Precision vs Confidence Tradeoffs for ℓ_2-Based Frequency Estimation in Data Streams

Sumit Ganguly

Indian Institute of Technology

Abstract. We consider the data stream model where an n-dimensional vector x is updated coordinate-wise by a stream of updates. The frequency estimation problem is to process the stream in a single pass and using small memory such that an estimate for x_i for any i can be retrieved. We present the first algorithms for ℓ_2-based frequency estimation that exhibit a tradeoff between the precision (additive error) of its estimate and the confidence on that estimate, for a range of parameter values. We show that our algorithms are optimal for a range of parameters for the class of matrix algorithms, namely, those whose state corresponding to a vector x can be represented as Ax for some $m \times n$ matrix A. All known algorithms for ℓ_2-based frequency estimation are matrix algorithms.

1 Introduction

The problem of estimating frequencies is one of the most basic problems in data stream processing. It is used for tracking heavy-hitters in low space and real time, for example, finding popular web-sites accessed, most frequently accessed terms in search-engines, popular sale items in supermarket transaction database, etc.. In the general *turnstile* data streaming model, an n-dimensional vector x is updated by a sequence of update entries of the form (i, v). Each update (i, v) transforms $x_i \leftarrow x_i + v$. The frequency estimation problem is to design a data structure and an algorithm $\mathcal{A}$ that (i) processes the input stream *in a single pass using as little memory as possible*, and, (ii) given any $i \in [n]$, uses the structure to return an estimate $\hat{x}_i$ for x_i satisfying, $|\hat{x}_i - x_i| \le Err_{\mathcal{A}}$, with confidence $1 - \delta$, where, C is a space parameter of $\mathcal{A}$ and $Err_{\mathcal{A}}$ denotes the *precision* or the additive error of the estimation. We consider frequency estimation algorithms whose error guarantees are in terms of the ℓ_2-norm. The COUNTSKETCH algorithm by Charikar et. al. [1] is the most well-known ℓ_2-based frequency estimation and has precision $Err_{\mathsf{CSK}} = \|x^{\mathrm{res}(C)}\|_2/\sqrt{C}$ and confidence $1-n^{-\Omega(1)}$. Here, $\|x^{\mathrm{res}(C)}\|_2$ is the second norm of x calculated after removing the top-C absolute frequencies from it. The residual norm is often smaller than the standard norm, since in many scenarios, much of the energy of x may concentrate in the top few frequencies.

Precision-Confidence Trade-offs. Let us associate with a randomized estimation algorithm $\mathcal{A}$ running on an input x, a pair of numbers namely, (1) its

Work	Precision	Failure Probability	Space O(words)	Update time $O(\cdot)$	Estimation Time $O(\cdot)$
COUNTSKETCH[1]	$\|x^{\mathrm{res}(C)}\|_2/\sqrt{C}$	$n^{-\Omega(1)}$	$C \log n$	$\log n$	n
ACSK-I $4 \le d \le O(\log n)$	$\|x^{\mathrm{res}(C)}\|_2$ $\times\sqrt{d/(C\log n)}$	$n^{-\Omega(1)}$ $+2^{-d}$	$(C+\log n)$ $\times \log n$	$\log^2 n$	n
ACSK-II $4 \le d \le O(\log(n/C))$	$\|x^{\mathrm{res}(C)}\|_2$ $\times\sqrt{d/(C\log(n/C))}$	$1/16$ $+2^{-d}$	$(C+\log n)$ $\times \log \frac{n}{C}$	$\log^{O(1)} n$	$C\times$ $\log^{O(1)} n$

Fig. 1. Precision-Confidence tradeoffs for ℓ_2-based frequency estimation. For ACSK-I and ACSK-II , the parameter $d \ge 4$ controls the precision-confidence tradeoff.

precision $Err_{\mathcal{A}}(x)$, and (2) the confidence denoted $1-\delta_{\mathcal{A}}$ with which the precision holds. We say that $\mathcal{A}$ exhibits a precision-confidence tradeoff if for each fixed input x, the set of feasible non-dominating $(Err_{\mathcal{A}}(x), \delta_{\mathcal{A}})$ pairs is at least 2 and preferably, is a large set. A point $(Err_{\mathcal{A}}(x), \delta_{\mathcal{A}})$ dominates $(Err'_{\mathcal{A}}(x), \delta'_{\mathcal{A}})$ if $Err_{\mathcal{A}}(x) < Err'_{\mathcal{A}}(x)$ and $\delta_{\mathcal{A}}(x) < \delta'_{\mathcal{A}}(x)$. For example, COUNTSKETCH has the single point $(\|x^{\mathrm{res}(C)}\|_2/\sqrt{C}, 1-n^{-\Omega(1)})$ and does not exhibit a tradeoff. Why are algorithms with precision-confidence tradeoffs useful? To illustrate, suppose that an application requires frequency estimation of items in some input set H of a-priori unknown size t with high constant probability. Using Algorithm ACSK-I (see Figure 1) with $d = \log(t) + O(1)$ gives a precision of $\|x^{\mathrm{res}(C)}\|_2\sqrt{\log t/(C\log n)}$ and confidence of $1 - t2^{-c\log t} = 1 - t^{1-c}$. If $t = O(1)$, the precision is superior to that of COUNTSKETCH by a factor of $\sqrt{\log n}$. If $t = n$ this matches the COUNTSKETCH guarantees. The important property is that *no changes or re-runs of the algorithm are needed*. The same output simultaneously satisfies all the precision-confidence pairs in its tradeoff set.

Contributions. We present a frequency estimation algorithm ACSK-I (Averaged CountSKetch-I) that has precision $O(\|x^{\mathrm{res}(C)}\|_2\sqrt{d/(C\log n)})$ and confidence $1-2^{-d}$, where, $4 \le d \le \Theta(\log n)$. A second frequency estimation algorithm ACSK-II has precision $O(\|x^{\mathrm{res}(C)}\|_2\sqrt{d/(C\log(n/C))})$ and confidence $1-2^{-d}$. Both algorithms show precision-confidence tradeoff by tuning the value of d in the allowed range. Figure 1 compares the algorithms along different measures. We also show that the algorithms are optimal up to constant factors for a wide range of the parameters among the class of algorithms whose state on input x can be represented as Ax, for some $m \times n$ matrix A.

Summary. We build on the COUNTSKETCH algorithm of Charikar et.al. in [1]. Instead of taking the median of estimates for x_i from the individual tables, we take the averages over the estimates for x_i from those tables where a set of heavy-hitters do not collide with i. The analysis uses the $2d$th moment method which requires $O(d)$-wise independence of the random variables. This degree of independence d parameterizes the precision-confidence tradeoff.

2 The ACSK Algorithms

Notation. Let COUNTSKETCH(C, s) denote the structure consisting of s hash tables $T_1, \ldots, T_s$, each having $8C$ buckets, using independently chosen pair-wise independent hash functions $h_1, \ldots, h_s$ respectively. The bucket $T_l[b]$ is the sketch: $T_l[b] = \sum_{h_l(i)=b} x_i \xi_{il}$, where the family $\{\xi_{il}\}_{i \in [n]}$ for each $l \in [s]$ is four-wise independent and the families use independent seeds across the tables. The estimated frequency is the median of the table estimates, that is, $\hat{x}_i = \mathit{median}_{l=1}^{s} T_l[h_l(i)]\xi_{il}$. Then, $|\hat{x}_i - x_i| \leq \|x^{\text{res}(C)}\|_2/\sqrt{C}$, with probability $1 - 2^{-\Omega(s)}$.

For an n-dimensional vector x and $H \subset [n]$, let x_H denote the sub-vector of x with coordinates in H.

The ACSK-I (C, s_0, s, d) structure with space parameter C, number of tables parameters s_0 and s, and degree of independence parameter d, maintains two structures, namely, (1) COUNTSKETCH$(2C, s_0)$, where, $s_0 = c \log n$ for some constant $c > 0$, and, (2) COUNTSKETCH(C', s), where, $C' = \lceil 3eC \rceil$, that uses (a) $2d + 1$-wise independent Rademacher families $\{\xi_{il}\}_{i \in [n]}$ for each $l \in [s]$, and, (b) the hash functions $h_1, \ldots, h_s$ corresponding to the tables $T_1, \ldots, T_s$ are independently drawn from a $d + 3$-wise independent hash family that maps $[n]$ to $[C']$. Both structures are updated as in the classical case. The frequency estimation algorithm is as follows.

1. Use the first COUNTSKETCH structure to obtain a set H of the top-$2C$ items by absolute values of their estimated frequencies (by making a pass over $[n]$).
2. Let $S(i, H)$ be the set of table indices in the second COUNTSKETCH structure where i does not collide with any item in $H \setminus \{i\}$. Return the average of the estimates for x_i obtained from the tables in $S(i, H)$.

$$\hat{x}_i = \text{average}_{l \in S(i,H)} T_l[h_l(i)] \cdot \xi_{il} \ .$$

Analysis. Let x'_i denote the estimated frequency obtained from the first structure. By property of precision of COUNTSKETCH[1] we have, $|x'_i - x_i| \leq \Delta$, where, $\Delta = \|x^{\text{res}(2C)}\|_2/\sqrt{2C}$. Let GoodH denote the event GoodH $\equiv \forall i \in [n], |x'_i - x_i| \leq \Delta$. So by union bound, $\Pr[\text{GoodH}] \geq 1 - n2^{-\Omega(s_0)}$. We first prove simple upper bounds for (a) the maximum frequency of an item in $\bar{H}$, and, (b) $\|x^{\text{res}(H)}\|_2^2 = \sum_{j \in [n] \setminus H} x_j^2$. Let T_H denote the maximum absolute frequency of an item not in H. Lemma 1 (a) is proved in Appendix A. Lemma 1 (b) follows variants proved in [3, 2].

Lemma 1. *Conditional on GoodH, (a)* $T_H \leq (1 + \sqrt{2})\|x^{res(C)}\|_2/\sqrt{C}$*, and, (b)* $\|x^{res(H)}\|_2^2 \leq 9\|x^{res(2C)}\|_2^2$.

Consider the second COUNTSKETCH structure of ACSK-I. Let $p = 1/(8C') = 1/(8\lceil 3eC \rceil) \leq 1/(24eC)$, which is the probability that a given item maps to a given bucket in a hash table. For $i, j \in [n], j \neq i$ and table index $l \in [s]$, let χ_{ijl} be 1 if $h_l(i) = h_l(j)$ and 0 otherwise. Lemma 2 shows that given sufficient independence of the hash functions, $S(i, H) = \Theta(s)$ with high probability.

Lemma 2. *Suppose the hash functions $h_1, \ldots, h_s$ of a* COUNTSKETCH *structure are each chosen from a pair-wise independent family. Let $C' \geq \lceil 1.5et \rceil + 1$. Then, for any given set H with $|H| = t$, $|S(i, H)| \geq 3s/5$ with probability $1 - e^{-s/3}$.*

Lemma 3 presents an upper bound on the $2d$th moment for the sum of $2d$-wise independent random variables, each with support in the interval $[-1, 1]$ and having a symmetric distribution about 0. Its proof, given in the Appendix, uses ideas from the proof of Theorem 2.4 in [6] but gives a slightly stronger result in comparison.

Lemma 3. *Suppose $X_1, X_2, \ldots, X_n$ are $2d$-wise independent random variables such that the X_i's have support in the interval $[-1, 1]$ and have a symmetric distribution about 0. Let $X = X_1 + X_2 + \ldots + X_n$. Then,*

$$E\big[X^{2d}\big] \leq \sqrt{2} \left(\frac{2d\mathsf{Var}\,[X]}{e} \right)^d \left(1 + \frac{d}{\mathsf{Var}\,[X]} \right)^{d-1} .$$

For a suitable normalization value T_1 and $j \in [n] \setminus (H \cup \{i\})$, let $X_{ijl} = (x_j/T_1)\xi_{jl}\xi_{il}\chi_{ijl}$ and let

$$X_i = (\hat{x}_i - x_i)|S(i, H)|/T_1 = \sum_{l \in S(i,H)} \sum_{j \notin H \cup \{i\}} X_{ijl} .$$

We wish to calculate $\mathbb{E}\big[X_i^{2d}\big]$ and use it to obtain a concentration of measure for X_i. However, the X_{ijl}'s contributing to X_i are conditioned on the event that $l \in S(i, H)$, a direct application of Lemma 3 is not possible. Lemma 4 gives an approximation for $\mathbb{E}\big[X_i^{2d}\big]$ in terms of $E[X_i^{2d}]$, where, $E[X_i^{2d}]$ is the $2d$th moment of the same random variable but under the assumption that the ξ_{jl}'s and the hash functions h_l's for each l are fully independent.

Lemma 4. *Let $C' = \lceil 3eC \rceil$, the h_l's be $d + 1 + t$-wise independent, $t \geq 2$ and $\{\xi_{il}\}_{i \in [n]}$ be $2d + 1$-wise independent. Then $\mathbb{E}\big[X_i^{2d}\big] \leq (1 + 8(12t)^{-t})^d E[X_i^{2d}]$.*

The proof of Lemma 4 requires the following Lemma 5, which is an application of the principle of inclusion-exclusion and Bayes' rule.

Lemma 5. *For any $s \geq 1$ and $t \geq 2$, let $X_1, \ldots, X_n$ be $s + t$-wise independent and identically distributed Bernoulli (i.e., 0/1) random variables with $t \geq 2$ and $p = \mathsf{Pr}\big[X_i = 1\big] \leq 1/(12e)$. Then, for disjoint sets $S, H \subset [n]$, with $|S| = s$ and $|H| \leq 1/(12pe)$, $\big|\mathsf{Pr}\big[\forall j \in S, X_j = 1 \mid \forall j \in H, X_j = 0\big] - p^s\big| \leq 8(12t)^{-t}$.*

The proof of Lemma 5 is given in the Appendix. We can now prove Lemma 4.

Proof (Of Lemma 4.).

$$\mathbb{E}\big[X_i^{2d}\big] = \mathbb{E}\Bigg[\Big(\sum_{l \in S(i,H), j \neq i} (x_j/T_1)\xi_{jl}\xi_{il}\chi_{ijl} \Big)^{2d}\Bigg]$$

$$= \sum_{\substack{\sum_{l \in S(i,H), j \neq i} e_{jl} = 2d \\ e_{jl}\text{'s even}}} \binom{2d}{e_{11}, \ldots, e_{ns}} \prod_{l \in S(i,H)} \mathbb{E}\Big[\prod_{j : e_{jl} > 0} (x_j/T_1)^{e_{jl}} \chi_{ijl} \mid l \in S(i, H) \Big]$$

Let e denote the vector $(e_{11}, \ldots, e_{ns})$ that satisfies the constraints in the summation, that is, (1) $\sum_{l \in S(i,H), j \neq i} e_{jl} = 2d$, (2) $e_{jl} = 0$ for each $l \in [s] \setminus S(i,H), j \in [n]$, and, (3) each e_{jl} is even. Let $S_{ile} = \{j : e_{jl} > 0\}$. Define the events:

$$E_1(i,l,e): \ \forall j \in S_{ile}, \chi_{ijl} = 1 \quad \text{and} \quad E_2(i,l,H): \ \forall j \in H \setminus \{i\}, \ \chi_{ijl} = 0.$$

Then,

$$\mathbb{E}\Big[\prod_{j: e_{jl} > 0} \chi_{ijl} \mid l \in S(i,H)\Big] = \mathsf{Pr}\big[E_1(i,l,e) \mid E_2(i,l,H)\big] \ .$$

Since the product is taken over positive e_{jl}'s, for each such l, S_{ile} is non-empty. A bound on $\mathsf{Pr}\big[E_1(i,l,e)) \mid E_2(i,l,H)\big]$ can now be obtained using Lemma 5, where, $p = \mathsf{Pr}\left[\chi_{ijl} = 1\right] = 1/C' \leq 1/(24eC)$. Further, $|S_{ile}| \leq d$ and $|H| = 2C \leq 1/(12pe)$. So the premises of Lemma 5 are satisfied. Also, since the hash function h_l is drawn from a $d+1+t$-wise independent family, the family of random variables $\{\chi_{ijl} : j \in [n], j \neq i\}$, for each fixed i and l, is $d+t$-wise independent, and across the l's is fully independent. Applying Lemma 5, we obtain $\mathsf{Pr}\big[E_1(i,l,e) \mid E_2(i,l,H)\big] \in p^{|S_{ile}|}(1 \pm 8(12t)^{-t})$. Hence,

$$\begin{aligned}
&\mathbb{E}\big[X_i^{2d}\big] \\
&\leq \sum_{\substack{\sum_{l \in S(i,H), j \neq i} e_{jl} = 2d \\ e'_{jl} s \text{ even}}} \binom{2d}{e_{11} \ldots e_{ns}} \prod_{l \in S(i,H)} \Big[\big(p^{|S_{ile}|}(1 + 8(12t)^{-t})\big) \prod_{j: e_{jl} > 0} (x_j/T_1)^{e_{jl}}\Big] \\
&\leq (1 + 8(12t)^{-t})^d \sum_{\substack{\sum_{l \in S(i,H), j \neq i} e_{jl} = 2d \\ e_{jl}\text{'s even}}} \binom{2d}{e_{11} \ldots e_{ns}} \prod_{l \in S(i,H)} p^{|S_{ile}|} \prod_{j: e_{jl} > 0} (x_j/T_1)^{e_{jl}} \\
&\leq (1 + 8(12t)^{-t})^d E[X_i^{2d}]
\end{aligned}$$

since, the *RHS*, discounting the multiplicative factor of $(1 + 8(12t)^{-t})^d$, is the expansion of $E[X_i^{2d}]$. □

We now prove the main theorem regarding the ACSK-I algorithm.

Theorem 6. *For $C \geq 2$, $s_0 = \Theta(\log n)$ and $s \geq 20d$, there is an algorithm that for any $i \in [n]$ returns $\hat{x}_i$ satisfying $|\hat{x}_i - x_i| \leq \|x^{res(C)}\|_2 \sqrt{3d/(sC)}$ with probability at least $1 - 2^{-\Omega(s)} - 2^{-d} - n2^{-s_0}$. Moreover, $\mathbb{E}\big[\hat{x}_i\big] = x_i$. The algorithm uses space $O(C(s + s_0))$ words.*

Proof. Consider the ACSK-I algorithm. For $l \in [s]$, $\mathbb{E}\big[T_l[h_l(i)] \cdot \xi_{il}\big] = x_i$. Hence the average of $T_l[h_l(i)] \cdot \xi_{il}$'s over some subset of the l's has the same expectation.

Fix $i \in [n]$. Let $T_1 \geq T_H$ which will be chosen later. Recall that for $j \in [n] \setminus (H \cup \{i\})$ and $l \in S(i,H)$, $X_{ijl} = (x_j/T_1)\xi_{il}\xi_{jl}\chi_{ijl}$. Since, $j \notin H$, $|X_{ijl}| \leq 1$ and X_{ijl} has 3-valued support $\{-x_j/T_1, 0, x_j/T_1\}$ with a symmetric distribution over it. Let $p = \mathsf{Pr}\big[\chi_{ijl} = 1\big] = 1/(8C') = 1/(24eC)$. By direct calculation,

$$\mathsf{Var}\left[X_i\right] = \sum_{l \in S(i,H)} \sum_{j \neq i} \Big(\frac{x_j}{T_1}\Big)^2 p = |S(i,H)| \frac{\|x^{\mathrm{res}(H \cup \{i\})}\|_2^2}{24eCT_1^2} \qquad (1)$$

By Lemma 3 and assuming full independence we have,

$$E[X_i^{2d}] \le \sqrt{2}\left(\frac{2d\mathsf{Var}\,[X_i]}{e}\right)^d \left(1 + \frac{2d}{9\mathsf{Var}\,[X_i]}\right)^{d-1} .$$

Let $t = 2$. Sine the hash functions are $d + 3 = d + t + 1$-wise independent and the Rademacher variables are $2d + 1$-wise independent, by Lemma 4 we have,

$$\mathbb{E}\big[X_i^{2d}\big] \le (1 + 8(12t)^{-t})^d E[X_i^{2d}] \le (1 + 1/72)^d E[X_i^{2d}], \quad \text{for } t = 2.$$

By $2d$th moment inequality, $\mathsf{Pr}\left[|X_i| > \sqrt{2}\big(\mathbb{E}\big[X_i^{2d}\big]\big)^{1/(2d)}\right] \le 2^{-d}$. Therefore,

$$\mathsf{Pr}\left[|X_i| > \sqrt{2(1+1/72)}\left(\frac{2d\mathsf{Var}\,[X_i]}{e}\Big(1 + \frac{d}{\mathsf{Var}\,[X_i]}\Big)\right)^{1/2}\right] \le 2^{-d} \qquad (2)$$

Let $E_{d,i}$ denote the event whose probability is given in (2). Consider the intersection of the following three events: (1) GoodH, (2) $|S(i,H)| \ge 3s/5$, and, (3) $E_{d,i}$. By union bound, the above three events hold with probability $1 - n2^{-\Omega(s_0)} - e^{-s/3} - 2^{-d} = 1 - \delta$ (say). Since, GoodH holds, we can choose $T_1 = (1+\sqrt{2})\|x^{\mathrm{res}(C)}\|_2/\sqrt{C}$. Then, (1) $T_H \le T_1$, by Lemma 1, and, (2) $\|x^{\mathrm{res}(H\cup\{i\})}\|_2^2 \le 9\|x^{\mathrm{res}(2C)}\|_2^2$, by Lemma 1 (b). Substituting in (1),

$$\mathsf{Var}\,[X_i] \le \frac{|S(i,H)|\|x^{\mathrm{res}(H\cup\{i\})}\|_2^2}{(24eC)T_1^2} \le \frac{s\cdot 9\|x^{\mathrm{res}(2C)}\|_2^2}{(24eC)(1+\sqrt{2})^2(\|x^{\mathrm{res}(C)}\|_2^2/C)} \le \frac{s}{20} \quad (3)$$

The deviation for $|X_i|$ in (2) is an increasing function of $\mathsf{Var}\,[X]$. Hence, replacing $\mathsf{Var}\,[X_i]$ by its upper bound gives us an upper bound on the deviation for the same tail probability. Hence, with probability $1 - \delta$, we have from (2) that

$$|X_i| \le \sqrt{2.5}\big(\tfrac{2ds}{20e}\big(1 + \tfrac{20d}{s}\big)\big)^{1/2} \le \sqrt{\tfrac{ds}{2e}}$$

since, $s \ge 20d$. Since, $|\hat{x}_i - x_i| = |X_i|T_1/|S(i,H)|$, we have,

$$|\hat{x}_i - x_i| \le \sqrt{\frac{ds}{2e}} \cdot \frac{(1+\sqrt{2})\|x^{\mathrm{res}(C)}\|_2}{\sqrt{C}} \cdot \frac{1}{(3s/5)} \le \sqrt{\frac{3d}{sC}}\|x^{\mathrm{res}(C)}\|_2 \ . \quad \square$$

Precision-Confidence Tradeoff. Theorem 6 can be applied using *any value of d in the range* $4 \le d \le s/4 = \Theta(\log n)$ (even after the estimate has been obtained). One can choose d to match the confidence to the desired level and minimize the precision (for e.g., choose $d = O(\log r)$, where r is the number of estimates taken).

The ACSK-II Algorithm. The ACSK-II algorithm uses the heavy-hitter algorithm by Gilbert et. al. in [4], denoted by$\mathsf{HH}^{\text{GLPS}}$, to find the heavy hitters.

Theorem 7 ([4]). *There is an algorithm and distribution on matrices Φ such that, given Φx and a concise description of Φ, the algorithm returns $\hat{x}$ such that $\|x - \hat{x}\|_2^2 \le (1+\epsilon)\|x^{res(C)}\|_2^2$ holds with probability $3/4$. The algorithm runs in time $C\log^{O(1)} n$ and Φ has $O((C/\epsilon)\log(n/C))$ rows.*

The only difference in the ACSK-II (C, s) algorithm is that it uses an $\mathsf{HH}^{\text{GLPS}}(2C, 1/2)$ structure to obtain a set H of heavy-hitters. The second COUNTSKETCH(C', s) structure of ACSK-I , and the estimation algorithm is otherwise identical. Here, $C' = \lceil 6eC \rceil$ and $s = O(\log(n/C))$. ACSK-II has significantly faster estimation time than ACSK-I due to the efficiency of Gilbert et. al.'s algorithm. However its guarantee holds only with high constant probability. We have the following theorem.

Theorem 8. *For each $C \geq 2$, $s \geq 20d$ and $r \geq 1$, there is an algorithm that given any set of distinct indices $i_1, \ldots, i_r$ from $[n]$, returns $\hat{x}_{i_j}$ corresponding to x_{i_j} satisfying $|\hat{x}_{i_j} - x_{i_j}| \leq \|x^{res(C)}\|_2 \sqrt{2d/(C\log(n/C))}$ for all $j \in [r]$, with probability $15/16 - r2^{-d}$. Moreover, $\mathbb{E}[\hat{x}_{i_j}] = x_{i_j}$, $j \in [r]$. The algorithm uses space $O(C\log(n/C))$ words and has update time $O(\log^{O(1)} n)$. The estimation time is $O(C\log^{O(1)}(n) + rCd\log(n))$.*

Proof. It follows from Theorem 7 that $\|x^{\text{res}(H)}\|_2^2 \leq (1+1/2)\|x^{\text{res}(C)}\|_2^2$. Further, the Loop Invariant in [4] ensures that upon termination, (a) the largest element not in H has frequency at most $T_H^2 < \|x^{\text{res}(C)}\|_2^2/C$, and, (b) $|H| = \|\hat{x}\|_0 \leq 4C$. We have upper bounds on all the parameters as needed, and the proof of Theorem 6 can be followed. $\square$

3 Lower Bound on Frequency Estimation

We say that a streaming algorithm has a *matrix representation with m rows* if the state of the structure on any input vector x can always be represented as Ax, where, A is some $m \times n$ matrix. All known data streaming algorithms for ℓ_2-based frequency estimation have a matrix representation. We show a lower bound on the number of rows in the matrix representation of a frequency estimation algorithm.

Theorem 9. *Suppose that a frequency estimation algorithm has a matrix representation with m rows. Let it have precision $\|x^{res(C)}\|_2\sqrt{d/(C\log(n/C))}$ such that for any number r of estimations, all the estimates satisfy the precision with probability $15/16 - r \cdot 2^{-d}$. Then, for $d = \Omega(1)$, $2 + \log C \leq d \leq \log \frac{n}{C}$ and $n = \Omega(C\log(\frac{n}{C})\log(C\log\frac{n}{C}))$, $m = \Omega(C\log(\frac{n}{C}) \cdot (1 - \frac{\log C}{d}))$.*

Proof. Let $D = [2^{d-3}]$ and $C = 4k$. Given a vector x with coordinates in D we make a pass over D and obtain the estimated frequency vector $\hat{x}$. Let H be the set of the top-$2k$ coordinates by absolute values of estimated frequency. Then, $\forall i \in D$, $|\hat{x}_i - x_i| \leq \|x^{\text{res}(4k)}\|_2 \sqrt{\frac{d}{4k\log(n/C)}}$ holds with probability $15/16 - 2^{d-3}2^{-d} > 2/3$. Following the proof of Theorem 3.1 in [5]), the resulting vector satisfies $\|x - \hat{x}_H\|_2^2 \leq \left(1 + \frac{d}{\log(n/C)}\right)\|x^{\text{res}(k)}\|_2^2$. Thus we have an ℓ_2/ℓ_2 k-sparse recovery algorithm with approximation factor $1 + d/\log(n/C)$ that succeeds with probability $2/3$. Since, $n = \Omega(C\log(\frac{n}{C})\log(C\log\frac{n}{C}))$ and

$n = \Omega(C \log^2(n/C)(\frac{1}{d} - \frac{\log(C)}{d}))$, by the Price-Woodruff lower bound for $(1+\epsilon)$-approximate k-sparse recovery [5], such a matrix A has number of rows

$$m = \Omega\left(\frac{k}{\epsilon} \log \frac{2^{d-3}}{k}\right) = \Omega\left(C \log\left(\frac{n}{C}\right) \cdot \left(1 - \frac{\log C}{d}\right)\right) \ . \ \square$$

Clearly, both ACSK algorithms have a matrix representation. Also ACSK-II satisfies the premise regarding precision and confidence of Theorem 9 and uses $O(C \log(n/C))$ rows. ACSK-I does too provided $C = n^{1-\Omega(1)}$. Hence, they are optimal up to constant factors in the range $\frac{d}{100} \leq \log C \leq d-2$ and $d \leq \log \frac{n}{C}$ along with the other constraints of Theorem 9 on d, n and C.

References

1. Moses Charikar, Kevin Chen, and Martin Farach-Colton. "Finding frequent items in data streams". *Theoretical Computer Science*, 312(1):3–15, 2004.
2. Graham Cormode and S. Muthukrishnan. "Combinatorial Algorithms for Compressed Sensing". In *Proceedings of International Colloquium on Structural Information and Communication Complexity (SIROCCO)*, 2006.
3. S. Ganguly, D. Kesh, and C. Saha. "Practical Algorithms for Tracking Database Join Sizes". In *Proceedings of Foundations of Software Technoogy and Theoretical Computer Science (FSTTCS)*, pages 294–305, Hyderabad, India, December 2005.
4. A. C. Gilbert, Y. Li, E. Porat, and M. J. Strauss. "Approximate sparse recovery: optimizing time and measurements". In *Proceedings of ACM Symposium on Theory of Computing STOC*, pages 475-484, 2010.
5. Eric Price and David Woodruff. "$(1+\epsilon)$-approximate Sparse Recovery". In *Proceedings of IEEE Foundations of Computer Science (FOCS)*, 2011.
6. J. Schmidt, A. Siegel, and A. Srinivasan. "Chernoff-Hoeffding Bounds with Applications for Limited Independence". In *Proceedings of ACM Symposium on Discrete Algorithms (SODA)*, pages 331–340, 1993.

A Proofs

Proof (Of Lemma 1). Assume GoodH holds. Let $|x_i| = T_H = \max_{j \notin H} |x_j|$. So if $|x_j| < T_H - 2\Delta$, then, $j \notin H$. Hence, $H \subset J = \{j : x_j \geq T_H - 2\Delta\}$. Now, $|J \setminus \textsc{Top}(C)| \geq |H \setminus \textsc{Top}(C)| \geq C$. Thus,

$$\|x^{\text{res}(C)}\|_2^2 \geq \textstyle\sum_{j \in J \setminus \textsc{Top}(C)} x_j^2 \geq |J \setminus \textsc{Top}(C)|(T_H - 2\Delta)^2 \geq C(T_H - 2\Delta)^2$$

or, $T_H \leq \left(\frac{\|x^{\text{res}(C)}\|_2^2}{C}\right)^{1/2} + 2\Delta = (1+\sqrt{2})\|x^{\text{res}(C)}\|_2/\sqrt{C}$. $\square$

Proof (Of Lemma 2.). Assume $t > 0$, otherwise the lemma trivially holds. Since $8C' \geq 8\lceil 1.5et \rceil \geq 12et$, we have, $\mathsf{Pr}\big[\chi_{ijl} = 1\big] = p = 1/(8C') \leq 1/(12et)$. Let $w = |H \setminus \{i\}|$. Denote by $Pr[\cdot]$ the probability measure under the assumption that the hash functions are fully independent. By inclusion-exclusion applied for $\mathsf{Pr}\big[\bigvee_j (\chi_{ijl} = 1)\big]$ and $Pr[\bigvee_j (\chi_{ijl} = 1)]$ respectively, where, j runs over $H \setminus \{i\}$), $d+1$-wise independence of the hash function h_l for $\mathsf{Pr}[\cdot]$ and using triangle inequality once, we have, $\big|\mathsf{Pr}\big[\bigvee_j \chi_{ijl} = 1\big] - Pr\big\{\bigvee_j \chi_{ijl} = 1\big\}\big| \leq 2\binom{w}{d} p^d$.

Since, $\mathsf{Pr}\big[\bigwedge_j(\chi_{ijl} = 0)\big] = 1 - \mathsf{Pr}\big[\bigvee_j(\chi_{ijl} = 1)\big]$, and $Pr[\bigwedge_j(\chi_{ijl} = 0)] = (1-p)^w$, we have, $\big|\mathsf{Pr}\big[\bigwedge_j \chi_{ijl} = 0\big] - (1-p)^w\big| \leq 2\binom{w}{d}p^d$. Further since $w \leq t$, we have, $\binom{w}{d}p^d \leq (pet/d)^d \leq (12d)^{-d}$. Also $(1-p)^w \geq 1 - tp \geq 1 - 1/(12e)$.

Therefore, $\mathsf{Pr}\big[\bigwedge_j \chi_{ijl} = 0\big] \geq 1 - 1/(12e) - 2(12d)^{-d} \geq 24/25$, for $d \geq 2$. Since the hash functions are independent across the tables, applying Chernoff's bounds, we have, $\mathsf{Pr}\big[|S(i,H)| \geq (3/5)s\big] \geq 1 - \exp\{-s/3\}$. $\square$

Proof (of Lemma 3.). We have, $X_i^{2j} \leq X_i^2$ and so $\mathbb{E}\big[X_i^{2j}\big] \leq \mathbb{E}\big[X_i^2\big]$. Also $\mathsf{Var}\,[X] = \sum_{j=1}^n \mathbb{E}\left[X_i^2\right]$. So for $X = X_1 + \ldots + X_n$, and since all odd moments of X_i's are 0, by symmetry of the individual distributions, we have,

$$\begin{aligned}
\mathbb{E}\big[X^{2d}\big] &= \sum_{r=1}^d \sum_{\substack{t_1+\ldots+t_r=d\\ t_j\text{'s}>0}} \binom{2d}{2t_1,2t_2,\ldots,2t_r} \sum_{1\leq j_1<\ldots<j_r\leq n} \prod_{u=1}^r \mathbb{E}\big[X_{j_u}^{2t_u}\big]\\
&= \sum_{r=1}^d \sum_{\substack{t_1+\ldots+t_r=d\\ t_j\text{'s}>0}} \binom{2d}{2t_1,2t_2,\ldots,2t_r} \sum_{1\leq j_1<\ldots<j_r\leq n} \prod_{u=1}^r \mathbb{E}\big[X_{j_u}^{2}\big]\\
&\leq \sum_{r=1}^d \sum_{\substack{t_1+\ldots+t_r=d\\ t_j\text{'s}>0}} \binom{2d}{2t_1,2t_2,\ldots,2t_r} \frac{(\mathsf{Var}[X])^r}{r!}\\
&= \sum_{l=0}^{d-1} T_l,\ \text{where, } T_l = \sum_{t_1+\ldots+t_{d-l}=d,\ t_j\text{'s}>0} \binom{2d}{2t_1,2t_2,\ldots,2t_{d-l}} \frac{(\mathsf{Var}[X])^{d-l}}{(d-l)!}
\end{aligned}$$

Since $\binom{2d}{2t_1,2t_2,\ldots,2t_{d-l}} \leq \binom{2d}{2,2,\ldots,2}$, we have,

$$\begin{aligned}
T_l &\leq \sum_{t_1+\ldots+t_{d-l}=d,\ t_j\text{'s}>0} \binom{2d}{2,2,\ldots,2} \frac{(\mathsf{Var}[X])^{d-l}}{(d-l)!} \leq \binom{d-1}{d-l-1}\binom{2d}{2,2,\ldots,2} \frac{(\mathsf{Var}[X])^{d-l}}{(d-l)!}\\
&= \binom{d-1}{d-l-1}\left(\frac{1}{\mathsf{Var}[X]}\right)^l \frac{d!}{(d-l)!} T_0
\end{aligned}$$

since, there are $\binom{d-1}{d-l-1}$ assignments for $t_1, \ldots, t_{d-l}$, all positive with sum d. Therefore,

$$\begin{aligned}
\mathbb{E}\left[X^{2d}\right] &\leq \sum_{l=0}^{d-1} T_l \leq \sum_{l=0}^{d-1} \binom{d-1}{d-l-1}\left(\frac{1}{\mathsf{Var}[X]}\right)^l \frac{d!}{(d-l)!} T_0\\
&\leq T_0 \sum_{l=0}^{d-1} \binom{d-1}{l}\left(\frac{1}{\mathsf{Var}[X]}\right)^l d^l = T_0\left(1 + \frac{d}{\mathsf{Var}[X]}\right)^{d-1}
\end{aligned}$$

Since,

$$T_0 = \binom{2d}{2,2,\ldots,2}\frac{(\mathsf{Var}[X])^d}{d!} = \frac{(2d)!}{2^d d!}(\mathsf{Var}\,[X])^d \leq \frac{2^{d+1/2}d^d}{e^d}(\mathsf{Var}\,[X])^d$$

by Stirling's approximation, we have,

$$\mathbb{E}\left[X^{2d}\right] \leq \sqrt{2}\left(\frac{2d\mathsf{Var}[X]}{e}\right)^d \left(1 + \frac{d}{\mathsf{Var}[X]}\right)^{d-1} . \quad \square$$

Proof (Of Lemma 5.). Define events $E_1 \equiv \forall j \in S, X_j = 1$ and $E_2 \equiv \forall j \in H, X_j = 0$. We have to bound the probability $\mathsf{Pr}\big[E_1 \mid E_2\big]$. Let $|H| = w$. Since, $|S| = s$, $\mathsf{Pr}\big[E_1\big] = p^s$. By inclusion and exclusion,

$$\begin{aligned}
&\left|\mathsf{Pr}\big[\exists j \in H, X_j = 1 \mid E_1\big] - \sum_{r=1}^{t-1}(-1)^{r-1} \sum_{\substack{j_1,\ldots,j_r\in H\\ j_1<\ldots<j_r}} \mathsf{Pr}\big[X_{j_1} = 1 \wedge \ldots \wedge X_{j_r} = 1 \mid E_1\big]\right|\\
&\qquad \leq \sum_{\substack{j_1,\ldots,j_t\in H\\ j_1<\ldots<j_t}} \mathsf{Pr}\big[X_{j_1} = 1 \wedge \ldots X_{j_t} = 1 \mid E_1\big]
\end{aligned}$$

Since the X_j's are $s+t$-wise independent and the event E_1 is a property of the X_j's for $j \in S$ and $|S| = s$, we have for distinct elements $j_1, \ldots, j_r$ from H (given $H \cap S$ is empty) and $1 \le r \le t$, $\mathsf{Pr}\big[X_{j_1} = 1 \wedge \ldots X_{j_r} = 1 \mid E_1\big] = \mathsf{Pr}\big[X_{j_1} = 1\big] \cdot \ldots \cdot \mathsf{Pr}\big[X_{j_r} = 1\big] = p^r$. Let $|H| = w$. The above equation is equivalently,

$$\left|\mathsf{Pr}\big[\exists j \in H, X_j = 1 \mid E_1\big] - \sum_{r=1}^{t-1}(-1)^{r-1}\binom{w}{r}p^r\right| \le \binom{w}{t}p^t \tag{4}$$

Suppose we denote by $Pr[E]$ the probability of an event $E = E(X_1, \ldots, X_n)$ assuming that the X_j's are fully independent. Then, by inclusion-exclusion, we have

$$\left|Pr\left[\exists j \in H, X_j = 1 \mid E_1\right] - \sum_{r=1}^{t-1}(-1)^{r-1}\binom{w}{r}p^r\right| \le \binom{w}{t}p^t \tag{5}$$

Since, $\mathsf{Pr}\big[X_j = 1\big] = Pr[X_j = 1] = p$, combining (4) and (5), we have by triangle inequality,

$$\left|\mathsf{Pr}\big[\exists j \in H, X_j = 1 \mid E_1\big] - Pr\left[\exists j \in H, X_j = 1 \mid E_1\right]\right| \le 2\binom{w}{t}p^t$$

Also, $\mathsf{Pr}\big[E_2 \mid E_1\big] = 1 - \mathsf{Pr}\big[\exists j \in H, X_j = 1 \mid E_1\big]$ and $Pr[E_2 \mid E_1] = 1 - Pr[\exists j \in H, X_j = 1 \mid E_1] = (1-p)^w$. Hence,

$$\left|\mathsf{Pr}\big[E_2 \mid E_1\big] - (1-p)^w\right| \le 2\binom{w}{t}p^t \tag{6}$$

Further, $\mathsf{Pr}\big[E_1\big] = Pr[\forall j \in S, X_j = 1] = p^s$. Using $s+t$-wise independence of the X_j's for $j \in H$, we can show similarly that

$$\left|\mathsf{Pr}\big[E_2\big] - (1-p)^w\right| \le 2\binom{w}{s+t}p^{s+t} \ .$$

Combining,

$$\mathsf{Pr}\big[E_1 \mid E_2\big] = \frac{\mathsf{Pr}\big[E_2 \mid E_1\big]\mathsf{Pr}\big[E_1\big]}{\mathsf{Pr}\big[E_2\big]} \in p^s\left(1 \pm \frac{2\binom{w}{t}p^t + 2\binom{w}{s+t}p^{s+t}}{(1-p)^w - 2\binom{w}{s+t}p^{s+t}}\right) \tag{7}$$

Since, $pw \le 1/(12e)$, $(1-p)^w \ge 1 - wp \ge 1 - 1/(12e)$, $\binom{w}{t}p^t \le (wep/t)^t \le 1/(12t)^t$ and $\binom{w}{s+t}p^{s+t} \le 1/(12(s+t))^{s+t}$. Thus, for $t \ge 2$, we have,

$$\frac{2\binom{w}{t}p^t + 2\binom{w}{s+t}p^{s+t}}{(1-p)^w - 2\binom{w}{s+t}p^{s+t}} \le \frac{2(12t)^{-t} + 2(12(s+t))^{-t-s}}{(1 - 1/(12e)) - 2(12(s+t))^{-s-t}} \le 8(12t)^{-t} \ .$$

since $t \ge 2$. Hence, (7) becomes

$$\mathsf{Pr}\big[E_1 \mid E_2\big] \in p^s\big[1 \pm 8(12t)^{-t}\big] \ . \quad \square$$