Estimating first frequency moment of data stream in nearly optimal space and time

Sumit Ganguly
IIT Kanpur, India

Purushottam Kar
IIT Kanpur, India

Introduction. A class of modern applications processes data that arrives rapidly and continuously, generically called data streams, for the purpose of integrity-monitoring or early warning of critical developments in a system. Data stream processing algorithms are (a) highly space-time efficient, (b) provide guaranteed errors and (c) are single-pass or online algorithms. An input stream is viewed as a potentially unbounded sequence of records of the form (pos, i, v), where, pos is the sequence number, $i \in \{1, 2, \ldots, n\} = [n]$, v is an integer, such that $|v| \leq M$, which reflects the change to the *frequency* f_i of item i, that is, corresponding to the input record (pos, i, v) changes $f_i \leftarrow f_i + v$. Hence $f_i = \sum_{(\text{pos},i,v)} v$. The vector $f = [f_1, f_2, \ldots, f_n]^T$ is called the frequency vector of the stream. Let m be the number of records appearing in the stream. The pth moment of the frequency vector of a stream is defined as $F_p = \sum_{i\in[n]} |f_i|^p$. The problem of estimating F_p has been fundamental to the development of data stream algorithms [1, 9, 2, 5]. It also has applications, for example, in network monitoring [4], approximate histogram maintenance for database query optimization and computing document similarities [11], etc..

Previous Work. Indyk in [9] pioneered random linear p-stable sketches and showed that a $1 \pm \epsilon$-approximation of F_1 with probability 15/16 is obtained by $\text{median}_{r=1}^{s} |X_r|$, where, X_r are *i.i.d.* 1-stable sketches and $s = O(\epsilon^{-2})$. A p-stable sketch is a linear combination $X = \sum_{i=1}^{n} a_i s_i$ where the s_i's are drawn independently from p-stable distribution $\text{St}(p, 1)$ with scale factor 1 (see [14]). For $p = 1$, the Cauchy distribution (density function $f(x) = 1/(\pi(1 + x^2))$, $x \in \mathbb{R}$) is 1-stable. Indyk uses Nisan's pseudorandom generator for fooling space bounded computations to reduce randomness requirement of full independence of the stable variables. Li in [11] shows that for any $0 < p \leq 2$ and $k \geq 3$, there is a function $C(p, k)$ such that the geometric means estimator $\hat{F}_p^{\text{GM}}(X_1, \ldots, X_k) = (C(p, k))^{-1} |X_1|^{p/k} \cdot |X_2|^{p/k} \cdots |X_k|^{p/k}$ satisfies $\mathsf{E}[\hat{F}_p] = F_p$ and $\mathsf{Var}\big[\hat{F}_p\big] = (K_{p,k} - 1)F_{2p}$, where, $K_{p,k} = 1 + \pi^2(2 + p^2)/(12k) + O(1/k^2)$. The space requirement of both Indyk's and Li's algorithm is $O(\epsilon^{-2} \log^2(mM))$. This is reduced to $O(\epsilon^{-2} \log(mM))$ by Kane et.al. [10] and is shown to be tight [16, 10]. The algorithms of Indyk, Li and Kane et.al. require time $\Omega(\epsilon^{-2} \cdot \text{polylog}(mMn))$ to process each stream update. Since data stream updates arrive very rapidly and ϵ can be small (.01 to 0.001), it is essential to reduce the time required to process each stream update. In this measure, the Hss based algorithm in [7] with update time $O(\log^2(mM))$ qualifies, although it has sub-optimal space usage $O(\epsilon^{-3} \log^2(mM))$.

This work. We present an algorithm for estimating F_1 that is nearly optimal with respect to space $O(\epsilon^{-2} \log^2(mM))$ and has update processing time that is $O(\log^2(mnM) \log(\epsilon^{-1}))$. After submission of this work in arXiv [8], we were made aware of the work in [12] (personal communication from Jelani Nelson) which was not available publicly till then. Their work presents an algorithm that is similar to the one presented in this paper.

Algorithm. The algorithm separates items based on some estimate of its frequency into "heavy" and "light" items, and then to separately estimate the contributions of heavy and light items to F_1 and return the sum. *Notation.* If $|f_{s_1}| \geq |f_{s_2}| \geq \ldots \geq |f_{s_n}|$ is an ordering of the items by non-increasing values of absolute frequencies, then, $F_p^{\text{res}}(k) = \sum_{j=k+1}^{n} |f_{s_j}|^p$. Let $B = 150/\epsilon^2$ and $C = 64B$. The heavy items are identified as follows. Keep a CountSketch structure [3] denoted as HH_2, consisting of $O(\log n)$ hash tables where each hash table has $64C$ buckets. The structure returns an estimate $\hat{f}_i$ satisfying $|\hat{f}_i - f_i| \leq (F_2^{\text{res}}(B)/C)^{1/2}$, for all $i \in [n]$, with joint probability of success 63/64. The algorithm of [6] is applied to HH_2 to obtain $\hat{F}_2^{\text{res}}(B)$ that is accurate to within $1 \pm 1/8$ of $F_2^{\text{res}}(B)$ with probability 63/64. An item is said to be heavy if $\hat{f}_i \geq (4/3)\hat{F}_2^{\text{res}}(B)/B$. The set of heavy items is denoted by H; the light items form the set $L = [n] \setminus H$. A family of functions $\mathcal{H}$ mapping $[n]$ to $[q]$ is S-uniform [15] for a given $S \subset [n]$ if $\mathsf{Pr}_{h\in\mathcal{H}}[\wedge_{j\in S} h(j) = y_j] = 1/q^{|S|}$ for any choice of $y_j \in [q]$, $j \in S$. The family $\mathcal{H}$ is said to be S-uniform with probability $1 - \delta$, if there is a subset of $\mathcal{H}' \subset \mathcal{H}$ of "good" hash functions such that $\mathcal{H}'$ is S-uniform, and $|\mathcal{H}'| \geq (1 - \delta)|\mathcal{H}|$. An implication of Siegel's construction [15] is reproduced below.

Theorem 1 ([15]). *For every $n > 0$, $0 < k < n$, $r \geq 0$, $q = 2k$, $d \geq (r + 1) \log n + \log k + 1$, all integral, there exists a family of functions $\mathcal{H}(n, k, q, r, d)$ mapping $[n]$ to $[q]$ such that, (1) a random choice $h \in_R \mathcal{H}$ is S-uniform for any subset $S \subset [n]$ of size k with probability more than $1 - n^{-rd}$, (2) each $h \in H$ can be represented using $dq(\log q)$ bits, and (3) $h(x)$ can be computed in time $O(d^2)$ operations over $\log q$-bit numbers.*

The light estimator uses the following structure. Keep a hash table U having C buckets numbered 1 to C and a hash function $h : [n] \to [C]$ chosen randomly from $\mathcal{H}(n, k, q, r, d)$, where, $k = C/2, q = C$, $r = 1$ and $d = 2 \log n + \log C + 1$.

Each bucket $U[b]$ maintains three p-stable sketches denoted by $X_{b,1}$, $X_{b,2}$ and $X_{b,3}$ of the sub-stream hashing to b: $X_{b,r} = \sum_{i:h(i)=b} f_i s_{b,r}(i), b \in [C], r \in \{1,2,3\}$. For each value of b and r, the variables $\{s_{b,r}(i)\}_{i\in[n]}$ are obtained by using Nisan's PRG [13] using a seed size of $T = O((\log((mM\epsilon^{-1}))(\log n))$. For each value of b, the seeds for $s_{b,r}(i)$ and $s_{b,r'}(i')$, $r \neq r'$ are three-wise independent. Across buckets in the same table, the seeds for the random variables $s_{b,r}(i)$ and $s_{b',r'}(i')$, for $b \neq b'$ are pair-wise independent. The light estimator is the following. For bucket index $b \in [C]$ say that the event NoCollision(b, H) holds if no heavy item maps to bucket b in table U, that is, NoCollision$(b, H) \equiv \forall k \in H, h(k) \neq b$. For a light item $j \in [n] \setminus H$, define NoCollision(j, H) to be the event that j does not map to any of the buckets to which H maps, that is, $h(j) \neq h(i), \forall i \in H$. Then,

$$\hat{F}_p^L = C_L \sum_{b\in[C]:\text{NoCollision}(b,H)} \hat{F}_p^{\text{GM}}(X_{b,1}, X_{b,2}, X_{b,3}), \quad \text{where, } C_L = 1/\mathsf{Pr}\left[\text{NoCollision}(j,H)\right] = (1-1/C)^{-|H|} \ .$$

The heavy estimator uses the following structure. Keep a collection of $g = O(\log \epsilon^{-1})$ hash tables $T_1, T_2, \ldots, T_g$, each consisting of C buckets and corresponding hash function $h_t : [n] \to [C]$ for $t \in [g]$. Each bucket of a table contains a single AMS sketch of the sub-stream of items mapping to that bucket, that is, $T_t[b] = \sum_{h_t(i)=b} f_i \xi_t(i)$, where, $\xi_t(i) \in_R \{1, -1\}$, the family $\{\xi_t(i)\}_{i\in[n]}$ for each fixed t is pair-wise independent and the seeds generating $\{\xi_t(i)\}_{t\in[g]}$ are pair-wise independent. The h_t's are chosen uniformly at random and independently from $\mathcal{H}(n, k, q, r, d)$ with the same parameter settings as before. The estimate for F_1 is obtained as follows. Say that the event NoHvyColl(H) holds if for each $i \in H$, there is a table index $\theta(i) \in [g]$ such that no other heavy item maps to the same bucket as i in that table, that is,

$$\text{NoHvyColl}(H) \equiv \forall i \in H, \exists \theta(i) \in [g] \text{ s.t. } \forall k \in H \backslash \{i\}, h_{\theta(i)}(i) \neq h_{\theta(i)}(k) \ .$$

The heavy estimate is defined assuming NoHvyColl(H) holds; otherwise it returns 0 (i.e., fails).

$$\hat{F}_1^H = \sum_{i\in H} T_{\theta(i)}[h_{\theta(i)}(i)] \cdot \text{sgn}(\hat{f}_i) \cdot \xi_{\theta(i)}(i) \qquad \text{if NoHvyColl}(H) \text{ holds.}$$

The final estimator is the sum of heavy and light estimators, namely, $\hat{F}_1 = \hat{F}_1^H + \hat{F}_1^L$.

Analysis. Define GoodEst to be the event that (a) frequencies of heavy items are estimated accurately within additive error of $(F_B^{\text{res}}/C)^{1/2}$ using HH_2, and, (b) $F_2^{\text{res}}(B)$ is estimated to within error of $1 \pm 7/8$. By property of the CountSketch structure used for HH_2 [3], the event GoodEst holds with probability $1-1/64$. The rest of the analysis is conditioned on GoodEst. An item i is heavy if $\hat{f}_i > (4/3)(\hat{F}_B^{\text{res}}/B)^{1/2}$. Since GoodEst holds, $|f_i - \hat{f}_i| \leq (F_B^{\text{res}}/C)^{1/2}$ and $\hat{F}_B^{\text{res}} > (7/8)F_2^{\text{res}}(B)$. So $|f_i| > (4/3)[(7/8)F_2^{\text{res}}(B)/B]^{1/2} - (F_2^{\text{res}}(B)/C)^{1/2} > 1.22(F_2^{\text{res}}(B)/B)^{1/2}$. Hence, $|H| \leq B + B/(1.22)^2 \leq (5/3)B$.

We first analyze the light estimator. Let $L = [n] \setminus H$ be the set of the light items. There are two sets of independent random bits: the random seed for the hash function $h \in \mathcal{H}$ and the stable sketch random seed $\bar{s}$. For bucket index b, denote by $\hat{F}_p^{\text{GM}}(b)$ the geometric means estimator applied to the sketches in $U[b]$. For simplicity of notation, let $\mathcal{E}_{b,h} = \mathcal{E}_{b,h,H}$ denote the event NoCollision(b, H) for $b \in [C]$ and similarly let $\mathcal{E}_{j,h}$ denote the event NoCollision(j, H) for a light item j. The two events are related: $\mathcal{E}_{j,h} \equiv \mathcal{E}_{h(j),h}$, for each $j \in L$. Given any event say $\mathcal{E}$, we define the boolean variable $I(\mathcal{E})$ that is 1 if $\mathcal{E}$ holds and $I(\mathcal{E})$ is 0 if $\mathcal{E}$ does not hold.
By property of $\hat{F}_p^{\text{GM}}$, $\mathsf{E}[\hat{F}_p^{\text{GM}}(b)] = \sum_{j:h(i)=b} |f_i|$. The analysis for the light estimator is done for $p \in (0, 2]$.

$$\mathsf{E}[\hat{F}_p^L] = C_L \mathsf{E}_h\Big[\sum_{b:\mathcal{E}_{b,h}} \mathsf{E}_{\bar{s}}\Big[\hat{F}_p^{\text{GM}}(b) \mid h\Big]\Big] = C_L \mathsf{E}_h\Big[\sum_{b:\mathcal{E}_{b,h}} \sum_{h(j)=b} |f_j|^p\Big] = C_L \mathsf{E}_h\Big[\sum_{j\in L} |f_j|^p I(\mathcal{E}_{j,h})\Big] = C_L \sum_{j\in L} |f_j|^p \cdot \mathsf{Pr}\left[\mathcal{E}_{j,h}\right] = \sum_{j\in L} |f_j|^p \ .$$

The last equality follows from the fact that $\mathsf{Pr}\left[\mathcal{E}_{j,h}\right] = 1/C_L$ under S-uniformity, which may fail with probability at most n^{-rd}. The variance calculation is as follows.

$$\mathsf{Var}_{h,\bar{s}}\left[\hat{F}_p^L\right] = C_L^2 \mathsf{E}\big[(\hat{F}_p^L)^2\big] - (F_p^L)^2 = C_L^2 \mathsf{E}_{h,\bar{s}}\Big[\sum_{b:\mathcal{E}_{bh}} (\hat{F}_p^{\text{GM}}(b))^2\Big] + C_L^2 \mathsf{E}_{h,\bar{s}}\Big[\sum_{b\neq b', \mathcal{E}_{bh}\wedge\mathcal{E}_{b'h}} \hat{F}_p^{\text{GM}}(b)\hat{F}_p^{\text{GM}}(b')\Big] - (F_p^L)^2 \quad (1)$$

The first expectation term above is bounded using the property of $\hat{F}_p^{\text{GM}}$ (see Introduction).

$$\mathsf{E}_{h,\bar{s}}\big[(\hat{F}_p^L)^2\big] = \mathsf{E}_h\Big[\mathsf{E}_{\bar{s}}\Big[\sum_{b:\mathcal{E}_{bh}} (\hat{F}_p^{\text{GM}}(b))^2 \mid h\Big]\Big] = \mathsf{E}_h\Big[\sum_{b:\mathcal{E}_{bh}} \sum_{h(j)=b} K_p |f_j|^{2p}\Big] = K_p \mathsf{E}_h\Big[\sum_{j\in L} |f_j|^{2p} \cdot I(\mathcal{E}_{j,h})\Big] = \sum_{j\in L} K_p |f_j|^{2p} \cdot (1/C_L) \ .$$

where the last equality follows from the (notational) fact $\mathsf{E}_h\big[I(\mathcal{E}_{j,h})\big] = \mathsf{Pr}_h\left[\mathcal{E}_{j,h}\right] = 1/C_L$.

Consider the second expectation term of (1). Let $\mathcal{D}(h,b,b') \equiv b \neq b' \wedge \mathcal{E}_{b,h} \wedge \mathcal{E}_{b'h}$. Since the seeds for stable sketches of $\hat{F}_p^{\mathrm{GM}}(b)$ and $\hat{F}_p^{\mathrm{GM}}(b')$ are pair-wise independent, and independent of the hash function seed, we have,

$$C_L^2 \mathsf{E}_{h,\bar{s}}\Big[\sum_{\mathcal{D}(h,b,b')} \hat{F}_p^{\mathrm{GM}}(b)\hat{F}_p^{\mathrm{GM}}(b')\Big] = C_L^2 \mathsf{E}_h\Big[\sum_{\mathcal{D}(h,b,b')} \mathsf{E}_{\bar{s}}\big[\hat{F}_p^{\mathrm{GM}}(b)\hat{F}_p^{\mathrm{GM}}(b') \mid h\big]\Big] = C_L^2 \mathsf{E}_h\Big[\sum_{D(h,b,b')} \sum_{h(j)=b} |f_j|^p \sum_{h(j')=b'} |f_{j'}|^p\Big]$$
$$= C_L^2 \sum_{j,j' \in L, j \neq j'} |f_j|^p |f_{j'}|^p \mathsf{Pr}_h\left[\mathcal{E}_{j,h} \wedge \mathcal{E}_{j'}^h \wedge h(j) \neq h(j')\right] = C_L^2 \sum_{j,j' \in L, j \neq j'} |f_j|^p |f_{j'}|^p (1/C_L)^2 (1 - 1/C)$$

where the last equality follows from the S-uniform property of $\mathcal{H}$. Note that this term is smaller than the cross terms' contribution coming from $(F_p^L)^2$. Substituting in (1), we have $\mathsf{Var}\big[\hat{F}_p^L\big] = (K_p C_L - 1)\sum_{j \in L} |f_j|^{2p}$.
The following lemma is proved using standard techniques.

Lemma 2. *Suppose $|f_{s_1}| \geq |f_{s_2}| \geq \ldots \geq |f_{s_n}|$. Then, for $0 < p \leq q$, $\sum_{j=B+1}^{n} |f_{s_i}|^q \leq (1/B^{q/p-1})\big(\sum_{j=1}^{n} |f_{s_i}|^p\big)^{q/p}$.* □

Since light items satisfy $f_i \leq (1.22)F_2^{\mathrm{res}}(B)/B$, hence, by Lemma 2,

$$\sum_{j \in L} |f_j|^{2p} \leq (1.22)^{p/2}\big(F_2^{\mathrm{res}}(B)/B\big)^{p/2} \sum_{j \in L} |f_j|^p \leq (1.22)^{p/2}(F_p/B)F_p \leq (1.22)^{p/2} F_p^2/B \ .$$

Since, $K_p \leq 3$ for $p = 1$ (see Introduction) and $C_L < 1$, $\mathsf{Var}[\hat{F}_1^L] \leq (K_1 C_L - 1)\sqrt{1.22}\epsilon^2 F_1^2/B \leq F_1^2\epsilon^2/64$, since $B = 150\epsilon^{-2}$. Applying Chebychev inequality, $\big|\hat{F}_1^L - F_1^L\big| \leq (\epsilon/2)F_1$ with probability 1-1/16.

Consider the heavy estimator. Since GoodEst holds, $\mathrm{sgn}(\hat{f}_i) = \mathrm{sgn}(f_i)$ and hence $f_i \cdot \mathrm{sgn}(\hat{f}_i) = |f_i|$. Denote the random seed vector of the hash functions $h_1, \ldots, h_g$ by $\bar{h}$ and the random seed vector of the AMS sketches by $\bar{\xi}$. For brevity, denote $\textsc{NoHvyColl}(H) = \mathcal{E}_{\bar{h},H}$. Let y_{ijr} denote the indicator variable that is 1 if items i and j collide to the same bucket in table r and is 0 otherwise, that is, $y_{ijr} \equiv h_r(i) = h_r(j)$. By independence of $\bar{h}$ and $\bar{\xi}$ and the independence of the ξ's across the tables,

$$\mathsf{E}\big[\hat{F}_1^H\big] = \mathsf{E}_{\bar{h}}\Big[\sum_{i \in H} \mathsf{E}_{\bar{\xi}}\Big[\sum_{j : h_{\theta(i)}(i) = h_{\theta(i)}(j)} f_j \xi_{\theta(i)}(j)\xi_{\theta(i)}(i) \mid \mathcal{E}_{\bar{h},H}\Big]\Big] = \mathsf{E}_{\bar{h}}\Big[\sum_{i \in H} \mathsf{E}_{\xi_{\theta(i)}}\Big[\sum_{j \in [n]} f_j \xi_{\theta(i)}(j)\xi_{\theta(i)}(i) y_{i,j,\theta(i)} \mid \mathcal{E}_{\bar{h},H}\Big]\Big]$$
$$= \mathsf{E}_{\bar{h}}\Big[\sum_{i \in H} |f_i| + \sum_{i \in H} \sum_{j : h_{\theta(i)}(i) = h_{\theta(i)}(j), j \neq i} f_j \mathsf{E}_{\xi_{\theta(i)}}\left[\xi_{\theta(i)}(j)\right] \mathsf{E}_{\xi_{\theta(i)}}\left[\xi_{\theta(i)}(i)\right] y_{i,j,\theta(i)} \mid \mathcal{E}_{\bar{h},H}\Big] = \mathsf{E}_{\bar{h}}\Big[\sum_{i \in H} |f_i|\Big] = \sum_{i \in H} |f_i| = F_1^H.$$

The third to last equality follows from the pair-wise independence of the family $\{\xi_r(j)\}_{j \in [n]}$ and so for $i \neq j$, $\mathsf{E}_{\xi_{\theta(i)}}\big[\xi_{\theta(i)}(i)\xi_{\theta(i)}(j)y_{ij\theta(i)}\big] = \mathsf{E}_{\xi_{\theta(i)}}\left[\xi_{\theta(i)}(j)\right] \mathsf{E}_{\xi_{\theta(i)}}\left[\xi_{\theta(i)}(i)\right] y_{ij\theta(i)} = 0 \cdot 0 \cdot y_{ij\theta(i)} = 0$, since $y_{ij\theta(i)}$ does not depend on $\bar{\xi}$. We now calculate $\mathsf{Var}\big[\hat{F}_1^H\big]$. Since, $(\xi_r(j))^2 = 1$,

$$\mathsf{E}\big[(\hat{F}_1^H)^2\big] = \mathsf{E}_{\bar{h},\bar{\xi}}\Big[\Big(\sum_{i \in H} \sum_{j \in [n]} f_j \xi_{\theta(i)}(j)\xi_{\theta(i)}(i) y_{ij\theta(i)}\Big)^2 \mid \mathcal{E}_{\bar{h},H}\Big] = \mathsf{E}_{\bar{h},\bar{\xi}}\Big[\sum_{i \in H, j \in [n]} f_j^2 y_{i,j,\theta(i)} \mid \mathcal{E}_{\bar{h},H}\Big] +$$
$$\mathsf{E}_{\bar{h},\bar{\xi}}\Big[\sum_{i \in H, j \in [n], i' \in H, j' \in [n], (i,j) \neq (i',j')} f_j f_{j'} \xi_{\theta(i)}(j)\xi_{\theta(i)}(i)\xi_{\theta(i')}(j')\xi_{\theta(i')}(i') y_{ij\theta(i)} y_{i'j'\theta(i')} \mid \mathcal{E}_{\bar{h},H}\Big] \qquad (2)$$

Consider the second term in the *RHS* of (2): $\mathsf{E}_{\bar{\xi}}\big[\xi_{\theta(i)}(j)\xi_{\theta(i)}(i)\xi_{\theta(i')}(j')\xi_{\theta(i')}(i')\big]$. There are two cases: (1) $\theta(i) = \theta(i')$ and (2) $\theta(i) \neq \theta(i')$. If $\theta(i) = \theta(i') = r$ (say), the term is $\mathsf{E}_{\xi_r}\left[\xi_r(i)\xi_r(j)\xi_r(i')\xi_r(j')\right]$. Since $(i,j) \neq (i',j')$, by 4-wise independence, this expectation is 0. In the other case, letting $r' = \theta(i')$, the term becomes $\mathsf{E}_{\xi_r,\xi_{r'}}\left[\xi_r(j)\xi_r(i)\xi_{r'}(i)\xi_{r'}(j')\right] = \mathsf{E}_{\xi_r}\left[\xi_r(i)\xi_r(j)\right] \mathsf{E}_{\xi_{r'}}\left[\xi_{r'}(i')\xi_{r'}(j')\right]$ by pair-wise independence among the seeds of $\{\xi_r\}_{r \in [g]}$. Since $(i,j) \neq (i',j')$ at least one of two expectations is 0 by 4-wise independence of $\{\xi_r(i)\}_{i \in [n]}$, for each $r \in [g]$. So $\mathsf{E}_{\bar{\xi}}\big[\xi_{\theta(i)}(j)\xi_{\theta(i)}(i)\xi_{\theta(i')}(j')\xi_{\theta(i')}(i')\big] = 0$. Since, $\bar{h}$ is independent of $\bar{\xi}$, therefore, $\mathsf{E}_{\bar{\xi}}\big[\xi_{\theta(i)}(j)\xi_{\theta(i)}(i)\xi_{\theta(i')}(j')\xi_{\theta(i')}(i') \mid \mathcal{E}_{\bar{h},H}\big] = 0$, for each $i \in H$, implying that the second expectation term is 0. Now consider the first term in the *RHS* of (2).

$$\mathsf{E}_{\bar{h},\bar{\xi}}\Big[\sum_{i \in H, j \in [n]} f_j^2 y_{i,j,\theta(i)} \mid \mathcal{E}_{\bar{h},H}\Big] = \mathsf{E}_{\bar{h}}\Big[\sum_{i \in H, j \in [n]} f_j^2 y_{i,j,\theta(i)} \mid \mathcal{E}_{\bar{h},H}\Big] = \sum_{i \in H, j \in [n]} f_j^2 \mathsf{Pr}_{\bar{h}}\left[y_{ij\theta(i)} = 1 \mid \mathcal{E}_{\bar{h},H}\right]$$

The hash family is assumed to be S-uniform for $|S| = 32B$ and $|H| \leq 1.5B$. So, for $i \in H$ and $j \in ([n] \setminus H)$, $\mathsf{Pr}_{h_r}\big[h_r(i) = h_r(j) \mid \mathcal{E}_{\bar{h},H}\big] = \mathsf{Pr}\left[h_r(i) = h_r(j)\right] = 1/C$. If $i = j$, then the above probability is 1, and if $i, j \in H$ and $i \neq j$, then, the above probability is 0, since $\mathcal{E}_{\bar{h},H}$ excludes this condition. Therefore,

$$\mathsf{Var}\big[\hat{F}_1^H\big] = \sum_{i \in H} f_i^2 + \sum_{i \in H, j \in [n] \setminus H} (1/C) f_j^2 - (F_1^H)^2 \leq (|H|/C) \sum_{j \in [n] \setminus H} f_j^2 = (|H|/C) F_2^{\mathrm{res}}(H)$$
$$\leq (|H|/C)(3/2) F_2^{\mathrm{res}}(|H|) \leq (3/2)(|H|/C) F_1^2/|H|^{2-1} = 3\epsilon^2 F_1^2/2048 \ .$$

Here $F_2^{\text{res}}(H) = \sum_{j \in [n] \setminus H} f_j^2$. The first inequality of the second line above uses Lemma 3 from [6] by setting $C = 64B$ and $|H| \leq (5/3)B$. The second inequality in second line follows using Lemma 2. Applying Chebychev's inequality, $|\hat{F}_1^H - F_1^H| \leq (\epsilon/4)F_1$ with probability at least 15/16.

Lemma 3. *Let H be the set of top-k items with respect to estimated frequencies using a* CountSketch *structure with C buckets per table. If $k \leq 8B$, then, $F_2^{res}(k) \leq F_2^{res}(K) \leq F_2^{res}(k)\big(1 + 2\sqrt{k/C} + (k/C)\big)$.* □

Adding the errors of the light and heavy estimator using triangle inequality, $|\hat{F}_1 - F_1| \leq |\hat{F}_1^H - F_1^H| + |\hat{F}_1^L - F_1^L| \leq (\epsilon/2 + \epsilon/4)F_1 \leq (3\epsilon/4)F_1$. The failure probability of the algorithm is 1/16 each for the application of Chebychev's bound for light and heavy estimation respectively plus the failure probability of 1/64 of GoodEst plus the failure probability of $((\log(1/\epsilon))+1)n^{-(2\log n)}$ for S-uniform hashing. The total failure probability is at most 1/7. The random seed length for storing the hash functions is $O(\epsilon^{-2}(\log(n\epsilon^{-1}))(\log(\epsilon^{-1}))$, for the AMS sketches is $O(\epsilon^{-2}(\log(\epsilon^{-1}))(\log(mM))$ bits. The seed for PRG is $O((\log(mM))(\log n)))$ bits long. The time taken to evaluate the hash functions is $O((\log \epsilon^{-1})(\log(n\epsilon^{-1}))^2)$ and the time required to generate the relevant portion of the generated seed is $O(\log n)$. We have now proved the following theorem.

Theorem 4. *For each $0 < \epsilon < 1$, there is an algorithm that returns a $1 \pm \epsilon$-factor accurate estimate for F_1 with probability 6/7 using space $O(\epsilon^{-2}(\log(n\epsilon^{-1}mM))(\log(\epsilon^{-1}))$ bits and can process each stream update in time $O((\log \epsilon^{-1})(\log(n\epsilon^{-1}))^2)$.* □

References

[1] Noga Alon, Yossi Matias, and Mario Szegedy. "The space complexity of approximating frequency moments". *J. Comp. Sys. and Sc.*, 58(1):137–147, 1998. Preliminary version appeared in Proceedings of *ACM STOC* 1996, pp. 1-10.

[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 ACM Symposium on Theory of Computing STOC*, pages 209–218, Princeton, NJ, 2002.

[3] Moses Charikar, Kevin Chen, and Martin Farach-Colton. "Finding frequent items in data streams". *Theoretical Computer Science*, 312(1):3–15, 2004. Preliminary version appeared in Proceedings of ICALP 2002, pages 693-703.

[4] Joanne Feigenbaum, Sampath Kannan, Martin Strauss, and M. Viswanathan. "An Approximate L^1-Difference Algorithm for Massive Data Streams". In *Proceedings of IEEE FOCS*, pages 501–511, New York, NY, October 1999.

[5] P. Flajolet and G.N. Martin. "Probabilistic Counting Algorithms for Database Applications". *J. Comp. Sys. and Sc.*, 31(2):182–209, 1985.

[6] S. Ganguly, D. Kesh, and C. Saha. "Practical Algorithms for Tracking Database Join Sizes". In *Proceedings of Foundations of Software Technoogy and Theoretical Computer Science (FSTTCS)*, pages 294–305, Hyderabad, India, December 2005.

[7] Sumit Ganguly and Graham Cormode. "On Estimating Frequency Moments of Data Streams". In *Proceedings of International Workshop on Randomization and Computation (RANDOM)*, 2007.

[8] Sumit Ganguly and Purushottam Kar. "On Estimating the First Frequency Moment of Data Streams". http://arxiv.org/abs/1005.0809, 2010.

[9] Piotr Indyk. Stable distributions, pseudorandom generators, embeddings, and data stream computation. *J. ACM*, 53(3):307–323, 2006. Preliminary Version appeared in Proceedings of IEEE FOCS 2000, pages 189-197.

[10] Daniel M. Kane, Jelani Nelson, and David P. Woodruff. "On the Exact Space Complexity of Sketching and Streaming Small Norms". In *Proceedings of ACM Symposium on Discrete Algorithms (SODA)*, 2010.

[11] Ping Li. Estimators and tail bounds for dimension reduction in ℓ_α $(0 < \alpha \leq 2)$ using stable random projections. In *Proceedings of ACM Symposium on Discrete Algorithms (SODA)*, pages 10–19, 2008.

[12] Jelani Nelson and David P. Woodruff. "Fast Manhattan Sketches in Data Streams". In *Proceedings of ACM International Symposium on Principles of Database Systems (PODS)*, June 2010.

[13] N. Nisan. "Pseudo-Random Generators for Space Bounded Computation". In *Proceedings of ACM Symposium on Theory of Computing STOC*, pages 204–212, May 1990.

[14] J. P. Nolan. *Stable Distributions - Models for Heavy Tailed Data.* Birkhäuser, Boston, To be published. In progress, Chapter 1 online at academic2.american.edu/∼jpnolan.

[15] Alan Siegel. "On universal classes of extremely random constant-time hash functions and their time-space trade-off". Technical Report, Courant Institute, New York University, 1995. Available from http://www.cs.nyu.edu/siegel/.

[16] David P. Woodruff. "Optimal space lower bounds for all frequency moments". In *Proceedings of ACM Symposium on Discrete Algorithms (SODA)*, pages 167–175, 2004.