Hierarchical Sampling from Sketches: Estimating Functions over Data Streams

Sumit Ganguly[1] and Lakshminath Bhuvanagiri[2]

[1] Indian Institute of Technology, Kanpur
[2] Google Inc., Bangalore

Abstract. We present a randomized procedure named Hierarchical Sampling from Sketches (HSS) that can be used for estimating a class of functions over the frequency vector f of update streams of the form $\Psi(\mathcal{S}) = \sum_{i=1}^{n} \psi(|f_i|)$. We illustrate this by applying the HSS technique to design nearly space-optimal algorithms for estimating the pth moment of the frequency vector, for real $p \geq 2$ and for estimating the entropy of a data stream. [3]

1 Introduction

A variety of applications in diverse areas, such as, networking, database systems, sensor networks, web-applications, share some common characteristics, namely, that data is generated rapidly and continuously, and must be analyzed in real-time and in a single-pass over the data to identify large trends, anomalies, user-defined exception conditions, etc.. Furthermore, it is frequently sufficient to continuously track the “big picture”, or, an aggregate view of the data. In this context, efficient and approximate computation with bounded error probability is often acceptable. The data stream model presents a computational model for such applications, where, incoming data is processed in an online fashion using sub-linear space.

1.1 The data stream model

A data stream $\mathcal{S}$ is viewed as a sequence of records of the form (pos, i, v), where, *pos* is the index of the record in the sequence, i is the identity of an item in $[1, n] = \{1, \ldots, n\}$, and v is the *change* to the frequency of the item. $v > 0$ indicates an insertion of multiplicity v, while $v < 0$ indicates a corresponding deletion. The frequency of an item i, denoted by f_i, is the sum of the changes to the frequency of i since the inception of the stream, that

[3] Preliminary version of this paper appeared as the following conference publications. “Simpler algorithm for estimating frequency moments of data streams”, Lakshminath Bhuvanagiri, Sumit Ganguly, Deepanjan Kesh and Chandan Saha, *Proceedings of the ACM Symposium on Discrete Algorithms*, 2006, pp. 708-713 and “Estimating Entropy over Data Streams”, Lakshminath Bhuvanagiri and Sumit Ganguly, *Proceedings of the European Symposium on Algorithms, Springer LNCS Volume 4168*, pp. 148-159, 2006.

is, $f_i = \sum_{(pos,i,v) \text{ appears in } \mathcal{S}} v$. The following variations of the data stream model have been considered in the research literature.

1. The *insert-only* model, where, data streams to not have deletions, that is, $v > 0$ for all records. The *unit insert-only* model is a special case of the insert-only model, where, $v = 1$ for all records.
2. The *strict update model*, where, insertions and deletions are allowed, subject to the constraint that $f_i \geq 0$, for all $i \in [1, n]$.
3. The *general update model* where no constraints are placed on insertions and deletions.
4. The *sliding window model*, where, a window size parameter W is given and only the portion of the stream that has arrived within the last W time units is considered to be relevant. Records that are not part of the the current window are deemed to have expired.

In the data stream model, an algorithm must perform its computations in an online manner, that is, the algorithm gets to view each stream record exactly once. Further, the computation is constrained to use sub-linear space, that is, $o(n \log F_1)$ bits, where, $F_1 = \sum_{i=1}^{n} |f_i|$. This implies that the stream cannot be stored in its entirety and summary structures must be devised to solve the problem.

We say that an algorithm estimates a quantity C with ϵ-accuracy and probability $1 - \delta$ if it returns an estimate that satisfies $|\hat{C} - C| \leq \epsilon C$ with probability $1 - \delta$. The probability is assumed to hold for every instance of input data and is taken over the random coin tosses used by the algorithm. More simply, we say that a randomized algorithm estimates C with ϵ-accuracy if it returns an estimate satisfying $|\hat{C} - C| \leq \epsilon C$ with constant probability greater than $\frac{1}{2}$ (for e.g., $\frac{7}{8}$). Such an estimator can be used to obtain another estimator for C that is ϵ-accurate and is correct with probability $1 - \delta$ by following the standard procedure of returning the the median of $s = O(\log \frac{1}{\delta})$ independent such estimates $\hat{C}_1, \ldots, \hat{C}_s$.

In this paper, we consider two basic problems in the data streaming model, namely, estimating the moment of the frequency vector of a data stream and estimating the entropy of a data stream. We first define the problems and review the research literature.

1.2 Previous work on estimating F_p for $p \geq 2$

For any real $p \geq 0$, the p^{th} moment of the frequency vector of the stream $\mathcal{S}$ is defined as

$$F_p(\mathcal{S}) = \sum_{i=1}^{n} |f_i|^p \ .$$

The problem of estimating F_p has led to a number of advancements in the design of algorithms and lower bound techniques for data stream computation. It was first introduced in

[1] that also presented the first sub-linear space randomized algorithm for estimating F_p, for $p > 1$ with ϵ-accuracy and using space $O(\frac{1}{\epsilon^2} n^{1-1/p} \log F_1)$ bits. For the special case of F_2, the seminal sketch technique was presented in [1], that uses space $O(\frac{1}{\epsilon^2} \log F_1)$ bits for estimating F_2 with ϵ-accuracy. The work in [9, 13] reduced the space requirement for ϵ-accurate estimation of F_p, for $p > 2$, to $O\left(\frac{1}{\epsilon^2} n^{1-1/(p-1)} (\log F_1)\right)$.[4] The space requirement was reduced in [12] to $O(n^{1-2/(p+1)} (\log F_1))$, for $p > 2$ and p integral. A space lower bound of $\Omega(n^{1-\frac{2}{p}})$ for estimating F_p, for $p > 2$, was shown in a series of contributions [1, 2, 7] (see also [20]). Finally, Indyk and Woodruff [18] presented the first algorithm for estimating F_p, for real $p > 2$, that matched the above space lower bound up to poly-logarithmic factors. The 2-pass algorithm of Indyk and Woodruff requires space $O(\frac{1}{\epsilon^{12}}\, n^{1-2/p} (\log^2 n)(\log^6 F_1))$ bits. The 1-pass data streaming algorithm derived from the 2-pass algorithm further increases the constant and poly-logarithmic factors. The work in [24] presents an $\Omega(\frac{1}{\epsilon^2})$ space lower bound for ϵ-accurate estimation of F_p, for any real $p \neq 1$ and $\epsilon \geq \frac{1}{\sqrt{n}}$.

1.3 Previous work on estimating entropy

The entropy H of a data stream is defined as

$$H = \sum_{i \in [1,n] : f_i \neq 0} \frac{|f_i|}{F_1} \log \frac{F_1}{|f_i|} \quad .$$

It is a measure of the information theoretic *randomness* or the *incompressibility* of the data stream. A value of entropy close to $\log n$, is indicative that the frequencies in the stream are randomly distributed, whereas, low values are indicative of "patterns" in the data. Monitoring changes in the entropy of a network traffic stream has been used to detect anomalies [15, 22, 25]. The work in [16] presents an algorithm for estimating H over unit insert-only streams to within α-approximation (i.e., $\frac{H}{a} \leq \hat{H} \leq bH$ and $a \cdot b \leq \alpha$) over insert-only streams using space $O\big(\frac{n^{1/\alpha}}{H}\big)$. The work in [5] presents an algorithm for estimating H to within α-approximation over insert-only streams using space $O\left(\frac{1}{\epsilon^2} F_1^{2/\alpha} (\log F_1)^2 (\log F_1 + \log n)\right)$. Subsequent to the publication of the conference version [3] of our algorithm, a different algorithm for estimating the entropy of insert-only streams was presented in [6]. The algorithm of [6] requires space $O\big(\frac{1}{\epsilon^2} (\log n)(\log F_1)(\log F_1 + \log(1/\epsilon))\big)$ respectively. The work in [6] also shows a space lower bound of $\Omega\left(\frac{1}{\epsilon^2 \log(1/\epsilon)}\right)$ for estimating entropy over a data stream.

[4] The algorithm of [13] assumes p to be integral.

1.4 Contributions

In this paper we present a technique called *hierarchical sampling from sketches* or HSS, that is a general randomized technique for estimating a variety of functions over update data streams. We illustrate the HSS technique by applying it to derive near-optimal space algorithms for estimating frequency moments F_p, for real $p > 2$. The HSS algorithm estimates F_p, for any real $p \geq 2$ to within ϵ-accuracy using space

$$O\left(\frac{p^2}{\epsilon^{2+4/p}}\, n^{1-2/p} \;\cdot (\log F_1)^2 \cdot (\log n + \log\log F_1)^2\right) \quad .$$

Thus, the space upper bound of these algorithms match the lower bounds up to factors that are poly-logarithmic in F_1, n and polynomial in $\frac{1}{\epsilon}$. The expected time required to process each stream update is $O(\log n + \log\log F_1)$ operations. The algorithm essentially uses the idea of Indyk and Woodruff [18] to classify items into groups, based on frequency. However, Indyk and Woodruff define groups whose boundaries are randomized; in our algorithm, the group boundaries are deterministic. The HSS technique is used to design an algorithm that estimates the entropy of general update data streams to within ϵ-accuracy using space

$$O\left(\frac{(\log F_1)^2}{\epsilon^3 \log(1/\epsilon)} \cdot (\log n + \log\log F_1)^2\right) \quad .$$

Organization. The remainder of the paper is organized as follows. In Section 2, we review relevant algorithms from the research literature. In Section 3, we present the HSS technique for estimating a class of data stream metrics. Sections 4 and 5 use the HSS technique to estimate frequency moments and entropy, respectively, of a data stream. Finally, we conclude in Section 6.

2 Preliminaries

In this section, we review the COUNTSKETCH and the COUNT-MIN algorithms for finding frequent items in a data stream. We also review algorithms to estimate the residual second moment of a data stream [14]. For completeness, we also present an algorithm for estimating the residual first moment of a data stream.

Given a data stream, *rank*(r) is an item with the r^{th} largest absolute value of the frequency, where, ties are broken arbitrarily. We say that an item i has rank r if *rank*$(r) = i$. For a given value of k, $1 \leq k \leq n$, the set *top*(k) is the set of items with rank $\leq k$. The residual second moment [8] of a data stream, denoted by $F_2^{res}(k)$, is defined as the second moment of the stream after the top-k frequencies have been removed, that is, $F_2^{res}(k) = \sum_{r>k} f^2_{rank(r)}$.

The residual first moment [10] of a data stream, denoted by F_1^{res}, is analogously defined as the F_1 norm of the data stream after the top-k frequencies have been removed, that is, $F_1^{res} = \sum_{r>k} |f_{rank(r)}|$.

Sketches. A *linear sketch* [1] is a random integer $X = \sum_i f_i \cdot x_i$, where, $x_i \in \{-1,+1\}$, for $i \in [1,n]$ and the family of variables $\{x_i\}_{i\in[1,n]}$ is either pair-wise or 4-wise independent, depending on the use of the sketches. The family of random variables $\{x_i\}_{i\in\mathcal{D}}$ is referred to as the *linear sketch basis*. For any $d \geq 2$, a d-wise independent linear sketch basis can be constructed in a pseudo-random manner from a truly random seed of size $O(d \log n)$ bits as follows. Let F be field of characteristic 2 and of size at least $n+1$. Choose a degree $d-1$ polynomial $g : F \rightarrow F$ with coefficients that are randomly chosen from F [4, 23]. Define x_i to be 1 if the first bit of $g(i)$ is 1, and define x_i to be -1 otherwise. The d-wise independence of the x_i's follows from Wegman and Carter's universal hash functions [23].

Linear sketches were pioneered by [1] to present an elegant and efficient algorithm for returning an estimate of the second moment F_2 of a stream to within a factor of $(1 \pm \epsilon)$ with probability $\frac{7}{8}$. Their procedure keeps $s = O(\frac{1}{\epsilon^2})$ independent linear sketches (i.e., using independent sketch bases) $X_1, X_2, \ldots, X_s$, where, the sketch basis used for each X_i is four-wise independent. The algorithm returns $\hat{F}_2$ as the average of the square of the sketches, that is, $\hat{F}_2 = \frac{1}{s}\sum_{j=1}^{s} X_j^2$. A crucial property observed in [1] is that $\mathsf{E}\left[X_j^2\right] = F_2$ and $\mathsf{Var}\left[X_j^2\right] \leq 5F_2^2$. In the remainder of the paper, we abbreviate linear sketches as simply, sketches.

COUNTSKETCH **algorithm for estimating frequency.** Pair-wise independent sketches are used in [8] to design the COUNTSKETCH algorithm for estimating the frequency of any given items $i \in [1,n]$ of a stream. The data structure consists of a collection of $s = O(\log \frac{1}{\delta})$ independent hash tables $U_1, U_2, \ldots, U_s$, each consisting of $8k$ buckets. A pair-wise independent hash function $h_j : [1,n] \rightarrow \{1, 2, \ldots, 8k\}$ is associated with each hash table that maps items randomly to one of the $8k$ buckets. Additionally, for each table index $j = 1, 2, \ldots, s$, the algorithm keeps a pair-wise independent family of random variables $\{x_{ij}\}_{i\in[1,n]}$, where, each $x_{ij} \in \{-1,+1\}$ with equal probability. Each bucket keeps a sketch of the sub-stream that maps to it, that is, $U_j[r] = \sum_{i:h_j(i)=r} f_i x_{ij}$, $i \in \{1,2,\ldots,s\}$ and $j \in \{1,2,\ldots,s\}$. An estimate $\hat{f}_i$ is returned as follows: $\hat{f}_i = \text{median}_{j=1}^{s}\, U_j[h_j(i)]x_{ij}$. The accuracy of estimation is stated as a function Δ of the residual second moment defined as [8]

$$\Delta(s, A) \overset{\text{def}}{=} 8\left(\frac{F_2^{res}(s)}{A}\right)^{1/2} .$$

The space versus accuracy guarantees of the COUNTSKETCH algorithm is presented in Theorem 1.

Theorem 1 ([8]). *Let $\Delta = \Delta(k, 8k)$. Then, for $i \in [1, n]$, $\Pr\left\{|\hat{f}_i - f_i| \le \Delta\right\} \ge 1 - \delta$. The space used is $O(k(\log\frac{1}{\delta})(\log F_1))$ bits and the time taken to process a stream update is $O(\log\frac{1}{\delta})$.* □

COUNT-MIN **Sketch for estimating frequency.** The COUNT-MIN algorithm [10] for estimating frequencies keeps a collection of $s = O(\log\frac{1}{\delta})$ independent hash tables $T_1, T_2, \ldots, T_s$, where, each hash table T_j is of size $b = 2k$ buckets and uses a pair-wise independent hash function $h_j : [1, n] \to \{1, \ldots, 2k\}$, for $j = 1, 2, \ldots, s$. The bucket $T_j[r]$ is an integer counter that maintains the following sum: $T_j[r] = \sum_{i:h_j(i)=r} f_i$. The estimated frequency $\hat{f}_i$ is obtained as $\hat{f}_i = \text{median}_{r=1}^{s} T_j[h_j(i)]$. The space versus accuracy guarantees for the COUNT-MIN algorithm is given in terms of the quantity $F_1^{res}(k) = \sum_{r>k} |f_{\mathit{rank}(r)}|$.

Theorem 2 ([10]). $\Pr\left\{|\hat{f}_i - f_i| \le \frac{F_1^{res}(k)}{k}\right\} \ge 1 - \delta$. *The space used is $O\left(k\left(\log\frac{1}{\delta}\right)(\log F_1)\right)$ bits and time $O\left(\log\frac{1}{\delta}\right)$ to process each stream update.* □

Estimating F_2^{res}. The work in [14] presents an algorithm to estimate $F_2^{res}(s)$ to within an accuracy of $(1 \pm \epsilon)$ with confidence $1 - \delta$ using space $O(\frac{s}{\epsilon^2}\log(F_1)\log(\frac{n}{\delta}))$ bits. The data structure used is identical to the COUNTSKETCH structure. The algorithm basically removes the top-s estimated frequencies from the COUNTSKETCH structure and then estimates F_2. The COUNTSKETCH structure is used to find the top-k items with respect to the absolute values of their estimated frequencies. Let $|\hat{f}_{\tau_1}| \ge \ldots \ge |\hat{f}_{\tau_k}|$ denote the top-k estimated frequencies. Next, the contributions of these estimates are removed from the structure, that is, $U_j[r]$:=$U_j[r] - \sum_{t:h_j(\tau_t)=r} f_{\tau_t} x_{j\tau_t}$. Subsequently, the FASTAMS algorithm [21], a variant of the original sketch algorithm [1], is used to estimate the second moment as follows: $\hat{F}_2^{res} = \text{median}_{j=1}^{s} \sum_{r=1}^{8k} (U_j[r])^2$. If $k = O(\epsilon^{-2}s)$, then, $|\hat{F}_2^{res} - F_2^{res}| \le \epsilon\hat{F}_2^{res}$ [14].

Lemma 1 ([14]). *For a given integer $k \ge 1$ and $0 < \epsilon < 1$, there exists an algorithm for update streams that returns an estimate $\hat{F}_2^{res}(k)$ satisfying $|\hat{F}_2^{res}(k) - F_2^{res}(k)| \le \epsilon F_2^{res}(k)$ with probability $1 - \delta$ using space $O(\frac{k}{\epsilon^2}(\log\frac{F_1}{\delta})(\log F_1))$ bits.* □

Estimating F_1^{res}. An argument similar to the one used to estimate $F_2^{res}(k)$ can be applied to estimate $F_1^{res}(k)$. We will prove the following property in this subsection.

Lemma 2. *For $0 < \epsilon \leq 1$, there exists an algorithm for update streams that returns an estimate $\hat{F}_1^{res}$ satisfying* $\Pr\left\{|\hat{F}_1^{res} - F_1^{res}(k)| \leq \epsilon F_1^{res}(k)\right\} \geq \frac{3}{4}$. *The algorithm uses* $O\left(\frac{k(\log F_1)}{\epsilon} + \frac{(\log F_1)(\log n + \log(1/\epsilon))}{\epsilon^2}\right)$ *bits.* □

The algorithm is the following. Keep a COUNT-MIN sketch structure with height b, where, b is a parameter to be fixed, and width $s = O(\log \frac{n}{\delta})$. In parallel, we keep a set of $s = O(\frac{1}{\epsilon^2})$ sketches based on a 1-stable distribution [17]. A one-stable sketch is a linear sketch $Y_j = \sum_i f_i z_{j,i}$, where, the $z_{j,i}$'s are drawn from a 1-stable distribution, $j = 1, 2, \ldots, s$. As shown by Indyk [17], the random variable

$$\hat{Y} = \text{median}_{j=1}^{s} |Y_j| \text{ satisfies } \Pr\left\{|\hat{Y} - F_1| \leq \epsilon F_1\right\} \geq \frac{7}{8} \ .$$

The COUNT-MIN structure is used to obtain the top-k elements with respect to the absolute value of their estimated frequencies. Suppose these items are $i_1, i_2, \ldots, i_k$ and $|\hat{f}_{i_1}| \geq |\hat{f}_{i_2}| \geq \ldots \geq |\hat{f}_{i_k}|$. Let $I = \{i_1, i_2, \ldots, i_k\}$. Each one-stable sketch Y_j is updated to remove the contribution of the estimated frequencies. That is,

$$Y_j' = Y_j - \sum_{r=1}^{k} \hat{f}_{i_r} z_{j,i_r} \ .$$

Finally, the following value is returned.

$$\hat{F}_1^{res}(k) = \text{median}_{j=1}^{k} |Y_j'| \ .$$

We now analyze the algorithm. Let $T = T_k = \{t_1, t_2, \ldots, t_k\}$ denote the set of indices of the top-k (true) frequencies, such that $|f_{t_1}| \geq |f_{t_2}| \geq \ldots \geq |f_{t_k}|$.

Lemma 3.

$$F_1^{res}(k) \leq \sum_{i \in [1,n],\ i \notin I} |f_i| \leq F_1^{res}(k)\left(1 + \frac{k}{b}\right) \ .$$

Proof. Since both T and I are sets of k elements each, therefore, $|T - I| = |I - T|$. Let $i \mapsto i'$ be an arbitrary 1-1 map from $i \in T - I$ to an element i' in $I - T$. Since, i is a top-k frequency and i' is not, therefore, $f_i \geq f_{i'}$. Further, $\hat{f}_{i'} \geq \hat{f}_i$, otherwise, i would be among the top-k items with respect to estimated frequencies, that is i would be in I, contradicting that $i \in T - I$. The condition $\hat{f}_{i'} \geq \hat{f}_i$ implies the following.

$$f_i - \Delta \leq \hat{f}_i \leq \hat{f}_{i'} \leq f_{i'} + \Delta$$

with probability $1 - \frac{\delta}{4n}$, or, that $f_i \leq f_{i'} + 2\Delta$, with probability $1 - \frac{\delta}{4n}$. Therefore,

$$f_{i'} \leq f_i \leq f_{i'} + \Delta \ .$$

Thus, using union bound for all items in $[1, n]$, we have with probability $1 - \frac{\delta}{2}$,

$$\sum_{i \in I-T} f_{i'} \le \sum_{i \in T-I} f_i \le \sum_{i \in I-T} f_{i'} + \Delta = \sum_{i \in I-T} f_{i'} + k\Delta \ . \tag{1}$$

Let

$$G = \sum_{i \in [1,n]-I} |f_i| = \sum_{i \in T-I} f_i + \sum_{i \notin (T \cup I)} f_i$$

By (1), it follows that

$$\sum_{i \in I-T} f_i \quad + \quad \sum_{i \notin (T \cup I)} f_i \quad \le \quad G \quad \le \quad \sum_{i \in I-T} f_i \quad + \quad k\Delta \quad + \quad \sum_{i \notin (T \cup I)} f_i \ .$$

Since, $\sum_{i \in I-T} f_i + \sum_{i \notin (T \cup I)} f_i = F_1^{res}(k)$, we have,

$$F_1^{res}(k) \le G \le F_1^{res}(k) + k\Delta \ .$$

By Theorem 2, we have, $\Delta \le \frac{F_1^{res}(k)}{b}$. Therefore,

$$F_1^{res}(k) \le G \le F_1^{res}(k)\left(1 + \frac{k}{b}\right), \text{ with prob. } 1 - \frac{\delta}{2} \ . \qquad \square$$

Proof (Of Lemma 2.). Let f' denote the frequency vector after the estimated frequencies are removed. Then,

$$f'_r = \begin{cases} f_r & \text{if } r \in [1, n] - I \\ f_r - \hat{f}_r & \text{if } r \in I. \end{cases}$$

Let F'_1 denote $\sum_{i \in [1,n]} |f'_i|$. Then, by Lemma 3, it follows that

$$F'_1 = \sum_{i \in [1,n]-I} f_r + \sum_{i \in I} (f_i - \hat{f}_i) \le G + k\Delta \le F_1^{res}(k)\left(1 + \frac{2k}{b}\right)$$

since, $\Delta \le \frac{F_1^{res}(b)}{b} \le \frac{F_1^{res}(k)}{b}$. Further,

$$F'_1 = \sum_{i \in [1,n]-I} f_r + \sum_{i \in I} (f_i - \hat{f}_i) \ge G - k\Delta \ge F_1^{res}(k)\left(1 - \frac{k}{b}\right)$$

Combining,

$$F_1^{res}(k)\left(1 - \frac{k}{b}\right) \le F'_1 \le F_1^{res}(k)\left(1 + \frac{2k}{b}\right)$$

with probability $1 - 2^{-\Omega(w)}$. By construction, the 1-stable sketches technique returns $\hat{F}_1^{res} = \text{median}_{j=1}^{s} |Y'_j|$ satisfying $\Pr\left\{ |\hat{F}_1^{res} - F'_1| \leq \epsilon F'_1 \right\} \geq \frac{7}{8}$. Therefore, using union bound

$$|\hat{F}_1^{res} - F_1^{res}(k)| \leq \frac{2k}{b} F_1^{res}(k) + \epsilon F_1^{res}(k) \left(1 + \frac{2k}{b}\right) \text{ with prob. } \frac{7}{8} - n2^{-\Omega(w)}$$

If $b \geq \lceil \frac{2k}{\epsilon} \rceil$, then,

$$|\hat{F}_1^{res} - F_1^{res}(k)| \leq 3\epsilon F_1^{res}(k) \quad \text{with prob. } \frac{3}{4} \enspace .$$

Replacing ϵ by $\epsilon/3$ yields the first statement of the lemma.

The space requirement is calculated as follows. We use the idea of a pseudo-random generator of Indyk [17] for streaming computations. The state of the data structure (i.e., stable sketches) is the same as if the items arrived in sorted order. Therefore, as observed by Indyk [17], Nisan's pseudo-random generator [19] can be used to simulate the randomized computation using $O(S \log R)$ random bits, where, S is the space used by the algorithm (not counting the random bits used) and R is the running time of the algorithm. We apply Indyk's observation to the portion of the algorithm that estimates F_1. Thus, $S = O\left(\frac{(\log F_1)}{\epsilon^2}\right)$ and the running time $R = O\left(\frac{F_0}{\epsilon^2}\right)$, assuming the input is presented in sorted order of items. Here, F_0 is the number of distinct items in the stream with non-zero frequency. Then the total random bits required is $O(S \log R) = O\left(\frac{(\log F_1)(\log n + \log(1/\epsilon)}{\epsilon^2}\right)$. □

3 The HSS algorithm

In this section, we present a procedure for obtaining a *representative sample* over the input stream, which we refer to as *Hierarchical Sampling over Sketches* (HSS) and use it for estimating a class of metrics over data-streams of the following form

$$\Psi(\mathcal{S}) = \sum_{i: f_i \neq 0} \psi(|f_i|) \enspace . \tag{2}$$

Sampling sub-streams. The HSS algorithm uses a sampling scheme as follows. From the input stream $\mathcal{S}$, we create sub-streams $\mathcal{S}_0, \ldots, \mathcal{S}_L$ such that $\mathcal{S}_0 = \mathcal{S}$ and for $1 \leq l \leq L$, $\mathcal{S}_l$ is obtained from $\mathcal{S}_{l-1}$ by sub-sampling each distinct item appearing in $\mathcal{S}_{l-1}$ independently with probability $\frac{1}{2}$. At level 0, $\mathcal{S}_0 = \mathcal{S}$. The stream $\mathcal{S}_1$, corresponding to level 1, is obtained by sampling choosing each distinct value of i with independently, with probability $\frac{1}{2}$. Since, the sampling is based on the identity of an item i, either all records in $\mathcal{S}$ with identity i are present in $\mathcal{S}_1$, or, none are–each of these cases holds with probability $\frac{1}{2}$. The stream $\mathcal{S}_2$, corresponding to level 2 is obtained by sampling each distinct value of i appearing in the sub-stream $\mathcal{S}_1$, with probability $\frac{1}{2}$ and independently of the other items in $\mathcal{S}_1$. In this manner, $\mathcal{S}_l$ is a randomly sampled sub-stream of $\mathcal{S}_{l-1}$, for $l \geq 1$, based on the identity of the items.

The sub-sampling scheme is implemented as follows. We assume that n is a power of 2. Let $h : [1, n] \rightarrow [0, \max(n^2, W)]$ be a random hash function drawn from a pair-wise independent hash family and $W \geq 2F_1$. Let $L_{\max} = \lceil \log(\max(n^2, W)) \rceil$. Define the random function level : $[1, n] \rightarrow [1, L_{\max}]$ as follows.

$$\text{level}(i) = \begin{cases} 1 & \text{if } h(i) = 0 \\ \text{lsb}(h(i)) & 2 \leq \text{level}(i) \leq L_{\max} \end{cases} .$$

where, $lsb(x)$ is the position of the least significant "1" in the binary representation of x. The probability distribution of the random level function is as follows.

$$\mathsf{Pr}\left\{\text{level}(i) = l\right\} = \begin{cases} \frac{1}{2} + \frac{1}{n} & \text{if } l = 1 \\ \frac{1}{2^l} & \text{otherwise.} \end{cases}$$

All records pertaining to i are included in the sub-streams $\mathcal{S}_0$ through $\mathcal{S}_{\text{level}(i)}$. The sampling technique is based on the original idea of Flajolet and Martin [11] for estimating the number of distinct items in a data stream. The Flajolet-Martin scheme maps the original stream into disjoint sub-streams $\mathcal{S}'_1, \mathcal{S}'_2, \ldots, \mathcal{S}'_{\lceil \log n \rceil}$, where, $\mathcal{S}'_l$ is the sequence of records of the form $(\textit{pos}, i, v)$ such that $\text{level}(i) = l$. The HSS technique creates a monotonic decreasing sequence of random sub-streams in the sense that $\mathcal{S}_0 \supset \mathcal{S}_1 \supset \mathcal{S}_2 \ldots \supset \mathcal{S}_{L_{\max}}$ and $\mathcal{S}_l$ is the sequence of records for item i such that $\text{level}(i) \geq l$.

At each level $l \in \{0, 1, \ldots, L_{\max}\}$, the HSS algorithm keeps a frequency estimation data-structure denoted by $\mathcal{D}_l$, that takes as input the sub-stream $\mathcal{S}_l$, and returns an approximation to the frequencies of items that map to $\mathcal{S}_l$. The $\mathcal{D}_l$ structure can be any standard data structure such as the COUNT-MIN sketch structure or the COUNTSKETCH structure. We use the COUNT-MIN structure for estimating entropy and the COUNTSKETCH structure for estimating F_p. Each stream update $(\textit{pos}, i, v)$ belonging to $\mathcal{S}_l$ is propagated to the frequent items data structures $\mathcal{D}_l$ for $0 \leq l \leq \text{level}(i)$. Let $k(l)$ denote a space parameter for the data structure $\mathcal{D}_l$, for example, $k(l)$ is the size of the hash tables in the COUNT-MIN or COUNTSKETCH structures. The values of $k(l)$ are the same for levels $l = 1, 2, \ldots, L_{\max}$ and is twice the value for $k(0)$. That is, if $k = k(0)$, then, $k(1) = \ldots = k(L_{\max}) = 2k$. This non-uniformity is a technicality required by Lemma 4 and Corollary 1. We refer to $k = k(0)$ as the space parameter of the HSS structure.

Approximating f_i. Let $\mathit{\Delta}_l(k)$ denote the additive error of the frequency estimation by the data structure $\mathcal{D}_l$) at level l and using space parameter k. That is, we assume that

$$|\hat{f}_{i,l} - f_i| \le \mathit{\Delta}_l(k) \text{ with probability } 1 - 2^{-t}$$

where, t is a parameter and $\hat{f}_{i,l}$ is the estimate for the frequency of f_i obtained using the frequent items structure $\mathcal{D}_l(k)$. By Theorem 2, if $\mathcal{D}_l$ is instantiated by the COUNT-MIN sketch structure with height k and width $\lceil \log t \rceil$, then, $|\hat{f}_{i,l} - f_i| \le \frac{F_1^{res}(k,l)}{k}$ with probability $1-2^{-t}$. If $\mathcal{D}_l$ is instantiated using the COUNTSKETCH structure with height $8k$ and width $O(\log t)$, then, by Theorem 1, it follows that $|\hat{f}_{i,l} - f_i| \le 8\left(\frac{F_2^{res}(k,l)}{k}\right)^{1/2}$ with probability $1 - 2^{-t}$. We first relate the random values $F_1^{res}(k,l)$ and $F_2^{res}(k,l)$ to their corresponding non-random values $F_1^{res}(k)$ and $F_2^{res}(k)$, respectively.

Lemma 4. *For $l \ge 1$ and $k \ge 2$, $\Pr\left\{F_1^{res}(k,l) \le \frac{F_1^{res}(2^{l-1}k)}{2^{l-1}}\right\} \ge 1 - 2e^{-k/6}$.*

Proof. For item $i \in [1,n]$, define an indicator variable x_i to be 1 if records corresponding to i are included in the stream at level l, namely, $\mathcal{S}_l$, and let x_i be 0 otherwise. Then, $\Pr\{x_i = 1\} = \frac{1}{2^l}$. Define the random variable $T_{l,k}$ as the number of items in $[1,n]$ with rank at most $2^{l-1}k$ in the original stream $\mathcal{S}$ and that are included in $\mathcal{S}_l$. That is,

$$T_{l,k} = \sum_{1 \le \mathit{rank}(i) \le 2^{l-1}k} x_i \enspace .$$

By linearity of expectation, $\mathsf{E}\left[T_{l,k}\right] = \frac{2^{l-1}k}{2^l} = \frac{k}{2}$. Applying Chernoff's bounds to the sum of indicator variables $T_{l,k}$, we obtain

$$\Pr\{T_{l,k} > k\} < e^{-k/6} \enspace .$$

Say that the event $\text{SPARSE}(l)$ occurs if the element with rank $k+1$ in $\mathcal{S}_l$ has rank larger than $2^{l-1}k$ in the original stream $\mathcal{S}$. The argument above shows that $\Pr\{\text{SPARSE}(l)\} \ge 1 - e^{-k/6}$. Thus,

$$F_1^{res}(k,l) \le \sum_{\mathit{rank}(i) > 2^{l-1}k} |f_i| x_i, \quad \text{assuming SPARSE}(l) \text{ holds.} \tag{3}$$

By linearity of expectation

$$\mathsf{E}\left[F_1^{res}(k,l) \mid \text{SPARSE}(l)\right] \le \frac{F_1^{res}(2^{l-1}k)}{2^l} \enspace .$$

Suppose u_l is the frequency of the item with rank $k+1$ in $\mathcal{S}_l$. Applying Hoeffding's bound to the sum of non-negative random variables in (3), each upper bounded by u_l, we have,

$$\mathsf{Pr}\left\{F_1^{res}(k,l) > 2\mathsf{E}\left[F_1^{res}(k,l)\right] \mid \text{SPARSE}(l)\right\} < e^{-\mathsf{E}\left[F_1^{res}(k,l)\right]/(3u_l)}$$

or,

$$\mathsf{Pr}\left\{F_1^{res}(k,l) < \frac{F_1^{res}(2^{l-1}k}{2^{l-1}} \mid \text{SPARSE}(l)\right\} < e^{-F_1^{res}(2^{l-1}k)/(3\cdot 2^l\cdot u_l)} \quad . \tag{4}$$

Assuming the event $\text{SPARSE}(l)$, it follows that $u_l \le f_{rank(2^{l-1}k+1)}$.

$$u_l \le \frac{1}{2^{l-2}k} \sum_{2^{l-2}k+1 \le rank(i) \le 2^{l-1}(k)} |f_{rank(i)}| \le \frac{F_1^{res}(2^{l-2}k)}{2^{l-2}k}$$

or,

$$\frac{F_1^{res}(2^{l-1}k)}{u_l} \le 2^{l-2}k \quad .$$

Substituting in (4) and taking the probability of the complement event, we have,

$$\mathsf{Pr}\left\{F_1^{res}(k,l) < \frac{F_1^{res}(2^{l-1}k)}{2^{l-1}} \mid \text{SPARSE}(l)\right\} > 1 - e^{-k/6} \quad .$$

Since, $\mathsf{Pr}\left\{\text{SPARSE}(l)\right\} > 1 - e^{-k/6}$,

$$\begin{aligned}\mathsf{Pr}\left\{F_1^{res}(k,l) < \frac{F_1^{res}(2^{l-1}k)}{2^{l-1}}\right\} &> (1-e^{-k/6}) \cdot \mathsf{Pr}\left\{\text{SPARSE}(l)\right\} \\ &= (1-e^{-k/6})^2 > 1 - 2e^{-k/6} \quad .\end{aligned}$$

Corollary 1. *For* $l \ge 1$, $F_2^{res}(k,l) \le \frac{F_2^{res}(2^{l-1}k)}{2^{l-1}}$ *with probability* $\ge 1 - 2^{-\frac{k}{6}}$.

Proof. Apply Lemma 4 to the frequency vector obtained by replacing f_i by f_i^2, for $i \in [1,n]$.

3.1 Group definitions

Recall that at each level l, the sampled stream $\mathcal{S}_l$ is provided as input to a data structure $\mathcal{D}_l$, that when queried, returns an estimate $\hat{f}_{i,l}$ for any $i \in [1,n]$ satisfying

$$|\hat{f}_{i,l} - f_i| \le \Delta_l, \quad \text{with prob. } 1 - 2^{-t} \quad .$$

Here, t is a parameter that will be fixed in the analysis and the additive error Δ_l is a function of the algorithm used by $\mathcal{D}_l$ (e.g., $\Delta_l = F_1^{res}(k)/(2^{l-1}k)$ for COUNT-MIN sketches and

$\Delta_l = F_2^{res}(k)/(2^{l-1}k)$ for COUNTSKETCH). Fix a parameter $\bar{\epsilon}$ which will be closely related to the given accuracy parameter ϵ, and is chosen depending on the problem. For example, in order to estimate F_p, $\bar{\epsilon}$ is set to $\frac{\epsilon}{4p}$. Therefore,

$$\hat{f}_{i,l} \in (1 \pm \bar{\epsilon})f_i, \quad \text{provided, } f_i > \frac{\Delta_l}{\bar{\epsilon}}, \quad \text{and } i \in \mathcal{S}_l, \text{ with prob. } 1 - 2^{-t} \ .$$

Define the following event

$$\text{GOODEST} \equiv |\hat{f}_{i,l} - f_i| < \Delta_l, \text{ for each } i \in \mathcal{S}_l \text{ and } l \in \{0, 1, \ldots, L\} \ .$$

By union bound,

$$\Pr\{\text{GOODEST}\} \geq 1 - n(L+1)2^{-t} \ . \tag{5}$$

Our analysis will be conditioned on the event GOODEST.
Define a sequence of geometrically decreasing thresholds $T_0, T_1, \ldots, T_L$ as follows.

$$T_l = \frac{T_0}{2^l}, \ l = 1, 2, \ldots, L \text{ and } \frac{1}{2} < T_L \leq 1 \ . \tag{6}$$

In other words, $L = \lceil \log T_0 \rceil$. Note that L and $L_{\max}$ are distinct parameters. $L_{\max}$ is a data structure parameter and is decided prior to the run of the algorithm. L is a dynamic parameter that is dependent on T_0 and is instantiated at the time of inference. In the next paragraph, we discuss how T_0 is chosen. The threshold values T_l's are used to partition the elements of the stream into groups $G_0, \ldots, G_L$ as follows.

$$G_0 = \{i \in \mathcal{S} : \ |f_i| \geq T_0\} \quad \text{and} \quad G_l = \{i \in \mathcal{S} : \ T_l < |f_i| \leq T_{l-1}\}, \ l = 1, 2, \ldots, L \ .$$

An item i is said to be *discovered as frequent* at level l, provided, i maps to $\mathcal{S}_l$ and $\hat{f}_{i,l} \geq Q_l$, where, Q_l, $l = 0, 1, 2 \ldots, L$, is a parameter family. The values of Q_l are chosen as follows.

$$Q_l = T_l(1 - \bar{\epsilon}) \tag{7}$$

The space parameter $k(l)$ is chosen at level l as follows.

$$\Delta_0 = \Delta_0(k) \leq \bar{\epsilon}Q_0, \qquad \Delta_0 = \Delta_l(2k) \leq \bar{\epsilon}Q_l, l = 1, 2, \ldots, L \ . \tag{8}$$

The choice of T_0. The value of T_0 is a critical parameter for the HSS parameter and its precis choice depends on the problem that is being solved. For example, for estimating F_p, T_0 is chosen as $\frac{1}{\bar{\epsilon}(1-\bar{\epsilon})}\left(\frac{\hat{F}_2}{k}\right)^{1/2}$. For estimating the entropy H, it is sufficient to choose T_0 as

$\frac{1}{\bar{\epsilon}(1-\bar{\epsilon})}\frac{\hat{F}_1^{res}(k')}{k}$, where, k' and k are parameters of the estimation algorithm. T_0 must be chosen as small as possible subject to the following property: $\Delta_l \le \bar{\epsilon}(1-\bar{\epsilon})\frac{T_0}{2^l}$. Lemma 4 and Corollary 1 show that for the COUNT-MIN structure and the COUNTSKETCH structure, T_0 can be chosen to be as small as $\frac{F_1^{res}(k')}{k}$ and $\left(\frac{F_2^{res}k}{k}\right)^{1/2}$, respectively. Since, neither $F_1^{res}(k)$ nor $F_2^{res}(k)$ can be exactly computed in sub-linear space, therefore, the algorithms of Lemmas 1 and Lemma 2 are used to obtain $\frac{1}{2}$-approximations[5] to the corresponding quantities. By replacing k by $2k$ at each level, it suffices to define T_0 as $\frac{\hat{F}_1^{res}(k)}{2k}$ or as $\left(\frac{\hat{F}_2^{res}k}{2k}\right)^{1/2}$, respectively.

3.2 Hierarchical samples

Items are sampled and placed into sampled groups $\bar{G}_0, \bar{G}_1, \ldots, \bar{G}_L$ as follows. The estimated frequency of an item i is defined as

$$\hat{f}_i = \hat{f}_{i,r}, \text{ where, } r \text{ is the lowest level such that } \hat{f}_{i,r} > Q_r \quad .$$

The sampled groups are defined as follows.

$$\bar{G}_0 = \{i : |\hat{f}_i| \ge T_0\} \text{ and } \bar{G}_l = \{i : T_{l-1} < |\hat{f}_i| \le T_l \text{ and } i \in \mathcal{S}_l\}, 1 \le l \le L \quad .$$

The choices of the parameter settings satisfy the following properties. We use the following standard notation. For $a, b \in \mathbb{R}$ and $a < b$, (a, b) denotes the open interval defined by the set of points between a and b (end points not included), $[a, b]$ represents the closed interval of points between a and b (both included) and finally, and $[a, b)$ and $(a, b]$ respectively, represent the two half-open intervals. Partition a frequency group G_l, for $1 \le l \le L-1$, into three adjacent sub-regions:

$$\begin{aligned} \text{lmargin}(G_l) &= [T_l, T_l + \bar{\epsilon}Q_l], \quad l = 0, 1, \ldots, L-1 \text{ and is undefined for } l = L. \\ \text{rmargin}(G_l) &= [Q_{l-1} - \bar{\epsilon}Q_{l-1}, T_{l-1}), \quad l = 1, 2, \ldots, L \text{ and is undefined for } l = 0. \\ \text{mid}(G_l) &= (T_l + \bar{\epsilon}Q_l, Q_{l-1} - \bar{\epsilon}Q_l), \quad 1 \le l \le L-1 \end{aligned}$$

These regions respectively denote the *lmargin* (left-margin), *rmargin* (right-margin) and *middle-region* of the group G_l. An item i is said to belong to one of these regions if its true frequency lies in that region. The middle-region of groups G_0 and G_l is extended to include the right and left margins, respectively. That is,

$$\begin{aligned} \text{lmargin}(G_0) &= [T_0, T_0 + \bar{\epsilon}Q_0) \text{ and } \text{mid}(G_0) = [T_0 + \bar{\epsilon}Q_0, F_1] \\ \text{rmargin}(G_L) &= (Q_{L-1} - \bar{\epsilon}Q_{L-1}, T_{L-1}) \text{ and } \text{mid}(G_0) = (0, Q_{L-1} - \bar{\epsilon}Q_{L-1}] \quad . \end{aligned}$$

[5] More accurate estimates of F_2^{res} and F_1^{res} can be obtained using Lemmas 1 and Lemma 2, but in our applications, a constant factor accuracy suffices.

Important Convention. For clarity of presentation, from now on, the description of the algorithm and the analysis throughout uses the frequencies f_i instead of $|f_i|$. However, the analysis remains unchanged if the frequencies are negative and $|f_i|$ is used in terms of f_i. The only reason for making this notational convenience is to avoid writing $|\cdot|$ in many places. An equivalent way of viewing this is to assume that the actual frequencies are given by an n-dimensional vector g. The vector f is defined as the absolute value of g, taken coordinate wise, (i.e., $f_i = |g_i|$ for all i). It is important to note that the HSS technique is only designed to work with functions of the form $\sum_{i=1}^{n} \psi(|g_i|)$. All results in this paper and their analysis, hold for general update data streams, where, item frequencies could be positive, negative or zero.

We would now like to show that the following properties hold, with probability $1 - 2^{-t}$ each.

1. Items belonging to the middle region of any G_l may be discovered as frequent, that is, $\hat{f}_{i,r} \geq Q_r$, only at a level $r \geq l$. Further, $\hat{f}_i = \hat{f}_{i,l}$, that is, the estimate of its frequency is obtained from level l. These items are never misclassified, that is, if i belongs to some sampled group $\bar{G}_r$, then, $r = l$.
2. Items belonging to the right region of G_l may be discovered as frequent at level $r \geq l-1$, but not at levels less than $l - 1$, for $l \geq 1$. Such items may be misclassified, but only to the extent that i may be placed in either $\bar{G}_{l-1}$ or $\bar{G}_l$.
3. Similarly, items belonging to the left-region of G_l may be discovered as frequent only at levels l or higher. Such items may be misclassified, but only to the extent that i is placed either in $\bar{G}_l$ or in $\bar{G}_{l+1}$.

Lemma 5 states the properties formally.

Lemma 5. *Let $\bar{\epsilon} \leq \frac{1}{6}$. The following properties hold conditional on the event* GOODEST.

1. *Suppose $i \in$ mid(G_l). Then, i is classified into $\bar{G}_l$ iff $i \in \mathcal{S}_l$ and $\hat{f}_i = \hat{f}_{i,l}$. If $i \notin \mathcal{S}_l$, then, $\hat{f}_i$ is undefined and i is unclassified.*
2. *Suppose $i \in$ lmargin(G_l), for some $l \in \{0, 1, \ldots, L-1\}$. If $i \notin \mathcal{S}_l$, then, i is not classified into any group. Suppose $i \in \mathcal{S}_l$. Then, (1) i is classified into $\bar{G}_l$ iff $i \in \mathcal{S}_l$ and $\hat{f}_{i,l} \geq T_l$, and, (2) i is classified into $\bar{G}_{l+1}$ iff $i \in \mathcal{S}_{l+1}$ $\hat{f}_{i,l} < T_l$. In both cases, $\hat{f}_i = \hat{f}_{i,l}$.*
3. *Suppose $i \in$ rmargin(G_l) for some some $l \in \{1, 2, \ldots, L\}$. If $i \notin \mathcal{S}_{l-1}$, then, $\hat{f}_i$ is undefined and i is unclassified. Suppose $i \in \mathcal{S}_{l-1}$. Then,*
   - *(a) i is classified into $\bar{G}_{l-1}$ iff (1) $\hat{f}_{i,l-1} \geq T_{l-1}$, or, (2) $\hat{f}_{i,l-1} < Q_l$ and $i \in \mathcal{S}_l$ and $\hat{f}_{i,l} \geq T_{l-1}$. In case (1), $\hat{f}_i = \hat{f}_{i,l-1}$ and in case (2), $\hat{f}_i = \hat{f}_{i,l}$.*
   - *(b) i is classified into $\bar{G}_l$ iff $i \in \mathcal{S}_l$ and either (1) $\hat{f}_{i,l-1} \geq Q_{l-1}$ and $\hat{f}_i < T_{l-1}$, or, (2) $\hat{f}_{i,l-1} < Q_{l-1}$ and $\hat{f}_i = \hat{f}_{i,l}$. In case (1), $\hat{f}_i = \hat{f}_{i,l-1}$ and in case (2) $\hat{f}_i = \hat{f}_{i,l}$.*

Proof. We prove the statements in sequence. Assume that the event GOODEST holds.

Let $i \in \text{mid}(G_l)$. If $l = 0$ then the statement is obviously true, so we consider $l \geq 1$. Suppose $i \in \mathcal{S}_r$, for some $r < l$, and i is discovered as frequent at level r, that is, $\hat{f}_{i,r} \geq Q_r$. Since, GOODEST holds, therefore, $f_i \geq Q_r - \Delta_r$. Since, $i \in \text{mid}(G_l)$, $f_i < Q_{l-1} - \bar{\epsilon}Q_{l-1}$. Combining, we have

$$Q_{l-1} - \bar{\epsilon}Q_{l-1} > f_i \geq Q_r - \Delta_r = Q_r - \bar{\epsilon}Q_r$$

which is a contradiction for $r \leq l-1$. Therefore, i is not discovered as frequent in any level $r < l$. Hence, if $i \notin \mathcal{S}_l$, i remains unclassified. Now suppose that $i \in \mathcal{S}_l$. Since, GOODEST holds, $\hat{f}_{i,l} \leq f_i + \Delta_l$. Since, $i \in \text{mid}(G_l)$, $f_i < Q_{l-1} - \bar{\epsilon}Q_{l-1}$. Therefore,

$$\hat{f}_{i,l} < Q_{l-1} - \bar{\epsilon}Q_{l-1} + \Delta_l < Q_{l-1} \tag{9}$$

since, $\Delta_l = \bar{\epsilon}Q_l = \bar{\epsilon}Q_{l-1}/2$. Further, $i \in \text{mid}(G_l)$ implies that $f_i > T_l + \bar{\epsilon}Q_l$. Since, GOODEST holds,

$$\hat{f}_{i,l} \geq f_i - \Delta_l > T_l + \bar{\epsilon}Q_l - \Delta_l = T_l \tag{10}$$

since, $\Delta_l = \bar{\epsilon}Q_l$. Combining (9) and (10), we have,

$$T_l < \hat{f}_{i,l} < Q_{l-1} < T_{l-1} \quad .$$

Thus, i is classified into $\bar{G}_l$.

We now consider statement (2) of the lemma. Assume that GOODEST holds. Suppose $i \in \text{lmargin}(G_l)$, for $l \in \{0, 1, \ldots, L-1\}$. Then, $T_l \leq f_i \leq T_l + \bar{\epsilon}Q_l = T_l + \Delta_l$. Suppose $r < l$. We first show that if $i \in \mathcal{S}_r$, then, i cannot be discovered as frequent at level r, that is, $\hat{f}_{i,r} < Q_r$. Assume to the contrary that $\hat{f}_{i,r} \geq Q_r$. Since, GOODEST holds, we have, $\hat{f}_{i,r} \leq f_i + \Delta_r$. Further, $\Delta_r = \bar{\epsilon}Q_r$ and $Q_r = (1-\bar{\epsilon})T_r$. Therefore,

$$T_r(1-\bar{\epsilon})^2 = Q_r - \Delta_r \leq f_i \leq T_l + \Delta_l = T_1(1 + \bar{\epsilon}(1-\bar{\epsilon})) \quad .$$

Since, $T_r = T_l \cdot 2^{l-r}$,

$$2^{l-r} \leq \frac{1 + \bar{\epsilon}(1-\bar{\epsilon})}{(1-\bar{\epsilon})^2} < 2, \quad \text{if } \bar{\epsilon} \leq \frac{1}{6} \quad .$$

This is a contradiction, if $l > r$. We conclude that i is not discovered as frequent at level $r < l$. Therefore, if $i \notin \mathcal{S}_l$, then, i is not classified into any of the $\bar{G}_r$'s. Now suppose that

$i \in \mathcal{S}_l$. We first show that i is discovered as frequent at level l. Since, $i \in \text{lmargin}(G_l)$, therefore, $f_i \geq T_l$ and hence,

$$\hat{f}_{i,l} > T_l - \Delta_l = \frac{Q_l}{1-\bar{\epsilon}} - \bar{\epsilon} Q_l > Q_l \ . \tag{11}$$

Thus, i is discovered as frequent at level l. There are two cases, namely, either $\hat{f}_{i,l} \geq T_l$ or $\hat{f}_{i,l} < T_l$. In the former case, i is classified into $\bar{G}_l$ and $\hat{f}_i = \hat{f}_i$. In the latter case, $\hat{f}_{i,l} < T_l$, the decision regarding the classification is made at the next level $l+1$. If $i \notin \mathcal{S}_{l+1}$, then, i remains unclassified. Otherwise, suppose $i \in \mathcal{S}_{l+1}$. The estimate $\hat{f}_{i,l+1}$ is ignored in favor of a lower level estimate $\hat{f}_{i,l}$, which is deemed to be accurate, since it is at least Q_l. By (11), $\hat{f}_{i,l} > Q_l > T_{l+1}$. By assumption, $\hat{f}_{i,l} < T_l$. Therefore, i is classified into $\bar{G}_{l+1}$. This proves statement (2) of the lemma.

Statement (3) is proved in a similar fashion. □

Estimator. The sample is used to compute the estimate $\hat{\Psi}$. We also define an idealized estimator $\bar{\Psi}$ that assumes that the frequent items structure is an oracle that does not make errors.

$$\hat{\Psi} = \sum_{l=0}^{L} \sum_{i \in \bar{G}_l} \psi(\hat{f}_i) \cdot 2^l \qquad \qquad \bar{\Psi} = \sum_{l=0}^{L} \sum_{i \in \bar{G}_l} \psi(f_i) \cdot 2^l \tag{12}$$

3.3 Analysis

For $i \in [1,n]$ and $r \in [0,L]$, define the indicator variable $x_{i,r}$ as follows.

$$x_{i,r} = \begin{cases} 1 & \text{if } i \in \mathcal{S}_r \text{ and } i \in \bar{G}_r \\ 0 & \text{otherwise.} \end{cases}$$

In this notation, equation (12) can be written as follows.

$$\hat{\Psi} = \sum_{i \in [1,n]} \psi(\hat{f}_i) \sum_{r=0}^{L} x_{i,r} \cdot 2^r \qquad \qquad \bar{\Psi} = \sum_{i \in [1,n]} \psi(f_i) \sum_{r=0}^{L} x_{i,r} \cdot 2^r \ . \tag{13}$$

Note that for a fixed i, the family of variables $x_{i,r}$'s is not independent, since, each item belongs to at most one sampled group $\bar{G}_r$ (i.e., $\sum_{r=0}^{L} x_{i,r}$ is either 0 or 1). We now prove a basic property of the sampling procedure.

Lemma 6. *Let $i \in G_l$.*

1. *If $i \in$ mid(G_l), then,*
   (a) $\mathsf{Pr}\left\{x_{i,l} = 1 \mid \textsc{GoodEst}\right\} = \frac{1}{2^l}$*, and, (b)* $\mathsf{Pr}\left\{x_{i,r} = 1 \mid \textsc{GoodEst}\right\} = 0$ *for* $l \neq r$*.*
2. *If $0 \le l \le L-1$ and $i \in$ lmargin(G_l), then,*
   (a) $\mathsf{Pr}\left\{x_{i,l} = 1 \mid \textsc{GoodEst}\right\} \cdot 2^l + \mathsf{Pr}\left\{x_{l+1} = 1 \mid \textsc{GoodEst}\right\} \cdot 2^{l+1} = 1$*, and,*
   (b) $\mathsf{Pr}\left\{x_{i,r} = 1\right\} = 0$*, for* $r \in \{0, \ldots, L\} - \{l, l+1\}$*.*
3. *If $1 \le l \le L$ and $i \in$ rmargin(G_l), then,*
   (a) $\mathsf{Pr}\left\{x_{i,l} = 1 \mid \textsc{GoodEst}\right\} \cdot 2^l + \mathsf{Pr}\left\{x_{i,l-1} = 1 \mid \textsc{GoodEst}\right\} \cdot 2^{l-1} = 1$*, and,*
   (b) $\mathsf{Pr}\left\{x_{i,r} = 1 \mid \textsc{GoodEst}\right\} = 0$*, for* $r \in \{0, \ldots, L\} - \{l-1, l\}$*.*

Proof. We first note that the part (b) of each of the three statements of the lemma is a restatement of parts of Lemma 5. For example, suppose $i \in \text{lmargin}(G_l)$, $l < L$ and assume that the event $\textsc{GoodEst}$ holds. By Lemma 5, part (2), either $i \in \bar{G}_l$ or $i \in \bar{G}_{l+1}$. Thus, $x_{i,r} = 0$, if r is neither l nor $l+1$. In a similar manner, parts (b) of the other statements of the Lemma can be seen as a restatement of parts of Lemma 5. We now consider part (a) of the statements. Assume that the event $\textsc{GoodEst}$ holds.

Suppose $i \in \text{mid}(G_l)$. The probability that i is sampled into $\mathcal{S}_l = \frac{1}{2^l}$, by construction of the sampling technique. By Lemma 5, part (1), if $i \in \mathcal{S}_l$, then, $i \in \bar{G}_l$. Therefore, $\mathsf{Pr}\left\{x_{i,l} = 1 \mid \textsc{GoodEst}\right\} = \frac{1}{2^l}$.

Suppose $i \in \text{lmargin}(G_l)$ and $l < L$. Then, $\mathsf{Pr}\left\{i \in \mathcal{S}_l\right\} = \frac{1}{2^l}$ and $\mathsf{Pr}\left\{i \in \mathcal{S}_{l+1} \mid i \in \mathcal{S}_l\right\} = \frac{1}{2}$. By Lemma 5, part (2), (a) $i \in \bar{G}_l$, or, $x_{i,l} = 1$ iff $i \in \mathcal{S}_l$ and $\hat{f}_{i,l} \ge T_l$ and, (b) $i \in \bar{G}_{l+1}$, or, $x_{i,l+1} = 1$ iff $i \in \mathcal{S}_{l+1}$ and $\hat{f}_{i,l} < T_l$. Therefore,

$$\begin{aligned}
&\mathsf{Pr}\left\{x_{i,l+1} = 1 \mid i \in \mathcal{S}_l \text{ and } \textsc{GoodEst}\right\}\\
&= \mathsf{Pr}\left\{i \in \mathcal{S}_{l+1} \text{ and } \hat{f}_{i,l} < T_l \mid i \in \mathcal{S}_l \text{ and } \textsc{GoodEst}\right\}\\
&= \mathsf{Pr}\left\{\hat{f}_{i,l} < T_l \mid i \in \mathcal{S}_l \text{ and } \textsc{GoodEst}\right\} \cdot \mathsf{Pr}\left\{i \in \mathcal{S}_{l+1} \mid i \in \mathcal{S}_l \text{ and } \textsc{GoodEst} \text{ and } \hat{f}_{i,l} < T_l\right\}
\end{aligned} \tag{14}$$

We know that $\mathsf{Pr}\left\{\hat{f}_{i,l} < T_l \mid i \in \mathcal{S}_l \text{ and } \textsc{GoodEst}\right\} = 1 - \mathsf{Pr}\left\{x_{i,l} = 1 \mid i \in \mathcal{S}_l \text{ and } \textsc{GoodEst}\right\}$. The specific value of the estimate $\hat{f}_{i,l}$ is a function solely of the random bits employed by $\mathcal{D}_l$ and the sub-stream $\mathcal{S}_l$. By full-independence of the hash function mapping items to the levels, we have that

$$\mathsf{Pr}\left\{i \in \mathcal{S}_{l+1} \mid i \in \mathcal{S}_l \text{ and } \textsc{GoodEst} \text{ and } \hat{f}_{i,l} < T_l\right\} = \mathsf{Pr}\left\{\in \mathcal{S}_{l+1} \mid i \in \mathcal{S}_l\right\} = \frac{1}{2} \ .$$

Substituting in (14), we have,

$$\mathsf{Pr}\left\{x_{i,l+1} = 1 \mid i \in \mathcal{S}_l \text{ and } \textsc{GoodEst}\right\} = \frac{1}{2}\left(1 - \mathsf{Pr}\left\{x_{i,l} = 1 \mid i \in \mathcal{S}_l \text{ and } \textsc{GoodEst}\right\}\right) \ .$$

By definition of conditional probabilities (and multiplying by 2),

$$\frac{2\mathsf{Pr}\left\{x_{i,l+1}=1 \mid \textsc{GoodEst}\right\}}{\mathsf{Pr}\left\{i \in \mathcal{S}_l\right\}} = 1 - \frac{\mathsf{Pr}\left\{x_{i,l}=1 \mid \textsc{GoodEst}\right\}}{\mathsf{Pr}\left\{i \in \mathcal{S}_l\right\}} \quad .$$

Since, $\mathsf{Pr}\left\{i \in \mathcal{S}_l\right\} = \frac{1}{2^l}$, we obtain,

$$2^{l+1}\mathsf{Pr}\left\{x_{i,l+1}=1 \mid \textsc{GoodEst}\right\} = 1 - 2^l\mathsf{Pr}\left\{x_{i,l}=1 \mid \textsc{GoodEst}\right\}$$

or,

$$\mathsf{Pr}\left\{x_{i,l}=1 \mid \textsc{GoodEst}\right\} \cdot 2^l + \mathsf{Pr}\left\{x_{i,l+1}=1 \mid \textsc{GoodEst}\right\} \cdot 2^{l+1} = 1 \quad .$$

This proves the statement (2) of the lemma. Statement (3) regarding the right-margin of G_l can be proved analogously. □

A useful corollary of Lemma 6 is the following.

Lemma 7. *For* $i \in [1,n]$, $\sum_{r=0}^{L} \mathsf{E}\left[x_{i,r} \mid \textsc{GoodEst}\right] \cdot 2^r = 1$.

Proof. If $i \in \text{mid}(G_l)$, then, by Lemma 6, $\mathsf{Pr}\left\{x_{i,l}=1 \mid \textsc{GoodEst}\right\} = \frac{1}{2^l}$ and $\mathsf{Pr}\left\{x_{i,r}=1 \mid \textsc{GoodEst}\right\} = 0$, if $r \neq l$. Therefore,

$$\sum_{r=0}^{L} \mathsf{E}\left[x_{i,r} \mid \textsc{GoodEst}\right] = \sum_{r=0}^{L} \mathsf{Pr}\left\{x_{i,r}=1 \mid \textsc{GoodEst}\right\} \cdot 2^l = \mathsf{Pr}\left\{x_{i,l}=1\right\} \cdot 2^l = 1 \quad .$$

Suppose $i \in \text{lmargin}(G_l)$, for $0 \le l < L$. By Lemma 6, $\mathsf{Pr}\left\{x_{i,l}=1 \mid \textsc{GoodEst}\right\} \cdot 2^l + \mathsf{Pr}\left\{x_{i,l+1}=1 \mid \textsc{GoodEst}\right\} \cdot 2^{l+1} = 1$ and $\mathsf{Pr}\left\{x_{i,r}=1 \mid \textsc{GoodEst}\right\} = 0$, for $r \notin \{l, l+1\}$. Therefore,

$$\sum_{r=0}^{L} \mathsf{E}\left[x_{i,r} \mid \textsc{GoodEst}\right] \cdot 2^r = \sum_{r=0}^{L} \mathsf{Pr}\left\{x_{i,r}=1 \mid \textsc{GoodEst}\right\} \cdot 2^r = 1 \quad .$$

The case for $i \in \text{rmargin}(G_l)$ is proved analogously. □

Lemma 8 shows that the expected value of $\bar{\Psi}$ is Ψ, assuming the event GoodEst holds.

Lemma 8. $\mathsf{E}\left[\bar{\Psi} \mid \textsc{GoodEst}\right] = \Psi$.

Proof. By (13), $\bar{\Psi} = \sum_{i \in [1,n]} \psi(f_i) \sum_{r=0}^{L} x_{i,r} \cdot 2^r$. Taking expectation and using linearity of expectation,

$$\begin{aligned} \mathsf{E}\left[\bar{\Psi} \mid \textsc{GoodEst}\right] &= \sum_{i \in [1,n]} \psi(f_i) \sum_{r=0}^{L} \mathsf{E}\left[x_{i,r} \cdot 2^r \mid \textsc{GoodEst}\right] \\ &= \sum_{i \in [1,n]} \psi(f_i), \quad \text{since, } \sum_{r=0}^{L} \mathsf{E}\left[x_{i,r} \cdot 2^r \mid \textsc{GoodEst}\right] = 1, \text{ by Lemma 7} \\ &= \Psi \quad . \end{aligned}$$

□

The following lemma is useful in the calculation of the variance of $\bar{\Psi}$.

Notation. Let $l(i)$ denote the index of the group G_l such that $i \in G_l$.

Lemma 9. *For $i \in [1, n]$ and $i \notin G_0 - lmargin(G_0)$, $\sum_{r=0}^{L} \mathsf{E}\left[x_{i,r} \cdot 2^{2r} \mid \textsc{GoodEst}\right] \le 2^{l(i)+1}$. If $i \in G_0 - lmargin(G_0)$, then, $\sum_{r=0}^{L} \mathsf{E}\left[x_{i,r} \cdot 2^{2r} \mid \textsc{GoodEst}\right] = 1$.*

Proof. We assume that all probabilities and expectations in this proof are conditioned on the event GOODEST. For brevity, we do not write the conditioning event. Let $i \in \text{mid}(G_l)$ and assume that GOODEST holds. Then, $x_{i,l} = 1$ iff $i \in \mathcal{S}_l$, by Lemma 6. Thus,

$$\mathsf{Pr}\left\{x_{i,l} = 1\right\} \cdot 2^{2l} = \frac{1}{2^l} \cdot 2^{2l} = 2^l \quad .$$

If $i \in \text{lmargin}(G_l)$, then, by the argument in Lemma 7,

$$\mathsf{Pr}\left\{x_{i,l+1} = 1\right\} \cdot 2^{l+1} + \mathsf{Pr}\left\{x_{i,l} = 1\right\} \cdot 2^l = 1 \quad .$$

Multiplying by 2^{l+1},

$$\mathsf{Pr}\left\{x_{i,l+1} = 1\right\} 2^{2(l+1)} + \mathsf{Pr}\left\{x_{i,l} = 1\right\} 2^{2l} \le 2^{l+1} \quad .$$

Similarly, if $i \in \text{rmargin}(G_l)$, then,

$$\mathsf{Pr}\left\{x_{i,l} = 1\right\} 2^l + \mathsf{Pr}\left\{x_{i,l-1} = 1\right\} 2^{l-1} = 1 \quad .$$

Therefore,

$$\mathsf{Pr}\left\{x_{i,l} = 1\right\} 2^{2l} + \mathsf{Pr}\left\{x_{i,l-1} = 1\right\} 2^{2(l-1)} \le 2^{2l}$$

Since, $l(i)$ denotes the index of the group G_l to which i belongs, therefore,

$$\sum_{r=0}^{L} \mathsf{E}\left[x_{i,r} \cdot 2^{2r}\right] < 2^{l(i)+1} \quad .$$

In particular, if $i \in G_0 - \text{lmargin}(G_0)$ or if $i \in \text{mid}(G_l)$, then, the above sum is $2^{l(i)}$. □

Lemma 10.

$$\mathsf{Var}\left[\bar{\Psi} \mid \textsc{GoodEst}\right] \le \sum_{\substack{i \in [1,n] \\ i \notin (G_0 - lmargin(G_0))}} \psi^2(f_i) \cdot 2^{l(i)+1} \quad .$$

Proof. We assume that all expressions for probability and expectations in this proof are conditioned on the event GOODEST. For brevity, it is not written explicitly.

$$\begin{aligned}
\mathsf{E}\big[\bar{\Psi}^2\big] &= \mathsf{E}\big[\big(\sum_{i\in[1,n]}\psi(f_i)\sum_{r=0}^{L}x_{i,r}\cdot 2^r\big)^2\big] \\
&= \mathsf{E}\big[\sum_{i\in[1,n]}\psi^2(f_i)\big(\sum_{r=0}^{L}x_{i,r}\cdot 2^r\big)^2\big] + \mathsf{E}\big[\sum_{i\neq j}\psi(f_i)\cdot\psi(f_j)\sum_{r_1=0}^{L}x_{i,r_1}\cdot 2^{r_1}\sum_{r_2=0}^{L}x_{j,r_2}\cdot 2^{r_2}\big] \\
&= \mathsf{E}\big[\sum_{i\in[1,n]}\psi^2(f_i)\sum_{r=0}^{L}x_{i,r}^2\cdot 2^{2r}\big] + \mathsf{E}\big[\sum_{i\in[1,n]}\psi^2(f_i)\sum_{r_1\neq r_2}x_{i,r_1}\cdot x_{i,r_2}\cdot 2^{r_1+r_2}\big] \\
&\qquad + \mathsf{E}\big[\sum_{i\neq j}\psi(f_i)\cdot\psi(f_j)\sum_{r_1=0}^{L}x_{i,r_1}\cdot 2^{r_1}\sum_{r_2=0}^{L}x_{j,r_2}\cdot 2^{r_2}\big]
\end{aligned}$$

We note that, (a) $x_{i,r}^2 = x_{i,r}$, (b) an item i is classified into a unique group G_r, and therefore, $x_{i,r_1}\cdot x_{i,r_2} = 0$, for $r_1 \neq r_2$, and, (c) for $i \neq j$, x_{i,r_1} and x_{j,r_2} are assumed to be independent of each other, regardless of the values of r_1 and r_2. Thus,

$$\begin{aligned}
\mathsf{E}\big[\bar{\Psi}^2\big] &= \sum_{i\in[1,n]}\mathsf{E}\big[\psi^2(f_i)\sum_{r=0}^{L}x_{i,r}\cdot 2^{2r}\big] + \sum_{i\neq j}\mathsf{E}\big[\psi(f_i)\sum_{r_1=0}^{L}x_{i,r_1}\cdot 2^{r_1}\big]\mathsf{E}\big[\psi(f_j)\sum_{r_2=0}^{L}x_{j,r_2}\cdot 2^{r_2}\big] \\
&= \sum_{i\in[1,n]}\psi^2(f_i)\mathsf{E}\big[\sum_{r=0}^{L}x_{i,r}\cdot 2^{2r}\big] + \Psi^2 - \sum_{i\in[1,n]}\psi^2(f_i)
\end{aligned}$$

since, by Lemma 8, $\mathsf{E}\big[\bar{\Psi}\big] = \Psi = \sum_{i\in[1,n]}\psi(f_i)\sum_{r=0}^{L}\mathsf{E}\big[x_{i,r}\cdot 2^r\big]$. As a result, the expression for $\mathsf{Var}\big[\bar{\Psi}\big]$ simplifies to

$$\begin{aligned}
\mathsf{Var}\big[\bar{\Psi}\big] &= \mathsf{E}\big[\bar{\Psi}^2\big] - (\mathsf{E}\big[\bar{\Psi}\big])^2 = \sum_{i\in[1,n]}\mathsf{E}\big[\psi^2(f_i)\sum_{r=0}^{L}x_{i,r}\cdot 2^{2r}\big] - \sum_{i\in[1,n]}\psi^2(f_i) \\
&\leq \sum_{\substack{i\in[1,n]\\ i\notin G_0-\mathrm{lmargin}(G_0)}}\psi^2(f_i)2^{l(i)+1} + \sum_{i\in G_0-\mathrm{lmargin}(G_0)}\psi^2(f_i) - \sum_{i\in[1,n]}\psi^2(f_i) \\
&\leq \sum_{\substack{i\in[1,n]\\ i\notin G_0-\mathrm{lmargin}(G_0)}}\psi^2(f_i)2^{l(i)+1}, \quad \text{by Lemma 9.} \qquad \square
\end{aligned}$$

For any subset $S \subset [1,n]$, denote by $\psi(S)$ the expression $\sum_{i\in S}\psi(f_i)$. Let $\Psi^2 = \Psi^2(\mathcal{S})$ denote $\sum_{i=1}^{n}\psi^2(|f_i|)$.

Corollary 2. *If the function $\psi(\cdot)$ is non-decreasing in the interval $[0 \ldots T_0 + \varDelta_0]$, then,*

$$\mathsf{Var}\big[\bar{\varPsi} \mid \textsc{GoodEst}\big] = \sum_{l=1}^{L} \psi(T_{l-1})\psi(G_l)2^{l+1} + \psi(T_0 + \varDelta_0)\psi(\mathit{lmargin}(G_0)) \tag{15}$$

Proof. If the monotonicity condition is satisfied, then $\psi(T_{l-1}) \geq \psi(f_i)$ for all $i \in G_l$, $l \geq 1$ and $\psi(f_i) \leq \psi(T_0 + \varDelta_0)$ for $i \in \text{lmargin}(G_0)$. Therefore, $\psi^2(f_i) \leq \psi(T_{l-1}) \cdot \psi(f_i)$, in the first case and $\psi^2(f_i) \leq \psi(T_0 + \varDelta_0)$ in the second case. By Lemma 10,

$$\begin{aligned}\mathsf{Var}\big[\bar{\varPsi} \mid \textsc{GoodEst}\big] &\leq \sum_{l=1}^{L} \sum_{i \in G_l} \psi(T_{l-1})\psi(f_i)2^{l(i)+1} + \sum_{i \in \text{lmargin}(G_0)} \psi(T_0 + \varDelta_0)\psi(f_i) \\ &= \sum_{l=1}^{L} \psi(T_{l-1})\psi(G_l)2^{l+1} + \psi(T_0 + \varDelta_0)\psi(\text{lmargin}(G_0)) \ . \qquad \square\end{aligned}$$

3.4 Error in the estimate

The error incurred by our estimate $\hat{\varPsi}$ is $|\hat{\varPsi} - \varPsi|$, and can be written as the sum of two error components using triangle inequality.

$$|\hat{\varPsi} - \varPsi| \leq |\bar{\varPsi} - \varPsi| + |\hat{\varPsi} - \bar{\varPsi}| = \mathcal{E}_1 + \mathcal{E}_2$$

Here, $\mathcal{E}_1 = |\varPsi - \bar{\varPsi}|$ is the error due to sampling and $\mathcal{E}_2 = |\hat{\varPsi} - \bar{\varPsi}|$ is the error due to the estimation of the frequencies. By Chebychev's inequality

$$\mathsf{Pr}\left\{\mathcal{E}_1 \leq 3(\mathsf{Var}\big[\bar{\varPsi}\big])^{1/2} \mid \textsc{GoodEst}\right\} \geq \frac{8}{9} \ .$$

Substituting the expression for $\mathsf{Var}\big[\bar{\varPsi}\big]$ from (15),

$$\mathsf{Pr}\left\{\mathcal{E}_1 \leq 3\left(\sum_{l=1}^{L} \psi(T_{l-1})\psi(G_l)2^{l+1} + \psi(T_0 + \varDelta_0)\psi(\text{lmargin}(G_0))\right)^{1/2} \middle| \textsc{GoodEst}\right\} \geq \frac{8}{9} \ . \tag{16}$$

We now present an upper bound on $\mathcal{E}_2$. Define a real valued function $\pi : [1, n] \rightarrow \mathbb{R}$ as follows.

$$\pi_i = \begin{cases} \varDelta_l \cdot |\psi'(\xi_i(f_i, \varDelta_l))| & \text{if } i \in G_0 - \text{lmargin}(G_0) \text{ or } i \in \text{mid}(G_l) \\ \varDelta_l \cdot |\psi'(\xi_i(f_i, \varDelta_l))| & \text{if } i \in \text{lmargin}(G_l), \text{ for some } l > 1 \\ \varDelta_{l-1} \cdot |\psi'(\xi_i(f_i, \varDelta_{l-1}))| & \text{if } i \in \text{rmargin}(G_l) \end{cases}$$

where, the notation $\xi_i(f_i, \varDelta_l)$ returns the value of t that maximizes $|\psi'(t)|$ in the interval $[f_i - \varDelta_l, f_i + \varDelta_l]$.

Lemma 11. *Assume that* GOODEST *holds. Then,* $\mathcal{E}_2 \le \sum_{i\in[1,n]} \pi_i \cdot \sum_{r=0}^{L} x_{i,r}2^r$.

Proof. Assume that the event GOODEST holds. By triangle inequality,

$$\mathcal{E}_2 \le \sum_{l=0}^{L}\sum_{i\in\bar{G}_l} |\psi(\hat{f}_i) - \psi(f_i)| x_{i,l}\cdot 2^l = \sum_{i\in[1,n]} |\psi(\hat{f}_i) - \psi(f_i)| \sum_{r=0}^{L} x_{i,r}\cdot 2^r \ .$$

Case 1: $i \in$ $\text{mid}(G_l)$ or $i \in G_0 - \text{lmargin}(G_0)$. Then, i is classified only in group $\bar{G}_l$ with probability $\frac{1}{2^l}$, (or remains unclassified), and $\hat{f}_{i,l} = \hat{f}_i$, by Lemma 6. By Taylor's series

$$|\psi(\hat{f}_i) - \psi(f_i)| \le \Delta_l \cdot |\psi'(\xi_i)|$$

where, $\xi_i = \xi_i(f_i, \Delta_l)$ maximizes $\psi'(t)$ for $t \in [f_i - \Delta_l, f_i + \Delta_l]$.
Case 2: $i \in \text{lmargin}(G_l)$ and $l < L$. Then, $\hat{f}_i = \hat{f}_{i,l}$ or $\hat{f}_i = \hat{f}_{i,l+1}$. Therefore, $|\hat{f}_i - f_i| \le \Delta_l$ and by Taylor's series, $|\psi(\hat{f}_i) - \psi(f_i)| \le \Delta_l |\psi'(\xi_i)|$. Finally, *Case 3:* $i \in \text{rmargin}(G_l)$ and $l > 0$. Then, $i \in \bar{G}_{l-1}$ or $i \in \bar{G}_l$. Similarly, it can be shown that $|\psi(\hat{f}_i) - \psi(f_i)| \le \Delta_{l-1} \cdot |\psi'(\xi_i)|$. Adding,

$$\mathcal{E}_2 \le \sum_{\substack{i\in G_0-\text{lmargin}(G_0)\\ \text{or } i\in\text{mid}(G_l)}} \Delta_l |\psi'(\xi_i(f_i,\Delta_l))| \sum_{r=0}^{L} x_{i,r}2^r + \sum_{l=0}^{L-1}\sum_{i\in\text{lmargin}(G_l)} \Delta_l |\psi'(\xi_i(f_i,\Delta_l))| \sum_{r=0}^{L} x_{i,r}2^r$$
$$+ \sum_{l=1}^{L}\sum_{i\in\text{rmargin}(G_l)} \Delta_{l-1}\cdot |\psi'(\xi_i(f_i,\Delta_{l-1}))| \sum_{r=0}^{L} x_{i,r}2^r2^l)$$

Using the notation of π_i's, we have,

$$\mathcal{E}_2 \le \sum_{i\in[1,n]} \pi_i \cdot \sum_{r=0}^{L} x_{i,r}2^r \ . \qquad \square$$

To abbreviate the statement of the next few lemmas, we introduce the following notation.

$$\Pi_1 = \sum_{i\in[1,n]} \pi_i \tag{17}$$

$$\Pi_2 = 3\Bigg(\sum_{\substack{i\in[1,n]\\ i\notin G_0-\text{lmargin}(G_0)}} \pi_i^2 \cdot 2^{l(i)+1}\Bigg)^{1/2}, \text{ and} \tag{18}$$

$$\Lambda = 3\Bigg(\sum_{l=1}^{L} \psi(T_{l-1})\psi(G_l)2^{l+1} + \psi(T_0+\Delta_0)\psi(\text{lmargin}(G_0))\Bigg)^{1/2} \tag{19}$$

Lemma 12.

$$\mathsf{E}\big[\mathcal{E}_2 \mid \textsc{GoodEst}\big] \leq \Pi_1, \textit{ and } \mathsf{Var}\big[\mathcal{E}_2 \mid \textsc{GoodEst}\big] \leq \frac{\Pi_2^2}{9} \enspace .$$

Therefore, $\mathsf{Pr}\left\{\mathcal{E}_2 \leq \Pi_1 + \Pi_2 \mid \textsc{GoodEst}\right\} \geq \frac{8}{9}$.

Proof. Assume that GOODEST holds and define $\mathcal{E}_2' = \sum_{i \in [1,n]} \pi_i \sum_{l=0}^{L} x_{i,l} 2^l$. By Lemma 11, $\mathcal{E}_2' \leq \mathcal{E}_2$. Applying Lemmas 8 and 10 to $\mathcal{E}_2'$ gives

$$\mathsf{E}\big[\mathcal{E}_2' \mid \textsc{GoodEst}\big] \leq \Pi_1, \text{ and } \mathsf{Var}\big[\mathcal{E}_2' \mid \textsc{GoodEst}\big] \leq \frac{\Pi_2^2}{9} \enspace .$$

By Chebychev's inequality, $\mathsf{Pr}\left\{\mathcal{E}_2' \leq \Pi_1 + \Pi_2 \mid \textsc{GoodEst}\right\} \geq \frac{8}{9}$. Thus,

$$\mathsf{Pr}\left\{\mathcal{E}_2 \leq \Pi_1 + \Pi_2 \mid \textsc{GoodEst}\right\} \geq \frac{8}{9} \enspace .$$

□

Lemma 13 presents the overall expression of error and its probability.

Lemma 13. *Let* $\bar{\epsilon} \leq \frac{1}{3}$. *Suppose* $\psi()$ *is a monotonic function in the interval* $[0, T_0 + \Delta_0]$.

$$\mathsf{Pr}\left\{|\hat{\Psi} - \bar{\Psi}| \leq \Lambda + \Pi_1 + \Pi_2\right\} > \frac{7}{9}(1 - (n(L+1))2^{-t}) \enspace .$$

Proof. Combining Lemma 12 and equation (16), and using the notation of equations (17), (18) and (19), we have,

$$\mathsf{Pr}\left\{|\hat{\Psi} - \bar{\Psi}| \leq \Lambda + \Pi_1 + \Pi_2 \mid \textsc{GoodEst}\right\} \geq 1 - \frac{1}{9} - \frac{1}{9} = \frac{7}{9}$$

Since, $\mathsf{Pr}\left\{\textsc{GoodEst}\right\} > 1 - (n(L+1))2^{-t}$, therefore,

$$\begin{aligned}
&\mathsf{Pr}\left\{|\hat{\Psi} - \bar{\Psi}| \leq \Lambda + \Pi_1 + \Pi_2\right\} \\
&= \mathsf{Pr}\left\{|\hat{\Psi} - \bar{\Psi}| \leq \Lambda + \Pi_1 + \Pi_2 \mid \textsc{GoodEst}\right\} \mathsf{Pr}\left\{\textsc{GoodEst}\right\} \\
&\geq \frac{7}{9}(1 - (n(L+1))2^{-t}).
\end{aligned}$$

□

Reducing Randomness by using Pseudo-random Generator. The analysis has assumed that the hash function mapping items to levels is completely independent. A pseudo-random generator can be constructed along the lines of Indyk in [17] and Indyk and Woodruff in [18], to reduce the required randomness. This is illustrated for each of the two estimations that we consider in the following sections, namely, estimating F_p and estimating H.

4 Estimating F_p for $p \geq 2$

In this section, we apply the HSS technique to estimate F_p for $p \geq 2$. Let $\bar{\epsilon} = \frac{\epsilon}{4p}$. For estimating F_p for $p \geq 2$, we use the HSS technique instantiated with the COUNTSKETCH data structure that uses space parameter k. The value of k will be fixed during the analysis. We use a standard estimator such as sketches [1] or FASTAMS [21] for estimating $\hat{F}_2$ to within accuracy of $1 \pm \frac{1}{4}$ and with confidence $\frac{26}{27}$ using space $O(\frac{\log F_1}{\epsilon^2})$. We extend the event GOODEST to include this event, that is,

$$\text{GOODEST} \cong |\hat{F}_2 - F_2| < \frac{F_2}{4} \text{ and } |\hat{f}_{i,l} - f_i| \leq \Delta_l, \forall i \in [1,n] \text{ and } l \in \{0, 1, \ldots, L\} \enspace .$$

The width of the COUNTSKETCH structure at each level is chosen to be $s = O(\log n + \log\log F_1)$ so that $\Pr\{\text{GOODEST}\} \geq \frac{26}{27}$.

Let F_0 denote the number of distinct items in the data stream, that is, the number of items with non-zero frequency.

Lemma 14. *Let* $\bar{\epsilon} = \min\left(\frac{\epsilon}{4p}, \frac{1}{6}\right)$. *Then,* $\Pi_1(F_p) \leq \epsilon F_p$.

Proof. Let $i \in G_l$ and $l > 0$. Then, $\pi_i \leq \Delta_{l(i)-1} |\psi'(\xi(f_i, \Delta_{l-1}))|$. Since, F_p is monotonic and increasing, we have, $\xi(f_i, t) = f_i + t$, for $t > 0$. For $i \in G_l$ and $l > 0$, $\Delta_{l-1} < 2\bar{\epsilon} f_i$. Further,

$$\psi'(\xi(f_i, \Delta_{l-1}) \leq p(f_i + \Delta_{l-1})^{p-1} \leq p f_i^{p-1}(1 + 2\bar{\epsilon})^{p-1} \leq p f_i^{p-1} e^{2(p-1)/4p} \leq 2p f_i^p \enspace .$$

Thus,

$$\pi_i \leq \Delta_{l-1} \psi'(\xi(f_i, \Delta_{l-1})) \leq 4\bar{\epsilon} f_i p f_i^{p-1} < \epsilon f_i^p, \quad \text{since, } \bar{\epsilon} = \frac{\epsilon}{4p}. \tag{20}$$

For $i \in G_0$, $\Delta_0 < \bar{\epsilon} T_l \leq \bar{\epsilon} f_i$. Therefore,

$$\pi_i \leq \Delta_0 \psi'(\xi(f_i, \Delta_0)) \leq 2\bar{\epsilon} f_i p (f_i + \Delta_0)^{p-1} \leq \epsilon f_i^p \enspace . \tag{21}$$

Combining this with (20), we have,

$$\Pi_1(F_p) \leq \sum_{i \in [1,n]} \epsilon f_i^p = \epsilon F_p \enspace . \qquad \square$$

Lemma 15. *Let* $\bar{\epsilon} \leq \min\left(\frac{\epsilon}{4p}, \frac{1}{6}\right)$. *Then,*

$$\Pi_2 \leq \epsilon F_p, \textit{ provided, } k \geq \frac{32 \cdot (72)^{2/p} \cdot p^2}{\epsilon^2} n^{1-2/p} \enspace .$$

Proof. $\mathit{\Pi}_2$ is defined as

$$(\mathit{\Pi}_2)^2 = 9 \left(\sum_{i \in [1,n], i \notin G_0 - \text{lmargin}(G_0)} \pi_i^2 \cdot 2^{l(i)+1} \right)$$

Suppose $i \in G_l$ and $l \geq 1$, or, $i \in \text{lmargin}(G_0)$. Then, by (20) and (21), it follows that, $\pi_i \leq \epsilon f_i^p$. Therefore,

$$\frac{(\mathit{\Pi}_2)^2}{9} \leq \left(\sum_{i \in [1,n], i \notin G_0 - \text{lmargin}(G_0)} \epsilon^2 f_i^{2p} \cdot 2^{l(i)+1} \right) \tag{22}$$

For $i \in G_l$ and $l \geq 1$, $f_i < T_{l-1}$ and therefore, $f_i^{2p} \leq T_{l-1}^p f_i^p$.
For $i \in \text{lmargin}(G_0)$, $f_i \leq T_0 + \mathit{\Delta}_0 < T_0(1 + \bar{\epsilon})$ and therefore,

$$f_i^{2p} \leq (T_0 + \mathit{\Delta}_0)^p f_i^p \leq T_0^p (1 + \bar{\epsilon})^p, \text{ for } i \in \text{lmargin}(G_0) \enspace .$$

Combining, we have,

$$\frac{(\mathit{\Pi}_2)^2}{9} \leq \epsilon^2 \sum_{i \in \text{lmargin}(G_0)} T_0^p (1 + \bar{\epsilon})^p f_i^p + \sum_{l=1}^{L} T_{l-1}^p \sum_{i \in G_l} f_i^p 2^{l+1} \enspace . \tag{23}$$

By construction,

$$\mathit{\Delta}_0 \leq \left(\frac{2F_2}{k} \right)^{1/2} = \bar{\epsilon} Q_0 = \bar{\epsilon}(1 - \bar{\epsilon}) T_0 \enspace .$$

and $T_l = \frac{T_0}{2^l}$. Substituting in (23), we have,

$$\frac{(\mathit{\Pi}_2)^2}{9} \leq \frac{\epsilon^2}{(1 - \bar{\epsilon})^p} \left(\frac{2F_2}{\bar{\epsilon}^2 k} \right)^{p/2} \left(\sum_{i \in \text{lmargin}(G_0)} (1 + \bar{\epsilon})^p f_i^p + \sum_{l=1}^{L} \sum_{i \in G_l} f_i^p \cdot \frac{2^{l+1}}{2^{(l-1)(p-1)}} \right) \tag{24}$$

Since, $p \geq 2$, $\frac{2^{l+1}}{2^{(l-1)(p-1)}} \leq 4$. Further, since, $\bar{\epsilon} \leq \frac{1}{4p}$, $(1 + \bar{\epsilon})^p \leq 2$ and $(1 - \bar{\epsilon})^p > \frac{1}{2}$. Therefore, (24) can be simplified as

$$\begin{aligned} \frac{(\mathit{\Pi}_2)^2}{9} &\leq 8\epsilon^2 \left(\frac{2F_2}{\bar{\epsilon}^2 k} \right)^{p/2} \left(\sum_{i \in \text{lmargin}(G_0)} f_i^p + \sum_{l=1}^{L} \sum_{i \in G_l} f_i^p \right) \\ &\leq 8\epsilon^2 \left(\frac{2F_2}{\bar{\epsilon}^2 k} \right)^{(p-1)/2} F_p \end{aligned} \tag{25}$$

since, $\left(\sum_{i \in \text{lmargin}(G_0)} f_i^p + \sum_{l=1}^{L} \sum_{i \in G_l} f_i^p\right) = F_p - \sum_{i \in G_0 - \text{lmargin}(G_0)} f_i^p$. We now use the identity

$$\left(\frac{F_2}{F_0}\right)^{1/2} \leq \left(\frac{F_r}{F_0}\right)^{1/r} \text{ or, } F_2^{r/2} \leq F_r \cdot F_0^{r/2-1}, \text{ for any } r \geq 2 \ . \tag{26}$$

Letting $r = p$, we have $F_2^p < F_p F_0^{p/2-1}$. Substituting in (25), we have,

$$(\mathit{\Pi}_2)^2 \leq \frac{72 \cdot 2^{p/2} \cdot \epsilon^2 \cdot F_p^2 \cdot F_0^{p/2-1}}{(\bar{\epsilon}^2 k)^{p/2}} \cdot F_p \leq 72\epsilon^2 \cdot 2^{p/2} \cdot F_p^2 \left(\frac{n^{1-2/p}}{\bar{\epsilon}^2 k}\right)^{p/2}$$

Letting $k = \frac{2 \cdot (72)^{2/p}}{\bar{\epsilon}^2} n^{1-2/p} = \frac{32 \cdot (72)^{2/p} \cdot p^2}{\epsilon^2} n^{1-2/p}$, we have,

$$(\mathit{\Pi}_2)^2 \leq \epsilon^2 F_p^2, \quad \text{or, } \mathit{\Pi}_2 \leq \epsilon F_p \ . \qquad \square$$

Lemma 16.

$$\textit{If } \bar{\epsilon} \leq \min\left(\frac{\epsilon}{4p}, \frac{1}{6}\right) \textit{ and } k \geq \frac{32 \cdot (72)^{2/p} \cdot p^2}{\epsilon^{2+4/p}} n^{1-2/p}, \textit{ then, } \mathit{\Lambda} < \epsilon F_p \ .$$

Proof. The expression (19) for $\mathit{\Lambda}$ can be written as follows.

$$\frac{\mathit{\Lambda}^2}{9} = \left(\sum_{i \in \text{lmargin}(G_0)} (T_0 + \mathit{\Delta}_0)^p f_i^p + \sum_{l=1}^{L} (T_{l-1})^p \sum_{i \in G_l} f_i^p 2^{l+1}\right) \tag{27}$$

Except for the factor of ϵ^2, the expression on the *RHS* is identical to the expression in the *RHS* (23), for which an upper bound was derived in Lemma 15. Following the same proof and modifying the constants, we obtain that

$$\mathit{\Lambda}^2 \leq \epsilon^2 F_p^2 \text{ if } k \geq \frac{32 \cdot (72)^{2/p} \cdot p^2}{\epsilon^{2+4/p}} n^{1-2/p} \ .$$

Recall that s is the width of the COUNTSKETCH structures kept at each level $l = 0, \ldots, L$.

Lemma 17. *Suppose* $p \geq 2$. *Let* $k \geq \frac{32 \cdot (72)^{2/p} \cdot p^2}{\epsilon^{2+4/p}} n^{1-2/p}$, $\bar{\epsilon} = \min\left(\frac{\epsilon}{4p}, \frac{1}{6}\right)$ *and* $s = O(\log n + \log\log F_1)$ *, then,* $\mathsf{Pr}\left\{|\hat{F}_p - F_p| < 3\epsilon F_p\right\}$ *with probability* $\frac{6}{9}$.

Proof. By Lemma 13,

$$\mathsf{Pr}\left\{|\hat{F}_p - F_p| \leq \mathit{\Pi}_1 + \mathit{\Pi}_2 + \mathit{\Lambda}\right\} \geq (1 - \frac{7}{9})(\mathsf{Pr}\left\{\text{GOODEST}\right\}) \ .$$

By Lemma 14, we have, $\mathit{\Pi}_1 \leq \epsilon F_p$. By Lemma 15, we have, $\mathit{\Pi}_2 \leq \epsilon F_p$. Finally, by Lemma 16, $\mathit{\Lambda} \leq \epsilon F_p$. Therefore,

$$\mathit{\Pi}_1 + \mathit{\Pi}_2 + \mathit{\Lambda} \leq 3\epsilon F_p \ .$$

By choosing the number s of independent tables to be $O(\log n + \log\log F_1)$ in the COUNTSKETCH structure at each level, we have, $\Pr\{\textsc{GoodEst}\} \geq 1 - \frac{1}{27}$. This includes the error probability of estimating $\hat{F}_2$ to within factor of $1 \pm \frac{1}{2}$ of F_2. Thus,

$$\Pr\left\{|\hat{F}_p - F_p| \leq \mathit{\Pi}_1 + \mathit{\Pi}_2 + \mathit{\Lambda}\right\} \geq \left(1 - \frac{2}{9}\right)\left(\Pr\{\textsc{GoodEst}\}\right) > \frac{6}{9} \ . \qquad \square$$

The space requirement for the estimation procedure whose guarantees are presented in Lemma 17 can be calculated as follows. There are $L = O(\log F_1)$ levels, where, each level uses a COUNTSKETCH structure with hash table size size $k = O\left(p^2 \cdot n^{1-2/p} \cdot \frac{1}{\epsilon^{1+2/p}}\right)$ and number of independent hash tables $s = O(\log n + \log\log F_1)$. Since, each table entry requires space $O(\log F_1)$ bits, the space required to store the tables is

$$S = O\left(\frac{p^2}{\epsilon^{2+4/p}} n^{1-\frac{2}{p}} (\log^2 F_1)(\log n + \log\log F_1)\right) \ .$$

The number of random bits can be reduced from $O(n(\log n + \log F_1))$ to $O(S \log R)$ using the pseudo-random generator of [17, 18]. Since the final state of the data structure is the same if the input is presented in the order of item identities, therefore, $R = O(n \cdot L \cdot s)$. Therefore, the number of random bits required is $O(S(\log n + \log\log F_1)$. The expected time required to update the structure corresponding to each stream update of the form (pos, i, v) is $O(s) = O(\log n + \log\log F_1)$. This is summarized in the following theorem.

Theorem 3. *For every $p \geq 2$ and $0 < \epsilon \leq 1$, there exists a randomized algorithm that returns an estimate $\hat{F}_p$ satisfying $\Pr\left\{|\hat{F}_p - F_p| \leq \epsilon F_p\right\} \geq \frac{2}{3}$ using space $O\left(\frac{p^2}{\epsilon^{2+4/p}} \cdot n^{1-2/p} \cdot (\log F_1)^2 \cdot (\log n + \log\log F_1)^2\right)$. The expected time to process each stream update is $O(\log n + \log\log F_1)$.* □

5 Estimating Entropy

In this section, we apply the HSS technique to estimate the entropy

$$H = \sum_{i: f_i \neq 0} \frac{|f_i|}{F_1} \log \frac{F_1}{|f_i|}$$

of a data stream. Let

$$h(x) = x \log \frac{1}{x}, \quad \frac{1}{F_1} \leq |x| \leq 1 \ \ .$$

By convention, we let $h(0) = 0$. In this notation,

$$H = \sum_{i : f_i \neq 0} h\left(\frac{f_i}{F_1}\right) \ \ .$$

In this section, the function $\psi(x) = h(x/F_1)$ and the statistic $\Psi = \sum_{i=1}^{n} h(f_i/F_1) = H$.

The HSS algorithm is instantiated using the COUNT-MIN sketch [10] as the frequent items structure $\mathcal{D}_l$ at each level. We assume that there are $8k$ buckets in each hash table, where, k is a parameter that we fix in the analysis. The parameter $\bar{\epsilon} = \bar{\epsilon}(\epsilon)$ is fixed in the analysis. We use Lemma 4 to estimate $F_1^{res}(k')$ to within factor of $1 \pm \frac{1}{2}$ and with constant probability. The parameter $k' \leq k$ will also be fixed in the analysis. The thresholds $T_0, T_1, \ldots, T_L$ are defined as follows.

$$T_0 = \frac{\hat{F}_1^{res}(k')}{\bar{\epsilon} k}, \quad \text{and } T_l = \frac{T_0}{2^l}, \quad \text{for } l \geq 1.$$

The rest of the parameters are defined in terms of T_l in the manner described in Section 3. Thus, $\Delta_l = \bar{\epsilon}(1 - \bar{\epsilon}) T_l$ and $Q_l = (1 - \bar{\epsilon}) T_l$.

The derivative h' of the function h is

$$h'(x) = \log \frac{1}{ex} \qquad\qquad\qquad \frac{1}{F_1} \leq x \leq 1 \ \ . \tag{28}$$

The function $h'(x)$ is concave in the interval $[F_1^{-1}, 1]$ and attains a unique local maximum at $x = \frac{1}{e}$. The absolute value of the derivative $|h'(\cdot)|$ decreases from $\log F_1 - 1$ to 0 in the interval $[\frac{1}{F_1}, \frac{1}{e}]$ and increases from 0 to 1 in the interval $[\frac{1}{e}, 1]$. We choose the parameters k and k' so that

$$T_0 + \Delta_0 < \frac{F_1}{e^2} \ \ . \tag{29}$$

We assume that $\bar{\epsilon} \leq \frac{1}{2}$ in the rest of the discussion.

Lemma 18. *For $i \in G_l$, $\pi_i \leq \frac{2\Delta_l}{F_1} \log \frac{F_1}{T_l} \leq 2\bar{\epsilon} h\left(\frac{f_i}{F_1}\right)$.*

Proof. Suppose $i \in G_l$ and $l \geq 1$, or $i \in \text{lmargin}(G_0)$. Due to the choice of k, k' as per (29), $|\psi'(t)|$ is maximized at $t = T_l - \Delta_l$. Therefore,

$$|\psi'(t)| \leq \psi'(T_l - \Delta_l) \leq \frac{1}{F_1} \log \frac{F_1}{T_l - \Delta_l} = \frac{1}{F_1}\left(\log \frac{F_1}{T_l} - \log 1 - \bar{\epsilon})\right) \leq \frac{2}{F_1} \log \frac{F_1}{T_l} \ \ .$$

since, $\bar{\epsilon} \leq \frac{1}{2}$, $-\log(1-\bar{\epsilon}) \leq \log 2 = 1$. Therefore, $\pi_i \leq \frac{2\Delta_l}{F_1} \log \frac{F_1}{T_l}$. Further, for $i \in G_l$, $l \geq 1$ or, $i \in \text{lmargin}(G_0)$,

$$\frac{\Delta_l}{F_1} \log \frac{F_1}{T_l} = \frac{\bar{\epsilon} T_l}{F_l} \log \frac{F_l}{T_l} < \bar{\epsilon} h \left(\frac{f_i}{F_1} \right)$$

Now suppose $i \in G_0 - \text{lmargin}(G_0)$. Then, by (29), $|\psi'(t)|$ has a maximum value at $t = T_0$. Therefore, $|\psi'(t)| \leq \frac{1}{F_1} \log \frac{F_1}{T_0}$ as explained above. The argument now proceeds in the same manner as above. □

Define the event GOODRESEST to be $F_1^{res}(k') \leq \hat{F}_1^{res}(k') \leq 2F_1^{res}(k')$. By Lemma 2, GOODRESEST holds with constant probability and can be accomplished using space $O\left(k'(\log F_1) + \log n\right)$. We assume that the event GOODEST is broadened to include the event GOODRESEST as well, such that $\Pr\{\text{GOODEST}\} \geq \frac{26}{27}$.

Lemma 19. $\frac{T_0}{F_1} \leq \frac{2H}{\bar{\epsilon} k (\log k')}$.

Proof. Since, GOODEST holds, $T_0 = \frac{\hat{F}_1^{res}(k')}{\bar{\epsilon} k} \leq \frac{2F_1^{res}(k)}{\bar{\epsilon} k}$. Therefore, if there are at most k distinct items in the stream (i.e., $F_0 \leq k$), then, $T_0 = 0$ and the lemma follows. Otherwise,

$$H \geq \sum_{rank(i) > k'} \frac{|f_i|}{F_1} \log \frac{F_1}{f_i} \geq \sum_{rank(i) > k'} \frac{|f_i|}{F_1} \log k' = \frac{\log k'}{F_1} F_1^{res}(k') \geq \bar{\epsilon} k (\log k') \frac{T_0}{2F_1} \quad . \qquad □$$

Lemma 20. $\Pi_1 \leq 2\bar{\epsilon} H$.

Proof. By definition, $\Pi_1 = \sum_{i \in [1,n]} \pi_i$. By Lemma 18, we have, $\pi_i \leq 2\bar{\epsilon} h \left(\frac{f_i}{F_1} \right)$. Thus,

$$\Pi_1 = \sum_{i \in [1,n]} \pi_i \leq \sum_{i \in [1,n]} 2\bar{\epsilon} h \left(\frac{f_i}{F_1} \right) = 2\bar{\epsilon} H. \qquad □$$

Lemma 21. *Then,*

$$\Pi_2^2 \leq \frac{36\bar{\epsilon}(\log(2F_1))H^2}{k(\log k')} \quad .$$

Proof. By definition of Π_2 (equation (18))

$$\frac{\Pi_2^2}{9} = \sum_{l=1}^{L} \sum_{i \in G_l} \pi_i^2 2^{l+1} + \sum_{i \in \text{lmargin}(G_0)} \pi_i^2 \quad .$$

If $F_0 \leq k'$, then, $F_1^{res}(k') = 0$ and therefore, $T_0 = 0$ and so $\Pi_2 = 0$¿ Therefore, without loss of generality, we may assume that $F_0 > k'$. We first consider the summation over elements in G_l, $l \geq 1$. In this region,

$$|h'(f_i/F_1)| \leq |h'(T_l/F_1)| \leq \frac{1}{F_1} \log \frac{F_1}{eT_l} \leq \frac{1}{F_1} \log F_1$$

and therefore,

$$\pi_i \le \mathit{\Delta}_l |h'(T_l/F_1)| < \frac{\bar{\epsilon} T_l}{F_1} \log F_1 .$$

Further, by Lemma 18, $\pi_i \le 2\bar{\epsilon} h\left(\frac{f_i}{F_1}\right)$. Therefore,

$$\sum_{l=1}^{L} \sum_{i \in G_l} \pi_i^2 2^{l+1} \le \sum_{l=1}^{L} \sum_{i \in G_l} \frac{\bar{\epsilon} T_l}{F_1} (\log F_1) \cdot (2\bar{\epsilon}) h\left(\frac{f_i}{F_1}\right) \cdot 2^{l+1} \le \frac{T_0 (4\bar{\epsilon}^2 \log F_1)}{F_1} \sum_{l=1}^{L} \sum_{i \in G_l} h\left(\frac{f_i}{F_1}\right)$$

In a similar manner,

$$\sum_{i \in lmargin(G_0)} \pi_i^2 \le 2(\bar{\epsilon})^2 \frac{T_0}{F_1} \log F_1 \sum_{i \in \text{lmargin}(G_0)} h(f_i/F_1) \enspace .$$

Adding,

$$\frac{\mathit{\Pi}_2^2}{9} \le 4\bar{\epsilon}^2 \frac{T_0}{F_1} H \le \frac{4\bar{\epsilon} H^2}{k(\log k')}$$

since, by Lemma 19, $\frac{T_0}{F_1} \le \frac{H}{\bar{\epsilon} k (\log k')}$.

Lemma 22. *Then,*

$$\mathit{\Lambda}^2 \le \frac{2(\log(2F_1)) H^2}{\bar{\epsilon} k (\log k')} \enspace .$$

Proof. By equation (29), $h(x)$ is monotonic increasing for $0 \le x \le \frac{T_0 + \mathit{\Delta}_0}{F_1}$. Therefore, from the definition of $\mathit{\Lambda}^2$,

$$\begin{aligned}
\frac{\mathit{\Lambda}^2}{9} &= \sum_{i \in \text{lmargin}(G_0)} h\left(\frac{T_0 + \mathit{\Delta}_0}{F_1}\right) \cdot h\left(\frac{f_i}{F_1}\right) + \sum_{l=1}^{L} \frac{T_{l-1}}{F_1} \log \frac{F_1}{T_{l-1}} \sum_{i \in G_l} h\left(\frac{f_i}{F_1}\right) \cdot 2^{l+1} \\
&\le \sum_{i \in \text{lmargin}(G_0)} \frac{2T_0}{F_1} (\log(2F_1)) h\left(\frac{f_i}{F_1}\right) + \sum_{l=1}^{L} \frac{T_0}{F_1} \log(2F_1) \sum_{i \in G_l} h\left(\frac{f_i}{F_1}\right) \\
&\le \frac{2T_0}{F_1} \log(2F_1) \sum_{i \notin G_0 - \text{lmargin}(G_0)} h\left(\frac{f_i}{F_1}\right) \\
&\le \frac{2T_0 (\log(2F_1)) H}{F_1} \le \frac{2(\log(2F_1)) H^2}{\bar{\epsilon} k (\log k')}
\end{aligned}$$

since, by Lemma 19, $\frac{T_0}{F_1} \le \frac{H}{\bar{\epsilon} k (\log k')}$. □

Lemma 23. *Let* $\epsilon \le \frac{1}{2}$, $k' \ge \frac{1}{\epsilon^2}$, $k \ge \frac{8(\log(2eF_1))}{\epsilon^3 \log(1/\epsilon)}$ *and* $\bar{\epsilon} = \frac{\epsilon}{4}$. *Choose the width of the* COUNT-MIN *structure at each level to be* $s = O(\log n + \log\log F_1)$. *Then,* $\Pr\left\{|\hat{H} - H| \le 3\epsilon H\right\} \ge \frac{2}{3}$.

Proof. By the above choice of k, $k \geq 2$ and therefore, $\log k \geq 1$. Therefore, by Lemma 20, $\Pi_1 \leq 2\bar{\epsilon}H \leq \epsilon H$. By the choice of k', $\log k' \geq \log \frac{1}{\epsilon}$. By Lemma 21,

$$\Pi_2^2 \leq \frac{36\bar{\epsilon}^2(\log(2F_1))H^2}{k(\log k')} \leq \frac{36\epsilon^2}{16}\frac{\epsilon^2\log(1/\epsilon)}{8\log(1/\epsilon)} \leq \epsilon^2 H^2 \ .$$

Therefore, $\Pi_2 \leq \epsilon H$. By Lemma 22,

$$\Lambda^2 \leq \frac{2(\log(2F_1))H^2}{k(\log k')} \leq \epsilon^2 H^2, \text{ or, } \Lambda \leq \epsilon H$$

By Lemma 13,

$$\Pr\left\{|\hat{H} - H| \leq \Pi_1 + \Pi_2 + \Lambda\right\} \geq \frac{7}{9} \cdot \Pr\left\{\textsc{GoodEst}\right\}$$

By choosing the width of each Count-Min structure to be $s = O(\log n + \log\log F_1)$, $\Pr\left\{\textsc{GoodEst}\right\} \geq \frac{26}{27}$.

$$\Pr\left\{|\hat{H} - H| \leq 3\epsilon H\right\} \geq \frac{6}{9} \ . \qquad \square$$

There are $L + 1$ levels, where each level keeps a Count-Min structure with height $k = O\left(\frac{1}{\epsilon^3\log(1/\epsilon)}\right)$ and width $s = O(\log n + \log\log F_1)$. Therefore, the space used by the data structure is $S = O\left(\frac{(\log F_1)^2(\log n + \log\log F_1)}{\epsilon^3\log(1/\epsilon)}\right)$, not counting the random bits required. The random bits can be reduced using the techniques of Indyk [17] that adapts the pseudo-random generator of Nisan [19] for space bounded computations. Using this technique, the number of random bits required becomes $O(S\log R)$, where, S is the space used by the algorithm and R is the running time of the algorithm. Since, the state of the data structure is the same if the input were presented in a sorted order of the item identities, therefore, $R = O(nL\log n)$ and thus, the number of random bits is

$$O(S\log R) = O\left(1/(\epsilon^3\log(1/\epsilon))(\log F_1)^2(\log n + \log\log F_1)^2\right) .$$

Finally, we calculate the total number of bits required to estimate $\hat{F}_1^{res}(k')$, for $k' = \lceil\frac{1}{\epsilon^2}\rceil$, to within relative accuracy of $1 \pm \frac{1}{2}$. By Lemma 4, this can be done using space $O\left(\frac{(\log F_1)(\log n + \log(1/\epsilon))}{\epsilon^2}\right)$. This is dominated by the space requirement of the Hss structure. This completes the proof of the main theorem of this section, stated below.

Theorem 4. *There exists an algorithm that returns an estimate $\hat{H}$ satisfying*

$$\Pr\left\{|\hat{H} - H| \leq \epsilon H\right\} \geq \frac{2}{3} \text{ using space } O\left(\frac{(\log F_1)^2}{\epsilon^3\log(1/\epsilon)}(\log n + \log\log F_1)^2\right) \ .$$

The expected time required to process each stream update is $O(\log n + \log\log F_1)$. □

6 Conclusions

We present Hierarchical Sampling from Sketches (HSS), a technique that can be used for estimating a class of functions over update streams of the form $\Psi(\mathcal{S}) = \sum_{i=1}^{n} \psi(f_i)$ and use it to design nearly space-optimal algorithms for estimating the p^{th} frequency moment F_p, for real $p \geq 2$, and for estimating the entropy of a data stream.

Acknowledgements

The first author thanks Igor Nitto for pointing out an omission in the calculation of the space requirement of the HSS algorithm for estimating entropy.

References

1. 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.
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 ACM Symposium on Theory of Computing*, 2002.
3. Lakshminath Bhuvanagiri and Sumit Ganguly. “Estimating Entropy over Data Streams”. In *Proceedings of the European Symposium on Algorithms*, pages 148–159, 2006.
4. J.L. Carter and M.N. Wegman. “Universal Classes of Hash Functions”. *Journal of Computer Systems and Sciences*, 18(2):143–154, 1979.
5. Amit Chakrabarti, D.K. Ba, and S. Muthukrishnan. “Estimating Entropy and Entropy Norm on Data Streams”. In *Proceedings of the Symposium on Theoretical Aspects of Computer Science*, 2006.
6. Amit Chakrabarti, Graham Cormode, and Andrew McGregor. “A Near-Optimal Algorithm for Computing the Entropy of a Stream”. In *Proceedings of the ACM Symposium on Discrete Algorithms*, 2007.
7. Amit Chakrabarti, Subhash Khot, and Xiaodong Sun. “Near-Optimal Lower Bounds on the Multi-Party Communication Complexity of Set Disjointness”. In *Proceedings of the Conference on Computational Complexity*, 2003.
8. Moses Charikar, Kevin Chen, and Martin Farach-Colton. “Finding frequent items in data streams”. In *Proceedings of the International Colloquium on Automata, Languages and Programming, 2002*, pages 693–703.
9. Don Coppersmith and Ravi Kumar. “An improved data stream algorithm for estimating frequency moments”. In *Proceedings of the ACM Symposium on Discrete Algorithms*, pages 151–156, 2004.
10. Graham Cormode and S. Muthukrishnan. “An Improved Data Stream Summary: The Count-Min Sketch and its Applications”. *Journal of Algorithms*, 55(1):58–75, April 2005.
11. P. Flajolet and G.N. Martin. “Probabilistic Counting Algorithms for Database Applications”. *Journal of Computer System and Sciences*, 31(2):182–209, 1985.
12. Sumit Ganguly. “A hybrid technique for estimating frequency moments over data streams”. Manuscript, July, 2004.
13. Sumit Ganguly. “Estimating Frequency Moments of Update Streams using Random Linear Combinations”. In *Proceedings of the International Workshop on Randomization and Computation (RANDOM)*, 2004.

14. Sumit Ganguly, Deepanjan Kesh, and Chandan Saha. “Practical Algorithms for Tracking Database Join Sizes”. In *Proceedings of the FSTTCS*, pages 294–305, December 2005.
15. Y. Gu, A. McCallum, and D. Towsley. “Detecting Anomalies in Network Traffic Using Maximum Entropy Estimation”. In *Proceedings of of Internet Measurement Conference*, pages 345–350, 2005.
16. Sudipto Guha, Andrew McGregor, and S. Venkatsubramanian. “Streaming and Sublinear Approximation of Entropy and Information Distances”. In *Proceedings of the ACM Symposium on Discrete Algorithms*, 2006.
17. Piotr Indyk. “Stable Distributions, Pseudo Random Generators, Embeddings and Data Stream Computation”. In *Proceedings of the IEEE Foundations of Computer Science*, pages 189–197, 2000.
18. Piotr Indyk and David Woodruff. “Optimal Approximations of the Frequency Moments”. In *Proceedings of the ACM Symposium on Theory of Computing*, pages 202–298, 2005.
19. Noam Nisan. “Pseudo-Random Generators for Space Bounded Computation”. In *Proceedings of the ACM Symposium on Theory of Computing*, 1990.
20. Michael Saks and Xiaodong Sun. “Space lower bounds for distance approximation in the data stream model”. In *Proceedings of the ACM Symposium on Theory of Computing*, 2002.
21. Mikkel Thorup and Yin Zhang. “Tabulation based 4-universal hashing with applications to second moment estimation”. In *Proceedings of the ACM Symposium on Discrete Algorithms*, pages 615–624, January 2004.
22. A. Wagner and B Plattner. “Entropy based worm and anomaly detection in fast IP networks”. In *14th IEEE WET ICE, STCA Security Workshop*, 2005.
23. M.N. Wegman and Carter J. L. “New Hash Functions and their Use in Authentication and Set Equality”. *Journal of Computer Systems and Sciences*, 22:265–279, 1981.
24. David P. Woodruff. “Optimal space lower bounds for all frequency moments”. In *Proceedings of the ACM Symposium on Discrete Algorithms*, pages 167–175, 2004.
25. K. Xu, Z. Zhang, and S. Bhattacharyya. “Profiling internet backbone traffic: behavior models and applications”. *SIGCOMM Comput. Commun. Rev.*, 35(4):169–180, 2005.