Mean Estimation in Low and High Dimensions

August 5, 2022

Paul Valiant

joint work with Jasper C.H. Lee

The Mean Estimation Problem

Given data, how to estimate mean of underlying distribution?

$\boxed{\text{Sample mean } \frac{1}{n}\sum_{i=1}^{n} x_i}$ ← Great for Gaussians, nice distributions

"I saw $\frac{2}{10}$ outliers" $\sim$ E[outliers] = $\frac{1}{10}$, or $\frac{4}{10}$

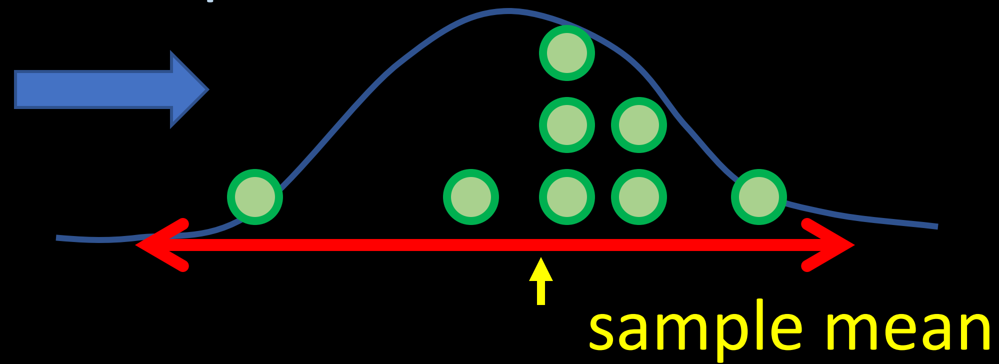
$\sim_{10^{-6}}$ E[outliers] =0.0001

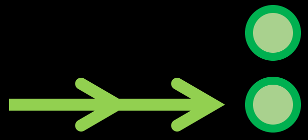
A) Most extreme distributions that "could have" led to data?

B) Find estimate that is accurate enough for all such distributions

Optimal alg must depend on desired confidence

iid samples





The Goal

"Given n samples from an unknown distribution p, estimate the mean of p to ±ε, with probability 1-δ"

What is the optimal tradeoff between n, ε, and δ?

#samples accuracy confidence

(computation time isn't the concern)

Today: Low and High Dimensions

Thm 1: when $d = 1$

Thm 2: when $d_{eff} = \omega(\log^2 \frac{1}{\delta})$

Thm 2b: new "vector Bernstein inequality"

My Perspective

Algorithms/Efficiency	Lower Bounds /Complexity
Time Efficiency	Time Complexity
Space Efficiency	Space Complexity
Data Efficiency (Sample Efficiency)	Data Complexity (Sample Complexity)

Low Dimensions: The Goal

"Given n samples from an unknown distribution p, estimate the mean of p to ±ε, with probability 1-δ"

What is the optimal tradeoff between n, ε, and δ?

#samples accuracy confidence

(computation time isn't the concern)

Benchmark against the ideal case: sample mean, on a Gaussian distribution

unknown

Given n samples from distribution of variance σ^2…?

$\boxed{\text{our algorithm}}$ returns, with probability $\geq 1-\delta$, answer

with accuracy $\pm\sigma\sqrt{\frac{(2+o(1))\log\frac{1}{\delta}}{n}}$

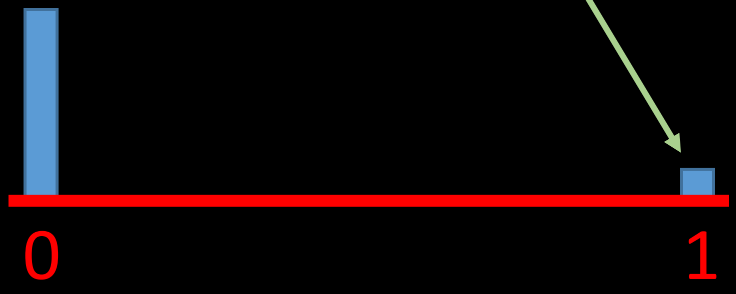
???

Algorithm 1: the Sample Mean

Works great for Gaussians, but…

n samples from distribution: $\frac{1}{1000\,n}$ probability of drawing 1

otherwise 0



mean = $\frac{1}{1000\,n}$

99.9% of time: n samples → all 0; sample mean 0; small error

0.1% of time: we see a 1; sample mean = $\frac{1}{n}$; error 999x as big!

Sample mean is unbiased, but not "robust"

Algorithm 2: Median of Means

Nemirovsky, Yudin (1983), Jerrum, Valiant, and Vazirani (1986),
Alon, Matias, and Szegedy (2002).

1. Blindly split data into $8\log\frac{1}{\delta}$ groups
2. Compute mean of each group
3. Return median of the means

Intuition: Median is robust; sample mean is unbiased; combine to get "best of both worlds" – robustness and accuracy.

Error $\pm\sigma\sqrt{\frac{20\log\frac{1}{\delta}}{n}}$

Algorithm 3: Catoni (2012)

Warmup:

Given data $x_1, \ldots, x_n$,

Its mean is the point u minimizing $\sum_i (x_i - u)^2$

or, solving $\sum_i x_i - u = 0$

Its median is the point u minimizing $\sum_i |x_i - u|$

or, solving $\sum_i sign(x_i - u) = 0$

Idea: pick a function ψ that is ≈linear near 0, and $\approx sign(x)$ away from 0

Algorithm: solve for u such that $\sum_i \psi(x_i - u) = 0$

Let $\psi(y) = f\left(\boxed{\frac{1}{\sigma}}\sqrt{\frac{2\log\frac{1}{\delta}}{n}}\,y\right)$; let $f(z) = \log\frac{1}{1-z+z^2/2}$ for $0 \le z \le 1$, and $f(z) = \log 2$ for $z \ge 1$, with odd symmetry about z=0.

Thm: if you know the variance σ; or, if p has bounded 4[th] moment

Thm:

Error $\pm\sigma\sqrt{\frac{(2+o(1))\log\frac{1}{\delta}}{n}}$

The Challenge:

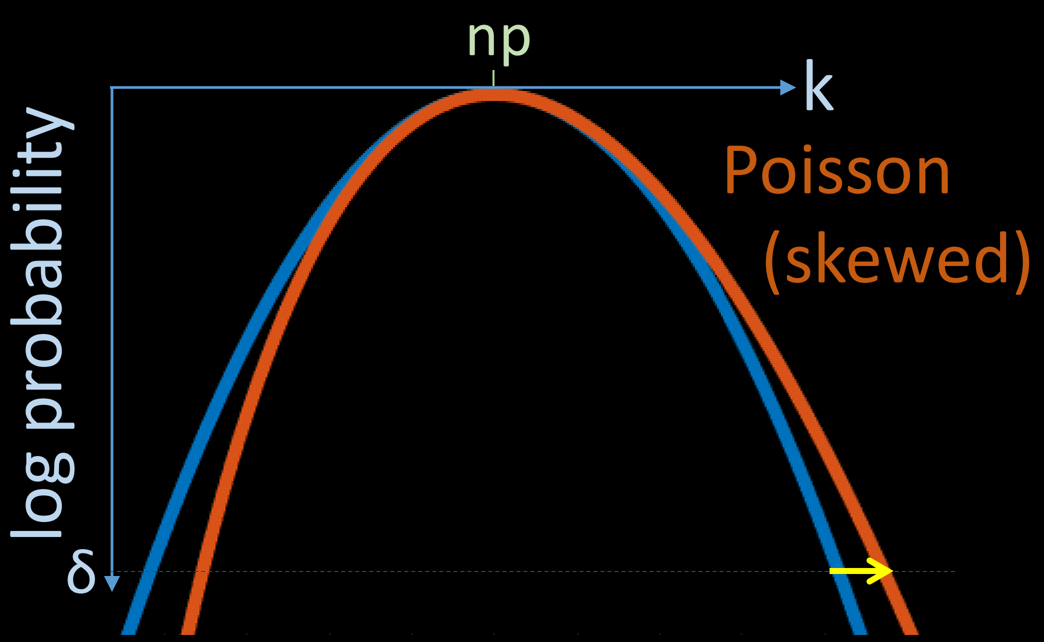
Catoni’s mean estimator needs to know the “width” of the distribution. Can we succeed without this?

The Bernoulli Case

(Poisson Case)

Suppose we get n draws from a coin of bias p   Bin(n,p) ≈ Poi(np)

Given as input k 1's, and n-k 0's, and parameter δ, what do we do?



Gaussian
(quadratic graph)

Gap of $\frac{1}{3}\log\frac{1}{\delta}$ at δ probability

Algorithm (sketch):

Recenter at a rough estimate (e.g. median of means)

in a weighted manner

Throw out $\frac{1}{3}\log\frac{1}{\delta}$ most extreme samples

with $weight(x) = \min(\alpha x^2, 1)$

Return mean of what remains

Punchlines: we can do mean estimation on *any* distribution as well as on a Gaussian of matching variance. We thought Gaussians were the best distribution; but they're actually the *worst.*

Techniques: duality; implicit ψ-estimator representation → i.i.d. sum

Next Steps:

- What can we do relative to α moments for $1 < \alpha < 2$ (instead of variance)?
- Maybe we shouldn't use Gaussians as a benchmark → *instance-optimal algs*
- Many new models: "robust" statistics - algs robust to outliers and weird distributions; however proof techniques often extend to "robust to adversarial data contamination", allowing for positive results outside the garden of "i.i.d. data"

Higher dimensions…

High Dimensions:

Two Problems

Mean estimation: Given samples from a high-dimensional distribution, estimate its mean, optimally

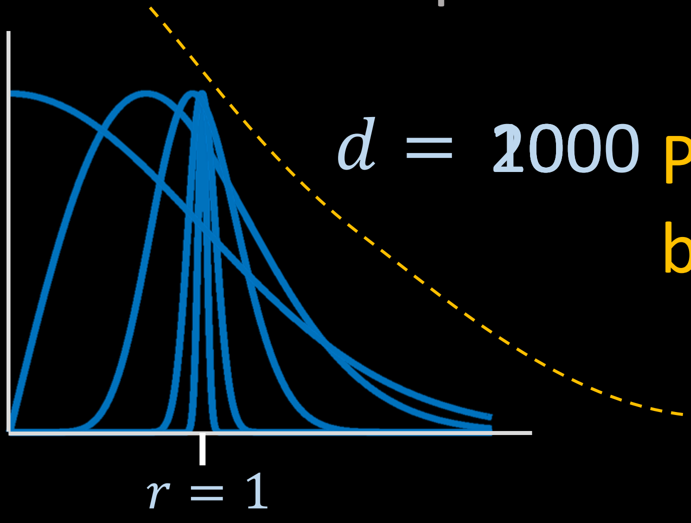
---

Tail bounds: For a general bounded distribution – "even when it does not look like a spherical shell, the sum of many samples does"

(seeking "vector Bernstein inequality")

– Guiding idea: a high-d Gaussian "looks like a spherical shell"

Distribution of $\frac{1}{\sqrt{d}}\|\mathcal{N}(\mathbf{0}, I_d)\|$:



Previous tail bounds:
Don't improve with d
[Matrix Chernoff Bounds]

The Goal

"Given n samples from an unknown distribution p, estimate the mean of p to L2 dist ε, with probability 1-δ"

Optimal tradeoff between n, ε, and δ?

#samples accuracy confidence

High dimensional case:

$d_{eff} = \omega(\log^2 \frac{1}{\delta})$

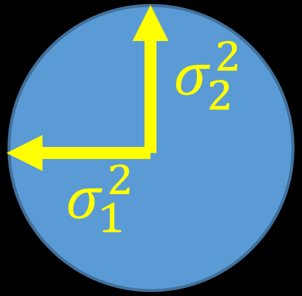
$\Sigma_{i,j} = E_{x \leftarrow p}[(x_i - \mu_i)(x_j - \mu_j)]$

$$d_{eff} = \frac{\mathrm{Tr}(\Sigma)}{\sigma_{max}^2}$$

d×d covariance matrix

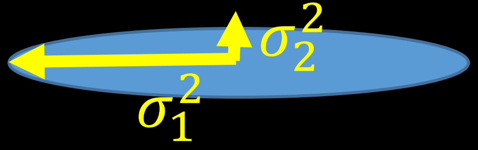
max directional variance

"effective dimension"



For spheres:

$d_{eff} = d$



For pancakes:

$d_{eff} < d$

Main Goal / The Result

"Given n samples from an unknown distribution p, estimate the mean of p to L2 dist ε, with probability 1-δ"

Optimal tradeoff between n, ε, and δ?

#samples accuracy confidence

Benchmark against the ideal case: sample mean, on a Gaussian distribution

unknown $d_{eff} = \omega(\log^2 \frac{1}{\delta})$

Given n samples from distribution of covariance Σ

our algorithm …? returns, with probability $\geq 1 - \delta$, answer

with error $(1 + o(1)) \sqrt{\frac{Tr(\Sigma)}{n}}$??? for $d_{eff} = \omega(\log \frac{1}{\delta})$

Prior Work on Constant-Factor Optimal Mean Est.

1d: median-of-means; Catoni (2012); Devroye et al. (2016); Lee-V (2022)

High-d: many generalizations of "median", tricky

	Time:	Sample-optimality:
Lugosi-Mendelson (2019)	Exp	$\Theta(1)$
Hopkins (2020)	Poly (SDP)	$\Theta(1)$
Cherapanamjeri, Flammarion, Bartlett (2020) Lei, Luh, Venkat, Zhang (2020)	$\tilde{O}(n^2 d)$	480000^2

Robust statistics approach: [Diakonikolas, Kane, Pensia'20]

Problem is scary from CS perspective (computational complexity) AND statistics perspective (sample complexity)

Today: linear-time; 1+o(1) optimal; extremely simple; but only in very high-d

Motivation

The good performance of the sample mean for Gaussian distributions comes from the fact that, in high dimensions, “Gaussians adhere to a spherical shell”

Natural to ask: can we extend “spherical shell tail bounds” beyond Gaussians?

The Bernstein Bound

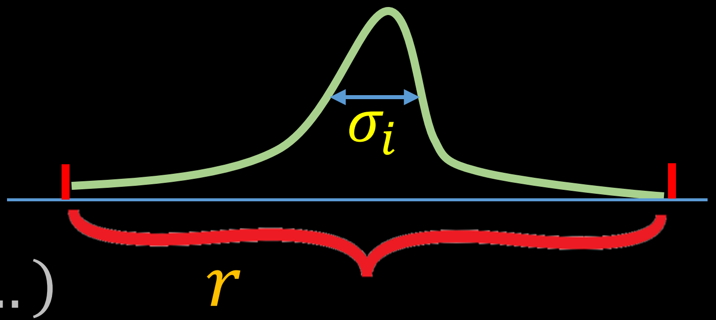
1d: Let $X_1, \dots, X_n$ be independent mean 0 random variables in $\mathbb{R}$, each bounded as $|X_i| \le r$. Then for any $t \ge 0$, $\Pr\left[\sum_i X_i \ge t\right] \le \exp\left(-\frac{\frac{1}{2}t^2}{\sigma^2 + \frac{1}{3}rt}\right)$

Gaussian term

Interaction shows how Gaussian bounds gracefully degrade in presence of outliers at radius r

New: Let $X_1, \dots, X_n$ be independent mean 0 random **vectors**, each bounded as $\|X_i\| \le r$. Then for any $t \in (0, \sqrt{Tr(\Sigma)}]$,

$$\Pr\left[\left\|\sum_i X_i\right\| \ge t + \sqrt{Tr(\Sigma)}\right] \le \exp\left(-\frac{\frac{1}{2}t^2}{\sigma_{max}^2 + \frac{1}{2}r\sqrt{Tr(\Sigma)}}\right) \cdot poly(\dots)$$



We want – for general distributions – to tightly match the ideal Gaussian performance; thus we seek a general tail bound that tightly matches the Gaussian's "spherical shell" behavior

Algorithm

Goal: come up with algorithm with 1+o(1) sample-optimal mean estimation, for all (high-dimensional) distributions

Tool/ Thm:

$$d_{eff} = \frac{Tr(\Sigma)}{\sigma^2_{max}}$$

Let $X_1, \ldots, X_n$ be independent mean 0 random vectors, each bounded as $\|X_i\| \leq r$. Then for any $\gamma \in (0,1]$, $\Pr\left[\left\|\sum_i X_i\right\| \geq (1+\gamma)\sqrt{Tr(\Sigma)}\right] \leq \exp\left(-\frac{\frac{1}{2}\gamma^2}{\frac{1}{d_{eff}}+\frac{1}{2}\frac{r}{\sqrt{Tr(\Sigma)}}}\right) \cdot poly\left(d_{eff}, \frac{\sqrt{Tr(\Sigma)}}{r}\right)$

Alg throws out $\omega(\log^2 \frac{1}{\delta})$ samples; in our optimal 1d algorithm [FOCS2021] we threw out $\frac{1}{3}\log\frac{1}{\delta}$ samples; this gives a sense of why our approach here can't extend transparently to low-d

s can be (almost) any upper bound on $\log^2 \frac{1}{\delta}$. "Multiple δ estimator"; impossible in 1d

1. Roughly estimate the mean with classical coordinate-wise median-of-means alg.
2. Throw out the $s = \omega(\log^2 \frac{1}{\delta})$ farthest samples. Return mean of what remains.

Linear time!

Given tail bound, our algorithm is simple, works for simple, robust reasons

…This new style of bound might be broadly useful

Contributions

Simple, linear-time, 1+o(1)-optimal mean estimation in “very high-d”

Vector Bernstein inequality that reproduces “spherical shell” tails

Next Steps:

- Bridging the gap between low and high dimensions
- Extending 1+o(1)-tight analysis to more regimes
- Instance-optimal algorithms
- New models, extending “robust” estimation

THANKS!

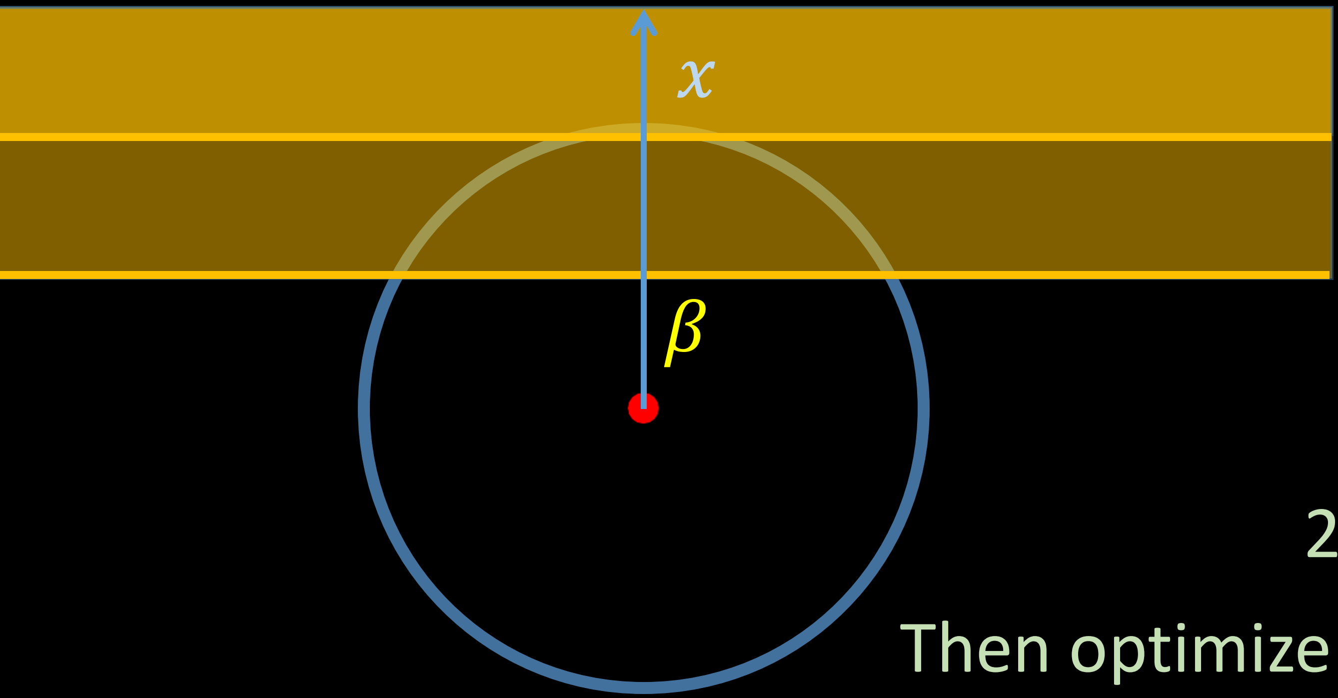
Vector Bernstein Proof Techniques

Thm: Let $X_1, \dots, X_n$ be independent mean 0 random vectors, each bounded as $\|X_i\| \leq r$. Then for any $\gamma \in (0,1]$, $\Pr\left[\left\|\sum_i X_i\right\| \geq\right.$

Proof ideas:

1. Average 1d Chernoff bounds in every direction x, at distance β

   "if every point outside sphere is hit by $\geq c$ fraction of Chernoff bounds: tail pr $\leq \frac{1}{c}$ (avg Chernoff bound)"



Issue: distributions can be "spiky", e.g. supported on axes; if x aligned with "spike", bounds blow up

2. Set aside support points y: $y \cdot x \geq p$

Then optimize over β, p to yield best bound

Same intuition for why Chernoff bounds are so tight: exponential nature of MGF means typically very narrow regime of distribution contributes most of bound; enough to pick β, p to perform well on narrow

Vector Bernstein: Tight?

Thm: Let $X_1, \dots, X_n$ be independent mean 0 random vectors, each bounded as $\|X_i\| \le r$. Then for any $\gamma \in (0,1]$, $\Pr\left[\left\|\sum_i X_i\right\| \geq\right.$

$$\exp\left(-\Omega\left(\gamma^2 \min\left(d_{eff}, \frac{\sqrt{Tr(\Sigma)}}{r}\right)\right)\right)$$

Tail lower bounds:

1. Gaussian: $\exp\left(-O\left(\gamma^2 d_{eff}\right)\right)$

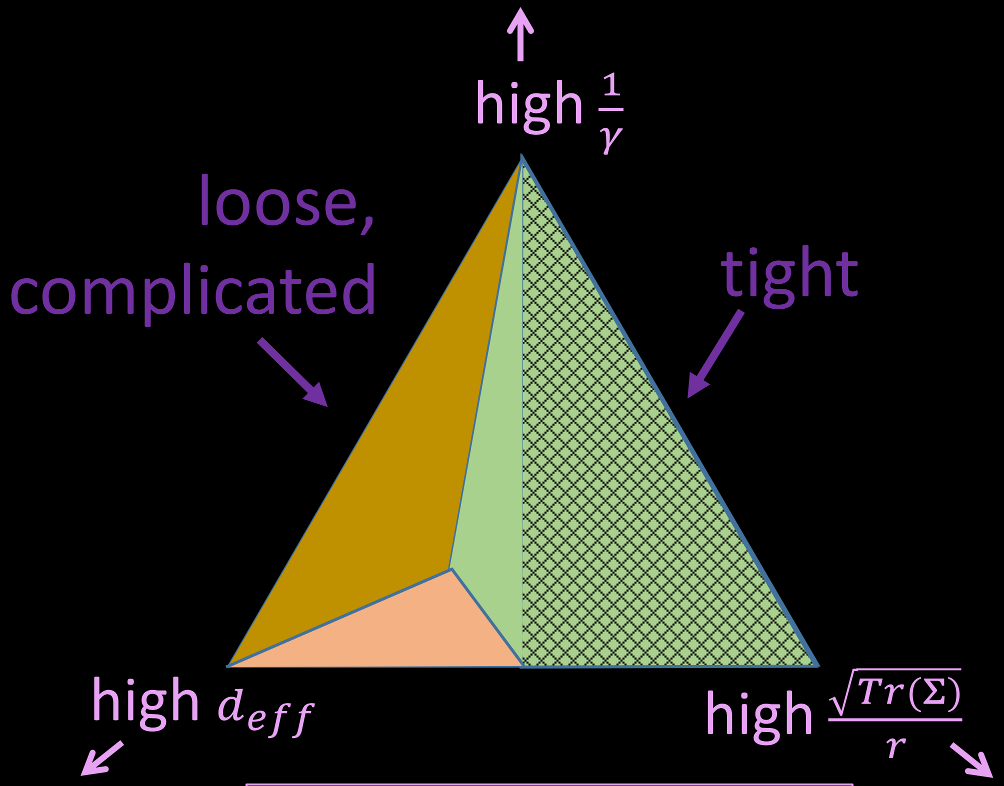
2. Axis-aligned Poisson at radius r:

   a) Tail likely along axis:

   $$\exp\left(-\tilde{O}\left(\sqrt{\gamma}\frac{\sqrt{Tr(\Sigma)}}{r}\right)\right)$$

   b) Tail likely in intermediate direction:

   $$\exp\left(-\tilde{O}\left(\gamma^2 \frac{Tr(\Sigma)}{r^2}\right)\right)$$



Parameter space – in homogeneous coords

Analysis

Algorithm:

1. Assume mean 0, variance 1, κ=0
2. Let $\sum_i \min\left(\alpha x_i^2, 1\right) - \frac{1}{3}\log\frac{1}{\delta} \equiv \psi_\alpha(x_i,\alpha,u) = 0$
3. Let $u - \frac{1}{n}\sum_i x_i\left(1-\min\left(\alpha x_i^2,1\right)\right) \equiv \psi_u(x_i,\alpha,u) = 0$

We have a 2-parameter "psi estimator".

Goal: show that, with probability 1-δ over sampling process, for all pairs (α, u) with $|u| > \sqrt{\frac{(2+o(1))\log\frac{1}{\delta}}{n}}$, the pair (α, u) will not satisfy $\vec{\psi}(x_i,\alpha,u) = 0$

Idea: show stronger statement, $\exists \vec{d}(\alpha,u)$ s.t. … w.p 1-δ, $\vec{\psi}(x_i,\alpha,u)\cdot\vec{d}(\alpha,u) > 0$

Standard technique: 1) pick a finite mesh of M points;

2) show $\vec{\psi}(x_i,\alpha,u)\cdot\vec{d}(\alpha,u)$ is Lipschitz between mesh points and monotonic beyond; 3) show that for each mesh point, $\Pr\left[\vec{\psi}(x_i,\alpha,u)\cdot\vec{d}(\alpha,u) \le 0\right] \le \frac{\delta}{M}$

→structural properties let us essentially move the "for all pairs" outside the probability