A simple sketching algorithm for entropy estimation over streaming data

Peter Clifford
Department of Statistics
University of Oxford
Oxford, UK. OX1 3TG
peter.clifford@jesus.ox.ac.uk

Ioana Ada Cosma
Department of Mathematics and Statistics
University of Ottawa
Ottawa, ON, Canada. K1N 6N5
icosma@uottawa.ca

Abstract

We consider the problem of approximating the empirical Shannon entropy of a high-frequency data stream under the relaxed strict-turnstile model, when space limitations make exact computation infeasible. An equivalent measure of entropy is the Rényi entropy that depends on a constant α. This quantity can be estimated efficiently and unbiasedly from a low-dimensional synopsis called an α-stable data sketch via the method of compressed counting. An approximation to the Shannon entropy can be obtained from the Rényi entropy by taking α sufficiently close to 1. However, practical guidelines for parameter calibration with respect to α are lacking. We avoid this problem by showing that the random variables used in estimating the Rényi entropy can be transformed to have a proper distributional limit as α approaches 1: the maximally skewed, strictly stable distribution with $\alpha = 1$ defined on the entire real line. We propose a family of asymptotically unbiased *log-mean estimators* of the Shannon entropy, indexed by a constant $\zeta > 0$, that can be computed in a single-pass algorithm to provide an additive approximation. We recommend the log-mean estimator with $\zeta = 1$ that has exponentially decreasing tail bounds on the error probability, asymptotic relative efficiency of 0.932, and near-optimal computational complexity.

Appearing in Proceedings of the 16^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2013, Scottsdale, AZ, USA. Volume 31 of JMLR: W&CP 31. Copyright 2013 by the authors.

1 INTRODUCTION

Streaming data is ubiquitous in a wide range of areas from engineering, and information technology, finance, and commerce, to atmospheric physics, and earth sciences (Muthukrishnan, 2005; Aggarwal, 2007). The term *streaming data* refers to the situation where data is continuously generated at high speed, and must be processed in real time to facilitate data analysis and decision making. The Shannon entropy (Shannon and Weaver, 1949) provides an important characterization of a data stream with many areas of application, e.g., network traffic monitoring for the purpose of anomaly detection or traffic clustering, analysis of commercial search logs, and signal processing. In network traffic monitoring, Lall et al. (2006) show that the empirical Shannon entropy is an appropriate summary statistic for capturing changes in the underlying traffic distribution. Changes in the distribution of the number of packets observed at different ports can be indicative of port scanning attacks.

In recent years, several algorithms have been developed for estimating the Shannon entropy over streaming data (Bhuvanagiri and Ganguly, 2006; Chakrabarti et al., 2006; Zhao et al., 2007; Harvey et al., 2008b; Chakrabarti et al., 2010; Li and Zhang, 2011). Many of these algorithms are based on the approach of α-stable *data sketching* (Indyk, 2006). Sketches are low-dimensional data structures, usually in vector or matrix format; when a new element in the stream is observed, the sketch is potentially updated, and the element is discarded. This update is handled in the same way, irrespective of the order of arrival of past data. The idea is to construct and maintain on-the-fly a compact synopsis of the data stream such that summary statistics of interest can be accurately approximated from the synopsis; in general, synopsis construction is specific to the statistic of interest.

This paper considers the problem of estimating the

Shannon entropy of a data stream under the assumption that the number of distinct elements observed in the stream is prohibitively large, so that the vector of cumulative quantities cannot be stored on main computer memory for fast and efficient access. We employ the method of α-stable data sketching, i.e., transforming distinct stream elements online to distinct realizations of a stable variable of index α (called α-stable hereafter), and storing weighted linear combinations of these realizations, independently replicated k times. These weighted linear combinations, known as *random projections*, form a k-dimensional synopsis of the data stream, called a data sketch hereafter, where k is determined by the accuracy desired in approximating the entropy. We consider data streams under the relaxed strict-turnstile model, which allows deletions, provided that, whenever the entropy is estimated, all cumulative quantities are non-negative.

We sketch distinct stream elements to pseudo-random variates following the maximally skewed α-stable distribution with $\alpha = 1$ via the *method of seeding*; the stream elements effectively index the random variates. This is in contrast to existing approaches that involve sketching with α close to 1, thus introducing an additional source of error as explained in Section 1.3. We present a family of *log-mean estimators* of the Shannon entropy whose construction is simple and direct, requiring a single pass over the data stream. We give explicit algorithms for implementing this estimation procedure, and analyze their computational complexity in terms of the length of the stream, and the additive approximation error.

1.1 Notation and terminology

A *data stream* S_T of length T is a transiently observed sequence of data elements (i_t, d_t) that arrive unordered, with repetition, and at very high rate of transmission. The item type i_t belongs to a large or possibly infinite set $\mathcal{D} = \{c_1, c_2, \ldots, c_N\}$ and the associated quantity is $d_t \in \mathbb{R}$, for $t = 1, 2, \ldots, T$. If $d_t < 0$, then the data element (i_t, d_t) is a deletion from the stream; otherwise, it is an insertion. For simplicity, we assume that $T \geq N$. The empirical probability distribution is given by

$$p_j = \frac{a_j}{\sum_{i=1}^{N} a_i},\ j = 1, \ldots, N,$$

where $a_j = \sum_{t=1}^{T} d_t \mathbb{I}(i_t = c_j)$ is the cumulative quantity of elements of type c_j at stage T, and $a_j \geq 0\ \forall j$ at every stage T of interest. So, the empirical distribution is well-defined. This is called the relaxed strict-turnstile model.

The *empirical Shannon entropy* is defined by

$$H(p) = -\sum_{j=1}^{N} p_j \log p_j, \tag{1}$$

where, by convention, $p \log p$ is defined to be 0 when $p = 0$, and log is the logarithm function to the base e. Equivalent measures of entropy are the Rényi (Rényi, 1961) and Tsallis (Tsallis, 1988) entropies, given, respectively, by

$$H_\alpha(p) = \frac{1}{1-\alpha} \log \left(\sum_{j=1}^{N} p_j^\alpha \right),$$

$$S_\alpha(p) = \frac{1}{1-\alpha} \left(\sum_{j=1}^{N} p_j^\alpha - 1 \right),$$

for $0 \leq \alpha$ and $\alpha \neq 1$. $H_\alpha(p)$ and $S_\alpha(p)$ equal the Shannon entropy in the limit as α tends to 1. Both quantities $H_\alpha(p)$ and $S_\alpha(p)$ are functions of the αth frequency moment, denoted by $F_\alpha(p)$, and defined as

$$F_\alpha(p) = \sum_{j=1}^{N} p_j^\alpha,$$

a connection that is exploited by many algorithms for estimating the Shannon entropy, as explained in Section 1.3.

1.2 Data sketching and the stable distribution

We employ the method of data sketching to the α-stable distribution. Following Zolotarev (1986), the stable distribution has four parameters: index $\alpha \in (0, 2]$, skewness $\beta \in [-1, 1]$, location $\delta \in \mathbb{R}$, and scale $\gamma > 0$, denoted by $F(x; \alpha, \beta, \gamma, \delta)$. If X has distribution $F(x; \alpha, \beta, \gamma, \delta)$ (written as: $X \sim F(x; \alpha, \beta, \gamma, \delta)$), then its characteristic function (c.f.) $\phi(\theta) = \mathbb{E} \exp(i\theta X)$, $\theta \in \mathbb{R}$, is given by

$$\phi(\theta) = \begin{cases} \exp\left(\gamma^\alpha [-|\theta|^\alpha + i\theta|\theta|^{\alpha-1}\beta \tan(\frac{\pi\alpha}{2})] + i\delta\theta\right), \\ \qquad \text{if } \alpha \neq 1 \\ \exp\left(\gamma[-|\theta| - i\theta\beta(\frac{2}{\pi}) \log|\theta|] + i\delta\theta\right), \\ \qquad \text{if } \alpha = 1, \end{cases}$$

where $\mathbb{E}$ denotes expected value, and $i = \sqrt{-1}$. If $\beta = \pm 1$, the distribution is called maximally skewed. In particular, we sketch to the maximally skewed distribution $F(x; 1, -1, \pi/2, 0)$ by simulating independent draws using the algorithm in Table 1 (Zolotarev, 1986).

Randomized algorithms for data sketching are probabilistic, in the sense that data stream elements are

Table 1: Algorithm to simulate from the maximally skewed stable distribution $F(x; 1, -1, \pi/2, 0)$. Unif$(0, 1)$ denotes the uniform distribution on $(0, 1)$

1: Simulate $U_1, U_2 \sim \text{Unif}(0, 1)$ independently.
2: Let $W_1 = \pi(U_1 - \frac{1}{2})$ and $W_2 = -\log U_2$.
3: Return $\tan(W_1)[\frac{\pi}{2} - W_1] + \log\left(W_2 \frac{\cos W_1}{\pi/2 - W_1}\right)$.

mapped deterministically to copies of pseudo-random variables, and the variables are transformed to form a synopsis representation of the data stream. From this representation, an estimate of the Shannon entropy is derived whose accuracy can be guaranteed, to within a specified level ϵ, with probability exceeding $1 - \rho$. In particular, a randomized algorithm for estimating $H(p)$ will return an (ϵ, ρ)-approximation $\hat{H}(p)$ that satisfies: $\mathbb{P}(|\hat{H}(p) - H(p)| \leq \epsilon H(p)) \geq 1 - \rho$ for a multiplicative approximation, and $\mathbb{P}(|\hat{H}(p) - H(p)| \leq \epsilon) \geq 1 - \rho$ for an additive approximation.

In fact, it suffices to have a randomized algorithm that returns an approximation $\hat{H}(p)$ to within accuracy ϵ with probability greater than 0.5. An application of Chernoff's bounds (Hoeffding, 1963) shows that from $n = \log(1/\rho)$ independent repetitions of the algorithm, the median of $\hat{H}_1(p), \ldots, \hat{H}_n(p)$ is an (ϵ, ρ)-approximation of $H(p)$, where $\hat{H}_i(p)$ is the approximation from the ith repetition. Hence, in general we speak of ϵ-additive and ϵ-multiplicative approximations.

1.3 Related work on Shannon entropy estimation

Approximating the empirical Shannon entropy from estimates of Rényi or Tsallis entropies started with the work of Zhao et al. (2007). These authors show that the function $x \log(x)$ can be well approximated by a linear combination of two functions of the form x^p, $p \in (0, 2]$, for x less than an upper bound. Summing over distinct data types, they obtain an estimate of the entropy from a linear combination of two frequency moments, effectively interpolating the entropy from two distinct values of $S_\alpha(p)$. Zhao et al. (2007) estimate the frequency moments by the random projections method of Indyk (2006). For data types whose cumulative quantity exceeds the upper bound, they estimate the contribution to the entropy separately.

More generally, Harvey et al. (2008b) estimate the Shannon entropy via interpolation from several Tsallis entropy estimates, computed at optimal values of α to minimize the approximation error. For arbitrary accuracy parameter $\epsilon > 0$, they present additive and multiplicative approximations in space $O(\epsilon^{-2} \log T(\log\log T + \log(1/\epsilon))^{O(1)})$ for the relaxed strict-turnstile model. The multiplicative approximation algorithm has near-optimal space complexity in terms of its dependence on ϵ, compared to the lower bound of $\Omega(\epsilon^{-2}/\log^2(\epsilon^{-1}))$ (Chakrabarti et al., 2010). Extensions to the general update model with no restrictions on deletions are possible, but the space bounds increase typically by a factor of $O(\log T)$ (Harvey et al., 2008b).

Li presents two estimators (the geometric and harmonic mean estimators) of the αth frequency moment, based on random projections to the symmetric, α-stable distribution (Li, 2008), or the positive, α-stable distribution (Li, 2009a,b). The latter method is called *compressed counting*, and it improves over symmetric stable random projections in terms of the asymptotic variance of the estimator around $\alpha = 1$. Compressed counting was recently applied to estimate the cardinality of a data stream in Clifford and Cosma (2012). Li (2009a,b) suggests that the Shannon entropy can be estimated from expressions $S_\alpha(p)$ or $H_\alpha(p)$ with α close to 1 using the geometric mean estimator. Unfortunately, the resulting algorithm is impractical since it has complexity of order $O(1/\Delta)$ with $\Delta = 1 - \alpha$. Li and Zhang (2011) offer a marked improvement with a new compressed counting algorithm that provides an ϵ-additive estimate of the Shannon entropy with complexity $O(1/\epsilon^2)$. In particular, they estimate the Shannon entropy with $\alpha \approx 1$ by

$$H_\alpha(p) = -\log J_\alpha(p) - \frac{1}{\Delta} \log F_1^\alpha(p),$$

where $J_\alpha(p) = F_\alpha^{-1/\Delta}(p)$, and the first frequency moment is computed exactly. Moreover, the estimator of $J_\alpha(p)$ has near-optimal efficiency properties in estimating $F_\alpha(p)$, and exponentially decreasing tail bounds.

In addition, Harvey et al. (2008a) analyze the rate of convergence of the Rényi entropy estimate of $H_\alpha(p)$ to the Shannon entropy as $\alpha \to 1^+$, and provide an explicit formula for $\alpha > 1$ that guarantees an ϵ-additive approximation. They estimate $F_\alpha(p)$ by the method of symmetric stable random projections (Li, 2008), where $0 < \alpha \leq 2$. However, a value of α exceeding 1 is not appropriate for maximally skewed stable random projections that require $\alpha < 1$. Another disadvantage is the prohibitively large space complexity, of order $\tilde{O}(\epsilon^{-4} \log^4 T)$, ignoring logarithmic terms.

The problem of estimating the Shannon entropy is related to that of measuring pairwise independence via the Kullback-Leibler divergence (Kullback and Leibler, 1951), where the latter has received a lot of attention in recent literature (Indyk and McGregor, 2008; Guha

et al., 2008; Braverman and Ostrovsky, 2010). Indyk and McGregor (2008) present a single-pass algorithm for an ϵ-additive approximation of the mutual information between two data streams. The empirical mutual information is the Kullback-Leibler divergence of the joint distribution and the product of the marginals. Their algorithm has space complexity $\tilde{O}(\epsilon^{-2})$, but an $(1+\epsilon)$-multiplicative approximation of the mutual information does not exist in small space (Indyk and McGregor, 2008). Since the mutual information can be expressed as the sum of the empirical Shannon entropies of the marginals minus the empirical Shannon entropy of the joint, our estimation approach can provide an additive approximation. The same holds for the conditional entropy represented in terms of Shannon entropies.

1.4 Our contributions

Let δ denote $-H(p)$, the negative of the empirical Shannon entropy. In Section 2.2 we present a family of log-mean estimators, denoted by $\hat{\delta}_{lm}(\zeta)$ and indexed by $\zeta > 0$, for the additive approximation of δ. The algorithm in Table 2 implements the estimation procedure with $\zeta = 1$, and has the following properties:

- It requires a single pass over the data stream.
- It constructs a k-dimensional data sketch by projecting to maximally skewed stable random variables with distribution $F(x; 1, -1, \pi/2, 0)$.
- It returns the log-mean estimator that avoids the problem of parameter calibration with respect to the index α (Harvey et al., 2008b; Li and Zhang, 2011), by going directly to the limit with $\alpha = 1$.

Section 2.1 provides the motivation, and the details are in Lemmas 2.1 and 2.2. The proposed estimator with $\zeta = 1$ has the following properties:

- It is asymptotically unbiased as $k \to \infty$, and we show in an empirical study that it has good small-sample performance.
- By estimating the entropy directly, rather than the Rényi entropy $H_\alpha(p)$ with $\alpha \approx 1$, we can make precise statements about the efficiency of our estimator: it is near-optimal with asymptotic relative efficiency (ARE) (Lehmann, 1998) of 0.932.
- Lemma 3.1 shows that the estimator has exponentially decreasing tail bounds; in particular, for arbitrary $\epsilon > 0$ and fixed $\zeta \leq 1$,

$$\mathbb{P}\left(\hat{\delta}_{lm}(\zeta) - \delta \geq \epsilon\right) < \exp\left(-k\frac{\epsilon^2}{G_R}\right)$$
$$\mathbb{P}\left(\hat{\delta}_{lm}(\zeta) - \delta \leq -\epsilon\right) < \exp\left(-k\frac{\epsilon^2}{G_L}\right),$$

  where G_L and G_R are small constants that tend in value to 6 as $\epsilon \to 0$. For $\epsilon \in [0.1, 1]$, numerical approximations show that these constants fall in ranges $(4.0, 6.0)$ and $(6.0, 9.5)$, respectively.
- It follows from Lemma 3.1 that for fixed $\rho \in (0, 1)$, the data sketch size k must be of order $O(1/\epsilon^2)$.
- The space complexity of the algorithm is $O\left(1/\epsilon^2 \log T \log(T/\epsilon)\right)$ bits of space, which is near-optimal in terms of dependence on ϵ; in particular, it is optimal up to $\log(1/\epsilon)$ (Kane et al., 2011).

Table 2: Algorithm to approximate the empirical Shannon entropy of a data stream $\mathcal{S}_T$ via the log-mean estimator $\hat{\delta}_{lm}(1)$

1: Initialize data sketch $(y_1, \ldots, y_k) = (0, \ldots, 0)$.
2: Set the counter $Y = 0$.
3: For $t = 1$ to T
4: Update the counter $Y = Y + d_t$.
5: Seed the PRNG with i_t.
6: For $j = 1$ to k
7: Generate $R_j(i_t) \sim F(x; 1, -1, \pi/2, 0)$
8: Update $y_j = y_j + R_j(i_t) \times d_t$.
9: At time $t = T$, set $y_j = y_j / Y$ for $j = 1, \ldots, k$.
10: Return $\hat{H}(p) = -\log\left(k^{-1} \sum_{j=1}^{k} \exp(y_j)\right)$.

2 THE LOG-MEAN ESTIMATOR

2.1 The method of random projections

The method of random projections requires that each element type $c_j \in \mathcal{D}$ that appears in the data stream can be transformed into a distinct random variable $R(c_j)$. In practice, this is achieved "to adequate approximation" by the method of seeding as follows: (i) map c_j to an integer (or vector of integers), (ii) use these integers to seed a pseudo-random number generator (PRNG), and (iii) use the seeded PRNG to simulate the random variable $R(c_j)$.

Nisan (1992) shows that there exists an explicit implementation of a PRNG that converts a random seed to a

sequence of bits, indistinguishable from truly random bits. So we assume that our PRNG produces truly random variables $R(c_j)$.

The projection is then accumulated online as $\sum_{t=1}^{T} R(i_t)d_t = \sum_{j=1}^{N} R(c_j)a_j$. This sum is the dot product of the vector of cumulative quantities with a vector of N independent random variables, each drawn from the maximally skewed stable distribution with $\alpha = 1$. This provides a single element of the data sketch. A further $k-1$ elements are generated independently in parallel to form the k-dimensional data sketch.

We now motivate the use of the maximally skewed stable distribution with $\alpha = 1$ in the random projections method by showing how the problem of estimating the Shannon entropy reduces to that of approximating a location parameter.

Define the quantity

$$B_\alpha = \left(\sum_{j=1}^{N} p_j^\alpha\right)^{1/\alpha} = F_\alpha^{1/\alpha}(p).$$

Let

$$Z_\alpha \sim F\left(z; \alpha, 1, \left(\cos\left(\frac{\pi\alpha}{2}\right)\right)^{1/\alpha}, 0\right),$$

for fixed $0 < \alpha < 1$; this is the positive, strictly stable distribution with Laplace transform $e^{-\lambda^\alpha}$ for $\lambda \geq 0$. Let $\left(Z_\alpha^{(1)}, \ldots, Z_\alpha^{(N)}\right)$ be a vector of independent copies of Z_α and let $p = (p_1, \ldots, p_N)$ be a vector of frequencies that satisfy $\sum_{j=1}^{N} p_j = 1$. From Zolotarev (1986), we have that

$$\sum_{j=1}^{N} Z_\alpha^{(j)} p_j \sim Z_\alpha \left(\sum_{j=1}^{N} p_j^\alpha\right)^{1/\alpha} = Z_\alpha B_\alpha. \qquad (2)$$

Projecting to the positive, strictly stable distribution and maintaining weighted linear combinations as in (2) is precisely the method of compressed counting (Li, 2009a,b). Compressed counting reduces the problem of Shannon entropy estimation to that of estimating the scale parameter $B_\alpha = J_\alpha^{-\Delta/\alpha}(p)$.

Instead, we project to the maximally skewed stable distribution with $\alpha = 1$ and $\beta = -1$, defined on the entire real axis. Starting from the Rényi entropy

$$H_\alpha(p) = \frac{\alpha}{1-\alpha}\log B_\alpha,$$

it is easy to show that as $\alpha \to 1$,

$$\frac{1-B_\alpha}{1-\alpha} = \frac{1}{1-\alpha}\left[1 - e^{(1-\alpha)H_\alpha(p)/\alpha}\right] \to \delta,$$

where $-\delta = -\sum_{j=1}^{N} p_j \log p_j$ is the Shannon entropy. Next, we define

$$Y_\alpha^{(j)} = \frac{1 - Z_\alpha^{(j)}}{1-\alpha} + \log(1-\alpha),$$

and, using (2), we obtain

$$\sum_{j=1}^{N} Y_\alpha^{(j)} p_j = \sum_{j=1}^{N}\left[\frac{1 - Z_\alpha^{(j)}}{1-\alpha} + \log(1-\alpha)\right] p_j \sim \left[\frac{1-Z_\alpha}{1-\alpha} + \log(1-\alpha)\right] + Z_\alpha \frac{(1-B_\alpha)}{1-\alpha}. \qquad (3)$$

Taking limits, $\sum_{j=1}^{N} Y_1^{(j)} p_j \sim Y_1 + \delta$, provided Y_α has a proper limit as $\alpha \to 1$, and using the fact that $Z_\alpha \to 1$ as $\alpha \to 1$. The following lemma provides the details.

Lemma 2.1. *The random variable Y_α has a proper limit Y_1 as $\alpha \to 1$. The variable Y_1 has a maximally skewed stable distribution with $\alpha = 1$, and c.f.*

$$\phi(\theta) = \exp\left(-\frac{1}{2}\pi|\theta| + i\theta\log|\theta|\right) = (i\theta)^{i\theta},$$

i.e., $Y_1 \sim F(y; 1, -1, \pi/2, 0)$. Moreover, the kth moment of the random variable $\exp(Y_1)$ is k^k for all $k > 0$.

Proof. See the supplementary material. □

The heart of our algorithm is contained in the following result; it shows that by sketching to the $F(y; 1, -1, \pi/2, 0)$ distribution, the negative of the Shannon entropy is recovered as the location parameter of the distribution of a linear combination weighted by the empirical probability mass function.

Lemma 2.2. *Let $X_1, \ldots, X_N \sim F(x; 1, -1, \pi/2, 0)$ i.i.d., and let $p_1, \ldots, p_N$ be positive constants satisfying $\sum_{j=1}^{N} p_j = 1$. Then,*

$$\sum_{j=1}^{N} p_j X_j \sim F\left(x; 1, -1, \frac{\pi}{2}, \sum_{j=1}^{N} p_j \log p_j\right).$$

Proof. The c.f. of $\sum_{j=1}^{N} p_j X_j$, for $\theta \in \mathbb{R}$, is given by

$$\mathbb{E}\exp\left(i\theta\sum_{j=1}^{N} p_j X_j\right) = \prod_{j=1}^{N}\mathbb{E}\exp\left(i\theta p_j X_j\right)$$
$$= \prod_{j=1}^{N}\exp\left(-\frac{\pi}{2}p_j|\theta| + i\theta p_j \log|\theta p_j|\right)$$
$$= \exp\left(\frac{\pi}{2}\left[-|\theta| + i\theta\log|\theta|\frac{2}{\pi}\right] + i\theta\sum_{j=1}^{N} p_j \log p_j\right).$$

The first equality follows from properties of characteristic functions of sums of independent random variables, the second follows from Lemma 2.1, and the third equality uses the fact that $\sum_{j=1}^{N} p_j = 1$. Since the distribution of a random variable is specified by its characteristic function (Grimmett and Stirzaker, 2001), the result follows by comparison to expression $\phi(\theta)$ in Section 1.2. □

2.2 Derivation of the family of log-mean estimators

Lemma 2.3. *Let $y_1, \ldots, y_k$ be independent samples from the $F(y; 1, -1, \pi/2, \delta)$ distribution, and let $\zeta > 0$ be a constant. The log-mean estimator of δ is*

$$\hat{\delta}_{lm}(\zeta) = \zeta^{-1} \log \left(\zeta^{-\zeta} k^{-1} \sum_{j=1}^{k} \exp(\zeta y_j) \right).$$

As the sample size k increases to ∞, the estimator is asymptotically unbiased; in particular, as $k \to \infty$,

$$\sqrt{k} \left(\hat{\delta}_{lm}(\zeta) - \delta \right) \to \text{Normal} \left(0, \frac{4^{\zeta} - 1}{\zeta^2} \right).$$

Moreover, the Fisher information about δ contained in a single random variable from the $F(y; 1, -1, \pi/2, \delta)$ distribution is approximately 0.3578, so the ARE of $\hat{\delta}_{lm}(\zeta)$ is

$$ARE(\hat{\delta}_{lm}(\zeta)) = \frac{\zeta^2}{0.3578(4^{\zeta} - 1)}.$$

Hence, the estimator $\hat{\delta}_{lm}(1.15)$ is near-optimal with largest ARE of 0.942, and $\hat{\delta}_{lm}(1)$ has ARE of 0.932.

Proof. See the supplementary material. □

In Section 3, we show that the log-mean estimator has exponentially decreasing tail bounds only for $\zeta \leq 1$. For this reason, we recommend the estimator with $\zeta = 1$ that attains the maximum ARE over the range $\zeta \leq 1$. The algorithm in Table 2 returns $\hat{H}(p) = -\hat{\delta}_{lm}(1)$.

3 PERFORMANCE, TAIL BOUNDS, AND SPACE COMPLEXITY

Figure 1 compares the log-mean estimator $\hat{\delta}_{lm}(\zeta)$ to the Rényi entropy estimator $\hat{H}_{\alpha}(p)$ (Li and Zhang, 2011) in terms of asymptotic relative efficiency over the range of values $\zeta \in (0, 2]$ and $\alpha \in [0.95, 0.99]$. The Fisher information about $H_{\alpha}(p)$ contained in a positive, α-stable random variable with scale parameter $F_{\alpha}^{1/\alpha}(p)$ is given by the expression $\mathrm{I}_2(\alpha) - 1$ in Li and Zhang (2011). The solid, straight line in Figure 1 is the ARE of $\hat{H}_{\alpha}(p)$, given by

$$ARE(\hat{H}_{\alpha}(p)) = \frac{1}{(\mathrm{I}_2(\alpha) - 1)(1 + 2\alpha)},$$

since $\sqrt{k}(\hat{H}_{\alpha}(p) - H_{\alpha}(p)) \to \text{Normal}(0, 1 + 2\alpha)$ as $k \to \infty$ (Li and Zhang, 2011).

Figure 2 compares the performances of the log-mean estimator with $\zeta = 1, 1.5$ and the estimator $\hat{H}_{\alpha}(p)$ with $\alpha = 0.97$ in terms of relative mean square error (MSE) in small samples. Plotted for comparison is the Cramér-Rao lower bound, defined by $(k \times 0.3578)^{-1}$, and giving a lower bound on the variance of any unbiased estimator of δ. The MSE is given by $\mathbb{E}(\hat{\delta}_{lm}(\zeta) - \delta)^2$ for the log-mean estimator, and by $\mathbb{E}(\hat{H}_{\alpha}(p) + \delta)^2$ for the Rényi estimator, since $\hat{H}_{\alpha}(p)$ estimates $H(p) = -\delta$ for $\alpha \approx 1$. The expectation is estimated from 10^5 replicates.

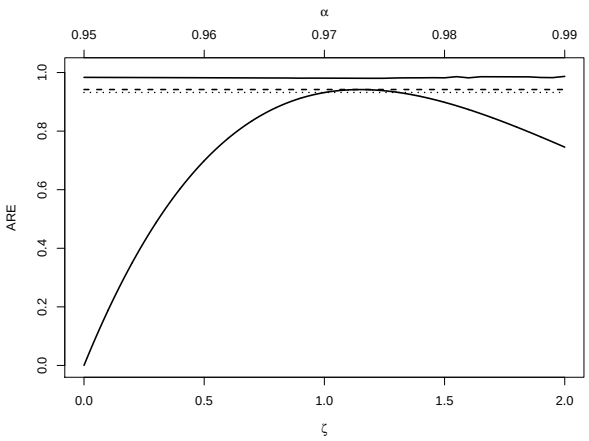

Figure 1: Comparison in terms of asymptotic relative efficiency of the log-mean estimator $\hat{\delta}_{lm}(\zeta)$ for $\zeta \in (0, 2]$ (curved, solid line, bottom axis) to the Rényi entropy estimator $\hat{H}_{\alpha}(p)$ for $\alpha \in [0.95, 0.99]$ (straight, solid line, top axis). Horizontal lines are drawn at ARE = 0.942 (long dashed line) and ARE=0.932 (dotted line), and a vertical line at $\zeta = 1$, or equivalently, $\alpha = 0.97$. The ARE of $\hat{\delta}_{lm}(\zeta)$ refers to Shannon entropy estimation, whereas the ARE of $\hat{H}_{\alpha}(p)$ refers to Rényi entropy estimation.

All computations were performed using the statistical software `R` (http://www.r-project.org/), and the same string of random numbers was employed in the computations of each estimator. The estimator $\hat{H}_{\alpha}(p)$ is derived from a random sample of positive, α-stable random variables, raised to the power $-\alpha/\Delta$ and scaled by $F_{\alpha}(p)^{-1/\Delta}$. With $\alpha \approx 1$, $F_{\alpha}(p)$ is approximately equal to $\sum_{i=1}^{N} a_i$, the total cumulative quantity, and, if this quantity is large, then the scaling factor is effec-

tively zero. We experience this problem for our simulated data stream with values of $\alpha \geq 0.98$, hence our choice of $\alpha = 0.97$ for the small-sample comparison in Figure 2. The estimator $\hat{H}_{0.97}(p)$ has ARE of 0.981.

Figure 2 shows that the three estimators have comparable performance in small samples. Compared to the Cramér-Rao lower bound, the performance of the log-mean estimators is particularly good for $k \geq 20$. However, we do not have a corresponding lower bound for the performance of $\hat{H}_\alpha(p)$ as an estimator of the Shannon entropy, since an analysis of the rate of convergence of the Rényi entropy to the Shannon entropy as $\alpha \to 1^-$ is lacking.

The length of the data sketch vector, k, is determined by the behaviour of the tail bounds of the additive approximation error. Lemma 3.1 shows that for $\zeta \leq 1$, the log-mean estimator has exponentially decreasing tail bounds.

Lemma 3.1. *Exponentially decreasing tail bound exist for $\zeta \leq 1$ and arbitrary $\epsilon > 0$, with*

$$\mathbb{P}\left(\hat{\delta}_{lm}(\zeta) - \delta \geq \epsilon\right) < \exp\left(-k\frac{\epsilon^2}{G_R}\right),$$
$$\mathbb{P}\left(\hat{\delta}_{lm}(\zeta) - \delta \leq -\epsilon\right) < \exp\left(-k\frac{\epsilon^2}{G_L}\right),$$

where

$$G_R = \frac{\epsilon^2}{\sup_{t>0} Q_\zeta(t,\epsilon)},\ G_L = \frac{\epsilon^2}{\sup_{t>0} Q_\zeta(-t,-\epsilon)},$$

and

$$Q_\zeta(t,\epsilon) = -\log\left(\sum_{j=0}^{\infty} t^j \frac{j^{\zeta j}}{j!}\right) + te^{\zeta\epsilon}.$$

Furthermore as $\epsilon \to 0$ both G_R and G_L tend to $2(4^\zeta - 1)/\zeta^2$.

Proof. See the supplementary material. □

Given $\epsilon > 0$ and $0 < \rho < 1$, bounding the additive approximation error by

$$\mathbb{P}\left(|\hat{\delta}_{lm}(1) - \delta| \geq \epsilon\right) < \rho,$$

requires that

$$k > -\frac{G}{\epsilon^2}\log\left(\frac{\rho}{2}\right) = O\left(\frac{1}{\epsilon^2}\right),$$

where $G = \max\{G_L, G_R\}$. For $\epsilon \in [0.1, 1]$, numerical approximations show that the constants G_R and G_L fall in ranges (4.0, 6.0) and (6.0, 9.5), respectively, so the hidden constant in the big O notation, ignoring the term $\log(1/\rho)$, is small, for small ϵ. Hence, the algorithm requires $O(\epsilon^{-2}\log T)$ random bits of space for a data stream with $d_t \in \{-1, 1\}$. The space complexity increases to $O\left(\epsilon^{-2}\log T\log(T/\epsilon)\right)$ bits after applying Nisan's PRNG (Nisan, 1992; Indyk, 2006). In the general case that allows insertions and deletions with $d_t \in \{-M, \ldots, M\}$, it suffices to increase T by a factor M (Harvey et al., 2008b).

4 CONCLUSION

This paper joins a growing body of literature on estimating the empirical Shannon entropy over streaming data efficiently, with small memory usage and fast updates. In particular, we adopt the method of random projections to the maximally skewed, strictly stable distribution with parameters $\alpha = 1$ and $\beta = -1$, thus avoiding the problem of the choice of parameter α (Harvey et al., 2008b; Li and Zhang, 2011). We derive properties of this distribution, showing that it has a surprisingly simple characteristic function $(i\theta)^{i\theta}$ and that the kth moment of the exponential of such a variable is k^k for all positive real values of k. These properties enable the Shannon entropy to be estimated directly from the associated data sketch as the logarithm of a simple average.

We recommend the asymptotically unbiased log-mean estimator with $\zeta = 1$ to provide an additive approximation of the Shannon entropy. By estimating the entropy directly, rather than via the Rényi entropy with $\alpha \approx 1$, we can determine the asymptotic relative efficiency of our estimator: 0.932 with $\zeta = 1$. Moreover, the probability of the estimator having an additive error greater than ϵ decreases exponentially with $k\epsilon^2$ for small ϵ, where k is the size of the data sketch. This results in a near-optimal space complexity bound of $O\left(\epsilon^{-2}\log T\log(T/\epsilon)\right)$, where T is the length of the data stream observed.

References

C. C. Aggarwal. *Data streams: Models and Algorithms.* Springer-Verlag, New York, 2007.

L. Bhuvanagiri and S. Ganguly. Estimating entropy over data streams. *Lecture Notes in Computer Science*, 4168:148–159, 2006.

V. Braverman and R. Ostrovsky. Measuring independence of datasets. In *Proceedings of the 42nd ACM symposium on Theory of computing (STOC)*, pages 271–280, New York, 2010. ACM.

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

A. Chakrabarti, G. Cormode, and A. McGregor. A

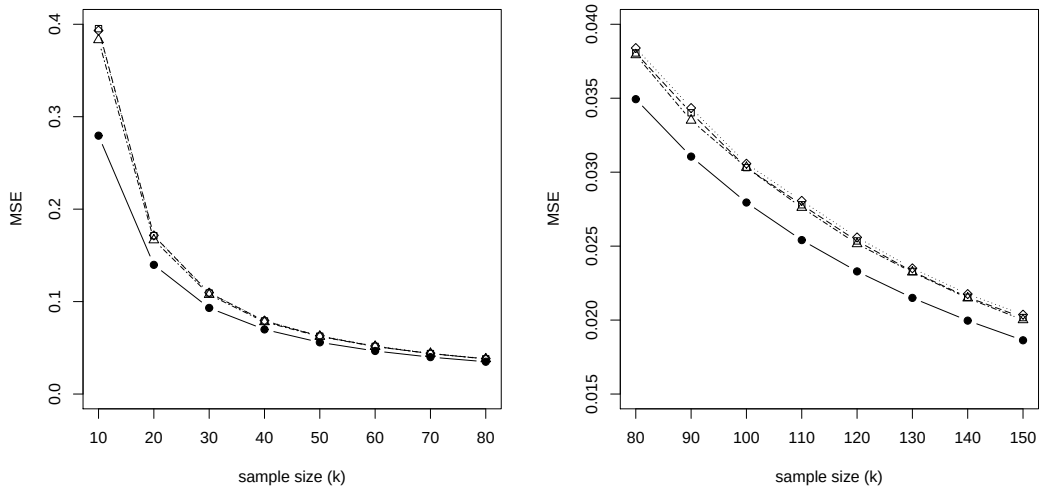

Figure 2: Comparison in terms of MSE of the log-mean estimators $\hat{\delta}_{lm}(1)$ (dotted line, $\Diamond$), $\hat{\delta}_{lm}(1.15)$ (long dash line, $\Box$), and the estimator $\hat{H}_{0.97}(p)$ (two dash line, $\triangle$) of Li and Zhang (2011). The solid line is the Cramér-Rao lower bound on the variance of an unbiased estimator of the Shannon entropy. The MSE is estimated from 10^5 replicates. The MSE lines are indistinguishable, and, for $k \geq 20$, the small-sample performance of the log-mean estimators is very good compared to the Cramér-Rao lower bound.

near-optimal algorithm for estimating the entropy of a stream. *ACM Transactions on Algorithms*, 6 (3), 2010.

P. Clifford and I. A. Cosma. A statistical analysis of probabilistic counting algorithms. *Scandinavian Journal of Statistics*, 39(1):1–14, 2012.

G. Grimmett and D. Stirzaker. *Probability and random processes*. Oxford University Press, USA, 3 edition, 2001.

S. Guha, P. Indyk, and A. McGregor. Sketching information divergences. *Machine Learning*, 72(1):5–19, 2008.

N. J. A. Harvey, J. Nelson, and K. Onak. Streaming algorithms for estimating entropy. In *IEEE Information Theory Workshop, 2008 (ITW)*, pages 227–231, 2008a.

N. J. A. Harvey, J. Nelson, and K. Onak. Sketching and streaming entropy via approximation theory. In *49th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 489–498, 2008b.

W. Hoeffding. Inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

P. Indyk. Stable distributions, pseudorandom generators, embeddings, and data stream computation. *Journal of the Association for Computing Machinery (ACM)*, 53(3):307–323, 2006.

P. Indyk and A. McGregor. Declaring independence via the sketching of sketches. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 737–745, 2008.

D. M. Kane, J. Nelson, E. Porat, and D. P. Woodruff. Fast moment estimation in data streams in optimal space. In *Proceedings of the 43rd annual ACM symposium on Theory of computing (STOC)*, pages 745–754, New York, NY, 2011. ACM.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

A. Lall, V. Sekar, M. Ogihara, J. Xu, and H. Zhang. Data streaming algorithms for estimating entropy of network traffic. *SIGMETRICS Performance Evaluation Review*, 34(1):145–156, 2006.

E. L. Lehmann. *Theory of Point Estimation*. Springer, New York, NY, 2 edition, 1998.

P. Li. Estimators and tail bounds for dimension reduction in l_α $(0 < \alpha \leq 2)$ using stable random variables. In *Proceedings of the nineteenth annual ACM-SIAM*

symposium on Discrete algorithms (SODA), pages 10–19, Philadelphia, PA, USA, 2008. Society for Industrial and Applied Mathematics.

P. Li. Compressed counting. In *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2009)*, pages 412–421. SIAM, 2009a.

P. Li. Improving compressed counting. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI 2009)*, pages 329–338. AUAI Press, 2009b.

P. Li and C.-H. Zhang. A new algorithm for compressed counting with applications in Shannon entropy estimation in dynamic data. *Journal of Machine Learning Research - Proceedings Track*, 19: 477–496, 2011.

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

N. Nisan. Pseudorandom generators for space-bounded computation. *Combinatorica*, 12(4):449–461, 1992.

A. Rényi. On Measures of Entropy and Information. *Proc. Fourth Berkeley Symp. Math. Stat. and Probability*, 1:547–561, 1961.

C. E. Shannon and W. Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, 1949.

C. Tsallis. Possible Generalization of Boltzmann-Gibbs Statistics. *Journal of Statistical Physics*, 52: 479–487, 1988.

H. C. Zhao, A. Lall, M. Ogihara, O. Spatscheck, J. Wang, and J. Xu. A data streaming algorithm for estimating entropies of OD flows. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 279–290, 2007.

V. M. Zolotarev. *One-dimensional stable distributions*. American Mathematical Society, Providence, RI, 1986.

APPENDIX - SUPPLEMENTARY MATERIAL

Proof of Lemma 2.1 Following Zolotarev (1986), the c.f. of Z_α equals

$$\mathbb{E}e^{i\theta Z_\alpha} = \exp\left\{-|\theta|^\alpha \cos\left(\frac{\pi\alpha}{2}\right) + i|\theta|^\alpha \mathrm{sgn}(\theta)\sin\left(\frac{\pi\alpha}{2}\right)\right\},$$

for $\theta \in \mathbb{R}$, where $\mathrm{sgn}(\theta) = \theta/|\theta|$ for $\theta \neq 0$, and 0 otherwise. As $\alpha \to 1$, $\mathbb{E}e^{i\theta Z_\alpha} \to e^{i\theta}$, so $\lim_{\alpha\to 1} Z_\alpha = 1$. It follows that the limit $Y_1 = \lim_{\alpha\to 1} Y_\alpha$ exists.

For $\theta \in \mathbb{R}$, the c.f. of Y_α is given by

$$\begin{aligned}\mathbb{E}e^{i\theta Y_\alpha} &= \exp\left\{i\theta\left(\frac{1}{1-\alpha} + \log(1-\alpha)\right)\right\}\mathbb{E}e^{-i\theta\frac{Z_\alpha}{1-\alpha}} \\ &= -\left|\frac{\theta}{1-\alpha}\right|^\alpha \exp\left\{i\theta\left(\frac{1}{1-\alpha} + \log(1-\alpha)\right)\right\} \\ &\quad \times \exp\left\{i\mathrm{sgn}(\theta)\frac{\pi\alpha}{2}\right\} \qquad (4)\end{aligned}$$

Letting $\alpha \to 1$ in (4), we have the desired limit $(i\theta)^{i\theta}$. Lastly, the kth moment of $\exp(Y_\alpha)$ for $k > 0$ is given by

$$\begin{aligned}\mathbb{E}e^{kY_\alpha} &= (1-\alpha)^k e^{k/(1-\alpha)}\mathbb{E}e^{-kZ_\alpha/(1-\alpha)} \\ &= (1-\alpha)^k e^{-[k/(1-\alpha)]^\alpha + k/(1-\alpha)}, \qquad (5)\end{aligned}$$

where the second equality follows from the Laplace transform of Z_α. Taking the limit as $\alpha \to 1$ in (5), we obtain the kth moment of $\exp(Y_1)$ for $k > 0$ as follows. Define $n = 1/(1-\alpha)$.

$$\begin{aligned}\mathbb{E}e^{kY_\alpha} &= \exp\left\{k\left[n - \log n - \frac{n}{(kn)^{1/n}}\right]\right\} \\ &= \exp\left\{-kn^{-1/n}\left(n\left[k^{-1/n} - 1\right]\right)\right\} \times \\ &\quad \exp\left\{k\left[n - n^{1-1/n} - \log n\right]\right\}.\end{aligned}$$

As $\alpha \to 1$, $n \to \infty$, and we have that $\lim_{n\to\infty} n^{-1/n} = 1$ and $\lim_{n\to\infty} n\left[k^{-1/n} - 1\right] = \log(1/k)$. It remains to show that $n - n^{1-1/n} - \log n \to 0$ as $n \to \infty$. By rewriting

$$n - \frac{n}{n^{1/n}} - \log n = n\left[1 - 1/n^{1/n} + \log\left(1/n^{1/n}\right)\right],$$

we use the fact that for $n > 1$, the Taylor expansion of $\log\left(1/n^{1/n}\right)$ is

$$\begin{aligned}\log\left(n^{-1/n}\right) &= \left(n^{-1/n} - 1\right) + \sum_{i=2}^{\infty} \frac{(-1)^{i+1}(n^{-1/n} - 1)^i}{i} \\ &= \left(n^{-1/n} - 1\right) + O\left((n^{-1/n} - 1)^2\right).\end{aligned}$$

So,

$$n - \frac{n}{n^{1/n}} - \log n = O\left(n^{1-2/n} - 2n^{1-1/n} + n\right),$$

and the right hand side converges to 0 as $n \to \infty$.

Proof of Lemma 2.3 Consider the following transformation:

$$w_j = e^{\zeta y_j} = e^{\zeta(\delta + z_j)} = e^{\zeta\delta} e^{\zeta z_j},$$

where $z_j \sim F(z; 1, -1, \pi/2, 0)$ i.i.d. have characteristic function $\phi(\theta) = \mathbb{E}\exp(i\theta z_j) = (i\theta)^{i\theta}$, for $\theta \in \mathbb{R}$. Then, from Lemma 2.1, $\mathbb{E}w_j = e^{\zeta\delta}\zeta^{\zeta}$. Let $\eta = e^{\zeta\delta}$. The estimator

$$\hat{\eta}(\zeta) = \zeta^{-\zeta} k^{-1} \sum_{j=1}^{k} w_j$$

is unbiased for η, i.e., $\mathbb{E}\hat{\eta}(\zeta) = \eta$, and has variance $\mathrm{var}(\hat{\eta}(\zeta)) = \eta^2 k^{-1}(4^{\zeta} - 1)$.

Moreover, by the Central Limit Theorem, as $k \to \infty$,

$$\sqrt{k}\eta^{-1}(\hat{\eta}(\zeta) - \eta) \to \mathrm{Normal}\left(0, 4^{\zeta} - 1\right).$$

The log-mean estimator of δ is

$$\hat{\delta}_{lm}(\zeta) = \zeta^{-1}\log\hat{\eta} = \zeta^{-1}\log\left(\zeta^{-\zeta}k^{-1}\sum_{j=1}^{k}\exp(\zeta y_j)\right).$$

By the Delta Method, as $k \to \infty$, $\sqrt{k}\big(\hat{\delta}_{lm}(\zeta) - \delta\big) \to \mathrm{Normal}\big(0, \zeta^{-2}(4^{\zeta}-1)\big)$, so $\hat{\delta}_{lm}(\zeta)$ is asymptotically unbiased for δ.

Finally, we want to find the optimal value of ζ that maximizes the ARE of $\hat{\delta}_{lm}(\zeta)$ relative to the MLE of δ. So, we begin by estimating the Fisher information about δ contained in a single random variable following the $F(y; 1, -1, \pi/2, \delta)$ distribution. Let $f(y; 1, -1, \pi/2, \delta)$ denote the corresponding density function. From Algorithm 1, it is possible to show that the density is given by

$$f(y; 1, -1, \pi/2, \delta) = \frac{e^{y-\delta}}{\pi}\int_0^{\pi} e^{-g(w)}e^{-e^{y-\delta-g(w)}}dw,$$

for $-\infty < y < \infty$, where

$$g(w) = \frac{w}{\tan(w)} + \log\left(\frac{\sin(w)}{w}\right).$$

And the Fisher information about δ is expressed as

$$\begin{aligned}\mathrm{I}_1(\delta) &= \mathbb{E}\left(\frac{\partial}{\partial\delta}\log f(y; 1, -1, \pi/2, \delta)\right)^2 \\ &= 1 - \frac{2}{\pi}\int_0^{\infty} s\mathrm{I}(2, s)ds + \frac{1}{\pi}\int_0^{\infty} s^2\frac{\mathrm{I}(2,s)^2}{\mathrm{I}(1,s)}ds,\end{aligned}$$

where

$$\mathrm{I}(l, s) = \int_0^{\pi} e^{-lg(w)}e^{-se^{-g(w)}}dw, \quad l = 1, 2.$$

We evaluate the integrals in $\mathrm{I}_1(\delta)$ numerically, and obtain that the Cramér-Rao lower bound for estimating δ is approximately $1/\mathrm{I}_1(\delta) = (0.3578)^{-1}$. So the ARE is

$$\frac{\zeta^2}{0.3578(4^{\zeta} - 1)}.$$

This is a concave function that attains a maximum value of 0.942 when $\zeta \approx 1.15$. When $\zeta = 1.0$, the ARE evaluates to 0.932.

Proof of Lemma 3.1 For $\epsilon > 0$ and $t > 0$,

$$\begin{aligned}\mathbb{P}\left(\hat{\delta}_{lm}(\zeta) - \delta \geq \epsilon\right) &= \mathbb{P}\left(\frac{\zeta^{-\zeta}}{k}\sum_{j=1}^{k}\exp(\zeta z_j) \geq e^{\zeta\epsilon}\right) \\ &\leq e^{-tke^{\zeta\epsilon}}\mathbb{E}\exp\left\{\sum_{j=1}^{k}\frac{te^{\zeta z_j}}{\zeta^{\zeta}}\right\},\end{aligned}$$

by the Chernoff bound (Grimmett and Stirzaker, 2001), provided the right hand side converges. Define $T_j = t\zeta^{-\zeta}e^{\zeta z_j}$, $j = 1, \ldots, k$. Then,

$$\begin{aligned}\mathbb{E}\exp\sum_{j=1}^{k}T_j = \left(\mathbb{E}\exp(T_1)\right)^k &= \left\{\sum_{j=0}^{\infty}\mathbb{E}T_1^j/j!\right\}^k \\ &= \left\{\sum_{j=0}^{\infty}t^j j^{\zeta j}/j!\right\}^k.\end{aligned}$$

By the Ratio Test, the series is absolutely convergent for all $t > 0$ if $0 < \zeta < 1$, and for $0 < t < e^{-1}$ if $\zeta = 1$. If $\zeta > 1$, the series is divergent. Define $T = \{t; t > 0\}$ for $0 < \zeta < 1$, and $T = \{t; 0 < t < e^{-1}\}$ for $\zeta = 1$. It follows that, if $\zeta \leq 1$, then $\hat{\delta}_{lm}(\zeta)$ has an exponentially decreasing right tail bound that satisfies

$$\mathbb{P}\left(\hat{\delta}_{lm}(\zeta) - \delta \geq \epsilon\right) < \exp\big(-k\epsilon^2/G_R\big),$$

where

$$\frac{\epsilon^2}{G_R} = \sup_{t\in T}\left\{-\log\Big(\sum_{j=0}^{\infty}\frac{t^j j^{\zeta j}}{j!}\Big) + te^{\zeta\epsilon}\right\}. \qquad (6)$$

It is straightforward to show that the function maximized in (6) is concave. The result follows similarly for the left tail bound. Furthermore by expanding the series in (6) for small values of t we can show that as $\epsilon \to 0$ both G_R and G_L converge to $2(4^{\zeta} - 1)/\zeta^2$. The details are as follows.

Define

$$M_{\zeta}(t) = \sum_{j=0}^{\infty}\frac{t^j j^{\zeta j}}{j!},$$

and consider

$$K_{\zeta}(s, \epsilon) = \big(M_{\zeta}(s\epsilon)\exp(-s\epsilon e^{\zeta\epsilon})\big)^{1/\epsilon^2}, \ s > 0.$$

$K_\zeta(s,\epsilon)$ is a convex function (Grimmett and Stirzaker, 2001), so it follows that $\inf_{s>0} K_\zeta(s,\epsilon) \to \inf_{s>0} K^\star_\zeta(s)$, where $K^\star_\zeta(s)$ is the pointwise limit of $K_\zeta(s,\epsilon)$ as $\epsilon \to 0$, provided this limit exists. Furthermore, since $1/G_R = -\log\big(\inf_{s>0} K_\zeta(s,\epsilon)\big)$, it follows that $\lim_{\epsilon\to 0} G_R = -\big[\log(\inf_{s>0} K^\star_\zeta(s))\big]^{-1}$. To establish the pointwise limit, first note that if $s\epsilon \in T$, then

$$\sum_{j=3}^{\infty} \frac{(s\epsilon)^j j^{\zeta j}}{j!} \leq \epsilon^3 \sum_{j=3}^{\infty} \frac{s^j j^{\zeta j}}{j!} = o(\epsilon^2).$$

So that expanding in powers of ϵ, we have that

$$\begin{aligned}\log(K_\zeta(s,\epsilon)) &= \frac{1}{\epsilon^2}\left[\log\left(1 + s\epsilon + \frac{(s\epsilon)^2 4^\zeta}{2!} + o(\epsilon^2)\right)\right] \\ &\quad + \frac{1}{\epsilon^2}\left[-s\epsilon(1+\zeta\epsilon) + o(\epsilon^2)\right] \\ &= \frac{s^2}{2}\left\{4^\zeta - 1\right\} = K^\star_\zeta(s)\end{aligned}$$

Differentiating with respect to s, we obtain $\inf_{s>0} K^\star_\zeta(s) = -\zeta^2/[2(4^\zeta - 1)]$, as required, where the convexity ensures a unique minimum.