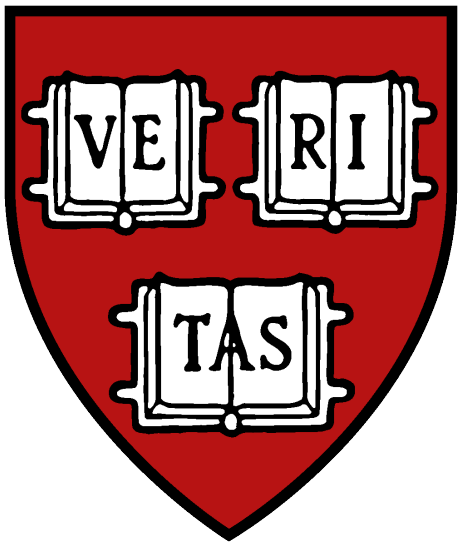
Suprema of Stochastic Processes a Survery in Estimating Frequency Moments of Streams

Suprema of stochastic processes
A survery in estimating frequency moments of streams

Ştefan Spătaru



A thesis submitted in partial fulfillment of the requirements for the degree of
Bachelor of Arts in Mathematics and Computer Science

Harvard University
Cambridge, MA
March 25^{th} 2019

Contents

1. Acknowledgements

I would like to thank Professor Jelani Nelson for the suggestion of the topic, references, and for the guidance throughout the thesis writing process. I would like to thank Jaroslaw Blasiok for helpful conversations about many of the results presented in this paper.

Abstract: In this presentation, we will explore the topic of streaming algorithms. In essence, streaming algorithms are just algorithm that provide low memory approximations for a variety of problems. In this presentation, I will be focusing on 2 related problems: the distinct elements problem and the problem of frequency moment estimation. We will start by providing the mathematical work underlying all of these results, and later delve into the specifics of the algorithms.

2. Introduction

The space of problems we will approach refers to insertion streams. Intuitively, an insertion stream is a data structure data supports insertion of elements, but does not support deletion of elements once they have been inserted. More formally, insertion streams can be seen as sequences of elements over a finite space, i.e., $a_1, ..., a_m \in \{1, ..., n\}$. One is interested in computing functions $f(a_1, ..., a_m)$. In general, computing these functions takes linear space in either the number of elements in the stream or the number of states in the finite space. Consequently, for memory efficiency purposes, one is interested in algorithms that achieve approximations to $f(a_1, ..., a_m)$ and that can fail with certain probabilities. In this presentation, we will be focusing on such algorithms. In particular, we will be interested in low-memory approximation algorithms for the frequency moment estimation problems and the distinct elements problems.

2.1. **Frequency moments.**

Definition 1. Given a stream of m integers a_i in the set $\{1, ..., n\}$, one can define the frequencies of the elements in the state space $\{1, ..., n\}$ as $f_i = |\{j \mid x_j = i\}|$. The kth frequency moment of stream a is defined as $l_k = (\sum_{i=1}^{n} f_i^k))^{\frac{1}{k}}$, i.e. the k-th moment of the frequency vector.

Example 2. A few examples worth mentioning are:

(1) $k = 1$. This is always m, as this counts the number of elements, with multiplicities, seen in the stream.
(2) $k = \infty$. This is just the problem of detecting the maximum frequency element.

2.2. **Distinct elements.**

The problem will be formally defined as seeing a stream of integers $i_1, ..., i_m \in [1, n]$. One's goal is to obtain the value of $F_0 = |\{i_1, ..., i_m\}|$. Given that an exact algorithm requires linear space [1], the problem one is trying to solve is finding an algorithm that approximates F_0 with high probability.

2.2.1. *Relation to frequency moments.*

There is a nice relation between the norms l_k and F_0. Remark that when one lets $F_k = l_k^k$, then $F_0 = \lim_{k \to 0} F_k$. While interesting enough, it is unclear how having estimation algorithms for F_k would translate in the limit to F_0.

2.3. **ε-approximation.**

Definition 3. Say that algorithm $\mathcal{A}$ provides an ϵ-approximation for f if the output s of the algorithm satisfies $|s - f(a_1, ..., a_m)| \leq \epsilon |f(a_1, ..., a_m)|$. Furthermore, say that algorithm $\mathcal{A}$ provides an ϵ-approximation with failure probability δ for f if the output s of the algorithm satisfies $\mathbb{P}\left(|s - f(a_1, ..., a_m)| \leq \epsilon |f(a_1, ..., a_m)|\right) \geq 1 - \delta$

Many of the algorithms we will provide are in fact ϵ-approximations with certain failure probability of the functions we presented. This is because often these

2.4. **Tracking.** Recently, there has been interest in the "tracking" aspect of streaming algorithms. Intuitively, one is interested in providing accurate answers to each one of $f(x_1), ..., f(x_1, ..., x_T)$. More formally, we will use the following definitions of weak and strong tracking:

Definition 4. Say that algorithm $\mathcal{A}$ provides ϵ-**weak tracking** for f with failure probability δ if the outputs s_t at time step $t \leq T$ of the algorithm satisfy

$$\mathbb{P}\left(\exists t \leq T \mid |s_t - f(a_1, ..., a_t)| \leq \epsilon \max_{i \leq T} |f(a_1, ..., a_i)|\right) \leq \delta$$

Furthermore, say that algorithm $\mathcal{A}$ provides ϵ-**strong tracking** for f with failure probability δ if the outputs s_t at time step t of the algorithm satisfy

$$\mathbb{P}\left(\exists t \leq T \mid |s_t - f(a_1, ..., a_t)| \leq \epsilon \max_{i \leq T} |f(a_1, ..., a_i)|\right) \leq \delta$$

Remark. Since we will be working with increasing functions, (frequency moments), $\max_{i \leq T} |f(a_1, ..., a_i)|$ will just be $|f(x_1, ..., x_T)|$ in the presented applications. For increasing functions, weak-tracking and strong-tracking are in fact related up to factors of accuracy. The following result summarizes the relationship:

Theorem 5. [5]
Let f be a positive function that is non-decreasing in t, i.e. $f(a_1, ..., a_{t+1}) \geq f(a_1, ..., a_t)$. Suppose $\mathcal{A}$ provides ϵ-weak tracking with failure probability δ. Then, $\mathcal{A}$ provides 2ϵ-strong tracking for f with failure probability $\delta \cdot \left(2 + \log_2 \frac{f(a_1, ..., a_T)}{f(a_1)}\right)$.

Proof:
Define a sequence t_i inductively. Let $t_0 = 1$ and let t_i be the first point at which $f(a_1, ..., a_{t_i}) \geq 2f(a_1, ..., a_{t_{i-1}})$. Let $t_i = T$ if not such i exists and stop the process. Suppose this finds $t_0 < t_1 \leq ... < t_k$ Then, $t_0 = 1$, and clearly $k \leq \left(1 + \log_2 \frac{f(a_1, ..., a_T)}{f(a_1)}\right)$.
Let s_t be the output of $\mathcal{A}$ at time t. By weak tracking,

$$\mathbb{P}(\exists t \leq t_i - 1; |f(a_1, ..., a_t) - s_t| \geq \epsilon |f(a_1, ..., a_{t_i - 1})|) \leq \delta$$

By union bound,

$$\mathbb{P}\left(\exists 1 \leq i \leq k; \exists t \leq t_i - 1; |f(a_1, ..., a_t) - s_t| \geq \epsilon |f(a_1, ..., a_{t_i - 1})|\right) \leq k\delta$$

Let E be the event $\exists i \leq k; \exists t \leq t_i$; One has that for $t_{i-1} < t \leq t_i$, $|f(a_1, ..., a_{t_{i-1}})| \leq |f(a_1, ..., a_t)|$ and for every $t < T$, there exists $1 \leq i \leq k$ such that $t_{i-1} \leq t < t_i$. Remark that under E, for any $t < T$, E implies that

$$\begin{aligned} |f(a_1, ..., a_t) - s_t| &\leq \epsilon |f(a_1, ..., a_{t_{i+1}-1})| \\ &\leq 2\epsilon |f(a_1, ..., a_{t_i})| \text{ by the choice of } t_i \\ &\leq 2\epsilon |f(a_1, ..., a_t)| \text{ by the monotonicity of } f \end{aligned} \tag{1}$$

Now, $\mathbb{P}(|f(a_1, ..., a_T) - s_T| \geq \epsilon |f(a_1, ..., a_T)|) < \delta$ and thus by union bound and the statement above, $E \cup (|f(a_1, ..., a_T) - s_T| \geq \epsilon |f(a_1, ..., a_T)|)$ cover the cases in which f does not provide 2ϵ-strong tracking. Thus, probability of failure is at most $(k+1)\delta \leq \left(2 + \log_2 \frac{f(a_1, ..., a_T)}{f(a_1)}\right)$.

Next, given that $|\frac{x}{l_p} - 1| = \Theta(|\frac{x^p}{F_p} - 1|)$, one can deduce the following:

Proposition 6. Any guarantee obtained for l_p tracking translates to a guarantee for F_p tracking up to an increase in ϵ by a constant factor.

3. Bounding tails of random variables

In general, bounding tails of random variables is well-understood. For the simple case of $X = X_1 + ... + X_n$, where X_i are independent identically distributed variables, methods to bound these variables are well-understood. The most common way to bound the tails of this distribution is through a Chernoff-bound. As such, $\mathbb{P}(X \geq \lambda) \leq \frac{\mathbb{E}(e^{tX_1})^n}{e^{t\lambda}}$, for any t. One can parameterize t to obtain different bounds on this. The most common application of Chernoff bounds has the following statement:

Theorem 7. *Chernoff bound.* [11]. *Let $(X_i)_{i=1,n}$ be a family of independent random variables, each having values between 0 and 1. Furthermore, let $X = \sum_{i=1}^n X_i$ and let $\mu = \mathbb{E}(X)$. Then,*

(a)

$$\mathbb{P}(|\sum_{i=1}^{n} X_i - \mu| \geq \gamma\mu) \leq e^{-\Omega(\min(\gamma, \gamma^2)\mu)}$$

(b)

$$\mathbb{P}(|\sum_{i=1}^{n} X_i - \mu| \geq a) \leq e^{-\Omega(\min(\frac{a^2}{\mu}, a))}$$

But similar results to Chernoff bounds hold in more general contexts.

3.1. **Negative association.** We will talk about the theory of negative association developed in [10].

Definition 8. Random variables $X_1, ..., X_n$ are said to be negatively associated if and only if for any disjoint sets I, J of $\{1, ..., n\}$ and any functions $f : \mathbb{R}^{|I|} \to \mathbb{R}$, $g : \mathbb{R}^{|J|} \to \mathbb{R}$ component-wise increasing or decreasing on the support of vector X spanned by coordinates $|I|$ and $|J|$, the following inequality holds:

$$\mathbb{E}(f(X_i)g(X_j)) \leq \mathbb{E}(f(X_i))\mathbb{E}(g(X_j))$$

Corollary 9. *If f is an increasing positive function, and $X_1, ..., X_n$ are negatively-associated random variables, then*

$$\mathbb{E}(\prod_{i=1}^{n} f(X_i)) \leq \prod_{i=1}^{n} \mathbb{E}(f(X_i))$$

Proof: Follows by induction from the definition of negatively associated random variables.

3.2. **Bounded independence.**

Definition 10. Say that random variables $X_1, ..., X_n$ are **k-independent** if for any k distinct indices $1 \leq i_1 < ... < i_k \leq n$, $X_{i_1}, ..., X_{i_k}$ are independent.

In the Chernoff bound, the variables are completely independent. There are cases in which one wants to consider variables that are k-independent for some k. The motivation for this type of processes comes from computer science. While generating k-independent hash-functions from set U to set V can be done using a seed of length $\mathcal{O}(k(\log U + \log V))$ as done in [7], generating classes of fully independent hash-functions is more costly in terms of the space needed to store the seed. (as they require $\mathcal{O}(n(\log U + \log V))$ bits of space to store the seed. The methods of proofs for such results will be based of a simple principle: For a degree k polynomial, P in variables $X_1, ..., X_n$, $P(X_1, ..., X_n)$ has the same distribution for k-independent $X_1, ..., X_n$ as it does for fully independent $X_1, ..., X_n$, as long as the 2 families of distributions share the same marginal distributions.

The following result gives exponentially bounded tails for r-independent random variables.

Result 11. [3] *Let $\{X_i\}_{i=1}^n$ be r-independent random variables such that $0 < X_i < 1$. Then, if $\mu = \mathbb{E}(X)$,*

$$\mathbb{P}\left(|X - \mu| \geq A\right) \leq t^{\mathcal{O}(1)} \cdot \left(\left(\frac{t}{A}\right)^t + \left(\frac{t\mu}{A^2}\right)^{\frac{t}{2}}\right)$$

Proof:

The proof will be based off the fact that if $Y = Y_1 + Y_2 + ... + Y_n$, where the Y_i are fully independent and have the same marginal distributions as X_i, the t-th moment $|Y - \mu|^t$ is the same as $|X - \mu|^t$. Now, remark that

$\mathbb{P}(|X-\mu|^t \geq r) = \mathbb{P}(|X-\mu| \geq r^{\frac{1}{t}}) \leq e^{-\Omega(\min(\frac{r^{2/t}}{\mu}, r^{\frac{1}{t}})} = \mathcal{O}(e^{-\Omega(r^{\frac{1}{t}})} + e^{-\Omega(e^{\frac{r^{2/t}}{\mu}})})$

Thus,

$$\begin{aligned}\mathbb{E}(|X-\mu|^t) &\leq \mathcal{O}\left(\int_0^\infty e^{-\Omega(r^{\frac{1}{t}})} dr\right) + \mathcal{O}\left(\int_0^\infty e^{-\Omega(e^{\frac{r^{2/t}}{\mu}})}\right) \\ &= t\mathcal{O}\left(\int_0^\infty e^{-\Omega(x)} x^{t-1} dx\right) + \mathcal{O}\left(\mu^{\frac{t}{2}} \cdot \frac{t}{2} \cdot \int_0^\infty x^{t/2-1} \cdot e^{-\Omega(x)}\right) \\ &= t^{\mathcal{O}(1)} \cdot \mathcal{O}\left(\Gamma(t) + \mu^{\frac{t}{2}} \cdot \Gamma(t/2)\right)\end{aligned} \tag{2}$$

, where Γ just stands for the well-known Gamma function, i.e. $\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt$. Thus,

$$\begin{aligned}\mathbb{P}(|X-\mu| \geq A) &\leq \frac{t^{\mathcal{O}(1)} \cdot \mathcal{O}\left(\Gamma(t) + \mu^{\frac{t}{2}} \cdot \Gamma(t/2)\right)}{A^t} \\ &\leq t^{\mathcal{O}(1)} \cdot \left(\left(\frac{t}{A}\right)^t + \left(\frac{t\mu}{A^2}\right)^{\frac{t}{2}}\right)\end{aligned} \tag{3}$$

Corollary 12. *In the same setup of 11,*

$$\mathbb{P}\left(|X-\mu| \geq \gamma\mu\right) \leq e^{-\Omega(t)} + e^{-\Omega(\lambda^2\mu)}$$

and for $\lambda \geq 2$,

$$\mathbb{P}\left(|X-\mu| \geq \gamma\mu\right) \leq e^{-\Omega(t\log\lambda)} + e^{-\Omega(\mu\lambda\log\lambda)} \leq e^{-\Omega(\min(t,\lambda\mu)\log\lambda)}$$

4. Distributions

4.1. **Throwing balls into bins.** Some of the distributions that will come up in the discussions will be the distributions related to throwing balls into bins.

Definition 13. The distribution $\mathcal{N}(A, I)$ of non-empty-bins is the distribution of bins that have no balls when throwing A balls into I bins. For this distirbution, one will let $X_1, ..., X_A$ be the result of the A balls thrown, and N be the number of non-empty bins after the throws are finalized. We will begin this section with a few preliminary properties of these distributions:

Property 14. $\mathbb{E}(\mathcal{N}(A, I)) = I(1-(1-\frac{1}{I})^A) \geq I(1-e^{-\frac{A}{I}})$. We will let this function be $\Phi_I(A)$ in further applications. We will state a few useful properties of Φ_A.

Property 15. The function Φ_A defined above verifies the following:

(1) Φ_I has Lipschitz properties for $0 \leq A \leq \frac{I}{20}$. In particular, $\frac{1}{2} \cdot |x-y| \leq |\Phi_A(x) - \Phi_A(y)| \leq |x-y|$.
(2) The inverse of Φ_A is $\frac{\ln\left(1-\frac{x}{A}\right)}{\ln(1-1/A)}$.

Property 16. $Var(\mathcal{N}(A,I)) = I \cdot (1-\frac{1}{I})^A) \cdot \left(1-(1-\frac{1}{I})^A\right) + 2\binom{I}{2} \cdot ((1-\frac{1}{2I})^A - (1-\frac{1}{I})^{2A})$.

Property 17. Let $X_1, ..., X_I$ be random Bernoulli variables, where X_i is tracking whether box i is non-empty when throwing A balls into I bins uniformly at random. Then, $X_1, ..., X_I$ are negatively associated.

Proof: One needs to show that for disjoints sets U, V, and non-decreasing functions f, g

$$\mathbb{E}(f(X_U)g(X_V)) \leq \mathbb{E}(f(X_U))\mathbb{E}(g(X_V))$$

, given that the non-increasing case follows from the non-decreasing one, from the fact that the inequalities are equivalent for the pairs of functions (f,g) and $(-f,-g)$. For notation, let $f_i(X_a)$ be the value of $f(Y)$, where Y is a vector equal to X_U, where the i-th position has been changed to 1 if it was 0. Make it analogous for j.

One will proceed by induction on A for all values of increasing f, g.

For 0 balls, the statement is voidly true, as both f and g are constants.

For the induction step, one can compute both the RHS and LHS in terms of the values obtained when throwing $A-1$ balls into I bins. Expectations below are over the process with $A-1$ balls.

$$LHS = \frac{1}{A}\left(\sum_{i\in U}\mathbb{E}(f_i(X_U)g(X_V)) + \sum_{j\in V}\mathbb{E}(f(X_U)g_j(X_V)) + (A-|U|-|V|)\mathbb{E}(f(X_U)g(X_V))\right)$$

$$RHS = \frac{1}{A^2}\left((A-|V|)\sum_{i\in U}\mathbb{E}(f_i(X_U))\mathbb{E}(g(X_V)) + (A-|U|)\sum_{j\in V}\mathbb{E}(f(X_U)g_j(X_V))\right)$$

$$+\frac{1}{A^2}\left(\sum_{i\in A}\sum_{j\in B}\mathbb{E}(f(X_U))\mathbb{E}(g(X_V)) + (A-|U|)(A-|V|)\mathbb{E}(f(X_U))\mathbb{E}(g(X_V))\right)$$

A simple calculation shows that

$$
\begin{aligned}
RHS = \frac{1}{A}\left(\sum_{i\in U}\mathbb{E}(f_i(X_U))\mathbb{E}(g(X_V)) + \sum_{j\in V}\mathbb{E}(f(X_U))\mathbb{E}(g_j(X_V))\right) \\
+ \frac{1}{A}\left((A-|U|-|V|)\mathbb{E}(f(X_U))\mathbb{E}(g(X_V))\right) \\
+ \frac{1}{A^2}\sum_{i\in U}\sum_{j\in V}(\mathbb{E}(f_i(X_U)) - \mathbb{E}(f(X_U)))(\mathbb{E}(g_j(X_V)) - \mathbb{E}(g(X_V))
\end{aligned} \tag{4}
$$

To finish the proof, remark that f_i, g_j are increasing. Also, by definition $f_i \geq f$, $g_j \geq g$ and thus

$$
\begin{aligned}
RHS \geq \frac{1}{A}\left(\sum_{i\in U}\mathbb{E}(f_i(X_U))\mathbb{E}(g(X_V)) + \sum_{j\in V}\mathbb{E}(f(X_U))\mathbb{E}(g_j(X_V))\right) + \\
\frac{1}{A}(A-|U|-|V|)\mathbb{E}(f(X_U))\mathbb{E}(g(X_V))
\end{aligned} \tag{5}
$$

and by the induction hypothesis

$$RHS \geq LHS$$

as every term in the RHS is bigger than the one in the LHS from the induction hypothesis.

Property 18. Upper tail bound: Let X be drawn over $\mathcal{N}(A, I)$. Then, for $\alpha \leq 1$,

$$\mathbb{P}(X - \mathbb{E}(X) \geq \alpha\mathbb{E}(X)) \leq e^{-\frac{\alpha^2}{3}\mathbb{E}(X)}$$

Proof: Let N_i be indicators as before. Then, from 17,

$$\mathbb{P}(X - \mathbb{E}(X) \geq \alpha\mathbb{E}(X)) \leq \frac{\mathbb{E}(e^{t(X-\mathbb{E}(X))})}{e^{t\cdot\alpha\mathbb{E}(X)}} = \frac{\mathbb{E}(e^{tX})}{e^{t(1+\alpha)\mathbb{E}(X)}}$$

and thus

$$\mathbb{P}(X - \mathbb{E}(X) \geq \alpha\mathbb{E}(X)) \leq e^{-\frac{\epsilon^2\mathbb{E}(X)}{3}}$$

, where the last step can be obtained through a method similar to the Chernoff bound. Analogously, one obtains an analogous bound for the lower tail. Thus, one obtains the following bound:

Corollary 19. *For X distributed on $\mathcal{N}(A, I)$, the following holds:*

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq \alpha\mathbb{E}(X)) \leq 2e^{-\frac{\alpha^2}{3}\mathbb{E}(X)}$$

4.1.1. *Limited independence.* One can imagine a family of distribution in which throwing bins into balls is no longer independent, but only k-independent, for some k. This has relevance to streaming algorithms, as k-independent hashing maps are less expensive to store, when compared to maps that are completely independent. One is thus interested in tail bounds of the balls into bins problem when the throws are only k-independent.

Definition 20. Let $\mathcal{N}_k(A, I)$ denote the family of distribution of non-empty bins when one throws A balls into I-bins and any k throws are independent. The most important result refers to the fact that the expectation of any distribution $\mathcal{D} \in \mathcal{N}_k(A, I)$ is very close to $\mathcal{N}(A, I)$ in expectation and variance.

Notation 21. [4]
One will let polynomial $I_d(x) = 1 - \sum_{i=0}^{d} (-1)^i \binom{x}{i}$

Proposition 22. The polynomial I_d has the following properties:

(a) $I_d(0) = 0$
(b) $I_d(x) = 1$ for x between 1 and d.
(c) $|I_d(x)| \leq \mathcal{O}\left(\binom{x+1}{d}\right)$

Proof:
The first 2 claims follow immediately from binomial coefficients identities. More specifically, they follow from the fact that the alternate sign sum of binomial coefficients is 0 for all numbers bigger than 0 and 1 for 0. For the third claim, one needs to remark that the sum $\sum_{i=0}^{k} (-1)^i \binom{x}{i}$ switches signs from k to $k+1$, which is exactly what we wanted to prove.

4.2. **Small number of bins regime.** In computer science, one is particular interested in the regime where the number of balls is small relative to the number of bins. This is because if one wants to create a family of hash functions for example, properties such as small number of collisions are very important and obtained in the regime of small relative number of bins. In the literature, this is taken as $I \geq 20A$. We will list the most important properties of this regime.

Property 23. Under the $I \geq 20A$ regime, $\mathcal{N}(A, I)$ has the following properties:

(a)
$$\mathbb{E}(\mathcal{N}(A, I)) \geq \frac{39}{40} A$$

(b)
$$Var(\mathcal{N}(A, I)) = \mathcal{O}(\frac{A^2}{I})$$

Proof:

(a)
$$\mathbb{E}(\mathcal{N}(A, I)) \geq I(1 - e^{-A/I}) \geq I(A/I - \frac{A^2}{2I^2}) \geq \frac{39}{40} A$$

(b) By Lagrange's theorem,

$$Var(\mathcal{N}(A,I)) \leq I\left(1-(1-\frac{1}{I})^A\right) - (I^2-I)\cdot\frac{1}{I^2}\cdot A\cdot(1-x)^A$$

, where x is between $2/I$ and $2/I - \frac{1}{I^2}$. Thus,

$$Var(\mathcal{N}(A,I)) \leq I\left(1-(1-\frac{1}{I})^A\right) - (I^2-I)\cdot\frac{1}{I^2}\cdot A\cdot\left(1-\frac{2}{I}\right)^A$$

Because $\left(1-\frac{1}{I}\right)^A = 1-\frac{A}{I}+\mathcal{O}(\frac{A^2}{I^2})$, then one has that $Var(\mathcal{N}(A,I)) \leq \mathcal{O}(\frac{A^2}{I})$

Proposition 24. [12] Let X' be drawn over $\mathcal{D} \in \mathcal{N}_{(k+1)}(A,I)$ and X drawn over $\mathcal{N}(A,I)$. Then, in the regime $I \leq \frac{A}{20}$ there exists c such that for $k = \frac{c\log\frac{1}{\epsilon}}{\log\log\frac{1}{\epsilon}}$

$$|\mathbb{E}(X') - \mathbb{E}(X)| \leq \epsilon\mathbb{E}(X)$$

Proof:
One uses the polynomial $I_d(x)$. Remark that $N_i = I_d(A_i) + \mathcal{O}(\binom{A_i}{k+1})$.

$$\mathbb{E}(X-X') = \sum_{i=1}^{I}\mathcal{O}\left(\mathbb{E}\left(\binom{A_i}{k+1}\right)\right)$$

Now, $\mathbb{E}(\binom{A_i}{k+1})$ is the expected number of ways to choose $k+1$ numbers is bin A_i. But this is exactly $\binom{A}{k+1}\cdot\left(\frac{1}{I}\right)^{k+1} \leq \frac{A}{I}\cdot(\frac{e}{20(k+1)})^{-k} = \frac{A}{I}\cdot e^{-k\cdot(\log(20)+\log(k+1))}$.

Now, by appropriately choosing c, one can make this smaller than $\frac{A}{K}\cdot\epsilon^2$ for small enough ϵ. Adding up all of the I bins gives one the bound $|\mathbb{E}(X-X')| \leq \mathcal{O}(\epsilon^2 A)$, which can be made smaller than $\epsilon\mathbb{E}(X)$ for ϵ small enough, since $\mathbb{E}(X) = \Theta(A)$.

Proposition 25. [12]

There exists c such that for $k = \mathcal{O}(\frac{c\log\frac{K}{\epsilon}}{\log\log\frac{K}{\epsilon}})$, $|Var(X') - Var(X)| \leq \epsilon^2$, when X' is drawn over $\mathcal{N}_{2(k+1)}(A,I)$ and X is drawn over the full independent family.

Proof:
Remark that for a $2k$-independent family, by a similar argument as above,

$$X_iX_j = (1-I_d(A_i))(1-I_d(A_j)) + \mathcal{O}\left(\binom{A_i}{K} + \binom{A_j}{K} + \binom{A_i}{K}\cdot\binom{A_j}{K}\right)$$

We already know how to bound $\binom{A_i}{K}, \binom{A_j}{K}$ in expectation from the previous part. Now, the part that is left is bounding the term $\mathbb{E}\left(\binom{A_i}{K}\cdot\binom{A_j}{K}\right)$. This is just $\binom{A}{k+1,k+1}I^{-2(k+1)} \leq \left(\frac{eA}{I(k+1)}\right)^{k+1}$.

Now, one has that

$$|Var(X')-Var(X)| \leq 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} |\mathbb{E}(X_i'X_j'')-\mathbb{E}(X_iX_j)|+|\mathbb{E}(X-X')|+|\mathbb{E}(X^2)-\mathbb{E}(X'^2)|$$

, which can be bounded by $\mathcal{O}(\epsilon^3)$ and thus can be made smaller than ϵ^2 for small enough ϵ.

Corollary 26. [12]

For ϵ small enough, if $X' \sim \mathcal{N}_k(A, I)$, $X \sim \mathcal{N}(A, I)$ for $I = \frac{1}{\epsilon^2}$, $k = \frac{\log 1/\epsilon}{\log\log 1/\epsilon}$ for any $\delta > 0$,

$$\mathbb{P}(|X' - \mathbb{E}(X)| \geq \mathcal{O}(\epsilon\mathbb{E}(X))) \leq \delta$$

Property 27. [4]

Let M be the distribution of the maximum number of balls in any of the bins from a $\mathcal{N}_k(A, I)$ distribution under the $I \geq 20A$ regime. Then,

$$\mathbb{P}(M \geq \lambda) \leq e^{\log I - \Omega(\min(\lambda\mu,k)\log\lambda)}$$

Proof: One can union bound

$$\mathbb{P}(M \geq \lambda) \leq \sum_{i=1}^{I} \mathbb{P}(X_i \geq \lambda) \leq Ie^{-\Omega(\min(\lambda,k)\log\lambda}$$

as a result of 11

4.3. **Stable distributions.**

Definition 28. A distribution is called **stable** if X, Y are independent variables that follow this distribution, then for any constants a, b, there exist constants c, d such that $\frac{aX+bY-d}{c}$ follows the same distribution.

Example 29. Examples of stable distributions include normal distributions, as well as the Cauchy distribution.

Remark. Stability is maintained for a distribution under scaling and translation.

The most important result about stable distributions is summarized below. It concerns the existence of certain distributions.

Theorem 30. [16]

For any $p > 0$, there exists a distribution $\mathcal{D}_p$ will the following properties:

(1) $\mathcal{D}_p$ is stable, symmetric around 0, and, moreover, if $X_1, ..., X_k$ are independent variables drawn off $\mathcal{D}_p$ and $a = (a_1, ..., a_k)$, then $\sum_{i=1}^{k} a_iX_i \sim |a|_p \cdot X$, where X is drawn off $\mathcal{D}_p$.

(2) The distribution has polynomial decreasing tails. More formally,

$$\mathbb{P}(|Z| \geq \lambda) = \mathcal{O}(\frac{1}{\lambda^p})$$

Example 31. The most common examples of p-stable distributions are the following:

(1) The Cauchy distribution given by probability density function $f(x) = \frac{1}{\pi(1+x^2)}$ is 1-stable.
(2) The Gaussian distribution is 2-stable.

Remark. One can see that the bounds on the tails of this distributions are much weaker than the bounds one obtains on more known distribution such as the normal distribution or exponential distribution. Indeed, the tails of p-stable distributions are much less well-behaved. In fact, the 1-stable distribution is the Cauchy distribution, whose tails are not well-behaved. In fact, the Cauchy distribution does not even have a mean!

The relevance of the stable distribution actually lies in its ability to model heavy-tailed data, and, in streaming, Indyk's p-stable sketch [9]. We will delve into the specifics later in this presentation.

Remark that if $\mathcal{D}_p$ is p-stable, then $a\mathcal{D}_p$ is p-stable too for a constant a. To normalize, we will consider $\mathcal{D}_p$ to be the distribution Z such that $\mathbb{P}(|Z| \geq 1) = \frac{1}{2}$.

4.4. **Bounded independence.** Just like in the case of throwing balls into bins, we will again be interested in drawing from families of $\mathcal{D}_p$ distributions that are not fully independent, but k-independent for some k. One is interested in a result in the fashion of 50. Very similar to 24 and 25, the following results hold true:

Theorem 32. [13]
Let $\epsilon > 0$. Let $X_1, ..., X_n$ be k-independent random variables drawn off $\mathcal{D}_p$ and let $Y_1, ..., Y_n$ be fully independent random variables drawn off $\mathcal{D}_p$. Let $a < b \in \mathbb{R}$. Let Z be a vector of length n. Then,

$$|\mathbb{P}(\langle X, Z\rangle \in (a, b)) - \mathbb{P}(\langle Y, Z\rangle \in (a, b))| \leq \mathcal{O}(k^{-\frac{1}{p}})$$

, where the $\mathcal{O}$ term can hide dependency on p, but does not depend on k, a, b.

Remark. Unlike the throwing balls into bins distribution, the p-stable distribution is not as well behaved. Thus, one can't obtain bounds on the deviation of the k-independent model from the fully independent model through Chebyshev type inequalities, since expectancy and variance are not even guaranteed. Thus, one needs to give the direct bound above to control the loss given by k-independence.

Proposition 33. [5]
Let Z_i, $i = 1, k$ be drawn off $\mathcal{D}_p$, and assume they are k-independent. Then,

$$\mathbb{P}\left(\sum x_i Z_i^2 \geq \lambda ||x||_p^2\right) \leq \mathcal{O}(\lambda^{-\frac{p}{2}}) + \mathcal{O}(k^{-\frac{1}{p}})$$

, where the constant in the second $\mathcal{O}$ term can contain a dependency on p.

Proof:
Remark that every entry Z_i can be written as $|Z_i| \cdot \sigma_i$. Now, conditional on $|Z_i|$, one has that

$$\mathbb{E}_\sigma\left((\sum x_i|Z_i|\sigma_i)^2 \mid |Z_1|,...,|Z_n|\right) = \sum x_i Z_i^2$$

since the σ_i are k-wise independent and thus pairwise independent. Now, let E be the event that $\sum_{i=1}^n x_i Z_i^2 \geq \lambda||x||_p^2$. Then,

$$\mathbb{P}_\sigma(|(\sum x_i|Z_i|\sigma_i)| \geq \sqrt{\frac{2}{3}}\lambda||x||_p) \geq \frac{1_E}{27}$$

, where 1_E is an indicator variable for E. This follows from the Paley-Zigmund inequality [17]:

$$\mathbb{P}(Z \geq \alpha\mathbb{E}(Z)) \geq (1-\alpha)^2 \cdot \frac{\mathbb{E}(Z)^2}{\mathbb{E}(Z^2)}$$

and the fact that $\mathbb{E}_\sigma(|(\sum x_i|Z_i|\sigma_i)|^4) < 3\left(\mathbb{E}_\sigma(|(\sum x_i|Z_i|\sigma_i)|^2)\right)^2$. Thus,

$$\mathbb{P}(E) \leq 27 \cdot \mathbb{P}(|\sum x_i Z_i| \geq \sqrt{\frac{2}{3}}\lambda||x||_p) = \mathbb{P}(||x||_p \cdot Z \geq sqrt\frac{2}{3}\lambda||x||_p) + \mathcal{O}(k^{-\frac{1}{p}})$$

$$\begin{aligned}\mathbb{P}(E) &\leq 27 \cdot \mathbb{P}(|\sum x_i Z_i| \geq \sqrt{\frac{2}{3}}\lambda||x||_p) \\ &= \mathbb{P}(||x||_p \cdot Z \geq \sqrt{\frac{2}{3}}\lambda||x||_p) + \mathcal{O}(k^{-\frac{1}{p}}) \text{ from 32} \\ &= \mathcal{O}(\frac{1}{\lambda^p}) + \mathcal{O}(k^{-\frac{1}{p}}) \text{ from 30}\end{aligned} \tag{6}$$

Similar to 50, one is interested in looking at the process $\langle v_i, X\rangle$, where X v_i are increasing vectors component-wise and X is a k-wise independent vector of entries drawn off $\mathcal{D}_p$.

5. Stochastic processes under bounded independence

5.1. Martingales.

Definition 34. A **martingale** is a sequence of variables $X_1, ..., X_n$ such that for $2 \leq k \leq n$, $\mathbb{E}(X_k \mid X_1, ..., X_{k-1}) = X_{k-1}$

Definition 35. A **sub-martingale** is a sequence of variables $X_1, ..., X_n$ such that for $2 \leq k \leq n$, $\mathbb{E}(X_k \mid X_1, ..., X_{k-1}) \geq X_{k-1}$

Definition 36. [15]

A **Doob martingale** is based off a sequence of random variables $X_1, ..., X_n : \Omega \to \mathbb{R}$ and a function $f : \mathbb{R}^n \to \mathbb{R}$. Then, for a draw of $X_1, ..., X_n$ from Ω, define $A_i = \mathbb{E}_{X_{i+1},...,X_n}(f(X_1, ..., X_n) \mid X_1, ..., X_i)$

One is interested in general, for a martingale $X_1, ..., X_n$, in bounding the tails of the quantity $\sup_{t \leq T} |X_i|$. A few well-known results describe bound the tails of this distribution.

A particular class of stochastic processes that one is interested in is the case of the partial sums of a number of random variables of mean 0. When the variables are independent, the process is a martingale and methods to bound the tails of this distribution are well-known. A related type of problems refers to the case in which the variables X_i are not fully independent, but show k-wise independent for certain k.

5.1.1. *Inequalities under independence.*

Result 37. *(Doob martingale inequality) Let $(X_n)_{1 \leq n \leq T}$ be a sub-martingale, where X_i can only take non-negative values. Then,*

$$\mathbb{P}(\sup_{i \leq T} X_i \geq \lambda) \leq \frac{\mathbb{E}(X_T)}{\lambda}$$

A corollary of this result gives right tail bounds on the supremum of a martingale

Corollary 38. *(Kolmogorov's inequality)*
Let $(Y_i)_{1 \leq i \leq T}$ be a sequence of independent random variables of mean 0. Let $X_t = \sum_{i=1}^{t} Y_i$ Then,

$$\mathbb{P}(\sup_{i \leq T} |X_i| \geq \lambda) \leq \frac{\mathbb{E}(Y_T^2)}{\lambda^2}$$

Proof:
By Jensen's inequality, X_t^2 is a sub-martingale when X_i are independent, since X_i is a martingale. Consequently, the statement follows from 37 for the non-negative sub-martingale Y_i.

Another interesting result under independence is given in [2].

Theorem 39. *Azuma-Hoeffding inequality*
Let $S_1, ..., S_n$ correspond to a Doob's martingale for a function $f(X_1, ..., X_n)$, for independent $X_1, ..., X_n$. Moreover, let $s_i = \sup |X_i - X_{i-1}|$. Then,

$$\mathbb{P}(|S_n - S_0| \geq \lambda) \leq 2e^{-\frac{\lambda^2}{\sum_{i=1}^{n} c_i^2}}$$

Remark. While in general this does not say anything about the supremum of the process, this can be obtained through a union bound. As such,

$$\mathbb{P}(\sup_{t \leq n} |S_t - S_0| \geq \lambda) \leq \sum_{i=1}^{n} 2e^{-\frac{\lambda^2}{\sum_{j=1}^{i} c_j^2}}$$

5.1.2. 4-*wise independence.*

Result 40. [4]

Let $(Y_i)_{1\leq i\leq T}$ *be a sequence of* 4*-wise independent random variables such that* $\mathbb{E}(Y_i)=0$ *and* $\mathbb{E}(Y_i^2)=1$*. Let* $X_i=\sum_{j=1}^{i}Y_j$*. Then,*

$$\mathbb{P}(\sup_{i\leq T}|X_i|\geq\lambda)=\mathcal{O}(\frac{T}{\lambda^2})$$

Remark. Remark that when the variables are only 4-wise independent, the process X_i is no longer a martingale. The new method of proof gives bounds that fall short of the bound obtained under independence by just a constant factor.

Proof:

For $k\leq T$, consider the quantity $(X_i-X_{i-k})^2\cdot(X_{i+k}-X_i)^2$. Remark that this quantity is a weighted sum of degree 4 polynomials in $X_{i-k+1},...,X_{i+k}$. By 4-independence of Y_i, all of these sums will be 0 unless the term is of the form $Y_a^2Y_b^2$ or Y_a^4. There are k^2 terms of the first kind and 0 of the second and thus

$$\mathbb{E}\left((X_i-X_{i-k})^2\cdot(X_{i+k}-X_i)^2\right)=k^2$$

This implies the probability bound

$$\mathbb{P}(\min(|X_t-X_{t-k}|,|X_{t+k}-X_t|)\geq\lambda\sqrt{k})\leq\frac{\mathbb{E}\left((X_i-X_{i-k})^2\cdot(X_{i+k}-X_i)^2\right)}{(\lambda\sqrt{k})^4}=\frac{1}{\lambda^4}$$

Now, remark that one can assume WLOG that $T=2^t$ is a power of 2, since increasing the length T of the random walk can only increase the supremum. Now, let $E_{i,s}$, for $i\leq t$, $1\leq s\leq 2^{t-i}-1$ be the event that $\min(|X_t-X_{t-2^s}|,|X_{t+2^s}-X_t|)\geq\gamma\cdot 2^{\frac{s}{2}}\cdot 2^{\frac{t-s}{3}}$. From the previous remark, $\mathbb{P}(E_{i,s})\leq\frac{1}{\gamma^4\cdot 2^{\frac{4(t-s)}{3}}}$ and thus

$$\mathbb{P}(\exists i\mid E_{i,s})\leq\frac{1}{\gamma^4\cdot 2^{\frac{t-s}{3}}}$$

from union bound. Thus,

$$\mathbb{P}(\exists i,s\mid E_{i,s})\leq\sum_{i=0}^{t}\frac{1}{\gamma^4\cdot 2^{\frac{t-s}{3}}}\leq\sum_{i=0}^{\infty}\frac{1}{\gamma^4\cdot 2^{\frac{t}{3}}}=\mathcal{O}(\frac{1}{\gamma^4})$$

Now, if no $E_{i,s}$ happens, then

$$\sup_{i\leq T}|X_i|\leq|S_n|+\gamma\sum_{s=0}^{t}\gamma 2^{\frac{s}{2}}\cdot 2^{\frac{t-s}{3}}$$

To see why this is true, remark that if all $E_{i,t}$ are true, at any point i, consider its level l to be the biggest power of 2 that divides i. Remark that one can reach the next level $i+1$ by a cost of at most $\gamma 2^{\frac{s}{2}}\cdot 2^{\frac{t-s}{3}}$ to $|X_i|$, since $E_{i,l}$ holds. Now, the last

number can be either 0 or 2^t and since $X_0 = 0$, we have that $\max(|X_0|, |X_n|) = |X_n|$, hence the inequality. Now,

$$\sum_{s=0}^{T} 2^{\frac{s}{2}} \cdot 2^{\frac{t-s}{3}} = \sum_{s=0}^{T} 2^{\frac{t-s}{2}} \cdot 2^{\frac{s}{3}} = \sqrt{n} \cdot \sum_{s=0}^{T} 2^{-\frac{s}{6}} = \mathcal{O}(\sqrt{n})$$

Now, $\mathbb{P}(|X_n| \geq \gamma\sqrt{n}) \leq \frac{\mathbb{E}(S_n^2)}{n\gamma^2} = \frac{1}{\gamma^2}$

Thus,

$$\mathbb{P}\left(\sup_{i\leq n} |X_i| \geq 2\gamma\mathcal{O}(\sqrt{n})\right) \leq \mathbb{P}\left(|X_n| \geq \gamma\mathcal{O}(\sqrt{n})\right) + \mathbb{P}(\exists i, s \mid E_{i,s}) \leq \mathcal{O}(\frac{1}{\gamma^2}) + \mathcal{O}(\frac{1}{\gamma^4})$$

and the result is proven.

Remark. One can remark that the 4-wise independence condition is one of the conditions that make this proof work. Another one that could work is $\mathbb{E}((S_i - S_j))^4) \leq \mathcal{O}((i-j)^2)$. This is because the inequality

$$\mathbb{E}\left((X_i - X_{i-k})^2 \cdot (X_{i+k} - X_i)^2\right) = k^2$$

can be obtained up to a constant factor from the Cauchy-Schwarz inequality.

5.1.3. *Doob's martingale on thrown balls.* We are coming back to the process of throwing balls into bins under $I \geq 20A$ regime. While we were able to provide tight bounds on the tails of the distribution, one can see the process as dynamic, with the throws $X_1, ..., X_A$ being observed one after the other. In such a case, one is interested in the Doob's martingale associated with the number of non-empty bins. As such, if $D_1, ..., D_A$ is the Doob's martingale, then one is interested in bounding $\sup_{i\leq A} |D_i - \Phi_I(i)|$. In the case of completely independent variables, Doob's martingale inequality for the martingale $D_i - \Phi_I(i)$ would give:

$$\mathbb{P}\left(\sup_{i\leq A} |D_i - \Phi_I(i)| \geq \lambda\right) \leq \frac{Var(N)}{\lambda^2}$$

One is interested in a similar result in the case of k-independent throws. The rest of the section will be dedicated to proving that one can achieve the bound $\sup_{i\leq A} |D_i - \Phi_I(i)| = \mathcal{O}(\sqrt{A})$ with probability $> \frac{1}{2}$ for some $k = poly((\log A)^2)$. The general strategy is to express $X_1, ..., X_A$ using $\log I$ bits, and create a polynomial on these bits of degree at most k that closely approximates the number of non-empty bins. The motivation behind this is that working with a polynomial of degree at most k over $X_1, ..., X_A$ is the same as working with independent $X_1, ..., X_A$.

Proposition 41. [4]

For any k, and any $p \in \{0,1\}^k$ there exists a degree k polynomial in variables $X_1, ..., X_k$ such that $f(q) = 0$ for $q \in \{0,1\}^k \setminus \{p\}$ in $\{0,1\}$ not all equal to 0 and $f(p) = 1$.

Proof:
Consider the polynomial $f(X_1, ..., X_k) = \prod_{i=1}^{k} (-1)^{p_i} \cdot (X_i - p_i)$. We will let this polynomial be EQ_p.

Proposition 42. [4]
There exists a polynomial P_r on $\log(I) \cdot A$ bits of degree $\mathcal{O}((\log P)^2)$ such that there exists $k = poly(\log I)$, such that

$$||P_r(X_1, ..., X_I) - N||_q = o(1)$$

for all $q \leq r$, where $||x||_p$ stands for $\mathbb{E}(|X|^p)^{\frac{1}{p}}$.

Proof:
One looks at the polynomial $P = \sum_{i=1}^{I} I_d \left(\sum_{j=1}^{A} EQ_j \right)$. Then, remark that $P = N$ as long as $M(v) \leq d$. Now, if $M \geq d$, then

$$|P - N| \leq A \binom{M}{d} \leq \mathcal{O}\left(A \left(\frac{eM}{d} \right)^d \right)$$

Thus,

$$||P-N||_q \leq \sum_{r \geq \log d} \mathbb{P}(X \geq 2^r) \cdot \mathcal{O}\left(I \cdot \left(\frac{e2^r}{d} \right)^d \right)^q \leq \sum_{r \geq \log d} e^{\Omega(q \log I + qd \log \frac{2^r}{d}) - \Omega(\min(2^r, k) 2^r)}$$

Thus, one can pick $r = poly(d)$ so that the exponents are exponentially decreasing in r. Thus, the sum would be $e^{\Omega(q \log I + qd) - \Omega(d \log d)}$
And since q is fixed, one can choose d such that the above is $e^{-\Omega(d \log d)}$, which finishes the proof since one can choose $d = \mathcal{O}(\log P)$.
Now, the degree of the polynomial is $\log P \cdot d = \mathcal{O}((\log P)^2)$ indeed.

Notation 43. *We will let S_i to be the stochastic process given by $S_i = \mathbb{E}(N(X_1, ..., X_A) \mid X_1, ..., X_{i-1})$. Similarly, let $B_i = \mathbb{E}(P_4(X_1, ..., X_A) \mid X_1, ..., X_{i-1})$*

Proposition 44. [4]

$$\mathbb{P}_{X_1, ..., X_t} (S_i - B_i \geq \lambda) \leq \mathcal{O}\left(\frac{1}{\lambda^2} \right)$$

, where $\mathcal{O}(1)$ does not depend on t.

Proof: Remark that

(7)

$$\begin{aligned}\mathbb{P}_{X_1,\ldots,X_t}\left(S_i - B_i \geq \lambda\right) &\leq \frac{\mathbb{E}_{X_1,\ldots,X_t}\left(|S_i - B_i|^2\right)}{\lambda^2} \text{ by Chebyshev} \\ &\leq \frac{\mathbb{E}_{X_1,\ldots,X_t}\left(\mathbb{E}_{X_{t+1},\ldots,X_A}|(N - P_2)|^2 \mid X_1, \ldots, X_t\right)}{\lambda^2} \text{ by Cauchy-Schwarz} \\ &= ||N - P_2|\,|_2 \\ &\leq \mathcal{O}\left(\frac{1}{\lambda^2}\right) \text{ by 42}\end{aligned}$$

In turn, union bound over the last inequality can use non-dependence on t to give a bound on the probability that at any point in time the expected difference between the polynomial P_2 and the number of non-empty bins is ever greater than some λ

Corollary 45.

$$\mathbb{P}\left(\sup_t |S_t - B_t| \geq \lambda\right) \leq \mathcal{O}(\frac{A}{\lambda^2})$$

Proposition 46. [4]
If $X_1, \ldots, X_A$ are at least $4deg(P_4)$ independent, then

$$\mathbb{P}\left(\sup_t |B_t - B_0|\right) \leq \mathcal{O}\left(\frac{A}{\lambda^2}\right)$$

Proof
One considers $\Delta_t = B_{t+1} - B_t$. We are planning to use the Azuma-Hoeffding inequality. As such, remark that changing the result of one throw can change the result by at most 1. Further remark that the direct use does not lead to a good bound, since we are interested in the regime $\lambda = \Theta(\sqrt{A})$, the union bound is rendered useless. Remark that using the Azuma-Hoeffding inequality would be in order, as one can assume that the throws are independent, as one only deals with polynomials of degree at most the degree of independence of the throws.

However, Azuma-Hoeffding gives that $|B_i - B_j|$ has sub-Gaussian tails with variance at most $\mathcal{O}(\sqrt{i-j})$). Consequently, $\mathbb{E}(|S_i - S_j|^4) \leq (i-j)^2$ and one can apply the results from 40 to achieve the bound.

Proposition 47. [4]
One has that

$$\mathbb{P}(\sup_{t \leq T} |S_t - S_0| \geq \lambda) \leq \mathcal{O}\left(\frac{1}{\lambda^2}\right)$$

Proof: By the triangle inequality,

$$\sup|S_t - S_0| \leq |B_0 - S_0| \sup|B_t - B_0| + \sup|S_t - B_t| \leq \sup|B_t - B_0| + 2\sup|S_t - B_t|$$

For the event to hold, one needs either $\sup |S_t - B_t| \geq \frac{\lambda}{3}$ or $|\sup_t B_t - B_0| \geq \frac{\lambda}{3}$, both of which are $\mathcal{O}(\frac{1}{\lambda^2})$ events.

We will end this section with the following theorem:

Theorem 48. [4]

For any $\delta > 0$, one has that with probability at least $1 - \delta$,

$$\sup_{t \leq T} ||X_{1,t}| - \Phi(t)| \leq \mathcal{O}(\sqrt{A})$$

, where the $\mathcal{O}(\sqrt{R})$ is allowed to have dependencies on δ.

Proof:

Based on 47, one can say that with probability $1 - \delta$,

$$\sup |\mathbb{E}_{X_{t+1},...,X_A}|X_{1,A}| - \Phi(A)| \leq \mathcal{O}(\sqrt{A})$$

Conditional on this for any $t \leq T$

$$\Phi^{-1}(|\mathbb{E}_{X_{t+1},...,X_A}|X_{1,A}||) - A = \mathcal{O}(\sqrt{A})$$

$$\Phi^{-1}(\Phi(\Phi^{-1}(|X_{1,t})| + A - t)) = \mathcal{O}(\sqrt{A})$$

$$|X_{1,t}| = \Phi(t) + \mathcal{O}(\sqrt{A})$$

which is what we wanted to prove.

5.2. **Maximum inner product.** The general setup we will have in this section refers to a set of component-wise increasing vectors in $\mathbb{R}^n$, with $0 = v_0 \leq v_1 \leq ... \leq v_n$, where in this context $\leq$ means component-wise increasing. If σ is a vector taken from a certain distribution, one is interested in bounding the tails of the distribution $\sup_{i \leq n} \langle v_i, \sigma \rangle$. The motivation for such processes is given by insertion only streams. In such streams, the frequency vector is always component-wise increasing.

5.2.1. *ε-net.* The main ideas of proof are ε-nets. Here is a definition and the main result.

Definition 49. For a set S endowed with a metric d, S_ε is called an ε-net of the set S if for any $s \in S$, there exists $t \in S_\varepsilon$ such that $d(s,t) < \varepsilon$.

We will be interested in finding small ϵ-nets of finite sets

Proposition 50. Particular case of $p = 2$ proved in [5]

If $v_0, v_1, ..., v_n$ are components increasing vectors in $\mathbb{R}^n$, then there exists an ϵ-net of this set of size at most $\frac{|v_n - v_0|_p^p}{\epsilon^p} + 1$, when $\mathbb{R}^n$ is endowed with the $|| \cdot ||_p$ distance metric for $p \geq 1$.

Proof:

Consider the sequence t_i given by $t_0 = 0$, and $t_i = \min\{j \geq t_i \mid |v_j - v_{t_i}| \geq \epsilon\}$. Stop whenever t_i can't be defined, i.e. the last number defined is k. Now, $T = \{t_i\}$ is an ε-net. Moreover, because v_i are component-wise increasing,

$$\sum_{i=0}^{k-1} |v_{t_{i+1}} - v_{t_i}|_p^p \leq |v_n - v_0|_p^p$$

and thus given that $|v_{t_{i+1}} - v_{t_i}|_2 \geq \epsilon$, $k \leq \frac{|v_n - v_0|_2^2}{\epsilon^p}$. Since $|T| = k + 1$, the result is proven.

5.2.2. *Bounded independent Rademachers.* The first case we will treat will be the case of σ being a set of Rademachers (random sign variable that is 1 with probability $\frac{1}{2}$ and -1 with probability $\frac{1}{2}$) that have bounded independence.

Proposition 51. (Weaker bound given in [5])

Let $0 = v_0 \leq v_1 \leq ... \leq v_n$. Let $Z \in \{-1, 1\}^n$ be a 4-wise independent vector of signs. Then,

$$\mathbb{P}(\sup_{i=1,n} \langle v_i, \sigma \rangle \geq \lambda |v_n|_2) \leq \mathcal{O}(\frac{1}{\lambda^4})$$

To prove this, first remark that $\mathbb{E}(\langle Z, x \rangle) \leq C|x|_2^4$. Moreover, consider S_k to be $\frac{|v_n|_2}{2^k}$ nets of $V = \{v_1, ..., v_n\}$ under the Euclidian norm. For a point x, let $x^{(k)}$ be a point of distance at most $\frac{|v_n|_2}{2^k}$ from x in $S_{\frac{1}{2^k}}$. Remark that the first net S_0 just contains 0. Now, if one looks at the possible values of $x^{(k)} - x^{(k-1)}$, remark that for any $|x^{(k)} - x^{(k-1)}| \leq |x^{(k)} - x| + |x^{(k-1)} - x| \leq \frac{3|v_n|_2}{2^k}$. Now, once $x^{(k)}$ is fixed, $x^{(k-1)}$ is of distance at most $\frac{3}{2^k}$ from this. Now, take s and t to be the smallest number in V smaller and bigger than $x^{(k)}$ of distance at most $\frac{3|v_n|_2}{2^k}$. Now, since v_i are increasing, there are at most 3 choices for $x^{(k-1)}$ once $x^{(k)}$ is fixed. Thus, $D_k = |\{x^{(k)} - x^{(k-1)} \mid x \in S\}| \leq 3 \cdot |S_k| = \mathcal{O}(2^{2k})$ from 50.

Now,

$$\langle Z, x \rangle = \lim_{k \to \infty} \sum_{k=1}^{M} \langle Z, x^{(k-1)} - x^{(k)} \rangle$$

Now, for a set S, let $X = \sup_{x \in S} |\langle Z, x \rangle|$. Then,

$$\begin{aligned} \mathbb{P}(|X| \geq \lambda) &\leq \sum_{x \in S} \mathbb{P}(|\langle Z, x \rangle| \geq \lambda) \text{ by union bound} \\ &\leq \sum_{x \in S} \frac{\mathbb{E}((\langle Z, x \rangle)^4)}{\lambda^4} \qquad (8) \\ &= \mathcal{O}(|S| \cdot \frac{\sup_{x \in S} |x|_2^4}{\lambda^4}) \text{ by Khintchine's inequality [14]} \end{aligned}$$

Let X_k be the variable X corresponding to D_k. Then, $\mathbb{P}(X_k \geq \frac{\lambda}{2^{\frac{k}{3}}}) = \mathcal{O}(\frac{2^{\frac{-2k}{3}}}{\lambda^4})$, given that $|D_k| \leq \mathcal{O}(2^{2k})$, $\sup_{x \in D_k} |x|_2^4 = \mathcal{O}(2^{-4k})$.

Now,

$$\mathbb{P}(\sup_{v \in V} |\langle v_i, Z \rangle| \geq \lambda) \leq \sum_{k=0}^{\infty} \mathbb{P}(\sup_{v \in D_k} |\langle v_i, Z \rangle| \geq C\lambda 2^{-\frac{k}{3}}) = \sum_{k=0}^{\infty} \mathcal{O}(\frac{2^{-\frac{2k}{3}}}{\lambda^4}) = \mathcal{O}(\frac{1}{\lambda^4})$$

for the constant $C = \frac{1}{\sum_{k=0}^{\infty} 2^{-k/3}}$ and the result is proven.

5.2.3. *Matrix norms.* One other context in which one obtains bounds is using matrix norms. There are 2 very common norms that one uses for the metrics on $n \times n$ matrices:

(a) The first one is the Euclidian metric, treating the vector as an $n \times n$ matrix, also called the Frobenius norm. We will let this be $||\cdot||_F$. The analysis for the Frobenius norm is identical to the one done for the vector case in Euclidian space, since there is a vector space isomoprhism between $\mathbb{R}^{n \times n}$ and $\mathbb{R}^{n^2}$.

(b) The second norm will be the spectral norm, i.e. the value of the largest eigenvalue of the matrix. Remark that the spectral norm is always smaller than the Frobenius norm - this is because if λ_i are the eigenvalues of a matrix A, then

$$||A|| = \max |\lambda_i| \leq \sqrt{\sum \lambda_i^2} = \sqrt{Tr(A^T A)} = ||A||_F$$

This gives the following corollary:

Corollary 52. *If S is an ϵ-net for a set V under $||\cdot||_F$, then it is an ϵ-net under $||\cdot||$.*

5.2.4. *Stable distribution.* The next similar inequality we will be interested in will be the case of a stable distribution. Here a rigurous statement of our claim:

Proposition 53. [5]

Let $v_0, v_1, ..., v_n$ be vectors with the property that $0 = v_0 \leq v_1 \leq ... \leq v_n$. Then, if Z is an n-vector with k-independent entries and $\mathcal{D}_p$ marginal distribution, then

$$\mathbb{P}(\sup_{i=1,n} |\langle Z, v_i \rangle| \geq \lambda ||v_n||_p) \leq \mathcal{O}\left(\lambda^{-\frac{2p}{2+p}}\right) + \mathcal{O}\left(k^{-\frac{1}{p}}\right)$$

Proof: The proof combines 50 and 33. One can again use the trick of seeing $Z_i = |Z_i| \cdot \sigma_i$. Now, if one defines v_i' to be a vector whose j-th component is the jth component of v_i times Z_i, then $\langle Z, v_i \rangle = \langle \sigma, v_i' \rangle$. Now, for any β,

(9)

$$\mathbb{P}\left(\sup_{i=1,n}|\langle Z,v_i\rangle|\geq\lambda||v_n||_p\right)\leq\mathbb{P}\left(\sum_{i=1}^{n}v_{i,n}^2Z_i^2\geq\beta||v_n||_p\right)+\mathbb{P}\left(\sup_{i=1,n}|\langle Z,v_i\rangle|\geq\frac{\lambda}{\beta}||v_n'||_2\right)$$
$$=\mathcal{O}\left(\frac{1}{\beta^p}\right)+\mathcal{O}\left(k^{-\frac{1}{p}}\right)+\mathcal{O}\left(\frac{\beta^2}{\lambda^2}\right)\text{ from 50 and 33}$$
$$=\mathcal{O}\left(\frac{1}{\lambda^{\frac{2p}{p+2}}}\right)+\mathcal{O}\left(k^{-\frac{1}{p}}\right)\text{ by chooosing }\beta=\Theta(\lambda^{\frac{2}{p+2}})$$

6. Number of distinct elements

6.1. **Constant-factor approximation with constant failure probability.** The first step towards a solution is finding a constant-factor approximation. The following subroutine and analysis is due to [12], which they call *RoughEstimator*. The most important subroutine is presented below. It depends on the values of φ, m.

We will let $\mathcal{H}_k(A,B)$ denote a family of k-independent hash functions from A to B.

Algorithm 1 Rough-estimator(ϕ, m)

Initializing:
Take constant $K=\max(m,\frac{\log n}{\log\log n})$
Take constants $C_1,C_2,...,C_K$ initialized to -1.
Pick hash functions $h_1\in\mathcal{H}_2([n],[0,n-1]),h_2\in\mathcal{H}_2([n],[0,K^3]),h_3\in\mathcal{H}_{2K}(K^3,K)$
Update step:
for $i=1,K$ **do**
  $C_{h_3(h_2(i)}\leftarrow\max(C_{h_3(h_2(j))},lsb(h_1(i))$
Estimator:
Define $T_r=\{i\mid C_i\geq r\}$. Consider r^* the largest r for which $T_r\geq\varphi K$. Return -1 if no r is found and $2^{r^*}\cdot K$.

Notation 54. *Let $F_0(t)$ be the value of F_0 at point t in the stream. Let $I_r(t)$ be the elements i in the value space $[1,n]$ that verify $lsb(h_1(i))\geq r$.*

The proof will use a few observations.

Proposition 55. For $r<s$, $I_r(t)\subseteq I_s(t)$.

Proof: Follows immediately from the definition of the sets $I_r(t)$.

Proposition 56. Informally, $I_r(t)$ sub-samples $F_0(t)$ by a factor of 2^r. More formally, $\mathbb{E}(|I_r|)=\frac{F_0(t)}{2^r}$, $Var(I_r(t))=\frac{F_0(t)}{2^r}-\frac{F_0(t)}{2^{2r}}$. In particular $Var(I_r(t))\leq\mathbb{E}(I_r(t))$.

Proof: Every element that is seen in the stream is hashed to a value with $lsb\geq r$ with probability $\frac{1}{2^r}$. The result follows from linearity of expectations. The variance is computed in a similar way, using the 2-independence of the hashing map.

Corollary 57. *From Chebyshev's inequality, and given that* $Var(I_r(t)) \leq \mathbb{E}(I_r(t))$ *one can deduce that*

$$\mathbb{P}(|I_r(t) - \frac{F_0(t)}{2^r}| \geq c \cdot \frac{F_0(t)}{2^r}) \leq \frac{1}{c^2} \cdot \frac{1}{\frac{F_0(t)}{2^r}}$$

Proposition 58. Now, from a discrete continuity argument, one can find r' such that $\mathbb{E}(I_{r'}(t))$ is between $\frac{K}{2}$ and K. (as long as $F_0(t) \geq K$ The next step is to provide a high probability inequality for $I_{r'}(t)$ being around K and not only in expectation. Consider the event $\mathcal{E}$ that $I_{r'}(t)$ is between $\frac{K}{3}$ and $\frac{4K}{3}$. $\mathbb{P}(\mathcal{E}) = 1 - O(\frac{1}{K})$.

Analogously, one has upper bounds on $|I_{r'+2}(t)|$. In particular, just as in [12], let $\mathcal{E}'$ be the event that $|I_{r'+2}(t)| \geq \frac{7}{24}K$. $\mathbb{P}(\mathcal{E}') = 1 - O(\frac{1}{K})$

Proof: Follows from 57

Next step in the solution requires looking at the hashing map h_2.

Proposition 59. Let $\mathcal{A}$ be the event the hashing map h_2 will create no collisions on $I_{r'}(t))$. Then, $\mathbb{P}(\mathcal{A}) = 1 - O(\frac{1}{K})$

Proof: With probability $1 - O(\frac{1}{K})$, there are at most $O(K)$ elements in $I_{r'}(t)$ and thus there are $O(K^2)$ pairs and by union bound there is at most an $1 - O(\frac{1}{K})$ probability there exists no collision. Also, one can remark that because of 55, as long as $\mathcal{A}$ holds, there are no collisions of elements in $I_r(t)$ for $r \geq r'$.

Proposition 60. In the event that both $\mathcal{E}$ and $\mathcal{A}$ hold, $T_r(t)$ is distributed as the distribution 13, when throwing $|I_{r'}(t)|$ balls into K bins.

Explanation: This is because when $\mathcal{E}$ holds, $I_{r'}(t)$ has size less than $2K$, which is the degree of independence of h_3. Now, since h_2 has no collisions, T_r is indeed distributed as said. This also means that

$$\mathbb{E}(T_{r'}(t) \mid \mathcal{E} \wedge \mathcal{A}) = K(1 - (1 - \frac{1}{K})^{|I_{r'}(t)|})$$

Similarly,

$$\mathbb{E}(T_{r''}(t) \mid \mathcal{E}' \wedge \mathcal{A}) = K(1 - (1 - \frac{1}{K})^{|I_{r''}(t)|})$$

Just as before, one wants to obtain tail bounds of the $T_{r'}$ and $T_{r''}$ distributions.

Proposition 61.

$$\mathbb{P}(|T_{r'} - \mathbb{E}(T_{r'} \mid \mathcal{E} \wedge A)| \geq \epsilon \mathbb{E}(T_{r'} \mid \mathcal{E} \wedge \mathcal{A})) \mid \mathcal{E} \wedge \mathcal{A}) \leq 2e^{-\epsilon^2 \frac{\mathbb{E}(T_{r'} \mid \mathcal{E} \wedge \mathcal{A})}{3}}$$

Proof:
This follows from 19

Notation 62. *Let* $\mathcal{F}$ *be the event that* $T_{r'}(t) \geq \varphi K$. *Let* $\mathcal{F}'$ *be the event that* $T_r(t) \leq \varphi K$ *for* $r \geq r' + 2$.

Now, remark that conditional on $\mathcal{E}$, $|I_{r'}(t)| \leq \frac{K}{3}$ and thus

$$\mathbb{E}(|T_{r'}(t)| \mid \mathcal{E} \wedge \mathcal{A}) \geq (1 - e^{-\frac{1}{3}}) \cdot K$$

and thus a natural choice for φ becomes $\varphi = (1 - \varepsilon) \cdot (1 - e^{-\frac{1}{3}})$ for some small ε. Now, for ϵ small enough, 61 will give

$$\mathbb{P}(\mathcal{F} \mid \mathcal{E} \wedge \mathcal{A}) \geq 1 - e^{\Omega(\mathbb{E}(T_{r'} | \mathcal{E} \wedge \mathcal{A}))}$$

Now, $\mathbb{E}(|T_{r'}(t)| \mid \mathcal{E} \wedge \mathcal{A}) = \Omega(K)$ and thus

$$\mathbb{P}(\mathcal{F} \mid \mathcal{E} \wedge \mathcal{A}) \geq 1 - e^{-\Omega(K)}$$

Proposition 63.

$$\mathbb{P}(\mathcal{F}' \mid \mathcal{E}' \wedge \mathcal{A}) \geq 1 - e^{-\Omega(K)}$$

Proof:

For $n \geq 8$, $\mathbb{E}(I_{r'+2}(t)) \leq .99\left(1 - e^{-\frac{1}{3}}\right)$ as long as $m \geq 8$, and thus by the concentration bound developed earlier, the result follows.

Proposition 64.

$$\mathbb{P}(\mathcal{F} \wedge \mathcal{F}') = 1 - O(\frac{1}{K})$$

Proof

$$\mathbb{P}(\mathcal{F}) = \mathbb{P}(\mathcal{F} \mid \mathcal{E}, \mathcal{A}) \cdot \mathbb{P}(\mathcal{A} \mid \mathcal{E}) \cdot \mathbb{P}(\mathcal{E}) \geq 1 - O(\frac{1}{K})$$

Absolutely similarly, $\mathbb{P}(\mathcal{F}') \geq 1 - O(\frac{1}{K})$ and thus from $\mathbb{P}(\mathcal{F} \wedge \mathcal{F}') \geq \mathbb{P}(\mathcal{F}) + \mathbb{P}(\mathcal{F}') - 1$, one gets that $\mathbb{P}(\mathcal{F} \wedge \mathcal{F}') \geq 1 - O(\frac{1}{K})$.

Corollary 65. *An implication of this result is that Rough-estimator outputs the correct result with probability* $1 - O(\frac{1}{K})$*, given that when* $\mathcal{F} \wedge \mathcal{F}'$ *holds,* r^* *will be on of* r' *and* $r' + 1$*, both of which will make the estimator* $2^r \cdot |T_r|$ *between* F_0 *and* $4F_0$*.*

Notation 66. *Consider the algorithm 3-Rough-estimator, which runs Rough-estimator 3 times where the hashing maps are created independently and outputs the median result of the* 3 *runs.*

Proposition 67. *3-Rough-estimator* outputs a value V between $F_0(t)$ and $4F_0(t)$ with probability $1 - \frac{1}{K}$.

Proof: 3-Rough-estimator outputs a number between $F_0(t)$ and $4F_0(t)$ if and only if 2 of the *Rough-estimator* runs are. Thus, probability of failure is $O(\frac{1}{K^2})$.

Proposition 68. *3-Rough-Estimator* can be made to estimate $F_0(t)$ to a factor of 8, an any point $F_0(t) \geq K$ with probability $1 - O(\frac{(\log \log n)^2}{\log n})$. From 67, one gets the result.

Proof:
One can union bound over all the positions t_i at which F_0 grows by a factor 2. In between the intervals, one can keep the estimator. Thus, one has to union bound over $\log n$ positions for this to be true

Proposition 69. The estimate given by *3-Rough-estimator* is non-decreasing.

Proof: Because K in the *3-Rough-estimator* is not changed by updates, remark that C_i are non-decreasing functions on the number of updates, and thus $|T_i|$ are non-decreasing in the number of items seen. Consequently, the r^* is never decreasing, and thus the estimate is non-decreasing.

Proof: With probability $1 - O(\frac{1}{K^2})$, *3-Rough-Estimator* outputs $F_0(t_r)$ correctly, where t_r is the first point F_0 is at least 2^r. Then, between t_r and t_{r+1}, F_0 can increase by at most a factor of 2, while the estimate from the algorithm only increases. Thus, doubling the estimate by 2 would give one a number that is between $F_0(t)$ and $8F_0(t)$ for all times t as long as it is so for powers of 2. But the probability it holds for powers of 2, by a union bound, is at least $1 - O(\frac{\log n}{K^2}) = 1 - O(\frac{(\log\log n)^2}{\log n})$

Proposition 70. The estimate from an instance of *3-Rough-estimator* can only increase by a multiplication of a power of 2.

Proof: The estimate is give by $2^{r^*} \cdot K$, which always has the same prime factors other that 2. Thus, the estimator can only increase by a number of factors of 2.

Proposition 71. *3-Rough-estimator* uses $\log(n)/\log\log n$ bits of space, in addition to $\mathcal{O}(\log n)$ random bits for seeds to the hash functions.

Proof: The hash functions require $\mathcal{O}(\log n)$ random bits from [7]. Maintaining the counters takes $\mathcal{O}(\frac{\log n}{\log\log n})$ bits of space.

6.2. **Arbitrary accuracy-small F_0.** This algorithms are attributed to [12]. Unless otherwise noted, the analysis is attributed to [12] too. We will first deal with the case of small F_0. We will first analyze the following algorithm.

Algorithm 2 Accurate-estimator under small F_0

Initializing:
Take constant $K = 1/\epsilon^2$
Take $k = \Omega(\frac{\log\frac{1}{\epsilon}}{\log\log\frac{1}{\epsilon}})$
Take constants $B_1, B_2, ..., B_{2K}$ initialized to 0.
Pick hash functions $h_1 \in \mathcal{H}_2([n], [0, K^3]), h_2 \in \mathcal{H}_k(K^3, 2K)$
Update step(i):
$B_{h_2(h_1(i))} \leftarrow 1$
Estimator:
Define $T = \{j \mid B_j = 1\}$. Estimate F_0 by $\frac{\ln(1-\frac{T}{2K})}{\ln(1-\frac{1}{2K})} = \Phi_{2K}^{-1}(T)$.

Proposition 72. For any $\delta > 0$, there exists a threshold t_0 with $F_0(t_0) \geq \frac{1}{16\epsilon^2}$ such that for large enough ϵ, with failure probability at most δ, for any fixed $t \leq t_0$, 6.2 outputs a value that is $F_0(t)(1 + \mathcal{O}(\epsilon))$ with probability at least $1 - \delta$.

Proof:

Let $\tilde{t}$ be the first time F_0 breaks $\frac{1}{16\epsilon^2}$. If this doesn't happen, let $\tilde{t}$ be the end of the stream. Remark that the algorithm places at time t, $F_0(t)$ elements into one of the K^3 buckets. Thus, as long as $F_0(t) \leq \frac{K}{16}$, probability of collision between $\frac{K}{16}$ elements while put in K^3 buckets is at most $\binom{\frac{K}{16}}{2} \cdot \left(\frac{1}{K^3}\right) = \mathcal{O}(\frac{1}{K})$. Denote this event by $\mathcal{E}$.

Now, for a fixed $t \leq \tilde{t}$, T is the number of bins hit when throwing $F_0(t)$ into $2K$ bins. Now, remark that every distinct element represent a ball. Thus, the estimate only changes when new elements are seen, and, moreover, they represent additional balls thrown into bins. Thus, it would be enough to prove that the number of balls thrown into bins stays close to $F_0(t)$ and the difference never gets bigger than $\epsilon F_0(t)$. Thus, for sufficiently small ϵ, this will be true from 26.

Proposition 73. For any $\delta > 0$, there exists a threshold t_0 with $F_0(t_0) \geq \frac{1}{16\epsilon^2}$ such that for large enough ϵ, with failure probability at most δ, 6.2 output a value that is $F_0(t) \pm \mathcal{O}(\frac{1}{\epsilon})$ for all t with probability at least $1 - \delta$.

Proof:

The proof is analogous to the one for the proposition above. Now, one can use 48 to obtain that for any fixed $\delta > 0$, one can have that the algorithm output $F_0(t)$ with an error of at most $\mathcal{O}(\frac{1}{\epsilon})$ at all times, given that with probability $1 - \delta$ the number of non-empty bins differs from its expectation by at most $\mathcal{O}(\sqrt{K}) = \mathcal{O}(\frac{1}{\epsilon})$.

Corollary 74. *There exists a subroutine CheckBig that with probability $1 - \delta$ correctly determines whether F_0 at some point t is at least $\frac{1}{16\epsilon^2}$ and that used $\mathcal{O}(\frac{1}{\epsilon^2})$ bits of space.*

6.2.1. *Large F_0.* In the case of arbitrary accuracy, we will use a slightly modified subroutine. Assume that one wants an algorithm that produces a number between $(1 - \epsilon)F_0$ and $(1 + \epsilon)F_0$. For this,

Intuition: The algorithm is in essence very similar to *RoughEstimator*. A few changes are the appearance of the variables $b, est, \mathcal{A}$.

(1) est is used such that 2^{est} is an approximation of F_0 at all times when F_0 is high enough.
(2) At point t, b holds a value such that $F_0(t)/2^b$ is $\Omega(K)$.
(3) $\mathcal{A}$ is an algorithm that gives strong tracking constant factor approximation of F_0. More formally, we will assume that:
   - $\mathcal{A}$ outputs a number e at time step t that has the properties that $F_0^{(t)} \leq e \leq 2^l \cdot F_0^{(t)}$ with probability at least p, as long as $F_0^{(t)} \geq \frac{K}{32}$. This event will be denoted by $\mathcal{E}$.
   - We will also assume that the output $\mathcal{A}$ is increasing in the timestep t.

Algorithm 3 Accurate-estimator($\mathcal{A}$)

Initializing:
Take constant $K = 1/\epsilon^2$
Take $k = \Omega(\frac{\log \frac{1}{\epsilon}}{\log\log \frac{1}{\epsilon}})$
Take constants $C_1, C_2, ..., C_K$ initialized to -1.
Pick hash functions $h_1 \in \mathcal{H}_2([n], [0, n-1]), h_2 \in \mathcal{H}_2([n], [0, K^3]), h_3 \in \mathcal{H}_k(K^3, K)$
Initialize $b, est = 0$.
Initialize a $\mathcal{A}$ P.
Update step:
$C_{h_3(h_2(i)} \leftarrow \max(C_{h_3(h_2(i))}, h_1(i) - b)$
Update P with i.
Estimate F_0 from P. Let the estimate be R.
if $R > 2^{est}$ **then**
  $est \leftarrow \log(R)$
  $b_{temp} \leftarrow \max(0, est - \log(K/32))$
  for $j = 1, K$ **do**
    $C_j \leftarrow \max(-1, C_j + b - b_{temp})$
  $b \leftarrow b_{temp}$
Estimator:
Define $T = \{j \mid C_j \geq 0\}$. Estimate F_0 by $2^b \cdot \frac{\ln(1-\frac{T}{K})}{\ln(1-\frac{1}{K}))} = 2^b \cdot \Phi_K^{-1}(T)$.

- We will also assume that the estimate can only increase in power of 2 increments.

The final goal of this algorithm would be a proof that there exists an ϵ-approximation algorithm for F_0 with $< \frac{1}{2}$ failure probability that runs in space $\mathcal{O}(\frac{1}{\epsilon^2} + \log n)$. Under such conditions, remark that is enough to consider the "small" ϵ-case, i.e. one can choose a constant C such that $\epsilon \leq \sqrt{C \log n}$ and only consider this case. This is because for large ϵ, the $\frac{1}{\epsilon^2}$ terms is dominated by the $\log n$ factor. Thus, $\frac{\log n + l}{K}$ can be made arbitrarily small assymptotically.

Proposition 75. Conditional on $\mathcal{E}$, the estimate $2^b \cdot \frac{\ln(1-\frac{T}{K})}{\ln(1-\frac{1}{K})}$ is $(1 \pm O(\epsilon))F_0$ with probability more than $\frac{7}{9}$.

Proof:
The most important part is looking at $I_b(T)$, denoted for convenience by I_b. $\mathbb{E}(I_b) = \frac{F_0}{2^b}$ and $Var(I_b) \leq \mathbb{E}(I_b)$. Conditional on $\mathcal{E}$, $\mathbb{E}(I_b)$ is between $\frac{K}{256}$ and $\frac{K}{32}$. Then, Chebyshev inequality implies, similarly to the proof of 58 that:

$$\mathbb{P}(\frac{K}{300} < |I_b| < \frac{K}{20}) = 1 - \mathcal{O}(\frac{1}{K})$$

Similar to the proof of 59, conditional on the previous event, h_2 will have no collisions on I_b with probability $1 - \mathcal{O}(\frac{1}{K})$.

Just as in the rough estimator case, T is the number of non-empty bins when throwing $|I_b|$ balls into K bins with limited independence of $k = \frac{\log \frac{K}{\varepsilon}}{\log \log \frac{K}{\varepsilon}}$.

Now, let T' be the same number when the $|I_b|$ balls thrown into K bins do not have limited independence but are in fact fully independent. Now, from 24 and 25, one can bound the tails of the distribution. In particular, with $\frac{4}{5}$ probability,

$$1 - T/K = (1 - \frac{1}{K})^{|I_b|} \pm \mathcal{O}\left(\epsilon(1 - (1 - \frac{1}{K})^{|I_b|})\right)$$

Thus, for small enough ϵ, the above gives $\ln(1 - \frac{T}{K}) = |I_b| \ln(1 - \frac{1}{K}) + O(\epsilon)$ and thus the estimate P of this algorithm verifies the bound $P = |I_b| \cdot 2^b + O(\epsilon \cdot 2^b \cdot K)$. Now, since $2^b \cdot K \geq 300 F_0$, it means that $P = |I_b| \cdot 2^b + O(\epsilon F_0)$. Since $|I_b| \cdot 2^b$ is F_0 in expectancy, and since Chebyshev inequality implies bounds under $\mathcal{E}$ of

$$\mathbb{P}(||I_b| - \mathbb{E}(|I_b|) \geq \frac{c}{\sqrt{K}}) \leq \mathcal{O}(\frac{1}{c^2})$$

Now, if $||I_b| - \mathbb{E}(|I_b|) \geq \frac{c}{\sqrt{K}}$ holds, together with $\mathcal{E}$ and the 2 previous events, then the output of the algorithm is an ϵ-approximation. Thus, remark that the probability of the answer being correct is $\frac{4}{5} - \mathcal{O}(\frac{1}{K}) - \frac{16}{c^2}$. This can be made arbitrarily close to $\frac{4}{5}$ for sufficiently large K, since one has choice over c. For example, this can be made at least $\frac{7}{9}$ as desired.

Corollary 76. *The algorithm succeeds with probability $\frac{7}{9} - \kappa$ for t with the property that $F_0(t) \geq \frac{1}{32\epsilon^2}$, where κ is the failure probability of the oracle $\mathcal{A}$. In particular, the algorithm succeeds with probability at least $\frac{3}{4}$ when $F_0(t) \geq \frac{K}{32}$, when one chooses the instance $\mathcal{A}$ to be $3RE$.*

After the correctness is established, one needs to bound the amount of space the algorithm uses (in addition to the random bits required to store the hash function).

Notation 77. *Let $F_{temp}(t)$ be the $3RE$ estimator at time t. Let $X_i(t) = \max(0, lsb(h_1(i)) - b)$. Let $A(t)$ be the amount of space used by the algorithm at point t, without considering the amount of space required to store the keys to hash function $h_{1,2,3}$. Similarly define $C_i(t)$. Let t_i be the points at which the estimate from $3RE$ increases.*

Proof:

In the regime $F_0(t) \geq \frac{K}{32}$, $\mathcal{E}$ holds with probability $1 - o(1)$ by the fact that $\frac{1}{\epsilon^2}$ can be chosen at least $\frac{\log n}{\log \log n}$, and thus $\frac{7}{9} - \kappa$ can be made at least $\frac{7}{9}$.

Proposition 78. $A(x) = A(x + 1)$ for $x \neq t_i$.

Proof of proposition If $x \neq t_i$, then the algorithm does not enter the last loop and thus b and C_i remain constant. Thus, A, the storage space remains constant.

Proposition 79.

$$\mathbb{P}(X(t) \geq 2K) \leq \frac{\frac{F_0(t)}{2^b}}{(2K - \frac{F_0}{2^b})^2}$$

Proof: The first observation to make is that the distribution of $lsb(h_1(i))$ is that it equals k with probability $\frac{1}{2^{k+1}}$ for $0 \leq k < \log n - 1$ and $\frac{1}{n}$ otherwise. Thus, the distribution of X_i is that it equals s with probability $\frac{1}{2^{s+b+1}}$ for $0 \leq s < \log n - b$, equals $\log n - b$ with probability $\frac{1}{n}$ and has an additional mass at 0 for the rest of the mass, i.e. a mass of $1 - \frac{1}{2^b}$.

Now, define Y_i to be a random variable that is equal to X_i when $X_i \leq \log n$, and draws a number off a geometric distribution of parameter $\frac{1}{2}$, to which it adds $\log n - b$ when the $X_i = \log n - b$. Then, $Y_i \geq X_i$ and the distribution of Y_i is a mass of $1 - \frac{1}{2^b}$ at 0, and a draw from a geometric distribution of parameter $\frac{1}{2}$ otherwise. Now, $\mathbb{E}(Y_i) = \frac{1}{2^b}$, as the mean of the geometric distribution of parameter $\frac{1}{2}$ is 1. $\mathbb{E}(Y_i^2) = 0 + \frac{1}{2^{2b}} \cdot 2 = \frac{1}{2^{2b-1}}$ and thus $Var(Y_i^2) = \frac{1}{2^{2b}}$ and thus $Var(Y_i) \leq \mathbb{E}(Y_i)$. Now, one can bound the tail of the distribution.

$$\mathbb{P}(X(t) \geq 2K) = \mathbb{P}(\sum_{i \in I(t)} X_i \geq 2K) \leq \mathbb{P}(\sum_{i \in I(t)} Y_i \geq 2K)$$

and since X_i are independent, Y_i are too and thus $\sum_{i \in I(t)} Y_i$ has sum $\frac{|I(t)|}{2^b}$ and variance at most that. Thus, Chebyshev implies

$$\mathbb{P}(\sum_{i \in I(t)} Y_i \geq 2K) \leq \mathbb{P}(\sum_{i \in I(t)} |Y_i - \frac{|I(t)|}{2^b}| \geq 2K - \frac{|I(t)|}{2^b}) \leq \frac{\frac{F_0(t)}{2^b}}{(2K - \frac{F_0(t)}{2^b})^2}$$

and this is exactly what we wanted to prove.

Proposition 80. The probability that the space $A(t)$ required by the algorithm ever exceeds $3K$ is at most $(\log n + l) \cdot \frac{\frac{F_0(t)}{2^b}}{(2K - \frac{F_0(t)}{2^b})^2}$

Proof:

$$A(t) = \sum_{i=1}^{K} \lceil \log(C_i + 2) \rceil \leq K + \sum_{i=1}^{K} \log \frac{C_i + 2}{K} \leq K \left(1 + \log\left(2 + \frac{\sum_{i=1}^{K} C_i}{K}\right)\right)$$

from Jensen's inequality and thus from 81 and one has that

$$A(t) \leq K \left(1 + \log\left(2 + \frac{\sum_{i \in I(t)} X_i}{K}\right)\right)$$

and thus

$$\mathbb{P}(A(t) \geq 3K) \leq \mathbb{P}\left(\sum_{i \in I(t)} X_i \geq 2K\right) \leq \frac{\frac{F_0(t)}{2^b}}{(2K - \frac{F_0(t)}{2^b})^2}$$

from the tail bound obtained in 79. Now, union bound gives:

$$\mathbb{P}(\exists t; A(t) \geq 3K) \leq (\log n + l)\mathbb{P}\left(\sum_{i \in I(t)} X_i \geq 2K\right) \leq (\log n + l)\frac{\frac{F_0(t)}{2^b}}{(2K - \frac{F_0(t)}{2^b})^2}$$

Proposition 81.

$$\sum_{i=1}^{K} C_i(t) \leq \sum_{i \in I(t)} X_i(t)$$

Proof: Each of the indices i seen at time t can contribute to one of the counters. Remark that if for $j = 1, K$, there exists $i \in I(t)$ with $h_1(i) = j$, then $C_i(t) \leq \max_{j \in I(t) | h_1(j) = i} X_i(t) \leq \max_{j \in I(t) | h_1(j) = i} X_i(t)$, given that X_i are positive. Given that $C_j = -1 \leq 0$ for a class j that is not the hash of anything, the inequality is established.

Notation 82. *Let $\mathcal{S}$ be the event that the algorithm fails because of space.*

Proposition 83.

$$\mathbb{P}(\exists t \mid A(t) \geq 3K) \leq \frac{1}{32}$$

Proof:

$$\mathbb{P}(\exists t \mid A(t) \geq 3K) = \mathbb{P}(\exists i \mid A(t_i) \geq 3K) \leq (\log n + l) \cdot \frac{\frac{F_0(t)}{2^b}}{\left(2K - \frac{F_0(t)}{2^b}\right)^2} \leq \frac{\log n + l}{32K} \leq \frac{1}{32}$$

for sufficiently small ϵ.

Proposition 84. [4]
Conditional on the success of the constant-factor estimator, $A(t)$ has narrower tails for large λ. More formally, for fixed t,

$$\mathbb{P}(A(t) \geq \lambda K) \leq e^{-\Omega(e^{\Omega}(\lambda))}$$

Proof:

$$
\begin{aligned}
\mathbb{P}(A(t) \geq \lambda K) &\leq \mathbb{P}\left(\exists_{s \leq K} \log\lceil C_s + 2\rceil \geq \lambda\right) \\
&\leq \mathbb{P}\left(\exists t' \leq t; lsb(h_1(t')) - b(t) \geq 2^{\lambda t}\right) \\
&\leq F_0(t) \cdot 2^{-\Omega(2^\lambda)} \\
&= F_0(t) \cdot \mathbb{P}(lsb(h_1(1)) - b(t) \geq 2^{\lambda t}) \\
&\leq F_0(t) \cdot 2^{-2^{\lambda t} - b(t)} \leq \frac{2^{-2^{\lambda t}+l+5}}{\epsilon^2}
\end{aligned} \tag{10}
$$

We conclude this section with the following result:

Proposition 85. [4]
If $t_2 \geq t_1$ such that $F_0(t_2) \leq 2F_0(t_1)$, then $A(t_1) \leq A(t_2) + \mathcal{O}\left(\frac{1}{\epsilon^2}\right)$, conditional on the success of the constant factor estimator.

Proof of the proposition:
The space to keep all of the counters is given by $\sum\lceil C_i(t_1) + 2\rceil$. Remark that from t_1 to t_2, the counters increase up to the change in b. Thus, the only decrease can come from an increase in b, which is at most constant.

Proposition 86. [4]
There exists a constant C such that at the expense of an 8-independent hash function h_1,

$$\mathbb{P}(A(t) \geq \frac{C}{\epsilon^2}) \leq \frac{1}{\epsilon^4}$$

The main idea in the proof is to look at X_i, and $T_i = X_i - \mathbb{E}(X_i)$. Now, one is interested in bounding the tails of $X_1 + ... + X_n$. One also assumes full-independence of the hashing map, as we will only work with polynomials of degree at most 8. The proof has 2 steps, when one starts with a number p and assumes T_i are 8-independent:

(1) The first step involves remarking that $||\sum X_i||_p \leq \mathbb{E}(X_i) + ||T_i||_p$.
(2) The second step involves remarking that $||\sum T_i||_p \leq \mathcal{O}(\sqrt{||\sum S_i^2||_{p/2}})$. To see this, remark that if one takes σ_i to be a vector of signs, then:

$$||\sum T_i||_p \leq 2||\sum \epsilon_i T_i||_p \leq \mathcal{O}(\sqrt{||\sum T_i^2||_{p/2}}) \leq \mathcal{O}(\sqrt{||\sum X_i^2||_{p/2}})$$

and by induction this gives that for 8-independent signs,

$$||\sum T_i||_8 \leq \frac{1}{\epsilon^4}$$

as long as $\sum ||X_i||_p \leq \mathcal{O}(\frac{1}{\epsilon^2})$ for $p \leq 8$. The last holds true, and thus

$$\mathbb{P}(\sum T_i \geq \frac{\lambda}{\epsilon^2}) \leq \mathcal{O}(\frac{1}{\lambda^8} \cdot \frac{1}{\epsilon^4})$$

which finishes the claim.

Theorem 87. [12] [4]
For any $\kappa > 0$, there exists an algorithm that, when given a stream $a_1, ..., a_T \in [n]$, it output a number t that with probability at least $\frac{7}{9} - \kappa$, t is an ϵ-approximation to the number of distinct elements in a. The algorithm requires access to $\mathcal{O}(\log n + \log \frac{1}{\epsilon})$ random bits and uses $A(t)$ additional bits of memory at time t, where A satisfies the following conditions:

(1)

$$\mathbb{P}(A(t) \geq \frac{3}{\epsilon^2}) \leq \epsilon^2$$

(2)

$$\mathbb{P}(A(t) \geq \frac{C_1}{\epsilon^2}) \leq \epsilon^4$$

for some constant C_1.

(3)

$$\mathbb{P}(A(t) \geq \frac{\lambda}{\epsilon^2}) \leq e^{-\Omega(\lambda)}$$

(4) If $t_2 \geq t_1$ such that $F_0(t_2) \leq 2F_0(t_1)$, then

$$A(t_1) \leq A(t_2) + \mathcal{O}\left(\frac{1}{\epsilon^2}\right)$$

6.3. **High probability.** This subsection is attributed to [4].

6.3.1. *Naive amplification.* The first idea in solving this question would be a naive amplification. As such, one can consider $\Theta(\frac{1}{\delta})$ independent runs of this algorithm, and consider the median. By a Chernoff-bound, probability of failure will be $\mathcal{O}(\delta)$. As such, since these runs are independent, they will independently provide approximations to F_0 with $p > \frac{1}{2}$ probability. In the constant factor approximation case, this leads to a space of $\mathcal{O}\left(\log n \cdot \frac{1}{\log \delta}\right)$, and in the ϵ-approximation case a space of $\mathcal{O}\left((\log n + \frac{1}{\epsilon^2}) \cdot \frac{1}{\log \delta}\right)$. This section is dedicated to improving these bounds.

6.3.2. *Constant factor approximation.* The first important sub-routine is presented below. This allows low-memory representation of multiple runs of F_0 estimation through a number of trials.

Remark. By a similar analysis to 3-Rough-Estimator, one can show that the algorithm returns a constant factor approximation with probability $\geq \frac{2}{3}$.

Definition 88. [4] On a finite universe $\mathcal{U}$, define a positive function f doubly exponential tailed if X is a uniform distributed variable on $\mathcal{U}$, then $\mathbb{P}(f(X) \geq \lambda) \leq 2^{-2^{\Omega(\lambda)}}$

Algorithm 4 Space-efficient storage of independent runs (h_i, $i = 1, N$ pairwise independent hash functions

Pick a family $[N]$ of pairwise independent hash-functions $h_i : [n] \to [n]$.
Update(i):
Use estimator $B_j^{(t)} = \max_{s \in I(t)} lsb(h_j(t))$. Update it after every move.
Update the median of B_j
Store $B_j^{(t)} - median_{i=1}^{w}(B_j^{(t)})$ and update after move.
Store the amount of bits necessary to store the data.
Estimator:
Compute the median m. Return 2^m.

Definition 89. [4] On a finite universe M, endowed with functions $f_1, ..., f_R$ that are doubly exponential tailed, a sequence $S \in M^*$ is said to be C-small if $\sum_{x \in S} f_i(S) \leq C|S|$

The main purpose of a C-small sequence comes from the following low-representability result:

Theorem 90. [4]
There exists a universal constant C such that for any R, M, and any $w \geq \Theta(\sqrt{R})$ and $w \geq (\log M)^{\Omega(1)}$, there exist w_1, w_2 such that $w_1 w_2 = \mathcal{O}(w)$, $w_2 = \Theta(\log w)$ $s = \mathcal{O}(w + \log M)$ and a function $G : \{0,1\}^s \times [w_1] \to [M]^{w_2}$ such that for any $g_1, ..., g_R$ with doubly exponential bounds,

$$\mathbb{P}_{U \sim Unif(\{0,1\}^s}(|k \mid \{G(U,k) \text{ is } C\text{-small}\} > \frac{w_1}{2}) \geq 1 - e^{-\Omega(w)}$$

Call such a G a sampler for R, M

This algorithm is based of [4]. G in the algorithm below is taken to be a sampler for $R = \log n$ and M a space of seeds of pairwise independent hash functions $[n] \to [n]$. Then, $\log M = \Theta(\log n)$ If M is a universe for seeds for hash function w_1, then $G(s,k)$ gives w_2 seeds of pairwise independent hash-functions, for a seed string $U \in \{0,1\}^s$ and $k \in [w_1]$.

For the analysis of this algorithm one consider t_i to be a sequence of elements such that t_i is the first point when the number of elements becomes 2^i.

Notation 91. *Denote by $e_j^{(t)}$ the deviation of the number of bits $B_j^{(t)}$ from the answer* $\log |I(t)|$, *i.e.* $e_j^{(t)} = B_j^{(t)} - \log F_0(t)$

Proposition 92.

$$\mathbb{P}(|e_j^{(t)}| \geq \lambda) \leq 2^{1-\lambda}$$

Proof:
Consider the set $S_{k,i}^{(t)} = \{a \leq t \& h_i(x_t) \leq k\}$. Then, $\mathbb{E}(|S_{k,i}^{(t)}|) = \frac{F_0(t)}{2^k}$ and because of the fact that h_1 is an independent has, $Var(|S_{k,i}^{(t)}|) = F_0(t) \cdot \frac{1}{2^k} \cdot (1 - \frac{1}{2^k}) \leq \mathbb{E}(|S_{k,i}^{(t)}|)$.

Algorithm 5 High accuracy estimator(s, C_2)

Initialization
Pick a seed string s. Pick G a sampler corresponding to s. Initialize w_1 structures $ES[w_1]$ of type space efficient storage. Initialize the hash maps as the ones given by the seeds given by $G(s,k)$.
Update:
for $t \in [w_1]$ **do**
Run update in $ES[t]$. If for some t, the amount of storage necessary for that group is $\geq C_2 w_2$, free the memory of that group and disregard it moving on.
Estimation: Return the median of the estimators of the numbers still alive.

Now, for $k = \log F_0(t) - \lambda$, $\mathbb{E}(|S_{k,i}^{(t)}|) = 2^\lambda$ and thus, $\mathbb{P}(|S_{k,i}^{(t)}| = 0) \leq \frac{1}{2^\lambda}$ from Chebyshev's inequality. One has that

$$\mathbb{P}(e_j^{(t)} \geq \lambda) = \mathbb{P}(|S_{k,i}^{(t)}| = 0) \leq \frac{1}{2^\lambda}$$

Similarly, from Markov's inequality, one has:

$$\mathbb{P}(e_j^{(t)} \leq -\lambda) = \mathbb{P}(|S_{\log F_0(t)+\lambda,i}^{(t)}| \geq 1) \leq \mathbb{E}(S_{\log F_0(t)+\lambda,i}^{(t)}) = 2^{-\lambda}$$

and the lower and upper tail bounds give a proof to the proposition.

As we move on, endow our universe with the functions

$$g_k(i) = Z_i^{(t_k)} = \log(2 + \log B_i^{(t_k)} - \log F_0^{(t_k)})$$

, and so we now have a context for the use of the sampler G, as well as a definition for C-smallness. In what it will follow, C-small refers to a sequence with respect to the functions g_i. Remark that it has double exponential-tails up to constants. The interpretation for the function g_i is that it is proportional to the space necessary to write down the deviations.

Proposition 93. If H is a C-small group, then $|\log F_0(t) - \log median_{i\in H} Y_i^{t_k}| \leq C_3$, where C_3 is a universal constant.

Proof: Because the set is C-small, $\mathbb{E}_{i\sim Unif(H)}(Z_i^{(t_k)}) < C$, which means by Markov's inequality that

$$\mathbb{P}_{i\sim Unif(H)}(Z_i^{(t_k)} \geq 3C) \geq \frac{2}{3}$$

which implies that

$$\mathbb{P}_{i\sim Unif(H)}(e_i^{t_k} \geq 2^{3C}) \geq \frac{2}{3}$$

and thus $\mathbb{P}_{i\sim Unif(H)}(e_i^{t_k} \geq 2^{3C}) \geq \frac{2}{3}$
Now, this implies that $|\log F_0(t) - \log median_{i\in H} Y_i^{t_k}| \leq 2^{3C}$

Proposition 94. To store the deviations in a C-small group H at time t_k $\mathcal{O}(w_2)$ bits are enough.

Proof: The space cost comes from 2 sources - one of them is storing the median itself. This takes $\mathcal{O}(\log\log n)$ bits. The other one is keeping track of the deviations. To bound that, remark that this space is given by $\sum_{i\in H}\log(2+|Y_i^{(t_k)}-M(t_k)|)$, where $M(t_k)$ is the median in group H at time t_k. Now,

$$\sum_{i\in H}\log(2+|Y_i^{(t_k)}-M(t_k)|)\leq\sum_{i\in H}\log(2+|Y_i^{(t_k)}-\log F_0(t_k)|+\sum_{i\in H}\log(1+|\log F_0(t_k)-M(t_k)|)$$

$$\leq C|H|+2^{3C}\cdot|H|=\mathcal{O}(|H|)=\mathcal{O}(w_2)$$

Proposition 95. 94 holds for all times t.

Proof: Pick t between t_k and t_{k+1}. Now, because Y_i is a maximum over numbers seen by time t, it is increasing in t. Consequently, $M(t)$ is increasing in t. Then, one has for every $i\in H$

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

$$\leq|Y_i^{t_k}-M(t_k)|+|Y_i^{t_{k+1}}-M(t_{k+1})|+2|M(t_k)-M(t_{k+1})|$$

and thus

$$\sum_{i\in H}\log(2+|Y_i^{(t)}-M(t)|)\leq\sum_{i\in H}\log(2+|Y_i^{t_k}-M(t_k)|)+\sum_{i\in H}\log(2+|Y_i^{t_{k+1}}-M(t_{k+1})|)$$

$$+2\sum_{i\in H}\log(1+|M(t_k)-M(t_{k+1})|)$$

and it is clear that the first 2 terms are $\mathcal{O}(w_2)$. For the last term, trinagle inequality gives

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

, which concludes that the deviations are $\mathcal{O}(w_2)$ at all times t.

Proposition 96. The total storage necessary for this algorithm is $\mathcal{O}(\log n+\log\frac{1}{\delta})$.

Proof:

There are at most w_1 groups at any point, each of which takes $\mathcal{O}(w_2)$ space to store, so this is $\mathcal{O}(w)=\mathcal{O}(\log+\frac{1}{\log\delta})$. In addition, the seed necessary for the construction takes $\mathcal{O}(s+\log N)=\mathcal{O}(\log n+\log\frac{1}{\delta}+\log poly(n))=\mathcal{O}(\log n+\frac{1}{\delta})$.

6.4. **Accurate approximation; Low failure probability.** The goal of this section will be to provide an algorithm for streaming that uses $\mathcal{O}(\log n + \frac{1}{\epsilon^2} \cdot \frac{1}{\log \delta})$.

Definition 97. [18]
A function $\Gamma : \{0,1\}^s \times [w] \to [M]$ is an (ϵ, δ) average sampler if for any function $f : [M] \to [0,1]$, when one lets $Y_i = f(U,i)$, and $\mu = \mathbb{E}(\sum_i Y_i)$, then

$$\mathbb{P}(|\sum_i f(Y_i) - w\mu| \geq \epsilon w) < \delta$$

Theorem 98. [18]
For constant ϵ, there exists a (ϵ, δ) average sampler with $w = \mathcal{O}(\log \frac{1}{\delta})$ and seed length $s = \mathcal{O}(w + \log M + \log \frac{1}{\delta})$, where the notation can hide dependency on ϵ.

Algorithm 6 Accurate approximation algorithm in the high probability regime(Γ, G)

---

if $\varepsilon < C_1(\frac{1}{\log n})^{\frac{1}{4}}$ **then**
  Pick a (ε, δ) average sampler Γ.
  for $i = 1, w$ **do**
    Run 87 with seeds given by $\Gamma(i)$. While running these, keep track of the space used and stop the process as long as the space becomes used ever becomes larger than $\frac{C_2}{\epsilon^2}$, for some constant C_4. Keep track of the reported estimators.
  Report the median estimate from all of these.
else
  for $i = 1, w_1$ **do**
    Run 87 with seeds given by $G(i)$. Consider the median.
    Keep track of the memory usage in every group.
    Discard a group if the memory usage becomes $\geq \frac{C_3}{\epsilon^2}$.
    Report the median
  Report the median of all the groups still standing.

One needs to specify conditions for Γ and G. The conditions are listed below:

(a) Γ needs to be a $(\frac{1}{10}, \gamma)$ sampler.
(b) Consider $w = \Omega(\log \frac{1}{\delta} + \log n)$. Then, let G be an explicit sampler for $R = \log n + 1$, and M a space of seeds for an algorithm as in 87. $\log M = \Theta(\log n)$.
(c) Consider $g_i = \frac{\epsilon^2}{C} \cdot A_m^{(t_i)}$ for $i = 1, R$. Consider g_R to be a function that is 0 on instantiations m that give $1 + \epsilon$ approximations and a large constant C_0 otherwise. If C_0 is high enough, given that the algorihtm succeeds with probability $\frac{5}{6}$, naive amplification can make the algorithm fail with probability c_0, for some arbitrarily small constant c_0 maintaining the same asymptotic guarantees. Once c_0 is small enough, g_R will be doubly exponential bounded. From 87, g_i are also doubly exponential bounded.

First, remark that this is the exact opposite of the constant failure case in terms of assumptions on parameters. This is because one can assume $\frac{1}{\epsilon^2} \leq C \log n$, since if not, the naive amplification provides an asymptotically correct bound.

Proposition 99. (Correctness for small ε case.) In the $\epsilon < (\frac{1}{\log n})^{1/4}$ case described above, with probability $\mathcal{O}(\delta)$ at least $\frac{w}{2}$ of the seeds $\Gamma_s(i)$ will give ϵ-approximations to F_0.

Proof:

First, remark that in this regime of parameters, one can assume $\log \frac{1}{\delta} \geq \frac{1}{\sqrt{\log n}}$. This is because otherwise, we can work with a larger δ, and the term $\log n$ is dominating the term $\frac{1}{\epsilon^2} \cdot \log \frac{1}{\delta}$.

For $i = 1, w$, let Q_k be the estimator given by the algorithm when supplied seed k. Then, one can consider the function $f(k) = 1_{|Q_k - F_0^{(T)}| \geq \epsilon F_0^{(T)}}$. One is guaranteed that $\sum_{i \in M} f(i) \geq \frac{5}{6}M$. Let $\mu = \frac{\sum_{i=1}^{M} f(i)}{M} \geq \frac{5}{6}$. Because Γ is a $(\frac{1}{10}, \delta)$ average sampler, one has that

$$\mathbb{P}(|\sum_{i=1}^{w} f(\Gamma(i)) - (\frac{7}{9} + \kappa)w| \leq \frac{w}{6}) \leq \delta$$

By the triangle inequality, this means that

$$\sum_{i=1}^{w} f(\Gamma(i)) \geq \frac{2}{3}w$$

Now, denote the events $|Q_{\Gamma(i)} - F_0^{(T)}| \leq \epsilon F_0^{(T)}$ by $\mathcal{E}_k$. Now, at least $\frac{2}{3}w$ out of the w $\mathcal{E}_k$ happen and thus the median will lead to an ϵ-approximation, with failure probability δ. One needs to take care of the groups that get finished because of space usage. Remark that every instance of 87, when instantiated with 8-independent h_1, verifies that at any position, the space usage is smaller than $\mathcal{O}(\frac{1}{\epsilon^2})$ with probability $\frac{1}{\epsilon^4} \leq \frac{\kappa}{\log n}$, where κ can be made arbitrarily small based on $\log n$. Now, consider a sequence t_i of indices such that $F_0(t_i) = 2^i$. Then, there are $\log n$ such indices. By union bound, there are at most $\log n$ of these and thus probability of failure at some point t_i is at most κ. Thus, probability of failing at any point is at most κ, just by enlarging the term $\mathcal{O}(\frac{1}{\epsilon^2})$. Thus, with probability $1 - \kappa$, the space consumption at any point is at most C_2/ϵ^2 with probability $1 - \kappa$. Now, let g be an indicator of whether the instations uses more than $\frac{C_2}{\epsilon^2}$ space. It will follow that at most $(\frac{1}{10} + \kappa)w$ of the instatiations would have this property. Consequently, the algorithm's corectness will be proven, since at most $(\frac{1}{10} + \kappa)w$ of these groups die, and so at least $(\frac{2}{3} - \frac{1}{10} - \kappa)w > \frac{w}{2}$ for sufficiently large κ of the seeds survive and provide ϵ-approximations.

Thus, the failure of this algorithm is at most 2δ, which can be fixed by re-scaling to start with.

Proposition 100. (Space for the small ε case)
The space usage of this algorithm is $\mathcal{O}(\log n + \frac{1}{\epsilon^2} \cdot \log \frac{1}{\delta})$ with probability $\mathcal{O}(\delta)$.

Proof:
There are $\log \frac{1}{\delta}$ seeds, and each of them requires $\mathcal{O}(\frac{1}{\epsilon^2})$ space. The seed requires $\mathcal{O}(\log n + \log \frac{1}{\delta})$ bits to store, which finishes the result.

Proposition 101. (Correctness for the large ε case.)
For any i, if $G(U, i)$ is C-small, the median estimator in the group is a $(1+\epsilon)$ approximation to F_0.

Proof: One needs to pay attention to the function g_{R+1}. Applying the C-small condition for the g_{R+1} function will give that at most $\frac{C}{C_0}$ of the elements in a C-small group are not $1+\epsilon$ approximation. Choosing $C_0 \geq 2C$ will give at least half of the elements in each group the property that is a $1+\epsilon$ approximation to F_0, which shows the property.

Proposition 102. A C-small group does not use more than $\mathcal{O}(\frac{1}{\epsilon^2})$ bits of space. One needs a seed $s = \mathcal{O}(w + \log n)$, and clearly we need $w = \mathcal{O}(\log \frac{1}{\delta})$.

Proof:
One needs to show that a C-small group occupies space $\mathcal{O}(\frac{1}{\epsilon^2})$ at any point in time. First, remark that this is guaranteed at steps $t_1, ..., t_R$ from the C-smallness condition applied to $g_1, ..., g_R$. But now one can remark that the space usage verifies the condition that the diffference between points where F_0 at most doubles is at most $\mathcal{O}(\frac{1}{\epsilon^2})$, which means that the space consumptions is $\mathcal{O}(\frac{1}{\epsilon^2})$. Thus, for a sufficiently large C, the space consumption is at most $C_3 \frac{1}{\epsilon^2}$ for some constant C_3.

Now, the total space consumption is $\mathcal{O}\left(\frac{1}{\epsilon^2} \cdot \log \frac{1}{\delta}\right)$ for the instantiations of 87 and $\mathcal{O}(\log n + \log \frac{1}{\delta})$ for the random seed to the $R+1, M$ sampler.

6.5. **Tracking in the high probability regime.**

6.5.1. *Naive amplification.* The first idea would be naive amplification. Remark that because $(1+\epsilon)^2 - 1 = \mathcal{O}(\epsilon)$, it is enough to provide approximations at points t_i in the stream at which the number of distinct elements breaks $(1+\epsilon)^i$ for the first time. But there are $\frac{\log n}{\log(1+\epsilon)}$ such indices. Thus, dropping the probability of failure to $\delta \cdot \frac{\log(1+\epsilon)}{\log n}$ would guarantee by a union bound an $\mathcal{O}(\epsilon)$ approximation at any point. This algorithm would run in space $\mathcal{O}(\log n + \frac{\log \frac{1}{\delta}}{\epsilon^2} + \frac{\log \log n}{\epsilon^2} + \frac{\log \frac{1}{\epsilon}}{\epsilon^2})$. The goal of the following would be to provide a better bound. In particular, we will prove one can get rid of the $\frac{\log \frac{1}{\epsilon}}{\epsilon^2}$.

6.5.2. *Tracking.* For the high probability regime, one will need the following result, that is an improvement of 75. We will return to analyzing 87

Proposition 103. 87 as described above is able to provide an approximation to F_0 at all points t. More formally, the algorithm provides an estimate that is at most $\epsilon F_0(T)$ away from the actual answer with probability $\frac{7}{10}$ for all t simultaneously.

Proof:

First of all, denote $I_b(t)$ to be the set of elements i with $lsb(h_1(i)) \geq b$. One can clearly restrict the analysis to the times t when an element is added. Consider $Y_k = 2^b|I_b(t)| - 2^b|I_{b-1}(t-1)|$. Now, Y_k are 4-wise independent and have variance $\mathcal{O}(2^b)$ and thus from 11,

$$\mathbb{P}\left(\exists t \leq T \text{ such that } |2^b \cdot |I_b(t)| - t| \geq \mathcal{O}\left(\epsilon |F_0(T)|\right)\right) \leq \frac{9}{10}$$

which means from the Lipschitz properties of Φ implies that the approximations $2^b \cdot \Phi^{-1}(T(t))$ are all accurate.

Now, at point t, the result would be the number of I_b balls thrown independently into $\frac{1}{\epsilon^2}$ bins, as h_2 contains no collisions with high probability. Now, consider $T(t)$ to be the T from the algorithm at time t. Then, $|h_3(h_2(T))| = \Phi_{\frac{1}{\epsilon^2}}(|I_b(t)|) \pm \mathcal{O}(\epsilon K)$ with $\frac{9}{10}$ probability from 48 at all times t. From the Lipschitz properties of $\Phi_{\frac{1}{\epsilon^2})}$, one will have that $\Phi^{-1}_{\frac{1}{\epsilon^2})}(T)$ stays closer to F_0 than $\mathcal{O}(\epsilon F_0(T))$ at all times with probability at least $\frac{7}{10}$.

Theorem 104. [4]

There exists an ϵ-strong tracking algorithm for F_0 that uses $\mathcal{O}(\frac{\log\log n + \log\frac{1}{\delta}}{\epsilon^2} + \log n)$ bits of space.

Without going into the details of the analysis, an amplification method using average samplers similar to the one done for the one timestamp approximation can be analogously applied here. As such, one can use an amplification to $\frac{\delta}{\log n}$ to ensure, after a union bound, can ensure the above is true for any $b \leq \log n$ (i.e. union bounding at points t_k at which F_0 increases by a factor of 2). Thus, since the space consumption for the regular algorithm is $\mathcal{O}\left(\log n + \frac{\log\frac{1}{\delta}}{\epsilon^2}\right)$, the result follows.

7. Frequency moments

7.1. $p \in (0, 2)$**.** We will be interested in tracking the p-th frequency norm of the stream. We will give bounds for both weak and strong tracking. The algorithm is attributed to [9]. The analysis is attributed to [5].

7.1.1. *Weak tracking.*

Proposition 105. Correctness of the algorithm above

One can choose $s = \Theta\left(\epsilon^{-p}\right)$, $r = \Theta\left(\log\frac{1}{\epsilon} + \log\frac{1}{\delta}\right)$, $d = \Theta\left(\frac{1}{\epsilon^2}(\log\frac{1}{\delta} + \log\frac{1}{\epsilon})\right)$ such that if the estimator return m_t at time t, then

$$\mathbb{P}\left(\exists t \leq T; |m_t - |x^{(t)}|| \geq \epsilon |x^{(T)}|\right) \leq \epsilon ||x^{(T)}||_p$$

Proof:

Algorithm 7 Weak tracking of the pth moment of the frequency vector

Pick Π a $d \times n$ matrix with r-independent rows and s-independent entries on every row.
Initialize sketch r of size d to the 0 vector.
Update(i):
Update the sketch $r = r + \Pi^{(a_i)}$, where $\Pi^{(a_i)}$ is the column corresponding to a_i.
Estimator:
At point t in the stream, estimate the norm by considering the median entry in the sketch r.

Just as in the case of the distinct elements problem, one will be looking at a sequence of t_i. Consider S to be a set of indices $0 = t_0 < t_1 < ... < t_k$ such that $||x_{t_i} - x_{t_{i-1}}||_p \geq \epsilon^{\frac{4}{p}} \cdot ||x_T||_p$. One can break into cases:

(1) If $p \geq 1$, then t_i just constitutes the construction of an $\epsilon^{\frac{4}{p}}$-net described in part 1 for the metric given by $d(x, y) = ||x - y||_p$. By 50, $k \leq \epsilon^{-4}$.
(2) If $p \leq 1$, then

$$
\begin{aligned}
||x^{(T)}||_p^p &= ||\sum_{i=0}^{q-1} x^{(i+1)} - x^{(i)}||_p^p || \sum_{i=0}^{q-1} x^{(i+1)} - x^{(i)}||_p^p \\
&\geq q^{p-1} \sum_{i=0}^{q-1} ||x^{(i+1)} - x^{(i)}||_p^p \text{ by Holder's inequality} \\
&\geq q^{p-1} \cdot q \cdot \epsilon^4 ||x^{(T)}||_p^p
\end{aligned} \tag{11}
$$

and thus $q \leq \epsilon^{-\frac{4}{p}}$.

For an individual row π_i, and an individual frequency vector $x^{(t)}$, remark that if γ is a vector of length n of independent $\mathcal{D}_p$ distributions, then $\langle \gamma, x^{(t)} \rangle$ has a distribution of $||x^{(t)}|| \cdot \mathcal{D}_p$.

The first step in the proof would be to prove that one can choose d, s, r as stated and the probability of failure at points t_i is $\mathcal{O}(\delta)$. The strategy involves breaking the failure event into 2 cases - failure because the algorithm returns a number that is too high is too low or failure because the answer 32 implies that

$$
\begin{aligned}
\mathbb{P}(|\langle \pi_i, x^{(t)} \rangle| - ||x^{(t)}||_p \geq \epsilon ||x^{(T)}||_p) &= \mathbb{P}(||\langle \gamma, x^{(t)} \rangle| - ||x^{(t)}||_p| \geq \epsilon ||x^{(T)}||_p) + \mathcal{O}(s^{-\frac{1}{p}}) \\
&= \mathbb{P}(|Z| \geq 1 - \epsilon) + \mathcal{O}(s^{-\frac{1}{p}}) \text{ where } Z \text{ is sampled off } \mathcal{D}_p \\
&= \frac{1}{2} - \Omega(\epsilon) + \mathcal{O}(s^{-\frac{1}{p}}) \\
&= \frac{1}{2} - \Omega(\epsilon) \text{ for a correct dependence of } s \text{ on } \epsilon^{-p}
\end{aligned} \tag{12}
$$

Absolutely similarly, $\mathbb{P}(|\langle \pi_i, x^{(t)} \rangle| - ||x^{(t)}||_p \leq -\epsilon ||x^{(T)}||_p) = 1 - \Omega(\epsilon)$

Let U_j be a random Bernoulli variable tracking whether $|\langle \pi_i, x^{(t)} \rangle| - ||x^{(t)}||_p \geq \epsilon ||x^{(T)}||_p$ and L_j be a random Bernoulli variable tracking whether $|\langle \pi_i, x^{(t)} \rangle| - ||x^{(t)}||_p \leq -\epsilon ||x^{(T)}||_p$. Then, $\mathbb{E}\left(\sum_{i=1}^{d} L_j\right) = \frac{d}{2} - \Omega(d\epsilon) \leq \frac{d}{2} - Cd$ for some constant c. Then, by 11,

$$\mathbb{P}\left(\sum_{i=1}^{d} L_j \geq \frac{d}{2} - \frac{C}{2}d\right) \leq \mathcal{O}\left(e^{-\Omega(d\epsilon^2)}\right) + e^{-\Omega(r)}$$

Thus, one can choose $d = \Theta(\epsilon^{-2}(\log \frac{1}{\epsilon} + \log \frac{1}{\delta}))$, $r = \Theta(\log \frac{1}{\delta} + \log \frac{1}{\epsilon})$ such that the above is $\mathcal{O}\left(\delta \epsilon^{\frac{8}{p}}\right)$. A similar bound holds true for U_j. By a union bound, remark that

$$\mathbb{P}\left(\exists_{i \leq k} \sum_{i=1}^{d} L_j \geq \frac{d}{2} - \frac{C}{2}d \text{ or } \sum_{i=1}^{d} U_j \geq \frac{d}{2} - \frac{C}{2}d \text{ at } t_k\right) = \mathcal{O}(\delta)$$

Now, let $F_{i,j}$ be the indicator of the event that $\exists t \in [t_j, t_{j+1} - 1] \mid |\langle x^{(t)} - x^{(t_j)}, \pi_i \rangle \geq \epsilon ||x^{(T)}||_p$. Remark that $0 = x^{(t_j)} - x^{(t_j)} \leq x^{(t_j+1)} - x^{(t_j)} \leq ... \leq x^{(t_{j+1}-1)} - x^{(t_j)}$, and $||x^{t_{j+1}-1} - x^{(t_j)}|| \leq \epsilon^{\frac{4}{p}} \cdot ||x^{(T)}||_p$. Thus, 53 implies $\mathbb{E}(F_{i,j}) \leq \mathcal{O}(\left(\frac{\epsilon}{\epsilon^{\frac{4}{p}}}\right)^{-\frac{2p}{2+p}}) + \mathcal{O}(s^{-\frac{1}{p}}) = \mathcal{O}\left(\epsilon^{\frac{2(4-p)}{2+p}}\right) = \mathcal{O}(\epsilon)$ which can be made at most $\frac{C}{4}\epsilon$ for an appropriate choice of d and s. Thus, for fixed j, $\sum_{i=1}^{d} \mathbb{P}(F_{i,j}) \leq \frac{C}{4} d\epsilon$. Now, by 11, one can obtain again that:

$$\mathbb{P}(\sum_{i=1}^{n} F_{ij} \geq \frac{C}{2} d\epsilon) \leq e^{-\Omega(d\epsilon)} + e^{-\Omega(r)} \leq \delta \epsilon^{-\frac{8}{p}} = \mathcal{O}(\frac{\delta}{q})$$

for an appropriate choice of d and r. Now, this implies that

$$\mathbb{P}(\exists j \sum F_{ij} \leq \frac{C}{2} d\epsilon) \leq \mathcal{O}(\delta)$$

Assume that neither of $\left(\exists_{i \leq k} \sum_{i=1}^{d} L_j \geq \frac{d}{2} - \frac{C}{2}d \text{ or } \sum_{i=1}^{d} U_j \geq \frac{d}{2} - \frac{C}{2}d \text{ at } t_k\right)$ or $\left(\exists j \sum F_{ij} \leq \frac{C}{2} d\epsilon\right)$ holds. This happens with probability $1 - \mathcal{O}(\delta)$. Fix some i. Then, by the triangle inequality, $|\langle \pi_i, x^{(t)} \rangle| \leq |\langle \pi_i, x^{(t_j)} \rangle| + |\langle \pi_i, x^{(t_j)} - x^{(t)} \rangle|$, where t_j is the last element in the sequence t encountered before t. Now, this is smaller than $|x^{(t_j)}|_p + 2\epsilon |x^{(T)}|$ for more than $\frac{d}{2}$ of $i \in d$. Similarly, is is bigger than $|x_{(t_j)}|_p - 2\epsilon$ for at least half of them. Thus, the median estimator will return a number s_t between $|x^{(t_j)}|_p \pm 2\epsilon |x^{(T)}|_p$. But now $|x^{(t)} - x^{(t_j)}|_p \leq \epsilon^{\frac{4}{p}} |x^{(t_j)}|$ by construction. Up to constants depending on p, $|x^{(t)} - x^{(t_j)}|_p \geq \Omega(|x^{(t)}|_p - |x^{(t_j)}|_p)$, which proves that given $p \leq 2$, one has that:

$$|s_t - ||x^{(t)}||_p| \leq \mathcal{O}(\epsilon) \cdot ||x^{(T)}||_p$$

which proves the weak tracking claim.

Theorem 106. *The above algorithm can be implemented using* $\mathcal{O}(\epsilon^{-2}\log T(\log\frac{1}{\epsilon}+\log\frac{1}{\delta}))$ *for the sketch and* $\mathcal{O}(\epsilon^{-p}(\log\frac{1}{\epsilon}+\log\frac{1}{\delta})(\log T+\log n))$ *for maintaining the matrix* d*. In paricular, one needs space* $\mathcal{O}(\epsilon^{-2}(\log T+\log n)(\log\frac{1}{\epsilon}+\log\frac{1}{\delta}))$ *for weak-tracking of* F_p *of a stream.*

The main ingredient in proving this result is sampling off a distribution nearby this one that only requires $\tau=\mathcal{O}(\log m\epsilon^{-1})$ random bits to be described.

Theorem 107. *One can achieve strong tracking guarantees in space* $\mathcal{O}\left(\epsilon^{-2}(\log T+\log n)(\log\frac{1}{\epsilon}+\log\frac{1}{\delta}+\log\log T)\right)$

Proof: Remark that $f(a_1)\geq 1$, and $f(a_1,...,a_T)\leq T^2$, as the biggest norm one can see clearly happens for $p=2$, and for $p=2$ the norm is concave is the frequency vector, which means it is maximized when one of the frequency elements is T and the others are 0. 5 implies the result.

7.2. $p=2$**.** The results in and guarantee F_2 approximation bounds for similar asymptotic dependencies. This section will be dedicated to improving these bounds.

Theorem 108. [1]
There exists an algorithm $\mathcal{A}$ *that uses* $\mathcal{O}(\epsilon^{-2}(\log m+\log n)\log\frac{1}{\delta})$ *bits of space that provides an* ϵ*-approximation to* F_2 *with fialure probability* δ*.*

Proof:
We will first go to describe the original [1] method and an improvement suggested in [6] for F_2 estimation of a stream.

Algorithm 8 Sketch given in [1]

Initialization
Pick a matrix Π of size $k\times n$ matrix for $d=\Theta(\frac{1}{\epsilon^2})$ which is composed of 8-wise independent variables that are ± 1 each with probability $\frac{1}{2}$.
Initialize a sketch vector $v=0$ of dimension d.
Update step(t):
Make $v=v+\Pi^{(a_t)}$, where a_t is the element seen at time step t, and $\Pi^{(a_t)}$ is the column corresponding to a_t.
Evaluation
Return the average F_2 estimator given by the above.

A stronger result was proven in [6].

Theorem 109. [6]
The algorithm in this section can be implemented in $\mathcal{O}\left(\epsilon^{-2}(\log n+\log T)\log\frac{1}{\delta}\right)$ *to provide* ϵ*-weak tracking to* F_2 *with failure probability* δ*, at the cost of making entries in every row in* Π 8*-wise independent. That is by amplifying the algorithm* $\log(\frac{1}{\delta})$ *times and taking the median for amplifying accuracy.*

Theorem 110. *The algorithm in this section can be implemented in* $\mathcal{O}\left(\epsilon^{-2}(\log n + \log T)(\log \frac{1}{\delta} + \log\log T)\right)$ *to provide* ϵ*-strong tracking for* F_2 *with failure probability* δ*, at the cost of making entries in every row in* Π 8*-wise independent. This guarantees strong tracking for* l_2 *in the same assymptotic time.*

Proof: Follows as a corollary to 109 and 5. We will now give a proof to 109.

Proof of 109:

First, let $l_2(t)$ be the l_2 norm at time t. First, consider the sequence of frequency vectors $f_1, ..., f_T$. Let $v_1, ..., v_T$ be the normalized version i.e, $f_i = \frac{f_i}{|f_T|}$. Then, $v_0 = 0$ and $|v_T| = 1$. The first thing to note is that the sketch that is loaded in memory is $\Pi x^{(t)}$, where $x^{(t)}$ is the frequency vector at time t. One can remark that $\Pi x^{(t)}$ can be dually represented by $A^{(t)}\tilde{\Pi}$, where $\tilde{\Pi}$ is vectorizing the matrix Π, and A is a $d \times (nd)$ matrix, that has blocks of the vector $x^{(t)}$ on the rows . Thus, $\sqrt{k}|\Pi x^{(t)}| = |A^{(t)}Z|$. Now, define $B^{(t)} = \frac{1}{k}(A^{(t)})^T A^{(t)}$. Then the estimator is given by $p_t(Z) = Z^T(A^{(t)})^T A^{(t)} Z = Z^T B^{(t)} Z$. Remark that $\mathbb{E}(p_t(Z)) = l_2(t)^2$. Consequently, one is interested in bounding the quanity $\mathbb{P}(\sup_{t\leq T} |\sqrt{p_t(Z)} - l_2(t)| \geq \lambda l_2(T))$ to obtain tracking guarantees. For this, it would be sufficient to provide a bound of the form

$$\mathbb{P}\left(\sup_{v\in V} |Z^{(t)T} B_v Z^{(t)} - \mathbb{E}(Z^{(t)T} B_v Z^{(t)})| \geq \epsilon\lambda\right) = o(1)$$

, when one looks at the dependency on λ.

Now, the Cauchy-Schwarz inequality, for vectors x, y of n elements and norm at most 1, the following holds true:

$$||x^T x - y^T y||_F \leq 4|x - y|_2$$

Now, let E_l be a $\frac{1}{2^l}$-net for the set of vectors v_i. Let $T_l = \{B_v \mid v \in T_l\}$. Then, by the inequality, above, T_l is a $\frac{4}{\sqrt{k}\cdot 2^l}$ net in $||\cdot||_F$ and thus in $||\cdot||$ as well. Now, given a sample Z of 8-wise independent signs, consider $\gamma(B) = |Z^{(t)T} B Z^{(t)} - \mathbb{E}(Z^{(t)T} B Z^{(t)})$. Now, very similar to the proof of 51, one will consider $x^{(l)}$ to be the closest point to x inside the net E_l. Then, $|x - x^{(l)}| \leq \frac{1}{2^l}$ and thus $||B_x - B_{x^{(l)}}||_F \leq \frac{4}{\sqrt{k}\cdot 2^l}$. Let D_l be the set of differences $B_{x^{(l+1)}} - B_{x^{(l)}}$ for $v \in V$. Then, $|D_l| = \mathcal{O}(2^{2l})$ Now, for some constant c

$$\mathbb{P}(\sup_{v\in V} \gamma(B_v) \geq c\lambda) \leq \sum \mathbb{P}(\sup_{B\in D_l} \gamma(B) \geq \frac{\lambda}{2^{i/3}})$$

To finish, we will use the following corollary of the Hanson-Writght inequality:

Theorem 111. [6]

If Z_i *is a sequence of independent Rademachers, then*

$$||Z^T A Z - \mathbb{E}(Z^T A Z)||_p = \mathcal{O}(||B||_F)$$

, where the constant is allowed to dependent on p.

Remark. For integers p, remark that Z does not need to be independent, but could as well be just $2p$-independent, since it only concerns a polynomial of degree $2p$ in Z_i.

Given, this, for $p = 4$, remark that

$$\mathbb{P}(\sup_{v \in E_i} \gamma(v) \geq \frac{\lambda}{2^{i/3}}) \leq \mathcal{O}(|D_i| \cdot \frac{\max_{A \in D_l} \mathbb{E}(|Z^T AZ - \mathbb{E}(Z^T AZ)|^4)}{\frac{\lambda^4}{2^{4i/3}}}) = \mathcal{O}(\frac{1}{\sqrt{k} \cdot \lambda^4} \cdot 2^{-2i/3})$$

and thus we get that $\mathbb{P}(\sup \gamma(B_v) \geq \Omega(\lambda)) = \mathcal{O}(\frac{1}{\sqrt{k} \cdot \lambda^4})$. Now, given that we take $k = \Theta(\frac{1}{\epsilon^2})$, the result is proven.

To finish the result, remark that one can take $\Theta(\log \frac{1}{\delta})$ independent runs of this algorithm. By a Chernoff bound, probability of failure will be bounded by δ. The space usage follows immediately.

References

[1] Noga Alon, Yossi Matias, and Mario Szegedy. 1996. The space complexity of approximating the frequency moments. In Proceedings of the twenty-eighth annual ACM symposium on Theory of computing (STOC '96). ACM, New York, NY, USA, 20-29.

[2] K. Azuma, "Weighted sums of certain dependent random variables," Tohoku Mathematical Journal, vol. 19, pp. 357–367

[3] 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.

[4] Jarosław Błasiok. Optimal streaming and tracking distinct elements with high probability. Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms. 2018, 2432-2448

[5] Jaroslaw Blasiok, Jian Ding, and Jelani Nelson. Continuous monitoring of p norms in data streams. In Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2017, August 16-18, 2017, Berkeley, CA, USA, pages 32:1–32:13, 2017

[6] Vladimir Braverman, Stephen R. Chestnut, Nikita Ivkin, Jelani Nelson, Zhengyu Wang, and David P. Woodruff. BPTree: an 2 heavy hitters algorithm using constant memory. In Proceedings of the 36th SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS), 2017.

[7] Carter, and Wegman. "Universal Classes of Hash Functions." Journal of Computer and System Sciences, vol. 18, no. 2, 1979, pp. 143–154.

[8] Doob, J. L. Notes on Martingale Theory. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Contributions to Probability Theory, 95–102, University of California Press, Berkeley, Calif., 1961. https://projecteuclid.org/euclid.bsmsp/1200512595

[9] Piotr Indyk. Stable distributions, pseudorandom generators, embeddings, and data stream computation. J. ACM, 53(3):307–323, May 2006.

[10] Joag-Dev, Kumar, and Frank Proschan. "Negative Association of Random Variables with Applications." The Annals of Statistics, vol. 11, no. 1, 1983, pp. 286–295.

[11] Hoeffding, W. (1963). "Probability inequalities for sums of bounded random variables". Journal of the American Statistical Association. 58 (301): 13–30

[12] Kane, Daniel M., Nelson Jelani, Woodruff P. David. "An optimal algorithm for the distinct elements problem." PODS (2010).

[13] Daniel M. Kane, Jelani Nelson, and David P. Woodruff. On the exact space complexity of sketching and streaming small norms. In Proceedings of the 21st Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), pages 1161–1178, 2010.

[14] U. Haagerup. The best constants in the Khintchine inequality. Studia Math., 70(3):231–283, 1982.

[15] McDiarmid, Colin (1989). "On the Method of Bounded Differences". Surveys in Combinatorics. 141: 148–188.

[16] J. P. Nolan. Stable Distributions - Models for Heavy Tailed Data. Birkhauser, Boston, 2017. In progress, Chapter 1 online at http://fs2.american.edu/jpnolan/www/stable/stable.html.

[17] R. E. A. C. Paley and A. Zygmund, A note on analytic functions in the unit circle, Proc. Camb. Phil. Soc. 28 (1932), 266–272

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