Compressed Counting

and the Application in Estimating Entropy of Data Streams

Ping Li

Department of Statistical Science

Faculty of Computing and Information Science

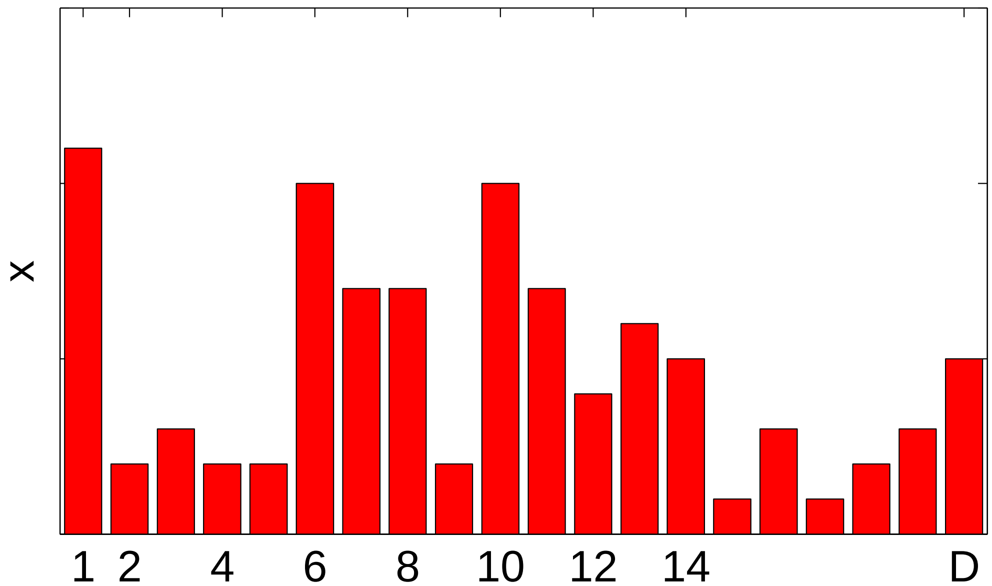
Cornell University

June 17, 2010

What is Counting in This Talk?

Assume a very long vector of D items: x_1, x_2, ..., x_D.
For example, $D = 2^{64}$, or $D = 2^{112}$.

This talk is about counting $\sum_{i=1}^{D} x_i^{\alpha}$, where $0 < \alpha \leq 2$.



The case $\alpha \rightarrow 1$ is particularly interesting and important (eg entropy estimation).

Isn't Counting a Simple (Trivial) Task?

Partially True!, if data are static. However

Real-world data are in general Massive and Dynamic —— Data Streams

- Databases in Amazon, Ebay, Walmart, and search engines
- Internet/telephone traffic, high-way traffic
- Finance (stock) data
- ...
- May need answers in real-time, eg anomaly detection (using entropy).

For example, the Turnstile data stream model for an online bookstore

t=0

0	0	0	0	0	0	....	0
IP 1	IP 2	IP 3	IP 4			....	IP D

t=1 arriving stream = (3, 10) user 3 ordered 10 books

0	0	10	0	0	0	....	0
IP 1	IP 2	IP 3	IP 4			....	IP D

t=2 arriving stream = (1, 5) user 1 ordered 5 books

5	0	10	0	0	0	....	0
IP 1	IP 2	IP 3	IP 4			....	IP D

t=3 arriving stream = (3, −8) user 3 cancelled 8 books

5	0	2	0	0	0	....	0
IP 1	IP 2	IP 3	IP 4			....	IP D

Turnstile Data Stream Model

At time t, an incoming element : $a_t = (i_t, I_t)$

$i_t \in [1, D]$ index, I_t: increment/decrement.

Updating rule : $A_t[i_t] = A_{t-1}[i_t] + I_t$

Goal : Count $F_{(\alpha)} = \sum_{i=1}^{D} A_t[i]^\alpha$

Counting: Trivial if $\alpha = 1$, but Non-trivial in General

Goal : Count $F_{(\alpha)} = \sum_{i=1}^{D} A_t[i]^\alpha$, where $\boxed{A_t[i_t] = A_{t-1}[i_t] + I_t}$.

When $\alpha \neq 1$, counting $F_{(\alpha)}$ exactly requires D counters. (but D can be 2^{64})

When $\alpha = 1$, however, counting the sum is trivial, using a simple counter.

$$F_{(1)} = \sum_{i=1}^{D} A_t[i] = \sum_{s=1}^{t} I_s,$$

The Intuition for $\alpha \approx 1$

There might exist an intelligent counting system which works like a simple counter when α is close 1; and its complexity is a function of how close α is to 1.

Our answer: Yes!

Two caveats:

(1) What if data are negative? Shouldn't we define $F_{(\alpha)} = \sum_{i=1}^{D} |A_t[i]|^{\alpha}$?

(2) Why the case $\alpha \approx 1$ is important ?

The Non-Negativity Constraint

"God created the natural numbers; all the rest is the work of man."

—- by German mathematician Leopold Kronecker (1823 - 1891)

Turnstile model, $a_t = (i_t, I_t), \quad A_t[i_t] = A_{t-1}[i_t] + I_t,$

$I_t > 0$: increment, insertion, eg place orders

$I_t < 0$: decrement, deletion, eg cancel orders,

This talk: Strict Turnstile model $A_t[i] \geq 0$, always.

One can only cancel an order if she/he did place the order!!

Suffices for almost all applications.

Sample Applications of αth Moments (Especially $\alpha \approx 1$)

1. $F_{(\alpha)} = \sum_{i=1}^{D} A_t[i]^\alpha$ itself is a useful summary statistic

   e.g., Rényi entropy, Tsallis entropy, are functions of $F_{(\alpha)}$.

2. Statistical modeling and inference of parameters using method of moments

   Some moments may be much easier to compute than others.

3. $F_{(\alpha)} = \sum_{i=1}^{D} A_t[i]^\alpha$ is a fundamental building element for other algorithms

   Eg., estimating Shannon entropy of data streams

Shannon Entropy of Data Streams

Definition of Shannon Entropy

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

Shannon entropy can be approximated by Rényi Entropy or Tsallis Entropy.

Rényi Entropy

$$H_\alpha = \frac{1}{1-\alpha} \log \frac{F_{(\alpha)}}{F_{(1)}^\alpha} \rightarrow H, \quad \text{as } \alpha \rightarrow 1$$

Tsallis Entropy

$$T_\alpha = \frac{1}{\alpha-1} \left(1 - \frac{F_{(\alpha)}}{F_{(1)}^\alpha}\right) \rightarrow H, \quad \text{as } \alpha \rightarrow 1$$

Algorithms for Estimating Shannon Entropy

- Many algorithms in theoretical CS and databases on estimating entropy.
- A recent trend: Using αth moments to approximate Shannon entropy.
  - Zhao et. al. (IMC07), used symmetric stable random projections (Indyk JACM06, Li SODA08) to approximate moments and Shannon entropy. Mainly an empirical paper.
  - Harvey et. al. (ITW08). A theoretical paper proposed a criterion on how close α is to 1. Used symmetric stable random projections as the underlying algorithm.
  - Harvey et. al. (FOCS08). They proposed refined criteria on how to choose α and cited both symmetric stable random projections and Compressed Counting as underlying algorithms.

Basic Ideas of Estimating Entropy Using Moments

Essentially, to achieve a ν-additive guarantee for the Shannon entropy, it suffices to estimate the αth frequency moment with an $\epsilon = \nu\Delta$-multiplicative guarantee (for sufficiently small Δ, e.g., $\Delta < 10^{-4}$ or even much smaller).

$$(1-\epsilon)F_{(\alpha)} \leq \hat{F}_{(\alpha)} \leq (1+\epsilon)F_{(\alpha)}$$

$$\Longrightarrow$$

$$H - \nu \leq \hat{H}_\alpha \leq H + \nu$$

if $\alpha = 1 - \Delta$ is extremely close to 1.

Recall the definition of Rényi entropy:

$$H_\alpha = \frac{1}{1-\alpha}\log\frac{F_{(\alpha)}}{F_{(1)}^\alpha}$$

Previous Methods for Estimating $F_{(\alpha)}$

- The pioneering work, [AMS STOC'96]
- A popular algorithm, symmetric stable random projections [Indyk JACM'06], [Li SODA'08]
  - Basic idea: Let $X = A_t \times \mathbf{R}$, where entries of $\mathbf{R} \in \mathbb{R}^{D \times k}$ are sampled from a symmetric α-stable distribution. Entries of $X \in \mathbb{R}^k$ are also samples from a symmetric α-stable distribution with the scale = $F_{(\alpha)}$.
  - $k = O\left(1/\epsilon^2\right)$, the large-deviation bound.
    k may be too large for real applications [GC RANDOM'07].
  - While it suggests an algorithm for estimating Shannon Entropy by letting α very close to 1 (Harvey et. al. [ITW08, FOCS08]). The required sample size $O\left(1/\epsilon^2\right)$ with (eg) $\epsilon < 10^{-5}$ can be prohibitive.

Compressed Counting: Skewed Stable Random Projections

Original data stream signal: $A_t[i]$, $i = 1$ to D. eg $D = 2^{64}$

Projected signal: $X_t = A_t \times \mathbf{R} \in \mathbb{R}^k$, k is small.

Projection matrix: $\mathbf{R} \in \mathbb{R}^{D \times k}$,

Sample entries of $\mathbf{R}$ i.i.d. from a skewed stable distribution.

Incremental Projection

Linear Projection: $X_t = A_t \times \mathbf{R}$, $A_t \in \mathbb{R}^D$, $\mathbf{R} \in \mathbb{R}^{D\times k}$.

$+$

Linear data model: $A_t[i_t] = A_{t-1}[i_t] + I_t$

$\Longrightarrow$

Conduct $\boxed{X_t = A_t \times \mathbf{R}}$ incrementally:

$$X_t[j] \leftarrow X_{t-1}[j] + r_{i_t,j} \times I_t, \quad j = 1 \text{ to } k.$$

Generate $r_{i,j}$, entries of $\mathbf{R}$, on-demand

Recover $F_{(\alpha)}$ from Projected Data

$$X_t = (x_1, x_2, ..., x_k) = A_t \times \mathbf{R}$$

$$\mathbf{R} = \{r_{ij}\} \in \mathbb{R}^{D\times k}, \;\; r_{ij} \sim S(\alpha, \beta, 1)$$

$S(\alpha, \beta, \gamma)$: α-stable, β-skewed distribution with scale γ

Then, by stability, at any t, x_j's are i.i.d. stable samples

$$x_j \sim S\left(\alpha, \beta, F_{(\alpha)} = \sum_{i=1}^{D} A_t[i]^\alpha\right)$$

$\Longrightarrow$ A statistical estimation problem.

Review of Skewed Stable Distributions

Z follows a β-skewed α-stable distribution if Fourier transform of its density

$$\mathscr{F}_Z(t) = \mathsf{E}\exp\left(\sqrt{-1}Zt\right) \qquad \alpha \neq 1,$$
$$= \exp\left(-F|t|^{\alpha}\left(1-\sqrt{-1}\beta\text{sign}(t)\tan\left(\frac{\pi\alpha}{2}\right)\right)\right),$$

$0 < \alpha \leq 2$, $-1 \leq \beta \leq 1$. The scale $F > 0$. $Z \sim S(\alpha, \beta, F)$

If $Z_1, Z_2 \sim S(\alpha, \beta, 1)$, independent, then for any $C_1 \geq 0$, $C_2 \geq 0$,

$$Z = C_1 Z_1 + C_2 Z_2 \sim S\left(\alpha, \beta, F = C_1^{\alpha} + C_2^{\alpha}\right).$$

The Statistical Estimation Problem

Task : Given k i.i.d. samples $x_j \sim S\left(\alpha, \beta, F_{(\alpha)}\right)$, estimate $F_{(\alpha)}$.

- No closed-form density in general, but closed-form moments exit.
- Two years ago (Li, SODA 2009):
  - A Geometric Mean estimator based on positive moments.
  - A Harmonic Mean estimator based on negative moments.
  - Their variances are proportional to $O\left(\Delta\right)$, $\Delta = |1-\alpha|$.
  - The complexity bound is $O\left(1/\epsilon\right)$, much better than $O\left(1/\epsilon^2\right)$.
  - To estimate entropy needs, for example, $\Delta < 10^{-4}$, $\epsilon = \nu\Delta < 10^{-5}$.
- Today: a new estimator (Unpublished)
  - The variance is proportional to $O\left(\Delta^2\right)$.
  - The complexity is essentially $O(1)$, or more precisely, $O\left(1/\nu^2\right)$.

The Moment Formula

If $Z \sim S(\alpha, \beta, F_{(\alpha)})$, then for any $\boxed{-1 < \lambda < \alpha}$,

$$\mathbf{E}\left(|Z|^\lambda\right) = F_{(\alpha)}^{\lambda/\alpha} \cos\left(\frac{\lambda}{\alpha}\tan^{-1}\left(\beta\tan\left(\frac{\alpha\pi}{2}\right)\right)\right)$$
$$\times\left(1+\beta^2\tan^2\left(\frac{\alpha\pi}{2}\right)\right)^{\frac{\lambda}{2\alpha}}\left(\frac{2}{\pi}\sin\left(\frac{\pi}{2}\lambda\right)\Gamma\left(1-\frac{\lambda}{\alpha}\right)\Gamma\left(\lambda\right)\right),$$

$\boxed{\lambda = \frac{\alpha}{k}} \Longrightarrow$ an unbiased geometric mean estimator.

The Moment Formula for $\beta = 1$

When $\beta = 1$, then, for $\alpha < 1$ and $\boxed{-\infty < \lambda < \alpha}$,

$$\mathbf{E}\left(|Z|^\lambda\right) = \mathbf{E}\left(Z^\lambda\right) = F_{(\alpha)}^{\lambda/\alpha} \frac{\Gamma\left(1 - \frac{\lambda}{\alpha}\right)}{\cos^{\lambda/\alpha}\left(\frac{\alpha\pi}{2}\right)\Gamma\left(1 - \lambda\right)}.$$

Nice consequence :

Estimators using negative moments will have infinite moments.

The Geometric Mean Estimator for $\beta = 1$

$$\hat{F}_{(\alpha),gm} = \frac{\prod_{j=1}^{k} |x_j|^{\alpha/k}}{D_{gm}}$$

$$\text{Var}\left(\hat{F}_{(\alpha),gm}\right) = \begin{cases} \frac{F_{(\alpha)}^2}{k}\frac{\pi^2}{6}\left(1-\alpha^2\right) + O\left(\frac{1}{k^2}\right), & \text{if } \alpha < 1 \\ \frac{F_{(\alpha)}^2}{k}\frac{\pi^2}{6}\left(\alpha-1\right)\left(5-\alpha\right) + O\left(\frac{1}{k^2}\right), & \text{if } \alpha > 1 \end{cases}$$

As $\alpha \to 1$, the asymptotic variance $\to 0$.

A Geometric Mean Estimator for Symmetric Projections $\beta = 0$

(Li, SODA'08)

Symmetric projections, ie $r_{ij} \sim S(\alpha, \beta = 0, 1)$.

Projected data: $x_j \sim S\left(\alpha, \beta = 0, F_{(\alpha)}\right),\ j = 1$ to k.

Geometric mean estimator:

$$\hat{F}_{(\alpha),gm,sym} = \frac{\prod_{j=1}^{k} |x_j|^{\alpha/k}}{D_{gm,sym}}$$

$$\text{Var}\left(\hat{F}_{(\alpha),gm,sym}\right) = \frac{F_{(\alpha)}^2}{k}\frac{\pi^2}{12}\left(2+\alpha^2\right) + O\left(\frac{1}{k^2}\right),$$

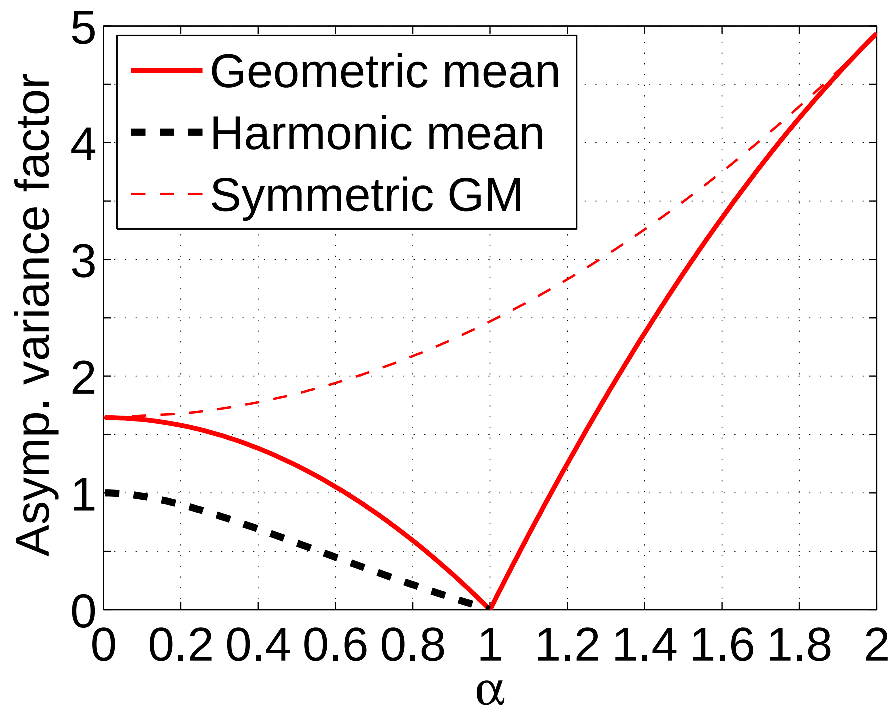
As $\alpha \rightarrow 1$, using skewed projections achieves an "infinite improvement".

A Better Estimator Using Harmonic Mean, for $\alpha < 1$

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

$$\text{Var}\left(\hat{F}_{(\alpha),hm}\right) = \frac{F_{(\alpha)}^2}{k}\left(\Delta + \Delta^2\left(2 - \frac{\pi^2}{6}\right) + O\left(\Delta^3\right)\right) + O\left(\frac{1}{k^2}\right).$$

Comparing Asymptotic Variances



Tail Bounds of the Geometric Mean Estimator

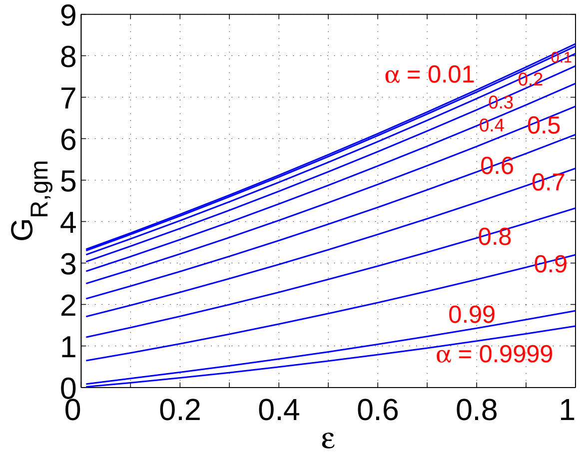
$$\mathbf{Pr}\left(\hat{F}_{(\alpha),gm} - F_{(\alpha)} \geq \epsilon F_{(\alpha)}\right) \leq \exp\left(-k\frac{\epsilon^2}{G_{R,gm}}\right), \ \ \epsilon > 0,$$

$$\mathbf{Pr}\left(\hat{F}_{(\alpha),gm} - F_{(\alpha)} \leq -\epsilon F_{(\alpha)}\right) \leq \exp\left(-k\frac{\epsilon^2}{G_{L,gm}}\right), \ 0 < \epsilon < 1,$$

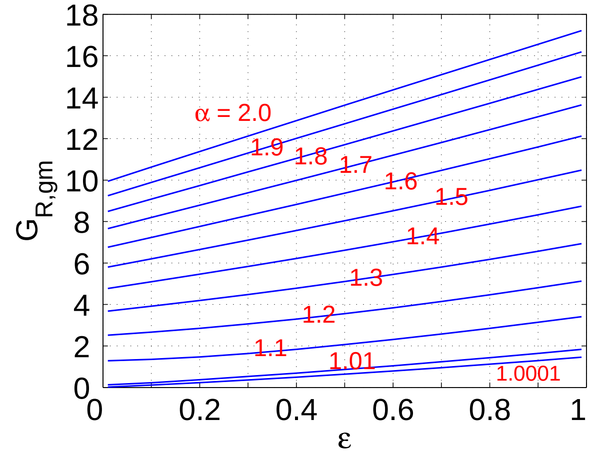
$$\frac{\epsilon^2}{G_{R,gm}} = C_R \log(1+\epsilon) - C_R\gamma_e(\alpha-1)$$
$$-\log\left(\cos\left(\frac{\kappa(\alpha)\pi C_R}{2}\right)\frac{2}{\pi}\Gamma\left(\alpha C_R\right)\Gamma\left(1-C_R\right)\sin\left(\frac{\pi\alpha C_R}{2}\right)\right)$$

C_R is the solution to to

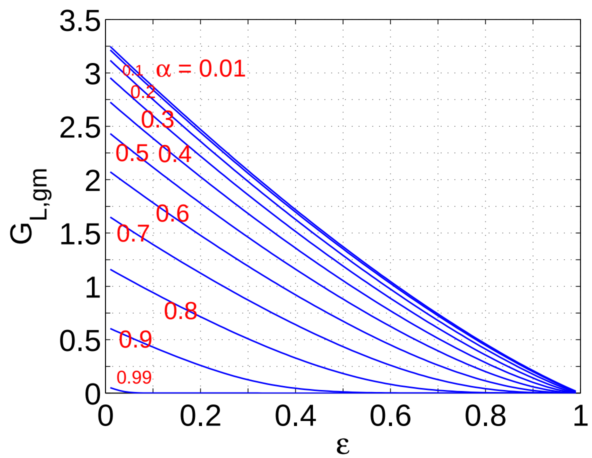
$$-\gamma_e(\alpha-1) + \log(1+\epsilon) + \frac{\kappa(\alpha)\pi}{2}\tan\left(\frac{\kappa(\alpha)\pi}{2}C_R\right)$$
$$-\frac{\alpha\pi/2}{\tan\left(\frac{\alpha\pi}{2}C_R\right)} - \frac{\Gamma'\left(\alpha C_R\right)}{\Gamma\left(\alpha C_R\right)}\alpha + \frac{\Gamma'\left(1-C_R\right)}{\Gamma\left(1-C_R\right)} = 0$$



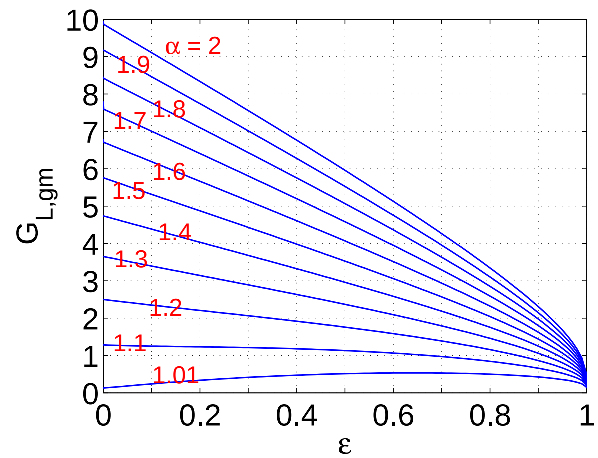
(a) Right bound, $\alpha < 1$



(b) Right bound, $\alpha > 1$



(c) Left bound, $\alpha < 1$



(d) Left bound, $\alpha > 1$

The Sample Complexity Bound

Let $G = \max\{G_{L,gm}, G_{R,gm}\}$.

Bound the error (tail) probability by δ, the level of significance (eg 0.05)

$$\mathbf{Pr}\left(|\hat{F}_{(\alpha),gm} - F_{(\alpha)}| \geq \epsilon F_{(\alpha)}\right) \leq 2\exp\left(-k\frac{\epsilon^2}{G}\right) \leq \delta$$

$$\Longrightarrow k \geq \frac{G}{\epsilon^2}\log\frac{2}{\delta}$$

Sample Complexity Bound (large-deviation bound):

If $k \geq \frac{G}{\epsilon^2}\log\frac{2}{\delta}$, then with probability at least $1-\delta$, $F_{(\alpha)}$ can be approximated within a factor of $1 \pm \epsilon$.

The Sample Complexity for $\alpha = 1 \pm \Delta$

For fixed ϵ, as $\alpha \rightarrow 1$ (i.e., $\Delta \rightarrow 0$),

$$G_{R,gm} = \frac{\epsilon^2}{\log(1+\epsilon) - 2\sqrt{\Delta \log\left(1+\epsilon\right)} + o\left(\sqrt{\Delta}\right)} = O\left(\epsilon\right)$$

If $\alpha > 1,$ then

$$G_{L,gm} = \frac{\epsilon^2}{-\log(1-\epsilon) - 2\sqrt{-2\Delta \log(1-\epsilon)} + o\left(\sqrt{\Delta}\right)} = O\left(\epsilon\right)$$

If $\alpha < 1,$ then

$$G_{L,gm} = \frac{\epsilon^2}{\Delta\left(\exp\left(\frac{-\log(1-\epsilon)}{\Delta} - 1 - \gamma_e\right)\right) + o\left(\Delta \exp\left(\frac{1}{\Delta}\right)\right)} = O\left(\epsilon \exp\left(-\frac{\epsilon}{\Delta}\right)\right)$$

For α close to 1, sample complexity is $O\left(G/\epsilon^2\right) = O\left(1/\epsilon\right)$ not $O\left(1/\epsilon^2\right)$.

New Algorithms/Estimators Are Needed

The geometric mean / harmonic mean estimators are inadequate for estimating Shannon entropy, using either Rényi Entropy or Tsallis Entropy

$$\hat{H}_\alpha = \frac{1}{1-\alpha}\log\frac{\hat{F}_{(\alpha)}}{F_{(1)}^\alpha}, \qquad \hat{T}_\alpha = \frac{1}{\alpha-1}\left(1-\frac{\hat{F}_{(\alpha)}}{F_{(1)}^\alpha}\right)$$

$$Var\left(\hat{H}_{(\alpha)}\right) \propto \frac{1}{(1-\alpha)^2}, \qquad Var\left(\hat{T}_{(\alpha)}\right) \propto \frac{1}{(1-\alpha)^2}.$$

The geometric mean / harmonic mean estimators are inadequate, becuase

- Their variances = $O\left(\Delta\right)$, $\Delta = |1-\alpha|$, are too large to cancel $\frac{1}{(1-\alpha)^2}$.

- The complexity $O\left(1/\epsilon\right)$ is too large as, for example, $\epsilon < 10^{-5}$.

A Recent New Algorithm/Estimator

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

$$x_j \sim S(\alpha, \beta = 1, F_{(\alpha)} \cos\left(\frac{\alpha\pi}{2}\right))$$

$$\Delta = 1 - \alpha$$

Variance and Bias of the New Estimator

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

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

Intuition Behind the New Estimator

Suppose a random variable $Z \sim S\left(\alpha < 1, \beta = 1, \cos\left(\frac{\pi}{2}\alpha\right)\right)$.

A popular way to sample from this distribution (Chambers-Mallows-Stuck method):

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

where $V \sim Uniform(0, \pi)$ and $W \sim Exp(1)$.

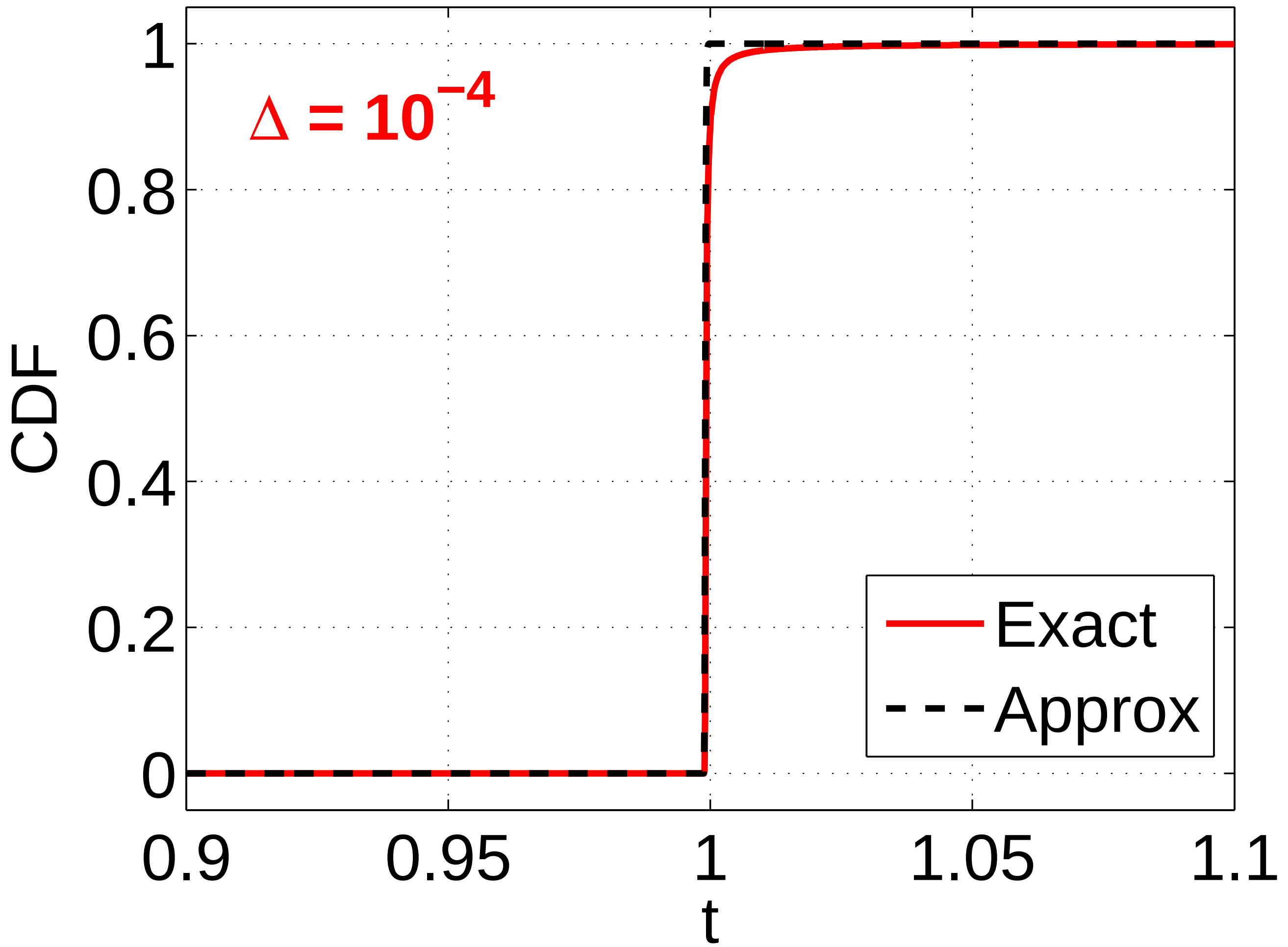
The Cumulative Distribution Function (CDF)

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

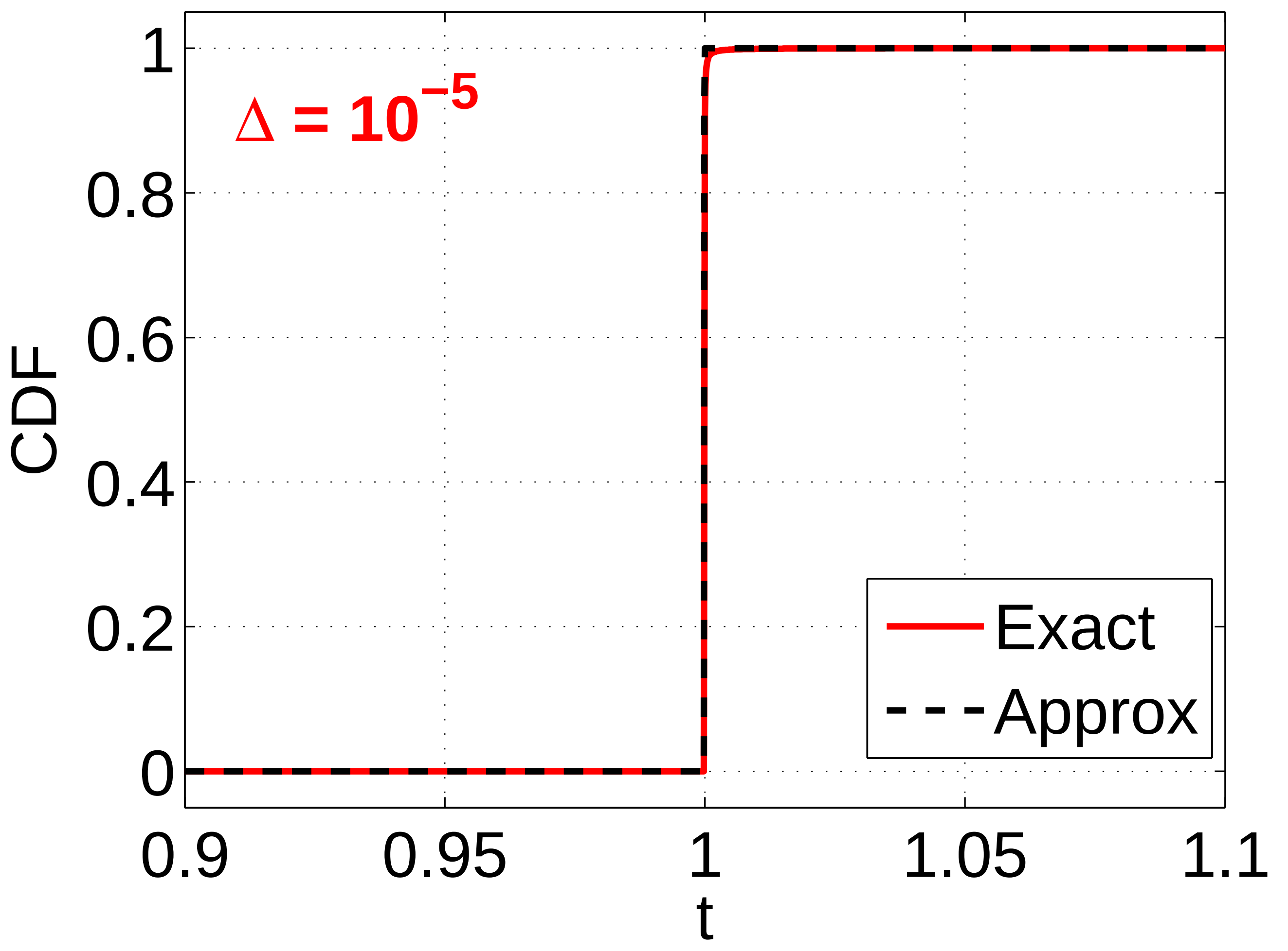
where

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

$$\lim_{\theta \to 0+} g(\theta;\Delta) = g\left(0+;\Delta\right) = \Delta\alpha^{\alpha/\Delta}.$$



Approximate CDF: replacing $g(\theta;\Delta)$ by $g\left(0+;\Delta\right)$



1
0.8
0.6
0.4
0.2
0
CDF
Δ = 10^−5
Exact
Approx
0.9
0.95
1
1.05
1.1
t

The MLE Using Approximate CDF

Consider a random variable Y whose cumulative distribution function (CDF) is

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

Consider an i.i.d. sample Y_j, $j = 1$ to k, and $x_j = cY_j$.

Here c^α is equivalent to our $F_{(\alpha)}$. $\Delta = 1 - \alpha$.

The maximum likelihood estimator (MLE) of c^α (equivalent to our $F_{(\alpha)}$) is

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

very similar to the proposed (guessed) new estimator $\hat{F}_{(\alpha)}$.

If $\Delta = 1-\alpha = 0.1$, then $\Delta^\Delta = 0.7943$, $\alpha^\alpha = 0.9095$.

If $\Delta = 1-\alpha = 0.01$, then $\Delta^\Delta = 0.9550$, $\alpha^\alpha = 0.9901$.

The New Estimator

$$x_j \sim S(\alpha, \beta = 1, F_{(\alpha)} \cos\left(\frac{\alpha\pi}{2}\right))$$

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

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

$$Var\left(\hat{F}_{(\alpha)}\right) = \frac{\Delta^2}{k}F^2_{(\alpha)}\left(3 - 2\Delta + O\left(\frac{1}{k}\right)\right).$$

Tail Bounds of the New Estimator

For any $\epsilon > 0$ and $0 < \Delta = 1 - \alpha < 1$, the right tail bound is

$$\mathbf{Pr}\left(\hat{F}_{(\alpha)} \geq (1+\epsilon)F_{(\alpha)}\right) \leq \exp\left(-k\frac{\epsilon^2}{G_R}\right)$$

$$\frac{\epsilon^2}{G_R} = -\left(\log\sum_{n=0}^{\infty}\frac{(-t_R)^n}{n!}\frac{\Gamma\left(1+\frac{n}{\Delta}\right)}{\Gamma\left(1+\frac{n\alpha}{\Delta}\right)} + \frac{t_R}{(1+\epsilon)^{1/\Delta}\Delta}\right)$$

where t_R is the solution to

$$\frac{\sum_{n=1}^{\infty}\frac{(-1)^n(t_R)^{n-1}}{(n-1)!}\frac{\Gamma\left(1+\frac{n}{\Delta}\right)}{\Gamma\left(1+\frac{n\alpha}{\Delta}\right)}}{\sum_{n=0}^{\infty}\frac{(-t_R)^n}{n!}\frac{\Gamma\left(1+\frac{n}{\Delta}\right)}{\Gamma\left(1+\frac{n\alpha}{\Delta}\right)}} + \frac{1}{(1+\epsilon)^{1/\Delta}\Delta} = 0$$

For any $0 < \epsilon < 1$ and $0 < \Delta = 1 - \alpha < 1$, the left tail bound is

$$\mathbf{Pr}\left(\hat{F}_{(\alpha)} \leq (1-\epsilon)F_{(\alpha)}\right) \leq \exp\left(-k\frac{\epsilon^2}{G_L}\right)$$

$$\frac{\epsilon^2}{G_L} = -\log \sum_{n=0}^{\infty} \frac{(t_L)^n}{n!}\frac{\Gamma\left(1+\frac{n}{\Delta}\right)}{\Gamma\left(1+\frac{n\alpha}{\Delta}\right)} + \frac{t_L}{(1-\epsilon)^{1/\Delta}\Delta}$$

where t_L is the solution to

$$-\frac{\sum_{n=1}^{\infty}\frac{(t_L)^{n-1}}{(n-1)!}\frac{\Gamma\left(1+\frac{n}{\Delta}\right)}{\Gamma\left(1+\frac{n\alpha}{\Delta}\right)}}{\sum_{n=0}^{\infty}\frac{(t_L)^n}{n!}\frac{\Gamma\left(1+\frac{n}{\Delta}\right)}{\Gamma\left(1+\frac{n\alpha}{\Delta}\right)}} + \frac{1}{(1-\epsilon)^{1/\Delta}\Delta} = 0$$

Exact Solution Exists When $\alpha \to 0$ ($\Delta \to 1$)

When $\Delta = 1$, i.e., $\alpha = 0$, then.

$$\frac{\epsilon^2}{G_R} = \log(1+\epsilon) - \frac{\epsilon}{1+\epsilon}, \qquad \epsilon > 0$$
$$\frac{\epsilon^2}{G_L} = \log(1-\epsilon) + \frac{\epsilon}{1-\epsilon}, \qquad 0 < \epsilon < 1.$$

If $\Delta = 1$ ($\alpha = 0$), then $\Gamma\left(1+\frac{n}{\Delta}\right) = n!$, $\Gamma\left(1+\frac{n\alpha}{\Delta}\right) = 1$:

$$\sum_{n=0}^{\infty} \frac{(-t_R)^n}{n!} \frac{\Gamma\left(1+\frac{n}{\Delta}\right)}{\Gamma\left(1+\frac{n\alpha}{\Delta}\right)} = \sum_{n=0}^{\infty} (-t_R)^n = \frac{1}{1+t_R}$$

A Numerically Stable Version of the Tail Bounds

$$\frac{\epsilon^2}{G_R} = -\log\left(1+\sum_{n=1}^{\infty}\left(-t_R\frac{e}{\Delta}\right)^n\prod_{j=0}^{n-1}\frac{n-j\Delta}{(n-j)e}\right) - \left(t_R\frac{e}{\Delta}\right)\frac{1}{e(1+\epsilon)^{1/\Delta}}$$

$$\frac{\epsilon^2}{G_L} = -\log\left(1+\sum_{n=1}^{\infty}\left(t_L\frac{e}{\Delta}\right)^n\prod_{j=0}^{n-1}\frac{n-j\Delta}{(n-j)e}\right) + \left(t_L\frac{e}{\Delta}\right)\frac{1}{e(1-\epsilon)^{1/\Delta}}.$$

————-

Always numerically stable if $\left|t\frac{e}{\Delta}\right| < 1$. Recall $n! \approx \sqrt{2\pi n}\frac{n^n}{e^n}$.

$\Longrightarrow \frac{\epsilon^2}{G} = \frac{\Delta^2\nu^2}{G} = O(1)$, i.e., $G_L = O\left(\Delta^2\right)$ and $G_R = O\left(\Delta^2\right)$.

Theoretical Limits when $\nu \rightarrow 0$

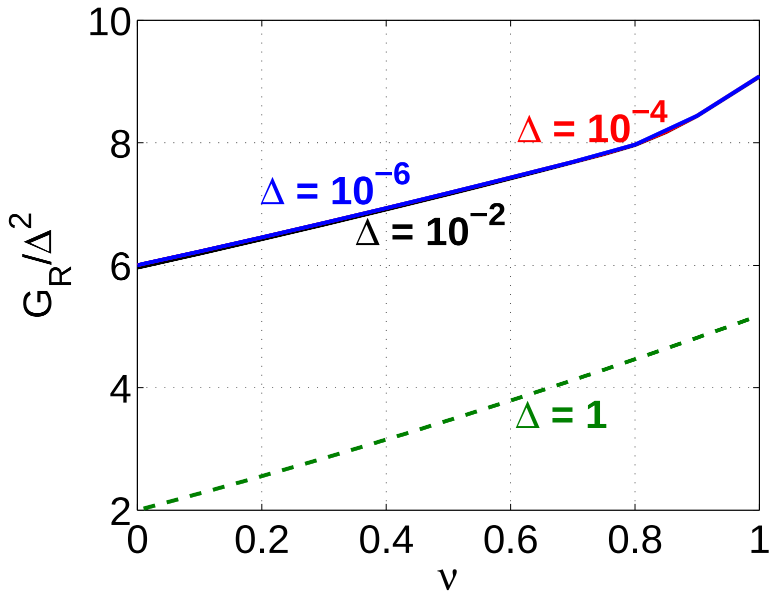
Recall $\epsilon = \nu\Delta$ and ν is the desired additive accuracy of entropy estimation.

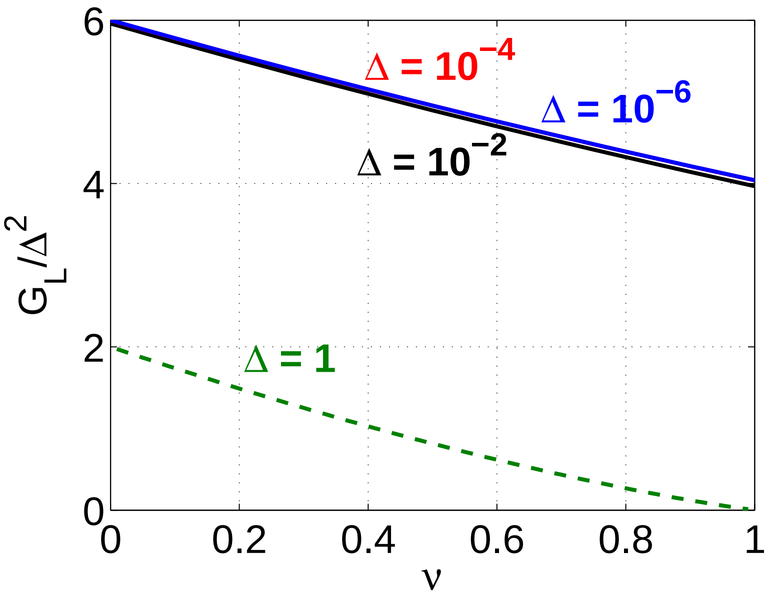
As $\nu \rightarrow 0$,

$$\frac{G_R}{\Delta^2} \rightarrow 6 - 4\Delta, \qquad\qquad \frac{G_L}{\Delta^2} \rightarrow 6 - 4\Delta.$$

Numerical Values of Tail Bound Constants

Right bound $\frac{G_R}{\Delta^2}$ Left bound $\frac{G_L}{\Delta^2}$





Complexity of Entropy Estimation Using the New Estimator

The new estimator provides a very satisfactory solution.

- The sample complexity for entropy estimation is $O\left(9/\nu^2\right)$.

  The constant 9 can be replaced by 6 when ν is small.

- Previous bound in FOCS08 is about $\left(10^6 \log M/\nu^2\right)$, where M is the "universe size." The constant, e.g., 10^6, may vary depending on a few parameters.

- Empirically, only $k = 10$ samples achieve good estimates.

An Empirical Study

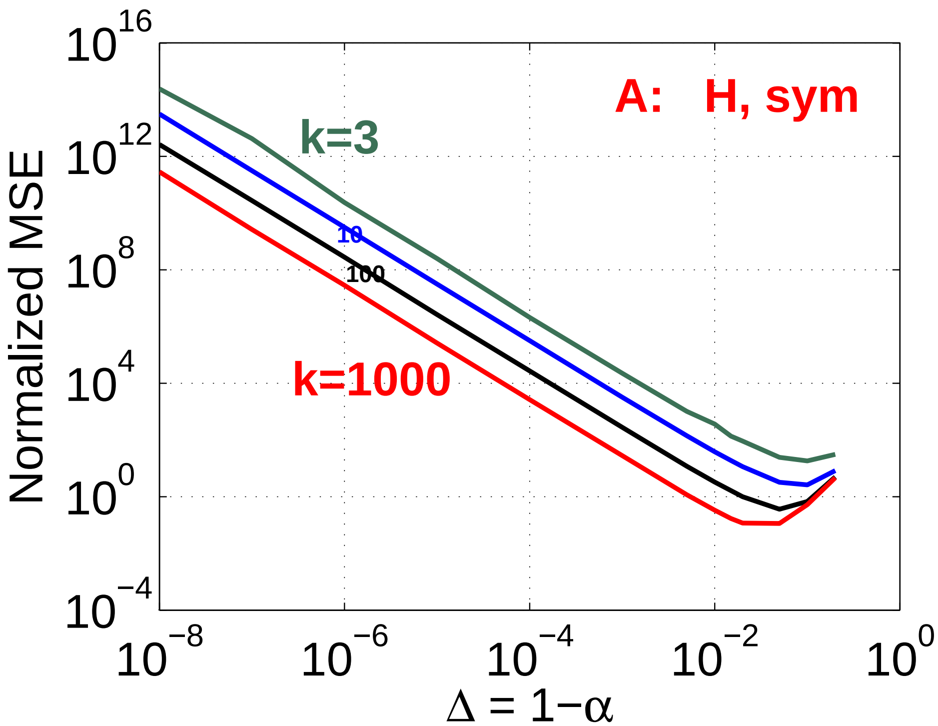
Data

Since estimation accuracy is what we care, we simply use static data instead of data streams. The projected vector $X = \mathbf{R}^{\mathsf{T}} A_t$ is the same, regardless whether it is computed at once (i.e., static) or incrementally (i.e., dynamic).

Eight English words are selected from a chunk of Web crawl data. Our data set consists of 8 vectors and the entries are the numbers of word occurrences in each document.

Word	Sparsity	Entropy H
TWIST	0.004	5.4873
FRIDAY	0.034	7.0487
FUN	0.047	7.6519
BUSINESS	0.126	8.3995
NAME	0.144	8.5162
HAVE	0.267	8.9782
THIS	0.423	9.3893
A	0.596	9.5463

Entropy Estimation Using Symmetric Stable Projections

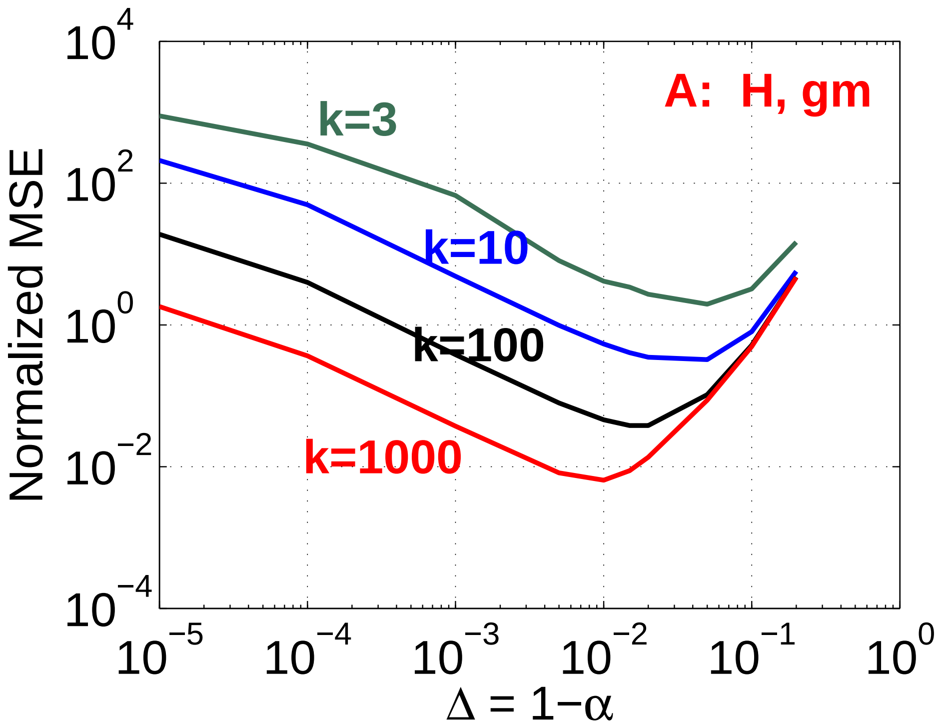


Y-axis: Normalized Mean Square Error (MSE)

The errors are huge if $\alpha = 1 - \Delta$ is too close to 1.

Even with $k = 1000$ samples, the smallest possible errors are still very large.

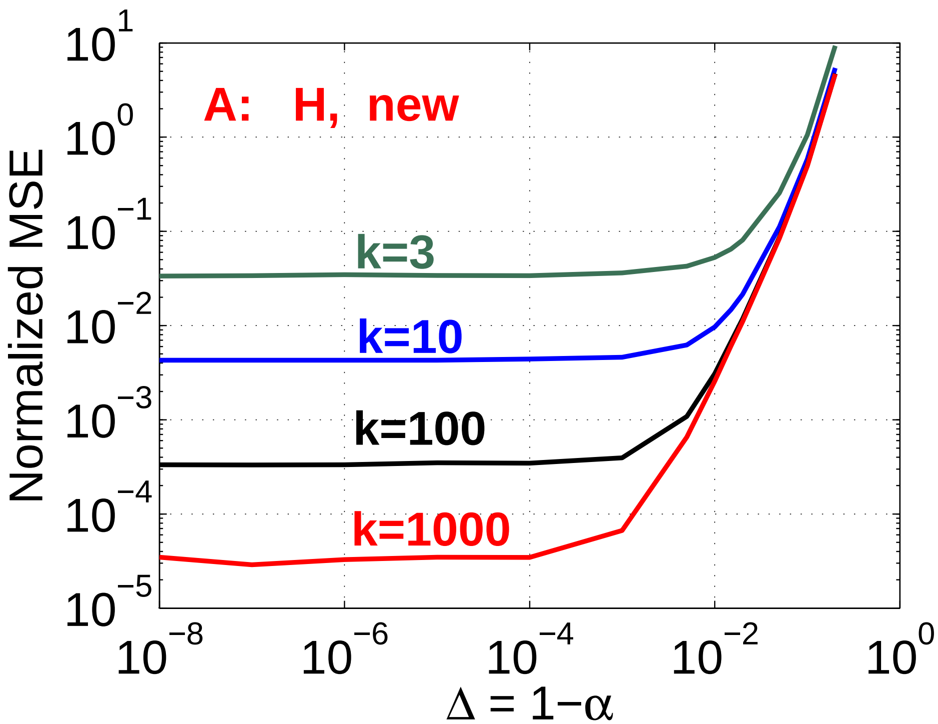
Entropy Estimation Using CC with Geometric Mean Estimator



Much smaller errors compared to using symmetric projections.

The errors still increase if $\alpha = 1 - \Delta$ is too close to 1. With $k = 1000$ samples, it is possible to obtain good estimates if α is chosen carefully.

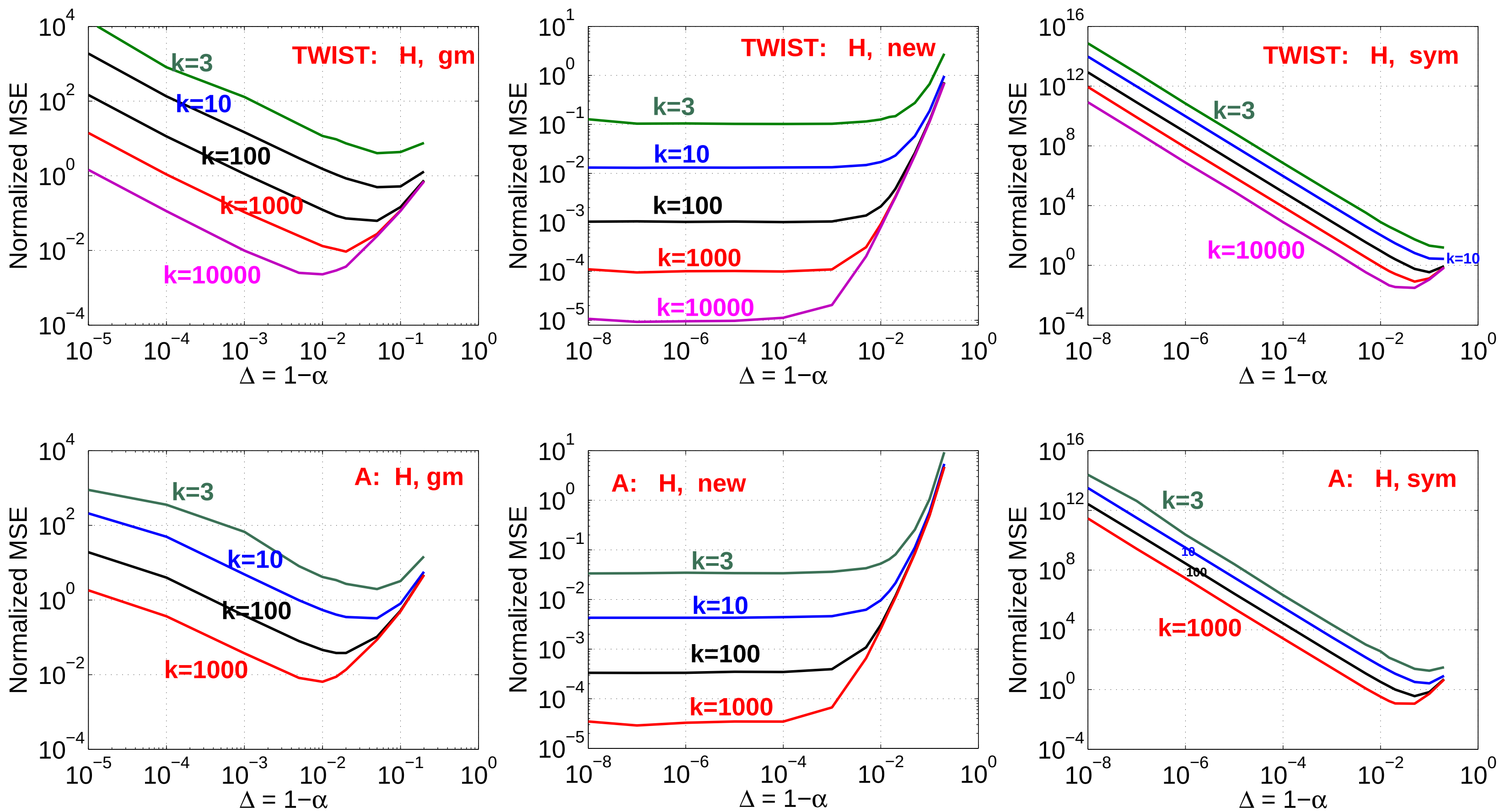
Entropy Estimation Using CC with the New Estimator

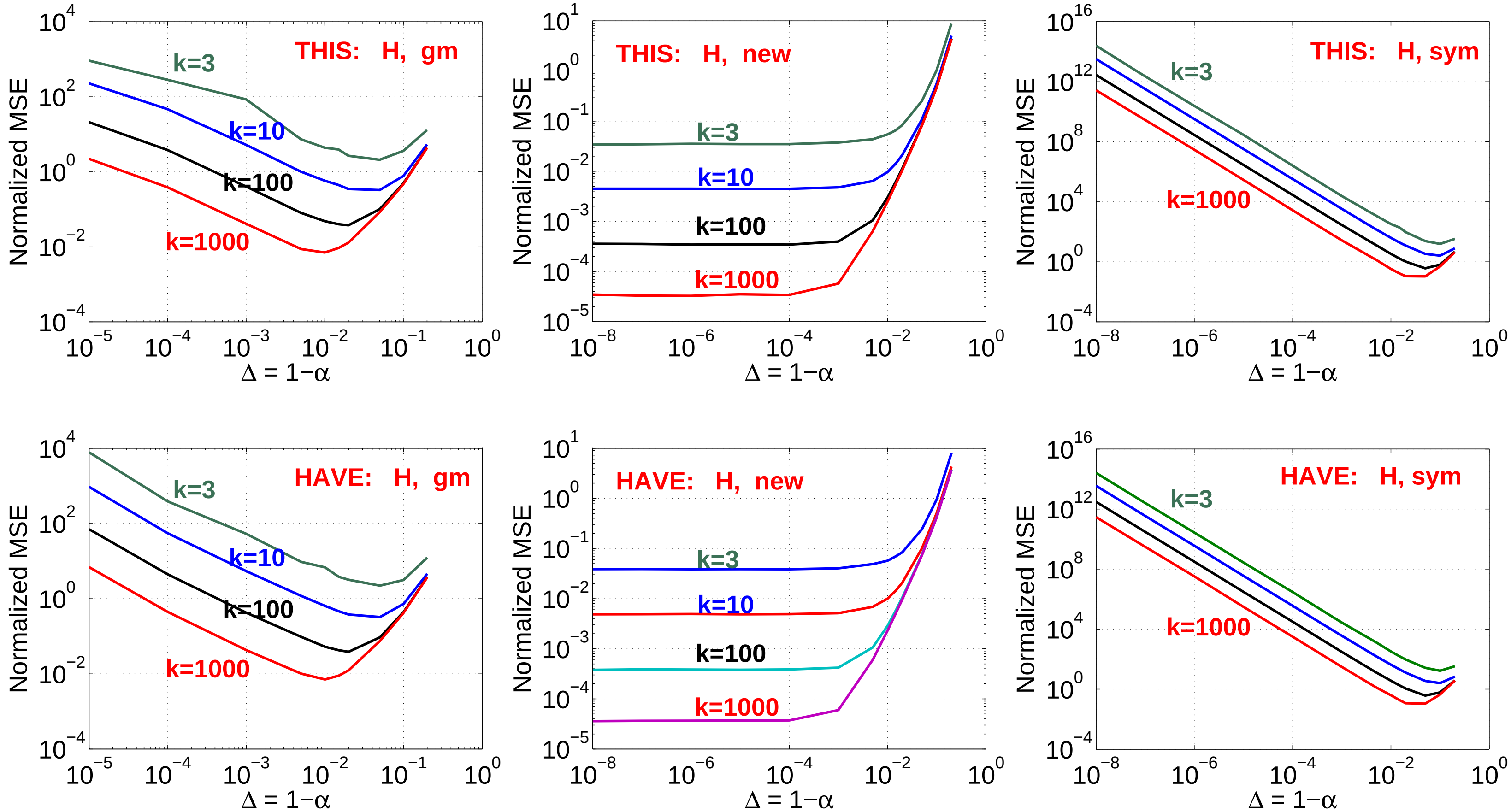


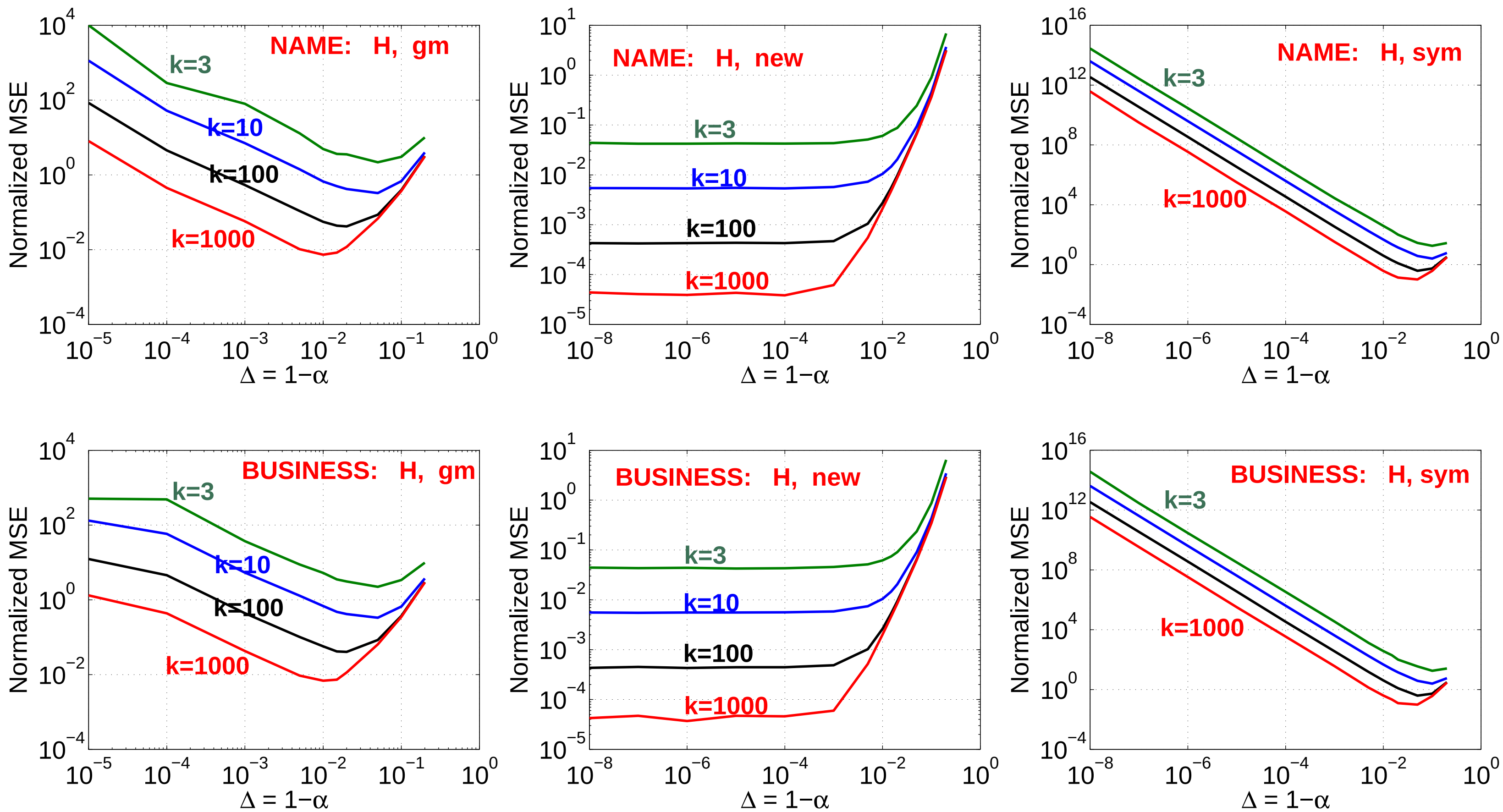
Only $k = 10$ (or even $k = 3$) samples are needed to produce good estimates.

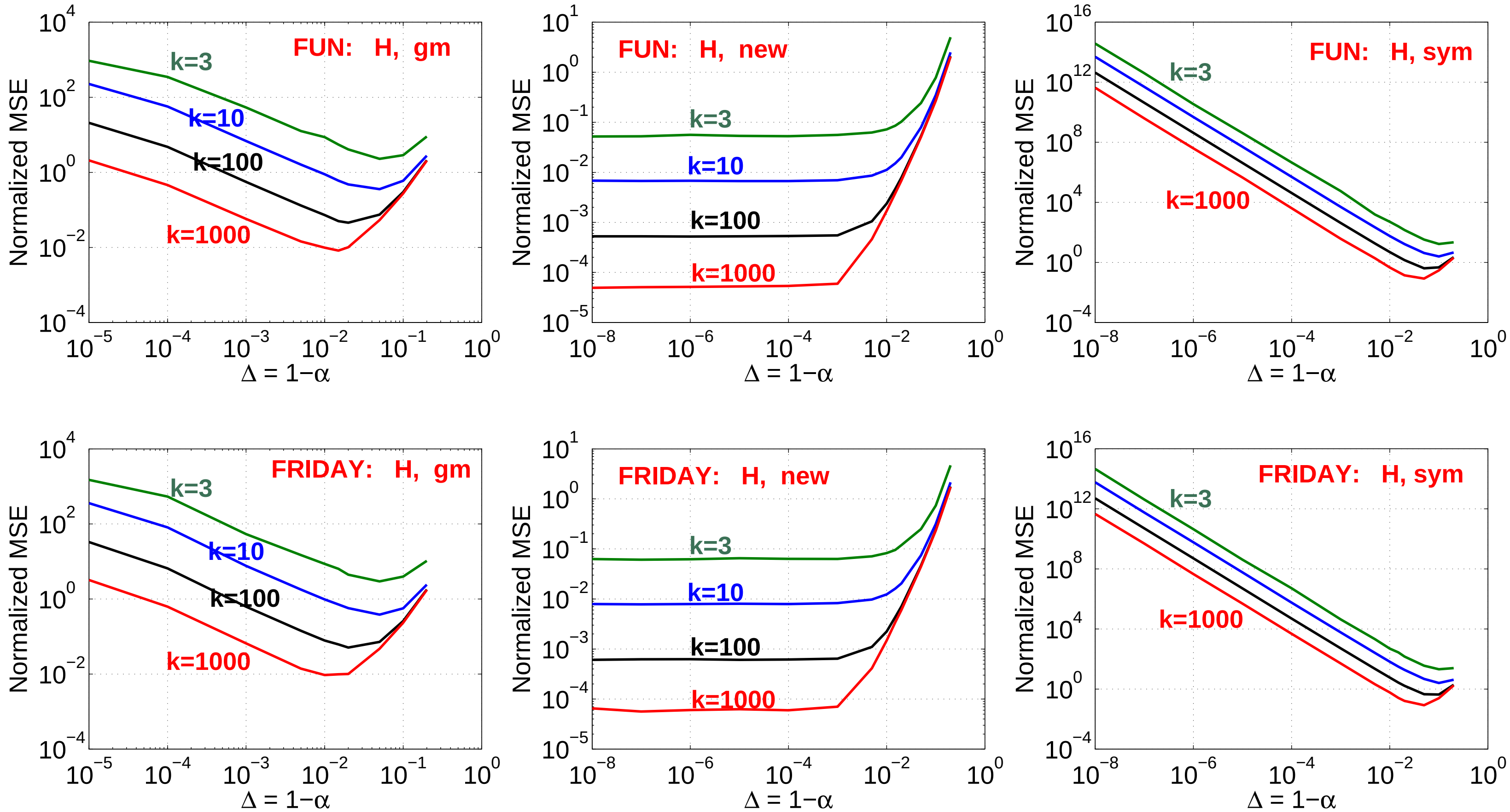
The errors do not increase as $\alpha = 1 - \Delta$ is closer and closer to 1.

Shannon Entropy Estimation Results for All Vectors




THIS: H, gm
k=3
k=10
k=100
k=1000
Normalized MSE
Δ = 1−α
THIS: H, new
k=3
k=10
k=100
k=1000
Normalized MSE
Δ = 1−α
THIS: H, sym
k=3
k=1000
Normalized MSE
Δ = 1−α
HAVE: H, gm
k=3
k=10
k=100
k=1000
Normalized MSE
Δ = 1−α
HAVE: H, new
k=3
k=10
k=100
k=1000
Normalized MSE
Δ = 1−α
HAVE: H, sym
k=3
k=1000
Normalized MSE
Δ = 1−α


NAME: H, gm
k=3
k=10
k=100
k=1000
Normalized MSE
Δ = 1−α
NAME: H, new
NAME: H, sym
BUSINESS: H, gm
BUSINESS: H, new
BUSINESS: H, sym


FUN: H, gm
k=3
k=10
k=100
k=1000
Normalized MSE
Δ = 1−α
FUN: H, new
FUN: H, sym
FRIDAY: H, gm
FRIDAY: H, new
FRIDAY: H, sym

Conclusions

- The α-th frequency moments of data streams have very important applications when $\alpha \approx 1$, eg. estimating entropy for anomaly detection.

- Well-known methods based on symmetric stable random projections do not capture the intuition that estimating α-th moments should be easy if $\alpha \approx 1$.

- Compressed Counting (CC) (maximally-skewed stable random projections) can provide the mechanism for dramatically improving estimates near $\alpha = 1$.

- To estimate Shannon entropy, the estimator of frequency moments should have variance decreasing to zero at the rate of $O\left(\Delta^2\right)$, $\Delta = |1-\alpha|$. Equivalently, the complexity should be essentially $O\left(1\right)$.

- The previous work on CC (two years ago) only achieved variances = $O\left(\Delta\right)$ and complexity =$O\left(1/\epsilon\right)$, but $\epsilon = O(\Delta)$ is extremely small.

- The new estimator (this talk) has achieved variance = $O\left(\Delta^2\right)$ and complexity = $O(1)$. It provides a practically satisfactory solution to the long-standing entropy estimation problem.

Acknowledgement

ONR-YIP: N000140910911

NSF-DMS: FODAVA-0808864