Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2004-044
AIM-2004-014

July 2, 2004

Optimal Approximations of the Frequency Moments

Piotr Indyk and David Woodruff

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

Optimal Approximations of the Frequency Moments

Piotr Indyk
MIT

David Woodruff*
MIT

Abstract

We give a 1-pass $\tilde{O}(m^{1-2/k})$-space algorithm for computing the k-th frequency moment of a data stream for any real $k > 2$. Together with the lower bounds of [1, 2, 3], this resolves the main problem left open by Alon et al in 1996 [1]. Our algorithm additionally works for streams with deletions and thus gives an $\tilde{O}(m^{1-2/p})$ space algorithm for the L_p difference problem for any $p > 2$. This essentially matches the known $\Omega(m^{1-2/p-o(1)})$ lower bound of [10, 2]. Finally the update time of our algorithms is $\tilde{O}(1)$.

1 Introduction

Computing over data streams is a recent phenomenon that is of growing interest in many areas of computer science, including databases, computer networks and theory of algorithms. In this scenario, it is assumed that the algorithm sees a stream of elements one-by-one in arbitrary order, and needs to compute certain statistics over the input. However, it does not have enough memory to store the whole stream. Therefore, it must maintain only a sketch of the data, which is small but nevertheless powerful enough to compute the desired statistics. For a survey, see [9].

Within the theory community, the recent interest in these problems can be traced to the paper of Alon, Matias and Szegedy [1]. In that influential paper, the authors give upper and lower bounds for the space complexity of streaming algorithms for the problem of estimating the *frequency moments* of a stream. In this problem, the stream is a sequence $\mathcal{S}$ of elements $a_1, \ldots, a_n$, $a_j \in [m] = \{1 \ldots m\}$. For each element $i \in [m]$, its frequency f_i in $\mathcal{S}$ is the number of times it occurs in $\mathcal{S}$. The k-th frequency moment (for real $k \geq 0$) of $\mathcal{S}$ is defined as

$$F_k = F_k(\mathcal{S}) = \sum_{i \in [m]} f_i^k$$

An algorithm A (ϵ, δ)-approximates F_k if A outputs a number $\tilde{F}_k$ for which $\Pr[|\tilde{F}_k - F_k| > \epsilon F_k] < \delta$. In [1], the authors present $\tilde{O}(1)$[1] space algorithms which (ϵ, δ)-approximate F_k for integers $k \leq 2$, and $\tilde{O}(m^{1-1/k})$ space algorithms for integers $k \geq 3$. For $k > 5$, they also show an $\Omega(m^{1-5/k})$ lower bound. Their ideas spurred vigorous research on problems in this area. Soon, their algorithmic results (for the case $k \leq 2$) were shown [6, 7] to hold for a more general problem of estimating the k-th norm of the vector $(f_1, \ldots, f_m)$; see the definition below. Later, a line of research [2, 3]

*Supported by an NDSEG fellowship.

[1]We write $f(m) = \tilde{O}(g(m))$ to denote that $f(m) = g(m) \log^{O(1)} m \log^{O(1)} n \left(\frac{1}{\epsilon}\right)^{O(1)} \log 1/\delta$; the $\tilde{\Omega}$ notation is defined similarly.

culminated in an $\tilde{\Omega}(m^{1-2/k})$ space lower bound for estimating F_k for any real $k > 2$. Very recently, a better upper bound of $\tilde{O}(m^{1-1/(k-1)})$ was shown in [5].

In this paper, we present a one-pass $\tilde{O}(m^{1-2/k})$-space streaming algorithm for estimating F_k for any real $k > 2$. By the aforementioned results, our space bound is tight (up to $\tilde{O}(1)$ factors). Thus, together with the lower bound of [2, 3], our result resolves the main problem left open by [1], and is therefore a classical "last step" result, resolving a long line of research on the frequency moments. We note that, in addition to matching the known space lower bound, the update time of our algorithm per stream element is exponentially smaller than the previous best-known space algorithm [5]. Indeed, the algorithm in [5] requires $m^{\Omega(1)}$ update time, while ours is $\tilde{O}(1)$, and the suppressed terms in the $\tilde{O}$ notation are small.

Our algorithm has the additional feature of being able to estimate F_k over streams with *deletions*, instead of just the update-only model defined above. Our algorithm can therefore be used to estimate L_p differences, $p > 2$, in one-pass with space $\tilde{O}(m^{1-2/p})$. The L_p difference problem is to approximate $\sum_{i=1}^{m} |a_i - b_i|^p$, where the a_i's and b_i's arrive in arbitrary order. Prior to our work the best known algorithm used space $\tilde{O}(m^{1-1/(p-1)})$ [5]. Since there is an $\Omega(m^{1-2/p-o(1)})$ space lower bound for this problem [10, 2], our algorithm is essentially optimal.

Our techniques: The earlier algorithms for estimating F_k were obtained by constructing a single estimator, which was shown to equal F_k in expectation, and to have small variance. Our algorithm departs from this approach. Instead, the main idea of our algorithm is as follows. First, we (conceptually) divide the elements into classes S_i, such that the elements in class S_i have[2] frequency $\approx 8^i$. We observe that, in order for the elements in S_i to contribute significantly to the value of F_k, it must be the case that the size s_i of S_i is comparable to the size of $S_1 \cup \ldots \cup S_{i-1}$. If this is the case, we will find an element from S_i if we restrict the stream to an $\approx 1/s_i$ fraction of universe elements, and then find the most frequent elements in the substream. The contribution of S_i to F_k can then be approximately estimated by $s_i \cdot (8^i)^k$. By summing up all estimated contributions, we obtain an estimator for F_k.

Unfortunately, finding the most frequent element in a general stream (even approximately) requires storage that is linear in the stream size. However, if the distribution of the stream elements is not very "heavy-tailed", then a more efficient algorithm is known [4]. This more efficient method is based on the sketching algorithm for F_2 given in [1], or equivalently, a version of the Johnson-Lindenstrauss [8] type dimension reduction for L_2. We show that the streams generated by our algorithm (for S_i's that contribute to F_k), satisfy this "tail property", and thus we can use the algorithm of [4] in our algorithm.

A somewhat technical issue that arises in implementing the above algorithm is a need to classify a retrieved element i into one of the classes. For this, we need to know f_i. This information is easy to obtain using a second pass. In the absence of that, we use the estimation of f_i provided by the algorithm of [4], but we define the thresholds defining the classes randomly, to ensure that the error in estimating the frequencies is unlikely to result in misclassification of a frequency. This technique, of dividing the items into frequency classes and estimating the size of each class, is of independent interest, and may prove useful for estimating other statistics in the streaming model.

[2] The constant 8 here is arbitrary.

2 The algorithm

Before presenting the full algorithm, we first walk through its important points. We use the term *substream* to refer to a restriction of the input stream to a subset of universe elements. Our main algorithm will (in a parallel, online fashion) generate substreams and use a helper algorithm called Summary to find and report frequent items in these substreams. Summary will in turn use the algorithm of [4] mentioned above to find these items. Intuitively, the algorithm of [4] uses small space to output approximations to all frequencies of items which occur in its input stream. We treat the algorithm of [4] as a black box whose properties are summarized by the following theorem:

Theorem 1 *There exists a* 1*-pass* $O(b \log m \log n \log 1/\delta)$*-space algorithm* $A(\mathcal{S}, b, \delta)$*, where* $\mathcal{S}$ *is an arbitrary stream of size* n *with elements drawn from* $[m]$*,* $b > 0$ *is any integer, and* $\delta > 0$ *is any real, such that, with probability at least* $1 - \delta$*,* A *outputs a string CountSketch*$(\mathcal{S}, b, \delta)$ *which can be used to compute an integer* $\tilde{f}_j$ *for every* $j \in [m]$ *with the property that*

$$f_j - \sqrt{F_2/b} \leq \tilde{f}_j \leq f_j + \sqrt{F_2/b}.$$

We note that A, as defined in this theorem, is not exactly the main algorithm given in [4], but rather a slight modification of it. Specifically, in lemma 3 of [4] they set a parameter $t = O(\log n + \log 1/\delta)$. Since we only need CountSketch$(\mathcal{S}, b, \delta)$ to be correct after the entire stream is processed, rather than after each position in the stream as required in [4], we can drop the $O(\log n)$ term from t. However, since we want it to be correct for all m universe elements, we need to add an $O(\log m)$ term. Hence we require $t = O(\log m + \log 1/\delta)$, which results in the claimed space bound. See [4] for more details.

We see that the output CountSketch of algorithm A gives a compact way of *summarizing* the frequency distribution of the stream $\mathcal{S}$. However, the approximations that CountSketch reports for items which occur relatively few times are heavily skewed by the additive error inherent in A. We thus would like a way of filtering out those frequencies which are likely to be far from their true value. For this, we introduce a helper algorithm Summary. Summary works by approximating the additive error of A itself by approximating $F_2 = F_2(\mathcal{S})$. This can be done in $\tilde{O}(1)$ space using the F_2 approximation algorithm of [1] or of [11]. Summary can then use its knowledge of the additive error to recognize if certain frequencies are on the same order of magnitude as the additive error, and filter these estimates out.

More precisely, Summary will be given a list of frequency ranges and its job will be, for each of these ranges, to report whether $\mathcal{S}$ has an item with frequency falling into the given range. To be compatible with our main algorithm and for technical reasons explained below, the frequency ranges will be specified by three parameters, the stream size n (which Summary easily determines from its input stream $\mathcal{S}$), and two real numbers $0 < \gamma, x < 1$. The ranges are then of the form $[x(1+\gamma)^i, x(1+\gamma)^{i+1}]$, for all $i \in \{0, \ldots, \lceil \log_{1+\gamma} n/x \rceil - 1\}$. Summary will also be given an integer b and a real δ which correspond to the accuracy and confidence parameters it will run algorithm A with. After running algorithm A and approximating A's additive error, Summary will filter out those estimated frequencies whose logarithm is smaller than a certain threshold. This threshold, denoted $z_{\min}$ in the algorithm description, is a function of the additive error of A and an input parameter τ, and its choice will be explained as we develop the intuition for the main algorithm. Finally, for those estimated frequencies which have not been filtered, Summary will set an indicator bit in a bitvector y for the corresponding range. Summary's output is y. The algorithm follows:

Summary$(\mathcal{S}, b, \gamma, x, \delta, \tau)$
Input: a stream $\mathcal{S}$, an integer b, and reals $\gamma, x, \delta, \tau > 0$
Output: $y \in \{0,1\}^{\lceil \log_{1+\gamma} n/x \rceil}$

1. In parallel,
   (a) Run $A(\mathcal{S}, b, \delta/2)$ and let CountSketch$(\mathcal{S}, b, \delta/2)$ be its ouput.
   (b) Compute a $(1 \pm 1/2)$-approximation $\tilde{F}_2$ to $F_2(\mathcal{S})$ with error probability $\delta/2$.
2. Initialize $y \leftarrow 0^{\lceil \log_{1+\gamma} n/x \rceil}$, where n is the length of $\mathcal{S}$, and set $z_{\min} = \lceil \log_{1+\gamma} \tau \sqrt{\frac{2\tilde{F}_2}{b}}/x \rceil$.
3. For $j = 1$ to m, or until $y_z = 1^{\lceil \log_{1+\gamma} n/x \rceil}$,
   - Compute $\tilde{f}_j$ from CountSketch$(\mathcal{S}, b, \delta/2)$.
   - Let $z = \lfloor \log_{1+\gamma} \tilde{f}_j/x \rfloor$.
   - If $z \geq z_{\min}$, set y_z to 1.
4. Output y.

2.1 Intuition for the main algorithm

We now give more intuition for our algorithm than that provided in the introduction. Recall that we want to estimate the size of each frequency range in the input stream, and then compute the sum over all ranges, of the kth power of a representative of the range times the range's size.

2.1.1 Randomizing the borders of frequency ranges

To estimate the size of a range, we restrict our input stream to successively smaller substreams until we find a substream containing only a few items in the range. The reason we do this is that it may be the case that a frequency range noticeably contributes to F_k but that we can't detect it in the original stream since it is dominated by more frequent items. That is, the algorithm Summary can only find the most frequent items in its input stream. However, if the range noticeably contributes, it will hold that when restricting to a substream containing only a few items in the range, these items will be amongst the most frequent.

To determine if a range has a representative in a given substream, we use Summary. One problem that occurs if the frequency ranges are fixed in advance is that an item whose frequency is on the border of its range may be incorrectly classified into a neighboring range. For example, suppose we had three frequency ranges $[a, b)$, $[b, c)$, and $[c, d)$ for reals $0 < a < b < c < d$, and we wanted to classify items into one of these three ranges. If we had three items r, s, t with $a \leq f_r < b$, $b \leq f_s < c$, and $c \leq f_t < d$, then we would want Summary to report the existence of items in all three of these ranges. However, if $f_r \approx b$ and $f_t \approx c$, both r and t may be misclassified because of the additive error of A: that is, we may have $\tilde{f}_r, \tilde{f}_s, \tilde{f}_t \in [b, c)$, causing Summary to only report the existence of items in $[b, c)$. This deteriorates our estimate of F_k.

A natural approach to avoid this problem is to *randomize* the borders of frequency ranges. This will ensure that the *estimated frequencies* $\tilde{f}$ that Summary finds are a good distance from their

nearest border. Observe that if $\tilde{f}$ is far from its nearest border, then if $\tilde{f}$ and f lie in different ranges, the distance $|\tilde{f} - f|$ must be at least the distance of $\tilde{f}$ to its nearest border. We will ensure that the distance from $\tilde{f}$ to its nearest border is larger than the additive error of A, and hence f and $\tilde{f}$ must lie in the same range.

To this end, we choose our borders randomly by choosing a random number x and define our ranges to be $S_i = \{j \in [m] \mid x(1+\epsilon')^i \leq f_j < x(1+\epsilon')^{i+1}\}$ for $0 \leq i < \log_{1+\epsilon'} n/x$ and $\epsilon' = \Theta(\epsilon)$. We use $\epsilon' = \epsilon/(3k)$ rather than ϵ itself since there will be two sources of error: (1) the estimated size of the frequency ranges, and (2) the contribution of a frequency range to F_k due to the fact that we fix a generic representative of the range and take its kth power.

To ensure we don't have too many ranges, we lower bound x by a constant by choosing $x \in [1/2, 1]$. The analysis shows we need only choose the first $O(\log \tau)$ bits of x for a parameter τ. We will set $\tau = O\left(\frac{1}{\epsilon^3} \log_{1+\epsilon} n \log m \log\log_{1+\epsilon} n\right)$, which we now justify. The idea is that, by choosing $O(\log \tau)$ bits of x we can ensure that a given estimated frequency $\tilde{f}$ is at a distance of at least f/τ from its borders with probability $1 - O(1/(\tau\epsilon))$. With our choice of parameters, f/τ will be less than the additive error of algorithm A, and by the reasoning above, $\tilde{f}$ and f must lie in the same range. As explained below, we will invoke Summary $O(\epsilon^{-2} \log m \log\log_{1+\epsilon} n)$ times, and therefore our choice of τ allows us to assume by a union bound that with constant probability, over all invocations of Summary and over all $O(\log_{1+\epsilon} n)$ frequency ranges, the *first* estimated frequency $\tilde{f}$ that Summary encounters in each range will be at a distance of at least f/τ from its borders. Since Summary marks its output bitvector based on the first estimated frequency it sees in each range, this will be sufficient. See section 3 for a more formal treatment of these ideas.

We say a range S_i *contributes* if $|S_i|(1+\epsilon')^{ik} = \Omega(\epsilon F_k / \log_{1+\epsilon} n)$. Hence, S_i contributes if its size times the kth power of a frequency in its range is an ϵ-fraction of the average contribution of a range to F_k. To estimate F_k we can restrict to contributing frequency ranges.

2.1.2 Finding items in substreams

After choosing x, we form $O(\log m)$ substreams $\mathcal{S}_j$ of the input stream $\mathcal{S}$. The jth substream $\mathcal{S}_j$ is formed by sampling roughly $m/8^j$ elements of the universe $[m]$, and restricting $\mathcal{S}$ to these elements. This is done by choosing $O(\log m)$ pairwise-independent hash functions $h_1, \ldots, h_{O(\log m)}$ mapping $[m]$ to $[8]$ and restricting $\mathcal{S}$ to elements in $\cap_{p=0}^{j} h_p^{-1}(0)$, where h_0 is defined to be the zero function. Pairwise independence will be enough to give constant probability that the smallest j for which an item in S_i appears in $\mathcal{S}_j$ is approximately $\log_8 |S_i|$. Note that $\mathcal{S}_j$ contains $\mathcal{S}_{j+1}$.

For each substream $\mathcal{S}_j$, we run Summary on $\mathcal{S}_j$, which returns a bit vector y summarizing the frequency distribution as explained in the previous section. To increase the accuracy and confidence of our estimates, we repeat the above sampling process $t' = O(\log\log_{1+\epsilon} n/\epsilon^2)$ times, denoting our hash functions by $h_{j,t}$, our streams by $\mathcal{S}_{j,t}$, and the outputs of Summary by $a^{j,t}$ for $0 \leq j \leq O(\log m)$ and $1 \leq t \leq t'$. To get a constant factor approximation to $|S_i|$ for each i, our algorithm first looks at the first $t'' = O(\log\log_{1+\epsilon} n)$ runs of the above process. For each i, it checks if there is some j for which there are more than $t''/3$ values of t for which $a^{j,t} = 1$ (clearly there can be at most two such j). We choose $t''/3$ instead of $t''/2$ since the smallest j for which an element of S_i appears in $\mathcal{S}_j$ may oscillate between two values over the different values of t. The algorithm then constructs a set U containing exactly those i for which there is such a value j, and for each such i it defines Z_i to be the maximum j for which $a^{j,t} = 1$ for more than $t''/3$ values of t.

By a Chernoff and a union bound (explaining our choice of t''), together with an analysis of

Summary (developed below), we can ensure that for all contributing S_i, it is likely that $i \in U$ and that 8^{Z_i} is a rough approximation to $|S_i|$. To get a better approximation we now look at those t between t'' and t'. We observe that for $i \in U$, the average T_i of $t' - t'' = O(\log \log_{1+\epsilon} n/\epsilon^2)$ indicator variables indicating whether or not an item in S_i appears in $\mathcal{S}_{Z_i+O(1),\,t}$ for $t \in \{t''+1, t''+2, \ldots, t'\}$, is tightly concentrated about its expectation. Choosing to sum the indicators $a_i^{Z_i+O(1),\,t}$ rather than the $a_i^{Z_i,\,t}$ is purely a matter of constants, simplifying the analysis. By a Chernoff and a union bound (explaining our choice of t'), we will have that $8^{Z_i+2}T_i$ is a $(1 \pm \Theta(\epsilon))$ approximation to $|S_i|$ if S_i contributes.

2.1.3 Analyzing the behavior of Summary

The crux of our analysis will be to show that with our parameters the following holds with constant probability: for all substreams, (1) Summary only reports the existence of elements with certain frequencies if they occur in the substream, and (2) Summary reports the existence of items in contributing frequency ranges when they do occur in the substream.

Property (1) follows from the fact that Summary filters frequencies which are too small. More precisely, recall that Summary only accepts frequencies $\tilde{f}_j$ if $z = \lfloor \log_{1+\gamma} \tilde{f}_j/x \rfloor \geq z_{min}$ for a parameter $z_{min} = \lceil \log_{1+\gamma} \frac{\tau}{x} \sqrt{\frac{2\tilde{F}_2}{b}} \rceil$. This inequality is equivalent to $\tilde{f}_v/\tau \geq \sqrt{\frac{F_2(\mathcal{S}_{j,t})}{b}}$, where the latter quantity is just the additive error of A. Since our randomization only ensures $\tilde{f}_j$ is at a distance of f_j/τ from its borders, if the additive error is large enough (so that $z < z_{min}$), $\tilde{f}_j$ and f_j may be in different ranges. In this case $\tilde{f}_j$ should be rejected, since it deteriorates our estimator.

Property (2) gives a converse: for all items v in contributing S_i, the additive error of A will be less than f_v/τ, and Summary will report them. To see this, note that the additive error is controlled by a parameter b, which we define to be

$$b = O(\tau^2 m^{1-2/k} \epsilon^{-2-2/k} \log m \log \log_{1+\epsilon} n \log_{1+\epsilon}^{2/k} n).$$

By a Markov and a union bound, it is likely that for every substream $\mathcal{S}_{j,t}$ we have $F_2(\mathcal{S}_{j,t}) \leq O(\epsilon^{-2} \log m \log \log_{1+\epsilon} n)\mathbf{E}[F_2(\mathcal{S}_{j,t})]$. The analysis will show $\mathbf{E}[F_2(\mathcal{S}_{j,t})]/(\epsilon^{-2/k} m^{1-2/k} \log_{1+\epsilon}^{2/k} n)$ is approximately f_v^2 if $v \in S_i$, $|S_i| \approx m/8^j$, and S_i contributes. Thus for v in contributing S_i, our choice of b makes f_v/τ less than A's additive error, and Summary will accept.

Parameters: We summarize the parameters used by our algorithm here. The only parameter left to motivate is our choice of δ', the confidence parameter we input to Summary. As explained above, the number of invocations of Summary is exactly $O(\epsilon\tau/(\log_{1+\epsilon} n))$, which we set to be the reciprocal of δ', and thus by a union bound, our choice of δ' allows us to assume that on all invocations, Summary does not internally fail.

- $\epsilon' = \epsilon/(3k)$
- $\tau = O(\epsilon^{-3} \log_{1+\epsilon} n \log m \log \log_{1+\epsilon} n)$
- $\delta' = O\left(\frac{\log_{1+\epsilon} n}{\epsilon\tau}\right)$
- $b = O(\tau^2 m^{1-2/k} \epsilon^{-2/k} \log m \log_{1+\epsilon}^{2/k} n \log \log_{1+\epsilon} n)$

2.2 The algorithm

The algorithm follows. It can be implemented in one pass by parallelizing all calls to Summary.

F_k-**Approximator**$(\mathcal{S}, \epsilon, \delta, k)$

Input: $\epsilon, \delta > 0, k > 2$, a stream $\mathcal{S}$ of elements $a_1, \ldots, a_n$ with $a_i \in [m]$ for all i.

Output: An (ϵ, δ) approximation to F_k.

1. Set ϵ', τ, δ', and b as defined above.
2. For $r = 1$ to $O(\log 1/\delta)$,
   (a) Choose the first $O(\log \tau)$ bits of a random $x \in [1/2, 1]$, and for $0 \leq i < \log_{1+\epsilon'} n/x$, define the ranges $S_i = \{j \in [m] \mid x(1+\epsilon')^i \leq f_j < x(1+\epsilon')^{i+1}\}$.
   (b) For $1 \leq t \leq t' = O(\epsilon^{-2} \log\log_{1+\epsilon} n)$, and for $0 \leq j \leq \lceil \log_8 m \rceil$,
      i. If $j = 0$ let $h_{0,t}$ be the zero function, otherwise let $h_{j,t}$ be a random 2-wise independent function from $[m]$ to $[8]$.
      ii. Let $\mathcal{S}_{j,t}$ be the restriction of $\mathcal{S}$ to $i \in \cap_{p=0}^{j} h_{p,t}^{-1}(0)$, and compute $a^{j,t} = \text{Summary}(\mathcal{S}_{j,t}, b, \epsilon', x, \delta', \tau)$.
   (c) Initialize U to $\emptyset$.
   (d) For $0 \leq i < \log_{1+\epsilon'} n/x$ and for $1 \leq t \leq t'' = O(\log\log_{1+\epsilon} n)$, if there is a j for which $a_i^{j,t} = 1$ for more than $t''/3$ values of t,
      i. Set $Z_i = \max\{j \mid a_i^{j,t}$ occurs more than $t''/3$ times$\}$.
      ii. Set $U \leftarrow U \cup \{i\}$.
   (e) For all $i \in U$, compute $T_i = \sum_{p=t''+1}^{t'} a_i^{Z_i+2,p}/(t' - t'')$.
   (f) Compute $\tilde{F}_{k,r} = x^k \sum_{i \in U} 8^{Z_i+2} T_i (1+\epsilon')^{k(i+1)}$.
3. Output the median $\tilde{F}_k$ of the values $\tilde{F}_{k,r}$ taken over all r.

3 Analysis

We divide the analysis into three sections. In the first section we show it is likely that for all substreams given to Summary, Summary only reports the existence of elements with certain frequencies in the substream if they in fact occur in the substream. In the second section we show it is likely that for all substreams given to Summary, Summary reports the existence of contributing frequencies when they do occur in a substream. Finally, in the last section we show that if the events of the first two sections occur, then $\tilde{F}_k$ is a good approximation to F_k.

3.1 Summary only reports frequencies if they occur

We first show that for all substreams, Summary only reports items of certain frequencies if such items actually occur in the substream. This is formalized by the following theorem:

Theorem 2 *With probability at least* $10/11$*, for all* $0 \le j \le \lceil \log_8 m \rceil$*, all* $1 \le t \le t'$*, and all* $0 \le i < \log_{1+\epsilon'} n/x$*,* $a_i^{j,t} = 1$ *only if there is some* $w \in S_i \cap \cap_{p=0}^{j} h_{p,t}^{-1}(0)$*.*

To prove this we need to ensure that the estimated frequencies found by Summary are not too close to $x(1+\epsilon')^i$ for any $0 \le i < \log_{1+\epsilon'} n/x$, as otherwise Summary may misclassify items. We need the following standard lemma and corollary for this. They will enable us to reason about how close the estimated frequencies can be to their borders when the borders are randomized. For readability, we defer the proofs of these two facts to the appendix.

Lemma 3 *Let* x *denote the number obtained by selecting the first* β *bits of a real number chosen uniformly at random from* $[1/2, 1]$*. Let* n *be a positive integer, let* y *be a real in* $[1, n]$*, and let* $r > 0$ *and* $0 < \gamma \le 1$ *be any reals. We allow* y, r*, and* γ *to be functions of* n*. Define the event* $\mathcal{E}$ *by:*

$$\min_{i \in \{0,\ldots,\lceil \log_{1+\gamma} 2n \rceil\}} |y - x(1+\gamma)^i| < r.$$

Then $\Pr_x[\mathcal{E}] = O\left(\frac{r}{\gamma y} + \frac{\log n}{\gamma 2^\beta}\right)$*.*

Corollary 4 *Let* $x, \beta, n, y, r, \gamma$*, and* $\mathcal{E}$ *be as before. Then for any* $i \in \{0, \ldots, \lceil \log_{1+\gamma} 2n \rceil - 1\}$*,*

$$\Pr_x[\mathcal{E} \mid y \in [x(1+\gamma)^i, x(1+\gamma)^{i+1})] = O\left(\frac{r}{\gamma y} + \frac{\log n}{\gamma 2^\beta}\right).$$

We can directly apply this corollary to the output of Summary via the following lemma:

Lemma 5 *For all* $0 \le j \le \lceil \log_8 m \rceil$*,* $1 \le t \le t'$*, and* $0 \le w < \log_{1+\epsilon'} n$*, let* $\alpha(w, j, t)$ *be the first element* [3] *of* $[m]$ *for which* $\lfloor \log_{1+\epsilon'} \tilde{f}_{\alpha(w,j,t)}/x \rfloor = w$*. Then:*

$$\Pr_x\left[\exists j, t, w \text{ s.t. } \min(\tilde{f}_{\alpha(w,j,t)} - x(1+\epsilon')^w, x(1+\epsilon')^{w+1} - \tilde{f}_{\alpha(w,j,t)}) \le \frac{\tilde{f}_{\alpha(w,j,t)}}{\tau}\right] \le 1/22.$$

Proof Consider the behavior of Summary$(\mathcal{S}_{j,t}, b, \epsilon', x, \delta')$ for some fixed j and t. In step 1, Summary will invoke A, and the output of A, CountSketch, will be independent of x. Let $\mathcal{E}_w$ be the event that $\min(\tilde{f}_{\alpha(w,j,t)} - x(1+\epsilon')^w, x(1+\epsilon')^{w+1} - \tilde{f}_{\alpha(w,j,t)}) \le \frac{\tilde{f}_{\alpha(w,j,t)}}{\tau}$.

Since $\tilde{f}_{\alpha(w,j,t)}$ is guaranteed to be an integer (see theorem 1), and since $w \ge 0$, the condition of the lemma ensures $\tilde{f}_{\alpha(w,j,t)} \ge 1$, so we may apply Corollary 4. By setting $\beta = O(\log \tau)$, $n = |\mathcal{S}_{j,t}|$, $y = \tilde{f}_{\alpha(w,j,t)}$, $r = y/\tau$, and $\gamma = \epsilon'$ in Corollary 4, since $\tau = O\left(\frac{1}{\epsilon^3} \log_{1+\epsilon} n \log m \log \log_{1+\epsilon} n\right)$ a union bound over frequency classes gives

$$\Pr_x\left[\exists w \in [\lceil \log_{1+\epsilon'} n/x \rceil \text{ s.t. } \mathcal{E}_w\right] < \frac{\epsilon^2}{O(\log m \log \log_{1+\epsilon} n)},$$

and the lemma follows by a union bound over the invocations of Summary. ■

We can now prove the main theorem of this section. The intuition is that if an item is misclassified, its estimated frequency must differ substantially from its true frequency since the previous

[3] Note that $\tilde{f}_{\alpha(w,j,t)}$ refers to the frequency algorithm A reports for α when processing substream $\mathcal{S}_{j,t}$.

lemma guarantees the estimated frequency is far from its borders, and thus far from frequencies in neighboring ranges (including its true frequency). This contradicts the guarantee on the additive error of A.

Proof of **Theorem** 2: With probability at least 21/22, both $A(\mathcal{S}_{j,t}, b, \delta')$ and step 1b of Summary succeed for all j and t. Indeed, since A and step 1b of Summary are each invoked $O(\epsilon\tau/\log_{1+\epsilon} n)$ times, each time with error probability set to $\delta'/2 = O\left(\frac{\log_{1+\epsilon} n}{\epsilon\tau}\right)$, the claim follows by a union bound. Hence with probability at least 10/11, Summary always succeeds and, using the notation of Lemma 5, $\mathcal{E}_w$ does not occur for any w during any invocation of Summary.

Now suppose $a_i^{j,t} = 1$ for some fixed i, j, and t. Then, using the notation of Summary, $i \geq \lceil \log_{1+\epsilon'} \tau\sqrt{\frac{2\tilde{F}_2}{b}}/x \rceil$, and there is some $v \in [m]$ for which $x(1+\epsilon')^i \leq \tilde{f}_v < x(1+\epsilon')^{i+1}$. In particular, $\tilde{f}_v \geq \tau\sqrt{\frac{2\tilde{F}_2(\mathcal{S}_{j,t})}{b}} \geq \tau\sqrt{\frac{F_2(\mathcal{S}_{j,t})}{b}}$, so $\frac{\tilde{f}_v}{\tau} \geq \sqrt{\frac{F_2(\mathcal{S}_{j,t})}{b}}$.

But if there is no $w \in S_i \cap \cap_{p=0}^{j} h_{p,t}^{-1}(0)$, then $f_v < x(1+\epsilon')^i$ or $f_v \geq x(1+\epsilon')^{i+1}$. Since we are assuming A is correct, we have $|f_v - \tilde{f}_v| \leq \sqrt{\frac{F_2(\mathcal{S}_{j,t})}{b}} \leq \frac{\tilde{f}_v}{\tau}$, contradicting our assumption that $\min(\tilde{f}_v - x(1+\epsilon')^i, x(1+\epsilon')^{i+1} - \tilde{f}_v) > \frac{\tilde{f}_v}{\tau}$. ■

3.2 Summary reports important frequencies when they do occur

The previous section just shows that $a_i^{j,t}$ is correct when it equals 1, without guaranteeing anything when $a_i^{j,t} = 0$. Fortunately, there is a partial converse in the case of certain, important S_i. We say the set S_i *contributes* if

$$|S_i| x^k (1+\epsilon')^{k(i+1)} \geq \frac{\epsilon F_k}{6144(1+\epsilon')^k \lceil \log_{1+\epsilon'} n \rceil}.$$

This exact expression for a contributing S_i is chosen by working backwards from the conclusion of the analysis, and will soon become clear. Observe that if $\mathcal{P}$ is the restriction of $\mathcal{S}$ to elements j which occur in some S_i that contributes, then

$$\left(1 - \frac{\epsilon}{6144(1+\epsilon')^k}\right) F_k \leq F_k(\mathcal{P}) \leq F_k.$$

We show that for certain values of j, $a_i^{j,t}$ is likely to be 1 if S_i contributes. For all i, let i^* be such that $1 \leq |S_i|/8^{i^*} < 8$. We formalize what we mean by Summary reporting the existence of important frequencies via the following theorem:

Theorem 6 *With probability at least* 9/11, *for all* $0 \leq j \leq \lceil \log_8 m \rceil$, *for all* $1 \leq t \leq t'$, *and for all* $0 \leq i < \log_{1+\epsilon'} n/x$, *if* S_i *contributes, if* $j \in \{i^*, i^*+1\}$, *and if there is some* $\alpha \in S_i \cap \cap_{p=0}^{j} h_{p,t}^{-1}(0)$, *then* $a_i^{j,t} = 1$.

To prove this theorem we first develop a property of contributing frequency ranges. The following lemma should give a glimpse of how the factor of $m^{1-2/k}$ appears in our space bound. Its proof is based on Hölder's inequality, and can be found in the appendix.

Lemma 7 *If S_i contributes, then* $|S_i| \geq (1+\epsilon')^{-2}\left(\frac{\epsilon}{6144}\right)^{2/k} \frac{F_2}{m^{1-2/k}x^2(1+\epsilon')^{2(i+1)}\lceil \log_{1+\epsilon'} n\rceil^{2/k}}$.

This lemma allows us to prove the main theorem. The intuition behind the theorem is that, since b is large and since items in contributing frequency ranges influence F_2 (at least more so than items not in contributing ranges), the additive error of algorithm A will translate into a small relative error to the frequencies of these items.

Two technicalities we have to account for are making sure (1) $F_2(\mathcal{S}_{j,t})$ on any substream is not abnormally large, and (2) the estimated frequencies found by Summary are far from their borders. The second technicality is what was used to prove theorem 2, and here we will derive a similar contradiction. Namely, it cannot simultaneously hold that the additive error of A is small, estimated frequencies are far from their borders, and an item is misclassified.

Proof of **Theorem** 6: As in the proof of Theorem 2, we add 1/11 to the error probability and assume both $A(\mathcal{S}_{j,t}, b, \delta')$ and step 1b of Summary succeed for all j, t, and using the notation of Lemma 5, $\mathcal{E}_w$ does not occur for any w on any invocation.

Before proceeding, we bound $F_2(\mathcal{S}_{j,t})$. Observe that $\mathbf{E}[F_2(\mathcal{S}_{j,t})] = F_2/8^j$, so by Markov's inequality,

$$\Pr[F_2(\mathcal{S}_{j,t}) > O(\log m \log\log_{1+\epsilon} n)F_2/8^j] < 1/O(\log m \log\log_{1+\epsilon} n).$$

Therefore by a union bound,

$$\Pr[\exists j,t \text{ s.t. } F_2(\mathcal{S}_{j,t}) > O(\log m \log\log_{1+\epsilon} n)F_2/8^j] < 1/22.$$

We assume $F_2(\mathcal{S}_{j,t}) = O(\log m \log\log_{1+\epsilon} n)F_2/8^j$ for all j, t and add 1/22 to the error probability. Fix j, t, i meeting the hypothesis of the lemma, and let α be the first element of $S_i \cap \cap_{p=0}^{j} h_{p,t}^{-1}(0)$.

There are two **conditions** needed for $a_i^{j,t} = 1$:

1. $\tilde{f}_\alpha \geq \tau\left(\frac{2\tilde{F}_2}{b}\right)^{1/2}(1+\epsilon')$ and
2. $x(1+\epsilon')^i \leq \tilde{f}_\alpha < x(1+\epsilon')^{i+1}$.

We use the following lemma to establish these conditions:

Lemma 8 $|f_\alpha - \tilde{f}_\alpha| \leq \frac{f_\alpha}{2\tau}$

Proof By our assumption on $F_2(\mathcal{S}_{j,t})$,

$$\tilde{f}_\alpha \geq f_\alpha - \left(\frac{F_2(\mathcal{S}_{j,t})}{b}\right)^{1/2} \geq f_\alpha - \left(\frac{O(\log m \log\log_{1+\epsilon} n)F_2}{b8^j}\right)^{1/2},$$

and since $b = O\left(\tau^2 m^{1-2/k} \log m \log\log_{1+\epsilon} n \left(\frac{\log_{1+\epsilon} n}{\epsilon}\right)^{2/k}\right)$,

$$\tilde{f}_\alpha \geq f_\alpha - \left(\frac{\epsilon}{\log_{1+\epsilon} n}\right)^{1/k} \frac{1}{\tau} \left(\frac{F_2}{O(m^{1-2/k})8^j}\right)^{1/2} = f_\alpha - \left(\frac{\epsilon}{\log_{1+\epsilon} n}\right)^{1/k} \frac{1}{\tau} \left(\frac{F_2}{O(m^{1-2/k}|S_i|)}\right)^{1/2},$$

since $j \in \{i^*, i^*+1\}$. Thus by Lemma 7,

$$\tilde{f}_\alpha \geq f_\alpha - \frac{x(1+\epsilon')^i}{2\tau} \geq f_\alpha \left(1 - \frac{1}{2\tau}\right)$$

for a large enough constant in the big-Oh notation for b. Since we also have $\tilde{f}_\alpha \leq f_\alpha + \left(\frac{F_2(\mathcal{S}_{j,t})}{b}\right)^{1/2}$, reversing these inequalities shows $\tilde{f}_\alpha \leq f_\alpha \left(1 + \frac{1}{2\tau}\right)$. ■

Condition 1: For a large enough constant in the big-Oh notation for b, applying lemmas 7 and 8,

$$\begin{aligned}
\tau \left(\frac{2\tilde{F}_2(\mathcal{S}_{j,t})}{b}\right)^{1/2} (1+\epsilon') &\leq \left(\frac{\epsilon}{2\lceil \log_{1+\epsilon'} n \rceil}\right)^{1/k} \left(\frac{F_2\ O(\log m \log\log_{1+\epsilon} n)}{8^j\ O(m^{1-2/k} \log m \log\log_{1+\epsilon} n)}\right)^{1/2} \\
&= \left(\frac{\epsilon}{2\lceil \log_{1+\epsilon'} n \rceil}\right)^{1/k} \left(\frac{F_2}{O(m^{1-2/k}|S_i|)}\right)^{1/2} \\
&\leq f_\alpha/2 \leq f_\alpha \left(1 - \frac{1}{2\tau}\right) \leq \tilde{f}_\alpha
\end{aligned}$$

Condition 2: For all $0 \leq j \leq \lceil \log_8 m \rceil$, $1 \leq t \leq t'$, and all w for which S_w contributes, let $\beta(w,j,t)$ be the first element of $S_w \cap \cap_{p=0}^{j} h_{p,t}^{-1}(0)$ if such an element exists. Observe that by lemma 8, $|f_{\beta(w,j,t)} - \tilde{f}_{\beta(w,j,t)}| \leq f_{\beta(w,j,t)}/(2\tau)$, which means that $\tilde{f}_{\beta(w,j,t)} \neq 0$, as otherwise $f_{\beta(w,j,t)} = 0$ which is impossible. Thus, since $\tilde{f}_{\beta(w,j,t)}$ is an integer, we have $\tilde{f}_{\beta(w,j,t)} \geq 1$, and we can apply Lemma 3. Hence, with probability at least 21/22, for all $0 \leq j \leq \lceil \log_8 m \rceil$, $1 \leq t \leq t'$, and w for which S_w contributes,

$$\min_{i \in \{0,\ldots,\lceil \log_{1+\gamma} n/c \rceil\}} |\tilde{f}_{\beta(w,j,t)} - x(1+\epsilon')^i| < \frac{\tilde{f}_{\beta(w,j,t)}}{\tau}.$$

Adding 1/22 to the error probability, and recalling the definition of α, we conclude there is a Δ for which

$$\min(\tilde{f}_\alpha - x(1+\epsilon')^\Delta, x(1+\epsilon')^{\Delta+1} - \tilde{f}_\alpha) > \frac{\tilde{f}_\alpha}{\tau_1}.$$

Then if $i \neq \Delta$, $|f_\alpha - \tilde{f}_\alpha| > \frac{\tilde{f}_\alpha}{\tau_1} \geq \frac{f_\alpha}{2\tau_1}$ since $x(1+\epsilon')^i \leq f_\alpha < x(1+\epsilon')^{i+1}$ and $\tilde{f}_\alpha \geq f_\alpha/2$ by Lemma 8. But this contradicts Lemma 8. So the two conditions hold, and so does the theorem. ■

3.3 $\tilde{F}_k$ approximates F_k well

Equipped with theorems 2 and 6, we now show our estimate is indeed a good estimate to F_k. We first need a simple probabilistic lemma, enabling us to reason about step 3d of our algorithm. Its proof is given in the appendix.

For $0 \leq i < \log_{1+\epsilon'} n/x$ and $1 \leq t \leq t'$, let $A_{i,t} = \max\{j \mid S_i \cap \cap_{p=0}^{j} h_{p,t}^{-1}(0) \neq \emptyset\}$. Here $A_{i,t}$ denotes the smallest substream containing an element of S_i for a given t. We will characterize the variable Z_i in step 3d of our algorithm in terms of the $A_{i,t}$.

Lemma 9 *With probability at least* $10/11$*, for all* i*, there are more than* $2t'/3$ *values* t *for which* $A_{i,t} \in \{i^*, i^*+1\}$.

We can now relate the $A_{i,t}$ to the algorithm variables Z_i via a straightforward lemma.

Lemma 10 *With probability at least* $7/11$*, for all* i*, if* $i \in U$ *then* $Z_i \leq i^*+1$*, and whenever* S_i *contributes,* $i \in U$ *and* $Z_i \in \{i^*, i^*+1\}$.

Proof With probability at least $7/11$, the events in the previous lemma and theorems 2 and 6 occur. By definition of Z_i, $a_i^{Z_i,t}$ must occur more than $t''/3$ times for $i \in U$. But by the previous lemma there are more than $2t''/3$ values t for which $A_{i,t} \in \{i^*, i^*+1\}$. By definition of $A_{i,t}$ and Theorem 2, this means there can be no $j > i^*+1$ for which $a_i^{j,t} = 1$ for more than $t''/3$ values t. Hence $Z_i \leq i^*+1$. If S_i contributes for some i, then by Theorem 6, $a_i^{j,t} = A_{i,t}$ whenever $A_{i,t} \in \{i^*, i^*+1\}$, and thus $i \in U$ and $Z_i \in \{i^*, i^*+1\}$. ■

Using Lemma 10, we can get excellent approximations to $|S_i|$ for contributing S_i. The idea is that, now that we have a constant factor approximation to $|S_i|$, we can get a tight concentration result by Chernoff bounds and independent trials. Let $U = U_1 \cup U_2$, where U_1 is exactly the set of i for which S_i contributes. We have:

Lemma 11 *If the event in Lemma 10 occurs, then with probability at least* $10/11$*, for all* $i \in U_1$*,*

$$(1-\epsilon/2)|S_i| \leq 8^{Z_i+2} T_i \leq (1+\epsilon/2)|S_i|.$$

Proof If the event in Lemma 10 occurs, for all i and all $p \in \{t''+1, \ldots, t'\}$,

$$1/512 \leq \mathbf{E}[\text{Summary}(\mathcal{S}_{i,Z_i+2,p}, b, \epsilon', x, \delta', \tau)_i] = |S_i|/8^{Z_i+2} \leq 1/8,$$

since $Z_i \in \{i^*, i^*+1\}$. Since $\mathbf{E}[\text{Summary}(\mathcal{S}_{i,Z_i+2,p}, b, \epsilon', x, \delta', \tau)_i] = \Theta(1)$, it follows by Chernoff bounds that for T_i the average of $t' - t'' = O\left(\frac{\log\log_{1+\epsilon} n}{\epsilon^2}\right)$ independent trials of Summary,

$$\Pr[(1-\epsilon/2)|S_i| \leq 8^{Z_i+2} T_i \leq (1+\epsilon/2)|S_i|] \geq 1 - \frac{1}{11\lceil \log_{1+\epsilon'} n/x \rceil},$$

and a union bound gives the lemma. ■

The penultimate step is to show that the items with frequencies which don't contribute, but are

still included in the set U constructed by our algorithm (i.e., those in U_2), really don't contribute. That is, their contribution to F_k is small.

Fix some $r \leq O(\log 1/\delta)$, let

$$\tilde{F}_{k,r}(U_1) = x^k \sum_{i \in U_1} 8^{Z_i+2} T_i (1+\epsilon')^{k(i+1)},$$

and

$$\tilde{F}_{k,r}(U_2) = x^k \sum_{i \in U_2} 8^{Z_i+2} T_i (1+\epsilon')^{k(i+1)}.$$

Recall that $\mathcal{P}$ is the restriction of the input stream $\mathcal{S}$ to elements occurring in a contributing S_i (for a fixed r), and that $F_k(\mathcal{P}) \geq F_k \left(1 - \frac{\epsilon}{6144(1+\epsilon')^k}\right)$. Let $\mathcal{Q}$ denote the stream of elements in $\mathcal{S} \backslash \mathcal{P}$.

From here on out the analysis is just a matter of inequalities.

Lemma 12 *If the event of Lemma 10 occurs, then $\tilde{F}_{k,r}(U_2) \leq \epsilon F_k/12$.*

Proof Using the observation that $T_i \leq 1$, the definition of i^* and Lemma 10, and the fact that $F_k(\mathcal{Q}) \leq \frac{\epsilon}{6144(1+\epsilon')^k} F_k$,

$$\begin{aligned}
\tilde{F}_{k,r}(U_2) &= x^k \sum_{i \in U_2} 8^{Z_i+2} T_i (1+\epsilon')^{k(i+1)} \leq x^k \sum_{i \in U_2} 8^{Z_i+2} (1+\epsilon')^{k(i+1)} \\
&\leq x^k \sum_{i \in U_2} 8^{i^*+3} (1+\epsilon')^{k(i+1)} \leq 512 x^k \sum_{i \in U_2} |S_i| (1+\epsilon')^{k(i+1)} \\
&= 512(1+\epsilon')^k \sum_{i \in U_2} |S_i| x^k (1+\epsilon')^{ki} \leq 512(1+\epsilon')^k F_k(\mathcal{Q}) \leq \epsilon F_k/12
\end{aligned}$$

■

And now we're done.

Theorem 13 *If ϵ is less than a small constant (depending on k), then*

$$\Pr[(1-\epsilon)F_k \leq \tilde{F}_k \leq (1+\epsilon)F_l] \geq 1 - \delta.$$

Proof With probability at least 6/11, the events of lemmas 10, 11, and 12 occur. Let's look at $\tilde{F}_{k,r}$ for a fixed $r \leq O(\log 1/\delta)$.

$$\begin{aligned}
\tilde{F}_{k,r} &\geq \tilde{F}_{k,r}(U_1) = x^k \sum_{i \in U_1} 8^{Z_i+2} T_i (1+\epsilon')^{k(i+1)} \geq (1-\epsilon/2) x^k \sum_{i \in U_1} |S_i| (1+\epsilon')^{k(i+1)} \\
&\geq (1-\epsilon/2) F_k(\mathcal{P}) \geq (1-\epsilon/2) \left(1 - \frac{\epsilon}{6144(1+\epsilon')^k}\right) F_k > (1-\epsilon) F_k
\end{aligned}$$

On the other hand, applying Lemma 12,

$$\begin{aligned}
\tilde{F}_{k,r} &= \tilde{F}_{k,r}(U_1) + \tilde{F}_{k,r}(U_2) \leq x^k \sum_{i \in U_1} 8^{Z_i+2} T_i (1+\epsilon')^{k(i+1)} + \frac{\epsilon F_k}{12} \\
&\leq (1+\epsilon/2) x^k \sum_{i \in U_1} |S_i| (1+\epsilon/(3k))^{k(i+1)} + \frac{\epsilon F_k}{12} \leq (1+\epsilon/2)(1+\epsilon/(3k))^k F_k + \frac{\epsilon F_k}{12} \leq (1+\epsilon) F_k,
\end{aligned}$$

for small enough ϵ. The theorem then follows by a Chernoff bound since $\tilde{F}_k$ is the median of $O(\log 1/\delta)$ independent $\tilde{F}_{k,r}$. ■

3.4 Algorithm Complexity

Space Complexity: For simplicity, we omit factors of $\log\log n$, $\log\log m$, and $\log 1/\epsilon$ in our analysis. The space complexity of Summary is dominated by the call to A, which takes space $O\left(b \log m \log n \log 1/\delta'\right)$, which, ignoring the abovementioned factors, is $O(b \log m \log n)$. The space complexity of our F_k-approximator is dominated by the $O(\epsilon^{-2} \log m \log \log_{1+\epsilon} n \log \frac{1}{\delta})$ calls to Summary in step 3b, which we take to be $O(\epsilon^{-2} \log m \log 1/\delta)$. So $O(\epsilon^{-2} b \log^2 m \log n \log 1/\delta)$ is our total space. Plugging in our choice of b gives total space:

$$O\left(\frac{\tau^2 m^{1-2/k} \log^3 m \log^{1+2/k} n \log 1/\delta}{\epsilon^{2+4/k}}\right) = O\left(\frac{m^{1-2/k} \log^5 m \log^{3+2/k} n \log 1/\delta}{\epsilon^{10+4/k}}\right) = \tilde{O}(m^{1-2/k}).$$

Time Complexity: Although our main focus is on matching the $\Omega(m^{1-2/k})$ space lower bound, our algorithm also has essentially optimal update time: $\tilde{O}(1)$ time per stream element. Indeed, as can be seen from its description in [4], algorithm A has $\tilde{O}(1)$ update time. For each new stream element, we simply feed the element into each of $O(\epsilon^{-2} \log m \log \log_{1+\epsilon} n \log 1/\delta)$ calls to Summary in step 3b, each in turn feeding the element into a local copy of algorithm A and an F_2-approximator, both requiring only $\tilde{O}(1)$ update time. Therefore our update time is $\tilde{O}(1)$. In fact, the hidden terms in the $\tilde{O}(1)$ are small. Indeed, ignoring the aforementioned factors, our update time is $O(\epsilon^{-2} \log m)$, since both our algorithm A of [4] and the algorithm for approximating F_2 given in [11] can be updated with $O(1)$ arithmetic operations (on words of size $O(\log n)$).

4 Generalization to L_p differences

The idea is that algorithm A of Theorem 1 can be applied to streams $\mathcal{S}$ with both insertions and deletions with the same approximation guarantee. More precisely, the jth stream element a_j is now a pair $(i, s) \in [m] \times \{-1, +1\}$, and the (possibly negative) frequency of item i is given by $f_i = \sum_{(i,s)\in\mathcal{S}} s$. To approximate $\sum_{i=1}^{m} |a_i - b_i|^p$ for the L_p difference problem, upon encountering a_i, we feed a_i copies of $(i, +1)$ into A, whereas for b_i we feed b_i copies of $(i, -1)$ into A. In step 3 of Summary, we now compute $z = \lfloor \log_{1+\gamma} |\tilde{f}_j|/x \rfloor$. Note that step 1b of Summary also extends to streams with deletions (running, say, the F_2 algorithm of [1]), so it is as if our algorithm is running on a (possibly shorter) data stream with frequencies $f_i = |a_i - b_i|$.

5 Conclusions

In this paper, we give an essentially tight upper bound for the space complexity of estimating F_k in a data stream.

The algorithm, as given, is not practical, due to several logarithmic and poly$(1/\epsilon)$ factors in its space bound, which we did not attempt to optimize. We do believe that further simplifications should lead to a more natural and practical algorithm. In particular, a few factors could be

eliminated if we removed the need to estimate the frequencies of retrieved elements up to sub-constant accuracy (it seems that constant-factor estimation should suffice).

Ultimately, it would be of great interest to design an "estimator-based" algorithm matching the $\tilde{O}(m^{1-2/k})$ space bound (as in [1, 5]). Such algorithms would likely have smaller big-Oh constants in their space complexity.

References

[1] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. In *Proceedings of the 28th Annual ACM Symposum on the Theory of Computing (STOC)*, pages 20-29, 1996. Also; J. Comp. Sys. Sci. 58, pages 137–147, 1999.

[2] Z. Bar Yossef, T.S. Jayram, R. Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. In *Proceedings of the 43rd Symposium on Foundations of Computer Science (FOCS)*, pages 209–218, 2002.

[3] A. Chakrabarti, S. Khot, and X. Sun. Near-optimal lower bounds on the multiparty communication complexity of set-disjointness. In *Proceedings of the 18th IEEE Conference on Computational Complexity (CCC)*, pages 107–117, 2003.

[4] M. Charikar, K. Chen, and M. Farach-Colton. Finding Frequent Items in Data Streams. In *Theor. Comput. Sci.* **312(1)**, pages 3-15, 2004 (Extended Abstract appeared in *Proceedings of the 29th International Colloquium on Automata Languages and Programming (ICALP)*, 2002).

[5] D. Coppersmith and R. Kumar, An improved data stream algorithm for frequency moments. In *Proc of the 15th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 151-156, 2004.

[6] J. Feigenbaum, S. Kannan, M. Strauss, and M. Viswanathan. An Approximate L1-Difference Algorithm for Massive Data Streams. In *SIAM Journal on Computing* **32**, pages 131-151, 2002 (Extended Abstract appeared in *Proceedings of IEEE Symposium on Foundations of Computer Science (FOCS), 1999*).

[7] P. Indyk. Stable Distributions, Pseudorandom Generators, Embeddings and Data Stream Computation. In *Proceedings of the 41st Symposium on Foundations of Computer Science (FOCS)*, pages 187-197, 2000.

[8] W. Johnson and J. Lindenstrauss. Extensions of Lipschitz maps into a Hilbert space. In *Contemp. Math* **26**, pages 189–206, 1984.

[9] S. Muthukrishnan. Data Streams: algorithms and applications. In *Proc. of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, page 413, 2003.

[10] M. Saks and X. Sun. Space lower bounds for distance approximation in the data stream model. In *Proc of the 34th Annual ACM Symposium on Theory of Computing (STOC)*, pages 360-369, 2002.

[11] M. Thorup and Y. Zhang. Tabulation based 4-universal hashing with applications to second moment estimation. In *Proc of the 15th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 615-624, 2004.

6 Appendix

Here we give proofs of lemmas which were omitted from section 3. This first lemma and corollary are used to show that our randomization of borders of frequency classes is likely to place frequencies away from their borders:

Lemma 14 *Let x denote the number obtained by selecting the first β bits of a real number chosen uniformly at random from $[1/2, 1]$. Let n be a positive integer, let y be a real in $[1, n]$, and let $r > 0$ and $0 < \gamma \leq 1$ be any reals. We allow y, r, and γ to be functions of n. Define the event $\mathcal{E}$ by:*

$$\min_{i \in \{0,\ldots,\lceil \log_{1+\gamma} 2n \rceil\}} |y - x(1+\gamma)^i| < r.$$

Then $\Pr_x[\mathcal{E}] = O\left(\frac{r}{\gamma y} + \frac{\log n}{\gamma 2^\beta}\right)$.

Proof We first show that $\Pr_x[\mathcal{E}] = O(r/(\gamma y))$ when x is a real number, that is, when we don't restrict x to its first β bits. We have:

$$\Pr_x[\mathcal{E}] = \sum_{i=0}^{\lfloor \log_{1+\gamma} 2n \rfloor} \Pr[\mathcal{E} \mid y \in [x(1+\gamma)^i, x(1+\gamma)^{i+1})] \Pr[y \in [x(1+\gamma)^i, x(1+\gamma)^{i+1})]$$

Observe that

$$\Pr[y \in [x(1+\gamma)^i, x(1+\gamma)^{i+1})] = \Pr\left[\frac{y}{(1+\gamma)^{i+1}} < x \leq \frac{y}{(1+\gamma)^i}\right],$$

and since $x \in [1/2, 1]$ this quantity is non-zero only if

$$\lceil \log_{1+\gamma} y \rceil - 1 \leq i \leq \lfloor \log_{1+\gamma} 2y \rfloor.$$

We split the analysis into three cases:

Case 1: $\lceil \log_{1+\gamma} y \rceil - 1 < i < \lfloor \log_{1+\gamma} 2y \rfloor$. We have:

$$\Pr\left[\frac{y}{(1+\gamma)^{i+1}} < x \leq \frac{y}{(1+\gamma)^i}\right] = \frac{2\gamma y}{(1+\gamma)^{i+1}},$$

and

$$\Pr[\mathcal{E} \mid y \in [x(1+\gamma)^i, x(1+\gamma)^{i+1})] \leq \frac{r(2+\gamma)}{\gamma y}.$$

Case 2: $i = \lceil \log_{1+\gamma} y \rceil - 1$. We have:

$$\Pr\left[\frac{y}{(1+\gamma)^{i+1}} < x \leq \frac{y}{(1+\gamma)^i}\right] = \Pr\left[\frac{y}{(1+\gamma)^{\lceil \log_{1+\gamma} y \rceil}} < x \leq 1\right] \leq \frac{2\gamma}{(1+\gamma)},$$

and

$$\Pr[\mathcal{E} \mid y \in [x(1+\gamma)^i, x(1+\gamma)^{i+1})] \leq \frac{1+\gamma}{\gamma}\left(\frac{r}{(1+\gamma)^{i+1}} + 1 + \frac{r-y}{(1+\gamma)^i}\right) \leq \frac{r(2+\gamma)(1+\gamma)}{\gamma y}.$$

Case 3: $i = \lfloor \log_{1+\gamma} 2y \rfloor$. Since $0 < \gamma \leq 1$, we have:

$$\Pr\left[\frac{y}{(1+\gamma)^{i+1}} < x \leq \frac{y}{(1+\gamma)^i}\right] = \Pr\left[1/2 < x \leq \frac{y}{(1+\gamma)^{\lfloor \log_{1+\gamma} 2y \rfloor}}\right] \leq \gamma,$$

and

$$\Pr[\mathcal{E} \mid y \in [x(1+\gamma)^i, x(1+\gamma)^{i+1})] \leq \frac{2}{\gamma}\left(\frac{y+r}{(1+\gamma)^{i+1}} - \frac{1}{2} + \frac{r}{(1+\gamma)^i}\right) \leq \frac{r(2+\gamma)}{\gamma y}.$$

Summing up the three cases,

$$\begin{aligned}
\Pr_x[\mathcal{E}] &= \sum_{i=\lceil \log_{1+\gamma} y \rceil - 1}^{\lfloor \log_{1+\gamma} 2y \rfloor} \Pr[\mathcal{E} \mid y \in [x(1+\gamma)^i, x(1+\gamma)^{i+1})] \Pr\left[\frac{y}{(1+\gamma)^{i+1}} < x \leq \frac{y}{(1+\gamma)^i}\right] \\
&\leq 2r(2+\gamma) \sum_{i=\lceil \log_{1+\gamma} y \rceil}^{\lfloor \log_{1+\gamma} 2y \rfloor - 1} \frac{1}{(1+\gamma)^i} + \frac{2r(2+\gamma)}{y} + \frac{r(2+\gamma)}{y} \\
&\leq \frac{2r(1+\gamma)(2+\gamma)}{y\gamma} + \frac{3r(2+\gamma)}{y} = O\left(\frac{r}{\gamma y}\right).
\end{aligned}$$

Now consider what happens when we restrict x to its first β bits. In our proof we have used the following bound:

$$\Pr[\mathcal{E} \mid y \in [x(1+\gamma)^i, x(1+\gamma)^{i+1})] \leq \Pr\left[\frac{y}{(1+\gamma)^{i+1}} \leq x \leq \frac{y+r}{(1+\gamma)^{i+1}}\right] + \Pr\left[\frac{y-r}{(1+\gamma)^i} \leq x \leq \frac{y}{(1+\gamma)^i}\right].$$

When we restrict x to its first β bits, the right hand side of this expression can increase by at most $4/2^{\beta-1}$. Similarly, $\Pr[y \in [x(1+\gamma)^i, x(1+\gamma)^{i+1})]$ can increase by at most $2/2^{\beta-1}$. Expanding the expression for $\Pr_x[\mathcal{E}]$, we see that our bound on $\Pr_x[\mathcal{E}]$ can increase by at most $O\left(\frac{\log n}{\gamma 2^\beta}\right)$, which proves the lemma. ■

Corollary 15 *Let $x, \beta, n, y, r, \gamma$, and $\mathcal{E}$ be as in the lemma above. Then for any $0 \leq i < \lceil \log_{1+\gamma} 2n \rceil$,*

$$\Pr_x[\mathcal{E} \mid y \in [x(1+\gamma)^i, x(1+\gamma)^{i+1})] = O\left(\frac{r}{\gamma y} + \frac{\log n}{\gamma 2^\beta}\right).$$

Proof This is immediate from the proof of the previous lemma. ■

The following lemma shows a technical property of contributing frequency classes:

Lemma 16 *If S_i contributes, then $|S_i| \geq (1+\epsilon')^{-2} \left(\frac{\epsilon}{6144}\right)^{2/k} \frac{F_2}{m^{1-2/k} x^2 (1+\epsilon')^{2(i+1)} \lceil \log_{1+\epsilon'} n \rceil^{2/k}}$.*

Proof By Hölder's inequality,

$$F_2 = \sum_{i=1}^m f_i^2 \cdot 1 \leq \left(\sum_{i=1}^m f_i^k\right)^{2/k} \left(\sum_{i=1}^m 1\right)^{1-2/k} = F_k^{2/k} m^{1-2/k},$$

and thus,

$$\Big(\frac{\epsilon}{6144}\Big)^{2/k}\frac{(1+\epsilon')^{-2}F_2}{m^{1-2/k}x^2(1+\epsilon')^{2(i+1)}\lceil\log_{1+\epsilon'}n\rceil^{2/k}} \le \Big(\frac{\epsilon}{6144}\Big)^{2/k}\frac{(1+\epsilon')^{-2}F_k^{2/k}}{x^2(1+\epsilon')^{2(i+1)}\lceil\log_{1+\epsilon'}n\rceil^{2/k}} \le |S_i|,$$

since either $\big(\frac{\epsilon}{6144}\big)\frac{(1+\epsilon')^{-k}F_k}{x^k(1+\epsilon')^{k(i+1)}\lceil\log_{1+\epsilon'}n\rceil} \ge 1$, in which case

$$\Big(\frac{\epsilon}{6144}\Big)^{2/k}\frac{(1+\epsilon')^{-2}F_k^{2/k}}{x^2(1+\epsilon')^{2(i+1)}\lceil\log_{1+\epsilon'}n\rceil^{2/k}} \le \frac{(1+\epsilon')^{-k}\epsilon F_k}{6144x^k(1+\epsilon')^{k(i+1)}\lceil\log_{1+\epsilon'}n\rceil} \le |S_i|,$$

or $\big(\frac{\epsilon}{6144}\big)\frac{(1+\epsilon')^{-k}F_k}{x^k(1+\epsilon')^{k(i+1)}\lceil\log_{1+\epsilon'}n\rceil} < 1$, in which case

$$\Big(\frac{\epsilon}{6144}\Big)^{2/k}\frac{(1+\epsilon')^{-2}F_k^{2/k}}{x^2(1+\epsilon')^{2(i+1)}\lceil\log_{1+\epsilon'}n\rceil^{2/k}} < 1 \le |S_i|,$$

since S_i is non-empty. ■

The next lemma gives probabilistic guarantees on the minimal substream containing an element of a frequency range S_i for each i.

Lemma 17 *With probability at least* $10/11$*, for all* i*, there are more than* $2t'/3$ *values* t *for which* $A_{i,t} \in \{i^*, i^*+1\}$.

Proof Fix any i, t. Then $|S_i|/8^{i^*+2} < 1/8$, so

$$\mathbf{E}[|S_i \cap \cap_{p=0}^{i^*+2} h_{p,t}^{-1}(0)|] < 1/8,$$

so by Markov's inequality, $\Pr[A_{i,t} \ge i^*+2] < 1/8$. We also have $\mathbf{E}[|S_i \cap \cap_{p=0}^{i^*-1} h_{p,t}^{-1}(0)|] = |S_i|/8^{i^*-1}$, and by independence of the $h_{p,t}$'s and pairwise-independence of each individual $h_{p,t}$, the indicator variables V_q indicating if the qth element of S_i is in $\cap_{p=0}^{i^*-1} h_{p,t}^{-1}(0)$ are pairwise-independent, so $\mathbf{Var}[\sum_q V_q] = \mathbf{Var}[|S_i \cap \cap_{p=0}^{i^*-1} h_{p,t}^{-1}(0)|] = |S_i|/8^{i^*-1}$. Applying Chebyshev's inequality,

$$\Pr[A_{i,t} \le i^*-1] \le \Pr[|S_i \cap \cap_{p=0}^{i^*-1} h_{p,t}^{-1}(0)| = 0] < 8^{i^*-1}/|S_i| < 1/8.$$

By a union bound, $\Pr[i^* \le A_{i,t} \le i^*+1] > 3/4$. It follows by Chernoff bounds that by independently repeating the above process $t'' = O(\log\log_{1+\epsilon} n)$ times, there are more than $2/3$ values of t for which $A_{i,t} \in \{i^*, i^*+1\}$ with probability at least $1 - \frac{1}{11\lceil\log_{1+\epsilon'} n/x\rceil}$. A union bound gives the lemma. ■