributed according to μ and complexity $C(P)$, there is a one-way communication protocol for a distribution μ with communication complexity $C(P)$ and failure probability δ/k. Formally, for every relation $\mathcal{R}$ the following inequality holds*

$$D^{\rightarrow}_{\mu,\delta/k}(\mathcal{R}) \leq D_{\mu,\delta}(T(\mathcal{R}, k)).$$

Proof. Consider fixed protocol P with $C(P) \leq S$. By standard averaging argument we can assume that P is a deterministic protocol.

Consider event A_t given by $(x_t, y_t, z_t) \in \mathcal{R}$. A_t — such an event depends only on $\{(x_s, y_s)\}_{s \leq t}$. We have $\delta \geq \mathbb{P}(\bigvee_t \neg A_t) = \sum_t \mathbb{P}(\neg A_t | \bigwedge_{s<t} A_s)$, and therefore there is t_0 for which $\mathbb{P}(\neg A_{t_0} | \bigwedge_{s<t_0} A_s) \leq \frac{\delta}{k}$. In particular, there is setting $(\hat{x}_1, \hat{y}_1), \ldots (\hat{x}_{t_0-1}, \hat{y}_{t_0-1})$, such that

$$\mathbb{P}_{(x_{t_0}, y_{t_0}) \sim \mu}(\neg A_{t_0} | \forall i < t_0,\ (x_i, y_i) = (\hat{x}_i, \hat{y}_i)) \leq \frac{\delta}{k}$$

Now, Alice and Bob can fix those $(\hat{x}, \hat{y})$, and use the restriction of protocol P to the k-th round as a single round one way communication protocol for $\mathcal{R}$. As described above, failure probability of this protocol is bounded by $\frac{\delta}{k}$. □

In what follows we will use T-games associated with following relation.

Definition 29 (Approximate distinct elements relation)**.** *We define relation $F_0^\varepsilon \subset \{0,1\}^n \times \{0,1\}^n \times \mathbb{Z}$, to be $F_0^\varepsilon = \{(x, y, z) : (1-\varepsilon)|x \vee y| \leq z \leq (1+\varepsilon)|x \vee y|\}$.*

The one-way communication complexity of this relation, in the low failure probability range, can be lower bounded as follows.

Theorem 30 ([JW13])**.** *For every ε there is a distribution μ over $(x, y) \in \{0,1\}^n \times \{0,1\}^n$, such that $D^{\to}_{\mu,\delta}(F_0^\varepsilon) = \Omega(\frac{\lg \delta^{-1}}{\varepsilon^2})$. Moreover, this distribution is supported on vectors with $|x| = |y| = \frac{n}{2}$*

It is enough now to show that strong tracking algorithm for distinct elements can be leveraged to obtain efficient protocols for T-game based on relation F_0^ε.

Lemma 31. *If there is a randomized streaming algorithm using space S for $(1+\varepsilon)$-strong tracking distinct elements on the universe of size $O(n^2)$, which succeeds with probability $\frac{2}{3}$, then for any distribution μ supported on pairs $(x, y) \in \{0,1\}^n \times \{0,1\}^n$ of vectors with Hamming weight $\frac{n}{2}$ we have $D_{\mu,\delta}(F_0^{2\varepsilon}) \leq S$.*

Proof. Indeed, consider universe U partitioned into subsets $U_1 \cup U_2 \cup \ldots \cup U_k$, such that $|U_1| = n$, and $|U_i| = 8|U_{i-1}|$. We can take $k = \Theta(\lg n)$ such that $|U| \leq n^2$. Moreover, for each $t \leq k$, consider a partition of U_t into n sets $U_t = U_t^1 \cup \ldots \cup U_t^n$ with $|U_t^i| = 8^{t-1}$. The players are going to pass between each other the memory content of the streaming algorithm. On the t-th round, Alice takes her input x_k, and feeds to the algorithm all the elements $A_t := \bigcup_{i:(x_k)_i=1} U_t^i$, then she sends the memory content to Bob, who in turn feeds to the algorithm set $B_t := \bigcup_{i:(y_k)_i=1}$, and reads off the answer w.

Let $P_t := \bigcup_{s<t} A_t \cup B_t$, and $D_t = A_t \cup B_t$. Note that Bob knows $|P_t|$, and moreover $|P_t| \leq \sum_{s<t} n8^s \leq \frac{1}{4}n8^t \leq \frac{|D_t|}{4}$, where the last inequality follows from the fact that all vectors x_i under consideration have Hamming weight exactly $\frac{n}{2}$.

By the correctness guarantee of the tracking algorithm, w is a good approximation of $|P_t \cup D_t|$, i.e. $w = (1 \pm \varepsilon)(|P_t \cup D_t|)$. Bob can estimate $|D_t|$ by $w - |P_t|$. Indeed: $w - |P_t| \leq (1+\varepsilon)(|D_t|) + \varepsilon|P_t| \leq (1+2\varepsilon)|D_t|$. Bob can report this estimate to the judge, and send the memory content of the algorithm back to Alice. □

Theorem 32. *Any algorithm satisfying $(1+\varepsilon)$ strong tracking of F_0 with failure probability at most $\frac{1}{3}$, needs to use at least $\Omega(\frac{\lg \lg n}{\varepsilon^2})$ bits of space*

Proof. The statement of this theorem follows directly by composing Lemma 31, Lemma 28 and Theorem 30. □

8 Pseudorandom construction

In this section we will prove Lemma 10. Before we proceed with the proof, let us introduce a useful definition.

Definition 33. *A function* $\Gamma : \{0,1\}^s \times [w] \to [M]$ *is called* γ_0*-strong sampler, if for any function* $f : [M] \to [0,1]$ *and random variables* $Y_i := \Gamma(U, i)$ *generated by supplying uniformly random* U*, we have for any* $2 < \gamma$

$$\mathbb{P}(|\sum_{i \le w} f(Y_i) > \mu\gamma) \le \exp(-\Omega(\mu\gamma \lg \min\{\gamma, \gamma_0\}))$$

and moreover for any fixed i*, we have* $\Gamma(U, i) \sim \mathrm{Unif}([M])$*.*

The definition above is non-vacuous — as it has been recently shown, standard pseudorandom constructions of samplers actually satisfy our definition of the strong sampler.

Theorem 34 ([RR17, Wag08])**.** *Random walk over a finite regular undirected graph with second largest eigenvalue* λ*, yields a* λ^{-1}*-strong sampler. This implies explicit* γ*-strong samplers* $\Gamma : \{0,1\}^s \times [w] \to [M]$ *with seed length* $s = \lg M + \mathcal{O}(w \lg \gamma)$*.*

In [RR17] bounds on the moments generating functions of $\sum Y_i$ are proven, instead of the tail bounds that appear in our definition of strong-sampler. They proved

$$\mathbb{E} \exp(\theta \sum f(Y_i)) \le \exp\left(2\mu(e^\theta - 1)\right)$$

for $\theta \le \ln \lambda^{-1} - 1$. There is a standard way of deducing tail bounds of the form required for strong samplers from this MGF bound

$$\mathbb{P}(|\sum_{i \le w} f(Y_i)| > \mu\gamma) \le \exp(-\mu\gamma\theta)\mathbb{E}\exp(\theta \sum f(Y_i)) \le \exp(-\mu\gamma\theta + 2\mu(e^\theta - 1))$$

we can plug in $\theta := \ln \gamma^* - \ln 4$, where $\gamma^* := \min\{\gamma, \gamma_0\}$ to get $\mathbb{P}(\sum_{i \le w} f(Y_i) > \mu\gamma) \le \exp(-\frac{1}{2}\mu\gamma \ln \gamma^*)$.

We will now show that sums of strongly concentrated random variables, sampled according to a strong sampler still satisfy similar type of tail-bounds as if they were sampled independently at random.

Lemma 35. *If* $\Gamma : \{0,1\}^s \times [w] \to [M]$ *is* $\exp(\ln^2 w)$*-strong sampler, and* $f : [M] \to \mathbb{R}_+$ *satisfies doubly exponential tail bounds* $\mathbb{P}_{X \sim \mathrm{Unif}([M])}(X > \gamma) \le \exp(-e^\gamma)$*, then*

$$\mathbb{P}(\sum_{i \le w} f(Y_i) > Cw) < \exp(-\Omega(w))$$

for some universal constant C*.*

Proof. Take some T_0, sufficiently large constant, and consider a sequence $T_k = 2^k T_0$, together with functions $f_k : [M] \to \{0,1\}$ given by $f_k(x) := [x > T_k]$. Let $\mu_k := w \,\mathbb{E} f_k(X)$, and notice that because of the assumed tail bounds on function f we have $\mu_k \le w \exp(-e^{2^k T_0})$.

We can bound value of $\sum f(Y_i)$ in terms of f_k as follows

$$\sum f(Y_i) \le \mathcal{O}(w) + T_0 \sum_{k \ge 0} 2^k \sum_{i \le w} f_k(Y_i).$$

We shall bound all terms $\sum_i f_k(Y_i)$ separately. Let us take k_1 smallest such that $\mu_{k_1} \le 1$ (i.e. $T_{k_1} \approx \ln\ln w$) and k_2 smallest such that $\mu_{k_2} \le \exp(-w)$ — we have $T_{k_2} = \Theta(\lg w)$.

Firstly, by Markov inequality $\mathbb{P}(\sum_{i \le w} f_{k_2}(Y_i) \ge 1) \le \mu_{k_2} \le \exp(-w)$, so with probability $1 - \exp(-w)$, we have $\sum f(Y_i) \le \mathcal{O}(w) + \sum_{k \le k_2} 2^k \sum_{i \le w} f_k(Y_i)$.

We will bound terms between 0 and k_1, and terms in the range k_1 and k_2 separately. For $k_1 < k < k_2$, we can use the Chernoff-type inequality guaranteed by the sampler. Indeed, for $k > k_1$, we have $\mu_k < w \exp(-e^{2 \ln\ln w}) < \exp(-\frac{1}{2}(\ln w)^2)$, and therefore if we pick $\gamma_k := \frac{w}{\ln^2 w}\mu_k^{-1} > \exp((\ln w)^2)$ we have by the definition of strong sampler

$$\mathbb{P}(\sum_{i \le w} f_k(Y_i) > w/\lg^2 w) < \exp(-\Omega(\frac{w}{\lg^2 w} \lg^2 w)) = \exp(-\Omega(w)).$$

If this (exponentially unlikely) event does not hold for any k in this range, we have $\sum_{k_1<k\leq k_2}\sum_{i\leq w} T_k f_k(Y_i) \leq \sum_{k_1\leq k\leq k_2} T_0 \ln w \frac{w}{\ln^2 w} = \mathcal{O}(w)$, because $k_2 = \mathcal{O}(\lg w)$.

Let us now focus on the range $k < k_1$, and let us consider γ_k such that $\gamma_k \mu_k = 3^{-k} w$. For $k < k_1$ we have $\gamma_k < w < \exp(\ln^2 w)$, so the sampler guarantee gives us

$$\mathbb{P}(\sum_{i\leq w} f(Y_i) > \gamma_k \mu_k) \leq \exp(-3^{-k} w \lg \gamma_k) \tag{14}$$

Clearly, if neither of those events hold, we have

$$\sum_{k\leq k_1}\sum_{i\leq w} T_k f(Y_i) \leq \sum_{k\leq k_1} T_0 2^k 3^{-k} w = \mathcal{O}(w)$$

It is enough to bound the failure probability in Eq. (14). We have $\lg \gamma_k = -k\lg 3 + \lg w - \lg \mu_k = -k \lg 3 + \lg \mathbb{P}(f(X) > T_k) > k \lg 3$. As such, for any fixed k, we have $\mathbb{P}(\sum_i \leq w f(Y_i) > \gamma_k \mu_k \leq \exp(-w)$, and by union bound the failure probability is bounded by $\exp(-w + \lg k_1) < \exp(-\Omega(w))$. □

First, let us observe that Theorem 34 and Lemma 35 implies that

Lemma 36. *For $w_2 \geq K \lg R$, there exist an explicit function $\Xi_0 : \{0,1\}^{s_0} \times [w_2] \to [M]$ such that set $\{\Xi_0(U,1), \ldots \Xi_0(U, w_2)\}$ is C-small except with probability $\exp(-cw_2)$.*

The seed length is $s_0 = \mathcal{O}(\lg M + w_2 \lg^2 w_2)$.

Proof. Consider Ξ_0 as in Theorem 34 with parameter $\lambda = \mathcal{O}(\exp(\lg^2 w_2))$. We know that for every specific g_t, with probability $\exp(-\Omega(w_2))$, the sum over the generated sequence satisfies $\sum_{t\leq w_2} g_t(\Xi_0(U,i)) \leq Cw_2$. We can union bound over all $t \leq R$, so the probability that $\Xi_0(U,*)$ fails to be C-small is bounded by $\exp(-\Omega(w_2) + \lg R) = \exp(-\Omega(w_2))$, as long as K in the statement of the lemma is sufficiently large constant. □

In what follows we will use as a building block the construction guaranteed by the following theorem

Theorem 37. *[Zuc97, GUV09] There exist an explicit (ε, δ)-averaging sampler $\Gamma\{0,1\}^s \times [W] \to [M]$, with $s = \mathcal{O}(\lg M + \lg \delta^{-1})$ and $W = \mathrm{poly}(\varepsilon^{-1}, \lg \delta^{-1})$.*

We shall use such a sampler to subsample a set of seeds for the expander random walks discussed in Lemma 36. We can ensure that except with probability $\exp(-\Omega(w))$ the subsampled set of seeds has the same fraction of seeds generating C-small sets.

Lemma 38. *For any $w > KR$ and w_2 satisfying $K \lg R < w_2 < w$, there exist c and an explicit function $\Xi_1 : \{0,1\}^{s_1} \times [W] \to [M]^{[w_2]}$ such that*

$$\mathop{\mathbb{P}}_{x\in \mathrm{Unif}(\{0,1\}^{s_1})}(\#\{w : G_1(x,w) \text{ is not } C\text{-small}\} > 2\exp(-cw_2)W)$$

The seed length here is $s_1 := \mathcal{O}(\lg M + w)$ and $W = w^{\mathcal{O}(1)}$.

Proof. Take Γ to be $(\exp(-cw_2), \exp(-w))$-averaging sampler, where c is such that Ξ_0 from Lemma 36 provides a C-small set except with probability $\exp(-cw_2)$. Consider function $\Xi_1(S,i) = \Xi_0(\Gamma(S,i),*)$, i.e. $[\Xi_1(S,i)]_j = \Xi_0(\Gamma(S,i),j)$. The required properties follow from definition of the averaging sampler applied to the indicator function of $f : \{0,1\}^{s_0} \to \{0,1\}$ with $f(s) = 1$ if and only if $\Xi_0(S,*)$ yields a C-small sequence. □

Finally, we are ready to prove the main lemma in this section.

Proof of Lemma 10. Given $w > KR^{1/2}$, take $w_2 = \Theta(\lg w)$ large enough to apply Lemma 38, and w_1 large enough with $w_1 w_2 = \Theta(w)$. Take Ξ_1 as in the Lemma 38. Consider the decomposition of the seed $S \in \{0,1\}^s$ as $S = (S_1, S_2) \in \{0,1\}^s = \{0,1\}^{s_1+s_2}$, and let us focus on collection $\mathcal{A} = \Xi_0(S_1, *)$ of W sequences in $[M]$. We know that, except with probability $\exp(-\Omega(w))$ over choice of the seed S_1, we most of the sequences in $\mathcal{A}$ is C-small — only $2\exp(-cw_2)$ fraction of all sequences is not C-small. Let us use S_2 to pick a uniformly random sequence of indices from $[W]$ — to achieve this, we need $s_2 = \Theta(w_1 \lg W)$. Note that if $\mathcal{A}$ indeed satisfies that $\#\{i : [\mathcal{A}]_i \text{ is } C\text{-small} < 2\exp(-cw_2)\}$, then for a uniformly random indices $i_1, i_2, \ldots i_{w_1} \in [W]$, we have

$$\mathbb{P}(\#\{j : [\mathcal{A}]_{i_j}] \text{ is } C\text{-small}\} > w_1/2) \leq \binom{w_1}{w_1/2} \exp(-w_2)^{w_1/2} \leq 2^{w_1} \exp(-\Omega(w)) = \exp(-\Omega(w))$$

The total seed length is $s = s_1 + s_2 = \mathcal{O}(\lg N + w) + \mathcal{O}(w_1 \lg W) = \mathcal{O}(\lg N + w + \frac{w}{\lg w} \lg w) = \mathcal{O}(\lg N + w)$. □

Acknowledgement The author thanks Raghu Meka for answering questions about the [Mek17] sampler construction, which inspired the proof of Lemma 10, Preetum Nakkiran for helpful discussion during various stages of the work, Thibaut Horel for comments on parts of the writeup. The author is especially grateful to Jelani Nelson for many inspiring and helpful discussions and comments.

References

[AMS96] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. In Gary L. Miller, editor, *Proceedings of the Twenty-Eighth Annual ACM Symposium on the Theory of Computing, Philadelphia, Pennsylvania, USA, May 22-24, 1996*, pages 20–29. ACM, 1996.

[BC09] Joshua Brody and Amit Chakrabarti. A multi-round communication lower bound for gap hamming and some consequences. *Electronic Colloquium on Computational Complexity (ECCC)*, 16:15, 2009.

[BCI+17] Vladimir Braverman, Stephen R. Chestnut, Nikita Ivkin, Jelani Nelson, Zhengyu Wang, and David P. Woodruff. BPTree: an ℓ_2 heavy hitters algorithm using constant memory. In *Proceedings of the 36th SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*, 2017.

[BCWY16] Vladimir Braverman, Stephen R. Chestnut, David P. Woodruff, and Lin F. Yang. Streaming space complexity of nearly all functions of one variable on frequency vectors. In *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS)*, pages 261–276, 2016.

[BDN17] Jaroslaw Blasiok, Jian Ding, and Jelani Nelson. Continuous monitoring of ℓ_p norms in data streams. *CoRR*, abs/1704.06710, 2017.

[BJK+02] Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, D. Sivakumar, and Luca Trevisan. Counting distinct elements in a data stream. In José D. P. Rolim and Salil P. Vadhan, editors, *Randomization and Approximation Techniques, 6th International Workshop, RANDOM 2002, Cambridge, MA, USA, September 13-15, 2002, Proceedings*, volume 2483 of *Lecture Notes in Computer Science*, pages 1–10. Springer, 2002.

[BR94] Mihir Bellare and John Rompel. Randomness-efficient oblivious sampling. In *Proceedings of the 35th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 276–287, 1994.

[DF03] Marianne Durand and Philippe Flajolet. *Loglog Counting of Large Cardinalities*, pages 605–617. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003.

[EVF06] Cristian Estan, George Varghese, and Michael E. Fisk. Bitmap algorithms for counting active flows on high-speed links. *IEEE/ACM Trans. Netw.*, 14(5):925–937, 2006.

[FFGM07] Philippe Flajolet, Éric Fusy, Olivier Gandouet, and Frédéric Meunier. Hyperloglog: the analysis of a near-optimal cardinality estimation algorithm. In *AofA: Analysis of Algorithms*, pages 137–156. Discrete Mathematics and Theoretical Computer Science, 2007.

[FM83] Philippe Flajolet and G. Nigel Martin. Probabilistic counting. In *24th Annual Symposium on Foundations of Computer Science, Tucson, Arizona, USA, 7-9 November 1983*, pages 76–82. IEEE Computer Society, 1983.

[Gar07] D. J. H. Garling. *Inequalities: A Journey into Linear Analysis.* Cambridge University Press, 2007.

[GG81] Ofer Gabber and Zvi Galil. Explicit constructions of linear-sized superconcentrators. *Journal of Computer and System Sciences*, 22(3):407–420, 1981.

[Gib01] Phillip B. Gibbons. Distinct sampling for highly-accurate answers to distinct values queries and event reports. In *VLDB 2001, Proceedings of 27th International Conference on Very Large Data Bases, September 11-14, 2001, Roma, Italy*, pages 541–550. Morgan Kaufmann, 2001.

[Gil98] David Gillman. A chernoff bound for random walks on expander graphs. *SIAM J. Comput.*, 27(4):1203–1220, August 1998.

[GT01] Phillip B. Gibbons and Srikanta Tirthapura. Estimating simple functions on the union of data streams. In *SPAA*, pages 281–291, 2001.

[GUV09] Venkatesan Guruswami, Christopher Umans, and Salil P. Vadhan. Unbalanced expanders and randomness extractors from parvaresh-vardy codes. *J. ACM*, 56(4):20:1–20:34, 2009.

[HTY14] Zengfeng Huang, Wai Ming Tai, and Ke Yi. Tracking the frequency moments at all times. *CoRR*, abs/1412.1763, 2014.

[JW13] T. S. Jayram and David P. Woodruff. Optimal bounds for johnson-lindenstrauss transforms and streaming problems with subconstant error. *ACM Trans. Algorithms*, 9(3):26:1–26:17, 2013.

[KNW10] Daniel M. Kane, Jelani Nelson, and David P. Woodruff. An optimal algorithm for the distinct elements problem. In Jan Paredaens and Dirk Van Gucht, editors, *Proceedings of the Twenty-Ninth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2010, June 6-11, 2010, Indianapolis, Indiana, USA*, pages 41–52. ACM, 2010.

[Mek17] Raghu Meka. Explicit resilient functions matching ajtai-linial. In Philip N. Klein, editor, *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2017, Barcelona, Spain, Hotel Porta Fira, January 16-19*, pages 1132–1148. SIAM, 2017.

[RR17] Shravas Rao and Oded Regev. A sharp tail bound for the expander random sampler. *CoRR*, abs/1703.10205, 2017.

[Vad12] Salil P. Vadhan. Pseudorandomness. *Foundations and Trends in Theoretical Computer Science*, 7(1-3):1–336, 2012.

[Wag08] R. O. Y. Wagner. Tail estimates for sums of variables sampled by a random walk. *Combinatorics, Probability & Computing*, 17(2):307–316, 2008.

[Woo04] David Woodruff. Optimal space lower bounds for all frequency moments. In *Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '04, pages 167–175, Philadelphia, PA, USA, 2004. Society for Industrial and Applied Mathematics.

[Zuc97] David Zuckerman. Randomness-optimal oblivious sampling. *Random Struct. Algorithms*, 11(4):345–367, 1997.

A Appendix

In this section, for the sake of completeness, we will discuss space complexity of the optimal algorithm with constant probability proposed in [KNW10] — we use it as a building block in Section 5. The existence of the algorithm as described below was proven in [KNW10], as well as the fact that it returns correct answer with large constant probability. In what follows we describe the KNW algorithm, and provide more detailed analysis of the space complexity of this algorithm — in the original paper, it was shown only that for $\frac{1}{\varepsilon^2} > \lg n$, the total space consumption is $\mathcal{O}(\frac{1}{\varepsilon^2})$ with large constant probability. The condition $\frac{1}{\varepsilon^2} > \lg n$ could have been assumed without loss of generality in the original setting, it is not the case in our application.

The correctness of this algorithm (with large constant probability) has been shown in [KNW10], it also follows from the proofs in Section 6. In particular in the proof of Theorem 5 it is shown how to deduce strictly stronger statement. We do not discuss it in this appendix.

Theorem 12. *There is an algorithm $\mathcal{F}_\varepsilon$ which gives a $(1+\varepsilon)$-approximation to $F_0^{(t)}$ with probability at least $\frac{5}{6}$, assuming access to an oracle providing strong tracking of $F_0^{(t)}$ with constant factor approximation C, and oracle access to $\mathcal{O}(\lg n + \operatorname{poly}\lg(1/\varepsilon))$ additional random bits. The space usage of this algorithm at any given time t (excluding random bits mentioned above), denoted by $W^{(t)}$, satisfies*

$$\mathbb{P}(W^{(t)} > \frac{C_2}{\varepsilon^2}) \leq \varepsilon^4 \tag{2}$$

and

$$\mathbb{P}(W^{(t)} > \frac{\lambda}{\varepsilon^2}) \leq \exp(-e^{\Omega(\lambda)}). \tag{3}$$

Moreover for $t_1 < t_2$ such that $|S^{(t_1)}| \geq |S^{(t_2)}|/2$ we have

$$W^{(t_1)} \leq W^{(t_2)} + \mathcal{O}(\frac{1}{\varepsilon^2}) \tag{4}$$

for some universal constant C.

Proof. Consider some $P = \frac{C_0}{\varepsilon^2}$ with large constant C_0 (depending on C) and some constant D_0 to be specified later. We will pick a random pairwise independent hash function $h_3 : [n] \to [P^2]$, and random polylog(P)-wise independent hash function $h_4 : [P^2] \to [P]$. We set $h_2 := h_4 \circ h_3$, and we take $h_1 : [n] \to [n]$ to be 8-wise independent hash function. The total number of random bits necessary to access is $\mathcal{O}(\lg n + \operatorname{poly}\lg P) = \mathcal{O}(\lg n + \operatorname{poly}\lg \frac{1}{\varepsilon})$.

We assume access to $\tilde{F}_0^{(t)}$ such that for every t we have $F_0^t \leq \tilde{F}_0^{(t)} \leq CF_0^{(t)}$. For each $i \in [P]$ we consider $Z_i^{(t)} := \max\{\operatorname{lsb}(h_1(s)) : s \in S^{(t)}, h_2(s) = i\}$, and we store $\hat{Z}_i^{(t)} := \max\{-1, Z_i^{(t)} - D^{(t)}\}$, where $D^{(t)} := \lg \tilde{F}_0^{(t)} - \lg \frac{1}{\varepsilon^2} - D_0$. At time t, we consider $Q^{(t)} := \#\{i : Z_i^{(t)} \geq 0\}$, and we report $\hat{F}^{(t)} := \Phi_K^{-1}(Q^{(t)}) 2^{D^{(t)}}$.

Let us consider total space used by all the counters Z_i. We need to use $\mathcal{O}(P)$ bits to store all counters for which value of $Z_i^{(t)}$ is -1, and space necessary to store counters with $Z_i^{(t)} \geq 0$ is bounded by $\sum_{i \in S^{(t)}} \lg \max(1, \operatorname{lsb}(h_1(s)) - D^{(t)})$. Hence, the space used by the algorithm at time t is bounded as

$$W^{(t)} \leq \mathcal{O}(P) + \sum_{i \in S^{(t)}} \lg \max(1, \operatorname{lsb}(h_1(s)) - D^{(t)}) \leq \mathcal{O}(P) + \sum_{i \in S^{(t)}} \max(0, \operatorname{lsb}(h_1(s)) - D^{(t)})$$

Note that for fixed t, variables $K_s := \max(0, \operatorname{lsb}(h_1(s)) - D^{(t)})$ are 8-wise independent (because h_1 was). For fixed s the random variable K_s have strongly decaying tails, i.e. $\mathbb{P}(K_s > \lambda) = \mathbb{P}(\operatorname{lsb}(h_1(s)) > D^{(t)} + \lambda) \leq 2^{-D^{(t)} - \lambda} \leq \frac{2^{D_0}}{\varepsilon^2 \hat{F}^{(t)}} 2^{-\lambda}$. We can apply Lemma 43 to appropriately rescaled K_s, with $p = 8$ and $\mu_i = \frac{2^{D_0}}{\varepsilon^2 \hat{F}^{(t)}}$. With this choice of μ_i we have $\frac{1}{\varepsilon^2} \cdot \frac{2^{D_0}}{C} \leq \sum \mu_i \leq \frac{1}{\varepsilon^2} 2^{D_0}$, so the conditions of this lemma are satisfied, and the conclusion of the lemma yields

$$\mathbb{P}(\sum_{s \in S^{(t)}} K_s > C_1 \frac{1}{\varepsilon^2}) \leq C_2 \varepsilon^4 2^{-D_0}$$

We can now take D_0 such that that $2^{-D_0} C_2 \le 1$ to finish the proof of Eq. (2).

Let us now turn our attention to the proof of (3). Observe that $\mathbb{P}(W^{(t)} > \frac{\lambda}{\varepsilon^2}) \le \mathbb{P}(\exists i \in [P], s.t. \lg \hat{Z}_i > C_1 \lambda) = \mathbb{P}(\exists s \in S^{(t)}, s.t. \operatorname{lsb}(h_1(s)) > D^{(t)} + 2^{C_1 \lambda})$, where C_1 is a constant depending on C_0. Using union bound, this latter quantity is bounded as follows

$$\mathbb{P}(\exists s \in S^{(t)}, s.t. \operatorname{lsb}(h_1(s)) > D^{(t)} + 2^{C\lambda}) \le F_0^{(t)} \, \mathbb{P}(\operatorname{lsb}(h_1(s_0)) > P^{(t)} + 2^{C_1 \lambda}) \le \frac{2^{D_0} C}{\varepsilon^2} 2^{-2^{C_1 \lambda}}.$$

Which yields a bounds of form $\mathbb{P}(W^{(t)} \ge \frac{\lambda}{\varepsilon^2}) \le \frac{C_2}{\varepsilon^2} 2^{-2^{C_1 \lambda}}$. On the other hand, by Lemma 43 we have $\mathbb{P}(W^{(t)} \ge \frac{\lambda}{\varepsilon^2}) \le \frac{\varepsilon^4}{\lambda^8}$. We can combine those two bounds for different ranges of λ to get $\mathbb{P}(W^{(t)} \ge \frac{\lambda}{\varepsilon^2}) \le \exp(-e^{\Omega(\lambda)})$. Indeed, for $\lambda < \lg \lg \frac{1}{\varepsilon^2}$ we already have $\frac{1}{\varepsilon^4} < \exp(-e^{\Omega(\lambda)})$ whereas for $\lambda > \lg \lg \frac{1}{\varepsilon^2}$ we have $\frac{1}{\varepsilon^2} \exp(-e^{\lambda}) < \exp(-e^{\lambda/2})$.

Finally, to show (4), note that $W^{(t_1)} \le \sum_i \lceil \lg(Z_i^{(t_1)} - D^{(t_1)}) \rceil \le \sum_i \lceil \lg(Z_i^{(t_2)} - D^{(t_1)}$, because $Z_i^{(t_2)} \ge Z_i^{(t_1)}$. By subadditivity of logarithm, we have $W^{(t_1)} \le \sum_i \lceil \lg(Z_i^{(t_2)} - D^{(t_2)}) \rceil + P(D^{(t_2)} - D^{(t_1)}) \le W^{(t_2)} + \mathcal{O}(P)$, where $(D^{(t_2)} - D^{(t_1)} = \mathcal{O}(1)$ follows from the fact that $D^{(t_2)} - D^{(t_1)} = \lg \frac{\hat{F}_0^{(t_2)}}{\hat{F}_0^{(t_1)}}$, and since $\hat{F}^{(t)}$ is a constant approximation to $F_0^{(t)}$, this quantity is bounded by $\lg \frac{F_0^{(t_2)}}{F_0^{(t_1)}} + \mathcal{O}(1)$, and by assumption on t_2, t_1, we have $\lg \frac{F_0^{(t_2)}}{F_0^{(t_1)}} \le \mathcal{O}(1)$. □

A.1 Probabilistic inequalities

Lemma 39. *Let $Z_1, \ldots Z_k$ be a sequence of non-negative, p-wise independent random variables (for some even p), satisfying $\|Z\|_p \le C$. Then*

$$\| \sum (Z_i - \mathbb{E} Z_i) \|_p \lesssim C \sqrt{p} \sqrt{k}$$

Proof. Take Y_i independent, with marginal distribution $Y_i \sim Z_i - \mathbb{E} Z_i$. Because $\| \sum (Z_i - \mathbb{E} Z_i) \|_p^p$ is a polynomial of degree p in variables Z_i, and Y_i are independent, it follows that $\| \sum (Z_i - \mathbb{E} Z_i) \|_p = \| \sum Y_i \|_p$, and it is enough to bound this second quantity. We can use symmetrization argument, to deduce that $\| \sum Y_i \|_p \lesssim \| \sum \varepsilon_i Y_i \|_p$ where ε_i are independent random signs.

Indeed, consider $\tilde{Y}_i$ distributed identically as Y_i and independent from those, then

$$\begin{aligned} \| \sum Y_i \|_p = \| \sum (Y_i - \mathbb{E} \tilde{Y}_i) \|_p &\le \| \sum (Y_i - \tilde{Y}_i) \|_p \\ &= \| \sum \varepsilon_i (Y_i - \tilde{Y}_i) \|_p \le 2 \| \sum \varepsilon_i Y_i \|_p. \end{aligned}$$

We can now condition on Y_i and use Khnitchine inequality [Gar07, Theorem 12.3.1] to bound

$$\begin{aligned} \| \sum \varepsilon_i Y_i \|_p &= \left(\mathbb{E} (\sum \varepsilon_i Y_i)^p \right)^{1/p} \\ &\lesssim \sqrt{p} \left(\mathbb{E} (\sum Y_i^2)^{p/2} \right)^{1/p} \\ &= \sqrt{p} \sqrt{\| \sum Y_i^2 \|_{p/2}} \\ &\le \sqrt{p} \sqrt{\sum \| Y_i^2 \|_{p/2}} \\ &\le \sqrt{p} \sqrt{\sum \| Y_i \|_p^2} \\ &\le C \sqrt{p} \sqrt{k} \end{aligned}$$

□

Lemma 40. *For every p there exist $C_p, \tilde{C}_p$ such that if non-negative independent random variables $Z_1, \ldots Z_k$ satisfy $\mathbb{E}\, Z_i^s \le \mu_i$ for all $1 \le s \le p$, where $\sum \mu_i \ge 1$ then*

$$\|\sum Z_i\|_p \le C_p \sum \mu_i \tag{15}$$

and moreover

$$\|\sum (Z_i - \mathbb{E}\, Z_i)\|_p \le \tilde{C}_p \sqrt{\sum \mu_i} \tag{16}$$

Proof. It is enough to prove ineqialities (15) and (16) for all values p that are powers of two. We will proceed by showing (15) by induction over p. The case $p = 1$ is trivial: $\|\sum Z_i\|_1 = \sum \mathbb{E}\, Z_i \le \sum \mu_i$. For $p > 1$, let us take $Y_i := Z_i - \mathbb{E}\, Z_i$. We have

$$\|\sum Z_i\|_p \le \sum \mathbb{E}\, Z_i + \|\sum Y_i\|_p \tag{17}$$

Now we can use standard symmetrization argument to bound $\|\sum Y_i\|_p$. Let us take $\tilde{Y}_i$ to be independent random variables with the same distribution as Y_i, and ε_i to be independent uniform ± 1 random variables. We have

$$\begin{aligned}\|\sum Y_i\|_p &= \|\sum (Y_i - \mathbb{E}\, \tilde{Y}_i)\|_p \le \|\sum (Y_i - \tilde{Y}_i)\|_p \\ &= \|\sum \varepsilon_i (Y_i - \tilde{Y}_i)\|_p \le 2\|\sum \varepsilon_i Y_i\|_p \end{aligned} \tag{18}$$

We can now condition on Y_i and use Khintchine inequality to deduce

$$\|\sum \varepsilon Y_i\|_p = \left(\mathbb{E}(\sum \varepsilon_i Y_i)^p\right)^{1/p} \lesssim \sqrt{p}\|\sum Y_i^2\|_{p/2}^{1/2} \le \sqrt{p}\|\sum Z_i^2\|_{p/2}^{1/2}$$

By applying inductive hypothesis to random variables Z_i^2 we obtain

$$\|\sum Y_i\|_p \le 2\|\sum \varepsilon Y_i\|_p \lesssim \sqrt{p}\sqrt{C_{p/2}}\sqrt{\sum \mu_i}$$

proving inequality (16). Finally, we can can compose this last inequality with inequality (17), to deduce

$$\|\sum Z_i\|_p \le \sum \mu_i + K\sqrt{p}\sqrt{C_{p/2}}\sqrt{\sum \mu_i} \le (1 + K\sqrt{pC_{p/2}})(\sum \mu_i)$$

which completes the proof of inductive hypothesis with $C_p = (1 + K\sqrt{pC_{p/2}})$. ☐

Lemma 41. *Let Z be a non-negative random variable satisfying for some T, that $\mathbb{P}(Z > \lambda T) \le \mu \exp(-\lambda)$. Then $\mathbb{E}\, Z^p \le p! T^p \mu$.*

Proof. We can assume without loss of generality that $T = 1$. We can bound

$$\begin{aligned}\mathbb{E}\, Z^p &= \int_0^\infty t^{p-1}\, \mathbb{P}(Z > t)\ \mathrm{d}t \\ &\le \mu \int_0^\infty t^{p-1} e^{-t}\ \mathrm{d}t.\end{aligned}$$

Now by repeatedly applying integration by parts, we obtain

$$\begin{aligned}\int_0^\infty t^{p-1} e^{-t}\ \mathrm{d}t &= (p-1)\int_0^\infty t^{p-2} e^{-t}\ \mathrm{d}t \\ &= \cdots \\ &= (p-1)! \int_0^\infty e^{-t}\ \mathrm{d}t = (p-1)!,\end{aligned}$$

which completes the proof of the desired inequality. ☐

Corollary 42. *Let $Z_1, \dots Z_k$ be a sequence of non-negative random variables satisftying for some T that $\mathbb{P}(Z_i > T\lambda) \leq \mu_i \exp(-\lambda)$. Then $\| \sum (Z_i - \mathbb{E} Z_i) \|_p \lesssim \tilde{C}_p T \sqrt{\sum \mu_i}$, where $\tilde{C}_p$ is a constant that depends only on p.*

Proof. Follows directly from Lemma 41 and Lemma 40. □

Lemma 43. *Let $Z_1, \dots Z_k$ be a sequence of p-wise independent non-negative random variables, satisfying $\mathbb{P}(Z_i > \lambda) \leq \mu_i \exp(-\lambda)$. Then for some universal constant K, and constant $\tilde{C}_p$ depending only on p we have for all $\lambda > K$ following tail bound*

$$\mathbb{P}(\sum Z_i > \lambda \sum \mu_i) \leq \frac{1}{\lambda^p} \left(\frac{\tilde{C}_p}{\sum \mu_i} \right)^{p/2}.$$

Proof. Note that for random variables as above we have $\mathbb{E} Z_i \leq K_0 \mu_i$ for some universal constant μ_i. Let us pick $K = 2K_0$, such that $\mathbb{P}(\sum Z_i > \lambda \sum \mu_i) \leq \mathbb{P}(\sum (Z_i - \mathbb{E} Z_i) > \frac{\lambda}{2} \sum \mu_i)$. By Chebyshev inequality we have

$$\mathbb{P}(\sum (Z_i - \mathbb{E} Z_i) > \frac{\lambda}{2} \sum \mu_i) \leq \left(2 \frac{\| \sum (Z_i - \mathbb{E} Z_i) \|_p}{\lambda \sum \mu_i} \right)^p.$$

We can bound the numerator in this expression using Corollary 42, i.e. $\| \sum (Z_i - \mathbb{E} Z_i) \|_p \leq \tilde{C}_p \sqrt{\sum \mu_i}$, to deduce desired probability bound. □