Improved Light Tailed Sample Averages for Robust Estimation of the Mean

Mark Huber
Claremont McKenna College
July 4, 2018

Supported by NSF grant DMS 1418495

Goal

To create an (ϵ, δ)-randomized approximation scheme for the mean μ of Monte Carlo data where σ^2/μ^2 has a known bound c.

Inputs from user

Monte Carlo iid data $X_1, X_2, \ldots$ with mean μ and variance σ^2
$\epsilon > 0$ bound on relative error
$\delta > 0$ bound on probability of failure

Today

A new variant of an existing algorithm that uses fewer samples

Where does this problem arise?

Used for approximating $\#P$ complete problems in computer science that also arise in statistical inference

Examples

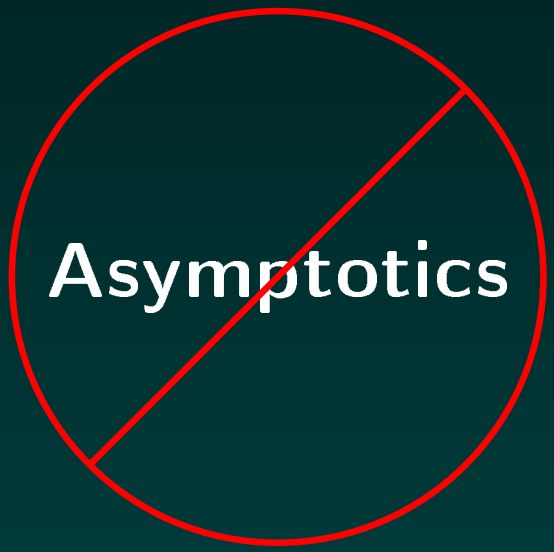
- Number of permutations with restricted positions [1]
- Volume of convex bodies [2]
- Normalizing constant for Gibbs distributions ($c = 2e$) [3]

---

[1] M.R. Jerrum, A. Sinclair, and E. Vigoda, A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries, *J. of the ACM*, 51(4):671–697, 2004

[2] L. Lovász and S. Vempala, Simulated annealing in convex bodies and an $O^*(n^4)$ volume algorithm, *J. Comput. Syst. Sci*, 72(2):392–417, 2006

[3] M. Huber, Approximation algorithms for the normalized constant of Gibbs distributions, *Ann. Appl. Probab*, arXiv:1206.2689, 51(1):92–105, 2015


Asymptotics


Order
notation

Current state of the art to low order

Rootfinding approach [4] (applies $X_i \in [0, \infty)$ or $X_i \in (-\infty, 0]$)

$$n = \lceil 2 \ln(2/\delta) c^2 \epsilon^{-2} [1 + 2\epsilon] \rceil$$

Light tailed sample averages [5] (applied $X_i \in \mathbb{R}$)

$$n = \lceil 2 \ln(4/\delta) c^2 \epsilon^{-2} [1 + 10.85\epsilon] \rceil$$

---

[4] Anonymous communication, 2018

[5] M. Huber, *An optimal (ϵ, δ)-approximation scheme for random variable with bounded relative variance*, arXiv:1706.01478?,2017

Today

Rootfinding approach (applies $X_i \in [0, \infty)$ or $X_i \in (-\infty, 0]$)

$$n = \lceil 2 \ln(1/\delta)[1 + \epsilon^{-2} c^2](1 - \epsilon^2)^{-1} \rceil$$

Light tailed sample averages (applied $X_i \in \mathbb{R}$)

$$n = \lceil 2 \ln(2/\delta)^2 c^2 \epsilon_\delta^{-2} [1 - 4.82\epsilon^2] \rceil$$

The tools for this

Smoothing

- By adding a random variable to data, can reduce the upper bounds obtains from Chernoff and Chebyshev by a factor of 2

Scaling

- For rootfinding, choosing scale parameter properly to equalize the bounds on the upper and lower tail

Where this problem arises

Monte Carlo approach

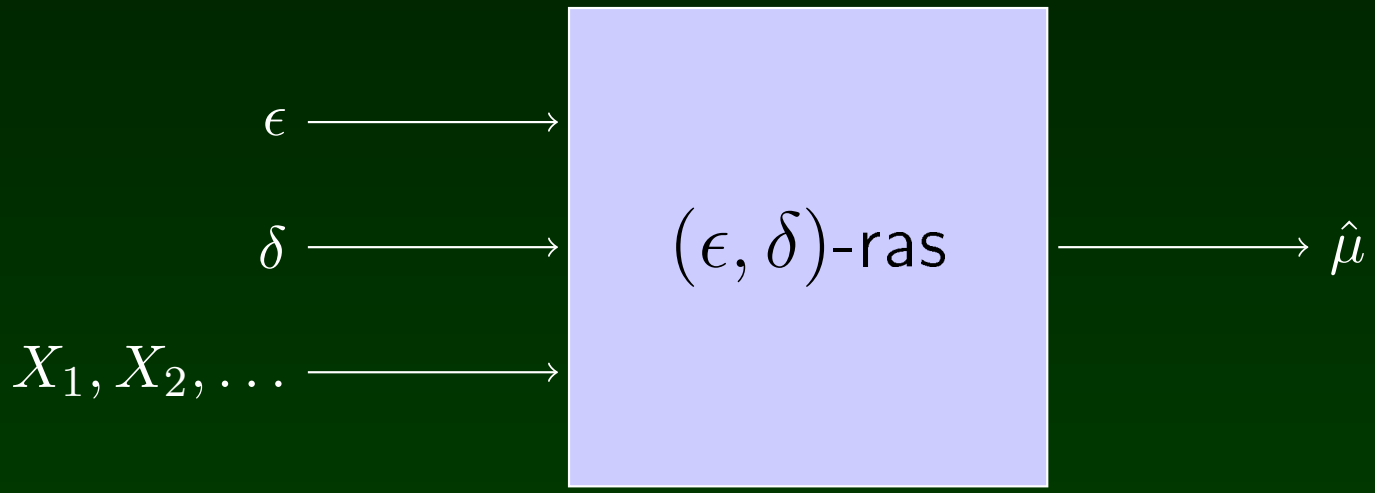
For many problems, it is possible to build a random variable X *such that* $\mathbb{E}[X] = a$*, the target result, and* $\mathbb{V}(X)/a^2 \leq c^2$ *where* c *is known.*

Monte Carlo approach

For many problems, it is possible to build a random variable X *such that* $\mathbb{E}[X] = a$*, the target result, and* $\mathbb{V}(X)/a^2 \leq c^2$ *where* c *is known.*

So the question is, how few samples are necessary to estimate a *within a target relative error with a bounded chance of failure that is given?*

Randomized Approximation Scheme



$$\mathbb{P}\left(\left|\frac{\hat{\mu}}{\mu} - 1\right| > \epsilon\right) < \delta$$

Broad generalization: CS vs. Stat point of view

Statistical error

Try to build estimate that draws as much information as possible out of data

Computer science error

Determine how much data needed to guarantee in worst case scenario that error is not too great

Our problem

Relative variance

The relative variance of a random variable X is

$$\frac{\mathbb{V}(X)}{\mathbb{E}[X]^2}$$

(Square of the coefficient of variation.)

Bounded relative variance

Suppose we know a constant c such that

$$\frac{\mathbb{V}(X)}{\mathbb{E}[X]^2} \leq c^2$$

for an (ϵ, δ)-ras

What if the random variables are normal?

Question

Suppose we knew

$$X_1, X_2, \ldots \sim \mathsf{N}(\mu, (c\mu)^2),$$

how many samples would we need then?

Answer

To first order, need

$$2c^2\epsilon^{-2}\ln(\delta^{-1})$$

for an (ϵ, δ)-ras

Simple unbiased estimate of μ

Sample average is easy to use

$$\hat{\mu}_n = \frac{X_1 + \cdots + X_n}{n}$$

Unfortunately does not give a (ϵ, δ)-ras

Problem with sample average

Fails for binary random variables

Even with bounded relative variance, binary random variables can have a large value that if seen, throws off the entire sample average

More details

- Normal random variables: error goes down exponentially in n
- Binary random variables over $\{a, b\}$, $a < b$,

$$\mathbb{P}(X = a) = 1 - \alpha, \quad \mathbb{P}(X = b) = \alpha,$$

For α polyonomially small in n, $\mathbb{E}[X]$ near μ and $\mathbb{V}(X)$ near $(c\mu)^2$, but

$$\mathbb{P}\left(\left|\frac{\hat{\mu}}{\mu} - 1\right| > \epsilon\right) \geq \mathbb{P}\left((\exists i)(X_i = b)\right) \approx n\alpha,$$

so only goes down polynomially in n, not exponentially

Binary random variable X with

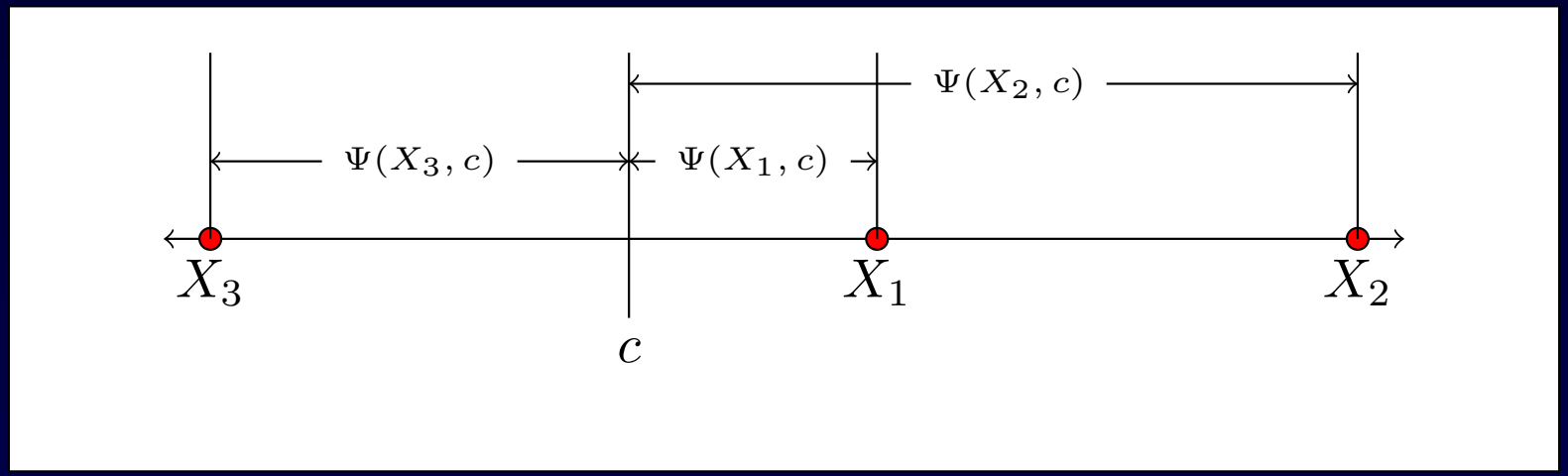
M-estimator approach

M-estimator

Ψ-type gives a measure of central tendency

- For c an estimate of the "center" of $\{X_i\}$, let $\Psi(X_i, c)$ be a signed distance of how far away X_i is from c
- Find $\hat{c}$ that balances data points above with those below, so

$$\hat{c} \text{ solves } f(c) = \sum_{i=1}^{n} \Psi(X_i, c) = 0$$



Two common examples

Sample average

$$\begin{aligned}\Psi_1(X_i, c) &= X_i - c \\ &= d_1(X_i - c),\ d_1(u) = u\end{aligned}$$

Sample median (n odd)

$$\Psi_2(X_i, c) = \begin{cases} +1 & X_i - c \geq 0 \\ -1 & X_i - c \leq 0 \end{cases}$$

$$= d_2(X_i - c),\ d_2(u) = \begin{cases} +1 & u \geq 0 \\ -1 & u \leq 0 \end{cases}$$

Catoni (2012)[6] M-estimator

Sample average

$$\Psi_1(X_i, x) = d_1(X_i - x) = \lambda^{-1} d_1(\lambda(X_i - x)),\ d_1(u) = u$$

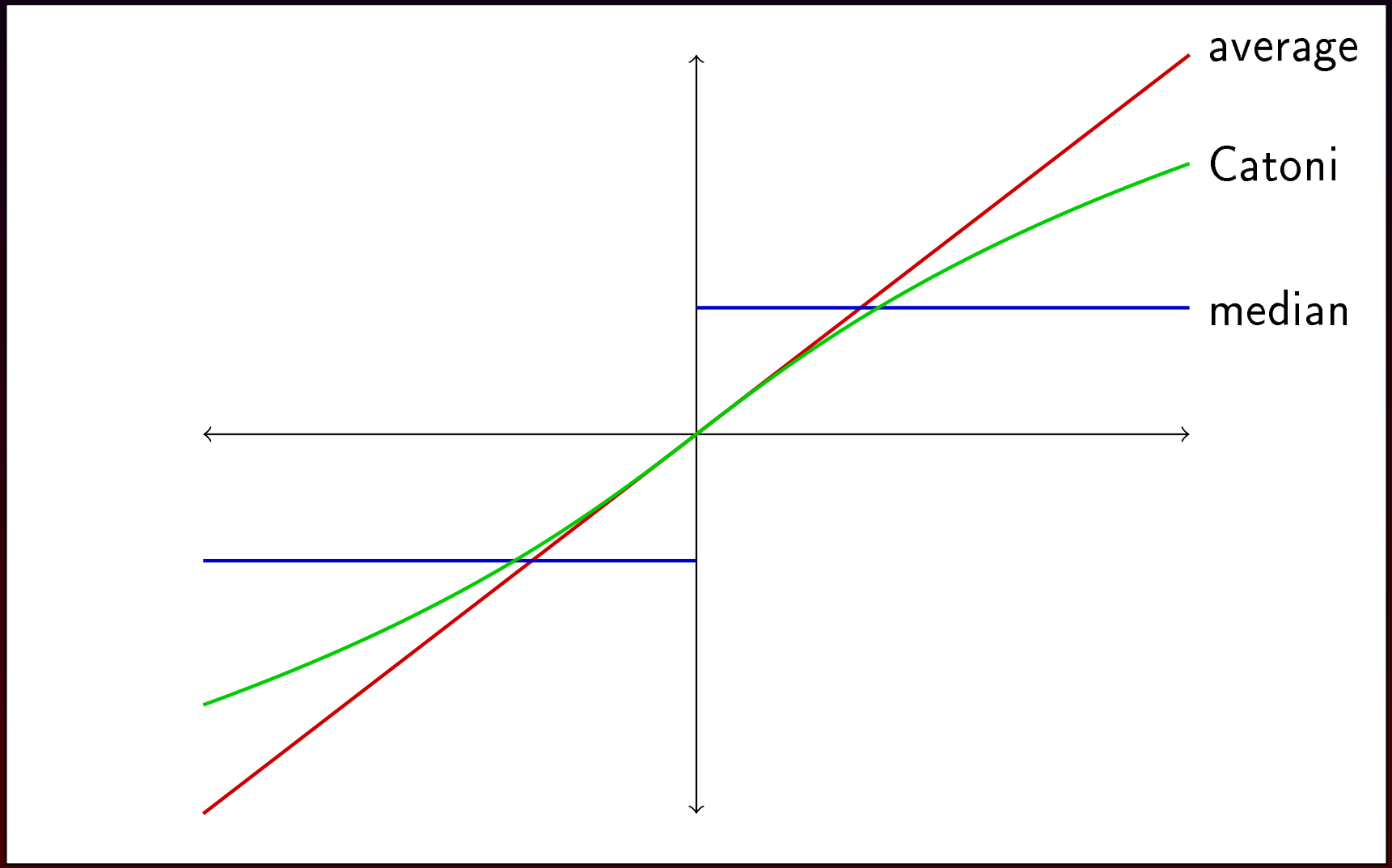
Want to discourage outliers

$$\Psi_3(X_i, x) = \lambda^{-1} d_3(\lambda(X_i - x))$$

$$d_3(u) = \begin{cases} \ln(1 + u + u^2/2) & u \geq 0 \\ -\ln(1 - u + u^2/2) & u \leq 0 \end{cases}$$

[6] O. Catoni, Challenging the empirical mean and empirical variance: A deviation study, *Ann. Inst. H. Poincaré Probab. Statist.*, 49(4):1148–1185, 2012

Picture



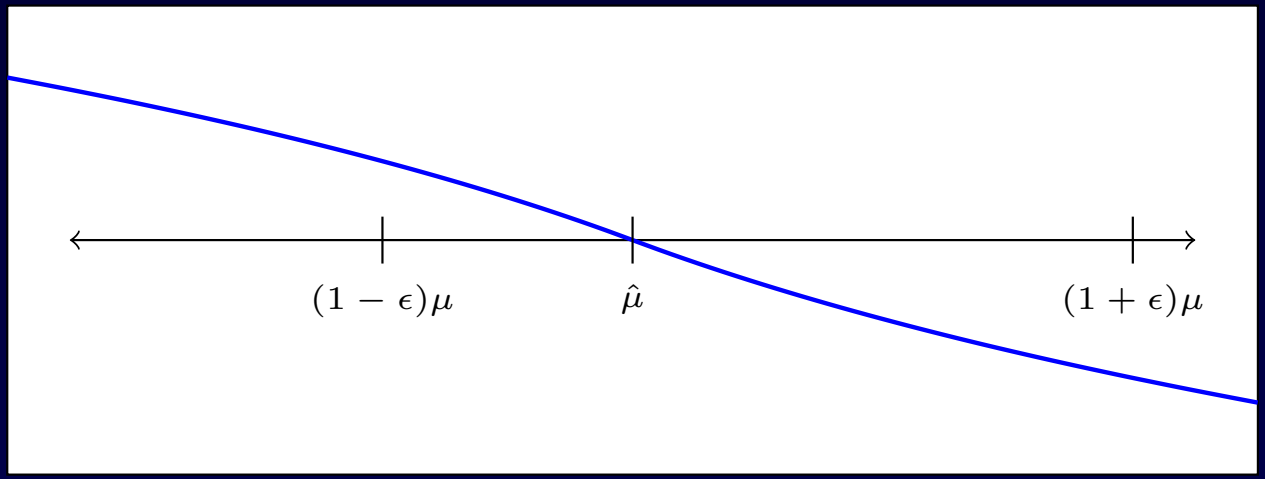
In practice

About the M-estimator

$$f_3(x) = \sum_{i=1}^{n} \lambda^{-1} d_3(\lambda(X_i - c))$$

is decreasing in c. So

$f_3((1-\epsilon)\mu) > 0$ and $f_3((1+\epsilon)\mu) < 0 \Rightarrow \hat{c} \in [(1-\epsilon)\mu, (1+\epsilon)\mu)]$



Getting relative error

Want to discourage relative outliers

$$\psi_4(X_i, x) = \lambda^{-1} d_3(\lambda(X_i/x - 1))$$

Now

$$f(\alpha\mu) = \lambda^{-1} d_3\left(\lambda\left(\alpha^{-1}\frac{X_i}{\mu} - 1\right)\right)$$

where

$$\mathbb{E}\left[\frac{X_i}{\mu}\right] = 1,\ \mathbb{V}\left[\frac{X_i}{\mu}\right] = \frac{\sigma^2}{\mu^2} \leq c^2$$

Let $Y_i = X_i/\mu$ for convenience

What we lose

Original M-estimator always decreasing in x

$$f_3(x) = \lambda^{-1} \sum_{i=1}^{n} d_3(\lambda(X_i - x))$$

New M-estimator decreasing only if X_i all same sign

$$f_4(x) = \lambda^{-1} \sum_{i=1}^{n} d_3(\lambda(X_i/x - 1))$$

Most applications always have $X_i \geq 0$

Improving Chernoff bounds

Why $\ln$ *in* d_3*?*

For all u

$$-\ln(1-u+u^2/2) \leq d_3(u) \leq \ln(1+u+u^2/2)$$

Designed to give control of moment generating function

$$\begin{aligned}
\mathrm{mgf}[d_3(\lambda(Y_i/\alpha-1))] &= \mathbb{E}\left[\exp\left(d_3\left(\lambda\left(\frac{Y_i}{\alpha}-1\right)\right)\right)\right] \\
&\leq \mathbb{E}\left[1+\lambda\left(\frac{Y_i}{\alpha}-1\right)+\frac{\lambda^2}{2}\left(\frac{Y_i}{\alpha}-1\right)^2\right] \\
&= 1+\left(\frac{\lambda}{\alpha}\right)(1-\alpha)+\left(\frac{\lambda}{\alpha}\right)^2[c^2+(1-\alpha)^2]
\end{aligned}$$

Why control mgf?

Fact (Chernoff's bound)

If $\mathbb{E}[\exp(X)]$ *exists,*

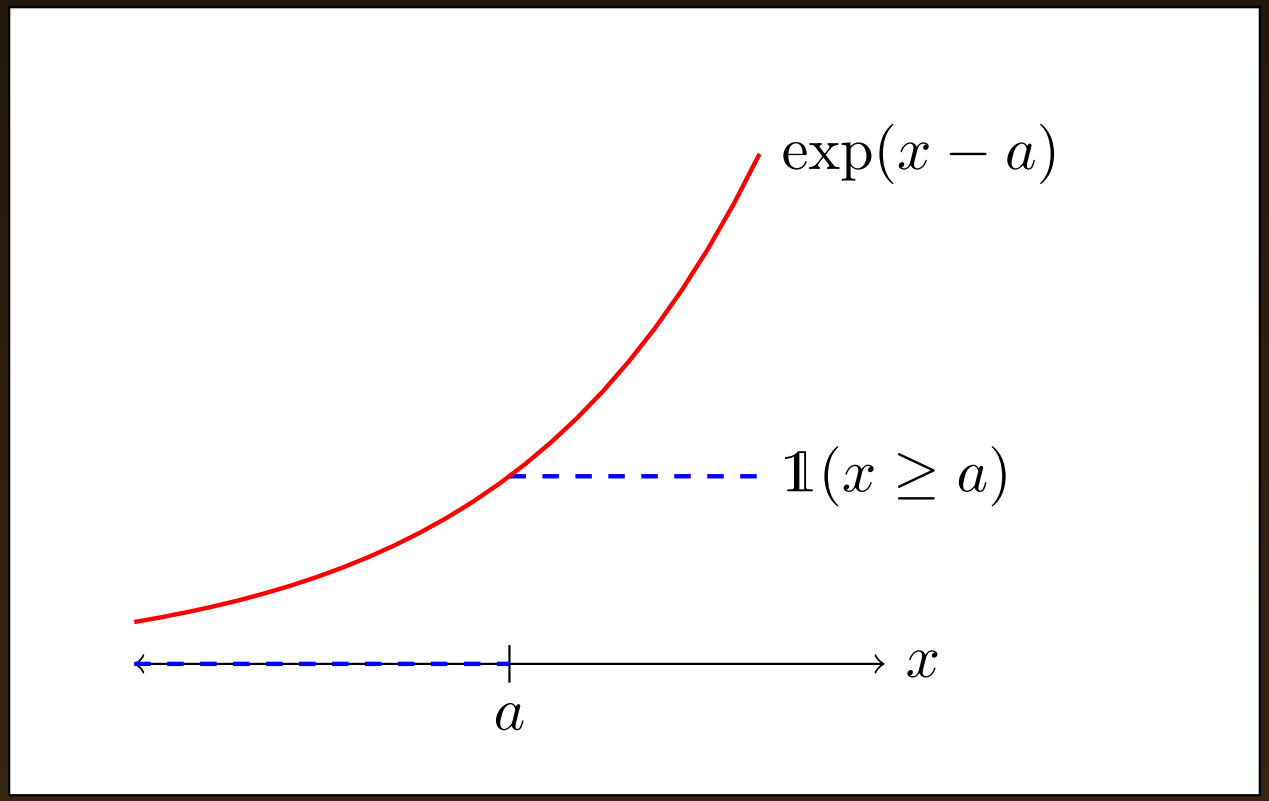
$$\mathbb{P}(X \geq a) \leq \mathbb{E}[\exp(X)] \exp(-a).$$

If $\mathbb{E}[\exp(-X)]$ *exists,*

$$\mathbb{P}(X \leq a) \leq \mathbb{E}[\exp(-X)] \exp(a).$$

Proof

$$\begin{aligned}\mathbb{P}(X \geq a) &= \mathbb{E}[\mathbb{1}(X \geq a)] \\ &\leq \mathbb{E}[\exp(X - a)]\end{aligned}$$



Reducing error by 1/2 using smoothing

Fact (Smoothed Chernoff bound)

Let X *with* $\mathbb{E}[\exp(X)]$ *finite and* $U \sim$ *Unif*$([-1,1])$ *be independent. Then*

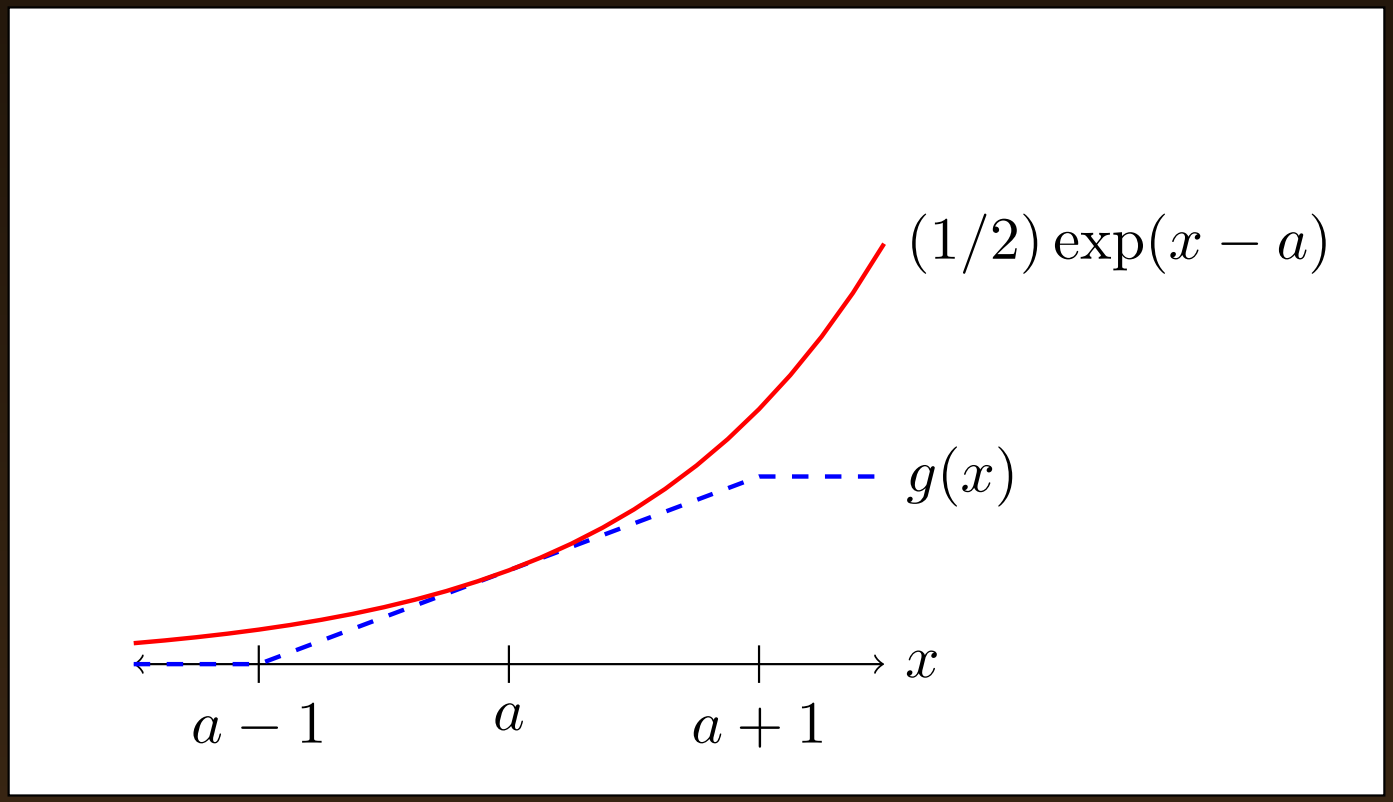
$$\mathbb{P}(X + U \geq a) \leq (1/2)\mathbb{E}[\exp(X)]\exp(-a).$$

Let X *with* $\mathbb{E}[\exp(X)]$ *finite and* $U \sim$ *Unif*$([-1,1])$ *be independent. Then*

$$\mathbb{P}(X + U \leq a) \leq (1/2)\mathbb{E}[\exp(-X)]\exp(a).$$

Proof

$$\begin{aligned}\mathbb{P}(X+U \geq a) &= \mathbb{E}[\mathbb{1}(X+U \geq a)] \\ &\leq \mathbb{E}[\mathbb{E}[\mathbb{1}(U \geq a-X)|X]] \\ &\leq \mathbb{E}[g(X)] \leq \mathbb{E}[(1/2)(\exp(X)-a)]\end{aligned}$$



Goal

Want to show

$$f_4((1-\epsilon)\mu) > 0, \;\; f_4((1+\epsilon)\mu) < 0$$

with probability at least

$$1-\delta$$

Know

$$\mathbb{P}(f_4((1+\epsilon)\mu) > 0) = \mathbb{P}(\lambda f_4((1+\epsilon)\mu) > 0)$$
$$\leq \left[1 - \frac{\lambda\epsilon}{1+\epsilon} + \left(\frac{\lambda}{1+\epsilon}\right)^2 [c^2 - (1-\epsilon^2)]\right]^n$$

Achieving our goal

From last slide

$$\mathbb{P}(f_4((1+\epsilon)\mu) > 0) = \mathbb{P}(\lambda f_4((1+\epsilon)\mu) > 0)$$
$$\leq \left[1 - \frac{\lambda\epsilon}{1+\epsilon} + \left(\frac{\lambda}{1+\epsilon}\right)^2 [c^2 - (1-\epsilon^2)]\right]^n$$

For lower tail

$$\mathbb{P}(f_4((1-\epsilon)\mu) < 0) = \mathbb{P}(\lambda f_4((1+\epsilon)\mu) > 0)$$
$$\leq \left[1 - \frac{\lambda\epsilon}{1-\epsilon} + \left(\frac{\lambda}{1-\epsilon}\right)^2 [c^2 - (1-\epsilon^2)]\right]^n$$

Proper scaling

Choose λ to make formulas inside brackets equal

$$\lambda = \frac{\epsilon(1-\epsilon^2)}{c^2+\epsilon^2}$$

Achieves scaling result

$$n = \lceil 2\ln(2/\delta)(c^2+\epsilon^2)\epsilon^{-2}(1-\epsilon^2)^{-1} \rceil$$

Smoothing piece

For $U \sim \text{Unif}([-1, 1])$ independent of the $\{X_i\}$

$$f_5(x) = \left[\sum_{i=1}^{n} \Psi_4(X_i, x)\right] + \frac{U}{\lambda}$$

Then only need

$$n = \lceil 2 \ln(1/\delta)(c^2 + \epsilon^2)\epsilon^{-2}(1 - \epsilon^2)^{-1} \rceil$$

The complete algorithm that uses rootfinding

1. Let $n \leftarrow \lceil 2\ln(1/\delta)c^2\epsilon^{-2}(1-\epsilon^{-2})^{-1} \rceil$,
   $\lambda \leftarrow \epsilon(1-\epsilon^2)(\epsilon^2+c^2)^{-1}$
2. Draw $X_1, X_2, \ldots, X_n \leftarrow X$
3. Draw $U \leftarrow \text{Unif}([-1,1])$
4. For the function

$$f_5(x) = \left[\sum_{i=1}^{n} \lambda^{-1} d_3(\lambda(X_i/x - 1)) \right] + \lambda^{-1}U,$$

   let $\hat{\mu}$ be the unique solution to $f_5(x) = 0$.

Light Tailed Sample Averages

Advantage

- Avoids need for rootfinding
- Works for $X_i \in \mathbb{R}$

Disadvantage

- Requires more samples

The complete LTSA algorithm

1. Use a median-of-means estimator to obtain $\hat{\mu}_0$ that is an $(\sqrt{\epsilon}), \delta/2)$-ras
2. Set $n \leftarrow \lceil 2\epsilon^{-2}(1-\epsilon)(\epsilon + c^2)\ln(\delta^{-1})\rceil$ Draw $X_1, \ldots, X_n$ iid
3. Set $\alpha \leftarrow \epsilon\hat{\mu}_0/(\epsilon + c^2)$. For each $i \in \{1, \ldots, n\}$,

$$W_i \leftarrow \hat{\mu}_0 + \alpha^{-1} d_3(\alpha(X_i - \hat{\mu}_0))$$

4. Draw $U \leftarrow \mathsf{Unif}([-1, 1])$. Final estimate

$$\hat{\mu}_{\mathsf{LTSA}} \leftarrow \frac{W_1 + \cdots + W_n + U/\alpha}{n}$$

Improvements

To median-of-means part

- Smooth to decrease chance of error
- Can get near $1/2$ factor for Chebyshev in fashion similar to Chernoff

To LTSA part

- Smooth as with rootfinding algorithm

Summary

- Want concentration around mean even for possibly heavy-tailed random draws
- By weighting appropriately, can use number of samples similar to that for normal data, but that work even when data is actually heavy-tailed
- Smoothing and scaling properly can assist in decreasing bounds on failure