A hybrid technique for estimating frequency moments over data streams

Sumit Ganguly

Indian Institute of Technology, Kanpur
e-mail: sganguly@iitk.ac.in

Abstract. The problem of estimating the k^{th} frequency moment F_k, for any non-negative integral value of k, over a data stream by looking at the items exactly once as they arrive, was considered in a seminal paper by Alon, Matias and Szegedy [1, 2]. They present a sampling based algorithm to estimate F_k where, $k \geq 2$, using space $\tilde{O}(n^{1-1/k})$. Coppersmith and Kumar [7] and [10], using different methods, present algorithms for estimating F_k with space complexity $\tilde{O}(n^{1-1/(k-1)})$. In this paper, we present an algorithm for estimating F_k with space complexity $\tilde{O}(n^{1-2/(k+1)})$, for $k > 2$, thereby, improving the space complexity compared to the algorithms in [1, 2, 7, 10] for $k \geq 4$.

1 Introduction

Data streams are characterized by large volumes of data that arrive rapidly and continuously. Due to the volume of the data, it is desirable to design algorithms that estimate metrics over the data streams, with *sub-linear* space complexity. One such metric is frequency moments, studied in a seminal paper by Alon, Matias and Szegedy [1, 2]. In this paper, we study and present novel algorithms for this problem.

We view a stream as a sequence of arrivals of items l, where, l is the identity of an item. The items are assumed to draw their identities from the domain $[N]$. The frequency of an item with identity l, denoted by f_l, is the number of occurrences of l since the inception of the stream. Thus, each arrival of an item increases its frequency by 1. The k^{th} frequency moment of the stream, denoted by F_k, is defined as $\sum_l f_l^k$, for $k > 2$ and k integral. This problem was first studied in a seminal paper by Alon, Matias and Szegedy [1, 2]. It is interesting for several reasons. Historically, it is one of the very first problems studied in the data streaming model and helped to consolidate the area of data streaming algorithms. Secondly, the techniques presented in [1, 2] have led to new solution techniques for several practical problems on data streams, for example, maintenance of histograms [12] and maintenance of the top-k frequent items [6].

1.1 Prior Work

The k^{th} moment estimation algorithm, presented in [1, 2] (which we call the AMS algorithm) is based on sampling and estimates F_k, for $k \geq 2$, to within any specified approximation parameter $0 < \epsilon < 1$, and with confidence exceeding a user-specified parameter $\delta < 1$. The space complexity of the AMS algorithm is $\tilde{O}(n^{1-1/k})$,

where, n is the number of distinct elements in the stream. Recently, Coppersmith and Kumar [7] have presented an algorithm for this problem whose space complexity is $\tilde{O}(n^{1-1/(k-1)})$, for $k \geq 3$. Independently, the work in [10] presents an algorithm for this problem with the same space complexity, using a different technique[1].

Lower bounds. The work in [1, 2] shows space lower bounds for this problem to be $\Omega(n^{1-5/k})$, for any $k > 5$. Subsequently, the space lower bounds have been strengthened to $\Omega(\epsilon^2 n^{1-(2+\epsilon)/k})$, for $k > 2, \epsilon > 0$ by Bar-Yossef, Jayram, Kumar and Sivakumar [3] and to $\Omega(n^{1-2/k})$ by Chakrabarti, Khot and Sun [5]. Saks and Sun [14] show that estimating the L_p distance d between two streaming vectors to within a factor of d^δ requires space $\Omega(n^{1-2/p-4\delta})$.

Other Related Work. For the special case of computing F_2, [1, 2] presents an $\tilde{O}(1)$ space and time complexity algorithm, where, m is the sum of the frequencies. Random linear combinations based on random variables drawn from stable distributions were considered by [13] to estimate F_p, for $0 < p \leq 2$. The work in [8] presents a sketch technique to estimate the difference between two streams based on the L_1 metric norm. There has been substantial work to estimate F_0 (i.e., the number of distinct items in a stream) and related metrics (i.e., set expression cardinalities) for general update streams [9, 1, 4, 11].

1.2 Contributions

In this paper, we present the *hybrid* algorithm for estimating F_k, for $k \geq 3$, to within any specified accuracy parameter $0 < \epsilon < 1/2$, and with confidence exceeding a user-specified confidence parameter $\delta < 1$. The space complexity of the algorithm is $\tilde{O}(n^{1-\frac{2}{k+1}})$ bits. Formally, the contribution of the paper is stated as follows.

Theorem 1. *For every $\epsilon < 1/2$ and $0 < \delta < 1$, there exists an algorithm that returns an estimate $\hat{F}_k$ such that $\mathbf{Pr}\left\{|\hat{F}_k - F_k| < \epsilon F_k\right\} \geq 1 - \delta$ using space $O\left(\frac{k}{\epsilon} \cdot n^{(1-\frac{2}{k+1})}\right) \cdot \left(\frac{1+\epsilon}{1-\epsilon}\right)^{2(k-1)} \log \frac{m}{\delta}\Big)$ bits.* □

The hybrid algorithm is based on a non-trivial combination of the classical algorithm for estimating moments [1, 2], and the COUNTSKETCH technique[6] for estimating frequent items over a stream. The space complexity of the hybrid algorithm, that is, $\tilde{O}(n^{1-\frac{2}{k+1}})$, is significantly superior to that of existing algorithms for this problem [7, 10], that is, $\tilde{O}(n^{1-\frac{1}{k-1}})$, for $k \geq 4$. Somewhat surprisingly, the hybrid technique uses completely different techniques from the ones used in [7] or [10].

1.3 Organization

The rest of the paper is organized as follows. Section 2 presents our algorithm to estimate F_k, and Section 3 presents its analysis. Finally, we conclude in Section 4.

[1] The algorithms in [7, 10] can handle insertions and deletions as well.

2 The Hybrid Estimator

In this section, we describe our estimator precisely. Before doing so, we first present an overview of the hybrid algorithm.

2.1 An overview of hybrid algorithm

The hybrid estimator uses two data structures in parallel, namely, the median of averages sampling based AMS estimator for F_k presented in [1, 2], and the COUNTSKETCH data structure of Charikar, Chen and Farach-Colton[6]. The COUNTSKETCH algorithm [6] is an application of the sketch technique [1, 2] to identify the top-k elements in the stream, in terms of their frequencies. It uses space $\tilde{O}(B)$, and returns all *dense* elements, that is, items with frequency $f_i^2 = \Omega(F_2/B)$. These estimates are then scaled appropriately to accurately estimate the dense component of F_k, defined as $F_k^{(d)} = \sum_{i \text{ dense}} f_i^k$.

Concurrently, we run the AMS algorithm and maintain $u \cdot v$ independent copies of the AMS estimator, divided into v groups of u estimators each (to enable a median of averages computation). Each AMS estimator is derived from a single element sample, and the hybrid estimator first discards all those samples that refer to dense elements. The result is a collection of *reduced samples*; each sample group retaining say, $u'(a) \leq u$ samples, $1 \leq a \leq v$. Next, the reduced samples are used to estimate the sparse component of F_k, defined as $F_k' = \sum_{i \text{ non-dense}} f_i^k$, using the median of averages calculation of the AMS estimator [1, 2]. Finally, the sum of the estimates for dense and sparse components is returned.

This summarizes the hybrid estimator except for a caveat. If there exists a small-sized sample group, that is, a group whose reduced size $u'(a) < \beta u$, (where, $\beta = \beta(\epsilon)$ is fixed later) then we return 0 as the estimate for the sparse component. In this case, we show that the sparse component is at most $\frac{\epsilon}{2}$ times the dense component, and therefore, the total error in estimation remains within the specified tolerance (with high probability).

Notation. The estimation procedure, at any time during the processing of the stream, classifies all items that have appeared in the stream into two categories, namely, *dense* and *sparse*. The dense component of the k^{th} moment, denoted by $F_k^{(d)}$, is the contribution made in the calculation of F_k the dense elements, that is, $F_k^{(d)} = \sum_{i \text{ dense}} f_i^k$. Analogously, the sparse component of F_k, denoted by F_k', is defined as $F_k' = \sum_{i \text{ not dense}} f_i^k$. Clearly, the dense and sparse components, that is, $F_k^{(d)}$ and F_k', respectively, are both random variables, and satisfy the relation, $F_k^{(d)} + F_k' = F_k$.

2.2 Dense Component Estimator

The dense component estimation procedure works by slightly modifying the COUNTSKETCH algorithm of Charikar, Chen and Farach-Colton [6], as follows.

The COUNTSKETCH algorithm maintains a heap of size $\tilde{O}(B)$, that keeps track of the potential top-B items and their frequencies. We modify this procedure slightly,

in order to have high confidence in the estimates, and return an item i from the heap, provided, its estimated frequency, $\hat{f}_i$, crosses a threshold, $T = \frac{c \cdot Y}{B}$. Here, Y is an estimate of F_2 to within a factor of $(1 \pm \frac{\epsilon}{3})$ obtained using the sketching algorithm of [1, 2], and $c = c(\epsilon)$ is a constant. An item is *defined* to be *dense* if it is returned by the COUNTSKETCH algorithm (subject to crossing the threshold). An item that is not dense is called *sparse*. The estimator D for the dense component is defined as follows.

$$D = \sum_{i \text{ dense}} \hat{f}_i^k$$

Since, the procedure is a minor modification of the COUNTSKETCH algorithm, we directly state its properties, without much elaboration or proof. The precise notion of when an item is dense or sparse is not crucial to the remainder of the paper, as long as the following properties hold.

First, D estimates $F_k^{(d)}$ closely and with high confidence. Secondly, the procedure gives an upper bound on the frequencies of the sparse values (with high confidence), denoted by T_s, and a lower bound on the frequencies of dense frequencies, denoted by T_d, as follows.

$$T_s = \frac{F_2}{B}, \quad T_d = \frac{F_2}{\lambda^2 \cdot B}, \quad \text{where, } \lambda^2 = \frac{T_s}{T_d} \leq \frac{(1+\epsilon)^2}{(1-\epsilon)^2} \tag{1}$$

The properties are formally stated in Lemma 2.

Lemma 2. *Let $0 < \epsilon < 1/3$ and $0 < \delta < 1$. There exists an algorithm that returns a set of dense items, and a value D, using space $O(k \cdot \frac{B}{\epsilon} \cdot \log \frac{m}{\delta} + k^2 \cdot \log \frac{m}{\delta})$, and satisfying the following properties.*

1. $\mathbf{Pr}\left\{|D - F_k^{(d)}| < \epsilon F_k^{(d)}\right\} > 1 - \delta$.
2. *if $f_i^2 > T_d$, then* $\mathbf{Pr}\{i \text{ is dense}\} > 1 - \frac{\delta}{m}$.
3. *if $f_i^2 < T_s$, then* $\mathbf{Pr}\{i \text{ is not dense}\} > 1 - \frac{\delta}{m}$. □

Threshold Consistency. In later analysis, several lemmas assume that the sparse frequencies are at most T_s and dense frequencies are at least T_d, and the results of these lemmas are later combined using the union bound, by adding the error probabilities. For ease of reference, we introduce the following notation. We say that the classification of items into dense and sparse categories is *threshold consistent*, if the following two conditions hold.

(1) For every sparse item i, $f_i^2 < T_s$, and, (2) for every dense item i, $f_i^2 > T_d$. By Lemma 2 parts (2) and (3), threshold consistency holds with probability at least $(1 - \delta)$.

2.3 Sparse Component Estimator and Hybrid Estimator

The sparse component estimator E keeps $u \cdot v$ independent copies of the AMS sample based estimator X, divided into groups of u copies each. The values of u and v will be fixed later. The samples are denoted as $X[a][b]$, $a \in \{1, \ldots, v\}$, and $b \in \{1, \ldots, u\}$. Let $S = S(a)$ denote the sample group $\{X[a][1], \ldots, X[a][u]\}$, $1 \leq a \leq v$. The estimator works in two steps, namely, it first *reduces* each sample group, and then, uses the reduced sample groups to estimate the sparse component, F_k'.

Reducing Sample Groups. Each AMS estimator $X[a][r]$ is calculated as $(g^k - (g-1)^k) \cdot m$, from an individual sample that is a pair of the form (i, g), where, i is the identity of an element and g is a count of the frequency of i from the (random) position in the stream where the counter was started. We remove all those sample entries in which the element i has been identified as a dense element. The resultant sample is called the *reduced sample group* $S' = S'(a)$. The size of the reduced sample group $|S'(a)|$ is denoted by $u'(a)$.

Sparse Estimator. For each $1 \leq a \leq v$, we first check to see whether $u'(a) \geq \beta \cdot u$, where, $\beta = \beta(\epsilon)$ is a fraction that is specified later. If this condition fails for one or more values of a, then the estimator E is set to 0. Otherwise, the algorithm of [1, 2] is run using each reduced sample group $S'(a)$ to obtain the average $E'(a)$ of the $u'(a)$ estimates, for $1 \leq a \leq v$. The estimator E returns the median of these estimates, that is, $E = \text{median}_{1 \leq a \leq v} E'(a)$.

Hybrid estimator. Finally, the hybrid estimator simply returns the sum of the dense component estimator and the sparse component estimator, that is, it returns $H = D + E$.

3 Analysis

In this section, we analyze the properties of the sparse component estimator and the hybrid estimator. The analysis proceeds in three parts. In the first part, we show that the process of reduction of samples by removing those samples whose items have been identified as dense items, reduces the variance, as compared to the variance of the original AMS estimator. In the second part, we show that the technique of reduction, together with simple choices of constants, ensures that an estimation of the sparse component, F'_k, can be performed to within an accuracy of $(1 \pm \epsilon)$ factor of F_k. Finally, we combine the results to prove accuracy and confidence properties of the hybrid estimator.

3.1 Reducing variance of AMS estimator

The variance of an (elementary) AMS estimator is $F_1 \cdot F_{2k-1}$. Therefore, after reducing the sample, each reduced sample has variance $F'_1 \cdot F'_{2k-1}$. In this section, we bound the expression $F'_1 \cdot F'_{2k-1}$. We begin with the following elementary fact about expectations.

Lemma 3. *Let X be a random variable and $g(\cdot)$ be a function that is convex over the set of values that X assumes. Then, $g(\mathbf{E}[X]) \leq \mathbf{E}[g(X)]$.*

Proof.

$$g(\mathbf{E}[X]) \;=\; g\Big(\sum_x \mathbf{Pr}\{X = x\} \cdot x\Big) \leq \sum_x \mathbf{Pr}\{X = x\} \cdot g(x) \;=\; \mathbf{E}[g(X)] \;. \quad \square$$

This simple fact has interesting and relevant corollaries.

Corollary 4. *If $j \leq k$, then, $F_j \leq n^{1-\frac{j}{k}} F_k^{\frac{j}{k}}$.*

Proof. Let S denote the set of items in the stream with non-zero frequency (i.e., items that have occurred) and let $n = |S|$. Let X denote the random variable that picks a random item i among these items, and returns its frequency. Let $Y = X^j$. By Lemma 3, and letting $g(Y) = Y^{\frac{k}{j}}$ (which is a convex function, for $k \geq j$), we have the following.

$$g(\mathbf{E}\big[Y\big]) \leq \mathbf{E}\big[g(Y)\big] \;\Leftrightarrow\; \left(\frac{F_j}{n}\right)^{\frac{k}{j}} \leq \frac{F_k}{n} \;\Leftrightarrow\; F_j \leq n^{1-\frac{j}{k}} F_k^{\frac{j}{k}} \;. \quad \square$$

Lemma 5 shows that if the classification of the items into sparse and dense categories is threshold consistent (which holds with high probability), then, the variance of the AMS estimator improves significantly, that is, $F_1' \cdot F_{2k-1}' \leq n^{1-\frac{2}{k+1}} \cdot F_k^2$, instead of, $F_1 F_{2k-1} \leq n^{1-\frac{1}{k}} F_k^2$, as proved in [1, 2].

Lemma 5. *Let $B \geq n^{1-\frac{2}{k+1}}$ and suppose that the classification of the items into sparse and dense categories is threshold consistent, with the sparse threshold $T_s = \frac{F_2}{B}$. Then,*

$$F_1' \cdot F_{2k-1}' \leq n^{1-\frac{2}{k+1}} \cdot F_k^2$$

Proof.

$$\begin{aligned}
F_{2k-1}' &\leq (\max_{i \text{ sparse}} f_i)^{k-1} F_k' \\
&\leq T_s^{\frac{k-1}{2}} F_k, \quad \text{by threshold consistency, } f_i \leq T_s, \text{ for } i \text{ sparse} \\
&= \left(\frac{F_2}{B}\right)^{(k-1)/2} F_k' \\
&\leq \left(\frac{F_2}{B}\right)^{\frac{k-1}{2}} F_k, \quad \text{since, } F_k' \leq F_k \\
&\leq \left(\frac{n^{1-\frac{2}{k}} F_k^{\frac{2}{k}}}{B}\right)^{\frac{k-1}{2}} F_k, \text{ since, } F_2 \leq n^{1-\frac{2}{k}} \cdot F_k^{\frac{2}{k}}, \text{ by Corollary 4,} \\
&\leq \left(\frac{n^{1-\frac{2}{k}}}{n^{1-\frac{2}{k+1}}}\right)^{\frac{k-1}{2}} F_k^{\frac{2}{k} \cdot \frac{(k-1)}{2} + 1}, \text{ since, } B \geq n^{1-\frac{2}{k+1}} \\
&= \left(n^{\frac{2}{k \cdot (k+1)}}\right)^{\frac{k-1}{2}} F_k^{2-\frac{1}{k}} \\
&= n^{\frac{k-1}{k \cdot (k+1)}} F_k^{2-\frac{1}{k}}
\end{aligned}$$

Substituting in the expression for $F_1' F_{2k-1}'$, and noting that $F_1' \leq F_1$, we have

$$\begin{aligned}
F_1' F_{2k-1}' &\leq F_1 \cdot n^{\frac{k-1}{k \cdot (k+1)}} \cdot F_k^{2-\frac{1}{k}} \\
&\leq n^{1-\frac{1}{k}} \cdot n^{\frac{k-1}{k \cdot (k+1)}} \cdot F_k^2, \qquad \text{since, } F_1 \leq n^{1-\frac{1}{k}} \cdot F_k^{\frac{1}{k}}, \text{ by Corollary 4} \\
&\leq n^{1-\frac{2}{k+1}} \cdot F_k^2 \;. \qquad \square
\end{aligned}$$

Tightness of the bound. We now show that the above bound is attained, to within constant factors, on the following input. Consider a distribution of frequencies over n items, where, there are $n^{1/(k+1)}$ items with frequency $n^{1/(k+1)}$, and the remaining items have frequency 1.

For this instance, $n < F_2 < 2n$. Therefore, $T_s = \frac{F_2}{B} > \frac{n}{n^{1-2/(k+1)}} = n^{\frac{2}{k+1}}$. It follows that there are no dense items. Therefore,

$$F'_{2k-1} = F_{2k-1} \sim n^{\frac{1}{k+1}} n^{\frac{2k-1}{k+1}} = n^{2-\frac{2}{k+1}}, \quad \text{and } F_k \sim n$$

where, $\sim$ denotes "within a factor of 2". Thus,

$$F'_1 F'_{2k-1} = F_1 F_{2k-1} \sim n \cdot n^{2-\frac{2}{k+1}} = n^{1-\frac{2}{k+1}} n^2 \sim n^{1-\frac{2}{k+1}} F_k^2,$$

showing that the bound is attained to within constant factors.

3.2 Analysis of Sparse Component Estimator

Overview of proof. The sparse component estimator E returns 0, if the size of a reduced sample is below a certain threshold $\beta \cdot u$. Clearly, such a strategy would not work if there is a significant probability that the reduced sample size becomes smaller than the threshold, and the sparse component remains a significant fraction of F_k. Our first goal is to eliminate this possibility. This is done in a few steps. First, we show that if the sparse component F'_k is at least $\epsilon \cdot F_k$, then, this implies that F'_1 is at least $c' \cdot F_1$, for some constant $c'(\epsilon)$. Next, observe that the size of each reduced sample group, on expectation, is $\frac{F'_1}{F_1} u$. Thus, by designing constants large enough, and using Chernoff's bounds, we can ensure that with high probability, if $F'_k \geq \epsilon \cdot F_k$, then there is a threshold $\beta = \beta(\epsilon)$, such that the size of each reduced sample is at least $\beta \cdot u$.

This line of argument is pursued in Lemmas 6, 7 and 8, and establishes that if the estimator encounters a small reduced sample (i.e., $u' < \beta \cdot u$), then, $F'_k < \epsilon \cdot F_k$, with high probability, implying that the dense component is a good approximation of the frequency moment. Since the dense component is always estimated closely, and with high confidence, an accurate estimation of F_k is returned.

Next, we assume that all reduced samples are large enough (i.e., $u'(a) \geq \beta \cdot u$, for every $1 \leq a \leq v$). Then, we use the bound on the variance of the AMS estimator, as proved in Lemma 5, together with Chebychev's inequality, to show that $\mathbf{Pr}\left\{|E'(a) - F'_k| \leq \epsilon \cdot F_k\right\} \geq \frac{4}{5}$, for every sample group a. Finally, the constant confidence is boosted by taking the median of $v = \Theta(\log \frac{1}{\delta})$ independent sample groups. This is proved in Lemmas 9 and 10.

We begin with the following elementary fact.

Lemma 6. *Suppose* $(a_1, a_2, \ldots, a_p)$ *and* $(b_1, b_2, \ldots, b_q)$ *are positive vectors such that* $\max_s a_s \leq \min_r b_r$. *Suppose that* $k \geq 2$, *and, for some* $\alpha > 0$, $\sum_{s=1}^p a_s^k \geq \alpha \sum_{r=1}^q b_r^k$. *Then* $\sum_{s=1}^p a_s \geq \alpha \sum_{r=1}^q b_r$.

Proof. Suppose $\sum_{s=1}^p a_s < \alpha \sum_{r=1}^q b_r$ and $\alpha > 0$. Then,

$$\frac{\sum_{s=1}^p a_s}{\sum_{s=1}^p a_s} > \frac{\alpha \sum_{r=1}^q b_r^k}{\alpha \sum_{r=1}^q b_r} = \frac{\sum_{r=1}^q b_r^k}{\sum_{r=1}^q b_r}$$

Thus, $\frac{\sum_{s=1}^{p} a_s}{\sum_{s=1}^{p} a_s} > \frac{\sum_{r=1}^{q} b_r^k}{\sum_{r=1}^{q} b_r}$. Cross multiplying and transferring the RHS to the LHS, we obtain,

$$\sum_{s,r} a_s^k b_r - a_s b_r^k > 0, \quad \text{or equivalently,} \quad \sum_{s,r} a_s \cdot b_r (a_s^{k-1} - b_r^{k-1}) > 0 \ .$$

This contradicts the assumption that $\max_{s=1}^{p} a_s \leq \min_{r=1}^{q} b_r$. □

The following lemma quantifies the following idea, that if the sparse component is larger than ϵ times the dense component (meaning, that it is large enough to be considered), that is, $F_k' > \epsilon \cdot F_k^{(d)}$, then, there is some function $c'(\epsilon)$ such that, the sum of the sparse frequencies is at least c' times the sum of the dense frequencies, that is, $F_1' > c' \cdot F_1^{(d)}$. A proof of this intuitive property assures us that with the appropriate choice of constants, the process of reduction retains a significant proportion of samples, provided, the sparse component was large enough to begin with (i.e., $F_k' > \epsilon \cdot F_k^{(d)}$).

Lemma 7. *Suppose that the classification of the items into dense and sparse categories is threshold consistent, with sparse threshold $T_s = \frac{F_2}{B}$. If $F_k' > \epsilon F_k^{(d)}$, then, $F_1' > \frac{\epsilon}{\lambda^{k-1}} F_1^{(d)}$.*

Proof. For a sparse element s, let $a_s = f_s^2$, and for a dense element r, let $b_r = \lambda \cdot f_r$. By threshold consistency, it follows that

$$a_s^2 \leq T_s = \frac{F_2}{B}, \quad \text{and} \quad b_r^2 \geq T_d = \frac{F_2}{\lambda^2 \cdot B}$$

We therefore have the following.

$$F_k' > \epsilon F_k^{(d)} \quad \Leftrightarrow \quad \sum_s a_s^k > \frac{\epsilon}{\lambda^k} \sum_r b_r^k \quad \Leftrightarrow \quad \sum_s a_s^k > \frac{\epsilon}{\lambda^k} \sum_r b_r^k$$

The right most relation allows us to use Lemma 6, to yield, $\sum_s a_s > \frac{\epsilon}{\lambda^k} \sum_r b_r$. By definitions of a_s and b_r, we have,

$$\sum_s a_s > \frac{\epsilon}{\lambda^k} \sum_r b_r \quad \Leftrightarrow \quad \sum_s f_s > \frac{\epsilon}{\lambda^k} \lambda (\sum_r f_r) \quad \Leftrightarrow \quad F_1' > \frac{\epsilon}{\lambda^{k-1}} F_1^{(d)} \ . \square$$

Lemma 8 observes that, a sample contains a sparse item with probability F_1'/F_1. Therefore, using appropriate constants, Chernoff's bounds and Lemma 7, Lemma 8 argues that if the sparse component is large enough, then, there are enough reduced samples (with high probability).

Lemma 8. *Let $0 < \epsilon < \frac{1}{2}$ and $0 < \delta < 1$. There exists a constant $\beta = \beta(\epsilon)$, such that, if $F_k' > \epsilon \cdot F_k^{(d)}$ and $u > \frac{8}{\beta} \log \frac{1}{\delta}$, then, $u' > \beta u$, with probability at least $1 - \delta$.*

Proof. Consider a single AMS estimator $X[a][c]$. The probability that this estimator refers to a sparse element is $p = F_1'/F_1$, since a given sample is equally likely to refer to each arrival over the stream. Since, $F_k' > \epsilon F_k^{(d)}$, it follows from Lemma 7

that $F_1' > \frac{\epsilon}{\lambda^{k-1}} F_1^{(d)}$. Let γ denote the constant $\frac{\epsilon}{\lambda^{k-1}}$. Thus, $p > \frac{\gamma}{1+\gamma}$, for brevity, let $\beta = \frac{\gamma}{2(1+\gamma)}$. In a group of u independent estimators, u' denotes the number of estimators that refer to sparse elements. Thus, $\mathbf{E}\big[u'\big] = p \cdot u > 2\beta u \geq 16 \cdot \log \frac{1}{\delta}$. By Chernoff's bounds,

$$\mathbf{Pr}\left\{u' > \beta u\right\} \geq \mathbf{Pr}\left\{u' > \frac{1}{2}\mathbf{E}\big[u'\big]\right\} \geq 1 - e^{-\frac{\mathbf{E}\left[u'\right]}{8}} \geq 1 - e^{-\log \frac{1}{\delta}} > 1 - \delta \qquad \square$$

Lemma 9 applies Lemmas 6 through Lemma 8, and presents a sufficient condition for the accurate estimation of F_k'. Note that, by accurate estimation of F_k', we mean an estimate E such that E is within $F_k' - \epsilon \cdot F_k \leq E \leq F_k' + \epsilon \cdot F_k$ (and not the usual $E \in (1 \pm \epsilon)F_k'$).

Lemma 9. *Let $0 < \epsilon < \frac{1}{2}$ and $0 < \delta < 1$ and suppose that the classification of the items into sparse and dense categories is threshold consistent. Then, there exists a constant $\beta = \beta(\epsilon)$, such that, if the following premise holds,*

$$F_k' > \epsilon F_k^{(d)} \text{ and } u > \frac{8}{\beta}\log\frac{1}{\delta}, u > \frac{8k \cdot F_1' \cdot F_{2k-1}'}{\beta \cdot \epsilon^2 F_k^2}, v > 5\log\frac{1}{\delta} \text{ and } v \cdot \delta \leq \frac{3}{40}$$

then, the following statement is true.

$$\mathbf{Pr}\left\{|E - F_k'| > \epsilon F_k \text{ and } u'(a) > \beta u, 1 \leq a \leq v\right\} < (v+1)\delta.$$

Proof. Suppose that $F_k' > \epsilon' F_k^{(d)}$. By Lemma 8, and using the value of β provided there, with probability at least $1-\delta$, the size of the reduced sample in an AMS estimator group is at least $u' > \beta u$. Using union bounds, it follows that this property holds for each of the the v estimator groups with a probability of at least $1 - v \cdot \delta$.

Fix a value of the group index a. Using Chebychev's inequality,

$$\mathbf{Pr}\left\{|E'(a) - F_k'| > \epsilon F_k\right\} < \frac{\mathbf{Var}\big[E'(a)\big]}{\epsilon^2 F_k^2} \tag{2}$$

It follows from [1, 2] that

$$\mathbf{Var}\big[E'(a)\big] = \frac{kF_1'F_{2k-1}'}{u'(a)}$$

Substituting in (2), we have,

$$\mathbf{Pr}\left\{|E'(a) - F_k'| > \epsilon F_k\right\} \leq \frac{k}{\epsilon^2} \cdot \frac{F_1'F_{2k-1}'}{u'(a) \cdot F_k^2} \leq \frac{k}{\beta \cdot \epsilon^2} \cdot \frac{F_1'F_{2k-1}'}{uF_k^2} < \frac{1}{8} + v\delta < \frac{1}{5}$$

assuming that $v\delta \leq \frac{1}{40}$. Since $v > 5\log\frac{1}{\delta}$, by taking the median E of the family $\{E'(a)\}_{1 \leq a \leq v}$, $\mathbf{Pr}\left\{|E(a) - F_k'| > \epsilon F_k\right\} < \delta$. By union bound, the total error probability is at most $(v+1)\delta$. $\square$

Lemma 10 summarizes the main property of the sparse estimator.

Lemma 10. *Let $\epsilon < 1/2$ and $\delta < \frac{1}{128}$. There exists a constant $\beta = \beta(\epsilon)$, such that, if $u \geq 8 \cdot k \cdot n^{1-\frac{2}{k+1}}/(\beta \cdot \epsilon^2)$ and $v \geq 5 \log \frac{1}{\delta}$, then the following statements hold.*

1. *If there exists a sample group such that $u' < \beta u$, then, $E = 0$ and $\mathbf{Pr}\left\{F'_k \leq \epsilon \cdot F_k^{(d)}\right\} \geq 1-(v+1)\cdot\delta$. Thus, $\mathbf{Pr}\left\{|E - F'_k| < \epsilon \cdot F_k\right\} \geq 1-(v+1)\cdot\delta$.*
2. *Otherwise, for any sample group, $\mathbf{Pr}\left\{|E - F'_k| < \epsilon \cdot F_k\right\} \geq 1 - (v+1) \cdot \delta$.*

Proof. Follows directly from Lemmas 9 and 5. □

3.3 Analysis of Hybrid Estimator

Recall that the *hybrid estimator* H returns the sum of the dense component estimator and the sparse component estimator, that is, it returns $H = D + E$. Lemma 11 presents theaccuracy and confidence guarantees of the hybrid estimator.

Lemma 11. *Let $\epsilon < 1/2$ and $\delta < \frac{1}{128}$. There exists constants $\beta = \beta(\epsilon)$ and $\lambda = \lambda(\epsilon)$, such that, if, $u \geq 8 \cdot k \cdot n^{1-\frac{2}{k+1}}/(\beta \cdot \epsilon^2)$, $v \geq 5 \log \frac{m}{\delta}$ and $B = O(\frac{k}{\epsilon} \cdot n^{1-\frac{2}{k+1}} \cdot \log \frac{m}{\delta})$, then, $\mathbf{Pr}\left\{|H - F_k| \leq \epsilon F_k\right\} > 1 - \delta$.* □

Proof. We invoke Lemmas 2 and 10, using accuracy parameter $\epsilon' = \frac{\epsilon}{2}$ and confidence parameter, $\delta' = \frac{\delta}{v+2}$, each. There are two cases, namely, (a) there exists a reduced sample group, whose size $u' < \beta \cdot u$, or, (b) all reduced groups have size at least $\beta \cdot u$.

Case 1. There is at least one value of a, for which $u'(a) < \beta u$. The estimator E is set to 0 in this case, and therefore $H = D$. Suppose that $F'_k > \epsilon' F_k^{(d)}$, then by Lemma 10, the probability of this observation is at most $(v+1)\delta'$. Therefore, $F'_k \leq \epsilon' F_k^{(d)}$ holds with probability at least $1 - (v+1) \cdot \delta'$. Since,

$$|H - F_k| = |D - F_k| = |D - (F_k^{(d)} + F'_k)| \leq |D - F_k^{(d)}| + F'_k$$

it follows that,

$$\begin{aligned}\mathbf{Pr}\left\{|H - F_k| \leq \epsilon F_k\right\} &\geq \mathbf{Pr}\left\{|D - F_k^{(d)}| \leq \epsilon' F_k \text{ and } F'_k \leq \epsilon' F_k\right\} \\ &\quad \mathbf{Pr}\left\{|D - F_k^{(d)}| \leq \epsilon' F_k^{(d)} \text{ and } F'_k \leq \epsilon' F_k\right\} \\ &\quad \geq 1 - \delta' - (v+1)\delta' \geq 1 - \delta,\end{aligned}$$

by Lemmas 2, 10 and the union bound.

Case 2. For every value $a, 1 \leq a \leq v$, $u'(a) \geq \beta u$. Then,

$$\begin{aligned}\mathbf{Pr}\left\{|H - F_k| \leq \epsilon F_k\right\} &= \mathbf{Pr}\left\{|D + E - F_k| \leq \epsilon F_k\right\} \\ &\geq \mathbf{Pr}\left\{|D - F_k^{(d)}| \leq \epsilon' F_k \text{ and } |E - F'_k| \leq \epsilon' F_k\right\} \\ &\geq 1 - \delta' - (v+1)\delta' \geq 1 - \delta,\end{aligned}$$

by Lemmas 2, 10 and the union bound. □

We can now prove the main theorem in the paper.

Proof (of Theorem 1.). By the argument in the proof of Lemma 7, it follows that $\beta = \frac{\epsilon}{2}\left(\frac{1-\epsilon}{1+\epsilon}\right)^{2(k-1)}$. The space complexity of the procedure is $\tilde{(}u \cdot v + \frac{B}{\epsilon})$. Using Lemma 11, together with the above value of β, and $B = n^{1-\frac{2}{k+1}}$, we obtain the statement of the theorem. □

4 Conclusions

The paper presents a hybrid method for estimating the k^{th} frequency moment, for $k > 2$, for data streams using space $\tilde{O}(n^{1-\frac{2}{k+1}})$ bits. It is based on a non-trivial combination of the classical algorithm for estimating moments [1, 2], and the COUNTSKETCH technique for estimating frequent items over a stream. The space complexity of the hybrid algorithm is better than the known space complexity of $\tilde{O}(n^{1-\frac{1}{k-1}})$ for this problem [7, 10]. Somewhat surprisingly, the hybrid technique does not use any ideas from [7, 10].

References

1. Noga Alon, Yossi Matias, and Mario Szegedy. "The Space Complexity of Approximating the Frequency Moments". In *Proceedings of the 28th Annual ACM Symposium on the Theory of Computing STOC, 1996*, pages 20–29, Philadelphia, Pennsylvania, May 1996.
2. Noga Alon, Yossi Matias, and Mario Szegedy. "The space complexity of approximating frequency moments". *Journal of Computer Systems and Sciences*, 58(1):137–147, 1998.
3. Ziv Bar-Yossef, T.S. Jayram, Ravi Kumar, and D. Sivakumar. "An information statistics approach to data stream and communication complexity". In *Proceedings of the 34th ACM Symposium on Theory of Computing (STOC), 2002*, pages 209–218, Princeton, NJ, 2002.
4. Ziv Bar-Yossef, T.S. Jayram, Ravi Kumar, D. Sivakumar, and Luca Trevisan. "Counting distinct elements in a data stream". In *Proceedings of the 6th International Workshop on Randomization and Approximation Techniques in Computer Science, RANDOM 2002*, Cambridge, MA, 2002.
5. Amit Chakrabarti, Subhash Khot, and Xiaodong Sun. "Near-Optimal Lower Bounds on the Multi-Party Communication Complexity of Set Disjointness". In *Proceedings of the 18th Annual IEEE Conference on Computational Complexity, CCC 2003*, Aarhus, Denmark, 2003.
6. Moses Charikar, Kevin Chen, and Martin Farach-Colton. "Finding frequent items in data streams". In *Proceedings of the 29th International Colloquium on Automata Languages and Programming*, 2002.
7. Don Coppersmith and Ravi Kumar. "An improved data stream algorithm for estimating frequency moments". In *Proceedings of the Fifteenth ACM SIAM Symposium on Discrete Algorithms*, New Orleans, LA, 2004.
8. Joan Feigenbaum, Sampath Kannan, Martin Strauss, and Mahesh Viswanathan. "An Approximate L^1-Difference Algorithm for Massive Data Streams". In *Proceedings of the 40th Annual IEEE Symposium on Foundations of Computer Science*, New York, NY, October 1999.
9. Philippe Flajolet and G.N. Martin. "Probabilistic Counting Algorithms for Database Applications". *Journal of Computer Systems and Sciences*, 31(2):182–209, 1985.

10. Sumit Ganguly. “Estimating Frequency Moments of Update Streams using Random Linear Combinations”. *To appear in the Proceedings of the 8th International Workshop on Randomization and Approximation Techniques in Computer Science*, RANDOM 2004.
11. Sumit Ganguly, Minos Garofalakis, and Rajeev Rastogi. “Processing Set Expressions over Continuous Update Streams”. In *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, San Diego, CA, 2003.
12. Anna Gilbert, Sudipto Guha, Piotr Indyk, Yannis Kotidis, S. Muthukrishnan, and Martin Strauss. “Fast Small-space Algorithms for Approximate Histogram Maintenance”. In *Proceedings of the 34th ACM Symposium on Theory of Computing (STOC), 2002*, Montreal, Canada, May 2002.
13. Piotr Indyk. “Stable Distributions, Pseudo Random Generators, Embeddings and Data Stream Computation”. In *Proceedings of the 41st Annual IEEE Symposium on Foundations of Computer Science*, pages 189–197, Redondo Beach, CA, November 2000.
14. M. Saks and X. Sun. “Space lower bounds for distance approximation in the data stream model”. In *Proceedings of the 34th ACM Symposium on Theory of Computing (STOC), 2002*, 2002.