JMLR: Workshop and Conference Proceedings 19 (2011) 477–496 24th Annual Conference on Learning Theory

A New Algorithm for Compressed Counting with Applications in Shannon Entropy Estimation in Dynamic Data

Ping Li PINGLI@CORNELL.EDU
Department of Statistical Science, Cornell University, Ithaca, NY 14853

Cun-Hui Zhang CZHANG@STAT.RUTGERS.EDU
Department of Statistics and Biostatatistics, Rutgers University, New Brunswick, NJ 08901

Editor: Sham Kakade, Ulrike von Luxburg

Abstract

Efficient estimation of the moments and Shannon entropy of data streams is an important task in modern machine learning and data mining. To estimate the Shannon entropy, it suffices to accurately estimate the α-th moment with $\Delta = |1 - \alpha| \approx 0$. To guarantee that the error of estimated Shannon entropy is within a ν-additive factor, the method of *symmetric stable random projections* requires $O\left(\frac{1}{\nu^2\Delta^2}\right)$ samples, which is extremely expensive. The first paper (Li, 2009a) in *Compressed Counting (CC)*, based on *skewed-stable random projections*, supplies a substantial improvement by reducing the sample complexity to $O\left(\frac{1}{\nu^2\Delta}\right)$, which is still expensive. The followup work (Li, 2009b) provides a practical algorithm, which is however difficult to analyze theoretically.

In this paper, we propose a new accurate algorithm for Compressed Counting, whose sample complexity is only $O\left(\frac{1}{\nu^2}\right)$ for ν-additive Shannon entropy estimation. The constant factor for this bound is merely about 6. In addition, we prove that our algorithm achieves an upper bound of the Fisher information and in fact it is close to 100% statistically optimal. An empirical study is conducted to verify the accuracy of our algorithm.

Keywords: Data Streams, Entropy Estimation, Maximally-Skewed Stable Random Projections

1. Introduction

The problem of "scaling up for high dimensional data and high speed data streams" is among the "10 challenging problems in data mining research" (Yang and Wu, 2006). This paper is devoted to estimating entropy of data streams. Mining data streams in (e.g.,) 100 TB scale databases has become an important area of research, e.g., (Henzinger et al., 1999; Domeniconi and Gunopulos, 2001; Aggarwal et al., 2004; Muthukrishnan, 2005), as the Web and network data can easily reach that scale (Yang and Wu, 2006).

Consider the *Turnstile* stream model (Muthukrishnan, 2005). The input stream $a_t = (i_t, I_t)$, $i_t \in [1,\ D]$ arriving sequentially describes the underlying signal A, meaning

$$A_t[i_t] = A_{t-1}[i_t] + I_t, \tag{1}$$

where the increment I_t can be either positive (insertion) or negative (deletion). For example, in network measurements, I_t can be the increment of the packet size at the location numbered by i_t.

In the model (1), restricting $A_t[i] \geq 0$ results in the *strict-Turnstile* model, which suffices for describing almost all natural phenomena (Muthukrishnan, 2005). This paper focuses on efficient algorithms for estimating *α-th frequency moment* of data streams

$$F_{(\alpha)} = \sum_{i=1}^{D} A_t[i]^{\alpha}. \tag{2}$$

We are interested in the case of $\alpha \to 1$, which is crucial for the estimation of *Shannon entropy*. Note that the first moment (i.e., the sum) $F_{(1)} = \sum_{s=0}^{t} I_s$ can be computed using a single counter.

1.1. Entropy, Moments, and Estimation Complexity

A widely useful summary statistic is the *Shannon entropy*

$$H = -\sum_{i=1}^{D} \frac{A_t[i]}{F_{(1)}} \log \frac{A_t[i]}{F_{(1)}}. \tag{3}$$

There are various generalizations of the Shannon entropy. The Rényi entropy (Rényi, 1961), denoted by H_α, and the Tsallis entropy (Havrda and Charvát, 1967; Tsallis, 1988), denoted by T_α, are

$$H_\alpha = \frac{1}{1-\alpha} \log \frac{F_{(\alpha)}}{F_{(1)}^{\alpha}}, \qquad T_\alpha = \frac{1}{1-\alpha} \left(\frac{F_{(\alpha)}}{F_{(1)}^{\alpha}} - 1 \right). \tag{4}$$

As $\alpha \to 1$, both Rényi entropy and Tsallis entropy converge to Shannon entropy:

$$\lim_{\alpha \to 1} H_\alpha = \lim_{\alpha \to 1} T_\alpha = H. \tag{5}$$

Thus, both Rényi entropy and Tsallis entropy can be computed from the α-th frequency moment; and one can approximate Shannon entropy using $\alpha \approx 1$. While this fact is well-known, it appears that (Zhao et al., 2007) is the first study that applied (5) to Shannon entropy estimation in data streams. Later (Harvey et al., 2008b,a) proposed criteria on theoretically (and conservatively) how close to 1 the α needs to be. One can numerically verify $\Delta = |1 - \alpha| < 10^{-7}$ in (Harvey et al., 2008b) or $\Delta < 10^{-5}$ in (Harvey et al., 2008a) are very likely.[1]

The difficulty in Shannon entropy estimation is reflected by the estimation variance. By the definitions of the Rényi and Tsallis entropies, we need estimators of $F_{(\alpha)}$ with variances proportional to $O\left(\Delta^2\right)$ in order to cancel the term $\frac{1}{(1-\alpha)^2} = \frac{1}{\Delta^2}$ (otherwise the sample size must be proportional to $\frac{1}{\Delta^2}$). In other words, the estimators of $F_{(\alpha)}$ must be extremely accurate.

1. In (Harvey et al., 2008b), $\Delta = \frac{c}{16 \log(1/c)}$, $c = \frac{\nu}{4 \log(D) \log(m)}$, where m is the number of streaming updates. If we let $D = 2^{64}$, $m = 2^{64}$, $\nu = 0.1$, then $\Delta \approx 7 \times 10^{-8}$. If we let $m = 10^6$, $\nu = 0.1$, then $\Delta \approx 2.5 \times 10^{-7}$. Harvey et al. (2008a) provides some improvements, to allow slightly larger Δ, which is still extremely small.

1.2. Some Applications of Shannon Entropy

Real-Time Network Anomaly Detection Network traffic is a typical example of high-rate data streams. An effective measurement of network traffic in real-time is crucial for anomaly detection and network diagnosis; and one such measurement metric is the Shannon entropy (Feinstein et al., 2003; Lakhina et al., 2005; Xu et al., 2005; Brauckhoff et al., 2006; Lall et al., 2006; Zhao et al., 2007). The *Turnstile* data stream model (1) is naturally suitable for describing network traffic, especially when the goal is to characterize the statistical distribution of the traffic. In its empirical form, a statistical distribution is described by histograms, $A_t[i]$, $i = 1$ to D. It is possible that $D = 2^{64}$ (or larger) if one is interested in measuring the traffic streams of all unique sources or destinations.

The Distributed Denial of Service (**DDoS**) attack, as a representative example of network anomalies, attempts to make computers unavailable to intended users, either by forcing users to reset the computers or by exhausting the resources of service-hosting sites. Since a DDoS attack often changes the statistical distribution of network traffic, a common practice to detect such an attack is to monitor the network traffic using certain summary statistics. As the Shannon entropy is well-suited for characterizing a distribution, a popular detection method is to measure the time-history of entropy and alarm anomalies when the entropy becomes abnormal (Feinstein et al., 2003; Lall et al., 2006).

Entropy measurements do not have to be "perfect" for detecting attacks. It is, however crucial that the algorithm should be computationally efficient at low memory cost, because the traffic data generated by large high-speed networks are enormous and transient. Algorithms should be real-time and one-pass, as the traffic data are unlikely to be stored permanently. Many algorithms have been proposed for "sampling" the traffic streaming data for estimating entropy (Lall et al., 2006; Zhao et al., 2007; Bhuvanagiri and Ganguly, 2006; Guha et al., 2006; Chakrabarti et al., 2006, 2007; Harvey et al., 2008b,a; Zhao et al., 2010).

Entropy of Query Logs in Web Search Mei and Church (2008) proposed to estimate the Shannon entropy of some commercial search logs, to help answer some basic problems in Web search, such as, *how big is the web?* The search logs can be viewed as data streams, and Mei and Church (2008) analyzed several "snapshots" of a sample of the search logs, which contained 10 million <Query, URL, IP> triples; each triple corresponded to a click from a particular IP address on a particular URL for a particular query. (Mei and Church, 2008) drew their important conclusions on this (hopefully) representative sample. Alternatively, one could apply new data stream algorithms on the entire history of the search logs.

Entropy in Neural Computations A workshop in NIPS'03 was devoted to entropy estimation (`www.menem.com/~ilya/pages/NIPS03`), owing to the wide-spread use of entropy in neural computations (Paninski, 2003), e.g., for studying the underlying structure of spike trains.

Graph Estimation As demonstrated in a recent paper (Gupta et al., 2010), Shannon entropy estimation plays a crucial role in graph estimation and density estimation in high dimensions.

1.3. Symmetric Stable Random Projections and Prior Work on Compressed Counting

The problem of estimating $F_{(\alpha)}$ has been heavily studied since the pioneering work of (Alon et al., 1996). For $0 < \alpha \leq 2$, the method of *symmetric stable random projections* (Indyk, 2006; Li, 2008; Li and Hastie, 2007) in many applications provides a practical algorithm, with a sample complexity of $O\left(\frac{1}{\epsilon^2}\right)$ (even for $\alpha = 1$), to estimate $F_{(\alpha)}$ within a $1 \pm \epsilon$ multiplicative factor.

Compressed Counting (CC) (Li, 2009a,b) is a recent breakthrough, which is based on *maximally-skewed stable random projections*. Li (2009a) provided two algorithms, using the *geometric mean* and *harmonic mean*. The *geometric mean* algorithm has the variance proportional to $O(\Delta)$ in the neighborhood of $\alpha = 1$, where $\Delta = |1 - \alpha|$. This is the first algorithm that reflected the intuition that, in the neighborhood of $\alpha = 1$, the moment estimation algorithms should work better and better as $\alpha \to 1$, in a continuous fashion. The *geometric mean* algorithm for CC, unfortunately, did not provide an adequate mechanism for entropy estimation. It only led to an entropy estimation algorithm with a complexity of $O\left(\frac{1}{\nu^2\Delta}\right)$, but (theoretically) Δ has to be extremely small.

Based on the *geometric mean* algorithm of CC (Li, 2009a), Harvey et al. (2008a) developed a complicated multi-point method for Shannon entropy estimation, with a sample (word) complexity of $O\left(\frac{1}{\nu^2}\log M\right)$ and a very large (like 10^7) constant[2], where (e.g.,) $M = \sum_{i=1}^{D} |A_t[i]|$ can be viewed as the "universe size." In comparison, our new algorithm is very simple with a sample (word) complexity of $O\left(\frac{1}{\nu^2}\right)$ and a small constant (about 6), without the $\log M$ term.

Li (2009b) proposed a practical algorithm based on numerical optimization and achieved very good performance. Since that estimator was complicated and implicit, Li (2009b) did not analyze the sample complexity and statistical efficiency and left them as open problems.

1.4. Another Perspective for Entropy Estimation

By the definition of Rényi entropy (4), instead of estimating $F_{(\alpha)}$, it suffices to estimate $J_{(\alpha)}$, where

$$J_{(\alpha)} = F_{(\alpha)}^{-1/\Delta} = \left[\sum_{i=1}^{D} A_t[i]^\alpha\right]^{-1/\Delta}, \tag{6}$$

because, if $\Delta = 1 - \alpha > 0$, then

$$H_\alpha = \frac{1}{1-\alpha}\log\frac{F_{(\alpha)}}{F_{(1)}^\alpha} = \frac{1}{\Delta}\log\frac{J_{(\alpha)}^{-\Delta}}{F_{(1)}^\alpha} = -\log J_{(\alpha)} - \frac{1}{\Delta}\log F_{(1)}^\alpha. \tag{7}$$

Since $\frac{1}{\Delta}\log F_{(1)}^\alpha$ is computed exactly, we only need to estimate $J_{(\alpha)}$. Our new algorithm will provide a ν-multiplicative estimate of $J_{(\alpha)}$ with a complexity of $O\left(\frac{1}{\nu^2}\right)$. For small ν, this translates into:

2. In Sec. 5.2 of (Harvey et al., 2008a), their sample complexity bound is $O\left(\left[200(z+1)^3\right]^2\frac{1}{\nu^2}\log M\right)$, where $z = \log(1/\nu) + \log\log M$. The constant $\left[200(z+1)^3\right]^2$ will exceed 10^7, even just for $z = 3$.

1. An $\epsilon = \nu\Delta$-multiplicative estimate of $F_{(\alpha)}$, with a sample complexity of $O\left(\frac{1}{\nu^2}\right)$. For example, denote the estimate of $J_{(\alpha)}$ by $\hat{J}_{(\alpha)}$, then

$$\mathbf{Pr}\left(\hat{J}_{(\alpha)} \geq (1+\nu)J_{(\alpha)}\right) = \mathbf{Pr}\left(\hat{J}_{(\alpha)}^{-\Delta} \leq (1+\nu)^{-\Delta}F_{(\alpha)}\right)$$
$$\mathbf{Pr}\left(\hat{J}_{(\alpha)} \leq (1-\nu)J_{(\alpha)}\right) = \mathbf{Pr}\left(\hat{J}_{(\alpha)}^{-\Delta} \geq (1-\nu)^{-\Delta}F_{(\alpha)}\right).$$

   For small ν, we have $(1+\nu)^{-\Delta} \approx 1-\nu\Delta = 1-\epsilon$ and $(1-\nu)^{-\Delta} \approx 1+\nu\Delta = 1+\epsilon$.

2. A ν-additive estimate of $\log J_{(\alpha)}$, with a sample complexity of $O\left(\frac{1}{\nu^2}\right)$. For example

$$\mathbf{Pr}\left(\hat{J}_{(\alpha)} \geq (1+\nu)J_{(\alpha)}\right) = \mathbf{Pr}\left(\log\hat{J}_{(\alpha)} \geq \log(1+\nu) + \log J_{(\alpha)}\right)$$
$$\mathbf{Pr}\left(\hat{J}_{(\alpha)} \leq (1-\nu)J_{(\alpha)}\right) = \mathbf{Pr}\left(\log\hat{J}_{(\alpha)} \leq \log(1-\nu) + \log J_{(\alpha)}\right).$$

   For small ν, we have $\log(1+\nu) \approx \nu$ and $\log(1-\nu) \approx -\nu$.

2. The Proposed Algorithm

Consider the *strict-Turnstile* data stream model (1). Conceptually, we multiply the data stream vector $A_t \in \mathbb{R}^D$ by a random matrix $\mathbf{R} \in \mathbb{R}^{D\times k}$, resulting in a vector $X = A_t \times \mathbf{R} \in \mathbb{R}^k$ with entries

$$x_j = [A_t \times \mathbf{R}]_j = \sum_{i=1}^{D} r_{ij}A_t[i], \quad j = 1, 2, ..., k$$

where r_{ij}'s are random variables generated as follows:

$$r_{ij} = \frac{\sin(\alpha v_{ij})}{[\sin v_{ij}]^{1/\alpha}}\left[\frac{\sin(v_{ij}\Delta)}{w_{ij}}\right]^{\frac{\Delta}{\alpha}}, \qquad \Delta = 1-\alpha > 0, \tag{8}$$

where $v_{ij} \sim Uniform(0,\pi)$ (i.i.d.) and $w_{ij} \sim Exp(1)$ (i.i.d.), an exponential distribution with mean 1. In data stream computations, the matrix $\mathbf{R}$ is not materialized. The standard procedure is to (re)generate entries of $\mathbf{R}$ on-demand (Indyk, 2006). Whenever a stream element $a_t = (i_t, I_t)$ arrives, one updates entries of X:

$$x_j \leftarrow x_j + I_t r_{i_t j}, \quad j = 1, 2, ..., k.$$

The cost of (re)generating (pseudo) random numbers is proportional to the sample size k. As our work substantially reduces the sample size, it also tremendously reduces the processing time.

Here, our goal is to estimate $J_{(\alpha)} = F_{(\alpha)}^{-1/\Delta}$ (and hence also $F_{(\alpha)}$). Our proposed algorithm is

$$\hat{J}_{(\alpha)} = \frac{\Delta}{k}\sum_{j=1}^{k} x_j^{-\alpha/\Delta}, \tag{9}$$

from which one can estimate the Shannon entropy, for example, by the Rényi entropy as

$$\hat{H}_\alpha = -\log \hat{J}_{(\alpha)} - \frac{1}{\Delta}\log F^\alpha_{(1)} \tag{10}$$

The following Lemma provides the moments of $\hat{J}_{(\alpha)}$.

Lemma 1

$$E\left(\hat{J}_{(\alpha)}\right) = J_{(\alpha)}, \tag{11}$$

$$Var\left(\hat{J}_{(\alpha)}\right) = \frac{J^2_{(\alpha)}}{k}\left(3-2\Delta\right), \tag{12}$$

$$E\left(\hat{J}_{(\alpha)} - J_{(\alpha)}\right)^3 = \frac{J^3_{(\alpha)}}{k^2}\left(17-21\Delta+6\Delta^2\right), \tag{13}$$

$$E\left(\hat{J}_{(\alpha)} - J_{(\alpha)}\right)^4 = 3\frac{J^4_{(\alpha)}}{k^2}(3-2\Delta)^2 + \frac{J^4_{(\alpha)}}{k^3}\left(142-252\Delta+140\Delta^2-24\Delta^3\right). \tag{14}$$

□

The first two moments immediately imply that the sample complexity of $\hat{J}_{(\alpha)}$ is $O\left(\frac{1}{\nu^2}\right)$ for a ν-multiplicative approximation of $J_{(\alpha)}$. The higher moments in Lemma 1 are also useful for the proof of Lemma 10.

The next Lemma provides the precise tail bounds.

Lemma 2 *1. The right tail bound: for $\nu > 0$,*

$$\mathbf{Pr}\left(\hat{J}_{(\alpha)} \geq (1+\nu)J_{(\alpha)}\right) \leq \exp\left(-k\frac{\nu^2}{G_R}\right) \tag{15}$$

$$\frac{\nu^2}{G_R} = -\log\left(1+\sum_{n=1}^{\infty} t_R^n e^{nH(n;\Delta)}\right) + t_R(1+\nu) \tag{16}$$

where t_R is the solution to

$$-\frac{\sum_{n=1}^{\infty} n t_R^{n-1} e^{nH(n;\Delta)}}{1+\sum_{n=1}^{\infty} t_R^n e^{nH(n;\Delta)}} + (1+\nu) = 0 \tag{17}$$

and

$$H\left(n;\Delta\right) = \prod_{i=0}^{n-1}\frac{n-i\Delta}{e(n-i)} \tag{18}$$

2. The left tail bound: for $0 < \nu < 1$,

$$\mathbf{Pr}\left(\hat{J}_{(\alpha)} \leq (1-\nu)J_{(\alpha)}\right) \leq \exp\left(-k\frac{\nu^2}{G_L}\right) \tag{19}$$

$$\frac{\nu^2}{G_L} = -\log\left(1+\sum_{n=1}^{\infty}(-t_L)^n e^{nH(n;\Delta)}\right) - t_L(1-\nu) \tag{20}$$

where t_L is the solution to

$$\frac{\sum_{n=1}^{\infty}(-1)^n n(t_L)^{n-1}e^n H(n;\Delta)}{1+\sum_{n=1}^{\infty}(-t_L)^n e^n H(n;\Delta)} + (1-\nu) = 0. \tag{21}$$

□

While the expressions for the tail bounds in Lemma 2 appear sophisticated, they are carefully formulated so that they can be accurately evaluated numerically; see Figure 1. The function $H(n;\Delta)$ in (18) approaches e^{-n} when $\Delta \to 1$, and it is always upper bounded by $\frac{1}{\sqrt{2\pi n}}$ even as $\Delta \to 0$, since

$$\prod_{i=0}^{n-1}\frac{n-i\Delta}{n-i} \leq \frac{n^n}{n!} \leq \frac{n^n}{(n-1)!} \leq \frac{e^n}{\sqrt{2\pi n}}$$

according to Stirling's series (Gradshteyn and Ryzhik, 1994, 8.327)

$$\Gamma(n) = (n-1)! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n\left[1+\frac{1}{12n}+\frac{1}{288n^2}-\frac{139}{51840n^3}-...\right].$$

Interestingly, we can obtain closed-form expressions when $\nu \to 0$.

Lemma 3 *As $\nu \to 0$, the constants G_R and G_L in (16) and (20), respectively, become*

$$G_R \to 6-4\Delta, \quad G_L \to 6-4\Delta. \tag{22}$$

□

In addition, when $\Delta \to 1-$, we can actually analytically express the tail bounds in closed-forms.

Lemma 4 *When $\Delta \to 1-$, i.e., $\alpha \to 0+$,*

$$\frac{\nu^2}{G_R} = -\log(1+\nu)+\nu, \quad \nu > 0 \tag{23}$$

$$\frac{\nu^2}{G_L} = -\log(1-\nu)-\nu, \quad 0 < \nu < 1. \tag{24}$$

□

We summarize the complexity bound of our algorithm in a theorem.

Theorem 5 *The proposed algorithm $\hat{J}_{(\alpha)}$ in (9) provides a ν-multiplicative approximation of $J_{(\alpha)}$ and a ν-additive approximation of the Shannon entropy with a probability at least $1-\delta$, using $\frac{C}{\nu^2}\log 2/\delta$ samples (words). The constant C approaches $6-4\Delta$ as $\nu \to 0$.*□

3. More Intuition and Explanation

The proposed algorithm (9) is based on the idea of *maximally-skewed stable random projections*.

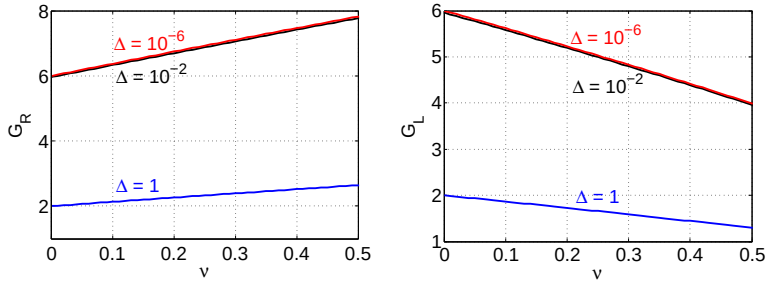

Figure 1: Numerical values of G_R (left panel) and G_L (right panel) in the tail bounds (16) and (20), for $\Delta = 10^{-2}$ and $\Delta = 10^{-6}$, together with the closed-form expressions for $\Delta = 1$ as obtained in Lemma 4. Note that as $\nu \to 0$, both G_R and G_L approach $6 - 4\Delta$, as proved in Lemma 3.

3.1. Review Maximally-Skewed Stable Random Projections and Estimators

The method for sampling from skewed stable distributions was proposed by Chambers et al. (1976). To sample from $S(\alpha, \beta = 1, 1)$, i.e., α-stable maximally-skewed ($\beta = 1$) with unit scale, one first generates an exponential random variable with mean 1, $W \sim Exp(1)$, and a uniform random variable $U \sim Uniform\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$. Then the following nonlinear transformation of W and U results in the desired random variable:

$$Z' = \frac{\sin\left(\alpha(U+\rho)\right)}{\left[\cos U \cos\left(\rho\alpha\right)\right]^{1/\alpha}} \left[\frac{\cos\left(U - \alpha(U+\rho)\right)}{W}\right]^{\frac{1-\alpha}{\alpha}} \sim S(\alpha, \beta = 1, 1), \tag{25}$$

where $\rho = \frac{\pi}{2}$ when $\alpha < 1$ and $\rho = \frac{\pi}{2}\frac{2-\alpha}{\alpha}$ when $\alpha > 1$. Note that $\cos\left(\frac{\pi}{2}\alpha\right) \to 0$ as $\alpha \to 1$. For convenience (and to avoid numerical problems), we use

$$Z = Z' \cos^{1/\alpha}\left(\rho\alpha\right) \sim S\left(\alpha, \beta = 1, \cos\left(\rho\alpha\right)\right).$$

It turns out, the random variable Z with $\alpha < 1$ has good properties. This study only considers $\alpha = 1 - \Delta < 1$, i.e, $\rho = \frac{\pi}{2}$. After simplification, we obtain

$$Z = \frac{\sin\left(\alpha V\right)}{\left[\sin V\right]^{1/\alpha}} \left[\frac{\sin\left(V\Delta\right)}{W}\right]^{\frac{\Delta}{\alpha}}, \tag{26}$$

where $V = \frac{\pi}{2} + U \sim Uniform(0, \pi)$. This explains (8).

Let $X = A_t \times \mathbf{R}$, where entries of $\mathbf{R}$ are i.i.d. samples of $S\left(\alpha, \beta = 1, \cos\left(\frac{\pi}{2}\alpha\right)\right)$. Then by properties of stable distributions, the entries of X are

$$x_j = [A_t \times \mathbf{R}]_j = \sum_{i=1}^{D} r_{i,j} A_t[i] \sim S\left(\alpha, \beta = 1, \cos\left(\frac{\pi}{2}\alpha\right) F_{(\alpha)}\right),$$

where $F_{(\alpha)} = \sum_{i=1}^{D} A_t[i]^\alpha$ as defined in (2). Li (2009a) provided two algorithms using on the *geometric mean* and *harmonic mean* estimators, based on the following basic moment formula.

Lemma 6 *(Li, 2009a).* *If $X \sim S(\alpha, \beta = 1, F_{(\alpha)} \cos\left(\frac{\alpha\pi}{2}\right))$, then $X > 0$, and for any $-\infty < \lambda < \alpha < 1$,*

$$E\left(X^\lambda\right) = F_{(\alpha)}^{\lambda/\alpha} \frac{\Gamma\left(1 - \frac{\lambda}{\alpha}\right)}{\Gamma\left(1 - \lambda\right)}.$$

□

3.1.1. The Geometric Mean Estimator

Assume x_j, $j = 1$ to k, are i.i.d. samples from $S(\alpha, \beta = 1, F_{(\alpha)} \cos\left(\frac{\alpha\pi}{2}\right))$. After simplifying the corresponding expression in (Li, 2009a), we obtain

$$\hat{F}_{(\alpha),gm} = \left[\frac{\Gamma\left(1 - \frac{\alpha}{k}\right)}{\Gamma\left(1 - \frac{1}{k}\right)}\right]^k \prod_{j=1}^{k} x_j^{\alpha/k}, \tag{27}$$

which is unbiased and has asymptotic variance

$$\mathrm{Var}\left(\hat{F}_{(\alpha),gm}\right) = \frac{F_{(\alpha)}^2}{k} \frac{\pi^2}{6} \Delta\left(1 + \alpha\right) + O\left(\frac{1}{k^2}\right) \tag{28}$$

As $\Delta = 1 - \alpha \to 0$, the asymptotic variance approaches zero at the rate of only $O\left(\Delta\right)$ (not $O\left(\Delta^2\right)$).

3.1.2. The Harmonic Mean Estimator

$$\hat{F}_{(\alpha),hm} = \frac{k \frac{1}{\Gamma(1+\alpha)}}{\sum_{j=1}^{k} x_j^{-\alpha}} \left(1 - \frac{1}{k}\left(\frac{2\Gamma^2(1+\alpha)}{\Gamma(1+2\alpha)} - 1\right)\right), \tag{29}$$

which is asymptotically unbiased and has variance

$$\mathrm{Var}\left(\hat{F}_{(\alpha),hm}\right) = \frac{F_{(\alpha)}^2}{k}\left(\frac{2\Gamma^2(1+\alpha)}{\Gamma(1+2\alpha)} - 1\right) + O\left(\frac{1}{k^2}\right). \tag{30}$$

3.2. Limitations of the Geometric Mean and Harmonic Mean Estimators

In order to estimate the Shannon entropy with a guaranteed ν-additive accuracy, the variance of the estimator of $F_{(\alpha)}$ should be $O\left(\Delta^2\right)$; or equivalently, the sample complexity should be $O\left(\frac{1}{\nu^2}\right)$.

The geometric mean estimator has variance proportional to only $O\left(\Delta\right)$; or equivalently, its complexity is $O\left(\frac{1}{\nu^2\Delta}\right)$, where Δ needs to be extremely small (e.g., $< 10^{-5}$). For the harmonic mean estimator in Li (2009a), the following Lemma says its variance is also proportional to $O\left(\Delta\right)$.

Lemma 7 *As* $\Delta = 1 - \alpha \to 0$,

$$\frac{2\Gamma^2(1+\alpha)}{\Gamma(1+2\alpha)} - 1 = \Delta + \Delta^2\left(2 - \frac{\pi^2}{6}\right) + O\left(\Delta^3\right). \tag{31}$$

□

In other words, the harmonic mean estimator improves the geometric mean estimator by reducing the variance by a factor of $\frac{\pi^2}{6}2 = 3.29$. Thus, we must develop significantly better algorithms.

3.3. The Distribution Function

This section provides the distribution function of $Z \sim S\left(\alpha < 1, \beta = 1, \cos\left(\frac{\pi}{2}\alpha\right)\right)$, which will be needed for better understanding the proposed estimator (9).

Lemma 8 *Suppose a random variable* $Z \sim S\left(\alpha < 1, \beta = 1, \cos\left(\frac{\pi}{2}\alpha\right)\right)$. *The cumulative distribution function (CDF) is*

$$F_Z(t) = \mathbf{Pr}\left(Z \leq t\right) = \frac{1}{\pi}\int_0^\pi \exp\left(-t^{-\alpha/\Delta} g\left(\theta;\Delta\right)\right) d\theta. \tag{32}$$

where

$$g(\theta;\Delta) = \frac{\left[\sin\left(\alpha\theta\right)\right]^{\alpha/\Delta}}{\left[\sin\theta\right]^{1/\Delta}} \sin\left(\theta\Delta\right), \qquad \theta \in (0,\pi)$$
$$g\left(0+;\Delta\right) = \lim_{\theta\to 0+} g(\theta;\Delta) = \Delta\alpha^{\alpha/\Delta}.$$

□

Note that $g\left(0+;\Delta\right) = \Delta\alpha^{\alpha/\Delta} \approx \Delta e^{-1}$ approaches zero as $\Delta \to 0$. Thus, one might be wondering if we replace $g\left(\theta;\Delta\right)$ by $g\left(0+;\Delta\right)$, the errors may be quite small, as seen in Figure 2.

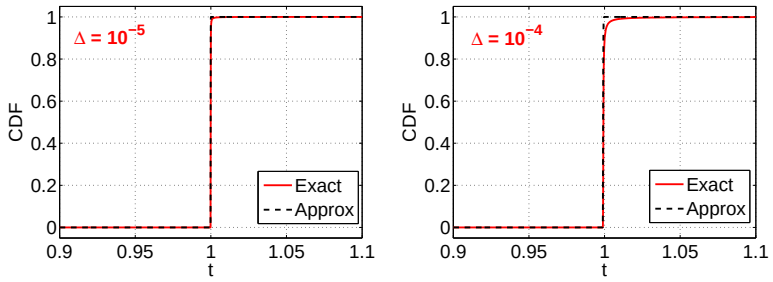

Figure 2: We plot the CDF curves as derived in Lemma 8, for $\Delta = 10^{-5}$ and 10^{-4}. As $\Delta \to 0$, the exact CDF (solid curves) is very close to the approximate CDF (dashed curves), which we obtain by replacing the exact $g(\theta; \Delta)$ function in Lemma 8 with the limit $g(0+; \Delta)$.

3.4. One Intuition Behind the Proposed Algorithm

The difficulty in developing accurate algorithms lies in that F_Z in (32) has no closed-form expression. From Lemma 8 and Figure 2, it appears that if one replaces the exact $g(\theta; \Delta)$ with its approximation $g(0+; \Delta)$, the error may be small. Thus, we consider a random variable Y with CDF

$$F_Y(t) = \mathbf{Pr}\left(y \le t\right) = \exp\left(-t^{-\alpha/\Delta}\Delta\alpha^{\alpha/\Delta}\right), \qquad t \in [0, \infty). \tag{33}$$

It is indeed a CDF because it is an increasing function of $t \in [0, \infty)$, $F_Y(0) = 0$, and $F_Y(\infty) = 1$.

Here, we are interested in estimating c^α from k i.i.d. samples $x_j = cy_j$, $j = 1$ to k. Statistics theory tells us that the maximum likelihood estimator (MLE) achieves the (asymptotic) optimality. Because F_Y has a closed-form expression, we can compute the MLE exactly.

Lemma 9 *Suppose y_j, $j = 1$ to k, are i.i.d. samples from a distribution whose CDF is given by (33). Let $x_j = cy_j$, where $c > 0$. Then the maximum likelihood estimator of c^α is given by*

$$\frac{1}{\Delta^\Delta \alpha^\alpha}\left[\frac{k}{\sum_{j=1}^k x_j^{-\alpha/\Delta}}\right]^\Delta. \tag{34}$$

□

In comparison, our proposed algorithm for estimating $J_{(\alpha)} = F_{(\alpha)}^{-1/\Delta}$ is defined in (9), which provides an estimator of $F_{(\alpha)}$:

$$\hat{F}_{(\alpha)} = \frac{1}{\Delta^\Delta} \left[\frac{k}{\sum_{j=1}^k x_j^{-\alpha/\Delta}} \right]^\Delta, \tag{35}$$

which is almost identical to (34). Note that, as $\Delta \to 0$, the extra term in (34), $\alpha^\alpha \to 1$, converges much faster than $\Delta^\Delta \to 1$. In other words, α^α is negligible.

Therefore, we should expect that our proposed estimator (35) is actually very close to the true MLE, even though we can not explicitly derive the MLE. Indeed, in the next section, Lemma 11 says that our algorithm is close to be 100% statistically optimal.

4. Additional Technical Results

4.1. The Moments of $\hat{F}_{(\alpha)}$

The following Lemma analyzes the mean square error: $\text{MSE} = E\left[\hat{F}_{(\alpha)} - F_{(\alpha)}\right]^2 = Var\left(\hat{F}_{(\alpha)}\right) + \text{Bias}^2$.

Lemma 10 *The estimator $\hat{F}_{(\alpha)}$ is asymptotically unbiased:*

$$E\left(\hat{F}_{(\alpha)}\right) = F_{(\alpha)}\left(1 + O\left(\frac{\Delta}{k}\right)\right). \tag{36}$$

The mean square error (MSE) is

$$E\left[\hat{F}_{(\alpha)} - F_{(\alpha)}\right]^2 = \frac{F_{(\alpha)}^2}{k}\Delta^2\left((3-2\Delta) + O\left(\frac{1}{k}\right)\right). \tag{37}$$

More precisely

$$0 \le E\left(\hat{F}_{(\alpha)} - F_{(\alpha)}\right) \le \frac{\Delta F_{(\alpha)}}{k} e^{2+\Delta}\left((1+\Delta)(3-2\Delta)/2 + \frac{k}{k-\Delta}\right). \tag{38}$$

and

$$\left| E\left(\frac{\hat{F}_{(\alpha)}}{F_{(\alpha)}} - 1\right)^2 - \frac{\Delta^2}{k}(3-2\Delta) - \frac{\Delta^2}{k^2}C_3^*(\Delta) \right| \le \frac{\Delta^2 C_4^*(\Delta)}{4k^2}(3-2\Delta)^2 + O(\Delta^2 k^{-3}\log k) \tag{39}$$

where $C_3^*(\Delta) = (1+\Delta)(17-21\Delta+6\Delta^2) = 17 + O(\Delta)$ and $C_4^*(\Delta) = e^{4+2\Delta}(11+18\Delta+7\Delta^2) + e^{5+2\Delta}(6+11\Delta+6\Delta^2+\Delta^3) = 11e^4 + 6e^5 + O(\Delta)$. □

4.2. Statistical Optimality

Recall we have k i.i.d. samples $x_j \sim S\left(\alpha, \beta = 1, \cos\left(\frac{\pi}{2}\alpha\right) F_{(\alpha)}\right)$. The goal is to estimate $J_{(\alpha)} = F_{(\alpha)}^{-1/\Delta}$. The classical theory of the Cramér-Rao lower bound tells us that the variance of the estimator is lower bounded by $\frac{1}{k}\frac{1}{I(J_{(\alpha)})}$, where $I(J_{(\alpha)})$ is the Fisher Information of $J_{(\alpha)}$.

A natural question is how much more improvement can we expect, after we have developed the estimator $\hat{J}_{(\alpha)}$ (9), whose variance is $\frac{J_{(\alpha)}^2}{k}(3 - 2\Delta)$? Lemma 11 provides the answer.

Lemma 11 *For a distribution $S\left(\alpha, \beta = 1, \cos\left(\frac{\pi}{2}\alpha\right) F_{(\alpha)}\right)$, the Fisher Information of $J_{(\alpha)} = F_{(\alpha)}^{-1/\Delta}$ is given by*

$$I(J_{(\alpha)}) = \frac{1}{J_{(\alpha)}^2}\left(I_2 - 1\right), \qquad I_2 = \int_0^\infty \frac{\left[\frac{1}{\pi}\int_0^\pi s g^2 e^{-sg} d\theta\right]^2}{\frac{1}{\pi}\int_0^\pi g e^{-sg} d\theta} ds \tag{40}$$

where $g = g(\theta; \Delta) = \frac{[\sin(\alpha\theta)]^{\alpha/\Delta}}{[\sin\theta]^{1/\Delta}} \sin(\theta\Delta)$. The Fisher Information of $F_{(\alpha)}$ is given by

$$I(F_{(\alpha)}) = \frac{1}{\Delta^2 F_{(\alpha)}^2}\left(I_2 - 1\right). \tag{41}$$

Furthermore, I_2 is bounded by $I_2 \leq 2$. Therefore the following bounds hold:

$$I\left(J_{(\alpha)}\right) \leq \frac{1}{J_{(\alpha)}^2}, \qquad I\left(F_{(\alpha)}\right) \leq \frac{1}{\Delta^2 F_{(\alpha)}^2}. \tag{42}$$

□

The Fisher information bounds (42) suggest that the optimal estimator (if one can find it) of $J_{(\alpha)}$ (or $F_{(\alpha)}$) exhibits variance of at least $\frac{J_{(\alpha)}^2}{k}$ (or $\frac{F_{(\alpha)}^2}{k}\Delta^2$). In this sense, our proposed estimator is statistically optimal (up to a constant factor) in the framework of CC. Furthermore, the integral I_2 in (40) can be numerically evaluated. Figure 3 plots $\frac{1}{I_2 - 1}$ (dashed curve) and $3 - 2\Delta$ (solid curve). Our proposed estimator is close to be 100% optimal and hence there is little room for improvement.

5. Experiments

This section demonstrates that the proposed estimator $\hat{J}_{(\alpha)}$ in (9) is a practical algorithm, while the previously proposed geometric mean algorithm (Li, 2009a) is inadequate for entropy estimation. We also demonstrate that algorithms based on *symmetric stable random projections* (Indyk, 2006; Li, 2009a; Li and Hastie, 2007) are not suitable for entropy estimation in practice. Note that Lemma 7 has shown that the harmonic mean algorithm proposed in (Li, 2009a) is only 3.29-fold better than the geometric mean algorithm and hence it makes no essential difference for entropy estimation.

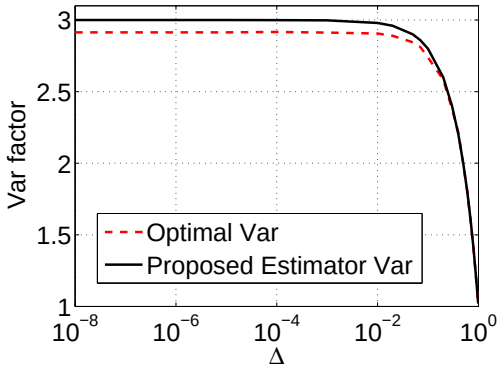

Figure 3: Dashed (red) curve: $\frac{1}{I_2-1}$ as in (40). Solid (black) curve: $3-2\Delta$.

5.1. Data

Since the estimation accuracy is what we are interested in, we simply use static data instead of real data streams, because the projected data vector $X = \mathbf{R}^{\mathrm{T}} A_t$ is the same at the end of the stream, regardless of whether it is computed at once (i.e., static) or incrementally (i.e., dynamic). As summarized in Table 1, 8 English words are selected from a chunk of Web crawl data, i.e., 8 vectors whose entries are the numbers of word occurrences in each document. The words are selected fairly randomly, although we make sure they cover a wide range of data sparsity, from function words (e.g., "A"), to common words (e.g., "FRIDAY") to rare words (e.g., "TWIST").

5.2. Estimating Shannon Entropies

We used the estimated frequency moments to estimate the Shannon entropies. For the data vector "TWIST", we present the results at sample sizes $k = 3, 10, 100, 1000$, and 10000. For all other vectors, we did not use $k = 10000$. Figure 4 presents the normalized mean square errors (MSEs).

Using our proposed algorithm (middle panels), only $k = 10$ samples already produces fairly accurate estimates. In fact, for some vectors (such as "A"), even $k = 3$ may provide reasonable estimates. We believe the performance of the new estimator is remarkable. Another nice property is that the estimation errors become stable after (e.g.,) $\Delta < 10^{-3}$ (or 10^{-4}). This essentially frees practitioners from specifying Δ.

Table 1: The data set consists of 8 English words selected from a corpus of Web pages, forming 8 vectors whose values are the word occurrences. The table lists their fractions of non-zeros (sparsity) and the Shannon entropies (H). The last column is the variance ratio for comparing CC with another algorithm named CRS; the details are in Section 6.

Word	Sparsity	Entropy H	Improvement over CRS
TWIST	0.004	5.4873	2.1
FRIDAY	0.034	7.0487	38.9
FUN	0.047	7.6519	23.1
BUSINESS	0.126	8.3995	48.7
NAME	0.144	8.5162	65.9
HAVE	0.267	8.9782	67.7
THIS	0.423	9.3893	84.4
A	0.596	9.5463	113.7

In comparison, the performance of the *geometric mean* algorithm (left panels) is not satisfactory. This is because its variance decreases only at the rate of $O(\Delta)$, not $O(\Delta^2)$. Also clearly, using *symmetric stable random projections* (right panels) would not provide good estimates of the Shannon entropy (unless the sample size is extremely large with a carefully chosen Δ).

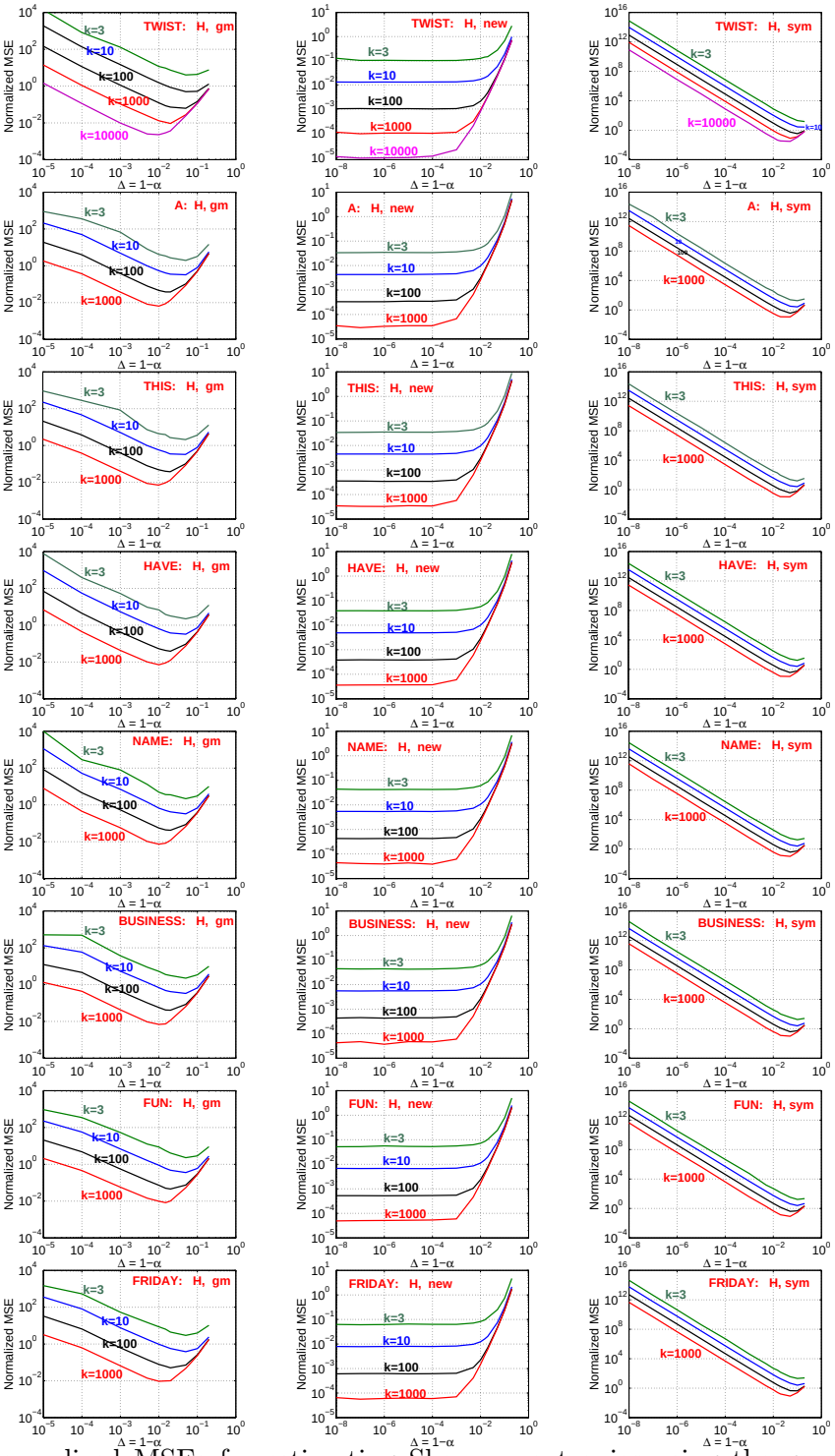

Figure 4: Normalized MSEs for estimating Shannon entropies using the geometric mean algorithm (left panels) proposed in (Li, 2009a), the proposed new algorithm $\hat{J}_{(\alpha)}$ (9) (middle panels) in this paper, and the geometric mean algorithm for *symmetric stable random projections* (right panels) in (Li, 2008).

6. Comparisons with Conditional Random Sampling (CRS)

Conditional Random Sampling (CRS), which is applicable to data stream computations, is another randomized algorithm (Li and Church, 2007; Li et al., 2008) particularly designed for sampling from sparse data. The significant advantage of CRS is that the method is "one sketch for all," meaning that the same set of "sketches" can be used to estimate a very wide range of summary statistics and distances including histograms, cross-entropy, χ^2 distances, inner products, general l_α distances (for any α). In comparison, the methods of (symmetric and skewed) α-stable random projections are generally limited to $0 < \alpha \leq 2$ and one has to re-do the projections (and keep multiple sets of samples) if the application requires to use multiple α values.[3] A recent manuscript (Zhao et al., 2010) compared CRS with a variety of other algorithms on the network data provided by ATT Labs.

It is interesting to compare CC (using the new estimator in this paper) with CRS for estimating Shannon entropy. Suppose we use the estimator (10) with sufficiently small Δ. Then the estimation variance is roughly just $\frac{3}{k}$, essentially independent of the original data. Using the generic approximate variance formula in Li et al. (2008), the variance for entropy estimation is denoted by $Var\left(\hat{H}_{CRS}\right)$:

$$Var\left(\hat{H}_\alpha\right) \approx \frac{3}{k} + O\left(\frac{1}{k^2}\right), \qquad \text{for sufficiently small } \Delta \tag{43}$$

$$Var\left(\hat{H}_{CRS}\right) \approx \frac{|\{i|A_t[i] > 0\}|}{k} \left\{ \sum_{i=1}^{D} \left[\frac{A_t[i]}{F_{(1)}} \log \frac{A_t[i]}{F_{(1)}}\right]^2 - \frac{1}{D}\left[\sum_{i=1}^{D} \frac{A_t[i]}{F_{(1)}} \log \frac{A_t[i]}{F_{(1)}}\right]^2 \right\} + O\left(\frac{1}{k^2}\right). \tag{44}$$

Table 1 (last column) already presents the variance ratios: $\frac{Var\left(\hat{H}_{CRS}\right)}{Var\left(\hat{H}_\alpha\right)}$ for the data used in our experiments. The ratios range from 2.1 to 113.7. The comparison further conforms that CC is extremely accurate for entropy estimation. On the other hand, CRS is actually also pretty good for entropy estimation, considering it is "one-sketch-for-all." Another significant advantage of CRS is that it is not limited to the strict-Turnstile data stream model, or even the general Turnstile model. It is particularly useful when applications require using nonlinearly transformed data (e.g., TF-IDF weighting in search and natural language processing) instead of the original data.

3. We should mention that the method of normal (l_2) random projections was recently extended (Li et al., 2010) to estimating l_α distances for $\alpha = 4, 6, 8, ...$ in massive (static) data matrices.

7. Conclusion

Many machine learning (e.g., neural computation, graph estimation) and data mining (e.g., anomaly detection) problems require estimating the Shannon entropy. When the data are dynamic (e.g., data streams), efficient estimation of the Shannon entropy using small space has been a challenging problem. It is known that we can approximate the Shannon entropy using the α-th frequency moment of the stream with α very close to 1, if the estimator of the moment is accurate enough with variance proportional to $O(\Delta^2)$, where $\Delta = |1 - \alpha|$. Our paper provides such a practical estimator. Our method is an ideal solution to the problem of entropy estimation when the data streams follow the strict-Turnstile model.

For ν-additive Shannon entropy estimation, the sample complexity of the algorithm is only $O\left(\frac{1}{\nu^2}\right)$. The constant factor for this bound is merely about 6. In addition, we prove that our algorithm achieves an upper bound of the Fisher information and in fact it is close to 100% statistically optimal. An empirical study is also conducted to verify the accuracy of our algorithm.

Further research: To further reduce the processing cost in order to better accommodate high-rate data streams, it is desirable to replace the dense matrix of skewed stable variables by a sparse matrix of Pareto-type variables. This is closely related to the prior study of *very sparse symmetric stable random projections* (Li, 2007). However, the extension to CC requires further work.

Acknowledgments

The authors thank the anonymous reviewers for their constructive comments. Ping Li's research is partially supported by the National Science Foundation (DMS-0808864), the Office of Naval Research (YIP-N000140910911), and a gift from Google. Cun-Hui Zhang's research is partially supported by the National Science Foundation (DMS-0804626, DMS-0906420) and the National Security Agency (H98230-09-1-0006, H98230-11-1-0205).

References

Charu C. Aggarwal, Jiawei Han, Jianyong Wang, and Philip S. Yu. On demand classification of data streams. In *KDD*, pages 503–508, Seattle, WA, 2004.

Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. In *STOC*, pages 20–29, Philadelphia, PA, 1996.

Lakshminath Bhuvanagiri and Sumit Ganguly. Estimating entropy over data streams. In *ESA*, pages 148–159, Zurich, Switzerland, 2006.

Daniela Brauckhoff, Bernhard Tellenbach, Arno Wagner, Martin May, and Anukool Lakhina. Impact of packet sampling on anomaly detection metrics. In *IMC*, pages 159–164, Rio de Janeriro, Brazil, 2006.

Amit Chakrabarti, Khanh Do Ba, and S. Muthukrishnan. Estimating entropy and entropy norm on data streams. *Internet Mathematics*, 3(1):63–78, 2006.

Amit Chakrabarti, Graham Cormode, and Andrew McGregor. A near-optimal algorithm for computing the entropy of a stream. In *SODA*, pages 328–335, New Orleans, Louisiana, 2007.

John M. Chambers, C. L. Mallows, and B. W. Stuck. A method for simulating stable random variables. *Journal of the American Statistical Association*, 71(354):340–344, 1976.

Carlotta Domeniconi and Dimitrios Gunopulos. Incremental support vector machine construction. In *ICDM*, pages 589–592, San Jose, CA, 2001.

Laura Feinstein, Dan Schnackenberg, Ravindra Balupari, and Darrell Kindred. Statistical approaches to DDoS attack detection and response. In *DARPA Information Survivability Conference and Exposition*, pages 303–314, 2003.

Izrail S. Gradshteyn and Iosif M. Ryzhik. *Table of Integrals, Series, and Products*. Academic Press, New York, fifth edition, 1994.

Sudipto Guha, Andrew McGregor, and Suresh Venkatasubramanian. Streaming and sublinear approximation of entropy and information distances. In *SODA*, pages 733 – 742, Miami, FL, 2006.

Anupam Gupta, John D. Lafferty, Han Liu, Larry A. Wasserman, and Min Xu. Forest density estimation. In *COLT*, pages 394–406, Haifa, Israel, 2010.

Nicholas J. A. Harvey, Jelani Nelson, and Krzysztof Onak. Sketching and streaming entropy via approximation theory. In *FOCS*, 2008a.

Nicholas J. A. Harvey, Jelani Nelson, and Krzysztof Onak. Streaming algorithms for estimating entropy. In *ITW*, 2008b.

M E. Havrda and F. Charvát. Quantification methods of classification processes: Concept of structural α-entropy. *Kybernetika*, 3:30–35, 1967.

Monika R. Henzinger, Prabhakar Raghavan, and Sridhar Rajagopalan. *Computing on Data Streams*. American Mathematical Society, Boston, MA, USA, 1999.

Piotr Indyk. Stable distributions, pseudorandom generators, embeddings, and data stream computation. *Journal of ACM*, 53(3):307–323, 2006.

Anukool Lakhina, Mark Crovella, and Christophe Diot. Mining anomalies using traffic feature distributions. In *SIGCOMM*, pages 217–228, Philadelphia, PA, 2005.

Ashwin Lall, Vyas Sekar, Mitsunori Ogihara, Jun Xu, and Hui Zhang. Data streaming algorithms for estimating entropy of network traffic. In *SIGMETRICS*, pages 145–156, Saint Malo, France, 2006.

Ping Li. Very sparse stable random projections for dimension reduction in l_α $(0 < \alpha \leq 2)$ norm. In *KDD*, San Jose, CA, 2007.

Ping Li. Estimators and tail bounds for dimension reduction in l_α $(0 < \alpha \leq 2)$ using stable random projections. In *SODA*, pages 10 – 19, San Francisco, CA, 2008.

Ping Li. Compressed counting. In *SODA*, New York, NY, 2009a.

Ping Li. Improving compressed counting. In *UAI*, Montreal, CA, 2009b.

Ping Li and Kenneth W. Church. A sketch algorithm for estimating two-way and multi-way associations. *Computational Linguistics (Preliminary results appeared in HLT/EMNLP 2005)*, 33(3):305–354, 2007.

Ping Li and Trevor J. Hastie. A unified near-optimal estimator for dimension reduction in l_α $(0 < \alpha \leq 2)$ using stable random projections. In *NIPS*, Vancouver, BC, Canada, 2007.

Ping Li, Kenneth W. Church, and Trevor J. Hastie. One sketch for all: Theory and applications of conditional random sampling. In *NIPS (Preliminary results appeared in NIPS 2006)*, Vancouver, BC, Canada, 2008.

Ping Li, Michael Mahoney, and Yiyuan She. Approximating higher-order distances using random projections. In *UAI*, 2010.

Qiaozhu Mei and Kenneth Church. Entropy of search logs: How hard is search? with personalization? with backoff? In *WSDM*, pages 45 – 54, Palo Alto, CA, 2008.

S. Muthukrishnan. Data streams: Algorithms and applications. *Foundations and Trends in Theoretical Computer Science*, 1:117–236, 2 2005.

Liam Paninski. Estimation of entropy and mutual information. *Neural Comput.*, 15(6): 1191–1253, 2003.

Alfred Rényi. On measures of information and entropy. In *The 4th Berkeley Symposium on Mathematics, Statistics and Probability 1960*, pages 547–561, 1961.

Constantino Tsallis. Possible generalization of boltzmann-gibbs statistics. *Journal of Statistical Physics*, 52:479–487, 1988.

Kuai Xu, Zhi-Li Zhang, and Supratik Bhattacharyya. Profiling internet backbone traffic: behavior models and applications. In *SIGCOMM '05: Proceedings of the 2005 conference on Applications, technologies, architectures, and protocols for computer communications*, pages 169–180, Philadelphia, Pennsylvania, USA, 2005.

Qiang Yang and Xindong Wu. 10 challeng problems in data mining research. *International Journal of Information Technology and Decision Making*, 5(4):597–604, 2006.

Haiquan Zhao, Ashwin Lall, Mitsunori Ogihara, Oliver Spatscheck, Jia Wang, and Jun Xu. A data streaming algorithm for estimating entropies of od flows. In *IMC*, San Diego, CA, 2007.

Haiquan Zhao, Nan Hua, Ashwin Lall, Ping Li, Jia Wang, and Jun Xu. Towards a universal sketch for origin-destination network measurements. Technical report, 2010.