An algebraic approach to complexity of data stream computations

Sumit Ganguly

Indian Institute of Technology, Kanpur

Abstract. We consider a family of practical problems in the *update* data stream processing model, that is, data streams where arbitrary insertions and deletions of items are allowed. A typical problem is that of approximately estimating item frequencies to within relative error ϵ. This problem has $\tilde{O}(\epsilon^{-1})$ [1] upper and lower randomized space bounds. The current deterministic algorithm for the same problem requires $\tilde{O}(\epsilon^{-2})$ space. In this paper, we derive a deterministic space lower bound of $\tilde{O}(\epsilon^{-2})$. This implies similar $\tilde{O}(\epsilon^{-2})$ deterministic space bounds for a number of related problems.

1 Introduction

A data stream can be viewed as a sequence of records of the form (pos, i, δ), where, *pos* is the current sequence index, i is an item from the domain $[1, n] = \{1, 2, \ldots, n\}$ and $\delta \in \{+1, -1\}$. Here, $\delta = 1$ refers to an insertion of an instance of i and $\delta = -1$ signifies a deletion. The data stream model of processing allows online computations over a stream of data using sub-linear space. For each data item $i \in [1, n]$, its frequency f_i is defined as $f_i = \sum_{(pos,i,\delta) \in \text{ stream}} \delta$. This model is called the *general update model*, where, item frequencies can be positive, negative or zero. A special case of this model is the *strict update model*, where, deletions always correspond to prior insertions. The data stream computation model has proved to be a viable model for a number of application areas, such as networking, databases, etc..

We consider the basic problem of ϵ-accurate frequency estimation over an update data stream: that is, given a data stream and $i \in [1, n]$, return $\hat{f}_i$ such that $|\hat{f}_i - f_i| \leq \epsilon \cdot \ell_1(f)$, where, $\ell_1(f) = \sum_{i \in [1,n]} |f_i|$ is the ℓ_1 norm of the frequency vector f. The problem of frequency estimation is a basic sub-routine and used in the solutions to related problems, for example, approximate quantiles, histograms, estimating range-sums, etc. [4]. The frequency estimation problem has a space lower bound of $\Omega(\epsilon^{-1} \log(n\epsilon))$ [3]. This can be derived as follows. There are $\binom{n}{\lfloor \epsilon^{-1} \rfloor}$ subsets of $[1, n]$ each consisting of $\lfloor \epsilon^{-1} \rfloor$ elements. Corresponding to each such subset one can construct a stream such that elements in the given subset have frequency 1 and the remaining elements have frequency 0. Each of these subsets can be distinguished by an $\frac{\epsilon}{4}$-accurate frequency estimator, since, if i belongs to the set, then, its frequency estimate is at least $3/4$

[1] The $\tilde{O}$ notation suppresses poly-logarithmic factors.

and otherwise, its frequency estimate is at most 1/4. The number of such subsets is $\Omega(\binom{n}{\lfloor \epsilon^{-1} \rfloor}) = \Omega((\epsilon^{-1})\log(n\epsilon))$, proving the lower bound. The randomized COUNT-MIN sketch technique [4] matches this bound up to logarithmic factors. However, the only known deterministic algorithm for this problem requires space that is quadratic in ϵ^{-1}, namely, $O(\epsilon^{-2}(\log^2 n)(\log^{-2} \epsilon^{-1})(\log mn))$ [6], where, $m = \ell_\infty(f) = \max_{i \in [1,n]} |f_i|$. Note that this observation applies to frequency estimation for update streams; for insert-only streams, the algorithms of Misra and Gries [9], Demaine et. al. [5] and Karp et. al. [8] provide deterministic ϵ-approximate frequency estimation using space $O((\epsilon^{-1})(\log(mn)))$ bits.

Contributions. We present a space lower bound of $\Omega(\epsilon^{-2}(\log \epsilon^{-1})^{-1}(\log m)(\log n))$ for deterministic ϵ-accurate frequency estimation over update streams, thus improving the current space lower bound of $\Omega(\epsilon^{-1}(\log n\epsilon)(\log m))$. The result implies analogous and near quadratic (in ϵ^{-1}) deterministic space lower bounds for related problems over update streams, for example, finding approximate frequent items, ϕ-quantiles, approximately optimal b-bucket histograms, estimating inner products, etc..

Organization. The remainder of the paper is organized as follows. In Section 2, we define stream automaton, an abstract machine model for data stream computations and establish basic properties. Section 3 proves lower bounds for stream automaton that estimate frequencies. Finally, we conclude in Section 4.

2 Stream Automaton

Definition 1 (Stream Automaton). *A deterministic stream automaton with space bound parameter $s = s(n, m)$ is a Turing machine with a finite state control that uses two tapes, namely, a* two-way read-write work-tape *and a* one-way read-only input tape*. The input tape is divided into three parts. The first part consists of n and indicates that the domain of items is $[1, n] = \{1, 2, \ldots, n\}$. The second part of the input tape consists of a sequence σ consisting of t records of the form (i, δ), where, t is potentially unbounded. The frequency vector of the stream σ is an element of $\mathbb{Z}^n$ whose ith coordinate $(\mathsf{freq}\ \sigma)_i$ is called the frequency of i and is defined as the sum of the δ's among the records of the form (i, δ). The third part of the input tape contains a query string. During the processing of the query string, the automaton may use any amount of space from the work-tape. Finally, the automaton writes a string as its output on the work-tape. The output of the automaton is a function only of the frequency vector of the entire input stream, the query string and the input n.* □

The *size* of the input stream is defined as the ℓ_∞ norm of the frequency vector of the stream. We say that a stream automaton uses space $s(n, m)$ if for all input streams σ over the domain $[1, n]$ and having size at most m, the number of cells (bits) on the work-tape currently in use, after having processed σ, is bounded by $s(n, m)$. Note that the size of an input stream is not proportional to the number of records in the input sequence. Indeed, a pair of insertion $(i, +1)$ and deletion $(i, -1)$ operations can be repeated any number of times to generate a stream with arbitrary number of records, but whose frequency vector is 0 and therefore its ℓ_∞ norm, or the size, is also 0. Since the sequence defining the stream may consist of arbitrary insertion and deletion operations, the ℓ_∞ norm (or, size) of the frequency vector of the stream may be less or more than the ℓ_∞ norm of some prefix sub-stream.

$s(n, m)$ is an upper bound on the number of cells in use on the work-tape *immediately after* processing a stream of size m over the domain $[1, n]$. The expression $s(n, m)$ does not count the number of cells used by the work-tape while processing a part of the given input stream. It is therefore possible that the work-tape uses more cells while processing a prefix of the input stream whose size is more than m. For naturalness of the metric, we will assume that $s(n, m)$ is monotonic in m, that is, $s(n, m) \leq s(n, m')$, for $m' > m$. The notion of the size of an update stream being defined in terms of its ℓ_∞ norm or in terms of the ℓ_1 norm is standard in the literature on streaming algorithms [4,6].

2.1 Preliminaries and Notation

Fix a value of $n \geq 2$. A stream record of the form $(i, 1)$ is equivalently viewed as e_i and $(i, -1)$ as $-e_i$, where, e_i is the i^{th} standard basis vector of $\mathbb{R}^n$. The set $\Delta = \{e_1, e_2, \ldots, e_n, -e_1, -e_2, \ldots, -e_n\}$ is referred to as the set of basic transition vectors. Any stream can be viewed as a sequence of these vectors. If transition vectors $u_1, u_2, \ldots, u_r$ belonging to Δ appear consecutively on the input tape, the notation $u_1 \circ u_2 \circ \ldots \circ u_r$ is used to denote the sequence. Similarly, the concatenation of two sequences of vectors (or sub-streams) σ and τ is denoted by $\sigma \circ \tau$. The *frequency* of a vector sequence $\sigma = u_1 \circ u_2 \circ \ldots \circ u_r$, where, $u_1, u_2, \ldots, u_r \in \Delta$, is defined as $\mathsf{freq}\ \sigma = u_1 + u_2 + \ldots + u_r$.

The configuration of an automaton is modelled as the triple (q, h, w), where, $q \in Q$ is the current state of the finite control, h is the index of the current cell of the work tape, and w is the current contents of the work-tape. The processing of each record can be viewed as a transition function $\oplus(a, v)$, where, a is the current configuration of A, and v is the current stream record on the input tape of the form $(i, \pm 1)$ and viewed as $\pm e_i$. The transition function $\oplus = \oplus_n$ is in general a function of n. It is written in infix form as $a \oplus v$ with the assumption that the $\oplus$ operator is left-associative.

2.2 Uniform Automaton

Definition 2 (Uniform Automaton). *A stream automaton A is* uniform*, provided, the configuration reached by A after starting from configuration s and after processing input stream σ, is dependent only on* $\mathsf{freq}\ \sigma$ *and s.* □

In this section, we study uniform stream automata. Let $K = (\mathbb{Z}^n, +)$ be the classical module of n-dimensional vectors of integers. Let A be a uniform automaton and let $\mathcal{C}$ denote the space of all configurations of A. Define the function

$+ : \mathbb{Z}^n \times \mathcal{C} \to \mathcal{C}$ as follows.

$$x + a = a \oplus \sigma \text{ where, } \mathsf{freq}\ \sigma = f \ .$$

Since A is a uniform automaton, by definition, the function $x+a$ is well-defined. The kernel M_A of a uniform automaton is defined as follows.

$$M_A = \{x \in \mathbb{Z}^n \mid x + o = o\}$$

The subscript A in M_A is dropped when the automaton A is clear from context.

Lemma 1. *The kernel of a uniform automaton is a sub-module of $\mathbb{Z}^n$.*

Proof. Let $x \in M$. Then, $o = -x + x + o = -x + o$ or $-x \in M$. If $x, y \in M$, then, $o = x + o = x + y + o$ or $x + y \in M$. Hence, M is a sub-module of $\mathbb{Z}^n$. □

The quotient set $\mathbb{Z}^n/M = \{x + M \mid x \in \mathbb{Z}^n\}$ together with the operation $(x+M)+(y+M) = (x+y)+M$ forms a module over $\mathbb{Z}$. It is of central interest from the point of view of characterizing the computations of the automaton.

Lemma 2. *Let M be the kernel of a uniform automaton A. The mapping $x + M \mapsto x + o$ is a set isomorphism between $\mathbb{Z}^n/M$ and the set of reachable configurations $\{x + o \mid x \in \mathbb{Z}^n\}$. A gives the same output for each $y \in x + M$, $x \in \mathbb{Z}^n$.*

Proof. $y \in x+M$ iff $x-y \in M$ or $-y+x+o = o$, or, $x+o = y+o$. A attains the same configuration after processing both x and y and therefore A gives the same output for both x and y. Consider the mapping $x + M \mapsto x + o$. $x + o = y + o$ iff $x - y \in M$, which implies that the mapping is an isomorphism. □

2.3 Free Automata

Let A be a uniform automaton for a fixed value of n with kernel $M = M_A$. Let $E = \mathbb{Z}^n/M$. From classical algebra, E is the direct sum of a free module and the quotient of E with its torsion module, that is, $E = E/E_{\text{tor}} \bigoplus F$, where, E_{tor} is the torsion module corresponding to E, and F is a free module. The torsion module corresponding to a module E over a ring R is $E_{\text{tor}} = \{x \in E \mid \exists a \in R \text{ s.t. } ax = 0\}$. What is $(\mathbb{Z}^n/M)_{\text{tor}}$ in our case? Let $\bar{x}$ denote the element $x+M$ of $\mathbb{Z}^n/M$. Then, $\bar{x} \in (\mathbb{Z}^n/M)_{\text{tor}}$ iff there exists $a \in \mathbb{Z}$ and $a \neq 0$ such that $a\bar{x} = \bar{0}$, or, equivalently, $ax \in M$.

Definition 3. *Let A be a uniform automaton and consider its execution on $[1, n]$. Let M be the kernel of A. Then, A is said to be free for the parameter value n, provided, the factor module $\mathbb{Z}^n/M$ is a free module, or equivalently, $x \in M$ whenever there exists $a \in \mathbb{Z} - \{0\}$ and $x \in \mathbb{Z}^n$ such that $ax \in M$.* □

We shall show that for problems of estimating frequencies, finding approximate frequent items, etc., it is sufficient to restrict attention to free automata. It is however worth noting in passing that there exist automata that are *torsion-full*, that is, $\mathbb{Z}^n/M = (\mathbb{Z}^n/M)_{\text{tor}}$, for example [7]. All other automata in the literature are free (to the best of the author's knowledge). Free automata afford the simplicity of vector spaces in their representation.

Lemma 3. *Let A be free automaton for input domain $[1, n]$ and let M be its kernel. There exists a unique vector subspace M^e of $\mathbb{R}^n$ of the smallest dimension containing M. The mapping $x + M \mapsto x + M^e$ is an injective mapping from $\mathbb{Z}^n/M$ to $\mathbb{R}^n/M^e$. If $\dim \mathbb{Z}^n/M = r$, then, there exists an orthonormal basis $V = [V_1, V_2]$ of $\mathbb{R}^n$ such that $\text{rank}(V_1) = r$, $\text{rank}(V_2) = n - r$, M^e is the linear span of V_2 and $\mathbb{R}^n/M^e$ is the linear span of V_1.*

Proof. $\mathbb{Z}$ is a principal and entire ring. Since $\mathbb{Z}^n$ is a module over $\mathbb{Z}$, its submodules are free modules. Therefore, M is a free module. Since $\mathbb{Z}^n/M$ is given to be free, $\mathbb{Z}^n$ is the direct sum of two free modules, $\mathbb{Z}^n = \mathbb{Z}^n/M \bigoplus M$. Therefore, both M and $\mathbb{Z}^n/M$ have bases, say B_1 and B_2 whose union is a basis for $\mathbb{Z}^n$. Since, $\mathbb{Z}^n$ is a free module and has the standard n-dimensional basis $e_1, \ldots, e_n$, therefore, all bases of $\mathbb{Z}^n$ have the same dimension. Without loss of generality, therefore, let $B = [b_1, b_2, \ldots, b_n]$ be a basis of $\mathbb{Z}^n$ such that $B_1 = [b_1, \ldots, b_r]$ is a basis for $\mathbb{Z}^n/M$ and $B_2 = [b_{r+1}, \ldots, b_n]$ is a basis for M.

Let M^e denote the span of $b_1, \ldots, b_r$ over $\mathbb{R}$. M^e is obviously the smallest vector space over $\mathbb{R}$ that contains M, since, every vector space over $\mathbb{R}$ containing M must contain the span of $b_1, \ldots, b_r$. Therefore, $\dim M^e \le r$ and therefore, $\dim \mathbb{R}^n/M^e \le n - r$ (same argument). However, the standard basis $\{e_1, \ldots, e_n\}$ is a basis of $\mathbb{Z}^n$ and therefore, $\dim M^e + \dim \mathbb{R}^n/M^e = n$. Hence, $\dim M^e = r$ and $\dim \mathbb{R}^n/M^e = n - r$. Further, $b_1, \ldots, b_n$ continues to be a basis for $\mathbb{R}^n$, of which $b_1, \ldots, b_r$ is a basis for M^e and $b_{r+1}, \ldots, b_n$ is a basis for $\mathbb{R}^n/M^e$.

Consider the mapping $x + M \mapsto x + M^e$. Let $\bar{x}$, $\bar{y}$ denote the elements $x + M$ and $y + M$ of $\mathbb{Z}^n/M$. Suppose that $\bar{x} \neq \bar{y}$. Then, $x - y \notin M$. $x - y$ can be expressed uniquely as a linear combination of the basis elements.

$$x - y = \sum_{j=1}^{n} \alpha_i b_i, \quad \alpha_i \in \mathbb{Z}$$

Hence, $x - y$ has the same unique representation in the vector space over $\mathbb{R}^n$. Further, at least one of the coordinates $\alpha_1, \ldots, \alpha_r$ is non-zero, otherwise, $x - y$ would belong to M. Since, $x - y$ has the same representation in the vector space $\mathbb{R}^n$, $x - y$ is not in M^e. The mapping $x + M \mapsto x + M^e$ is therefore injective. Using standard Gram-Schmidt orthonormalization of B_1 and B_2 respectively viewed as defining vector sub-spaces over $\mathbb{R}$, we get V_1 and V_2. By the previous argument, $\text{rank}(V_1) = r$ and $\text{rank}(V_2) = n - r$. □

The sub-matrix V_1 of the orthonormal basis $V = [V_1\ V_2]$ is called the *transition matrix* of the automaton A and defines the space $\mathbb{R}^n/M^e$. The sub-matrix V_2 defines the kernel space M^e. Obviously, neither V_1 nor V_2 are unique, though they serve as useful representations for the sub-spaces they denote. Lemma 3 can be used to characterize the space requirement of a uniform automaton.

Lemma 4. *For a free automaton A with kernel M, $s(n,m) = \Omega((n-r)\log m)$, where, $r = \dim M$.*

Proof. Consider the feasible space $F_m = [-m \ldots m]^n$. The number of distinct configurations that the automaton A goes through is exactly $F_m + M = \{x + M \mid x \in F_m\}$. The number of configurations using workspace of $s = s(n,m)$ is $|Q| \cdot s \cdot 2^s$. Therefore,

$$|Q| \cdot s \cdot 2^s \geq \left|\{x + M \mid x \in F_m\}\right| \ . \tag{1}$$

We first obtain an upper bound on the size $|M \cap F_m|$. Let $b_1, b_2, \ldots, b_r$ be a basis for M. The set

$$P_m = \{\alpha_1 b_1 + \ldots + \alpha_r b_r \mid |\alpha_i| \leq m, \text{ for } i = 1, 2, \ldots, n\}$$

defines the set of all integral points generated by $b_1, b_2, \ldots, b_r$ and that lie in a parallelopiped whose axes are $b_1, b_2, \ldots, b_r$ and length is $2m+1$ (i.e., $[-m \ldots m]$). Therefore, $|P_m| = (2m+1)^r$. Since, each b_i is integral, therefore,

$$|M \cap F_m| = |P_m \cap F_m| \leq |P_m| = (2m+1)^r \ .$$

It follows that

$$\left|\{x + M \mid x \in F_m\}\right| \geq \frac{|F_m|}{|M \cap F_m|} \geq (2m+1)^{n-r} \ .$$

Since, $r = \dim M^e$, substituting in (1), we have

$$s(n,m) = \Omega(\log\left|\{x + M \mid x \in F_{m,n}\}\right|) = \Omega((n-r)\log m) \ . \quad \square$$

3 Frequency estimation

For $x \in \mathbb{R}^n$, let $\|x\|_1$, $\|x\|_2$ and $\|x\|_\infty$ respectively denote the ℓ_1, ℓ_2 and the ℓ_∞ norm of x.

Definition 4. *We say that a stream automaton A approximately estimates the frequency of items with relative error ϵ, provided, after processing any input stream σ with* freq *$\sigma = x$, A returns a vector $\hat{x} \in \mathbb{R}^n$ satisfying*

$$err(\hat{x}, x) = \frac{\|\hat{x} - x\|_\infty}{\|x\|_1} \leq \epsilon \ .$$

The error metric $err(\hat{x}, x)$ is called the relative error of estimation. □

3.1 Frequency estimation using free automata

Suppose that a free automaton A estimates the frequency of the input stream with relative error ϵ. Suppose the kernel of A is M. As shown in Section 2, streams with frequency vector from the same coset of M map to the same configuration of A, and hence, yield the same output. Suppose the vector output by A when in the configuration corresponding to the coset $x + M$ is $h(x + M)$.

Given a coset $x + M$, let $\bar{x}_1 \in x + M$ have the smallest ℓ_1 value.

$$\bar{x}_1 = \min_{y \in x+M} \quad \|y\|_1 \enspace .$$

Lemma 5. *Suppose there exists $h \in \mathbb{R}^n$ such that $err(h, x + M) \leq \epsilon$. Then $err(\bar{x}_1, x + M) \leq 2\epsilon$.*

Proof. Let $\hat{x}$ denote $h(x + M)$. Let $y \in x + M$.

$$\begin{aligned} err(\bar{x}_1, y) &= \frac{\|\bar{x}_1 - y\|_\infty}{\|y\|_1} \leq \frac{\|\bar{x}_1 - h\|_\infty}{\|y\|_1} + \frac{\|h - y\|_\infty}{\|y\|_1} && \text{by triangle inequality} \\ &\leq \frac{\|\bar{x}_1 - h\|_\infty}{\|\bar{x}_1\|_1} + \frac{\|h - y\|_\infty}{\|y\|_1} && \text{since, } \|\bar{x}_1\|_1 \leq \|y\|_1 \\ &\leq \epsilon + \epsilon = 2\epsilon \end{aligned}$$

□

Lemma 6. *Suppose there exists h such that $err(h, M) \leq \epsilon$. Then, $err(0, M) \leq \epsilon$.*

Proof. For any $y_i \in \mathbb{Z}$, $\max(|h_i - y_i|, |h_i + y_i|) \geq |y_i|$. Therefore,

$$\max(\|h - y\|_\infty, \|h + y\|_\infty) \geq \|y\|_\infty$$

. Let $y \in M$. Since, M is a module, $-y \in M$. Thus,

$$\begin{aligned} err(0, y) = err(0, -y) = \frac{\|y\|_\infty}{\|y\|_1} &\leq \frac{1}{\|y\|_1} \max(\|h - y\|_\infty, \|h + y\|_\infty) \\ &= \max(err(h, y), err(h, -y)) \leq \epsilon \end{aligned}$$

□

We now show that if $err(0, y) \leq \epsilon$ for each $y \in M$, then, $err(0, z) \leq \epsilon$ for each $z \in M^e$. The proof is straightforward and is deferred to Appendix A.

Lemma 7. *If $err(0, M) \leq \epsilon$ then $err(0, M^e) \leq \epsilon$.* □

For a free automaton A with kernel M and corresponding $n \times r$ transition matrix V_1, the minimum ℓ_2 estimator is defined in terms of the transition matrix V_1 as follows.

$$est_2(x) = V_1 V_1^T x \enspace . \tag{2}$$

To show that the ℓ_2 estimator is well-defined, let $x + M^e = y + M^e$, or that, $x - y \in M^e$. Since, the columns of V_2 form a basis for M^e, $x - y = V_2 z$, for some z. Therefore,

$$V_1 V_1^T (x - y) = V_1 V_1^T V_2 z = V_1 \cdot 0 \cdot z = 0 \enspace .$$

or that $V_1V_1^Tx = V_1V_1^Ty$, or equivalently, $est_2(x) = est_2(y)$. It is called the ℓ_2 estimator since it returns a point in the set $x + M^e$ that is closest to the origin in terms of the ℓ_2 distance.

We now show that there is a subset J of the set of the standard unit vectors $\{e_1, e_2, \ldots, e_n\}$ such that $|J| = \Theta(n)$ and the minimum ℓ_2 estimator is nearly optimal for the unit vectors in J. We first prove a technical lemma.

Lemma 8. *For any real $C \geq 1$, let $J_C = \{i : 1 \leq i \leq n \text{ and } \|V_1V_1^Te_i\|_1 \geq C\}$. Then, $|J_C| \leq \frac{n}{C}$.*

Proof. Since, V_1 has orthonormal columns, $\|V_1\|_2 = \|V_1V_1^T\|_2 = 1$. By a standard identity between norms, we have

$$\|V_1V_1^T\|_F \leq \sqrt{n}\|V_1V_1^T\|_2 = \sqrt{n}$$

Therefore,

$$|J_C| \cdot C \leq \sum_{i \in J_C} \|V_1V_1^Te_i\|_1 \leq \|V_1V_1^T\|_F^2 \leq n, \text{ or, } |J_C| \leq \frac{n}{C} \quad . \qquad \square$$

Lemma 9. *Let A be a free automaton that estimates the frequency vector of the input stream over domain $[1, n]$ with relative error ϵ. Then there exists a subset $J' \subset \{1, 2, \ldots, n\}$ of size at least $\lceil n/2 \rceil$ such that for $i \in J'$, $err(est_2(x), e_i) \leq 3\epsilon$.*

Proof. For $C > 1$, let J'_C be the index set

$$J'_C = \{i : \|V_1V_1^Te_i\|_1 < C\} \quad .$$

For $i \in J'_C$,

$$err(est_2(e_i), e_i) = \frac{\|est_2(e_i) - e_i\|_\infty}{\|e_i\|_1} = \frac{\|est_2(e_i) - e_i\|_\infty}{\|est_2(e_i) - e_i\|_1} \cdot \|est_2(e_i) - e_i\|_1$$

The element $w = est_2(e_i) - e_i \in M^e$. By Corollary 6, $err(0, M) \leq \epsilon$. By Lemma 7,

$$err(0, w) \leq err(0, M^e) \leq err(0, M) \leq \epsilon \quad .$$

Further,

$$\|est_2(e_i) - e_i\|_1 \leq \|est_2(e_i)\|_1 + \|e_i\|_1 < C + 1$$

since, $i \in J'_C$ and $est_2(e_i) = V_1V_1^Te_i$. Combining, for $i \in J'_C$

$$err(est_2(e_i), e_i) < \epsilon(C + 1)$$

Choose $C = 2$. By Lemma 8 it follows that $|J'_C| = n - |J_C| \geq n - n/2 \geq \lceil \frac{n}{2} \rceil$. $\square$

Theorem 1 (Alon [2,1]). *There exists a positive constant c so that the following holds. Let B be an n by n real matrix with $b_{i,i} \geq \frac{1}{2}$ for all i and $|b_{i,j}| \leq \epsilon$ for all $i \neq j$, where, $\frac{1}{2\sqrt{n}} \leq \epsilon < \frac{1}{4}$. Then the rank of B satisfies*

$$\text{rank}(B) \geq \frac{c \log n}{\epsilon^2 \log(1/\epsilon)} \quad .$$

Lemma 10. *Let $\frac{1}{2\sqrt{n}} \leq \epsilon < \frac{1}{12}$ and let A be a free and uniform stream automaton over the domain $[1, n]$ such that A estimates the frequency vector with relative error ϵ. Then, the rank of the transition matrix of A satisfies*

$$\text{rank}(V_1) = \Omega(\frac{\log n}{\epsilon^2 \log(\epsilon^{-1})}) \ .$$

Proof. By Lemma 9, there exists $J' \subset \{1, 2, \ldots, n\}$ such that $|J'| \geq \lceil n/2 \rceil$ and $err_2(e_i) \leq 3\epsilon$ for $i \in J'$. Let V_1 be an $n \times r$ matrix. Define Y to be the $|J'| \times r$ sub-matrix of V_1 that includes the ith row of V_1 for each $i \in J'$. Let $U = Y^T$ which is an $r \times |J'|$ matrix. For $i \in J'$,

$$err(est_2(e_i), e_i) \leq 3\epsilon, \text{ or, } \frac{\|V_1 V_1^T e_i - e_i\|_\infty}{\|e_i\|_1} = \|V_1 V_1^T e_i - e_i\|_\infty \leq 3\epsilon \ .$$

Therefore, for $j, k \in \{1, 2, \ldots, |J'|\}$, $j \neq k$

$$|U_j^T U_k| \leq 3\epsilon \ \text{ if } j \neq k \text{ and } |U_j^T U_j - 1| \leq 3\epsilon$$

The matrix UU^T satisfies all the premises of Alon's Theorem [2] (reproduced above). Therefore,

$$\text{rank}(V_1) \geq \text{rank}(U) = \text{rank}(UU^T) = \Omega\left(\frac{\log n}{\epsilon^2 \log \epsilon^{-1}}\right) \ . \quad \square$$

Lemma 11. *Let A be a free automaton that estimates frequency with a relative error of $\frac{1}{2\sqrt{n}} \leq \epsilon < \frac{1}{24}$. Then, A requires space $\Omega(\frac{(\log n)(\log m)}{\epsilon^2 \log \epsilon^{-1}})$ on its work tape for stream of size at most m.*

Proof. By Lemma 10, the rank of the transition matrix of A is $\Omega\left(\frac{\log n}{\epsilon^2(\log \epsilon^{-1})}\right)$. Since, $\text{rank}(V_1) = n - \dim M^e$ and $\dim M^e = \dim M$, by Lemma 4, $s(n, m) = \Omega((n - \dim M^e) \log m) = \Omega(\text{rank}(V_1) \log m) = \Omega\left(\frac{(\log m)(\log n)}{\epsilon^2(\log \epsilon^{-1})}\right) \ . \ \square$

3.2 General uniform automata

We now show that it is sufficient to consider only free automata for estimating frequencies. Suppose that M is the module of the kernel of a uniform automaton A and $\mathbb{Z}^n/M$ is not necessarily free. Let M' be the module that removes the torsion from $\mathbb{Z}^n/M$, that is,

$$M' = \{x \in \mathbb{Z}^n \mid \exists a \in \mathbb{Z}, a \neq 0 \text{ and } ax \in M\} \tag{3}$$

Then, $\mathbb{Z}^n/M'$ is torsion-free. If $y \in M'$ is a torsion element, then $by = 0$ for some $b \in \mathbb{Z}$, or, $bax = 0$ for some $a \in \mathbb{Z}$ and $x \in M$. Since, $\mathbb{Z}$ has no zero divisors, $x = 0$, implying that $y = 0$. Thus, M' is torsion free. Therefore,

$$\mathbb{Z}^n = \mathbb{Z}^n/M' \bigoplus M'$$

where, both $\mathbb{Z}^n/M'$ and M' are free modules (i.e., torsion-free). Before proving the equivalence of free and uniform automaton for the problem of estimating frequencies, we need a few algebraic facts.

Fact 12 *Let $b_1, b_2, \ldots, b_r$ be a basis of M'. Then, there exist non-zero elements $\alpha_1, \alpha_2, \ldots, \alpha_r \in \mathbb{Z}$ such that $\alpha_1 b_1, \ldots, \alpha_r b_r$ is a basis for M.*

Proof. It follows from standard algebra that the basis of M is of the form $\alpha_1 b_1, \ldots, \alpha_r b_r$. It remains to be shown that the α_i's are non-zero. Suppose that $\alpha_1 = 0$. For any $a \in \mathbb{Z}$, $a \neq 0$, suppose $ax \in M$ and $x \in M'$. Then, x has a unique representation as $x = \sum_{j=1}^{r} x_j b_j$. Thus, $ax = \sum_{j=1}^{r} (ax_j) b_j \in M$ and has the same representation in the basis $\{\alpha_j b_j\}_{j=1,\ldots,n}$. Therefore, $ax_1 = 0$ or $x_1 = 0$ for all $x \in M'$, which is a contradiction. □

Corollary 1. $M^e = (M')^e$.

Proof. Let $\{b_1, b_2, \ldots, b_r\}$ be a basis for M'. Then, by Lemma 12, there exist non-zero elements $\alpha_1, \ldots, \alpha_r$ such that $\{\alpha_1 b_1, \alpha_2 b_2, \ldots, \alpha_r b_r\}$ is a basis for M. Therefore, over reals, $(b_1, \ldots, b_r) = (\alpha_1 b_1, \ldots, \alpha_r b_r)$. Thus, $M^e = (M')^e$. □

We can now show that frequency estimation by a uniform automaton with relative error ϵ can be simulated by a free automaton having relative error $O(\epsilon)$.

Lemma 13. *Suppose A is a uniform automaton that estimates the frequency with relative error ϵ and has kernel M. Then there exists a free automaton B with the following properties: (1) the kernel of B is M' such that $M' \supset M$ and $\mathbb{Z}^n / M'$ is free, and (2) B estimates frequency with relative error 4ϵ .*

Proof. Let M be the kernel of A and let M' be the torsion-free extension of M. Since A estimates the frequency with relative error ϵ, there exists for coset $x+M$, a value $\hat{x}$ such that $err(\hat{x}, x+M) \leq \epsilon$. Consider the set of cosets $\{x + M'\}_{x \in \mathbb{Z}^n}$ and for each coset return the element with the smallest ℓ_1 norm. Denote this as $h(x + M')$

$$h(x + M') = \min_{y \in x + M'} \quad \|y\|_1$$

Let $y \in x + M'$. Then, $y \in x_1 + M$ for some x_1. Let $\hat{y}$ denote the point returned by A for $x_1 + M$ and let y' denote the point in $x_1 + M$ with the minimum ℓ_1 value. Let h denote $h(x + M')$ and let $\hat{h}$ denote the point returned by A for the coset $h + M$. Therefore,

$$err(h, y) = \frac{\|y - h\|_\infty}{\|y\|_1} \leq \frac{\|y - \hat{y}\|_\infty}{\|y\|_1} + \frac{\|\hat{y} - y'\|_\infty}{\|y\|_1} + \frac{\|y' - h\|_\infty}{\|y\|_1} \tag{4}$$

The first and the second terms above are bounded by ϵ as follows. The first term $\frac{\|y - \hat{y}\|_\infty}{\|y\|_1} = err(\hat{y}, y) \leq \epsilon$, since, $y \in x_1 + M$ and $\hat{y}$ is the estimate returned by A for this coset. The second term

$$\frac{\|\hat{y} - y'\|_\infty}{\|y\|_1} \leq \frac{\|\hat{y} - y'\|_\infty}{\|y'\|_1} = err(\hat{y}, y') \leq \epsilon$$

since, $\|y'\|_1 \leq \|y\|_1$ and y' lies in the coset $x_1 + M$.

$y' - h \in M'$ and $M' \subset M^e$, by Corollary 1. The third term in (4) can be rewritten as

$$\begin{aligned}\frac{\|y'-h\|_\infty}{\|y\|_1} &\leq \frac{\|y'-h\|_\infty}{\|y'-h\|_1} \cdot \frac{\|y'-h\|_1}{\|y'\|_1}, \qquad \text{since, } \|y'\|_1 \leq \|y\|_1 \\ &\leq \epsilon \cdot \frac{\|y'\|_1 + \|h\|_1}{\|y'\|_1} \quad \text{by Lemma 7 and by triangle inequality} \\ &\leq 2\epsilon, \qquad \text{since, } \|h\|_1 \leq \|y'\|_1\end{aligned}$$

By (4), $err(h, y) \leq 4\epsilon$. □

By Lemma 2 and since $M' \supset M$, the above lemma shows that the space of reachable configurations of B which is isomorphic to $\mathbb{Z}^n + M'$ is no larger than the space of reachable configurations of A which is isomorphic to $\mathbb{Z}^n + M$. In particular, if we restrict the execution of A to input streams with size at most m, then, the space of reachable configurations if $F_m + M$, as compared to $F_m + M'$ for B, where, $F_m = [-m \dots m]^n$.

3.3 Non-uniform automata

We now show that given a non-uniform automaton that estimates frequency with relative error ϵ, one can construct a uniform automaton that estimates the frequency with relative error $O(\epsilon)$.

Lemma 14. *Suppose A is a possibly non-uniform automaton that estimates the frequency vector with relative error ϵ. Then there exists a uniform automaton that uses fewer configurations and has relative error at most 2ϵ.*

Proof. Let A be a possibly non-uniform automaton that estimates the frequency vector with relative error ϵ. For a reachable configuration a define

$$W_a = \{f \mid \exists \sigma \text{ s.t.} o \oplus \sigma = a \text{ and } \mathsf{freq}\ \sigma = f\}$$

Consider the set of configurations

$$O = \{s \mid \exists\ \sigma \text{ s.t. } o \oplus \sigma = s \text{ and } \mathsf{freq}\ \sigma = 0\}$$

Let

$$W_O = \cup_{s \in O} W_s$$

Suppose $f, g \in W_a$. Since, $g \in W_a$, $a \oplus \tau \in O$ where, $\mathsf{freq}\ \tau = -g$. Further, since, $f \in W_a$, there exists σ such that $o \oplus \sigma = a$ and $\mathsf{freq}\ \sigma = f$. Therefore

$$o \oplus \sigma \oplus \tau = a \oplus \tau \in O \text{ and } \mathsf{freq}\ \sigma \oplus \tau = f - g$$

In other words, $f - g \in W_O$. In particular, if $g \in W_O$ and since $0 \in W_O$ (trivially), so, $0 - g = -g \in W_O$. If $f, g \in W_O$, then, $f + g = f - (-g) \in W_O$. Therefore, W_O is a sub-module of $\mathbb{Z}^n$.

If $f \in W_a$, then, $W_a \subset f + W_O$. (Why? since, for any $g \in W_a$, $g - f \in W_O$ or $g \in f + W_O$). Therefore, there is an onto mapping from $\{W_a\}_{a \text{ reachable}}$ to the set $\{f + W_O \mid f \in \mathbb{Z}^n\}$. Each coset $f + W_O$ is the union of several W_a's (possibly non-disjoint.) Let x be the smallest ℓ_1 element in $f + W_O$ and y be any element from $f + W_O$. Then,

$$err(x, y) = \frac{\|x - y\|_\infty}{\|y\|_1} \leq \frac{\|x - y\|_\infty}{\|y - x\|_1} \cdot \frac{\|y - x\|_1}{\|y\|_1}$$

Since, $x - y \in W_O$ therefore, $err(0, x - y) = \frac{\|x-y\|_\infty}{\|y-x\|_1} \leq \epsilon$ (Lemma 6). Further

$$\frac{\|y - x\|_1}{\|y\|_1} \leq \frac{\|y\|_1 + \|x\|_1}{\|y\|_1} \leq 2$$

since, $\|x\|_1 \leq \|y\|_1$. Therefore, $err(x, y) \leq 2\epsilon$. □

We can now prove the main theorem of this paper.

Theorem 2. *A stream automaton A that estimates the frequency vector of its input stream with relative error ϵ requires space $s(n, m) = \Omega\left(\frac{(\log m)(\log n)}{\epsilon^2 \log(1/\epsilon)}\right)$.*

Proof. By Lemma 14, if there is a uniform or a non-uniform automaton A that estimates the frequency with error ϵ, then, there exists a uniform automaton B that estimates frequency with error 2ϵ and requiring fewer configurations. In particular, if we restrict the execution of A to input streams with size at most m, then, the space of reachable configurations of A is no smaller than those for B, which can be counted as $F_m + W_0$, where, $F_m = [-m \ldots m]^n$. By Lemma 13, for every such uniform automaton B, there is a free automaton C that estimates the frequency vector with relative error at most 4ϵ and whose space of reachable configurations is no larger than that of B. The space of reachable configurations of an automaton A , restricted to streams of size m, is $s_A(n, m) = \Omega($log of the number of reachable configurations of $A)$. By the preceding argument, $s_A(n, m) = \Omega($log of the number of reachable configurations of $C)$, which by Lemmas 4 and 11, is $\Omega\left(\frac{(\log m)(\log n)}{\epsilon^2 \log(1/\epsilon)}\right)$. □

4 Conclusion

We present a deterministic space lower bound of $\Omega\left(\frac{(\log m)(\log n)}{\epsilon^2 (\log \epsilon^{-1})}\right)$ for the problem of estimating frequencies within relative error ϵ over update streams. This improves the existing space bounds for a number of related problems over update streams via standard reductions, including, finding ϵ-approximate frequent items, finding ϵ-approximate quantiles, approximate range-sums, etc..

Acknowledgements

The author thanks Piotr Indyk for useful discussions.

References

1. Noga Alon. "Problems and results in extremal combinatorics, I". *Discrete Math.*, 273(1-3):31–53, 2003.
2. Noga Alon. "Perturbed identity matrices have high rank: proof and applications". Available from http://www.math.tau.ac.il/∼nogaa/identity.pdf, 2006.
3. P. Bose, E. Kranakis, P. Morin, and Y. Tang. "Bounds for Frequency Estimation of Packet Streams". In *Proc. SIROCCO*, pages 33–42, 2003.
4. Graham Cormode and S. Muthukrishnan. "An Improved Data Stream Summary: The Count-Min Sketch and its Applications". *J. Algorithms*, 55(1):58–75, April 2005.
5. E. D. Demaine, A. López-Ortiz, and J. I Munro. "Frequency estimation of internet packet streams with limited space". In *Proc. ESA*, pages 348–360, 2002.
6. S. Ganguly and Majumder A. "CR-precis: A Deterministic Summary Structure for Update Streams". In *Proc. ESCAPE*, 2007.
7. S. Ganguly and A. Majumder. "Deterministic K-set Structure". In *Proc. ACM PODS*, pages 280–289, 2006.
8. R.M. Karp, S. Shenker, and C.H. Papadimitriou. "A Simple Algorithm for Finding Frequent Elements in Streams and Bags". *ACM TODS*, 28(1):51–55, 2003.
9. J. Misra and Gries. D. "Finding repeated elements". *Sci. Comput. Programm.*, 2:143–152, 1982.

A Proofs

Proof (of Lemma 7). Let $z \in M^e$. Let $b_1, b_2, \ldots, b_r$ be a basis of the free module M. For $t > 0$, let tz be expressed uniquely as $tz = \alpha_1 b_1 + \ldots + \alpha_r b_r$, where, α_i's belong to $\mathbb{R}$. Consider the vertices of the parallelopiped P_{tz} whose sides are $b_1, b_2, \ldots, b_r$ and that encloses tz.

$$\begin{aligned} P_{tz} = {} & [\alpha_1]b_1 + [\alpha_2]b_2 + \ldots + [\alpha_n]b_n \\ & + \{\beta_1 b_1 + \beta_2 b_2 + \ldots + \beta_r b_r \mid \beta_j \in \{0,1\}, j = 1,2,\ldots,r\} \end{aligned}$$

where, $[\alpha]$ denotes the largest integer smaller than or equal to α. Since, ℓ_∞ is a convex function $\|tz\|_\infty \leq \|y\|_\infty$ for some $y \in P_{tz}$. Let $y = \sum_{j=1}^r \beta_j b_j$, for $\beta_j \in \{0,1\}$, $j = 1,2,\ldots,r$.

$$\|y - tz\|_1 = \|\sum_{j=1}^r (\beta_j - [\alpha_j])b_j\|_1 \leq \sum_{j=1}^r \|(\beta_j - [\alpha_j])b_j\|_1 \leq \sum_{j=1}^r \|b_j\|_1$$

$$\text{or,} \quad \|tz\|_1 \geq \|y\|_1 - \sum_{j=1}^r \|b_j\|_1$$

Therefore,

$$\begin{aligned} err(0, tz) = \frac{\|tz\|_\infty}{\|tz\|_1} &\leq \frac{\|y\|_\infty}{\|y\|_1 - \sum_{j=1}^r \|b_j\|_1} \\ &\leq \left(\frac{\|y\|_1}{\|y\|_\infty} - \frac{\sum_{j=1}^r \|b_j\|_1}{\|y\|_\infty} \right)^{-1} \leq \left(\frac{1}{\epsilon} - \frac{\sum_{j=1}^r \|b_j\|_1}{\|y\|_\infty} \right)^{-1} \end{aligned}$$

where, the last step follows from the assumption that $y \in M$ and therefore, $err(0, y) = \frac{\|y\|_\infty}{\|y\|_1} \le \epsilon$. The ratio $\frac{\sum_{j=1}^{r} \|b_j\|_1}{\|y\|_\infty}$ can be made arbitrarily small by choosing t to be arbitrarily large. Thus, $\lim_{t \to \infty} err(0, tz) \le \epsilon$. Since, $err(0, tz) = \frac{\|tz\|_\infty}{\|tz\|_1} = \frac{\|z\|_\infty}{\|z\|_1} = err(0, z)$, for all t, we have, $err(0, z) \le \epsilon$. □