Estimating Entropy over Data Streams

Lakshminath Bhuvanagiri and Sumit Ganguly

Indian Institute of Technology, Kanpur
{blnath,sganguly}@cse.iitk.ac.in

Abstract. We present an algorithm for estimating entropy of data streams consisting of insertion and deletion operations using $\tilde{O}(1)$ space.[1]

1 Introduction

Recently, there has been an emergence of *monitoring applications* in diverse areas, including, network traffic monitoring, network topology monitoring, sensor networks, financial market monitoring, web-log monitoring, etc.. In these applications, data is generated rapidly and continuously, and must be analyzed very efficiently, in real-time, to identify large trends, anomalies, user-defined exception conditions, etc.. The data streaming model [1, 12] has gained popularity as a computational model for such applications—where, incoming data (or updates) are processed very efficiently and in an online fashion using space that is much less than what is needed to store the data in its entirety.

A data stream $\mathcal{S}$ is viewed as a sequence of arrivals of the form (i, v), where, i is the identity of an item that is a member of the domain $\{0, 1, \ldots, N-1\}$, and v is the *update* 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 updates to i since the inception of the stream, that is, $f_i = \sum_{(i,v) \text{ appears in } \mathcal{S}} v$. In the *strict update model*, deletions are assumed to correspond to prior insertions, and, therefore the frequency of an item is always non-negative. In the *general update model*, item frequencies may be negative or positive. Let n denote the number of items with non-zero frequencies in the stream and let m denote the sum of the absolute values of the item frequencies.

The entropy H of a data stream is defined as $H = \sum_{i:f_i>0} \frac{|f_i|}{m} \log \frac{m}{|f_i|}$. We study the problem of continuously tracking the entropy of a data stream in low space. The entropy of a data stream, or that of a frequency distribution, measures its information theoretic *randomness* and *incompressibility*. 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. Further, monitoring changes in the entropy of a network traffic stream has been used detect anomalies [8, 14, 15].

[1] We use the $\tilde{O}$ notation to simplify complexity expressions. If N is the domain size of the stream items and m is the sum of the absolute values of the item frequencies, we say that $f(m, N) = \tilde{O}(g(m, N))$, if $f(m, N) = O\left(\frac{1}{\epsilon^{O(1)}}(\log^{O(1)} m)(\log^{O(1)} n)\, g(m, N)\right)$.

Prior Work. The work in [9] presents an algorithm for estimating H over insert-only streams to within a factor of $1 \pm \epsilon$ using space $\tilde{O}(n^{\frac{1}{1+\epsilon}})$. [9] also presents an algorithm for estimating H using space $\tilde{O}(1)$ bits in a *random*-streaming model, in which, the order of arrival of items is assumed to be completely random. [3] presents an algorithm for estimating entropy H over insert-only streams to within factors of $1 \pm \epsilon$ using $\tilde{O}(\min(m^{2/3}, m^{2(1-\epsilon)})$ space.

Contributions. In this paper we present an algorithm that estimates the entropy of strict update data streams to within factors of $1 \pm \epsilon$ using space $\tilde{O}(1)$. We also show how the algorithm can be generalized to the general update streaming model using space $\tilde{O}(1)$. We prove the following theorem.

Theorem 1. *There exists an algorithm for strict update streams that returns an estimate $\hat{H}$ of the entropy H of a stream such that $|\hat{H} - H| \leq \epsilon H$ with probability $1 - \delta$ using $O\left(\frac{(\log^4 m)}{\epsilon^3} \frac{(\log m + \log \frac{1}{\epsilon})}{(\log \frac{1}{\epsilon} + \log\log m)} (\log \frac{1}{\delta})\right)$ bits.*

Organization. The remainder of the paper is organized as follows. In Section 2, we present an abstract algorithm called HSS for estimating a class of data stream metrics and then use it in Section 3 to estimate entropy.

2 The HSS algorithm

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.

$$\Psi(\mathcal{S}) = \sum_{i:f_i>0} \psi(f_i) \tag{1}$$

Section 3 specializes this procedure to yield a straight forward algorithm for estimating the entropy of a data stream in $\tilde{O}(1)$ space. The algorithm can be naturally adapted for general update streams. A specialization of the HSS procedure was used in [2] to give an algorithm for finding the k^{th} frequency moment $F_k = \sum f_i^k$.

Preliminaries. Given a data stream, *rank*(r) returns an item with the r^{th} largest frequency (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 $\leq k$. We use the COUNT-MIN algorithm [6] for estimating the frequency $\hat{f}_i$ of an item i. Its guarantees are summarized in Theorem 2 and are given in terms of the quantity $m^{res}(k) = \sum_{i \notin \textit{ top}(k)} f_i = \sum_{r>k} f_{\textit{rank}(r)}$.

Theorem 2. [6] *For $0 < \epsilon < 1$, the* COUNT-MIN *algorithm uses space $O(\frac{k}{\epsilon} \log \frac{1}{\delta} \log m)$ bits and time $O(\log \frac{1}{\delta})$ to process each stream update. It returns an estimate $\hat{f}_i$ that satisfies $f_i \leq \hat{f}_i \leq f_i + \frac{\epsilon m^{res}(k)}{k}$ with probability $1 - \delta$.* □

The HSS *structure.* Let $T_0 > T_1 > \ldots > T_L$ (L will be fixed later) be a sequence of exponentially decreasing thresholds that partition the elements of the stream into groups $G_0 \ldots G_L$, where, $G_0 = \{i \in \mathcal{S} : \ f_i \geq T_0\}$ and $G_l = \{i \in \mathcal{S} : \ T_l \leq f_i < T_{l-1}\}, \ \ 1 \leq l \leq L$. Intuitively, the HSS algorithm works 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 $\frac{1}{2}$ (hence $L = O(\log m)$). The sub-stream $\mathcal{S}_l$ is referred to as the sub-stream at level l, for $l = 0, 1, \ldots, L$. Let k be a space parameter. At each level l, we keep a data-structure denoted by $\mathcal{D}_l$, that takes as input the sub-stream $\mathcal{S}_l$, and returns an *approximation to the top*(k) *items of its input stream and their frequencies*. This data-structure is typically instantiated using standard synopsis structures, such as, COUNT-MIN (for estimating entropy), COUNTSKETCH [4] (for estimating frequency moments [2] and for estimating entropy when frequencies may be negative), etc. We posit the following invariant.

C1: All items of G_l present in $\mathcal{S}_l$ must be discovered as frequent items by $\mathcal{D}_l$.

Approximating f_i. We assume that frequent items discovery and frequency estimation algorithms have an additive error of at most Δ in the estimated frequencies (usually, with high probability [4–6]), where Δ is a function of the space parameter k and some aggregate statistic of the input stream (e.g., $m^{res}(k)$ for the COUNT-MIN algorithm and $F_2^{res}(k)$ for the COUNTSKETCH algorithm). We use $\Delta_l = \Delta_l(k)$ to denote the error incurred by the estimates obtained from $\mathcal{D}_l$ operating on the sub-stream $\mathcal{S}_l$.

Let Q_l denote a frequency threshold defined as $Q_l = \frac{\Delta_l}{\bar{\epsilon}}$ and let $\hat{f}_{i,l}$ denote the estimate of the frequency of i as obtained from the data structure $\mathcal{D}_l$, assuming that the item i has been sampled at level l. It follows that if $\hat{f}_{i,l} > Q_l$, then, $|\hat{f}_{i,l} - f_i| \leq \bar{\epsilon} f_i$ with high probability. Lemma 1 establishes a relation between the values Δ_l for various l for a popular *top*(k) estimation algorithm, namely the COUNT-MIN sketch [2]. This relationship helps us to set the threshold Q_l to $\frac{\Delta_0(k)}{\bar{\epsilon} \cdot 2^l} = \frac{Q_0}{2^l}$ for $l > 0$.

Lemma 1. *For* COUNT-MIN *sketch algorithm,* $\Delta_l(k) \leq \frac{\Delta_0(2^{l-1}k)}{2^l}$ *with probability* $\geq 1 - 2^{-\Omega(k)}$ *for* $l \geq 1$ *and* $k \geq 36$.

Proof. By Theorem 2, $\Delta_l(k) = \frac{m^{res}(k,l)}{k}$, where, $m^{res}(k, l)$ is the (random) residual first moment of $\mathcal{S}_l$ when the top-k ranked items have been removed from $\mathcal{S}_l$. The expected number of the top-$2^{l-1}k$ ranked items in $\mathcal{S}$ appearing in $\mathcal{S}_l$ is $\frac{1}{2^l} 2^{l-1}k = \frac{k}{2}$. By Chernoff's bounds, the number of the top-$\frac{2^l k}{2}$ ranked items in $\mathcal{S}$ appearing in $\mathcal{S}_l$ is no more than $\frac{3k}{4}$, with probability at least $1 - 2^{-\Omega(k)}$. In other words, the non-top-k elements of $\mathcal{S}_l$ only includes the non-top-$2^{l-1}k$ elements of the original stream, with probability $1 - 2^{-\Omega(k)}$. Therefore, $\mathsf{E}\big[m^{res}(k, l)\big] \leq \frac{1}{2^l} m^{res}(2^{l-1}k)$. By a similar argument, the largest non-top-k frequency, u_l, in $\mathcal{S}_l$ has rank at least $\frac{3 \cdot 2^l k}{4}$ in $\mathcal{S}$. Thus, $u_l \leq f_{rank(3 \cdot 2^{l-2} \cdot k)} \leq \frac{m^{res}(2^{l-1}s)}{2^{l-2}k}$. Items with ranks between $\frac{3 \cdot 2^l k}{4}$ and $2^{l-1}k$ have frequencies at least u_l, and their sum is at most $m^{res}(2^{l-1}k)$. That is, $(\frac{3}{4})(2^l k) u_l \leq m^{res}(2^{l-1}k)$, or that, $u_l \leq \frac{m^{res}(2^{l-1}k)}{3 \cdot 2^{l-2}k}$.

[2] A similar result can also be shown for COUNTSKETCH .

Hence, with probability $\geq 1-\delta$, $m^{res}(k,l) \leq \max(2\mathsf{E}\left[m^{res}(k,l)\right], 3u_l \log\frac{1}{\delta})) \leq \max(2\frac{m^{res}(2^{l-1}k)}{2^l}, 3\frac{m^{res}(2^{l-1}k)}{2^{l-2}k}\log\frac{1}{\delta})$ (from Hoeffding's bounds). Let $\delta = 2^{-\frac{k}{16}}$. Since $k \geq 36$, we get $m^{res}(k,l) \leq \frac{m^{res}(2^{l-1}k)}{2^l}$ with probability $> 1-2^{-\Omega(k)}$. □

Disambiguating estimated frequency. It is possible for the estimate $\hat{f}_{i,l}$ of an item i obtained from the sub-stream $\mathcal{S}_l$ to exceed the threshold Q_l for multiple l. For example, consider an item with actual frequency larger than Q_0. It crosses the threshold at level 0 and thus is estimated accurately at level 0. However, it may be sub-sampled at level 1, and in this case, its frequency estimate also crosses the threshold Q_1 (with high probability). In this manner, this item may get estimated accurately at all the levels at which it has been successfully sub-sampled. Each of these estimates $\hat{f}_{i,l}$ may be different, though they are all within factors of $1 \pm \bar{\epsilon}$ to the actual value. We therefore apply the "disambiguation-rule" of using the estimate obtained from the *lowest level* at which it crosses the threshold for that level. The estimated frequency after disambiguation is denoted as $\hat{f}_i$.

Setting T_l. As per the invariant **C1**, all elements in $G_l \cap S_l$ must be discovered as frequent items by $\mathcal{D}_l$. Since, Q_l defines the threshold for "good estimation", fixing $T_l = Q_l$ might seem a possibility. However, the elements of G_l with frequency close to $T_l (= Q_l)$ might fail to even appear in the sample $\bar{G}_l$ due to errors in estimation, thereby violating **C1**. One way to solve this problem is to choose $T_l = \sqrt{Q_l \cdot Q_{l+1}} = 2^{1/2} \cdot Q_l$. Since $Q_l(1+\bar{\epsilon}) \geq Q_l + \Delta_l$ (by definition of Q_l), we choose $\bar{\epsilon}$ such that $(1+\bar{\epsilon}) < 2^{1/2}$, and hence $T_l \geq Q_l + \Delta_l$. Thus, any $i \in G_l$ appearing in $\mathcal{S}_l$ will be present in the *top*(k) set returned by $\mathcal{D}_l$, and hence included into the sample $\bar{G}_l$ since it can suffer an additive error of at most $\pm\Delta_l$.

2.1 Algorithm

Obtaining hierarchical samples. For every stream update (i, v), we use a hash-function $h : \{1 \ldots N\} \rightarrow \{1 \ldots N\}$ to map the item onto level $u = lsb(h(i))$ [3]. The update (i, v) is then propagated to the frequent items data structures $\mathcal{D}_l$ for $0 \leq l \leq u$, in effect, i is included in the sub-streams from level 0 to level u. The hash function is assumed to be chosen randomly from a fully independent family; later we reduce the number of random bits required.

At inference time, the algorithm collects samples as follows. From each level l, the set of items whose estimated frequency crosses the threshold Q_l are identified, using the frequent items structure $\mathcal{D}_l$. If an item crosses the threshold at multiple levels, then, the disambiguation rule is applied that sets $\hat{f}_i$ to $\hat{f}_{i,l}$, where, l is the smallest of the levels r such that $\hat{f}_{i,r} \geq Q_r$. Based on their disambiguated frequencies, the sampled items are sorted into their respective groups. In order to maintain the invariant **C1**, we include an item i in group G_l only if it hashes to level l. More, precisely, we form the sampled groups, $\bar{G}_0, \bar{G}_1, \ldots, \bar{G}_L$, as follows.

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

[3] $lsb(x)$ is the position of the least significant "1" in binary representation of x.

Note that any item belonging to G_l and $\mathcal{S}_l$ is "discovered" at level l, with high probability . However, if f_i is close to the right or the left boundary of G_l, the $\pm\bar{\epsilon}f_i$ estimation error could cause i to be misclassified into its adjacent group. We consider this issue in the next section.

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.

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

2.2 Analysis

Let $x_{i,r}$ denote a random variable which takes the value 1 iff $i \in \mathcal{S}_r$ and is also classified by the algorithm into $\bar{G}_r$. Thus, equation (2) can be written as follows.

$$\bar{\Psi} = \sum_{i \in \mathcal{S}} \psi(f_i) \sum_{r=0}^{L} x_{i,r} \cdot 2^r, \qquad \hat{\Psi} = \qquad \sum_{i \in \mathcal{S}} \psi(\hat{f}_i) \sum_{r=0}^{L} x_{i,r} \cdot 2^r$$

Lemma 2 shows that the expected value of $\bar{\Psi}$ is close to Ψ.

Lemma 2. *Suppose that for* $0 \le i \le N-1$ *and* $0 \le l \le L$, $|\hat{f}_{i,l} - f_i| \le \epsilon f_i$ *with probability* $\ge 1 - 2^{-t}$. *Then* $|\mathsf{E}[\bar{\Psi}] - \Psi| \le \Psi \cdot 2^{-t+\log L}$.

Proof. $\mathsf{E}[\bar{\Psi}] = \sum_{i \in \mathcal{S}} \mathsf{E}[\psi(f_i) \sum_{r=0}^{L} x_{i,r} \cdot 2^r] = \sum_{i \in \mathcal{S}} \psi(f_i) \sum_{r=0}^{L} \mathsf{E}[x_{i,r} \cdot 2^r]$, where, $\sum_{r=0}^{L} \mathsf{E}[x_{i,r} \cdot 2^r] = \sum_{r=0}^{L} \mathsf{Pr}\{x_{i,r} = 1\} \cdot 2^r$.

Consider an item $i \in G_l$. The frequency group (or interval) G_l is partitioned into three sub-regions, namely, $lr(G_l) = [T_l, T_l + \bar{\epsilon}Q_l]$, $rr(G_l) = [T_{l-1} - \bar{\epsilon}Q_l, T_{l-1}]$ and $mr(G_l) = [T_l + \bar{\epsilon}Q_l, T_{l-1} - \bar{\epsilon}Q_l]$, that, respectively denote the *left-region*, *right-region* 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. As seen earlier, if $i \in G_l$ and $i \in \mathcal{S}_l$, then $\hat{f}_{i,l} > Q_l$, and hence, it is discovered by $\mathcal{D}_l$ (with high probability). We now consider items that lie in the middle-region and the left-region respectively (the argument for right-region is analogous to that of the left-region).

Let $i \in mr(G_l)$. Then, $T_l \le f_i - \Delta_l \le \hat{f}_{i,l} \le f_i + \Delta_l \le T_{l-1}$, that is, the error Δ_l is not large enough to cause $\hat{f}_{i,l}$ to cross either T_l or T_{l-1}. Hence, with probability $1 - 2^{-t}$, if $i \in \mathcal{S}_l$, then, is correctly classified into group $\bar{G}_l$. Therefore $\frac{1}{2^t}(1 - 2^{-t}) \le \mathsf{Pr}\{x_{i,l} = 1\} \le \frac{1}{2^t}$, and $\mathsf{Pr}\{x_{i,r} = 1\} \le \frac{1}{2^r} \cdot 2^{-t}$ for $r \ne l$. Thus, $\sum_{r=0}^{L} \mathsf{E}[x_{i,r} 2^r] \le \frac{1}{2^l} \cdot 2^l + \sum_{r \ne l} 2^{-t} \cdot \frac{1}{2^r} \cdot 2^r \le 1 + L \cdot 2^{-t}$, and $\sum_{r=0}^{L} \mathsf{E}[x_{i,r} 2^r] \ge \frac{1}{2^l} \cdot (1 - 2^{-t}) \cdot 2^l$. Hence, $|\sum_{r=0}^{L} \mathsf{E}[x_{i,r} 2^r] - 1| \le 2^{-t+\log L}$.

Let $i \in lr(G_l)$. Then, with high probability, i will not be discovered at any level $l' < l$, since $\hat{f}_{i,l'} \le f_i + \bar{\epsilon}Q_{l'} \le (1+\bar{\epsilon})T_l + \bar{\epsilon}Q_{l'} < Q_{l'}$. Therefore, the estimate $\hat{f}_i = \hat{f}_{i,l}$

is obtained from level l. However, by virtue of its true frequency (f_i) being close to T_l, the estimate $\hat{f}_{i,l}$ might be on either side of T_l causing i to be classified into either $\bar{G}_l$ or $\bar{G}_{l+1}$. Let p be the probability that $\hat{f}_{i,l} > T_l$, that is, the item gets "correctly" classified into group $\bar{G}_l$. Therefore, $(1-2^{-t})\cdot p\cdot\frac{1}{2^l} \leq \mathsf{Pr}\{x_{i,l}=1\} \leq p\cdot\frac{1}{2^l}$. The probability of the same i getting classified into $\bar{G}_{l+1}$ at $\mathcal{D}_l$ is $1-p$, resulting in $(1-p)\cdot(1-2^{-t})\cdot 2^{-(l+1)} \leq \mathsf{Pr}\{x_{i,l+1}=1\} \leq (1-p)\cdot\frac{1}{2^{l+1}}$. Note that this argument assumes that the random sub-sampling choices for an item are made independent of the choices of the other items. Therefore, $\sum_{r=0}^{L}\mathsf{E}\big[x_{i,r}\ 2^r\big] \leq (p\ \frac{1}{2^l})\ 2^l + (1-p)\ \frac{1}{2^{l+1}}\ 2^{l+1} + \sum_{r\notin\{l,l-1\}}\frac{1}{2^r}\ 2^r\ 2^{-t}$. Therefore, $1-2^{-t} \leq \sum_{r=0}^{L}\mathsf{E}\big[x_{i,r}\ 2^r\big] < 1 + L\cdot 2^{-t}$. Therefore, $|\mathsf{E}\big[\sum_{r=0}^{L} x_{i,r}\ 2^r - 1\big]| \leq 2^{-t+\log L}$. A similar argument can be made for $rr(G_l)$. Combining, we get

$$|\mathsf{E}\big[\bar{\Psi}\big] - \Psi(\mathcal{S})| = \sum_{i\in\mathcal{S}}\psi(f_i)\ |\mathsf{E}\big[x_{i,r}\ 2^r\big] - 1| \leq \Psi(\mathcal{S})\cdot 2^{-t+\log L}\ .\ \square$$

We now present a bound on the variance of the idealized estimator. For any item i with non-zero frequency, we denote by $l(i)$ the group index l such that $i \in G_l$.

Lemma 3. *Suppose that for all $0 \leq i \leq N-1$ and $0 \leq l \leq L$, $|\hat{f}_{i,l} - f_i| \leq \epsilon f_i$ with probability $\geq 1-2^{-t}$. Then,*

$$\mathsf{Var}\big[\bar{\Psi}\big] \leq \sum_{i\in\mathcal{S}} 2^{-t+L+2}\cdot\psi^2(f_i) + \sum_{i\notin(G_0-lm(G_0))}\psi^2(f_i)\cdot 2^{l(i)+1}\ .$$

Proof. The proof is analogous to that of Lemma 2 and is given in Appendix A. $\square$

Corollary 1. *If the function $\psi(\cdot)$ is increasing in the interval $[0\ldots T_0+\Delta_0]$, then, choosing $t = L + \log\frac{1}{\epsilon^2} + 2$ we get*

$$\mathsf{Var}\big[\bar{\Psi}\big] \leq \sum_{i\in\mathcal{S}}\epsilon^2\psi^2(f_i) + \sum_{l=1}^{L}\sum_{i\in G_l}\psi(T_{l-1})\psi(f_i)2^{l+1} + 2\sum_{i\in lm(G_0)}\psi(T_l+\Delta_0)\psi(f_i) \tag{3}$$

Proof. If the monotonicity condition is satisfied, then $\psi(T_{l-1}) > \psi(f_i)$ for all $i \in G_l$, $l \geq 1$ and $\psi(f_i) \leq \psi(T_0+\Delta_0)$ for $i \in lm(G_0)$. Therefore, $\psi^2(f_i) \leq \psi(T_{l-1})\cdot\psi(f_i)$, in the first case and $\psi^2(f_i) \leq \psi(T_0+\Delta_0)$ in the second case. By Lemma 3 and the chosen value for t gives the desired result. $\square$

2.3 Error in the estimate

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

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

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

$$\mathcal{E}_1 = |\Psi-\bar{\Psi}| \leq |E[\bar{\Psi}]-\Psi| + 3\sqrt{\mathsf{Var}\big[\bar{\Psi}\big]} \text{ with probability } \frac{8}{9}\ .$$

Using Lemma 2 and Corollary 1, and choosing $t = L + \log \frac{1}{\epsilon^2} + 2$, the expression for $\mathcal{E}_1$ can be simplified as follows.

$$\mathcal{E}_1 \leq \frac{\epsilon^2 L\Psi}{m} + 3\Big(\sum_{i\in\mathcal{S}} \epsilon^2\psi^2(f_i) + \sum_{i\in G_l, l\geq 1} \psi(T_{l-1})\psi(f_i)2^{l+1} + \sum_{i\in lm(G_0)} 2\psi(T_l+\Delta_0)\psi(f_i)\Big)^{1/2} \tag{4}$$

with probability $\frac{8}{9}$. We now present an upper bound on $\mathcal{E}_2$.

Lemma 4. *Suppose that for $0 \leq i \leq N-1$ and $0 \leq l \leq L$, $|\hat{f}_{i,l} - f_i| \leq \epsilon f_i$ with probability $\geq 1 - 2^{-t}$. Then, $\mathcal{E}_2 \leq 16 \cdot \bar{\epsilon} \cdot Q_0 \sum_{l=0}^{L} \sum_{i\in G_l} \frac{|\psi'(\xi_i)|}{2^l}$ with probability $\geq \frac{9}{10} - 2^{-t}$, where for an $i \in G_l$, ξ_i lies between f_i and $\hat{f}_i$, and maximizes $\psi'()$.*

Proof. Let $y_{i,l}$ denote the indicator random variable that is 1 if $i \in \mathcal{S}_l$ and is 0 otherwise. Note that $y_{i,l+1}$ is 1 only if $y_{i,l}$ is 1. Let $i \in G_l$. By arguing similarly to that of Lemma 2, we have the following cases. Consider the set of items i such that $i \in mr(G_l)$, for some $l \geq 1$, or, $i \in G_0 - lr(G_0)$. If $r = l$, then, $x_{i,r} = y_{i,r}$, with probability $1 - 2^{-t}$, and otherwise, $x_{i,r} = 0$, with probability at most 2^{-t}. Therefore, for such an item i , $\sum_{r=0}^{L} x_{i,r}2^r = y_{i,l}2^l$, with probability $1 - 2^{-t}$.

Suppose $i \in lr(G_l)$, for some $l \geq 0$. Then, $x_{i,l} + x_{i,l+1} = y_{i,l}$, with probability $1 - 2^{-t}$, and $x_{i,r} = 0$, with probability at most 2^{-t}, for $r \notin \{l, l+1\}$. Therefore, for such items i, $\sum_{r=0}^{L} x_{i,r}2^r = x_{i,l}2^l + x_{i,l+1}2^{l+1} \leq (x_{i,l} + x_{i,l+1})2^{l+1} = y_{i,l}2^{l+1}$, with probability $1 - 2^{-t}$.

Finally, consider those items i such that $i \in rr(G_l)$ for some $l \geq 1$. Using a similar argument, we can show that $x_{i,l} + x_{i,l-1} = y_{i,l-1}$, with probability $1 - 2^{-t}$, and $x_{i,r} = 0$, with probability 2^{-t} for $r \notin \{l, l-1\}$. Therefore, $\sum_{r=0}^{L} x_{i,r}2^r \leq y_{i,l-1}2^l$. Since, $y_{i,l}$ is 1 only if $y_{i,l-1} = 1$, for $l \geq 1$, in all cases, we have,

$$\sum_{r=0}^{L} x_{i,r} \cdot 2^r \leq y_{i,l(i)-1} \cdot 2^{l+1}, \text{ with probability } \geq 1 - 2^{-t}, \text{ if } l(i) \geq 1 \ . \tag{5}$$

By triangle inequality, $\mathcal{E}_2 \leq \sum_{l=0}^{L} \sum_{i\in G_l} |\psi(\hat{f}_i) - \psi(f_i)| \cdot (\sum_{r=0}^{L} x_{i,r} \cdot 2^r)$. Using Taylor's expansion, $\mathcal{E}_2 \leq \sum_{l=0}^{L} \sum_{i\in G_l} |\Delta_l \cdot \psi'(\xi_i)| \cdot (\sum_{r=0}^{L} x_{i,r} \cdot 2^r)$, where ξ_i is the value between f_i and $\hat{f}_i$ at which $\psi'()$ takes its maximum absolute value. Using (5)

$$\mathcal{E}_2 \leq 2\Delta_0 \sum_{i\in G_0} |\psi'(\xi_i)| y_{i,0} + \sum_{l=1}^{L} \Delta_l \sum_{i\in G_l} |\psi'(\xi_i)| \ y_{i,l-1} \ 2^{l+1}.$$

Since $\mathcal{S}_0 = \mathcal{S}$, we have $y_{i,0} = 1, \ \forall i \in G_0$. Applying Hoeffding's bounds to the second term above , we obtain with probability $\geq \frac{9}{10} - 2^{-t}$,

$$\begin{aligned}\mathcal{E}_2 &\leq 2\Delta_0\Big(\sum_{i\in G_0} |\psi'(\xi_i)|\Big) + 2 \cdot \Big(4\sum_{l=1}^{L} \Delta_l \sum_{i\in G_l} |\psi'(\xi_i)| + 4 \max_{i\in\mathcal{S}-G_0} |\Delta_{l(i)} \ \psi'(\xi_i)|\Big) \\ &\leq 16\sum_{l=0}^{L} \Delta_l \sum_{i\in G_l} |\psi'(\xi_i)| \leq 16 \cdot \bar{\epsilon} \cdot Q_0 \sum_{l=0}^{L} \sum_{i\in G_l} \frac{|\psi'(\xi_i)|}{2^l}, \text{ since, } \Delta_l = \frac{\bar{\epsilon}Q_0}{2^l}. \ \square\end{aligned}$$

Reducing random bits. The number of random bits used by the algorithm can be reduced to $O(s \log m)$, where, s is the space used by the HSS structure, by using a classical result of Nisan [13] on pseudo-generators for space bounded computation, as adapted for use by data stream algorithms by Indyk [10]. Since, this approach has been adequately treated in [10] and [11], we do not discuss it in greater detail.

3 Estimating Entropy

In this section, we apply the HSS algorithm to estimate the entropy $H = \sum_{i:f_i>0} \frac{f_i}{m} \log \frac{m}{f_i}$ of a data stream. We assume that the stream follows the strict update model (i.e., $f_i \geq 0$). Later we remark how the algorithm can be modified for general update streams. For any $0 \leq x \leq m$, let $h(x)$ denote $\frac{x}{m} \log \frac{m}{x}$ (we assume that $h(0) = 0$). In this section, the function $\psi(x) = h(x)$ and the statistic $\Psi = \sum_i h(f_i) = H$.

We instantiate the HSS algorithm using COUNT-MIN sketch [6] as the frequent items structure $\mathcal{D}_l$ with $\frac{8k}{\bar{\epsilon}}$ buckets in each hash table, where $\bar{\epsilon} = \epsilon$ itself. We also estimate $m^{res}(k)$ to within accuracy factors of $1 \pm \epsilon$ with probability $1 - \delta$. This is done using an algorithm similar to that of estimating $F_2^{res}(k)$ presented in [7], and uses space $O(\frac{k}{\epsilon} \log \frac{m}{\delta} \log m)$ bits. For brevity, we state the theorem without proof.

Theorem 3. *For a given integer $k \geq 1$ and $0 < \epsilon < 1$, there exists an algorithm for strict update streams that returns an estimate $\hat{m}^{res}(k)$ satisfying $(1-\epsilon)m^{res}(k) \leq \hat{m}^{res}(k) \leq m^{res}(k)$ with probability $1 - \delta$ using $O(\frac{k}{\epsilon}(\log \frac{k}{\delta})(\log m))$ bits.* □

We use the algorithm of Theorem 3 with $\delta = (20mL)^{-1}$ to obtain an estimate $\hat{m}^{res}(k)$ and use it to compute the thresholds Q_l and T_l, for levels $l = 0, 1, \ldots, L$, as follows, where $L = \lceil \log \frac{m}{k} \rceil$: $Q_0 = \frac{\hat{m}^{res}(k)}{k}$, $Q_l = \frac{Q_0}{2^l}$ and $T_l = 2^{1/2} Q_l$, for $0 \leq l \leq L$. We now bound the errors $\mathcal{E}_1$ and $\mathcal{E}_2$.

Lemma 5. *Let $k = \frac{4\sqrt{2} \log m}{\epsilon^2 (\log \frac{1}{\epsilon} + \log \log m)}$ and $0 < \epsilon < \frac{1}{2}$. Then, $\mathcal{E}_1 \leq 5\epsilon H$.*

Proof. We use (4) to bound $\mathcal{E}_1$. For $l \geq 1$, $h(T_{l-1}) \cdot 2^{l+1} = \left(\frac{T_{l-1}}{m} \log \frac{m}{T_{l-1}}\right) \cdot 2^{l+1} = \left(\frac{\sqrt{2}\hat{m}^{res}(k)}{m \cdot (k \cdot 2^{l-1})} \log \frac{m}{T_l}\right) \cdot 2^{l+1} \leq \frac{4\sqrt{2} m^{res}(k)}{mk} \log m$. Further, since the frequency of a non-top k item is at most $\frac{m}{k}$, we have, $\hat{m}^{res}(k) = \sum_{i \notin \ top(k)} f_i \leq \frac{1}{\log k} \sum_{i \notin top(k)} f_i \cdot \log \frac{m}{f_i} < \frac{mH}{\log k}$. Therefore, $h(T_l) \cdot 2^{l+1} \leq \frac{4\sqrt{2} H \log m}{k \log k}$, for $l \geq 1$. Further, $2h(T_0 + \Delta_0) \leq \frac{2(T_0+\Delta_0)}{m} \log \left(\frac{m}{T_0+\Delta_0}\right) \leq \frac{2(1+\frac{\epsilon}{\sqrt{2}})T_0}{m} (\log m) \leq \frac{4m^{res}(k) \log m}{mk} \leq \frac{4H \log m}{k \log k}$, by the argument above. Therefore, the following summation from the expression for $\mathcal{E}_1$ in (4) is,

$$\sum_{i \in G_l, l \geq 1} h(T_{l-1}) h(f_i) 2^{l+1} + \sum_{i \in lm(G_0)} 2h(T_l + \Delta_0) h(f_i)$$

$$\leq \frac{4\sqrt{2} H \log m}{k \log k} \left(\sum_{i \in G_l, l \geq 1} h(f_i) + \sum_{i \in lm(G_0)} h(f_i) \right) \leq \frac{4\sqrt{2} H \log m}{k \log k} \sum_i h(f_i)$$

$$\leq \frac{4\sqrt{2} H^2 \log m}{k \log k} \leq \epsilon^2 H^2$$

by the choice of k as given in the statement. Substituting in (4), and using $L = \log \frac{m}{k}$, we obtain that $\mathcal{E}_1 \leq \frac{\epsilon^2 (\log m) H}{m} + 3(\epsilon^2 H^2 + \epsilon^2 H^2)^{1/2} < 5\epsilon H$. □

Lemma 6. *If* $0 < \epsilon \leq 1$ *and* $k \geq \lceil 8e \rceil$, *then* $\mathcal{E}_2 \leq 8\sqrt{2}\epsilon H$.

Proof. Since, $\bar{\epsilon} = \epsilon$, by Lemma 4, $\mathcal{E}_2 \leq 16\epsilon Q_0 \sum_{l=0}^{L} \sum_{i \in G_l} \frac{|h'(\xi_i)|}{2^l}$, with probability $\geq \frac{9}{10} - 2^{-t}$. Since we are using COUNT-MIN sketch , for an $i \in G_l$, ξ_i is the value which maximizes $|h'()|$ in the interval $(f_i, f_i + \Delta_l)$. By Theorem 3, $\hat{m}^{res}(k) < m^{res}(k)$, and therefore, $Q_0 = \frac{\hat{m}^{res}(k)}{k} < \frac{m^{res}(k)}{k}$. Let h_i denote $h(f_i)$, that is, the contribution of i to H. Let θ_i denote the contribution of i to $\mathcal{E}_2$, that is, $\theta_i = \frac{16\epsilon Q_0 |h'(\xi_i)|}{2^{l(i)}}$. Thus, $\mathcal{E}_2 = \sum_{i:f_i>0} \theta_i$.

Case 1: $f_i \leq \frac{m}{e} - \Delta_0$. Since $h'()$ is positive and non-increasing in $[1, \frac{m}{e}]$, the value of ξ_i maximizing $|h'()|$ in $[f_i, f_i + \Delta_0]$ is f_i. Therefore $h'(\xi_i) \leq h'(f_i) = \frac{1}{m}\big(\log \frac{m}{f_i} - 1\big) < \frac{1}{m} \log \frac{m}{f_i} = \frac{h_i}{f_i} < \frac{h_i}{T_{l(i)}}$. Therefore, $\theta_i \leq \frac{16\epsilon Q_0 h_i}{2^{l(i)} T_{l(i)}} \leq 8\sqrt{2}\epsilon h_i$, since, $2^{l(i)} T_{l(i)} = T_0$ and $T_0 = Q_0\sqrt{2}$.

Case 2: $\frac{m}{e} - \Delta_0 < f_i \leq \frac{m}{e}$. Since, $k \geq \lceil 8e \rceil$, $T_0 \leq \frac{m}{k} \leq \frac{m}{8e}$ and therefore, $i \in G_0$. In this case, we consider two possibilities. First, if $\hat{f}_i < \frac{m}{e}$, then the value of $\xi_i \in \{f_i, \hat{f}_i\}$ maximizing $h'()$ will be f_i, and the analysis proceeds as in Case 1. The second possibility is: $\hat{f}_i > \frac{m}{e}$. In this case, note that $|h'(f_i - y)| > |h'(f_i + y)|$ for $0 < y < \Delta_0$. Hence, $|h'(\xi_i)| < |h'(f_i - \Delta_0)| < \frac{1}{m} \log \frac{m}{f_i - \Delta_0} = \frac{1}{m}(\log \frac{m}{f_i} - \log(1 - \frac{\Delta_0}{f_i})) < \frac{h_i}{f_i} + \frac{2\Delta_0}{m f_i} = \frac{h_i}{f_i}(1 + \frac{2\Delta_0}{m h_i})$. Since $m h_i = f_i \log \frac{m}{f_i} > f_i$, and $\frac{\Delta_0}{f_i} < \frac{\Delta_0}{T_0} = \frac{\epsilon}{\sqrt{2}}$ we can write $h'(\xi_i) < \frac{h_i}{f_i}(1 + \epsilon\sqrt{2}) \leq \frac{h_i(1+\sqrt{2})}{f_i}$, since $\epsilon \leq 1$. Also, $f_i \geq \frac{m}{e} - \Delta_0 = \frac{m}{e} - \frac{\epsilon T_0}{\sqrt{2}} \geq \frac{m}{e} - T_0$. Since, $T_0 \leq \frac{m}{k} \leq \frac{m}{8e}$, therefore, $f_i \geq 7T_0$. Thus, $h'(\xi_i) < \frac{h_i}{7T_0}(1 + \sqrt{2}) \leq \frac{h_i}{2T_0}$. Therefore, $\theta_i \leq 8\epsilon h_i$.

Case 3: $f_i > \frac{m}{e}$. As argued in Case 2, $i \in G_0$. Also $|h'()|$ is increasing in the range $(\frac{m}{e}, m)$. Let $f_i = (1 - \alpha_i)m$, where, $0 \leq \alpha_i < 1 - \frac{1}{e}$. Therefore $h_i = (1 - \alpha_i) \log \frac{1}{(1-\alpha_i)} > \alpha_i(1 - \alpha_i)$. Further, $|h'(\xi_i)| < |h'(f_i + \Delta_0)|$. Let $f_i + \Delta_0 = (1 - \alpha_i')m$, that is, $\alpha_i' = \alpha_i - \frac{\Delta_0}{m}$. Then, $|h'((1 - \alpha_i')m)| = \frac{1}{m}(1 - \log \frac{1}{1-\alpha_i'}) < \frac{1-\alpha_i'}{m}$. Further, $Q_0 \leq \frac{m^{res}(k)}{\sqrt{2}k} \leq \frac{\alpha_i m}{\sqrt{2}k}$, which gives, $\theta_i = 16\epsilon Q_0 h'(\xi_i) < 16\epsilon \frac{\alpha_i m}{k} \frac{(1-\alpha_i')}{m} = \frac{16\epsilon(1-\alpha_i')\alpha_i}{k} \leq \frac{32\epsilon(1-\alpha_i)\alpha_i}{k}$, since, $\hat{f}_i = (1 - \alpha')m \leq f_i + \Delta_0 = f_i(1 + \frac{\Delta_0}{f_i}) = (1 - \alpha)m(1 + \frac{\epsilon Q_0}{T_0}) < 2(1 - \alpha)m$. Therefore, $\theta_i \leq \frac{32\epsilon h_i}{k} \leq 2\epsilon h_i$, for the given k.

In all cases, $\theta_i \leq 8\sqrt{2}\epsilon h_i$. Therefore, $\mathcal{E}_2 = \sum_{i:f_i>0} \theta_i \leq 8\sqrt{2}\epsilon \sum_{i:f_i>0} h_i = 8\sqrt{2}\epsilon H$. □

We can now prove the main theorem.

Proof. [Of Theorem 1] The estimation error is bounded by $\mathcal{E}_1 + \mathcal{E}_2$. By Lemmas 5 and 6, the total error is $(5 + 8\sqrt{2})\epsilon H \leq 17\epsilon H$. Replacing ϵ by $\frac{\epsilon}{17}$ and returning the median $\hat{H}^{\text{med}}$ of $O(\log \frac{1}{\delta})$ independent estimates gives $|\hat{H}^{\text{med}} - H| \leq \epsilon H$ with probability $1 - \delta$.

Let $k = O\left(\frac{\log m}{\epsilon^2(\log\frac{1}{\epsilon}+\log\log m)}\right)$. The space used by the COUNT-MIN sketch substructure at each level of the HSS structure is $O(\frac{k}{\epsilon}(\log m + \log\frac{1}{\epsilon})(\log m))$ bits. This is calculated as follows. The height of the COUNT-MIN structure is $O(\frac{k}{\epsilon})$, width = $O(\log m+\log\frac{1}{\epsilon})$, since, confidence of inference per item is $1-2^{-t}$ and $t = O(\log m+\log\frac{1}{\epsilon})$ and there are $O(\frac{k}{\epsilon})$ items retrieved at each level.[4] Finally, each counter requires $O(\log m)$ bits for storage. The number of levels is $L = \log\frac{m}{k} = O(\log m)$. The use of the pseudo-random generator contributes an additional factor of $\log m$ to the space requirement. A collection of $O(\log\frac{1}{\delta})$ copies are kept to return the median estimate. Therefore, the total space requirement is $O\left(\frac{k}{\epsilon}(\log m+\log\frac{1}{\epsilon})(\log^3 m)(\log\frac{1}{\delta})\right) = O\left(\frac{(\log^4 m)}{\epsilon^3}\frac{(\log m+\log\frac{1}{\epsilon})}{(\log\frac{1}{\epsilon}+\log\log m)}(\log\frac{1}{\delta})\right)$ bits. □

Generalizing to streams with negative frequencies. We briefly outline how the algorithm can be applied to the general update streaming model. First, the COUNTSKETCH algorithm for finding frequent items is used instead of COUNT-MIN sketch algorithm. The role of $m^{res}(k)$ is replaced by $F_2^{res}(k)$; otherwise, the algorithm and its analysis is quite similar. The space complexity of the algorithm is polynomial in $\frac{1}{\epsilon}$ and $(\log F_2+\log N)$.

References

1. B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom. "Models and Issues in Data Stream Systems". In *Proceedings of ACM PODS*, 2002.
2. L. Bhuvanagiri, S. Ganguly, D. Kesh, and C. Saha. "Simpler algorithm for estimating frequency moments of data streams". In *Proceedings of ACM SODA*, 2006.
3. A. Chakrabarti, D.K. Ba, and S. Muthukrishnan. "Estimating Entropy and Entropy Norm on Data Streams". In *Proceedings of STACS*, 2006.
4. M. Charikar, K. Chen, and M. Farach-Colton. "Finding frequent items in data streams". In *Proceedings of ICALP*, 2002.
5. G. Cormode and S. Muthukrishnan. "What's Hot and What's Not: Tracking Most Frequent Items Dynamically". In *Proceedings of ACM PODS*, May 2003.
6. G. Cormode and S. Muthukrishnan. "An improved data stream summary: The Count-Min sketch and its applications". In *Proceedings of LATIN, Springer LNCS Vol. 2976*, 2004.
7. S. Ganguly, D. Kesh, and C. Saha. "Practical Algorithms for Tracking Database Join Sizes". In *Proceedings of FSTTCS*, 2005.
8. Y. Gu, A. McCallum, and D. Towsley. "Detecting Anomalies in Network Traffic Using Maximum Entropy Estimation". In *Proceedings of of Internet Measurement Conference*, 2005.
9. S. Guha, A. McGregor, and S. Venkatsubramanian. "Streaming and Sublinear Approximation of Entropy and Information Distances". In *Proceedings of ACM SODA*, 2006.
10. P. Indyk. "Stable Distributions, Pseudo Random Generators, Embeddings and Data Stream Computation". In *Proceedings of IEEE FOCS*, 2000.
11. P. Indyk and D. Woodruff. "Optimal Approximations of the Frequency Moments". In *Proceedings of ACM STOC*, 2005.

[4] For efficient retrieval of frequent items, a COUNT-MIN structure can be kept for each of the $\log m$ dyadic levels; this would obviate the need for a sequential scan of the domain, at the expense of an additional factor of $O(\log m)$ space.

12. S. Muthukrishnan. *"Data Streams: Algorithms and Applications"*. Foundations and Trends in Theoretical Computer Science, *Vol. 1, Issue 2*, 2005.
13. N. Nisan. "Pseudo-Random Generators for Space Bounded Computation". In *Proceedings of ACM STOC*, 1990.
14. A. Wagner and B Plattner. "Entropy based worm and anomaly detection in fast IP networks". In *14th IEEE WET ICE, STCA Security Workshop*, 2005.
15. K. Xu, Z. Zhang, and S. Bhattacharyya. "Profiling internet backbone traffic: behavior models and applications". *SIGCOMM Comput. Commun. Rev.*, 35(4), 2005.

A Proof of Lemma 3

Proof.

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

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 independent of each other, regardless of the values of r_1 and r_2. Thus,

$$\mathsf{E}\big[\bar{\Psi}\big]^2 = \sum_i \mathsf{E}\big[\psi^2(f_i)\sum_{r=0}^{L} x_{i,r}\cdot 2^{2r}\big] + \sum_{i\neq j}\mathsf{E}\big[\psi(f_i)\sum_{r_1=0}^{L} x_{i,r_1}\cdot 2^{r_1}\big]\mathsf{E}\big[\psi(f_j)\sum_{r_2=0}^{L} x_{j,r_2}\cdot 2^{r_2}\big]$$

As a result, the expression for $\mathsf{Var}\big[\bar{\Psi}\big]$ simplifies to

$$\mathsf{Var}\big[\bar{\Psi}\big] = \mathsf{E}\big[\bar{\Psi}^2\big] - \mathsf{E}\big[\bar{\Psi}\big]^2 = \sum_i \mathsf{E}\big[\psi^2(f_i)\sum_{r=0}^{L} x_{i,r}\cdot 2^{2r}\big] - \sum_i \mathsf{E}\big[\psi(f_i)\sum_{r=0}^{L} x_{i,r}\cdot 2^{r}\big]^2$$

$\mathsf{E}\big[\psi(f_i)\sum_{r=0}^{L} x_{i,r}\cdot 2^r\big]$ is given by Lemma 2. $\mathsf{E}\big[\psi^2(f_i)\sum_{r=0}^{L} x_{i,r}\cdot 2^{2r}\big]$ is calculated in an almost similar manner; we briefly outline the calculation. Let $i \in G_l$. We decompose groups into the left-region (lr), middle-region (mr) and right regions (rr) as in Lemma 2.

Suppose $i \in G_0 \setminus lr(G_0)$: Then, $1 - 2^{-t} < \mathsf{Pr}\{\hat{f}_i > T_0\} \leq 1$ and probability of i being classified into any other $G_r, r \neq 0$ is at most $2^{-t} \cdot \frac{1}{2^r}$. Therefore, $\sum_r \mathsf{E}\big[x_{i,r} 2^{2r}\big] < 1 + \sum_{r>0} 2^{-t} \cdot 2^r < 1 + 2^{-t+L+1}$.

Suppose $i \in lr(G_0)$: In this case, i can get classified into either $\bar{G}_0$ or $\bar{G}_1$, with probability at least $1 - 2^{-t}$. Given that i is classified into one of $\bar{G}_0$ or $\bar{G}_1$, let p be the conditional probability that $\mathsf{Pr}\{i \in \bar{G}_0 \mid i \in \bar{G}_0 \cup \bar{G}_1\}$. Therefore $\mathsf{Pr}\{x_{i,0}\} \leq p$, $\mathsf{Pr}\{x_{i,1}\} \leq \frac{(1-p)}{2}$ and $\mathsf{Pr}\{x_{i,r}\} \leq 2^{-t}$ for $r \notin \{0,1\}$. Therefore, $\sum_r \mathsf{E}\big[x_{i,r} 2^{2r}\big] < p + \frac{1-p}{2} \cdot 2^2 + \sum_{r>1} 2^{-t} \cdot 2^r < 2 + 2^{-t+L+1}$.

Suppose $i \in lr(G_l)$ for $l > 1$: The analysis for this case is similar to that of $i \in lr(G_0)$, except that $\mathsf{Pr}\{x_{i,l}\} \leq \frac{p}{2^l}$, $\mathsf{Pr}\{x_{i,l+1}\} \leq \frac{1-p}{2^{l+1}}$. Therefore, $\sum_r \mathsf{E}\big[x_{i,r} 2^{2r}\big] < \frac{p}{2^l} \cdot 2^{2l} + \frac{1-p}{2^{l+1}} \cdot 2^{2(l+1)} + \sum_{r \notin \{l,l+1\}} 2^{-t} \cdot 2^r < 2^{l+1} + 2^{-t+L+1}$.

Suppose $i \in mr(G_l)$ for $l > 1$: Such elements will be classified into $\bar{G}_l$ with probability $\geq 1 - 2^{-t}$, resulting in $\sum_r \mathsf{E}\big[x_{i,r} 2^{2r}\big] < \frac{1}{2^l} \cdot 2^{2l} + \sum_{r \neq l} 2^{-t} \cdot 2^r < 2^l + 2^{-t+L+1}$.

Suppose $i \in rr(G_l)$ for $l > 1$: Using an argument similar to that for $ll(G_l)$, we get $\sum_i \mathsf{E}\big[x_{i,r} 2^{2r}\big] < 2^l + 2^{-t+L+1}$. Combining the above cases, we obtain

$$\sum_i \mathsf{E}\big[\psi^2(f_i) \sum_{r=0}^{L} x_{i,r} \cdot 2^{2r}\big] \leq \sum_{i \in G_0 \setminus lr(G_0)} (1 + 2^{-t+L+1}) \cdot \psi^2(f_i) + \sum_{i \notin G_0 \setminus lr(G_0)} \psi^2(f_i) \cdot (2^{l+1} + 2^{-t+L+1}) \ .$$

In conjunction with Lemma 2, we get

$$\mathsf{Var}\big[\bar{\Psi}\big] \leq \sum_{i \in G_0 \setminus lr(G_0)} (1 + 2^{-t+L+1}) \cdot \psi^2(f_i) + \sum_{i \notin G_0 \setminus lr(G_0)} \psi^2(f_i) \cdot (2^{l+1} + 2^{-t+L+1}) - \sum_i (1 - 2^{-t+\log L}) \cdot \psi^2(f_i) \leq \sum_{i \in \mathcal{S}} 2^{-t+L+2} \cdot \psi^2(f_i) + \sum_{i \notin G_0 \setminus lr(G_0)} \psi^2(f_i) \cdot 2^{l+1}.$$

□