Continuous Monitoring of ℓ_p Norms in Data Streams*

Jarosław Błasiok[1], Jian Ding[2], and Jelani Nelson[3]

1 **Harvard University, Cambridge, MA, USA**
jblasiok@g.harvard.edu
2 **University of Chicago, Chicago, MA, USA**
jianding@galton.uchicago.edu
3 **Harvard University, Cambridge, MA, USA**
minilek@seas.harvard.edu

Abstract

In insertion-only streaming, one sees a sequence of indices $a_1, a_2, \ldots, a_m \in [n]$. The stream defines a sequence of m frequency vectors $x^{(1)}, \ldots, x^{(m)} \in \mathbb{R}^n$ with $(x^{(t)})_i \stackrel{\text{def}}{=} |\{j : j \in [t], a_j = i\}|$. That is, $x^{(t)}$ is the frequency vector after seeing the first t items in the stream. Much work in the streaming literature focuses on estimating some function $f(x^{(m)})$. Many applications though require obtaining estimates at time t of $f(x^{(t)})$, for every $t \in [m]$. Naively this guarantee is obtained by devising an algorithm with failure probability $\ll 1/m$, then performing a union bound over all stream updates to guarantee that all m estimates are simultaneously accurate with good probability. When $f(x)$ is some ℓ_p norm of x, recent works have shown that this union bound is wasteful and better space complexity is possible for the continuous monitoring problem, with the strongest known results being for $p = 2$ [29, 10, 9]. In this work, we improve the state of the art for all $0 < p < 2$, which we obtain via a novel analysis of Indyk's p-stable sketch [30].

1998 ACM Subject Classification F.2 Analysis of Algorithms and Problem Complexity

Keywords and phrases data streams, continuous monitoring, moment estimation

Digital Object Identifier 10.4230/LIPIcs.APPROX/RANDOM.2017.32

1 Introduction

Estimating statistics of frequency vectors implicitly defined by insertion-only update streams, as defined in the abstract, was first studied by Flajolet and Martin in [24]. They studied the so-called *distinct elements problem*, in which $f(x)$ is the support size of x. In the insertion-only model, the support size of x is equivalent to the number of distinct a_i appearing in the stream. One goal in such streaming algorithms, both for this particular distinct elements problem as well as for many others function estimation problems studied in subsequent works, is to minimize the space consumption of the stream-processing algorithm, ideally using $o(n)$ words of memory (note there is always a trivial n space algorithm by storing x explicitly in memory).

For over two decades, work on estimating statistics of frequency vectors of streams remained dormant, until the work of [1] on estimating the p-norm $\|x\|_p = (\sum_{i=1}^n x_i^p)^{1/p}$ in

* J.B. supported by ONR grant N00014-15-1-2388. J.D. partially supported by NSF grant DMS-1455049 and an Alfred P. Sloan Research Fellowship. J.N. supported by NSF grant IIS-1447471 and CAREER award CCF-1350670, ONR Young Investigator award N00014-15-1-2388, and a Google Faculty Research Award.

Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2017).
Editors: Klaus Jansen, José D. P. Rolim, David Williamson, and Santosh S. Vempala; Article No. 32; pp. 32:1–32:13
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

streams for integer $p \geq 1$. Since then several works have studied these and several other problems, from the perspective of both upper and lower bounds, including estimating $\|x\|_p$ for all $0 < p \leq 2$ (not necessarily integral) [1, 30, 32, 49, 41, 42, 39, 44, 38, 37], $\|x\|_p$ for $p > 2$ [1, 4, 21, 33, 7, 26, 34, 2, 16, 13, 25], empirical entropy [19, 20, 6, 28] and other information-theoretic quantities [31, 27, 14], cascaded norms [22, 36, 35], and several others. There have also been general theorems classifying which statistics of frequency vectors admit space-efficient streaming estimation algorithms [15, 8, 17, 12, 11].

Taking a dynamic data structural viewpoint, "streaming algorithms" is simply a synonym for "dynamic data structures" but with an implied focus on minimizing memory consumption (typically striving for an algorithm using *sublinear* memory). Elements in the stream can be viewed as updates to the frequency vector x (seeing $a \in [n]$ in the stream can be seen as $\texttt{update}(a, 1)$, causing the change $x_a \to x_a + 1$), and the request for an estimate of some statistic of x is a query. In this data structural language, all the works cited in the previous paragraph provide Monte-Carlo guarantees of the following form for queries: starting from any fixed frequency vector and after executing any fixed sequence of updates, the probability that the output of a subsequent query then fails is at most δ. Here we say a query fails if, say, the output is not a good approximation to some particular $f(x)$ (this will be made more formal later). In many applications however, one does not simply want the answer to one query at the end of some large number of updates, but rather one wants to *continuously* monitor the data stream. That is, the sequence of data structural operations is an intermingling of updates and queries. For example, one may have a threshold T in mind, and if $f(x)$ ever increases beyond T some data analyst should be alerted. Such a goal could be achieved (approximately) by querying after every update to determine whether the updated frequency vector satisfies this property. Indeed, the importance of supporting continuous queries in append-only databases (analogous to the insertion-only model of streaming) was recognized 25 years ago in [47], with several later works focused on continuous stream monitoring with application areas in mind such as trend detection, anomaly detection, financial data analysis, and (bio)sensor data analysis [3, 18, 46].

If one assumes that a query is being issued after every update, then in a stream of m updates the failure probability should be set to $\delta \ll 1/m$ so that, by a union bound, all queries succeed. Most Monte-Carlo streaming algorithms achieve some space S to achieve failure probability $1/3$, at which point one can achieve failure probability δ by running $\Theta(\lg(1/\delta))$ instantiations of the algorithm in parallel and returning the median estimate (see for example [1]). This method increases the space from S to $\Theta(S \lg(1/\delta))$, and for many problems (such as ℓ_p-norm estimation) it is known that at least in the so-called strict turnstile model (i.e. $\texttt{update}(a, \Delta)$ is allowed for both positive and negative Δ, but we are promised $x_i \geq 0$ for all i at all times) this form of space blow-up is necessary [37]. Nevertheless, although improved space lower bounds have been given when desiring that the answer to a *single* query fails with probability at most δ, no such blow-up has been shown necessary for the continuous monitoring problem in which one wants, with failure probability $1/3$, to provide simultaneously correct answers for m queries intermingled with m updates. In fact to the contrary, in certain scenarios such as estimating distinct elements or the ℓ_2-norm in insertion-only streams, improved *upper bounds* have been given!

▶ **Definition 1.** We say a Monte-Carlo randomized streaming algorithm $\mathcal{A}$ provides ***strong tracking*** for f in a stream of length m with failure probability η if at each time $t \in [m]$, $\mathcal{A}$ outputs an estimate $\tilde{f}_t$ such that

$$\mathbb{P}(\exists t \in [m] : |\tilde{f}^t - f(x^{(t)})| > \varepsilon f(x^{(t)})) < \eta.$$

We say that $\mathcal{A}$ provides ***weak tracking*** for f if

$$\mathbb{P}(\exists t \in [m] : |\tilde{f}^t - f(x^{(t)})| > \varepsilon \sup_{t' \in [m]} f(x^{(t')})) < \eta.$$

Note if f is monotonically increasing, then for insertion-only streams $\sup_{t' \in [m]} f(x^{(t')})$ is simply $f(x^{(m)})$.

The first non-trivial tracking result we are aware of which outperformed the median trick for insertion-only streaming was the RoughEstimator algorithm given in [40] for estimating the number of distinct elements in a stream. RoughEstimator provided a strong tracking guarantee for $f(x) = |\mathit{support}(x)|$ (the distinct elements problem) for constant ε, η, using the same space as what is what is required to answer only a single query. This strong tracking algorithm was used as a subroutine in the main *non-tracking* algorithm of that work for approximating the number of distinct elements in a data stream up to $1 + \varepsilon$.

For ℓ_p-estimation for $p \in (0, 2]$, without tracking, it is known that $\mathcal{O}(\varepsilon^{-2} \lg(1/\delta))$ words of memory is achievable to return a $(1 + \varepsilon)$-approximate value of $f(x) = \|x\|_p$ with failure probability δ [1, 30, 39][1]. This upper bound thus implies a strong tracking algorithm with space complexity $\mathcal{O}(\varepsilon^{-2} \lg m)$ for tracking failure probability $\eta = 1/3$, by setting $\delta < 1/(3m)$ and performing a union bound. The work [29] considered the strong tracking variant of ℓ_p-estimation in insertion-only streams for for any p in the more restricted interval $(1, 2]$. They showed that the same algorithms of [1, 30], unchanged, provide strong tracking with $\eta = 1/3$ with space $\mathcal{O}(\varepsilon^{-2}(\lg n + \lg\lg m + \lg(1/\varepsilon)))$ words[2]. This is an improvement over the standard median trick and union bound when the stream length is very long ($m > n^{\omega(1)}$) and ε is not too small ($\varepsilon > 1/m^{o(1)}$). They also showed that in an update model which allows deletions of items ("turnstile streaming"), any algorithm which only maintains a linear sketch Πx of x must use $\Omega(\lg m)$ words of memory for constant ε, showing that the median trick is optimal for this restricted class of algorithms.

A different algorithm was given in [10] for strong tracking for ℓ_2 using space $\mathcal{O}(\varepsilon^{-2}(\lg(1/\varepsilon) + \lg\lg m))$. It was then most recently shown in [9] that the AMS sketch itself of [1] (though with 8-wise independent hash functions instead of the original 4-wise independence proposed in [1]) provides strong tracking in space $\mathcal{O}(\varepsilon^{-2} \lg\lg m)$, and weak tracking in space $\mathcal{O}(1/\varepsilon^2)$. That is, the AMS sketch provides weak tracking without any asymptotic increase in space complexity over the requirement to correctly answer only a single query.

Despite the progress in upper bounds for tracking ℓ_2, the only non-trivial improvement for tracking ℓ_p is the $\mathcal{O}(\varepsilon^{-2}(\lg n + \lg\lg m + \lg(1/\varepsilon)))$ upper bound of [29]. Although this bound provides an improvement for very long streams (m super-polynomial in n), it does not provide any improvement over the standard median trick for the case most commonly studied case in the literature of m, n being polynomially related.

Our contribution

We show that Indyk's p-stable sketch [30] for $0 < p \leq 2$, derandomized using bounded independence as in [39], provides weak tracking while using $\mathcal{O}(\lg(1/\varepsilon)/\varepsilon^2)$ words of space. It also provides strong tracking using $\mathcal{O}(\varepsilon^{-2}(\lg\lg m + \lg(1/\varepsilon))$ words of space. Our bounds thus both improve the space complexity achieved in [29] for ℓ_p-tracking, and well as the

---

1 For constant δ and $p = 2$, [1] shows that space $\mathcal{O}(\varepsilon^{-2}(\lg n + \lg\lg m))$ *bits* is achievable in insertion-only streams.

2 For $p = 2$ their space is as written including the space required to store all hash functions, but for $1 < p < 2$ this space bound assumes that the storage of hash functions is for free.

range of p supported from $p \in (1, 2]$ to all $p \in (0, 2]$ (note for $p > 2$, it is known that any algorithm requires polynomial space even to obtain a 2-approximation for a single query, i.e. the non-tracking variant of the problem [4]).

2 Notation

We use $[n]$ for integer n to denote $\{1, \ldots, n\}$. We measure space in words unless stated otherwise, where a single word is at least $\lg(nm)$ bits. For $p \in (0, 2]$, we let $\mathcal{D}_p$ denote the symmetric p-stable distribution, scaled so that for $Z \sim \mathcal{D}_p$, $\mathbb{P}(|Z| > 1) = \frac{1}{2}$. The distribution $\mathcal{D}_p$ has the property that it is supported on the reals, and for any fixed vector $v \in \mathbb{R}^n$ and $Z_1, \ldots, Z_n, Z$ i.i.d. from $\mathcal{D}_p$, $\sum_{i=1}^n Z_i x_i$ is equal in distribution to $\|x\|_p \cdot Z$. See [45] for further reading on these distributions.

For two vectors $u, v \in \mathbb{R}^n$ we write $u \preceq v$ to denote coordinatewise comparison, i.e. $u \preceq v$ iff $\forall_i u_i \le v_i$. For a finite set S, we write $\#S$ to denote cardinality of this set.

3 Preliminaries

The following lemma is standard. A proof with explicit constants can be found in [43, Theorem 42].

▶ **Lemma 2.** *If $Z \sim \mathcal{D}_p$, then $\mathbb{P}(Z > \lambda) \le \frac{C_p}{\lambda^p}$ for some explicit constant C_p depending only on p.*

We also state some other results we will need.

▶ **Lemma 3** (Paley-Zygmund). *If $Z \ge 0$ is a random variable with finite variance, then*

$$\mathbb{P}(Z > \theta\, \mathbb{E}\, Z) \ge (1-\theta)^2 \frac{(\mathbb{E}\, Z)^2}{\mathbb{E}(Z^2)}.$$

▶ **Corollary 4.** *For fixed vector $v \in \mathbb{R}^n$, if $\sigma \in \{\pm 1\}^n$ is a vector of 4-wise independent random signs, then*

$$\mathbb{P}(\langle \sigma, v\rangle^2 \ge \frac{2}{3}\|v\|_2^2) \ge \frac{1}{27}\,.$$

Proof. This follows from $\mathbb{E}\langle \sigma, v\rangle^4 < 3(\mathbb{E}\langle \sigma, v\rangle^2)^2$ and the Paley-Zygmund inequality. ◀

▶ **Theorem 5** ([10, 9, Theorem 15]). *Let $v^{(1)}, v^{(2)}, \ldots v^{(m)} \in \mathbb{R}^n$, be a sequence of vectors such that $0 \preceq v^{(1)} \preceq v^{(2)} \preceq \ldots \preceq v^{(m)}$. Let $\sigma \in \{\pm 1\}^n$ be a vector of 4-wise independent random signs. Then*

$$\mathbb{P}\left(\sup_{i \le m} |\langle \sigma, v^{(i)}\rangle| > \lambda \|v^{(n)}\|_2\right) < \frac{C}{\lambda^2}$$

for some universal constant C.

▶ **Theorem 6.** *[39, 23] If $Z_i \sim \mathcal{D}_p$ for $i \in [n]$ are k-wise independent random variables, then for every vector $x \in \mathbb{R}^n$ and every pair $a, b \in \mathbb{R} \cup \{\pm\infty\}$ we have*

$$\mathbb{P}(\langle Z, x\rangle \in (a, b)) = \mathbb{P}(\|x\|_p Z_1 \in (a, b)) \pm \mathcal{O}(k^{-1/p})\,.$$

▶ **Theorem 7.** *[5, Lemma 2.3] Let $X_1, \ldots X_n \in \{0, 1\}$ be a sequence of k-wise independent random variables, and let $\mu = \sum_{i=1}^n \mathbb{E}\, X_i$. Then*

$$\forall \lambda > 0,\ \mathbb{P}(\sum_{i=1}^n X_i \ge (1+\lambda)\mu) \le \exp(-\Omega(\min\{\lambda, \lambda^2\}\mu)) + \exp(-\Omega(k))\,.$$

4 Overview of approach

Indyk's p-stable sketch picks a random matrix $\Pi \in \mathbb{R}^{d\times n}$ such that each entry is drawn according to the distribution $\mathcal{D}_p$. It then maintains the sketch $\Pi x^{(t)}$ of the current frequency vector. This sketch can be easily updated as the frequency vector changes, i.e. after observing an index $a_j \in [n]$ we update the sketch by $\Pi x^{(t+1)} := \Pi x^{(t)} + \Pi e_{a_j}$. An $\|x\|_p$-estimate query is answered by returning the median of $|\Pi x^{(i)}|_j$ over $j \in [d]$. Since storing Π in memory explicitly is prohibitively expensive, we generate it so that the entries in each row are k-wise independent for $k = \mathcal{O}(1/\varepsilon^p)$ (as done in [39]), and the d seeds used to generate the rows of Π are $\mathcal{O}(\lg(1/(\varepsilon\delta)))$-wise independent. We also work with discretized p-stable random variables to take bounded memory. All together, the bounded independence and discretization, also performed in [39], allow us to store Π using low memory.

We then show that instantiating Indyk's algorithm with $d = \mathcal{O}(\varepsilon^{-2}\lg(1/(\varepsilon\delta)))$ provides the weak tracking guarantee with failure probability δ. The analysis of the correctness of this algorithm is as follows. Let π_i denote the ith row of Π. We first show a result resembling the Doob's martingale inequality – namely, in Section 5 we show that for a fixed i, if we look at the evolution of $\langle \pi_i, x^{(t)}\rangle$ as t increases, the largest attained value $(\sup_{t\le m}\langle \pi_i, x^{(t)}\rangle)$ is with good probability not much larger than the median of the distribution $|\langle \pi_i, x^{(m)}\rangle|$, which is the typical magnitude of the counter at the end of the stream. This fact resembles similar facts shown in [10, 9] for when the π_i have independent Rademachers as entries, though our situation is complicated by the fact that p-stable random variables have much heavier tails.

We then, discussed in Section 5.1, show how the previous paragraph implies a weak tracking algorithm with $d = \mathcal{O}(\varepsilon^{-2}\lg(1/(\varepsilon\delta)))$: we split the sequence of updates into $poly(1/\varepsilon)$ intervals such that the ℓ_p-norm of the frequency vector of updates in each of those intervals, i.e. $\|x^{(t+1)} - x^{(t)}\|_p$, is of the order $\varepsilon^{\Theta(1)}\|x^{(m)}\|_p$. We then union bound over the $poly(1/\varepsilon)$ intervals to argue that the algorithm's estimate is good at each of the interval endpoints. This is the source of the extra factor of $\lg(1/\varepsilon)$ in our space bound: to obtain $\varepsilon^{-\Omega(1)}$ failure probability to union bound over these intervals. On the other hand, within each of the intervals most of the counters do not change too rapidly by the argument developed in Section 5.

Finally, in Section 5.2 we show how given an algorithm satisfying a weak tracking guarantee, one can use it to get a strong-tracking algorithm with slightly larger space complexity. This argument was already present in [9]. One first identifies q points in the input stream at which the ℓ_p norm roughly doubles when compared to the previously marked point. There are only $\mathcal{O}(\lg m)$ such intervals. It is then enough to ensure that our algorithm satisfies weak tracking for all those $\mathcal{O}(\lg m)$ prefixes simultaneously, in order to deduce that the algorithm in fact satisfies strong tracking. This is done by union bound over $\mathcal{O}(\lg m)$ bad events (as opposed to standard union bound over $\mathcal{O}(m)$ bad events), which introduces an extra $\lg\lg m$ factor in the space complexity as when compared to weak tracking.

5 Analysis

We first show two lemmas that play a crucial role in our weak tracking analysis.

► **Lemma 8.** *Let $x \in \mathbb{R}^n$ be a fixed vector, and $Z \in \mathbb{R}^n$ be a random vector with k-wise independent entries drawn according to $\mathcal{D}_p$. Then*

$$\mathbb{P}(\sum_{i=1}^{n} x_i^2 Z_i^2 \ge \lambda^2\|x\|_p^2) \le \frac{C}{\lambda^p} + \mathcal{O}(k^{-1/p})$$

for some universal constant C.

Proof. Let E_0 be the event $\sum_{i=1}^n x_i^2 Z_i^2 \geq \lambda^2 \|x\|_p^2$. Note that E_0 depends only on $|Z_i|$, and does not depend on the signs of the Z_i. We write $Z_i = |Z_i|\sigma_i$, where σ_i are k-wise independent random signs. Conditioning on $|Z_i|$,

$$\mathop{\mathbb{E}}_{\sigma}\left(\left(\sum_{i=1}^n x_i|Z_i|\sigma_i\right)^2 \middle| |Z_1|, \dots |Z_n|\right) = \sum_{i=1}^n x_i^2 Z_i^2$$

and therefore for any $|Z_1|, \dots, |Z_m|$ for which E_0 holds, by Corollary 4

$$\begin{aligned}
&\mathop{\mathbb{P}}_{\sigma}\left(\left(\sum_{i=1}^n x_i|Z_i|\sigma_i\right)^2 \geq \frac{2}{3}\lambda^2\|x\|_p^2 \middle| |Z_1|, \dots, |Z_m|\right) \\
&\quad \geq \mathop{\mathbb{P}}_{\sigma}\left(\left(\sum_{i=1}^n x_i|Z_i|\sigma_i\right)^2 \geq \frac{2}{3}\sum_{i=1}^n x_i^2 Z_i^2 \middle| |Z_1|, \dots, |Z_m|\right) \\
&\quad \geq \frac{1}{27}
\end{aligned}$$

and thus

$$\mathop{\mathbb{P}}_{\sigma}\left(\left(\sum_{i=1}^n x_i|Z_i|\sigma_i\right)^2 \geq \frac{2}{3}\lambda^2\|x\|_p^2 \middle| |Z_1|, \dots |Z_n|\right) \geq \frac{\mathbf{1}_{E_0}}{27},$$

where $\mathbf{1}_{E_0}$ is an indicator random variable for event E_0. Integrating over $|Z_i|$,

$$\mathop{\mathbb{P}}_{\sigma,Z}\left(\left(\sum_{i=1}^n x_i|Z_i|\sigma_i\right)^2 \geq \frac{2}{3}\lambda^2\|x\|_p^2\right) \geq \frac{1}{27}\mathop{\mathbb{P}}_{Z}(E_0). \tag{1}$$

On the other hand $|Z_i|\sigma_i$ has the same distribution as Z_i, and moreover

$$\begin{aligned}
\mathop{\mathbb{P}}_{Z}\left(\left(\sum_{i=1}^n x_i Z_i\right)^2 \geq \frac{2}{3}\lambda^2\|x\|_p^2\right) &= \mathop{\mathbb{P}}_{Z}\left(|\langle x, Z\rangle| \geq \sqrt{\frac{2}{3}}\lambda\|v\|_p\right) \\
&\leq \mathop{\mathbb{P}}_{Z}\left(\|x\|_p\tilde{Z} \geq \sqrt{\frac{2}{3}}\lambda\|x\|_p\right) + \mathcal{O}(k^{-1/p}) \\
&\leq \frac{C}{\lambda^p} + \mathcal{O}(k^{-1/p})
\end{aligned} \tag{2}$$

where $\tilde{Z} \sim \mathcal{D}_p$. The inequalities are obtained via Theorem 6 and Lemma 2. Combining (1), (2) yields

$$\mathop{\mathbb{P}}_{Z}(E_0) \leq \frac{27C}{\lambda^p} + \mathcal{O}(k^{-1/p}).$$ ◀

▶ **Lemma 9.** *Let $x^{(1)}, x^{(2)}, \dots x^{(m)} \in \mathbb{R}^n$ satisfy $0 \preceq x^{(1)} \preceq x^{(2)} \preceq \dots \preceq x^{(m)}$. Let $Z \in \mathbb{R}^n$ have k-wise independent entries marginally distributed according to $\mathcal{D}_p$. Then for some C_p depending only on p,*

$$\mathbb{P}\left(\sup_{k \leq m} |\langle Z, x^{(k)}\rangle| \geq \lambda\|x^{(m)}\|_p\right) \leq C_p\left(\frac{1}{\lambda^{2p/(2+p)}} + k^{-1/p}\right).$$

Proof. Observe that for any β we have

$$\mathbb{P}\left(\sup_{k\le m}|\langle Z, x^{(k)}\rangle| \ge \lambda\|x^{(m)}\|_p\right) \le \mathbb{P}\left(\sum_{i=1}^n Z_i^2 (x^{(m)})_i^2 \ge \beta^2\|x^{(m)}\|_p^2\right) + \mathbb{P}\left(\sup_{k\le m}|\langle Z, x^{(k)}\rangle| \ge \lambda\|x^{(m)}\|_p \middle| \sum_{i=1}^n Z_i^2(x^{(m)})_i^2 < \beta^2\|x^{(m)}\|_p^2\right).$$

Lemma 8 directly implies that

$$\mathbb{P}\left(\sum_{i=1}^n Z_i^2(x^{(m)})_i^2 \ge \beta^2\|x^{(m)}\|_p^2\right) \le \frac{C}{\beta^p} + \frac{C}{k^{1/p}}. \tag{3}$$

On the other hand we can write $Z_i = |Z_i|\sigma_i$, where σ_i are k-wise independent Rademacher random variables, independent from $|Z_i|$. Let us define $w^{(k)} \in \mathbb{R}^n$ for $k \in [m]$ to be the vector with coordinates $(w^{(k)})_i := (x^{(k)})_i|Z_i|$, so that $\langle x^{(k)}, Z\rangle = \langle w^{(k)}, \sigma\rangle$, and in particular

$$\sup_{k\le m}\left|\langle Z, x^{(i)}\rangle\right| = \sup_{k\le m}\left|\langle \sigma, w^{(i)}\rangle\right|.$$

Now, if we condition on $|Z_1|, \dots |Z_n|$, then the sequence $w^{(1)}, \dots w^{(k)}$ of vectors satisfies the assumptions of Theorem 5, and we can conclude that

$$\mathbb{P}\left(\sup_{k\le m}\left|\langle\sigma, w^{(k)}\rangle\right| > \frac{\lambda}{\beta}\|w^{(m)}\|_2\right) \le \frac{C\beta^2}{\lambda^2}.$$

Moreover if $|Z_i|$ are such that $\sum_{i=1}^n Z_i^2(x^{(m)})_i^2 \le \beta^2\|x^{(m)}\|_p^2$, or equivalently $\|w^{(m)}\|_2^2 \le \beta^2\|x^{(m)}\|_p^2$, we have

$$\mathbb{P}\left(\sup_{k\le m}\left|\langle\sigma, w^{(k)}\rangle\right| > \lambda\|x^{(m)}\|_p\right) \le \frac{C\beta^2}{\lambda^2},$$

which implies

$$\mathbb{P}\left(\sup_{k\le m}|\langle Z, x^{(k)}\rangle| \ge \lambda\|x^{(m)}\|_p \middle| \sum_{i=1}^n (Z_i x_i^{(m)})^2 < \beta\|x^{(m)}\|_p^2\right) \le \frac{C\beta^2}{\lambda^2}.$$

This together with Equation (3) yields

$$\mathbb{P}\left(\sup_{k\le m}|\langle Z, x^{(k)}\rangle| \ge \lambda\|x^{(m)}\|_p\right) \le \frac{1}{\beta^p} + \frac{C\beta^2}{\lambda^2} + \frac{C}{k^{1/p}}.$$

We can take $\beta := \Theta(\lambda^{\frac{2}{2+p}})$, to have $\frac{1}{\beta^p} + \frac{C\beta^2}{\lambda^2} = \mathcal{O}(\lambda^{-\frac{2p}{2+p}})$. ◀

5.1 Weak tracking of $\|x\|_p$

In this section we upper bound the number of rows needed in Indyk's p-stable sketch with boundedly independent entries to achieve weak tracking.

▶ **Lemma 10.** *Let $x^{(1)}, \dots x^{(m)} \in \mathbb{R}^n$ be any sequence satisfying $0 \preceq x^{(1)} \preceq x^{(2)} \preceq \dots \preceq x^{(m)}$. Take $\Pi \in \mathbb{R}^{d\times n}$ to be a random matrix with entries drawn according to $\mathcal{D}_p$, and such that the rows are r-wise independent, and all entries within a row are s-wise independent.*

For every $k \in [m]$, define s_k to be median $\left(|(\Pi x^{(k)})_1|, \dots, |(\Pi x^{(k)})_d|\right)$. If $d = \Omega(\varepsilon^{-2}(\lg\frac{1}{\varepsilon} + \lg\frac{1}{\delta}))$, $r = \Omega(\lg\frac{1}{\varepsilon} + \lg\frac{1}{\delta})$ and $s = \Omega(\varepsilon^{-p})$, then with probability at least $1-\delta$ we have

$$\forall k \in [m],\ \|x^{(k)}\|_p - \varepsilon\|x^{(m)}\|_p \le s_k \le \|x^{(k)}\|_p + \varepsilon\|x^{(m)}\|_p.$$

Proof. Consider a sequence of indices $1 < t_1 < t_2 < \ldots < t_{q+1} = m$, constructed inductively in the following way. We take t_1 to be the smallest index with $\|x^{(t_1)}\|_p \geq \varepsilon^4 \|x^{(m)}\|_p$. Given t_k, we take t_{k+1} to be the smallest index such that $\|x^{(t_{k+1})} - x^{(t_k)}\|_p \geq \varepsilon^4 \|x^{(m)}\|_p$ if there exists one, and $t_{k+1} = m$ otherwise.

Observe $q \leq \varepsilon^{-8}$. Indeed, for $p \geq 1$ we have

$$\|x^{(m)}\|_p^p = \|x^{(t_1)} + \sum_{1 \leq i < q} (x^{(t_{i+1})} - x^{(t_i)})\|_p^p \geq \|x^{(t_1)}\|_p^p + \sum_{1 \leq i < q} \|x^{(t_{i+1})} - x^{(t_i)}\|_p^p \geq q\varepsilon^{4p} \|x^{(m)}\|_p^p$$

where the inequality $\|x^{(t_1)} + \sum_{i \geq 1} (x^{(t_{i+1})} - x^{(t_i)})\|_p^p \geq \|x^{(t_1)}\|_p^p + \sum_{1 \leq i < q} \|x^{(t_{i+1})} - x^{(t_i)}\|_p^p$ holds because all vectors $x^{(1)}$ and $x^{(t_{i+1})} - x^{(t_i)}$ for every i have non-negative entries – we can consider each coordinate separately, and use the fact that for $p \geq 1$ and nonnegative numbers a_i we have $(\sum a_i)^p \geq \sum a_i^p$ – or equivalently, $\|a\|_1^p \geq \|a\|_p^p$. After rearranging this yields $q \leq \varepsilon^{-4p}$.

Similarly, for $p \leq 1$, we have that for non-negative numbers a_i, $(\sum_{i \leq q} a_i)^p \geq q^{p-1} \sum i \leq q a_i^p$ (this is true because for fixed $\sum a_i$, the sum $\sum a_i^p$ is maximized when all a_i are equal), and therefore

$$\begin{aligned} \|x^{(m)}\|_p^p = \|x^{(t_1)} + \sum_{1 \leq i < q} (x^{(t_{i+1})} - x^{(t_i)})\|_p^p &\geq q^{p-1} \left(\|x^{(t_1)}\|_p^p + \sum_{1 \leq i < q} \|x^{(t_{i+1})} - x^{(t_i)}\|_p^p \right) \\ &\geq q^p \varepsilon^{4p} \|x^{(m)}\|_p^p \end{aligned}$$

which implies $q \leq \varepsilon^{-4}$.

For $j \in [m]$, let us define

$$\begin{aligned} l_j &:= \#\{i : |\langle \pi_i, x^{(j)} \rangle| < (1 - \varepsilon)\|x^{(j)}\|_p\} \\ u_j &:= \#\{i : |\langle \pi_i, x^{(j)} \rangle| > (1 + \varepsilon)\|x^{(j)}\|_p\}. \end{aligned}$$

Let $\tilde{\pi}_i$ be a vector of i.i.d. random variables drawn according to $\mathcal{D}_p$. We know that $\langle \tilde{\pi}_i, x^{(j)} \rangle \sim \|x^{(j)}\|_p \mathcal{D}_p$. Hence $\mathbb{P}(|\langle \tilde{\pi}_i, x^{(j)} \rangle| > \|x^{(j)}\|_p) = \frac{1}{2}$, and $\mathbb{P}(|\langle \tilde{\pi}_i, x^{(j)} \rangle| > (1 + \varepsilon)\|x^{(j)}\|_p) \leq \frac{1}{2} - 2C\varepsilon$ for some universal constant C. Similarly $\mathbb{P}(|\langle \tilde{\pi}_i, x^{(j)} \rangle| < (1 - \varepsilon)\|x^{(j)}\|_p) \leq \frac{1}{2} - 2C\varepsilon$.

Entries of π_i are s-wise independent, for $s \geq C_2 \varepsilon^{-p}$ with some large constant C_2 depending on C. Thus by Theorem 6, $\mathbb{P}(|\langle \pi_i, x^{(j)} \rangle| < (1 - \varepsilon)\|x^{(j)}\|_p) \leq \mathbb{P}(|\langle \tilde{\pi}_i, x^{(j)} \rangle| < (1 - \varepsilon)\|x^{(j)}\|_p) + C\varepsilon \leq \frac{1}{2} - C\varepsilon$, and analogously for $\mathbb{P}(|\langle \pi_i, x^{(j)} \rangle| > (1 + \varepsilon)\|x^{(j)}\|_p) < \frac{1}{2} - C\varepsilon$.

Hence

$$\begin{aligned} \mathbb{E}\, l_j &\leq d\left(\frac{1}{2} - C\varepsilon\right) \\ \mathbb{E}\, u_j &\leq d\left(\frac{1}{2} - C\varepsilon\right). \end{aligned}$$

For $j \in [q]$, let S_j be the event

$$\left\{ l_{t_j} \leq \frac{d}{2} - \frac{Cd}{2}\varepsilon \right\} \wedge \left\{ u_{t_j} \leq \frac{d}{2} - \frac{Cd}{2}\varepsilon \right\}$$

Note that for fixed j and varying i, indicator random variables for the events "$|\langle \pi_i, x^{(j)} \rangle| < (1 - \varepsilon)\|x^{(j)}\|_p$" are r-wise independent. Thus by Theorem 7, $\mathbb{P}(S_j) \geq 1 - C' \exp(-\Omega(d\varepsilon^2)) - \exp(-\Omega(r))$. Taking $d = \Omega(\varepsilon^{-2}(\lg \frac{1}{\varepsilon} + \lg \frac{1}{\delta}))$ and $r = \Omega(\lg \frac{1}{\varepsilon\delta})$ we obtain $\mathbb{P}(S_j) \geq 1 - \frac{\delta\varepsilon^8}{2}$, and hence by a union bound all S_j hold simultaneously except with probability at most $\frac{\delta}{2}$ since the number of events S_j is $q \le \varepsilon^{-8}$.

For $i \in [d]$ and $j \in [q]$, let $E_{i,j}$ be the event

$$\exists s \in [t_j, t_{j+1} - 1],\ |\langle x^{(s)} - x^{(t_j)}, \pi_i\rangle| > \varepsilon \|x^{(m)}\|_p.$$

By construction of the sequence t_j, all $x^{(s)} - x^{(t_j)}$ above have ℓ_p norm at most $\varepsilon^4\|x^{(m)}\|_p$, we can invoke Lemma 9 to deduce that $\mathbb{P}(E_{ij}) \le C_3 \left(\frac{\varepsilon^4}{\varepsilon}\right)^{2/3} + C_3 s^{-1/p}$. Again if we pick $s \ge C_4\varepsilon^{-p}$ for sufficiently large C_4 and small enough ε we have $\mathbb{P}(E_{ij}) \le \frac{C}{4}\varepsilon$. Therefore for any fixed j, we have

$$\mathbb{E}\sum_{i=1}^{d} \mathbf{1}_{E_{ij}} \le \frac{C}{4} d\varepsilon$$

And finally again by Theorem 7, for each j

$$\mathbb{P}(\sum_{i=1}^{d} \mathbf{1}_{E_{ij}} \ge \frac{C}{2} d\varepsilon) \lesssim \exp(-C' d\varepsilon) + \exp(-C' r)$$

We have $d \ge C_3\varepsilon^{-2} \lg \frac{1}{\delta\varepsilon}$, and $q \le \varepsilon^{-8}$, hence for sufficiently small ε, we have $\exp(-C'd\varepsilon) \le \frac{\delta}{2q}$. On the other hand if $r = \Omega(\lg \frac{1}{\delta\varepsilon})$ is sufficiently large, we have $\exp(-C'r) \le \frac{\delta}{2q}$. We invoke the union bound over all j to deduce that with probability at least $1 - \frac{\delta}{2}$ the following event V holds:

$$\forall j,\ \sum_{i=1}^{d} \mathbf{1}_{E_{ij}} \le \frac{C}{2} d\varepsilon.$$

We know that with probability at least $1 - \delta$ simultaneously V and all the events S_j hold. We will show now that, when these events all hold, then $\forall k$ $\|x^{(k)}\|_p - K\varepsilon\|x^{(m)}\|_p \le s_k \le \|x^{(k)}\|_p + K\varepsilon\|x^{(m)}\|_p$ for some universal constant K. Indeed, consider some k, and let us assume that $t_j \le k \le t_{j+1}$. With event S_j satisfied, we know that $\#\{i : |\langle \pi_i, x^{(t_j)}\rangle| \le \|x^{(t_j)}\|_p + \varepsilon\|x^{(m)}\|_p\} \ge d\left(\frac{1}{2} + \frac{C\varepsilon}{2}\right)$, and with event V satisfied, we know that for all but $\frac{C\varepsilon}{2}d$ of indices i we have $|\langle \pi_i, x^{(k)} - x^{(t_j)}\rangle| \le \varepsilon\|x^{(m)}\|$.

By the triangle inequality $|\langle \pi_i, x^{(k)}\rangle| \le |\langle \pi_i, x^{(t_j)}\rangle| + |\langle \pi_i, x^{(k)} - x^{(t_j)}\rangle|$, yielding

$$\#\{i : |\langle \pi_i, v_k\rangle| \le \|v_{t_j}\|_p + 2\varepsilon\|v_m\|_p\} \ge \frac{d}{2}.$$

With similar reasoning we can deduce that

$$\#\{i : |\langle \pi_i, x^{(k)}\rangle| \ge \|x^{(t_j)}\|_p - 2\varepsilon\|x^{(m)}\|_p\} \ge \frac{d}{2},$$

which implies the median of $|\langle \pi_i, x^{(k)}\rangle|$ over $i \in [d]$ is in the range $\|x^{(t_j)}\|_p \pm 2\varepsilon\|x^{(m)}\|_p$. In other words

$$\|x^{(t_j)}\|_p - 2\varepsilon\|x^{(m)}\|_p \le s_k \le \|x^{(t_i)}\|_p + 2\varepsilon\|x^{(m)}\|_p.$$

Finally we also have $\left|\|x^{(k)}\|_p - \|x^{(t_j)}\|_p\right| \le \varepsilon\|x^{(m)}\|_p$ by construction of the sequence $\{t_j\}_{j=1}^q$, so the claim follows up to rescaling ε by a constant factor. ◀

▶ **Lemma 11.** *The above algorithm can be implemented using $\mathcal{O}(\varepsilon^{-2}\lg(1/(\varepsilon\delta))\lg m)$ bits of memory to store fixed precision approximations of all counters $(\Pi x^{(k)})_i$, and $\mathcal{O}(\varepsilon^{-p}\lg(1/(\varepsilon\delta))\lg(nm))$ bits to store Π.*

Proof. Consider a sketch matrix Π as in Lemma 10 – i.e. $\Pi \in \mathbb{R}^{d\times n}$ with random $\mathcal{D}_p$ entries, such that all rows are r-wise independent and all entries within a row are s-wise independent. Moreover let us pick some $\gamma = \Theta(\varepsilon m^{-1})$ and consider discretization $\tilde{\Pi}$ of Π, namely each entry $\tilde{\Pi}_{ij}$ is equal to Π_{ij} rounded to the nearest integer multiple of γ. The analysis identical to the one in [39, A.6] shows that this discretization have no significant effect on the accuracy of the algorithm, and moreover that one can sample from a nearby distribution using only $\tau = \mathcal{O}(\lg m\varepsilon^{-1})$ uniformly random bits. Therefore we can store such a matrix succinctly using $\mathcal{O}\left(rs(\lg n + \tau) + r\lg d\right)$ bits of memory, by storing a seed for a random r-wise independent hash function $h : [d] \to \{0,1\}^{\mathcal{O}(s(\lg n+\tau))}$ and interpreting each $h(i)$ as a seed for an s-wise independent hash function describing the i-th row of $\tilde{\Pi}$ [48, Corollary 3.34]. Hence the total space complexity of storing the sketch matrix $\tilde{\Pi}$ in a succinct manner is $\mathcal{O}\left(\frac{\lg\delta^{-1}+\lg\varepsilon^{-1}}{\varepsilon^p}(\lg n+\lg m)\right)$ bits.

Additionally we have to store the sketch of the current frequency vector itself, i.e. for all $i \in [d]$ we need to store $\langle\tilde{\pi}_i, x^{(k)}\rangle$; for every such counter we need $\mathcal{O}(\lg m\varepsilon^{-1}) = \mathcal{O}(\lg m)$ bits, and there are $d = \mathcal{O}\left(\frac{\lg\varepsilon^{-1}+\lg\delta-1}{\varepsilon^{-2}}\right)$ counters. ◀

We thus have the following main theorem of this section.

▶ **Theorem 12.** *For any $p \in (0,2]$ there is an insertion-only streaming algorithm that provides the weak tracking guarantees for $f(x) = \|x\|_p$ with probability $1-\delta$ using at most $\mathcal{O}\left(\frac{\lg m+\lg n}{\varepsilon^2}(\lg\varepsilon^{-1}+\lg\delta^{-1})\right)$ bits of memory.*

5.2 Strong tracking of $\|x\|_p$

In this section we discuss achieving a strong tracking guarantee. The same argument for ℓ_2-tracking appeared in [9]. The reduction is in fact general, and shows that for any monotone function f the strong tracking problem for f reduces to the weak tracking version of the same problem with smaller failure probability.

▶ **Lemma 13.** *Let $f : \mathbb{R}^n \to \mathbb{R}_+$ be any monotone functon of $\mathbb{R}^n$ (i.e. $x \preceq y \implies f(x) \le f(y)$), such that $\min_i f(e_i) = 1$ (where e_i are standard basis vectors). Let $\mathcal{A}$ be an insertion-only streaming algorithm satisfying weak tracking for any sequence of updates with probability $1-\delta$ and accuracy ε. Then for a sequence of frequency vectors $0 \preceq x^{(1)} \preceq \ldots \preceq x^{(m)}$ algorithm $\mathcal{A}$ satisfies strong tracking with probability $1-\delta\lg f(x^{(m)})$ and accuracy 2ε.*

Proof. Define $t_1 < t_2 < \cdots < t_q$ so that t_i is the smallest index in $[m]$ larger than t_{i-1} with $f(x^{(t_i)}) \ge 2^i$ (if no such index exists, define $q = i$ and $t_q = m$). Note that $q \le \lg f(x^{(m)})$.

The algorithm will fail with probability at most δ to satisfy the conclusion of Theorem 12 for a particular sequence of vectors $x^{(1)}, x^{(2)}, \ldots x^{(t_j)}$. That is, for every j, with probability $1-\delta$, we have that

$$\forall i \le t_j,\ f(x^{(i)}) - \varepsilon f(x^{(t_j)}) \le \tilde{f}^i \le f(x^{(i)}) + \varepsilon f(x^{(t_j)}),$$

where $\tilde{f}^t$ is the estimate output by the algorithm at time t.

We can union bound over all $j \in [q]$ to deduce that except with probability $q\delta \le \delta\lg f(x^{(m)})$,

$$\forall i \le t_j,\ f(x^{(i)}) - \varepsilon f(x^{(t_j)}) \le \tilde{f}^i \le f(x^{(i)}) + \varepsilon f(x^{(t_j)}).$$

By construction of the sequence of t_j, we know that for every i, if we take t_j to be smallest such that $i \le t_j$, then $f(x^{(t_j)}) \le 2f(x^{(i)})$, and the claim follows. ◀

▶ **Theorem 14.** *For any $p \in (0,2]$ there is an insertion-only streaming algorithm that provides strong tracking guarantees for estimating the ℓ_p-norm of the frequency vector with probability $1-\delta$ and multiplicative error $1+\varepsilon$, with space usage in bits bounded by $\mathcal{O}\left(\frac{\lg m + \lg n}{\varepsilon^2}(\lg \varepsilon^{-1} + \lg \delta^{-1} + \lg\lg m)\right)$.*

Proof. This follows from Lemma 11 and Lemma 13 by observing that after a sequence of m insertions, the ℓ_p norm of the frequency vector is bounded by m^2, i.e. $\lg(\|x^{(m)}\|_p) = \mathcal{O}(\lg m)$. ◀

References

1. Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *J. Comput. Syst. Sci.*, 58(1):137–147, 1999.
2. Alexandr Andoni, Robert Krauthgamer, and Krzysztof Onak. Streaming algorithms via precision sampling. In *Proc. of the 52nd IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, pages 363–372, 2011.
3. Shivnath Babu and Jennifer Widom. Continuous queries over data streams. *SIGMOD Record*, 30(3):109–120, 2001.
4. Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. *Journal of Computer and System Sciences*, 68(4):702–732, 2004.
5. Mihir Bellare and John Rompel. Randomness-efficient oblivious sampling. In *Proc. of the 35th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 276–287, 1994.
6. Lakshminath Bhuvanagiri and Sumit Ganguly. Estimating entropy over data streams. In *Proc. of the 14th Annual European Symposium on Algorithms (ESA)*, pages 148–159, 2006.
7. Lakshminath Bhuvanagiri, Sumit Ganguly, Deepanjan Kesh, and Chandan Saha. Simpler algorithm for estimating frequency moments of data streams. In *Proc. of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 708–713, 2006.
8. Vladimir Braverman and Stephen R. Chestnut. Universal sketches for the frequency negative moments and other decreasing streaming sums. In *Proc. of the 18th International Workshop on Approximation, Randomization, and Combinatorial Optimization: Algorithms and Techniques (APPROX)*, pages 591–605, 2015.
9. Vladimir Braverman, Stephen R. Chestnut, Nikita Ivkin, Jelani Nelson, Zhengyu Wang, and David P. Woodruff. BPTree: an ℓ_2 heavy hitters algorithm using constant memory. In *Proc. of the 36th SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*, 2017.
10. Vladimir Braverman, Stephen R. Chestnut, Nikita Ivkin, and David P. Woodruff. Beating countsketch for heavy hitters in insertion streams. In *Proc. of the 48th Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pages 740–753, 2016.
11. Vladimir Braverman, Stephen R. Chestnut, Robert Krauthgamer, and Lin F. Yang. Streaming symmetric norms via measure concentration. In *Proc. of the 49th Annual ACM Symposium on Theory of Computing (STOC), to appear*, 2017.
12. Vladimir Braverman, Stephen R. Chestnut, David P. Woodruff, and Lin F. Yang. Streaming space complexity of nearly all functions of one variable on frequency vectors. In *Proc. of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS)*, pages 261–276, 2016.
13. Vladimir Braverman, Jonathan Katzman, Charles Seidell, and Gregory Vorsanger. An optimal algorithm for large frequency moments using $o(n^{1-2/k})$ bits. In *Proc. of the 17th International Workshop on Approximation, Randomization, and Combinatorial Optimization: Algorithms and Techniques (APPROX)*, pages 531–544, 2014.

14 Vladimir Braverman and Rafail Ostrovsky. Measuring independence of datasets. In *Proc. of the 42^{nd} ACM Symposium on Theory of Computing (STOC)*, pages 271–280, 2010.

15 Vladimir Braverman and Rafail Ostrovsky. Zero-one frequency laws. In *Proc. of the 42^{nd} ACM Symposium on Theory of Computing (STOC)*, pages 281–290, 2010.

16 Vladimir Braverman and Rafail Ostrovsky. Approximating large frequency moments with pick-and-drop sampling. In *Proc. of the 16^{th} International Workshop on Approximation, Randomization, and Combinatorial Optimization: Algorithms and Techniques (APPROX)*, pages 42–57, 2013.

17 Vladimir Braverman, Rafail Ostrovsky, and Alan Roytman. Zero-one laws for sliding windows and universal sketches. In *Proc. of the 18^{th} International Workshop on Approximation, Randomization, and Combinatorial Optimization: Algorithms and Techniques (APPROX)*, pages 573–590, 2015.

18 Donald Carney, Ugur Çetintemel, Mitch Cherniack, Christian Convey, Sangdon Lee, Greg Seidman, Michael Stonebraker, Nesime Tatbul, and Stanley B. Zdonik. Monitoring streams - A new class of data management applications. In *Proc. of the 28^{th} International Conference on Very Large Data Bases (VLDB)*, pages 215–226, 2002.

19 Amit Chakrabarti, Khanh Do Ba, and S. Muthukrishnan. Estimating entropy and entropy norm on data streams. *Internet Mathematics*, 3(1):63–78, 2006.

20 Amit Chakrabarti, Graham Cormode, and Andrew McGregor. A near-optimal algorithm for estimating the entropy of a stream. *ACM Trans. Algorithms*, 6(3):51:1–51:21, 2010.

21 Amit Chakrabarti, Subhash Khot, and Xiaodong Sun. Near-optimal lower bounds on the multi-party communication complexity of set disjointness. In *Proc. of the 18^{th} Annual IEEE Conference on Computational Complexity (CCC)*, pages 107–117, 2003.

22 Graham Cormode and S. Muthukrishnan. Space efficient mining of multigraph streams. In *Proc. of the 24^{th} ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS)*, pages 271–282, 2005.

23 Ilias Diakonikolas, Daniel M. Kane, and Jelani Nelson. Bounded independence fools degree-2 threshold functions. In *Proc. of the 51^{st} Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 11–20, 2010.

24 Philippe Flajolet and G. Nigel Martin. Probabilistic counting algorithms for data base applications. *J. Comput. Syst. Sci.*, 31(2):182–209, 1985.

25 Sumit Ganguly. Taylor polynomial estimator for estimating frequency moments. In *Proc. of the 42^{nd} International Colloquium on Automata, Languages, and Programming (ICALP)*, pages 542–553, 2015.

26 André Gronemeier. Asymptotically optimal lower bounds on the nih-multi-party information complexity of the and-function and disjointness. In *Proc. of the 26^{th} International Symposium on Theoretical Aspects of Computer Science (STACS)*, pages 505–516, 2009.

27 Sudipto Guha, Piotr Indyk, and Andrew McGregor. Sketching information divergences. *Machine Learning*, 72(1-2):5–19, 2008.

28 Nicholas J. A. Harvey, Jelani Nelson, and Krzysztof Onak. Sketching and streaming entropy via approximation theory. In *Proc. of the 49^{th} Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 489–498, 2008.

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

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

31 Piotr Indyk and Andrew McGregor. Declaring independence via the sketching of sketches. In *Proc. of the 19^{th} Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 737–745, 2008.

32 Piotr Indyk and David P. Woodruff. Tight lower bounds for the distinct elements problem. In *Proc. of the 44th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 283–, 2003.

33 Piotr Indyk and David P. Woodruff. Optimal approximations of the frequency moments of data streams. In *Proc. of the 37th Annual ACM Symposium on Theory of Computing (STOC)*, pages 202–208, 2005.

34 T. S. Jayram. Hellinger strikes back: A note on the multi-party information complexity of AND. In *Proc. of the 12th International Workshop on Randomization and Approximation Techniques (RANDOM)*, pages 562–573, 2009.

35 T. S. Jayram. On the information complexity of cascaded norms with small domains. In *IEEE Information Theory Workshop (ITW)*, pages 1–5, 2013.

36 T. S. Jayram and David P. Woodruff. The data stream space complexity of cascaded norms. In *Proc. of the 50th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 765–774, 2009.

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

38 Daniel M. Kane, Jelani Nelson, Ely Porat, and David P. Woodruff. Fast moment estimation in data streams in optimal space. In *Proc. of the 43rd ACM Symposium on Theory of Computing (STOC)*, pages 745–754, 2011.

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

40 Daniel M. Kane, Jelani Nelson, and David P. Woodruff. An optimal algorithm for the distinct elements problem. In *Proc. of the 29th SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*, pages 41–52, 2010.

41 Ping Li. Estimators and tail bounds for dimension reduction in ℓ_α $(0 < \alpha \leq 2)$ using stable random projections. In *Proc. of the 19th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 10–19, 2008.

42 Ping Li. Compressed counting. In *Proc. of the 20th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 412–421, 2009.

43 Jelani Nelson. *Sketching and streaming high-dimensional vectors.* PhD thesis, Massachusetts Institute of Technology, 2011.

44 Jelani Nelson and David P. Woodruff. Fast manhattan sketches in data streams. In *Proc. of the 29th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*, pages 99–110, 2010.

45 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`.

46 Chris Olston, Jing Jiang, and Jennifer Widom. Adaptive filters for continuous queries over distributed data streams. In *Proc. of the ACM SIGMOD International Conference on Management of Data (SIGMOD)*, pages 563–574, 2003.

47 Douglas B. Terry, David Goldberg, David A. Nichols, and Brian M. Oki. Continuous queries over append-only databases. In *Proc. of the ACM SIGMOD International Conference on Management of Data (SIGMOD)*, pages 321–330, 1992.

48 Salil P. Vadhan. Pseudorandomness. *Foundations and Trends in Theoretical Computer Science*, 7(1-3):1–336, 2012. `doi:10.1561/0400000010`.

49 David P. Woodruff. Optimal space lower bounds for all frequency moments. In *Proc. of the 15th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 167–175, 2004.