# Tracking the $\ell_2$ Norm with Constant Update Time


## Chi-Ning Chou
School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts, USA
http://cnchou.github.io
chiningchou@g.harvard.edu

## Zhixian Lei
School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts, USA
zhixianlei@seas.harvard.edu

## Preetum Nakkiran
School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts, USA
http://preetum.nakkiran.org
preetum@cs.harvard.edu



## Abstract

The $\ell_2$ *tracking problem* is the task of obtaining a streaming algorithm that, given access to a stream of items $a_1, a_2, a_3, \ldots$ from a universe $[n]$, outputs at each time $t$ an estimate to the $\ell_2$ norm of the *frequency vector* $f^{(t)} \in \mathbb{R}^n$ (where $f_i^{(t)}$ is the number of occurrences of item $i$ in the stream up to time $t$). The previous work [Braverman-Chestnut-Ivkin-Nelson-Wang-Woodruff, PODS 2017] gave a streaming algorithm with (the optimal) space using $O(\epsilon^{-2}\log(1/\delta))$ words and $O(\epsilon^{-2}\log(1/\delta))$ update time to obtain an $\epsilon$-accurate estimate with probability at least $1-\delta$. We give the first algorithm that achieves update time of $O(\log 1/\delta)$ which is independent of the accuracy parameter $\epsilon$, together with the nearly optimal space using $O(\epsilon^{-2}\log(1/\delta))$ words. Our algorithm is obtained using the Count Sketch of [Charilkar-Chen-Farach-Colton, ICALP 2002].



**2012 ACM Subject Classification** Theory of computation → Sketching and sampling

**Keywords and phrases** Streaming algorithms, Sketching algorithms, Tracking, CountSketch

**Digital Object Identifier** 10.4230/LIPIcs.APPROX-RANDOM.2019.2

**Category** APPROX

**Related Version** A full version of the paper is available at https://arxiv.org/abs/1807.06479.

**Funding** *Chi-Ning Chou*: Supported by NSF awards CCF 1565264 and CNS 1618026.
*Zhixian Lei*: Supported by NSF awards CCF 1565264 and CNS 1618026.
*Preetum Nakkiran*: Work supported in part by a Simons Investigator Award, NSF Awards CCF 1565641 and CCF 1715187, and the NSF Graduate Research Fellowship Grant No. DGE1144152.

**Acknowledgements** The authors wish to thank Jelani Nelson for invaluable advice throughout the course of this research. We also thank Mitali Bafna and Jarosław Błasiok for useful discussion and thank Boaz Barak for many helpful comments on an earlier draft of this article. We are also grateful to reviewers' comments.


## 1 Introduction

The *streaming model* considers the following setting. One is given a list $a_1, a_2, \ldots, a_m \in [n]$ as input where we think of $n$ as extremely large. The algorithm is only allowed to read the input once in a stream and the goal is to answer some predetermined queries using space of size logarithmic in $n$. For each $i \in [n]$ and time $t \in [m]$, define $f_i^{(t)} = |\{1 \leq j \leq t : \ a_j = i\}|$ as the frequency of $i$ at time $t$. Many classical streaming problems are concerned with approximating statistics of $f^{(m)}$ such as the distinct element problem (*i.e.,* $\|f^{(m)}\|_0$). One of the most well-studied problems is the one-shot $\ell_2$ estimation problem where the goal is to estimate $\|f^{(m)}\|_2^2$ within multiplicative error $(1 \pm \epsilon)$ and had been achieved by the seminal AMS sketch by Alon et al. [1].

We consider a streaming algorithm $A$ that maintains some logarithmic space and outputs an estimation $\sigma_t$ at the $t^{\text{th}}$ step of the computation. $A$ achieves $\ell_2$ $(\epsilon, \delta)$-tracking if for every input stream $a_1, a_2, \ldots, a_m \in [n]$

$$\Pr\left[\exists_{t\in[m]} \left|\sigma_t - \|f^{(t)}\|_2^2\right| > \epsilon\Delta_t\right] \leq \delta$$

where the "normalization factor" $\Delta_t$ differs between *strong* tracking and *weak* tracking. For $(\epsilon, \delta)$-*strong tracking*, $\Delta_t = \|f^{(t)}\|_2^2$ is the norm squared of the frequency vector up to the time $t$, while for $(\epsilon, \delta)$-weak tracking, $\Delta_t = \|f^{(m)}\|_2^2$ is the norm squared of the overall frequency vector. Note that strong tracking implies weak tracking and weak tracking implies one-shot approximation. In this work, we focus on $\ell_2$ tracking via linear sketching, where we specify a distribution $D$ on matrices $\Pi \in \mathbb{R}^{k\times n}$, and maintain a sketch vector at time $t$ as $\tilde{f}^{(t)} \triangleq \Pi f^{(t)}$. Then the estimate $\sigma_t$ is defined as $\|\tilde{f}^{(t)}\|_2^2$. The space complexity of $A$ is the number of machine words[1] required by $A$. The update time complexity of $A$ is the time to update $\sigma_t$, in terms of number of arithmetic operations.

Both weak tracking and strong tracking have been studied in different context [11, 5, 4] and the focus of this paper is on the *update time complexity*. Specifically, we are interested in the dependency of update time on the approximation factor $\epsilon$. The state-of-the-art result prior to our work is by Braverman et al. [4] showing that AMS provides weak tracking with $O(\epsilon^{-2}\log(1/\delta))$ update time and $O(\epsilon^{-2}\log(1/\delta))$ words of space.

Apart from tracking, there have been several sketching algorithms for one-shot approximation that have faster update time. Dasgupta et al. [8] and Kane and Nelson [16] showed that sparse JL achieves $O_\delta(\epsilon^{-1})$ [2] update time for $\ell_2$ one-shot approximation. Charikar, Chen, and Farach-Colton [6] designed the CountSketch algorithm for the heavy hitter problem and Thorup and Zhang [23] showed that it achieve $O_\delta(1)$ update time for $\ell_2$ one-shot approximation.

## Update time

Unlike the space complexity in streaming model, there have been less studies in the update time complexity though it is of great importance in applications. For example, the *packet passing problem* [21] requires the $\ell_2$ estimation in the streaming model with input arrival rate as high as $7.75 \times 10^6$ packets [3] per second. Thorup and Zhang [24] improved the update time from 182 nanoseconds to 50 nanoseconds and made the algorithm more practical.

While some streaming problems have algorithms with constant update time (*e.g.,* distinct elements [19] and $\ell_2$ estimation [24]), some other important problems do not ($\ell_p$ estimation for $p \neq 2$ [17], heavy hitters problems[4] [6, 7], and tracking problems [4]). Larsen et al. [22] systematically studies the update time complexity and showed lower bounds against heavy hitters, point query, entropy estimation, and moment estimation in the non-adaptive turnstile streaming model. In particular, they show that $O(\epsilon^{-2})$-space algorithms for $\ell_2$ estimation of vectors over $\mathbb{R}^n$, with failure probability $\delta$, must have update time roughly $\Omega(\log(1/\delta)/\sqrt{\log n})$. Note that their lower bound does not depend on $\epsilon$.

[1] Following convention, we assume the size of a machine word is at least $\Omega(\max(\log n, \log m))$ bits.
[2] $O_\delta(\cdot)$ is the same as the usual big O notation except treating $\delta$ as a constant.
[3] Each packet has 40 bytes (320 bits).
[4] There is a memory and update time tradeoff for heavy hitter from space $O(\epsilon^{-2}\log(n/\delta))$ to $O(\epsilon^{-2}(n/\delta))$ to get constant update time. However, achieving constant update time and logarithmic space simultaneously is unknown.

### Space lower bounds

For one-shot estimation of the $\ell_2$ norm, Kane et al. [20] showed that $\Theta(\epsilon^{-2}\log m + \log\log n)$ bits of space are required, for any streaming algorithm. This space lower bound is tight due to the AMS sketch. However, this only applies in the constant failure probability regime.

In the regime of sub-constant failure probability $\delta$, known tight lower-bounds on Distributional JL [15, 14] imply that $\Omega(\epsilon^{-2}\log(1/\delta))$ rows are necessary for the special case of linear sketching algorithms. [5] For linear sketches, this lower bound on number of rows is equivalent to a lower bound on the words of space.

For the regime of faster update time, Kane and Nelson [16] shows that CountSketch-type of constructions (with the optimal $\Omega(\epsilon^{-2}\log(1/\delta))$ rows) require sparsity i.e. number of non-zero elements $\widetilde{\Omega}(\epsilon^{-1}\log(1/\delta))$ [6] per column to achieve distortion $\epsilon$ and failure probability $\delta$. But, this does not preclude a sketch with suboptimal dependency on $\delta$ in the number of rows from having constant sparsity, for example a sketch with $\Omega_\delta(\epsilon^{-2})$ rows and constant sparsity – indeed, this is what CountSketch achieves. Note that in our setting, we can boost constant-failure probability to arbitrarily small failure probability by taking medians of estimators.[7] Thus, we may be able to bypass the lower-bounds for linear sketches.

To summarize the situation: for constant failure probability, it is only known that linear sketches require dimension $\Omega(\epsilon^{-2})$, and it is not known if super-constant sparsity is required for tracking with this optimal dimension. In particular, it was not known how to achieve say $(\epsilon, O(1))$-weak tracking for $\ell_2$, with $O(\epsilon^{-2})$ words of space and constant update time.

### Our contributions

In this paper, we show that there is a streaming algorithm with $O(\log(1/\delta))$ update time and space using $O(\epsilon^{-2}\log(1/\delta))$ words that achieves $\ell_2$ $(\epsilon,\delta)$-weak tracking.

▶ **Theorem 1** (informal). *For any $\epsilon > 0$, $\delta \in (0,1)$, and $n \in \mathbb{N}$. For any insertion-only stream over $[n]$ with frequencies $f^{(1)}, f^{(2)}, \dots, f^{(m)}$, there exists a streaming algorithm providing $\ell_2$ $(\epsilon,\delta)$-weak tracking with space using $O(\epsilon^{-2}\log(1/\delta))$ words and $O(\log(1/\delta))$ update time.*

Further, by applying a standard union bound argument in Lemma 13, the same algorithm can achieve $\ell_2$ strong tracking as well.

▶ **Corollary 2.** *For any $\epsilon > 0$, $\delta \in (0,1)$, and $n \in \mathbb{N}$. For any insertion-only stream over $[n]$ with frequencies $f^{(1)}, f^{(2)}, \dots, f^{(m)}$, there exists a streaming algorithm providing $\ell_2$ $(\epsilon,\delta)$-strong tracking with $O(\epsilon^{-2}\log(1/\delta)\log\log m)$ words and $O(\log(1/\delta)\log\log m)$ update time.*

The algorithm in the main theorem is obtained by running $O(\log(1/\delta))$ many copies of CountSketch and taking the median.

The main techniques used in the proof are the chaining argument and Hansen-Wright inequality which are also used in [4] to show the tracking properties of AMS. However, direct applications of these tools on the CountSketch algorithm would not give the desired bounds due to the sparse structure of the sketching matrix. To overcome this issue, we have to dig into the structure of sketching matrix of CountSketch. We will compare the difference between our techniques and that in [4] after presenting the proof of Theorem 1 (see Remark 12).

---

[5] Note that an $(\epsilon,\delta)$-weak tracking via linear sketch defines a distribution over matrices that satisfies the Distributional JL guarantee, with distortion $(1 \pm \epsilon)$ and failure probability $\delta$.

[6] $\tilde{\Omega}(\cdot)$ is the same as the $\Omega(\cdot)$ notation by ignoring extra logarithmic factor.

[7] This is not immediate for weak tracking.

The rest of the paper is organized as follows. Some preliminaries are provided in Section 2. In Section 3, we prove our main theorem showing that CountSketch with $O(\epsilon^{-2})$ rows achieves $\ell_2$ $(\epsilon, O(1))$-weak tracking with constant update time. As for the $\ell_2$ strong tracking, we discuss some upper and lower bounds in Section 4. In Section 5, we discuss some future directions and open problems.

## 2 Preliminaries

In the following, $n \in \mathbb{N}$ denotes the size of the universe, $k$ denotes the number of rows of the sketching matrix, $t$ denotes the time, and $m$ denote the final time. We let $[n] = \{1, 2, \ldots, n\}$ and use $\tilde{O}(\cdot)$ and $\tilde{\Omega}(\cdot)$ to denote the usual $O(\cdot)$ and $\Omega(\cdot)$ with some extra poly-logarithmic factor.

The input of the streaming algorithm is a list $a_1, a_2, \ldots, a_m \in [n]$. For each $i \in [n]$ and time $t \in [m]$, define $f_i^{(t)} = |\{1 \leq j \leq t : \ a_j = i\}|$ as the frequency of $i$ at time $t$. The one-shot $\ell_2$ approximation problem is to produce an estimate for $\|f^{(m)}\|_2^2$ with $(1 \pm \epsilon)$ multiplicative error and success probability at least $1 - \delta$ for $\epsilon > 0$ and $\delta \in (0, 1)$.

### 2.1 $\ell_2$ tracking

Here, we give the formal definition of $\ell_2$ tracking for sketching algorithm.

▶ **Definition 3** ($\ell_2$ tracking)**.** *For any $\epsilon > 0, \delta \in (0, 1)$, and $n, m \in \mathbb{N}$. Let $f^{(1)}, f^{(2)}, \ldots, f^{(m)}$ be the frequency of an insertion-only stream over $[n]$ and $\tilde{f}^{(1)}, \tilde{f}^{(2)}, \ldots, \tilde{f}^{(m)}$ be its (randomized) approximation produced by a sketching algorithm. We say the algorithm provides $\ell_2$ $(\epsilon, \delta)$-strong tracking if*

$$\Pr\left[\exists_{t \in [m]}, \ \left| \|\tilde{f}^{(t)}\|_2^2 - \|f^{(t)}\|_2^2 \right| > \epsilon \|f^{(t)}\|_2^2 \right] \leq \delta.$$

*We say the algorithm provides $\ell_2$ $(\epsilon, \delta)$-weak tracking if*

$$\Pr\left[\exists_{t \in [m]}, \ \left| \|\tilde{f}^{(t)}\|_2^2 - \|f^{(t)}\|_2^2 \right| > \epsilon \|f^{(m)}\|_2^2 \right] \leq \delta.$$

Note that the difference between the two tracking guarantee is that in strong tracking we bound the deviation of the estimate from the true norm squared by $\epsilon \|f^{(t)}\|_2^2$ while in the weak tracking we bound this deviation by $\epsilon \|f^{(m)}\|_2^2$.

### 2.2 AMS sketch and CountSketch

Alon *et al.* [1] proposed the seminal AMS sketch for $\ell_2$ approximation in the streaming model. In AMS sketch, consider $\Pi \in \mathbb{R}^{k \times n}$ where $\Pi_{j,i} = \sigma_{j,i}/\sqrt{k}$ and $\sigma_{j,i}$ is i.i.d. Rademacher for each $j \in [m], i \in [n]$. When $k = O(\epsilon^{-2})$, AMS sketch approximates $\ell_2$ norm within $(1 \pm \epsilon)$ multiplicative error. Note that the update time of AMS sketch is $k$ since the matrix $\Pi$ is dense.

Charikar, Chen, and Farach-Colton [6] proposed the following CountSketch algorithm for the heavy hitter problem and Thorup and Zhang [23] showed that CountSketch is also able to solve the $\ell_2$ approximation. Here, consider $\Pi \in \mathbb{R}^{k \times n}$ where we denote the $i^{\text{th}}$ column of $\Pi$ as $\Pi_i$ for each $i \in [n]$. $\Pi_i$ is defined as follows. First, pick $j \in [k]$ uniformly and set $\Pi_{j,i}$ to be an independent Rademacher. Next, set the other entries in $\Pi_i$ to be 0. Note that unlike AMS sketch, the normalization term in CountSketch is 1 since there is exactly one non-zero entry in each column. [6] showed that CountSketch provides one-shot $\ell_2$ approximation with $O(\epsilon^{-2})$ rows.

▶ **Lemma 4** ([6, 23]). *Let $\epsilon > 0$, $\delta \in (0,1)$, and $n \in \mathbb{N}$. Pick $k = \Omega(\epsilon^{-2}\delta^{-1})$, we have for any $x \in \mathbb{R}^n$,*

$$\Pr_{\Pi}\left[|\|\Pi \boldsymbol{x}\|_2^2 - \|\boldsymbol{x}\|_2^2| > \epsilon\|\boldsymbol{x}\|_2^2\right] \leq \delta.$$

#### Implement CountSketch in logarithmic space

Previously, we defined CountSketch using uniformly independent randomness, which requires space $\Omega(nk)$. However, one could see that in the proof of Theorem 8 we actually only need 8-wise independence. Thus, the space required can be reduced to $O(\log n)$ for each row. It is well known that CountSketch with $k$ rows can be implemented with 8-wise independent hash family using $O(k)$ words. We describe the whole implementation in Appendix A for completeness.

### 2.3 $\epsilon$-net for insertion-only stream

In our analysis, we will use the following existence of a small $\epsilon$-net for insertion-only streams.

▶ **Definition 5** ($\epsilon$-net). *Let $S \subseteq \mathbb{R}^n$ be a set of vectors. For any $\epsilon > 0$, we say $E \subseteq \mathbb{R}^n$ is an $\epsilon$-net for $S$ with respect to $\ell_2$ norm if for any $x \in S$, there exists $y \in E$ such that $\|x - y\|_2 \leq \epsilon$.*

▶ **Lemma 6** ([5]). *Let $\{x^{(t)}\}_{t\in[m]}$ be an insertion-only stream. For any $\epsilon > 0$, there exists a size $\left(1 + \epsilon^{-2} \cdot \|x^{(m)}\|_2\right)$ $\epsilon$-net for $\{x^{(t)}\}_{t\in[m]}$ with respect to $\ell_2$ norm. Moreover, the elements in the net are all from $\{x^{(t)}\}_{t\in[m]}$.*

**Proof Sketch.** The idea is to use a greedy algorithm, by scanning through the stream from the beginning and adding an element $x^{(t)}$ into the net if there does not already exist an element in the net that is $\epsilon$-close to $x^{(t)}$. ◀

### 2.4 Concentration inequalities

Our analysis crucially relies on the following Hanson-Wright inequality [10].

▶ **Lemma 7** (Hanson-Wright inequality [10]). *For any symmetric $B \in \mathbb{R}^{n\times n}$, $\sigma \in \{\pm 1\}^n$ being independent Rademacher vector, and integer $p \geq 1$, we have*

$$\|\sigma^\top B\sigma - \mathbb{E}_\sigma[\sigma^\top B\sigma]\|_p \leq O\left(\sqrt{p}\|B\|_F + p\|B\|\right) = O(p\|B\|_F),$$

*where $\|X\|_p$ is defined as $\mathbb{E}[|X|^p]^{1/p}$ and $\|\cdot\|_F$ is the Frobenius norm.*

Note that the only randomness in $\sigma^\top B\sigma - \mathbb{E}_\sigma[\sigma^\top B\sigma]$ is the Rademacher vector $\sigma$.

## 3 CountSketch with $O(\epsilon^{-2})$ rows provides $\ell_2$ weak tracking

In this section we will show that CountSketch with $O(\epsilon^{-2})$ rows provides $(\epsilon, O(1))$-weak tracking.

▶ **Theorem 8** (CountSketch with $O(\epsilon^{-2})$ rows provides $\ell_2$ weak tracking). *For any $\epsilon > 0$, $\delta \in (0,1)$, and $n \in \mathbb{N}$. Pick $k = \Omega(\epsilon^{-2}\delta^{-1})$. For any insertion-only stream over $[n]$ with frequency $f^{(1)}, f^{(2)}, \ldots, f^{(m)}$, the CountSketch algorithm with $k$ rows provides $\ell_2$ $(\epsilon, \delta)$-weak tracking.*

▶ Remark. Note that for linear sketches, the dependency of number of rows on $\epsilon$ is tight in Theorem 8. This is implied by known lower-bounds on Distributional JL [15, 14], which imply lower-bounds on one-shot $\ell_2$ approximation.

▶ Remark. Recall that the number of rows in linear sketches is proportional to the number of words needed in the algorithm.

Using the standard median trick, we can run $O(\log(1/\delta))$ copies of CountSketch with $k = O(\epsilon^{-2})$ in parallel and output the median. With this, Theorem 8 immediately gives the following corollary with better dependency on $\delta$.

▶ **Corollary 9.** *For any $\epsilon > 0$, $\delta \in (0, 1)$, and $n \in \mathbb{N}$. For any insertion-only stream over $[n]$ with frequency $f^{(1)}, f^{(2)}, \ldots, f^{(m)}$, there exists a streaming algorithm providing $\ell_2$ $(\epsilon, \delta)$-weak tracking with $k = O(\epsilon^{-2}\log(1/\delta))$ rows and update time $O(\log(1/\delta))$.*

The proof of Theorem 8 uses the Dudley-like chaining technique similar to other tracking proofs [4]. However, direct application of the chaining argument would not suffice and we have to utilize the structure of the sketching matrix of CountSketch (see Remark 12 for comparison). We will prove Theorem 8 in Subsection 3.1.

## 3.1 Proof of Theorem 8

In this subsection, we give a formal proof for our main theorem. Let us start with some notations for CountSketch. Recall that for any $i \in [n]$, the $i^{\text{th}}$ column of $\Pi$ is defined by (i) picking $j \in [k]$ uniformly and set $\Pi_{j,i}$ to be a Rademacher random variable and (ii) set the other entries in $\Pi_i$ to be 0. Denote $\Pi_{j,i} = \sigma_{j,i}\eta_{j,i}$, where $\sigma_{j,i}$ is a Rademacher random variable, and $\eta_{j,i}$ is the indicator for choosing the $j^{\text{th}}$ row in the $i^{\text{th}}$ column. Note that there is exactly one non-zero entry in each column and the probability distribution is uniform. The approximation error of $\Pi$ for a vector $\mathbf{x} \in \mathbb{R}^n$ is denoted as $\gamma(\mathbf{x}) := \left|\|\Pi\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\right|$. To show weak tracking, it suffices to upper bound the supremum of $\gamma(f^{(t)})$.

$$\mathbb{E}_\Pi \sup_{t\in[m]} \gamma(f^{(t)}) = \mathbb{E}_\Pi \sup_{t\in[m]} \left|\|\Pi f^{(t)}\|_2^2 - \|f^{(t)}\|_2^2\right|. \tag{1}$$

The first observation[8] is that one can rewrite the error $\gamma(\mathbf{x})$ as follows.

$$\gamma(\mathbf{x}) = \left|\mathbf{x}^\top\Pi^\top\Pi\mathbf{x} - \mathbf{x}^\top\mathbf{x}\right| = \left|\sigma^\top B_{\eta,\mathbf{x}}\sigma - \mathbf{x}^\top\mathbf{x}\right| = \left|\sigma^\top \tilde{B}_{\eta,\mathbf{x}}\sigma\right|,$$

where $\sigma \in \{-1, 1\}^n$ is an independent Rademacher random vector and for any $i, i' \in [n]$,

$$(\tilde{B}_{\eta,\mathbf{x}})_{i,i'} = \begin{cases} \mathbf{x}_i\mathbf{x}_{i'}, & i \neq i' \text{ and } \exists j \in [k],\ \eta_{j,i} = \eta_{j,i'} = 1 \\ 0, & \text{else.} \end{cases}$$

Note that the diagonals of $\tilde{B}_{\eta,\mathbf{x}}$ are all zero as follow.

$$\tilde{B}_{\eta,\mathbf{x}} = \begin{pmatrix} 0 & \mathbf{x}_1\mathbf{x}_2\langle\Pi_1,\Pi_2\rangle & \cdots & \mathbf{x}_1\mathbf{x}_n\langle\Pi_1,\Pi_n\rangle \\ \mathbf{x}_2\mathbf{x}_1\langle\Pi_2,\Pi_1\rangle & 0 & \cdots & \mathbf{x}_2\mathbf{x}_n\langle\Pi_2,\Pi_n\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n\mathbf{x}_1\langle\Pi_n,\Pi_1\rangle & \mathbf{x}_n\mathbf{x}_2\langle\Pi_n,\Pi_2\rangle & \cdots & 0 \end{pmatrix}.$$

[8] Note that the matrix $\tilde{B}_{\mathbf{x}}$ we are using is different from the matrix used in the previous analysis of [4]. This difference is crucial since the matrix of [4] does not work for CountSketch.

For convenience, for any matrix $B \in \mathbb{R}^{n\times n}$, we overload the notation $\gamma$ by denoting $\gamma(B) = \sigma^\top B\sigma$. That is, $\gamma(\tilde{B}_{\eta,\mathbf{x}}) = \gamma(\mathbf{x})$. One benefit of writing $\ell_2$ weak tracking error into the above quadratic form is that Hanson-Wright inequality (see Lemma 7) is now applicable.

The lemma below shows that the expectation of the weak tracking error is upper bounded by the Frobenius norm of $\tilde{B}_{\eta,f^{(m)}}$.

▶ **Lemma 10.** *Let $\{f^{(t)}\}_{t\in[m]}$ be the frequencies of an insertion-only stream. We have*

$$\mathbb{E}\left[\sup_{t\in[m]} \gamma(f^{(t)}) \mid \eta\right] = O(\|\tilde{B}_{\eta,f^{(m)}}\|_F).$$

The proof of Lemma 10 uses the Dudley-like chaining argument. For the smooth of presentation, we postpone the details to Subsection 3.2. Next, the following lemma shows that for any vector $x \in \mathbb{R}^n$, with high probability, $\|\tilde{B}_{\eta,x}\|_F = O(\|x\|_2^2/\sqrt{k})$.

▶ **Lemma 11.** *For any $\delta \in (0,1)$ and $x \in \mathbb{R}^n$,*

$$\Pr\left[\|\tilde{B}_{\eta,x}\|_F > \frac{\sqrt{2}\|x\|_2^2}{\sqrt{\delta\cdot k}}\right] \le \frac{\delta}{2}.$$

Lemma 11 has similar flavor as Lemma 4. The proof can be found in Subsection 3.2. Finally, Theorem 8 is an immediate corollary of Lemma 10 and Lemma 11. Here we provide a proof for completeness.

**Proof of Theorem 8.** Recall that to prove Theorem 8, it suffices to show that with probability at least $1-\delta$ over $\eta$, $\sup_{t\in[m]} \gamma(f^{(t)}) \le \epsilon$. From Lemma 10, for a fixed $\eta$, we have $\Pr\left[\sup_{t\in[m]} \gamma(f^{(t)}) > C_1\|\tilde{B}_{\eta,f^{(m)}}\|_F\right] \le \delta/2$ for some constant $C_1 > 0$. Next, from Lemma 11, we have $\|\tilde{B}_{\eta,f^{(m)}}\|_F \le \|f^{(m)}\|_2^2 \cdot k^{-1/2} \cdot \delta^{-1/2}$ with probability at least $1-\delta/2$ over the randomness in $\eta$ for some constant $C_2 > 0$. Pick $m \ge C_1C_2 \cdot \epsilon^{-2} \cdot \delta^{-1}$, we have $\Pr\left[\sup_{t\in[m]} \gamma(f^{(t)}) > \epsilon\|f^{(m)}\|_2^2\right] \le \delta$ and complete the proof. ◀

## 3.2 Proof of the two key lemmas

In this subsection, we provide the proofs for Lemma 10 and Lemma 11. Let us start with Lemma 10 which shows that the tracking error can be upper bounded by the Frobenius norm of $\tilde{B}_{\eta,f^{(m)}}$.

**Proof of Lemma 10.** Recall that we define $\tilde{B}_{\eta,x}$ such that $\gamma(x) = \sigma^\top\tilde{B}_{\eta,x}\sigma$ where $\sigma$ is 8-wise independent Rademacher random vector. An important trick here is that we think of *fixing*[9] $\eta$ in the following.

The starting point of chaining argument is constructing a sequence of $\epsilon$-nets with exponentially decreasing error for $\{\tilde{B}_{\eta,f^{(t)}}\}_{t\in[m]}$. Note that here $\{\tilde{B}_{\eta,f^{(t)}}\}_{t\in[m]}$ are matrices but one can view it as a vector and apply Lemma 6 where $\ell_2$ norm for a vector becomes Frobenius norm for a matrix. Namely, for any non-negative integer $\ell$, let $T_{\eta,\ell}$ be the $(\|\tilde{B}_{\eta,f^{(m)}}\|_F/2^\ell)$-net for $\{\tilde{B}_{\eta,f^{(t)}}\}_{t\in[m]}$ under Frobenius norm where $|T_{\eta,\ell}| \le 1 + 2^{2\ell}$. Note that here we fixed $\eta$ first and then constructed the nets. Thus, for each $t \in [m]$, one can rewrite $\tilde{B}_{\eta,f^{(t)}}$ into a *chain* as follows.

$$\tilde{B}_{\eta,f^{(t)}} = B^{(t)}_{\eta,0} + \sum_{\ell=1}^{\infty} B^{(t)}_{\eta,\ell} - B^{(t)}_{\eta,\ell-1}, \tag{2}$$

[9] We do this by conditioning on $\eta$.

where $B^{(t)}_{\eta,\ell} \in T_{\eta,\ell}$ and $\|\tilde{B}_{\eta,f^{(t)}} - B^{(t)}_{\eta,\ell}\|_F \leq 2^{-\ell} \cdot \|\tilde{B}_{\eta,f^{(m)}}\|_F$. Moreover, from Equation 2 we have

$$\mathbb{E} \sup_{t\in[m]} \gamma(f^{(t)}) \leq \mathbb{E} \sup_{t\in[m]} \gamma(B^{(t)}_{\eta,0}) + \sum_{\ell=1}^{\infty} \mathbb{E} \sup_{t\in[m]} \gamma(B^{(t)}_{\eta,\ell} - B^{(t)}_{\eta,\ell-1}). \tag{3}$$

To bound the first term of Equation 3, observe that $T_{\eta,0} = \{\tilde{B}_{\eta,f^{(1)}}\}$ where $\tilde{B}_{\eta,f^{(1)}}$ is the all zero matrix. Namely, the first term of Equation 3 is zero. As for the second term of Equation 3, we apply the chaining argument as follows. For any positive integer $\ell$, denote $\mathcal{A}_\ell = \{B^{(t)}_{\eta,\ell} - B^{(t)}_{\eta,\ell-1}\}_{t\in[m]}$. Note that from the construction of $\epsilon$-net in Lemma 6, we have $|\mathcal{A}_\ell| \leq 2|T_{\eta,\ell}| \leq 2^{2\ell+2}$ by triangle inequality.

$$\begin{aligned}\mathbb{E}\left[\sup_{t\in[m]} \gamma(B^{(t)}_{\eta,\ell} - B^{(t)}_{\eta,\ell-1})\right] &= \int_0^\infty \Pr\left[\sup_{A\in\mathcal{A}_\ell} \gamma(A) > u\right] du \\ &\leq u^*_\ell + \int_{u^*_\ell}^\infty \Pr\left[\sup_{A\in\mathcal{A}_\ell} \gamma(A) > u\right] du, \end{aligned}\tag{4}$$

where $u^*_\ell > 0$ will be chosen later. For any $A \in \mathcal{A}_\ell$ and integer $p \geq 2$, by Markov's inequality and Hanson-Wright inequality, we have

$$\Pr[\gamma(A) > u] \leq \frac{\mathbb{E}[\gamma(A)^p]}{u^p} = \frac{\|\sigma^\top A\sigma\|_p^p}{u^p} \leq \frac{\left(C \cdot \sqrt{p}\|A\|_F + C \cdot p\|A\|\right)^p}{u^p}$$

for some constant $C > 0$. Note that the randomness here is only in $\sigma$ and thus we can apply the Hanson-Wright inequality. Let $R_\ell = \sup_{A\in\mathcal{A}_\ell} \left(C \cdot \sqrt{p}\|A\|_F + C \cdot p\|A\|\right) \leq C'p \cdot \|\tilde{B}_{\eta,f^{(m)}}\|_F \cdot 2^{-\ell}$ for some $C' > 0$. The last inequality holds because of $\|\cdot\| \leq \|\cdot\|_F$ and the choice of $\epsilon$-net. Now, choose $u^*_\ell = 2S_\ell \cdot R_\ell$ where $S_\ell$ will be decided later, Equation 4 becomes

$$\begin{aligned}\mathbb{E}\left[\sup_{t\in[m]} \gamma(B^{(t)}_{\eta,\ell} - B^{(t)}_{\eta,\ell-1})\right] &\leq u^*_\ell + \int_{u^*_\ell}^\infty |\mathcal{A}_\ell| \cdot \frac{{R_\ell}^p}{u^p} du \qquad (5)\\ &\leq 2S_\ell R_\ell + |\mathcal{A}_\ell| \cdot \frac{{R_\ell}^p}{(2S_\ell R_\ell)^{p-1}} \\ &\leq 2S_\ell C'p \cdot \|\tilde{B}_{\eta,f^{(m)}}\|_F \cdot 2^{-\ell} + |\mathcal{A}_\ell| \cdot \frac{C'p \cdot \|\tilde{B}_{\eta,f^{(m)}}\|_F}{S_\ell^{p-1}} \cdot 2^{-\ell}\end{aligned}$$

where the second term of Equation 5 is due to union bound. Now, Equation 3 becomes

$$\begin{aligned}\mathbb{E} \sup_{t\in[m]} \gamma(f^{(t)}) &\leq \sum_{\ell=1}^{\infty} 2S_\ell C'p \cdot \|\tilde{B}_{\eta,f^{(m)}}\|_F \cdot 2^{-\ell} + |\mathcal{A}_\ell| \cdot \frac{C'p \cdot \|\tilde{B}_{\eta,f^{(m)}}\|_F}{S_\ell^{p-1}} \cdot 2^{-\ell} \\ &\leq \|\tilde{B}_{\eta,f^{(m)}}\|_F \cdot \left(\sum_{\ell=1}^{\infty} 2C'pS_\ell \cdot 2^{-\ell} + \frac{2^\ell C'p}{S_\ell^{p-1}}\right). \end{aligned}\tag{6}$$

Choose $S_\ell = 2^{3\ell/4}$ and $p \geq 4$, the summation term in Equation 6 can thus be upper bounded by a constant. We conclude that

$$\mathbb{E} \sup_{t\in[m]} \gamma(f^{(t)}) = O(\|\tilde{B}_{\eta,f^{(m)}}\|_F).$$

Note that this also means that 8-wise independence suffices and thus the sketching matrix can be efficiently stored (see Appendix A for more details). ◀

Next, we prove Lemma 11 which upper bounds the expectation of $\|\tilde{B}_{\eta,\mathbf{x}}\|$ for any $\mathbf{x} \in \mathbb{R}^n$.

**Proof of Lemma 11.** We first show that $\mathbb{E}_\eta \|\tilde{B}_{\eta,x}\|_F^2 \leq \frac{\|x\|_2^4}{k}$ and the lemma immediately holds due to Markov's inequality.

Let $\mathbf{1}_{ii'}$ be the indicator for whether there exists $j \in [k]$ such that $\eta_{ij} = \eta_{i'j} = 1$. Note that for $i \neq i'$, $\mathbb{E}[\mathbf{1}_{ii'}] = 1/k$ and the only randomness here is in $\eta$.

$$\begin{aligned}\mathbb{E}\|\tilde{B}_{\eta,x}\|_F^2 &= \mathbb{E} \sum_{i,i' \in [n]} (\tilde{B}_{\eta,x})_{i,i'}^2 = \mathbb{E} \sum_{(i,i') \in [n]^2,\ i \neq i'} x_i^2 x_{i'}^2 \mathbf{1}_{ii'} \\ &= \frac{1}{k} \sum_{(i,i') \in [n]^2,\ i \neq i'} x_i^2 x_{i'}^2 \leq \frac{\|x\|_2^4}{k},\end{aligned}$$

where the last inequality is by Cauchy-Schwarz. Note that 8-wise independence is sufficient in the above argument. ◀

▶ Remark 12. Here, let us briefly compare the difference between our techniques and that in [4]. There are two key observations on the structure of the sketching matrix of CountSketch. First, we observe that the Frobenius norm of $\Pi^\top \Pi$ is dominated by its diagonal and thus *removing* the diagonal would give us a more accurate analysis on the contribution from the off-diagonal term. However, removing the diagonal of $\Pi^\top \Pi$ destroys the symmetric structure and thus the standard $\epsilon$-net argument (e.g., in [4]) would not work. To overcome this, we observe that one can directly construct $\epsilon$-net for the matrix obtained by removing the diagonal from $\Pi^\top \Pi$. Combining these two observations and standard chaining argument, we are able to show that CountSketch provides $\ell_2$ weak tracking.

## 4 Strong tracking of AMS sketch and CountSketch

In this section, we are going to discuss the strong tracking of AMS sketch and CountSketch. We start with a standard reduction from weak tracking to strong tracking via union bound. This gives us an $O(\log m)$ blow-up in the dependency on $\delta$. Next, we show that this is essentially tight for both AMS sketch and CountSketch up to a logarithmic factor.

▶ **Lemma 13** (folklore). *For any $\epsilon > 0$, $\delta \in (0,1)$, and $n, m \in \mathbb{N}$. If a linear sketch provides $(\epsilon, \delta)$ weak tracking for length $m$ inputs having value from $[n]$, then it also provides $(2\epsilon, \delta')$ strong tracking where $\delta' = \min\{1, (\log m) \cdot \delta\}$.*

**Proof.** See Subsection B.1 for details. ◀

From Lemma 13, we immediate have the following corollaries.

▶ **Corollary 14.** *For any $\epsilon > 0$ and $\delta \in (0,1)$, AMS sketch with $O\left(\epsilon^{-2}(\log\log m + \log(1/\delta))\right)$ rows provides $\ell_2$ $(\epsilon, \delta)$-strong tracking.*

▶ **Corollary 15.** *For any $\epsilon > 0$ and $\delta \in (0,1)$, CountSketch with $O\left(\epsilon^{-2}\delta^{-1}\log m\right)$ rows provides $\ell_2$ $(\epsilon, \delta)$-strong tracking.*

▶ Remark. After applying median trick on CountSketch, the dependency of the number of rows on $\delta$ becomes $O(\log(1/\delta))$ and thus $O\left(\epsilon^{-2}(\log\log m + \log(1/\delta))\right)$ rows suffices to achieve $\ell_2$ $(\epsilon, \delta)$-strong tracking.

In the following, we are going to show that the above two upper bounds are essentially tight for these two algorithms.

▶ **Theorem 16.** *There exists constants $C > 0$ such that for any $\epsilon \in (0, 0.1)$ and $\delta \in (0, 1)$, there exists $N_0 \in \mathbb{N}$ such that if $k < C \cdot \left(\log \frac{\log m}{\log(1/\epsilon)} + \log(1/\delta)\right)$ and $N_0 \leq n \leq m$, then fully independent AMS sketch with $k$ rows does not provide $\ell_2$ $(\epsilon, \delta)$-strong tracking.*

That is, AMS sketch requires $\tilde{\Omega}\left(\epsilon^{-2}(\log \log m + \log(1/\delta))\right)$ rows to achieve $\ell_2$ $(\epsilon, \delta)$-strong tracking. Interestingly, the hard instance for AMS sketch to achieve strong tracking is simply the stream consisting all distinct elements. See Subsection B.2 for details.

▶ **Theorem 17.** *There exists a constant $C > 0$ such that for any $\epsilon \in (0, 0.5)$, and $\delta \in (0, 1)$, there exists $N_0 \in \mathbb{N}$ such that if $k \leq C \cdot \epsilon^{-2}\delta^{-1}\frac{\log m}{\log(1/\epsilon)}$ and $N_0 \leq n \leq O(\log m)$, then CountSketch with $k$ rows does not provide $\ell_2$ $(\epsilon, \delta)$-strong tracking.*

That is, CountSketch requires $\tilde{\Omega}(\epsilon^{-2}\delta^{-1}\log m)$ rows to achieve $\ell_2$ $(\epsilon, \delta)$-strong tracking. The hard instance for CountSketch is more complicated than that of AMS sketch. See Subsection B.3 for details.

## 5 Conclusion

In this work, we showed that CountSketch provides $\ell_2$ weak tracking with update time having no dependence on the error parameter $\epsilon$. We also give almost tight $\ell_2$ strong tracking lower bounds for AMS sketch and CountSketch.

An immediate open problem after this work would be tracking $\ell_p$ with faster update time for $0 < p < 2$. The $\ell_p$ estimation problem had been solved by Indyk [12] via *p-stable sketch* and was proven to provide weak tracking by Błasiok et al. [3]. However, same as AMS sketch, the $p$-stable sketch is dense and has update time $\Omega(\epsilon^{-2})$. Nevertheless, Kane et al. [18] gave a space-optimal algorithm for $\ell_p$ estimation problem with update time $O(\log^2(1/\epsilon)\log\log(1/\epsilon))$. It would be interesting to see if their algorithm also provides $\ell_p$ weak tracking.

### References

1 Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pages 20–29. ACM, 1996.

2 Andrew C Berry. The accuracy of the Gaussian approximation to the sum of independent variates. *Transactions of the american mathematical society*, 49(1):122–136, 1941.

3 Jaroslaw Blasiok, Jian Ding, and Jelani Nelson. Continuous Monitoring of l_p Norms in Data Streams. In *LIPIcs-Leibniz International Proceedings in Informatics*, volume 81. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

4 Vladimir Braverman, Stephen R Chestnut, Nikita Ivkin, Jelani Nelson, Zhengyu Wang, and David P Woodruff. BPTree: An $\ell_2$ heavy hitters algorithm using constant memory. In *Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pages 361–376. ACM, 2017.

5 Vladimir Braverman, Stephen R Chestnut, Nikita Ivkin, and David P Woodruff. Beating CountSketch for heavy hitters in insertion streams. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 740–753. ACM, 2016.

6 Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding frequent items in data streams. In *International Colloquium on Automata, Languages, and Programming*, pages 693–703. Springer, 2002.

7 Graham Cormode and Shan Muthukrishnan. An improved data stream summary: the count-min sketch and its applications. *Journal of Algorithms*, 55(1):58–75, 2005.

**8** Anirban Dasgupta, Ravi Kumar, and Tamás Sarlós. A sparse Johnson-Lindenstrauss transform. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 341–350. ACM, 2010.

**9** Carl-Gustaf Esseen. *On the Liapounoff limit of error in the theory of probability*. Almqvist & Wiksell Stockholm, 1942.

**10** David Lee Hanson and Farroll Tim Wright. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3):1079–1083, 1971.

**11** Zengfeng Huang, Wai Ming Tai, and Ke Yi. Tracking the Frequency Moments at All Times. *arXiv preprint*, 2014. `arXiv:1412.1763`.

**12** Piotr Indyk. Stable distributions, pseudorandom generators, embeddings, and data stream computation. *Journal of the ACM (JACM)*, 53(3):307–323, 2006.

**13** Tadeusz Inglot and Teresa Ledwina. Asymptotic optimality of new adaptive test in regression model. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 42(5):579–590, 2006.

**14** 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, June 2013. `doi:10.1145/2483699.2483706`.

**15** Daniel Kane, Raghu Meka, and Jelani Nelson. Almost optimal explicit Johnson-Lindenstrauss families. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 628–639. Springer, 2011.

**16** Daniel M Kane and Jelani Nelson. Sparser Johnson-Lindenstrauss transforms. *Journal of the ACM (JACM)*, 61(1):4, 2014.

**17** Daniel M Kane, Jelani Nelson, Ely Porat, and David P Woodruff. Fast moment estimation in data streams in optimal space. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 745–754. ACM, 2011.

**18** Daniel M Kane, Jelani Nelson, Ely Porat, and David P Woodruff. Fast moment estimation in data streams in optimal space. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 745–754. ACM, 2011.

**19** Daniel M Kane, Jelani Nelson, and David P Woodruff. An optimal algorithm for the distinct elements problem. In *Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 41–52. ACM, 2010.

**20** Daniel M Kane, Jelani Nelson, and David P Woodruff. On the exact space complexity of sketching and streaming small norms. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*, pages 1161–1178. Society for Industrial and Applied Mathematics, 2010.

**21** Balachander Krishnamurthy, Subhabrata Sen, Yin Zhang, and Yan Chen. Sketch-based change detection: methods, evaluation, and applications. In *Proceedings of the 3rd ACM SIGCOMM conference on Internet measurement*, pages 234–247. ACM, 2003.

**22** Kasper Green Larsen, Jelani Nelson, and Huy L Nguyên. Time lower bounds for nonadaptive turnstile streaming algorithms. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 803–812. ACM, 2015.

**23** Mikkel Thorup and Yin Zhang. Tabulation based 4-universal hashing with applications to second moment estimation. In *SODA*, volume 4, pages 615–624, 2004.

**24** Mikkel Thorup and Yin Zhang. Tabulation-based 5-independent hashing with applications to linear probing and second moment estimation. *SIAM Journal on Computing*, 41(2):293–331, 2012.

# A Implementation of CountSketch

Here, we present the implementation of CountSketch for the completeness. Note that the construction is standard and not new.

**Algorithm 1** Constructing CountSketch.

---

$k \leftarrow \lceil \frac{c}{\epsilon^2} \rceil$ for some constant $c > 0$.
$\tilde{f} \in \mathbb{Z}^k$ vector with initial value 0.
Sample $h : [n] \to [k]$ from a 8-wise independent hash family.
Sample $g : [n] \to \{\pm 1\}$ from a 8-wise independent hash family.
**for** $t = 1, 2, \ldots, m$ **do**
    On input $a_t = i$, set $\tilde{f}_{h(i)} = \tilde{f}_{h(i)} + g(i)$.

---

Note that both $h$ and $g$ can be stored in space $O(\log n + \log(1/\epsilon))$ and be evaluated in $O(1)$ many arithmetic operations. $\tilde{f}$ can be stored in space $O(\epsilon^{-2} \log m)$ bits. For the convenience of analysis, we define the sketching matrix $\Pi \in \{0, \pm 1\}^{k \times n}$ of CountSketch by $\Pi_{h(i),i} = g(i)$ for all $i \in [n]$.

# B Proofs for strong tracking

## B.1 From weak tracking to strong tracking

After applying union bound on all points $t = 1, 2, \ldots, m$, a streaming algorithm provides $\ell_2$ $(\epsilon, \delta)$-approximation also provides $\ell_2$ $(\epsilon, \delta')$-strong tracking where $\delta' = \min\{1, m\delta\}$. However, the blow-up in $\delta$ is $m$, which is undesirable. The following lemma shows that with a more delicate union bound argument, the reduction from weak tracking to strong tracking only has $O(\log m)$ blow-up in $\delta$. Note that the lemma is a folklore and we provide a proof for completeness.

**Proof.** Let $\{f^{(t)}\}_{t \in [m]}$ be the frequency of an insertion-only stream and let $\{\tilde{f}^{(t)}\}_{t \in [m]}$ be its (randomized) approximations produced by the linear sketch. Let $w = \lfloor \log m \rfloor + 1$ and $t_i = 2^i - 1$ for each $i \in [w]$. Note that for each $i \in [w]$ and $t_{i-1} < t \leq t_i$, $\frac{1}{2}\|f^{(t_i)}\|_2^2 \leq \|f^{(t)}\|_2^2 \leq \|f^{(t_i)}\|_2^2$. Define the event

$$E_i := \left\{ |\|\tilde{f}^{(t_i)}\|_2^2 - \|f^{(t_i)}\|_2^2| > \epsilon \|f^{(t_i)}\|_2^2 \right\}.$$

Observe that for each $t_{i-1} < t \leq t_i$, $||\|\tilde{f}^{(t)}\|_2^2 - \|f^{(t_i)}\|_2^2| > 2\epsilon \cdot \|f^{(t)}\|_2^2$ would imply $\neg E_i$. Namely, $\neg \cup_{i \in [w]} E_i$ implies strong tracking.

By the $\ell_2$ $(\epsilon, \delta)$-weak tracking property of the streaming algorithm, for each $i \in [w]$, we have $\Pr[E_i] \leq \delta$ and thus $\Pr[\cup_{i \in [w]} E_i] \leq w\delta$. We conclude that the streaming algorithm provides $\ell_2$ $(2\epsilon, w\delta)$-strong tracking. ◀

## B.2 Strong tracking lower bound for AMS sketch

The hard instance is simply the stream of all distinct elements, *i.e.*, $i_t = t$ for all $t \in [m]$.

**Proof of Theorem 16.** Consider the stream of all distinct elements as the hard instance, *i.e.*, $i_t = t$ for all $t \in [m]$. Thus, $\|f^{(t)}\|_2^2 = t$ and $\|\Pi f^{(t)}\|_2^2 = \sum_{i \in [k]} \left( \sum_{j \in [t]} \Pi_{i,j} \right)^2$ for all $t \in [m]$.

Define a sequence of time $\{t_j\}$ as follows. $t_0 = 0$ and $t_j = \sum_{i\in[j]} \Delta_i$ where $\Delta_i = \lceil 10/\epsilon \rceil^i$. Pick $\ell$ and $m$ properly such that $t_\ell \leq m$. Some quick facts about the choice of parameters here: (i) $|t_j - \Delta_j| \leq \frac{\epsilon}{5}\cdot t_j$. (ii) $\ell = \Theta(\frac{\log m}{\log(1/\epsilon)})$.

To show AMS sketch does not provide $(\epsilon,\delta)$-strong tracking for $\epsilon \in (0, 0.1)$ and $\delta \in (0, 1)$, it suffices to show that with probability at least $\delta$ there exists $j \in [\ell]$ such that $\|\Pi f^{(t_j)}\|_2^2 - t_j > (1+\epsilon)\cdot t_j$.

For the convenience of the analysis, for any $i \in [k]$ and $j \in [\ell]$, let $X_i^{(t_j)} = \sum_{s=t_{j-1}+1}^{t_j} \Pi_{i,s}$ which is the sum of $\Delta_j$ independent Rademacher random variables divided by $\sqrt{k}$. Also let $Z_j = \sum_{i\in[k]}(X_i^{(t_j)})^2$. Note that $\mathbb{E}[Z_j] = \Delta_j/\sqrt{k}$ and

$$\begin{aligned}
\|\Pi f^{(t_j)}\|_2^2 &= \sum_{i\in[k]}\left(\sum_{j'\in[j]} X_i^{(t_{j'})}\right)^2 \\
&= Z_j + \sum_{i\in[k]}\left(\sum_{j'\in[j-1]} X_i^{(t_{j'})}\right)^2 + 2\sum_{i\in[k]}\langle X_i^{(t_j)}, \sum_{j'\in[j-1]} X_i^{(t_{j'})}\rangle. \qquad (7)
\end{aligned}$$

Define an event $E_j := \{Z_j \geq (1+2\epsilon)\cdot\mathbb{E}[Z_j]\}$ for each $j\in[\ell]$. Observe that when conditioning on $\cap_{j'\in[j-1]}\neg E_{j'}$, the second term of Equation 7 is bounded by $O(t_{j-1})$ and the third term is bounded by $O(\sqrt{t_{j-1}Z_j})$ due to Cauchy-Schwarz. By the choice of parameters, both term can be bounded by $0.1t_j$. Furthermore, $E_j$ implies $\|\Pi f^{(t_j)}\|_2^2 - t_j > (1+\epsilon)\cdot t_j$. Note that $E_j$ is independent to $E_1,\dots,E_{j-1}$. The following lemma lower bound the probability of $E_j$ to happen.

▶ **Lemma 18.** *There exists a constant $c > 0$ such that $\Pr[E_j] \geq e^{-c\epsilon^2 k}$ for any $j = \Omega(\log\log k)$.*

**Proof of Lemma 18.** From the seminal *Berry-Esseen theorem* [2, 9], we know that when $t_j = e^{\Omega(k)} = \Omega(\frac{\log m}{\delta})$ then $X^{(t_j)}$ is point-wisely $e^{-\Omega(k)}$-close to a normal distribution with zero mean and variance $\Delta_j$. That is, $\frac{kZ_j}{\Delta_j}$ is also point-wisely $e^{-\Omega(k)}$-close to a *chi-square* distribution $\chi^2_{\Delta_j}$ with mean $\Delta_j$ and $\Delta_j$ degree of freedom[10].

Inglot and Ledwina [13] showed that the tail of chi-square random distribution can be lower bounded as $\Pr[\chi_k^2 \geq (1+2\epsilon)\cdot k] \geq \frac{1}{2}e^{-\epsilon^2 k/10}$ when $k$ large enough. Combine with the Berry-Esseen theorem, we have $\Pr[E_j] \geq e^{-c\epsilon^2 k}$ for some constant $c > 0$. ◀

Note that as $\{Z_j\}_{j\in[\ell]}$ are mutually independent, the events $\{E_j\}_{j\in[\ell]}$ are also mutually independent. That is,

$$\begin{aligned}
\Pr\left[\exists t\in[m],\ \left|\|\Pi f^{(t)}\|_2^2 - \|f^{(t)}\|_2^2\right| > 2\epsilon\|f^{(t)}\|_2^2\right] &\geq \Pr\left[\cup_{j\in[\ell]}E_j\right] \\
&\geq 1 - \prod_{j\in[\ell]}\Pr\left[\neg E_j \mid \neg E_{j'},\ \forall j'\in[j-1]\right] \\
&\geq 1 - \left(1 - e^{-c\epsilon^2 k}\right)^\ell \geq \ell e^{-c\epsilon^2 k}.
\end{aligned}$$

◀

Namely, there exists another constant $C > 0$ such that if $k < C\epsilon^{-2}\left(\log\frac{\log m}{\log(1/\epsilon)} + \log(1/\delta)\right) \leq \frac{1}{c}\epsilon^{-2}\log\frac{\ell}{\delta}$. Thus, AMS sketch does not provide $(\epsilon,\delta)$-strong tracking for all $\epsilon \in (0, 0.1)$.

[10] Recall that a *chi-square random variable* of $d$ degree of freedom is equivalent to the sum of $d$ squares of the standard normal random variable.

## B.3 Strong tracking lower bound for CountSketch

To prove Theorem 17, we are going to construct a stream such that any CountSketch does not provide strong tracking. Let's start from some observation. For any $i \neq i' \in [n]$ and $a > 0$, let $\mathbf{x} = a(\mathbf{e}_i + \mathbf{e}_{i'})$ such that $\|\mathbf{x}\|_2^2 = 2a^2$. Now, observe that If $\Pi_i = \Pi_{i'}$, then we have $\|\Pi\mathbf{x}\|_2^2 = 4a^2$. If $\Pi_i = -\Pi_{i'}$, then we have $\|\Pi\mathbf{x}\|_2^2 = 0$. Note that in both cases, the approximation $\|\Pi\mathbf{x}\|_2^2$ and the correct answer $\|\mathbf{x}\|_2^2$ has a huge gap $2a^2$, *i.e.*, $\big|\|\Pi\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\big| \geq \|\mathbf{x}\|_2^2$.

With the above observation, one can see that a collision (either $\Pi_i = \Pi_{i'}$ or $\Pi_i = -\Pi_{i'}$) is a sufficient condition for an estimation error. As a result, to show CountSketch does not provide strong tracking, it suffices to show the following two things: (i) there will be some collision with constant probability and (ii) construct a stream such that once a collision happens, the estimation error is large.

Note that (ii) is very specific to tracking since unlike $\ell_2$ estimation which only cares about the final estimation, we need to keep track of the estimation at any time. Thus, to show the impossibility of tracking, we have to show that the estimation fails at least once at some point.

**Proof of Theorem 17.** Let $n$ be the number of elements and $k$ be the number of rows of CountSketch. Let $\Delta = \lceil 100/\epsilon \rceil$ and $w = \lceil 1/\epsilon \rceil$. For any $j \in [\ell]$, define $t_j = \sum_{j' \in [j]} \Delta^{j'+1} = \frac{\Delta^{j+1} - \Delta^1}{\Delta - 1}$ and the stream at time $t_j$ as follows.

$$f^{(t_j)} = \left( \underbrace{\Delta, \ldots, \Delta}_{w}, \underbrace{\Delta^2, \ldots, \Delta^2}_{w}, \underbrace{\Delta^j, \ldots, \Delta^j}_{w}, 0, \ldots, 0 \right).$$

We have $\|f^{(t_j)}\|_2^2 = \sum_{j' \in [j]} w \cdot \Delta^{2j'+1} = \frac{w \cdot \Delta^{2j+2} - w \cdot \Delta^2}{\Delta^2 - 1}$. Note that one can easily complete rest of the stream $\{f^{(t)}\}_{t \in [m]}$ for any $m \geq t_\ell$. Note that here we can pick $\ell = \Theta(\frac{\log m}{\log(1/\epsilon)})$.

Define the event $E_j := \{\|\Pi f^{(t_j)}\|_2^2 - \|f^{(t_j)}\|_2^2 > \epsilon \cdot \|f^{(t_j)}\|_2^2\}$. To show that CountSketch does not provide $w_2$ $(\epsilon, \delta)$-strong tracking, it suffices to prove $\Pr[\cup_{j \in [\ell]} E_j] > \delta$. The following lemma lower bounds the probability of single $E_j$.

▶ **Lemma 19.** *For each $j \in \ell$, we have* $\Pr[E_j \mid \neg \cup_{j' \in [j]} E_{j'}] \geq \frac{1}{10k\epsilon^2}$.

**Proof.** First, let $\bar{f}^{(t_j)} = f^{(t_j)} - f^{(t_{j-1})}$ for each $j \in \ell$ where we define $f^{(0)} = \mathbf{0}$. Observe that

$$\begin{aligned}
\|\Pi f^{(t_j)}\|_2^2 - \|f^{(t_j)}\|_2^2 &= \|\Pi \bar{f}^{(t_j)} + \Pi f^{(t_{j-1})}\|_2^2 - \|\bar{f}^{(t_j)} + f^{(t_{j-1})}\|_2^2 \\
&= \|\Pi \bar{f}^{(t_j)}\|_2^2 - \|\bar{f}^{(t_j)}\|_2^2 + \|\Pi f^{(t_{j-1})}\|_2^2 - \|f^{(t_{j-1})}\|_2^2 \\
&\quad + 2\langle \Pi \bar{f}^{(t_j)}, \Pi f^{(t_{j-1})} \rangle - 2\langle \bar{f}^{(t_j)}, f^{(t_{j-1})} \rangle.
\end{aligned}$$

Further, condition on $\neg \cup_{j' \in [j-1]} E_{j'}$, we have $\|f^{(t_{j-1})}\|_2^2$, $\|\Pi f^{(t_{j-1})}\|_2^2$, $|\langle \Pi \bar{f}^{(t_j)}, \Pi f^{(t_{j-1})} \rangle|$, and $|\langle \bar{f}^{(t_j)}, f^{(t_{j-1})} \rangle|$ are all at most $(\epsilon/10) \cdot \|f^{(t_j)}\|_2^2$ by the choice of $\Delta$. Namely,

$$\|\Pi f^{(t_j)}\|_2^2 - \|f^{(t_j)}\|_2^2 \geq \|\Pi \bar{f}^{(t_j)}\|_2^2 - \|\bar{f}^{(t_j)}\|_2^2 - \frac{\epsilon}{2} \cdot \|f^{(t_j)}\|_2^2. \tag{8}$$

▶ **Lemma 20.** $\Pr\left[\|\Pi \bar{f}^{(t_j)}\|_2^2 - \|\bar{f}^{(t_j)}\|_2^2 > 3\epsilon \cdot \|f^{(t_j)}\|_2^2\right] > \frac{1}{10k\epsilon^2}$.

**Proof.** Let us consider the columns of $\Pi$ that correspond to the non-zero entries of $\bar{f}^{(t_j)}$. That is, column $\Delta\cdot(j-1)+1$ to $\Delta\cdot j$. Note that once there are exactly one collision happens among these columns and the both the value are the same, then $\|\Pi\bar{f}^{(t_j)}\|_2^2-\|\bar{f}^{(t_j)}\|_2^2>3\epsilon\cdot\|f^{(t_j)}\|_2^2$. The probability of the above to happen is at least the following.

$$\frac{1}{2}\cdot\frac{k\cdot\binom{w}{2}\cdot(k-1)\cdot(k-2)\cdots(k-w+2)}{k^w}\geq\frac{w^2}{5k}>\frac{1}{10k\epsilon^2}.$$ ◀

Now, Lemma 19 immediately follows from Equation 8 and Lemma 20. ◀

Let us wrap up the proof of Theorem 17 as follows.

$$\begin{aligned}\Pr\left[\exists t\in[m],\ \left|\|\Pi f^{(t)}\|_2^2-\|f^{(t)}\|_2^2>\epsilon\|f^{(t)}\|_2^2\right|\right]&\geq\Pr\left[\cup_{j\in[\ell]}E_j\right]\\&=\prod_{j\in[\ell]}\Pr\left[E_j\mid\neg\cup_{j'\in[j-1]}E_{j'}\right]\\&\geq\left(1-\frac{1}{10k\epsilon^2}\right)^{\ell}\geq1-\frac{\ell}{k\epsilon^2}.\end{aligned}$$

By the choice of parameters, the last quantity would be greater than $\delta$ and thus CountSketch with $k\leq C\cdot\epsilon^{-2}\delta^{-1}\frac{\log(m)}{\log(1/\epsilon)}$ rows does not provide $\ell_2$ $(\epsilon,\delta)$-strong tracking. ◀