Compressed Counting

Ping Li

Department of Statistical Science

Faculty of Computing and Information Science

Cornell University

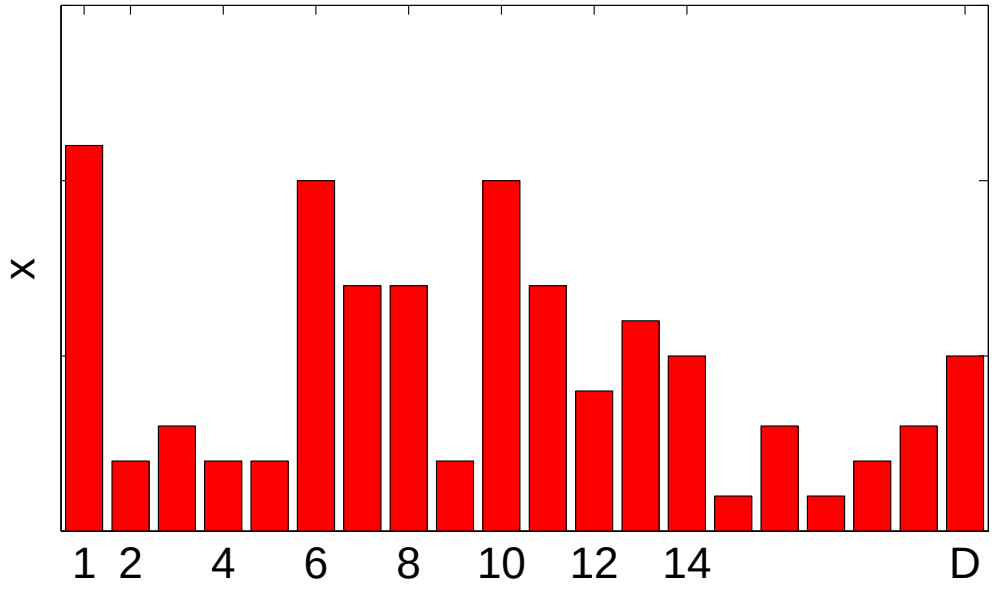
Ithaca, NY 14850

March, 2009

What is Counting in This Talk?

Assume a very long vector of D items: x_1, x_2, ..., x_D.

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



The case $\alpha \to 1$ is particularly interesting and important.

Related Summary Statistics

- The sum $\sum_{i=1}^{D} x_i$. The number of non-zeros, $\sum_{i=1}^{D} 1_{x_i \neq 0}$
- The αth moment $F_{(\alpha)} = \sum_{i=1}^{D} x_i^{\alpha}$

  $F_{(1)}$ =the sum, $F_{(2)}$ = the power/energy, $F_{(0)}$ = number of non-zeros.
- The future fortune, $\sum_{i=1}^{D} x_i^{1 \pm \Delta}$, Δ = interest/decay rate (usually small)
- The entropy moment $\sum_{i=1}^{D} x_i \log x_i$ and entropy $\sum_{i=1}^{D} \frac{x_i}{F_{(1)}} \log \frac{x_i}{F_{(1)}}$
- The Tsallis Entropy $\frac{1 - F_{(\alpha)}/F_{(1)}^{\alpha}}{\alpha - 1}$ The Rényi Entropy $\frac{1}{1-\alpha} \log \frac{F_{(\alpha)}}{F_{(1)}^{\alpha}}$

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 : $\boxed{a_t = (i_t, I_t)}$

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

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

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

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

Tsallis Entropy

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

Algorithms on 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.
  - 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.

Anomaly Detection in Large Networks Using Entropy of Traffic

Example: Laura Feinstein, Dan Schnackenberg, Ravindra Balupari, and Darrell Kindred. Statistical approaches to DDoS attack detection and response. In DARPA Information Survivability Conference and Exposition, 2003

General idea: Anomaly events (such as failure of service, distributed denial of service (DoS) attacks) change the the distribution of the traffic data.

The change of distribution can be characterized by the change of entropy.

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].

Compressed Counting: Skewed Stable Random Projections

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

Projected signal: $X_t = A_t \times \mathbf{R} \quad \in \mathbb{R}^k, \quad k$ is small (eg $k = 20 \sim 100$)

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

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

The Standard Data Stream Technique: Incremental Projection

Linear Projection: $X_t = A_t \times \mathbf{R}$

$+$

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.

Generate entries of $\mathbf{R}$ on-demand

Our method differs from previous algorithms in the choice of the distribution of $\mathbf{R}$.

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

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

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

$S(\alpha, \beta, \gamma)$: 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 $C_1 \geq 0, C_2 \geq 0$,

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

If C_1 and C_2 do not have the same signs, the "stability" does not hold.

Let $Z = C_1 Z_1 - C_2 Z_2$, with $C_1 \geq 0$ and $C_2 \geq 0$.

Because $\mathscr{F}_{-Z_2}(t) = \mathscr{F}_{Z_2}(-t)$,

$$\mathscr{F}_Z(t) = \exp\left(-|C_1 t|^\alpha \left(1 - \sqrt{-1}\beta \text{sign}(t) \tan\left(\frac{\pi\alpha}{2}\right)\right)\right)$$
$$\times \exp\left(-|C_2 t|^\alpha \left(1 + \sqrt{-1}\beta \text{sign}(t) \tan\left(\frac{\pi\alpha}{2}\right)\right)\right),$$

Does NOT represent a stable law, unless $\beta = 0$ or $\alpha = 2, 0+$.

Symmetric ($\beta = 0$) projections work for any data,

but if data are non-negative, benefits of skewed projection are enormous.

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.
- A Geometric Mean estimator based on positive moments.
- A Harmonic Mean estimator based on negative moments.
- Both estimators exhibit exponential error (tail) bounds.

The Moment Formula

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

$$\boldsymbol{E}\left(|Z|^\lambda\right) = F_{(\alpha)}^{\lambda/\alpha} \cos\left(\frac{\lambda}{\alpha}\tan^{-1}\left(\beta\tan\left(\frac{\alpha\pi}{2}\right)\right)\right)$$

$$\times\left(1+\beta^2\tan^2\left(\frac{\alpha\pi}{2}\right)\right)^{\frac{\lambda}{2\alpha}}\left(\frac{2}{\pi}\sin\left(\frac{\pi}{2}\lambda\right)\Gamma\left(1-\frac{\lambda}{\alpha}\right)\Gamma\left(\lambda\right)\right),$$

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

Nice things happen when $\beta = 1$.

Lemma 2 *When $\beta = 1$, then, for $\alpha < 1$ and* $\boxed{-\infty < \lambda < \alpha}$,

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

Nice consequence :

Estimators using negative moments will have infinite moments.

$\Longrightarrow$ Good statistical properties.

The Geometric Mean Estimator for all β

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

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

$$D_{gm,\beta} = \cos^k\left(\frac{1}{k}\tan^{-1}\left(\beta\tan\left(\frac{\alpha\pi}{2}\right)\right)\right) \times$$

$$\left(1+\beta^2\tan^2\left(\frac{\alpha\pi}{2}\right)\right)^{\frac{1}{2}} \left[\frac{2}{\pi}\sin\left(\frac{\pi\alpha}{2k}\right)\Gamma\left(1-\frac{1}{k}\right)\Gamma\left(\frac{\alpha}{k}\right)\right]^k.$$

Which β ? : Variance of $\hat{F}_{(\alpha),gm,\beta}$ is decreasing in $\beta \in [0, 1]$.

$$\text{Var}\left(\hat{F}_{(\alpha),gm,\beta}\right) = F_{(\alpha)}^2 V_{gm,\beta}$$

$$V_{gm,\beta} = \left[2 - \sec^2\left(\frac{1}{k}\tan^{-1}\left(\beta\tan\left(\frac{\alpha\pi}{2}\right)\right)\right)\right]^k \times \frac{\left[\frac{2}{\pi}\sin\left(\frac{\pi\alpha}{k}\right)\Gamma\left(1-\frac{2}{k}\right)\Gamma\left(\frac{2\alpha}{k}\right)\right]^k}{\left[\frac{2}{\pi}\sin\left(\frac{\pi\alpha}{2k}\right)\Gamma\left(1-\frac{1}{k}\right)\Gamma\left(\frac{\alpha}{k}\right)\right]^{2k}} - 1,$$

A decreasing function of $\beta \in [0, 1]$. $\Longrightarrow$ Use $\beta = 1$, maximally skewed

The Geometric Mean Estimator for $\beta = 1$

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

Lemma 3

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

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

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

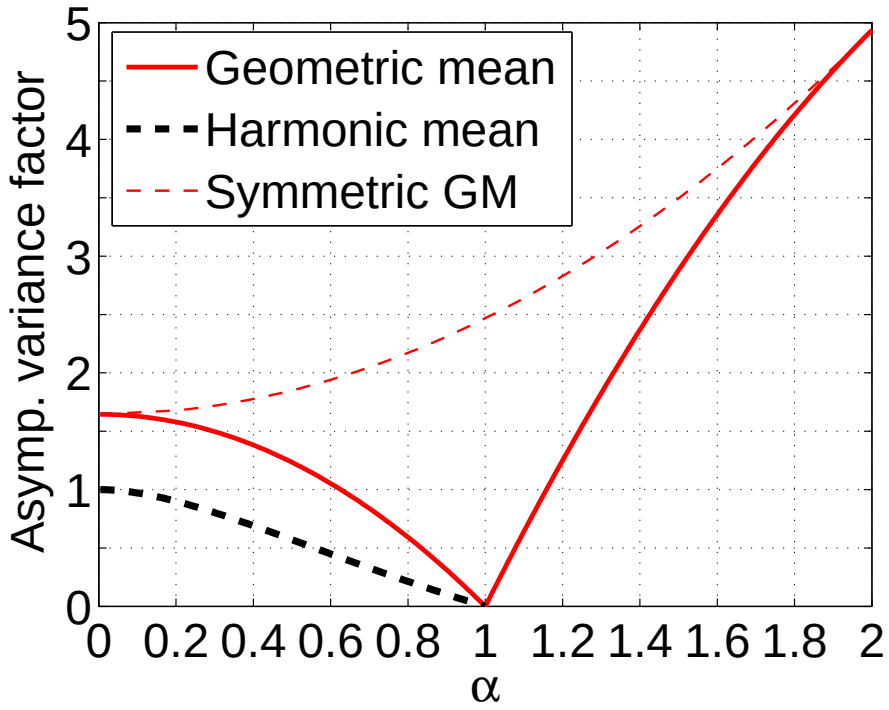
Skewed Projections $(\beta = 1)$

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

Advantages of $\hat{F}_{(\alpha),hm}$

- Smaller variance
- Smaller tail bound constant
- Moment generating function exits.

Comparing Asymptotic Variances



Tail Bounds of the Geometric Mean Estimator

Lemma 4

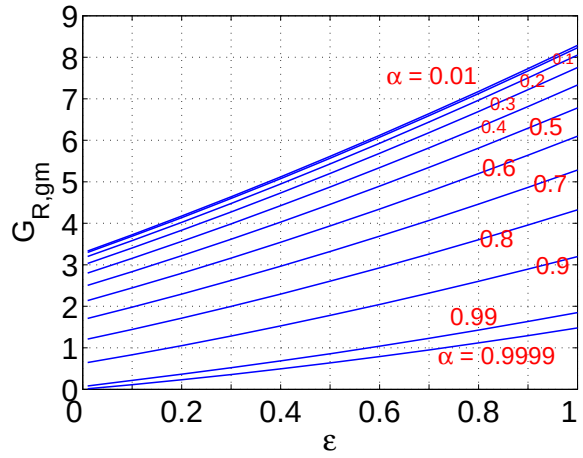
$$\mathbf{Pr}\left(\hat{F}_{(\alpha),gm} - F_{(\alpha)} \geq \epsilon F_{(\alpha)}\right) \leq \exp\left(-k\frac{\epsilon^2}{G_{R,gm}}\right), \quad \epsilon > 0,$$

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

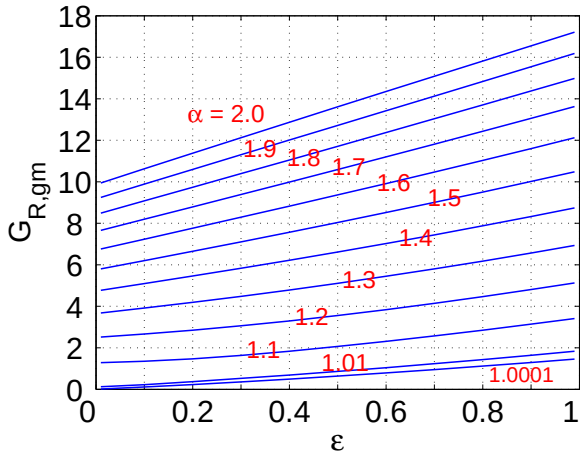
$$\frac{\epsilon^2}{G_{R,gm}} = C_R \log(1+\epsilon) - C_R\gamma_e(\alpha-1)$$
$$- \log\left(\cos\left(\frac{\kappa(\alpha)\pi C_R}{2}\right)\frac{2}{\pi}\Gamma\left(\alpha C_R\right)\Gamma\left(1-C_R\right)\sin\left(\frac{\pi\alpha C_R}{2}\right)\right)$$

C_R is the solution to to

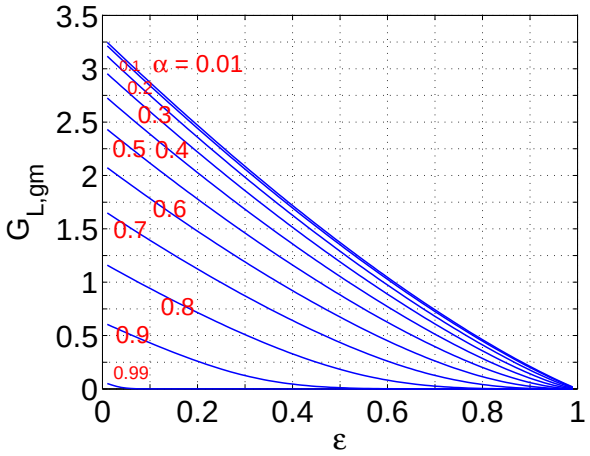
$$-\gamma_e(\alpha-1) + \log(1+\epsilon) + \frac{\kappa(\alpha)\pi}{2}\tan\left(\frac{\kappa(\alpha)\pi}{2}C_R\right)$$
$$-\frac{\alpha\pi/2}{\tan\left(\frac{\alpha\pi}{2}C_R\right)} - \frac{\Gamma'\left(\alpha C_R\right)}{\Gamma\left(\alpha C_R\right)}\alpha + \frac{\Gamma'\left(1-C_R\right)}{\Gamma\left(1-C_R\right)}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



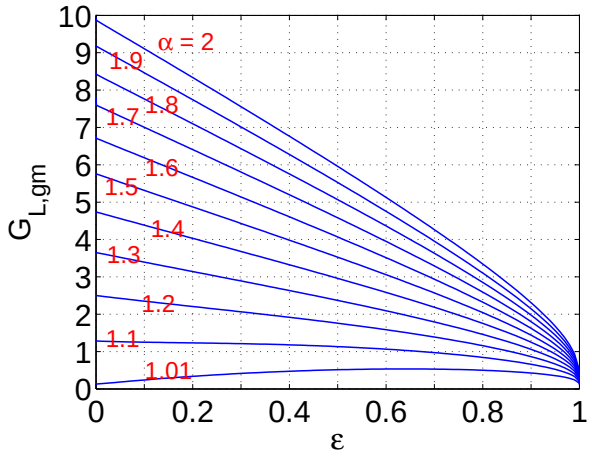
(a) Right bound, $\alpha < 1$



(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 $O\left(1/\epsilon^2\right)$ bound in general can not be improved — Central Limit Theorem

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

Lemma 5 *For fixed ϵ, as $\alpha \to 1$ (i.e., $\Delta \to 0$),*

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

If $\alpha > 1$, then

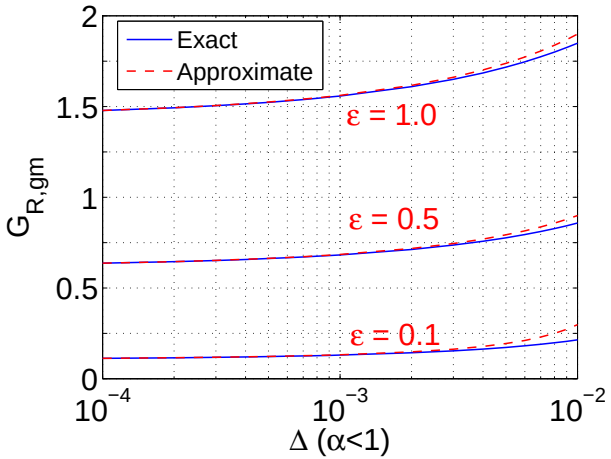
$$G_{L,gm} = \frac{\epsilon^2}{-\log(1-\epsilon) - 2\sqrt{-2\Delta \log(1-\epsilon)} + o\left(\sqrt{\Delta}\right)} = O(\epsilon)$$

If $\alpha < 1$, then

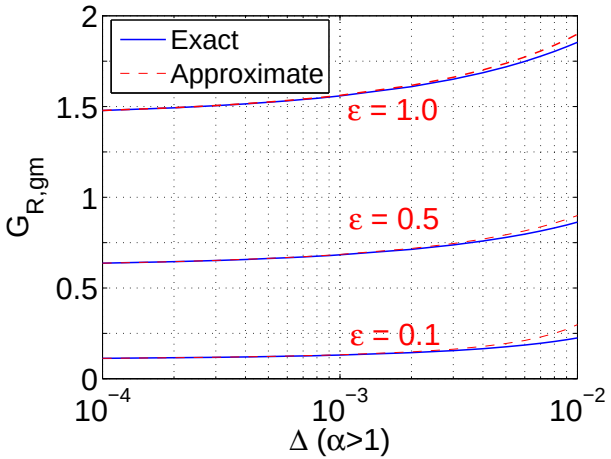
$$G_{L,gm} = \frac{\epsilon^2}{\Delta\left(\exp\left(\frac{-\log(1-\epsilon)}{\Delta} - 1 - \gamma_e\right)\right) + o\left(\Delta \exp\left(\frac{1}{\Delta}\right)\right)} = O\left(\epsilon \exp\left(-\frac{\epsilon}{\Delta}\right)\right)$$

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

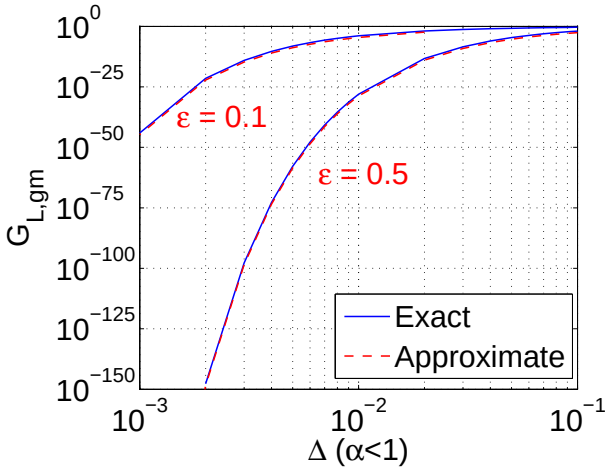
Not violating fundamental principles.



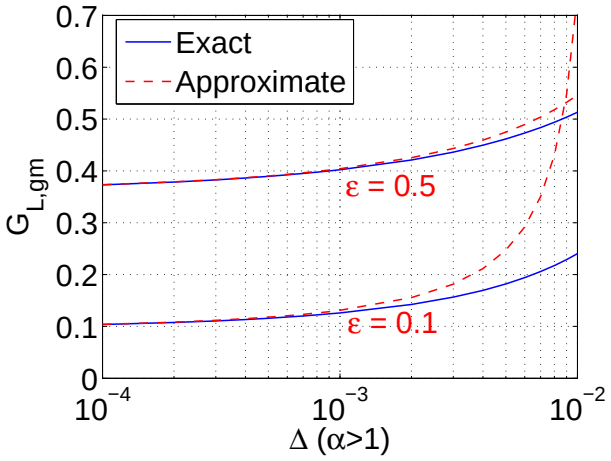
(e) Right bound, $\alpha < 1$



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



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



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

Sampling From Maximally-Skewed Stable Distributions

To sample from $Z \sim S(\alpha, \beta = 1, 1)$:

$$W \sim \exp(1) \qquad\qquad U \sim \text{Uniform}\left(-\frac{\pi}{2}, \quad \frac{\pi}{2}\right)$$

$$\rho = \begin{cases} \frac{\pi}{2} & \alpha < 1 \\ \frac{\pi}{2}\frac{2-\alpha}{\alpha} & \alpha > 1 \end{cases}$$

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

$\cos^{1/\alpha}\left(\rho\alpha\right)$ can be removed and later reflected in the estimators.

Sampling from Skewed distributions is as easy as from symmetric distributions.

Empirical Study of CC

Goals:

- Demonstrate the huge improvement of CC over symmetric projections.
- Illustrate that CC is highly efficient in estimating Shannon entropy.
  Exploiting the bias-variance trade-off is the key.

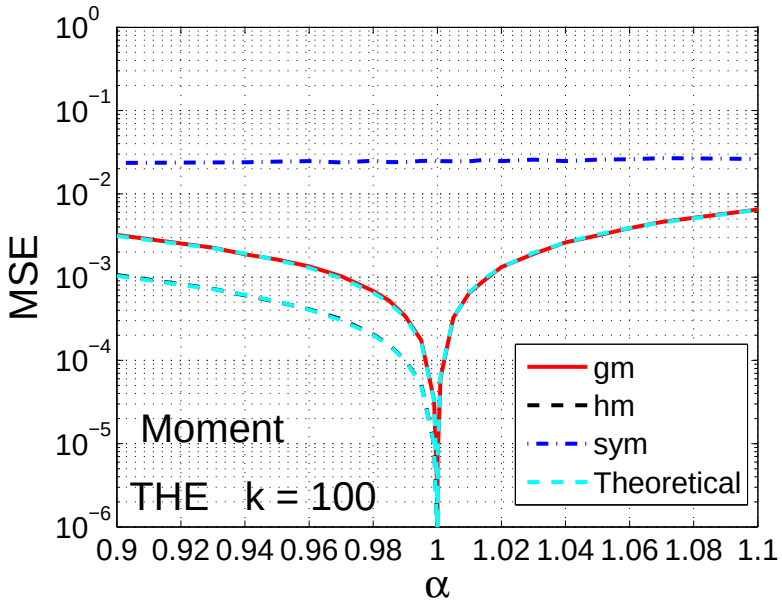
Data:

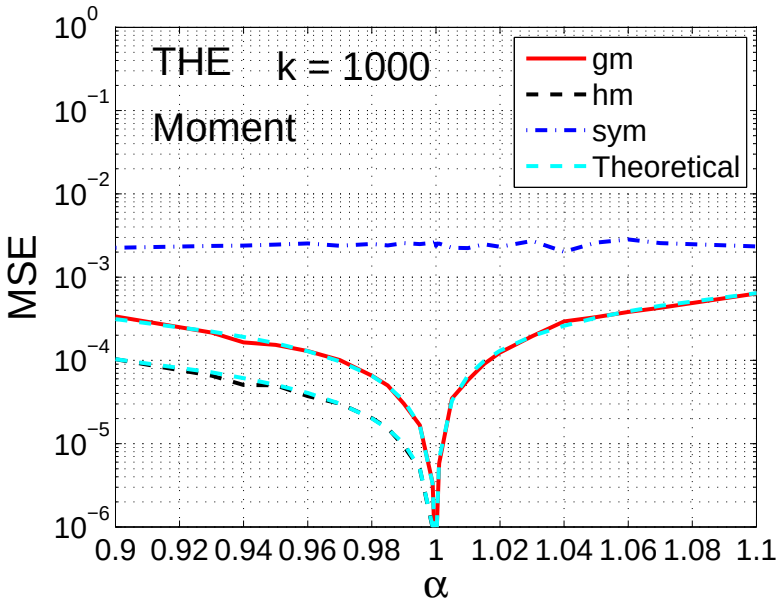
- 10 English words from a chuck of MSN Web crawl with $D = 2^{64}$ documents.
  Each word is a vector of length D whose entries are number of occurrences
- Static data suffice for comparing the estimation accuracy.
  $X_t = A_t \times \mathbf{R}$ is the same, whether it is computed in one time (static) or incrementally (dynamic).

Word	Nonzero	H	$H_{0.95}$	$H_{1.05}$	$T_{0.95}$	$T_{1.05}$
TWIST	274	5.4873	5.4962	5.4781	6.3256	4.7919
RICE	490	5.4474	5.4997	5.3937	6.3302	4.7276
FRIDAY	2237	7.0487	7.1039	6.9901	8.5292	5.8993
FUN	3076	7.6519	7.6821	7.6196	9.3660	6.3361
BUSINESS	8284	8.3995	8.4412	8.3566	10.502	6.8305
NAME	9423	8.5162	9.5677	8.4618	10.696	6.8996
HAVE	17522	8.9782	9.0228	8.9335	11.402	7.2050
THIS	27695	9.3893	9.4370	9.3416	12.059	7.4634
A	39063	9.5463	9.5981	9.4950	12.318	7.5592
THE	42754	9.4231	9.4828	9.3641	12.133	7.4775

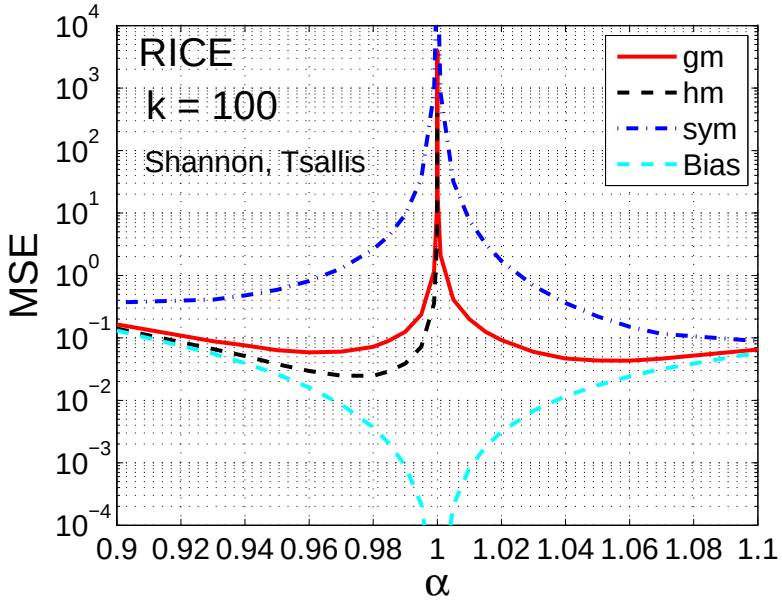
Results are similar across words, measured by normalized MSE = Bias^2 + Var.

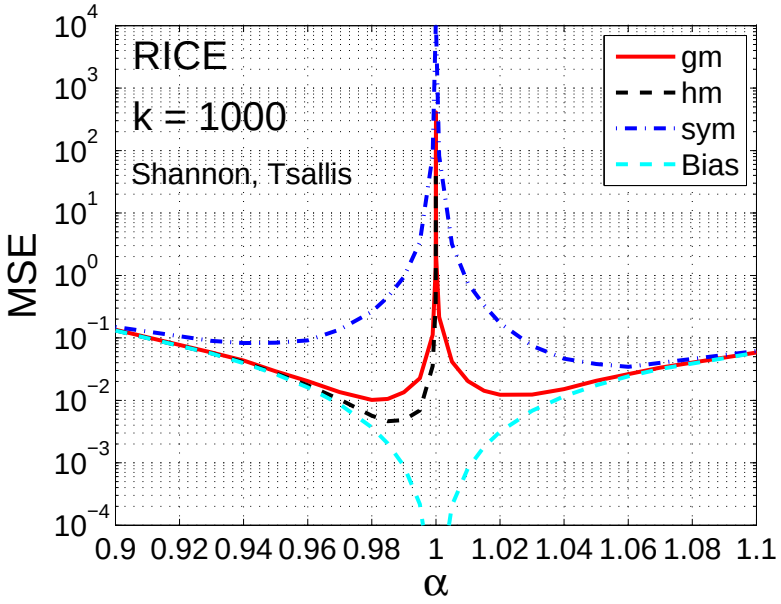
Estimating Frequency Moments



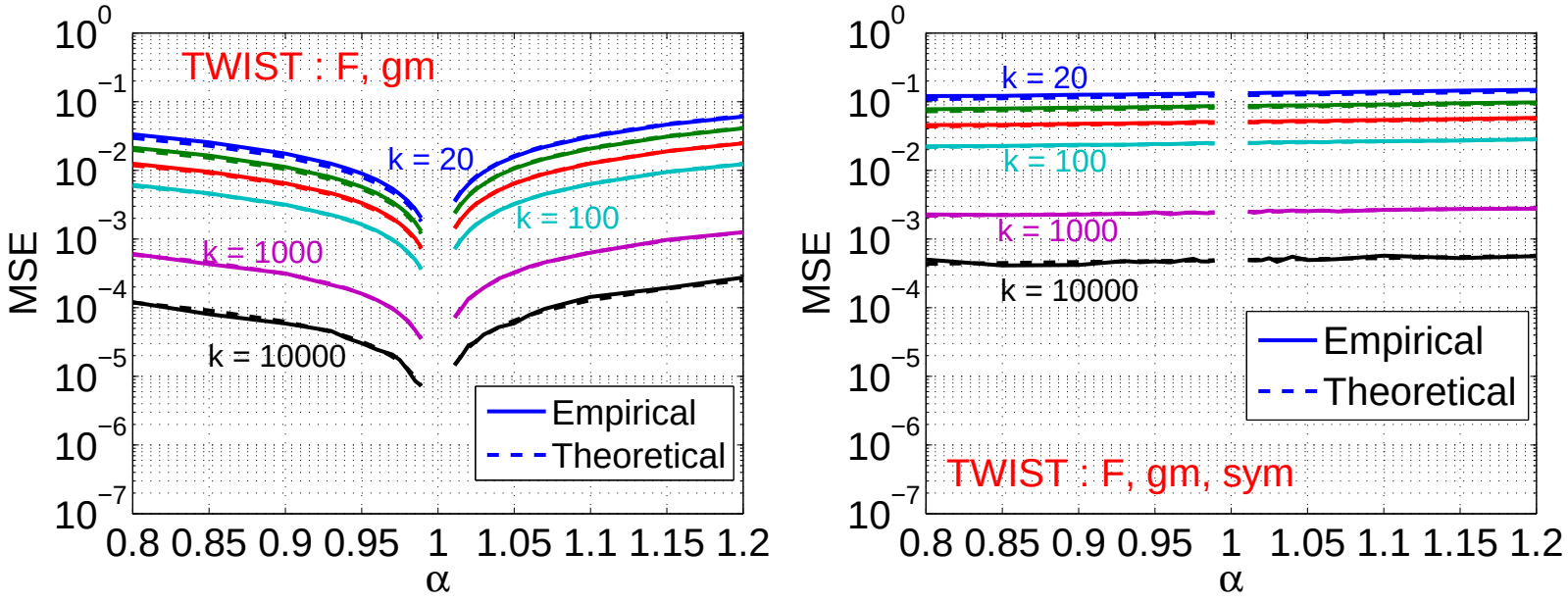


Estimating Shannon Entropy from Tsallis Entropy

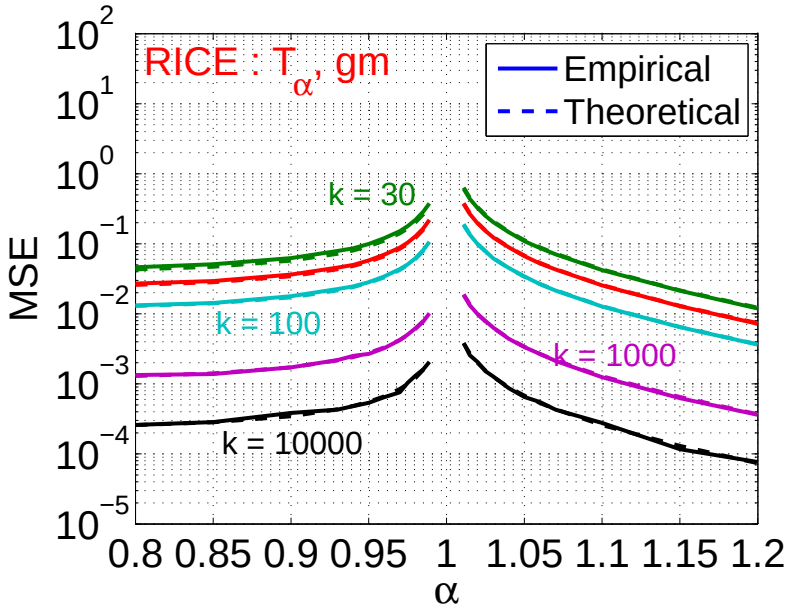


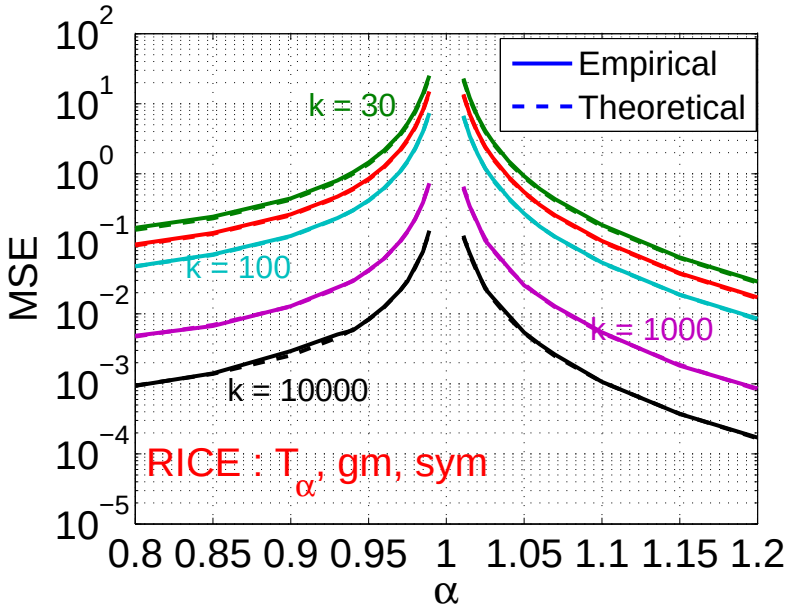


Estimating Frequency Moments

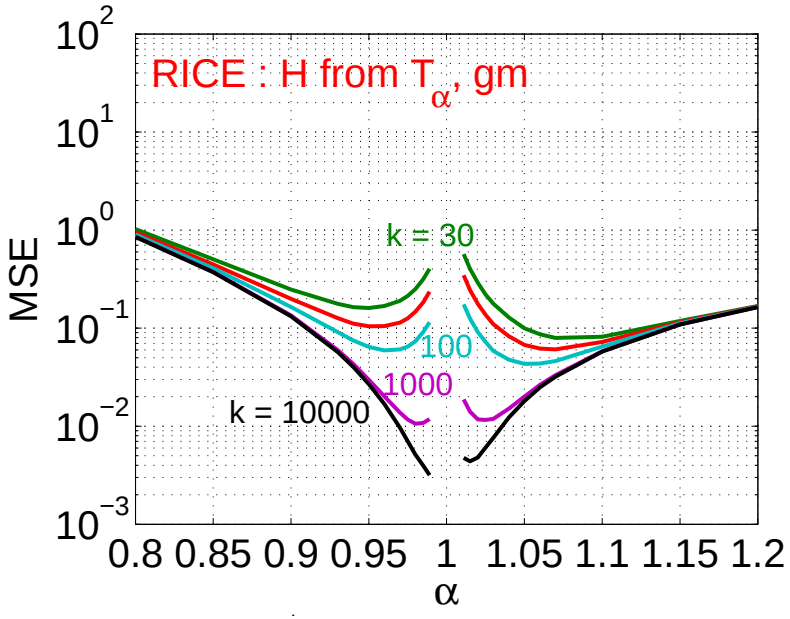


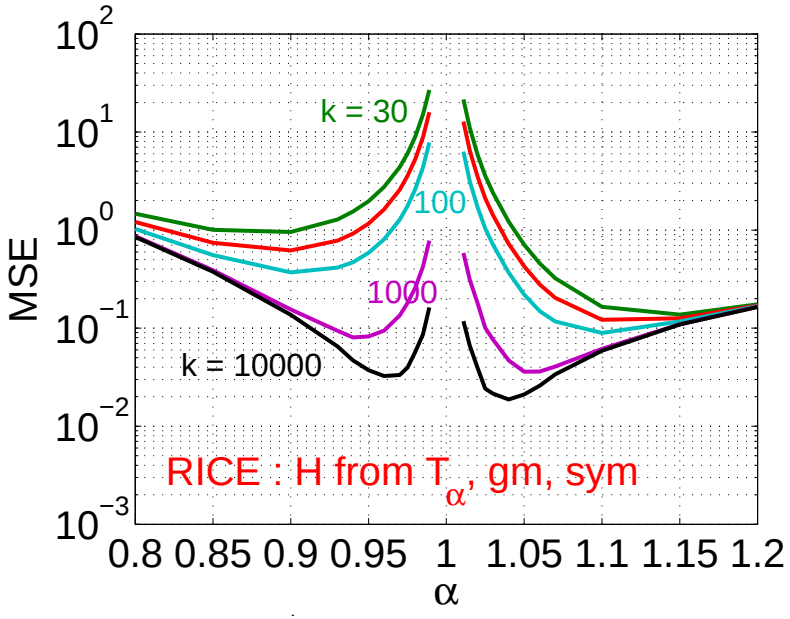
Estimating Tsalis Entropy





Estimating Shannon Entropy Using Tsallis Entropy





Applications in Method of Moments

For example, z_i, $i = 1$ to D are collected from data streams. z_i's follow a generalized gamma distribution $z_i \sim GG(\theta_1, \theta_2, \theta_3)$:

$$\mathrm{E}(z_i) = \theta_1\theta_2, \qquad \mathrm{Var}(z) = \theta_1\theta_2^2, \qquad \mathrm{E}\left(z - \mathrm{E}(z)\right)^3 = (\theta_3 + 1)\theta_1\theta_2^3$$

Estimate θ_1, θ_2, θ_3 using

- First three moments ($\alpha = 1, 2, 3$) $\Longrightarrow$ Computationally very expensive
- Fractional moments (eg. $\alpha = 0.95, 1.05, 1$) $\Longrightarrow$ Computationally cheap

Will this affect estimation accuracy? Not really, because D is large!

A Simple Example with One Parameter

Suppose $z_i \sim Gamma(\theta, 1)$. The data z_i's are collected from data streams.

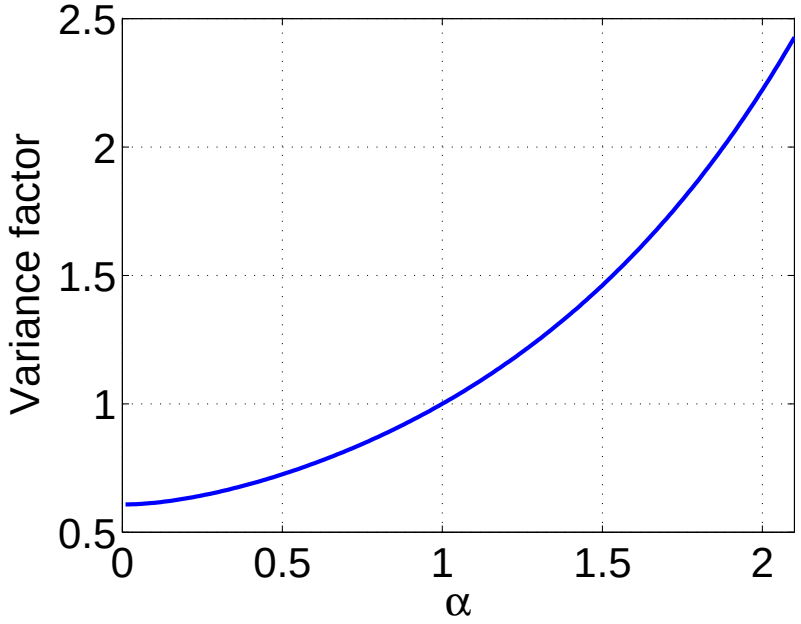
Estimate θ by αth moment: $\boxed{\mathrm{E}(z_i^\alpha) = \Gamma(\alpha+\theta)/\Gamma(\theta)}$.

Solve for $\hat{\theta}$ from the moment equation:

$$\frac{\Gamma(\alpha+\hat{\theta})}{\Gamma(\hat{\theta})} = \frac{1}{D}\sum_{i=1}^{D} z_i^\alpha$$

$$\mathrm{Var}\left(\hat{\theta}\right) \approx \frac{1}{D}\left(\frac{\Gamma(2\alpha+\theta)\Gamma(\theta)}{\Gamma^2(\alpha+\theta)} - 1\right) \frac{1}{\left(\frac{\Gamma'(\alpha+\theta)}{\Gamma(\alpha+\theta)} - \frac{\Gamma'(\theta)}{\Gamma(\theta)}\right)^2}$$

$$\text{Var}\left(\hat{\theta}\right)\big|_{\alpha=0} \approx \frac{0.608}{D}, \qquad \text{Var}\left(\hat{\theta}\right)\big|_{\alpha=1} \approx \frac{1}{D},$$



Trade-off:

$\alpha = 1$, higher variance, fewer counters

$\alpha = 0$, smaller variance, more counters

Since D is very large, the difference between $\frac{0.608}{D}$ and $\frac{1}{D}$ may not matter.

Summary

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

- Previous methods (eg. symmetric stable random projections) do not capture the intuition that estimating α-th moments should be easy if $\alpha \approx 1$.

- Compressed Counting (CC) improves symmetric stable random projections for all $0 < \alpha < 2$. The improvement is dramatic when $\alpha \to 1$.

- Using CC for estimating Shannon entropy is highly efficient.

Thank you!