Recent Advances in Algorithmic Heavy-Tailed Statistics

Sam Hopkins, UC Berkeley (→ MIT)

Based on joint work with Yeshwanth Cherapanamjeri, Tarun Kathuria, Prasad Raghavendra, and Nilesh Tripuraneni.

Measuring success probability in estimation (confidence intervals)

Given $X_1 \ldots X_n \sim P_\theta$, find $\hat{\theta}$ s.t. $\|\theta - \hat{\theta}\| \leq \underbrace{r(n, d, \delta)}_{\text{"rate"}}$

with prob. $\geq 1-\delta$

no. samples

ambient dimension

failure rate

For $P_\theta \in \mathcal{P}$

Class of distributions

Our game: large class $\mathcal{P}$ (e.g. all distributions with $O(1)$ bdd. moments), but similar guarantees as if $\mathcal{P}$ contains Gaussians.

Example: Estimating the mean in ℓ_2

$X_1, \dots, X_n \sim X$, $\mathrm{Cov}(X) \preceq I$

Gaussian: $\left\| \frac{1}{n}\sum X_i - \mathbb{E}X \right\| \le \overbrace{\sqrt{\frac{d}{n}}}^{\mathbb{E}\left\|\frac{1}{n}\sum X_i - \mathbb{E}X\right\|} + \overbrace{\sqrt{\frac{\log 1/\delta}{n}}}^{e^{-t^2 n}}$ w.p. $1-\delta$

Drop the Gaussian assumption?

Is there an estimator $\hat{\mu}_\delta(X_1, \dots, X_n)$ s.t.

$$\|\hat{\mu}_\delta - \mu\| \le O\left(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log 1/\delta}{n}}\right) \text{ w.p. } 1-\delta \text{ ?}$$

Example: Estimating the mean in ℓ_2

$d=1$: Median of means, truncated mean, ...

Lugosi-Mendelson '18: estimator based on high-dimensional quantiles/median of means (exp. time) $\mathrm{poly}(n,d,\frac{1}{\delta})$

H. '18: algorithmic approach to high-dimensional quantiles

Theorem: for every $\delta > 2^{-n}$ exists $\hat{\mu}_\delta(X_1 \dots X_n)$ s.t.
$\|\hat{\mu}_\delta - \mu\| \leq O\left(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log^1/\delta}{n}}\right)$ w.p. $1-\delta$, $\mathrm{poly}(n, d, \log^1/\delta)$ time

Agenda

1. Survey of recent (algorithmic) developments
2. Heavy-tailed covariance estimation
   - Algorithm of [Cherapanamjeri - H. - Kathuria - Raghavendra - Tripuraneni]
   - 2 Key techniques for concentration + convex prog.
     - bounded differences
     - SoS Bernstein

$$\|\sum M_i\|_{op} = \max_{\substack{Y \succeq 0 \\ \operatorname{tr} Y}} \langle Y, \sum M_i \rangle$$

$$\|\sum M_i\|_{SoS} = \max_{Y \in SoS} \langle Y, \sum M_i \rangle$$

Matching Gaussian Confidence w/ only $O(1)$ moments

- mean estimation, ℓ_2 – median of means, trimmed mean
- mean estimation, $\|\cdot\|$
- covariance estimation
- regression
- regularized regression/sparse recovery
- ...

[Lugosi-Mendelson '19, '19, '19, '20, '20, Mendelson-Zhivotovskiy '20, Minsker '15, '18, '20, Hsu-Sabato '16, Lerasale-Oliveira '12, Joly-Lugosi-Oliveira '17, ...]

Matching Gaussian Confidence w/ only O(1) moments

in polynomial time

Mean estimation, ℓ_2

authors	time
H.	$O(nd)^{28}$
Cherapanamjeri-Flammarion-Bartlett	$\tilde{O}(n^{3.5} + n^2 d)$
Depersin-Lecué, Lei-Luh-Venkat-Zhang	$\tilde{O}(n^2 d)$

Covariance, $\|\cdot\|_{op}$

authors	Suboptimal by
Minsker	$\sqrt{\log 1/\delta}$
CHKRT	$(\log 1/\delta)^{1/4}$

Similar for linear regression (Hsu-Sabato)

Connections to Robust Statistics

$X_1, \dots, X_n \longrightarrow$

$\longrightarrow Y_1, \dots, Y_n$

- related techniques – new notions of high-dimensional quantiles & algorithms to compute them
- Simultaneously robust and high-confidence algos

[Depersin-Lecue ‘19, Diakonikolas-Kane-Pensia ‘20, H.-Li-Zhang ‘20,...]

Agenda

1. ~~Survey of recent (algorithmic) developments~~
2. Heavy-tailed covariance estimation
   - Algorithm of [Cherapanamjeri - H. - Kathuria - Raghavendra - Tripuraneni]
   - 2 Key techniques for concentration + convex prog.
     - bounded differences
     - SoS Bernstein

Setup: $X_1, \dots, X_n \sim X$ on $\mathbb{R}^d$, covariance $\Sigma \in \mathbb{R}^{d \times d}$

for simplicity, $\mathrm{Tr}\,\Sigma = O(d)$, $\|\Sigma\|_{op} = O(1)$

Theorem: Can find $\hat{\Sigma}_\delta$ s.t. $\|\hat{\Sigma}_\delta - \Sigma\|_{op} \le \tilde{O}\left(\sqrt{\frac{d}{n}} \cdot (\log 1/\delta)^{1/4} + \sqrt{\frac{\log 1/\delta}{n}}\right)$ w.p. $1-\delta$ in time $\mathrm{poly}(n, d, \log 1/\delta)$ if X is Certifiably $(2,8)$-hypercontractive.

- Sub-optimal by $(\log 1/\delta)^{1/4}$ – can remove in:
  - exp. time [Mendelson-Zhivotovsky]
  - $\mathrm{poly}(n, d, \frac{1}{\delta})$ time [CHKRT]
- best previous: $\tilde{O}\left(\sqrt{\frac{d \log 1/\delta}{n}}\right)$ [Minsker]

Only require $(2,4)$-hypercontractivity

Theorem: Can find $\hat{\Sigma}_S$ s.t. $\|\hat{\Sigma}_S - \Sigma\|_{op} \le \tilde{O}\left(\sqrt{\frac{d}{n}} \cdot (\log 1/\delta)^{1/4} + \sqrt{\frac{\log 1/\delta}{n}}\right)$ w.p. $1-\delta$ in time $\mathrm{poly}(n, d, \log 1/\delta)$ if X is certifiably (2,8)-hypercontractive.

(2,8)-hypercontractivity: $\mathbb{E}\langle X, v\rangle^8 \le O(\mathbb{E}\langle X, v\rangle^2)^4$

Certifiable: $O(\mathbb{E}\langle X, v\rangle^2)^4 - \mathbb{E}\langle X, v\rangle^8 = \sum p_i(v)^2$

- products of heavy-tailed dist'ns
- mixtures thereof
- affine transformations thereof
- ...

The High-Dimensional Median-of-Means Paradigm

① Buckets

$$\left.\begin{matrix} X_1 \\ \vdots \\ X_n \end{matrix}\right\} \begin{matrix} B_1 \text{ — } Z_1 = \mathbb{E}_{i\sim B_1} X_i X_i^T \\ \vdots \\ B_{\log 1/\delta} \text{ — } Z_{\log 1/\delta} = \mathbb{E}_{i\sim B_{\log 1/\delta}} X_i X_i^T \end{matrix}$$

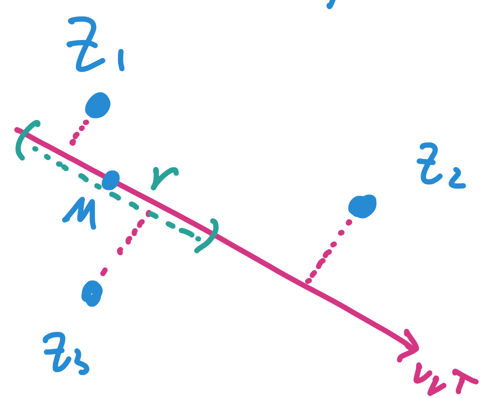
② Quantiles

"spectral r-center":

$\forall v, \ |\langle Z_i, vv^T\rangle - \langle M, vv^T\rangle| \le r$

for 60% of Z_i's.

$\|Z_i - M\|_{op} \le r$ for 60% of Z_i's

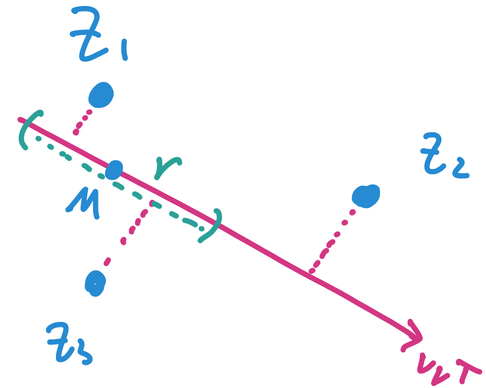


The High-Dimensional Median-of-Means Paradigm

"Spectral r-center":

$\forall v, \ |\langle Z_i, vv^T\rangle - \langle M, vv^T\rangle| \le r$

for 60% of Z_i's.



- Σ is a spectral $\sqrt{\frac{d}{n}} + \sqrt{\frac{\log 1/\delta}{n}}$ center of $Z_1 \dots Z_{\log 1/\delta}$ w.p. $1-\delta$

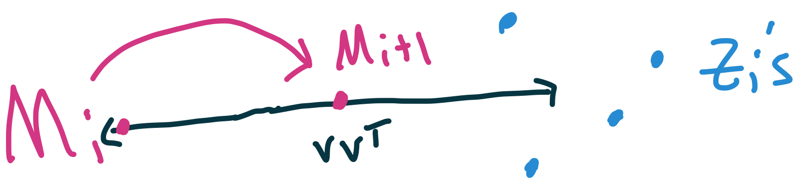
[Lugosi-Mendelson, Mendelson-Zhiva.]

- M, M' spectral r-centers $\rightarrow$ $\|M - M'\|_{op} \le 2r$

Just need to find a spectral r-center!

Our Strategy: $M_1 \to M_2 \to \dots \to M$

- Attempt to certify that M_i is a spectral r center
    - Success $\to$ output M_i
    - Failure $\to$ witness v, update $M_{i+1} = M_i \pm vv^T$



[CFB '19, CHKRT '20]

If GIVEN Σ and $z_1, \dots, z_{\log 1/\delta}$,

can you check that

$$|\langle z_i, vv^T\rangle - \langle \Sigma, vv^T\rangle| \leq r$$

for 60% of z_i's and all unit v?

Optimization Approach

Variables $b_1 \dots b_{\log' 1s}$, $v_1, \dots, v_d$

$$\max_{b, v} \sum_{i=1}^{\log' 1s} b_i$$

s.t. $b_i \in \{0,1\}$, $\|v\| = 1$,

$$b_i(\langle Z_i, vv^T \rangle - \langle M, vv^T \rangle) \geq b_i \cdot r$$

If GIVEN Σ and $Z_1, \dots, Z_{\log' 1s}$, can you check that

$$|\langle Z_i, vv^T \rangle - \langle \Sigma, vv^T \rangle| \leq r$$

for 60% of Z_i's and all unit v?

Optimization Approach

If GIVEN Σ and $z_1, \dots, z_{\log'|s|}$, can you check that

$$|\langle z_i, vv^T\rangle - \langle \Sigma, vv^T\rangle| \le r$$

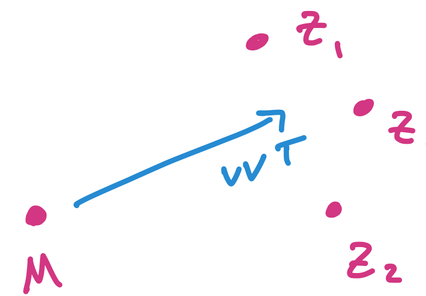
for 60% of z_i's and all unit v?

Variables $b_1 \dots b_{\log'|s|}$, $v_1, \dots, v_d = v$

$$\max_{b,v} \sum_{i=1}^{\log'|s|} b_i$$

s.t. $b_i \in \{0,1\}$, $\|v\| = 1$,

$$b_i(\langle z_i, vv^T\rangle - \langle M, vv^T\rangle) \ge b_i \cdot r$$



Optimization Approach

Variables $b_1 \ldots b_{\log 1/\delta}$, $v_1, \ldots, v_d$

$$\max_{b,v} \sum_{i=1}^{\log 1/\delta} b_i$$

s.t. $b_i \in \{0,1\}$, $\|v\| = 1$,

$$b_i\left(\langle Z_i, vv^T\rangle - \langle \mu, vv^T\rangle\right) \geq b_i \cdot r$$

If GIVEN Σ and $Z_1, \ldots, Z_{\log 1/\delta}$, can you check that

$$|\langle Z_i, vv^T\rangle - \langle \Sigma, vv^T\rangle| \leq r$$

for 60% of Z_i's and all unit v?

- Know: $\leq 0.6 \cdot \log 1/\delta$ for $r \approx \sqrt{\frac{d}{n}} + \sqrt{\frac{\log 1/\delta}{n}}$, w.p. $1-\delta$

- Convex relaxation: $\leq 0.6 \log 1/\delta$, for $r \approx (\log 1/\delta)^{1/4} \sqrt{\frac{d}{n}} + \sqrt{\frac{\log 1/\delta}{n}}$, w.p. $1-\delta$

Agenda

1. ~~Survey of recent (algorithmic) developments~~
2. Heavy-tailed covariance estimation
   - Algorithm ~~of [Cherapanamjeri - H. - Kathuria - Raghavendra - Tripuraneni]~~
   - 2 Key techniques for concentration + convex prog.
     - bounded differences
     - SoS Bernstein

Optimization Approach

Variables $b_1 \dots b_{\log 1/\delta}$, $v_1, \dots, v_d$

$$\max_{b,v} \sum_{i=1}^{\log 1/\delta} b_i$$

s.t. $b_i \in \{0,1\}$, $\|v\| = 1$,

$$b_i\left(\langle z_i, vv^T\rangle - \langle \mathbb{E} z_i, vv^T\rangle\right) \geq b_i \cdot r$$

If GIVEN Σ and $z_1, \dots, z_{\log 1/\delta}$, can you check that

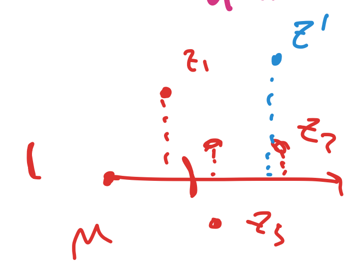
$$|\langle z_i, vv^T\rangle - \langle M, vv^T\rangle| \leq r$$

for 60% of z_i's and all unit v?

Any "reasonable" convex relaxation will satisfy bounded diffs.

$$\left|\widetilde{OPT}(z_1, \dots, z_{\log 1/\delta}) - \widetilde{OPT}(z_1', \dots, z_{\log 1/\delta})\right| \leq 1$$

$B_i \sim b_i$ $\quad 0 \leq B_i \leq 1$



New Goal: Convex relaxation of

$$z_i = \mathbb{E}_{j \sim B_i} x_j x_j^T$$

$$\max_{b,v} \sum_{i=1}^{\log 1/\delta} b_i$$

$$\text{s.t. } b_i \in \{0,1\}, \ \|v\| = 1,$$

$$b_i \left(\langle z_i, vv^T \rangle - \langle \Sigma, vv^T \rangle \right) \geq b_i \cdot r$$

s.t. $\mathbb{E}_{z_1 \dots z_{\log 1/\delta}} \mathrm{OPT} \leq 0.6 \log 1/\delta$, assuming only

Certifiable (2,8)-hypercontractivity

We use $\underbrace{\text{Sum of Squares Semidefinite Program}}$

hierarchy of convex relaxations for any polynomial optimization problem

$$\max_{b,v} \sum_{i=1}^{\log^{1/8}} b_i$$

$b_i(1-b_i)=0$ $\quad = \sum v_i^2$

Proxy variables for b_i, $b_i b_j$, $\underbrace{v_i v_j b_k}$

B_i $\quad$ . . .

$$\text{s.t.} \quad b_i \in \{\underline{\underline{0}},1\}, \quad \|v\|^2 = 1,$$

$$\underbrace{b_i(\langle Z_i, vv^T\rangle - \langle M, vv^T\rangle)} \geq b_i \cdot r$$

① $$\max_{b,v} \sum_{i=1}^{\log 1/\delta} b_i$$

s.t. $b_i \in \{0,1\}$, $\|v\|=1$,

$$b_i(\langle z_i, vv^T\rangle - \langle \Sigma, vv^T\rangle) \geq b_i \cdot r$$

$$\leq \max_{b,v} \frac{1}{r} \sum b_i(\langle z_i, vv^T\rangle - \langle \Sigma, vv^T\rangle)$$

$$\leq \max_{\|v\|=1} \frac{1}{r} \sqrt{\log\tfrac{1}{\delta}} \cdot \left(\sum (\langle z_i - \Sigma, vv^T\rangle^2\right)^{1/2}$$

Suffices to bound

$$\mathbb{E}_{z_1 \ldots z_{\log 1/\delta}} \operatorname{SoS}\left(\max_{\|v\|=1} \sum_{i=1}^{\log 1/\delta} \langle z_i - \Sigma, vv^T\rangle^2\right)$$

Suffices to bound

$$\mathop{\mathbb{E}}_{z_1 \ldots z_{\log 1/\delta}} \text{SoS}\left(\max_{\|v\|=1} \underbrace{\sum_{i=1}^{\log 1/\delta} \langle z_i - \Sigma, vv^T \rangle^2}_{\text{Sum of i.i.d. polynomials in } v} \right)$$

Suffices to bound

$$\underset{z_1 \ldots z_{\log 1/\delta}}{\mathbb{E}} \mathrm{SoS}\left(\max_{\|v\|=1} \underbrace{\sum_{i=1}^{\log 1/\delta} \langle z_i - \Sigma, vv^T \rangle^2}_{\text{Sum of i.i.d. polynomials in } v} \right)$$

$$\max_{\substack{Y \in \mathbb{R}^{d^2 \times d^2} \\ Y \in \mathrm{SoS} \\ \mathrm{Tr}\, Y = 1,\ Y \succeq 0}} \sum \left\langle (z_i - \Sigma) \otimes (z_i - \Sigma),\ Y \right\rangle$$

Suffices to bound

$$\mathbb{E}_{z_1 \ldots z_{\log 1/\delta}} \mathrm{SoS}\left(\max_{\|v\|=1} \underbrace{\sum_{i=1}^{\log 1/\delta} \langle z_i - \Sigma, vv^T \rangle^2}_{\text{sum of i.i.d. polynomials in } v} \right)$$

$$\mathbb{E} \max_{Y \in \mathbb{R}^{d^2 \times d^2}} \sum \langle (z_i - \Sigma) \otimes (z_i - \Sigma), Y \rangle$$

$\{Y \in \mathrm{SoS}$

$$= \mathbb{E} \left\| \sum (z_i - \Sigma) \otimes (z_i - \Sigma) \right\|_{\mathrm{SoS}}$$

② $$\mathbb{E}\left\| \sum (z_i - \Sigma) \otimes (z_i - \Sigma) \right\|_{\underline{SoS}}$$

SoS Bernstein: control via $$\Big\| \underbrace{\mathbb{E} \sum (z_i - \Sigma)^2 \otimes (z_i - \Sigma)^2}_{\text{degree 8 polynomial in } X} \Big\|_{\underline{SoS}}$$

$\Rightarrow$ control via (2,8) certifiable hypercontractivity:

$$\|\mathbb{E}\, X^{\otimes 8}\|_{SoS} \le O(1)$$

Yields covariance estimation algo. with $r = \tilde{O}\left((\log)^{1/4}\sqrt{\frac{d}{n}} + \sqrt{\frac{\log 1/\delta}{n}}\right)$

Open: remove $(\log 1/\delta)^{1/4}$?

Open: optimal heavy-tailed linear regression?

Applications of

- bdd. diffs for convex programs ?
- SoS Bernstein?

$\mathbb{E}\langle x, \underline{v} \rangle^8$

THANKS