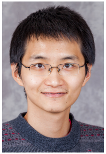
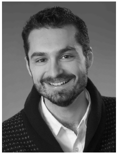
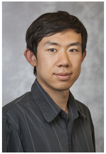
High-Dimensional Robust Mean Estimation in Nearly-Linear Time

Yu Cheng [1] Ilias Diakonikolas [2] Rong Ge [1]

[1]Duke University

[2]University of Southern California

Mean Estimation

Mean Estimation

- *Input:* N samples $\{X_1, \ldots, X_N\}$ drawn from $\mathcal{N}(\mu^\star, I)$ on $\mathbb{R}^d$.
- *Goal:* Learn $\mu^\star$.

Mean Estimation

Mean Estimation

- *Input:* N samples $\{X_1, \ldots, X_N\}$ drawn from $\mathcal{N}(\mu^\star, I)$ on $\mathbb{R}^d$.
- *Goal:* Learn $\mu^\star$.
- Empirical mean $\widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i$ works:

$$\|\widehat{\mu} - \mu^\star\|_2 \leq \epsilon \text{ when } N = \Omega(d/\epsilon^2).$$

Robust Mean Estimation

Definition (ϵ-Corruption)

- N samples are drawn i.i.d. from the ground-truth distribution D.

Robust Mean Estimation

Definition (ϵ-Corruption)

- N samples are drawn i.i.d. from the ground-truth distribution D.
- Adversary replaces ϵN samples with arbitrary points (after inspecting D, the samples, and the algorithm).

Robust Mean Estimation

Definition (ϵ-Corruption)

- N samples are drawn i.i.d. from the ground-truth distribution D.
- Adversary replaces ϵN samples with arbitrary points (after inspecting D, the samples, and the algorithm).

Robust Mean Estimation

- *Input:* an ϵ-corrupted set of N samples $\{X_1, \ldots, X_N\}$ drawn from an unknown distribution D on $\mathbb{R}^d$ with mean $\mu^\star$.
- *Goal:* Learn $\mu^\star$ in ℓ_2-norm.

Previous Work

Robustly learn $\mu^\star$ given ϵ-corrupted samples from $\mathcal{N}(\mu^\star, I)$:

Algorithm	Error Guarantee	Poly-Time?

Previous Work

Robustly learn $\mu^\star$ given ϵ-corrupted samples from $\mathcal{N}(\mu^\star, I)$:

Algorithm	Error Guarantee	Poly-Time?
Tukey Median	$O(\epsilon)$	No
Geometric Median	$O(\epsilon\sqrt{d})$	Yes
Tournament	$O(\epsilon)$	No
Pruning	$O(\epsilon\sqrt{d})$	Yes
RANSAC	∞	Yes

Previous Work

Robustly learn $\mu^\star$ given ϵ-corrupted samples from $\mathcal{N}(\mu^\star, I)$:

Algorithm	Error Guarantee	Poly-Time?
Tukey Median	$O(\epsilon)$	No
Geometric Median	$O(\epsilon\sqrt{d})$	Yes
Tournament	$O(\epsilon)$	No
Pruning	$O(\epsilon\sqrt{d})$	Yes
RANSAC	∞	Yes
[LRV'16]	$O(\epsilon\sqrt{\log d})$	Yes
[DKKLMS'16]	$O(\epsilon\sqrt{\log(1/\epsilon)})$	Yes

Previous Work

Robustly learn $\mu^\star$ given ϵ-corrupted samples from $\mathcal{N}(\mu^\star, I)$:

Algorithm	Error Guarantee	Poly-Time?
Tukey Median	$O(\epsilon)$	No
Geometric Median	$O(\epsilon\sqrt{d})$	Yes
Tournament	$O(\epsilon)$	No
Pruning	$O(\epsilon\sqrt{d})$	Yes
RANSAC	∞	Yes
[LRV'16]	$O(\epsilon\sqrt{\log d})$	Yes
[DKKLMS'16]	$O(\epsilon\sqrt{\log(1/\epsilon)})$	Yes

Previous Work

Robustly learn $\mu^\star$ given ϵ-corrupted samples from $\mathcal{N}(\mu^\star, I)$:

Algorithm	Error (δ)	Runtime

Previous Work

Robustly learn $\mu^\star$ given ϵ-corrupted samples from $\mathcal{N}(\mu^\star, I)$:

Algorithm	Error (δ)	Runtime
Dimension Halving [LRV'16]	$O(\epsilon\sqrt{\log d})$	$\Omega(Nd^2)$ + SVD

Previous Work

Robustly learn $\mu^\star$ given ϵ-corrupted samples from $\mathcal{N}(\mu^\star, I)$:

Algorithm	Error (δ)	Runtime
Dimension Halving [LRV'16]	$O(\epsilon\sqrt{\log d})$	$\Omega(Nd^2)$ + SVD
Convex Programming [DKKLMS'16]	$O(\epsilon\sqrt{\log(1/\epsilon)})$	Ellipsoid Algorithm

Previous Work

Robustly learn $\mu^\star$ given ϵ-corrupted samples from $\mathcal{N}(\mu^\star, I)$:

Algorithm	Error (δ)	Runtime
Dimension Halving [LRV'16]	$O(\epsilon\sqrt{\log d})$	$\Omega(Nd^2)$ + SVD
Convex Programming [DKKLMS'16]	$O(\epsilon\sqrt{\log(1/\epsilon)})$	Ellipsoid Algorithm
Filtering [DKKLMS'16]	$O(\epsilon\sqrt{\log(1/\epsilon)})$	$\Omega(Nd^2)$

Previous Work

Robustly learn $\mu^\star$ given ϵ-corrupted samples from $\mathcal{N}(\mu^\star, I)$:

Algorithm	Error (δ)	Runtime
Dimension Halving [LRV'16]	$O(\epsilon\sqrt{\log d})$	$\Omega(Nd^2)$ + SVD
Convex Programming [DKKLMS'16]	$O(\epsilon\sqrt{\log(1/\epsilon)})$	Ellipsoid Algorithm
Filtering [DKKLMS'16]	$O(\epsilon\sqrt{\log(1/\epsilon)})$	$\Omega(Nd^2)$
This paper	$O(\epsilon\sqrt{\log(1/\epsilon)})$	$\widetilde{O}(Nd/\epsilon^6)$

All these algorithms have sample complexity $N = O(d/\delta^2)$.

Our Results

Robustly learn $\mu^\star$ given ϵ-corrupted samples from D on $\mathbb{R}^d$.

Distribution	Error (δ)	# of Samples (N)	Runtime

Our Results

Robustly learn $\mu^\star$ given ϵ-corrupted samples from D on $\mathbb{R}^d$.

Distribution	Error (δ)	# of Samples (N)	Runtime
Sub-Gaussian	$O(\epsilon\sqrt{\log(1/\epsilon)})$	$O(d/\delta^2)$	$\widetilde{O}(Nd/\epsilon^6)$
Bounded Covariance ($\Sigma \preceq I$)	$O(\sqrt{\epsilon})$	$\widetilde{O}(d/\delta^2)$	

Our Results

Robustly learn $\mu^\star$ given ϵ-corrupted samples from D on $\mathbb{R}^d$.

Distribution	Error (δ)	# of Samples (N)	Runtime
Sub-Gaussian	$O(\epsilon\sqrt{\log(1/\epsilon)})$	$O(d/\delta^2)$	$\widetilde{O}(Nd/\epsilon^6)$
Bounded Covariance ($\Sigma \preceq I$)	$O(\sqrt{\epsilon})$	$\widetilde{O}(d/\delta^2)$	

When ϵ is constant, our algorithm has the best possible error guarantee, sample complexity, and running time (up to polylogarithmic factors).

Our Results

Robustly learn $\mu^\star$ given ϵ-corrupted samples from D on $\mathbb{R}^d$.

Distribution	Error (δ)	# of Samples (N)	Runtime
Sub-Gaussian	$O(\epsilon\sqrt{\log(1/\epsilon)})$	$O(d/\delta^2)$	$\widetilde{O}(Nd/\epsilon^6)$
Bounded Covariance ($\Sigma \preceq I$)	$O(\sqrt{\epsilon})$	$\widetilde{O}(d/\delta^2)$	

When ϵ is constant, our algorithm has the best possible error guarantee, sample complexity, and running time (up to polylogarithmic factors).

The ϵ^{-6} in runtime comes from packing/covering SDP solvers.

Our Results

Robustly learn $\mu^\star$ given ϵ-corrupted samples from D on $\mathbb{R}^d$.

Distribution	Error (δ)	# of Samples (N)	Runtime
Sub-Gaussian	$O(\epsilon\sqrt{\log(1/\epsilon)})$	$O(d/\delta^2)$	$\widetilde{O}(Nd/\epsilon^6)$
Bounded Covariance ($\Sigma \preceq I$)	$O(\sqrt{\epsilon})$	$\widetilde{O}(d/\delta^2)$	

When ϵ is constant, our algorithm has the best possible error guarantee, sample complexity, and running time (up to polylogarithmic factors).

The ϵ^{-6} in runtime comes from packing/covering SDP solvers.

Suppose we can solve one packing/covering SDP in time $T = T(N, d, \epsilon)$.

Our runtime is $O(Nd) + \widetilde{O}(\log^2(d/\epsilon))\left(T + d^2\right)$.

Our Results

Distribution	Error (δ)	# of Samples (N)	Runtime

Our Results

Distribution	Error (δ)	# of Samples (N)	Runtime
Sub-Gaussian	$O(\epsilon\sqrt{\log(1/\epsilon)})$	$O(d/\delta^2)$	$\widetilde{O}(Nd/\epsilon^6)$
Bounded Covariance ($\Sigma \preceq I$)	$O(\sqrt{\epsilon})$	$\widetilde{O}(d/\delta^2)$	

Our Results

Distribution	Error (δ)	# of Samples (N)	Runtime
Sub-Gaussian	$O(\epsilon\sqrt{\log(1/\epsilon)})$	$O(d/\delta^2)$	$\widetilde{O}(Nd/\epsilon^6)$
Bounded Covariance ($\Sigma \preceq I$)	$O(\sqrt{\epsilon})$	$\widetilde{O}(d/\delta^2)$	

Robust mean estimation under bounded covariance assumptions has been used as a subroutine to obtain robust learners for a wide range of supervised learning problems that can be phrased as stochastic convex programs.

Our Results

Distribution	Error (δ)	# of Samples (N)	Runtime
Sub-Gaussian	$O(\epsilon\sqrt{\log(1/\epsilon)})$	$O(d/\delta^2)$	$\widetilde{O}(Nd/\epsilon^6)$
Bounded Covariance ($\Sigma \preceq I$)	$O(\sqrt{\epsilon})$	$\widetilde{O}(d/\delta^2)$	

Robust mean estimation under bounded covariance assumptions has been used as a subroutine to obtain robust learners for a wide range of supervised learning problems that can be phrased as stochastic convex programs.

Our result provides a faster implementation of such a subroutine, hence yields faster robust algorithms for all these problems.

Intuition: Reweight the Samples

[DKKLMS'16]: To shift the empirical mean far from $\mu^\star$, the corrupted samples must introduce a large eigenvalue in a covariance-like matrix.

Intuition: Reweight the Samples

[DKKLMS'16]: To shift the empirical mean far from $\mu^\star$, the corrupted samples must introduce a large eigenvalue in a covariance-like matrix.

Good Weights

$$\begin{array}{ll} \text{minimize} & \lambda_{\max}\left(\sum_{i=1}^{N} w_i (X_i - \mu^\star)(X_i - \mu^\star)^\top\right) \\ \text{subject to} & w \in \Delta_{N,\epsilon} \qquad \left(\sum_i w_i = 1 \text{ and } 0 \le w_i \le \frac{1}{(1-\epsilon)N}\right) \end{array}$$

Intuition: Reweight the Samples

[DKKLMS'16]: To shift the empirical mean far from $\mu^\star$, the corrupted samples must introduce a large eigenvalue in a covariance-like matrix.

Good Weights

$$\begin{array}{ll} \text{minimize} & \lambda_{\max}\left(\sum_{i=1}^{N} w_i (X_i - \mu^\star)(X_i - \mu^\star)^\top\right) \\ \text{subject to} & w \in \Delta_{N,\epsilon} \qquad \left(\sum_i w_i = 1 \text{ and } 0 \le w_i \le \frac{1}{(1-\epsilon)N}\right) \end{array}$$

Lemma ([DKKLMS'16])

If we can find a near-optimal solution w, we can output $\widehat{\mu}_w = \sum_i w_i X_i$.

Intuition: Reweight the Samples

[DKKLMS'16]: To shift the empirical mean far from $\mu^\star$, the corrupted samples must introduce a large eigenvalue in a covariance-like matrix.

Good Weights

$$\begin{array}{ll} \text{minimize} & \lambda_{\max}\left(\sum_{i=1}^{N} w_i (X_i - \mu^\star)(X_i - \mu^\star)^\top\right) \\ \text{subject to} & w \in \Delta_{N,\epsilon} \qquad \left(\sum_i w_i = 1 \text{ and } 0 \le w_i \le \frac{1}{(1-\epsilon)N}\right) \end{array}$$

Lemma ([DKKLMS'16])

If we can find a near-optimal solution w, we can output $\widehat{\mu}_w = \sum_i w_i X_i$.

This looks like a packing SDP in w (which we can solve in nearly-linear time).

Except that ...

Intuition: Reweight the Samples

[DKKLMS'16]: To shift the empirical mean far from $\mu^\star$, the corrupted samples must introduce a large eigenvalue in a covariance-like matrix.

Good Weights

$$\begin{array}{ll} \text{minimize} & \lambda_{\max}\left(\sum_{i=1}^{N} w_i (X_i - \mu^\star)(X_i - \mu^\star)^\top\right) \\ \text{subject to} & w \in \Delta_{N,\epsilon} \qquad \left(\sum_i w_i = 1 \text{ and } 0 \le w_i \le \frac{1}{(1-\epsilon)N}\right) \end{array}$$

Lemma ([DKKLMS'16])

If we can find a near-optimal solution w, we can output $\widehat{\mu}_w = \sum_i w_i X_i$.

This looks like a packing SDP in w (which we can solve in nearly-linear time).
Except that ... we do not know $\mu^\star$.

Our Approach

Idea: guess the mean ν and solve the SDP with parameter ν.

Our Approach

Idea: guess the mean ν and solve the SDP with parameter ν.

Primal SDP (with parameter ν)

$$\begin{array}{ll} \text{minimize} & \lambda_{\max}\left(\sum_{i=1}^{N} w_i (X_i - \nu)(X_i - \nu)^\top\right) \\ \text{subject to} & w \in \Delta_{N,\epsilon} \end{array}$$

Our Approach

Idea: guess the mean ν and solve the SDP with parameter ν.

Primal SDP (with parameter ν)

$$\begin{array}{ll} \text{minimize} & \lambda_{\max}\left(\sum_{i=1}^{N} w_i (X_i - \nu)(X_i - \nu)^\top\right) \\ \text{subject to} & w \in \Delta_{N,\epsilon} \end{array}$$

We give a win-win analysis: either

- a near-optimal solution w to the primal SDP give a good answer $\widehat{\mu}_w$, or

Our Approach

Idea: guess the mean ν and solve the SDP with parameter ν.

Primal SDP (with parameter ν)

$$\begin{array}{ll} \text{minimize} & \lambda_{\max}\left(\sum_{i=1}^{N} w_i (X_i - \nu)(X_i - \nu)^\top\right) \\ \text{subject to} & w \in \Delta_{N,\epsilon} \end{array}$$

We give a win-win analysis: either

- a near-optimal solution w to the primal SDP give a good answer $\widehat{\mu}_w$, or
- a near-optimal solution to the dual SDP yields a new guess ν' that is closer to $\mu^\star$ by a constant factor.

Our Approach

Iteratively move ν closer to $\mu^\star$ using the dual SDP,

until primal SDP has a good solution and we can output $\widehat{\mu}_w$.

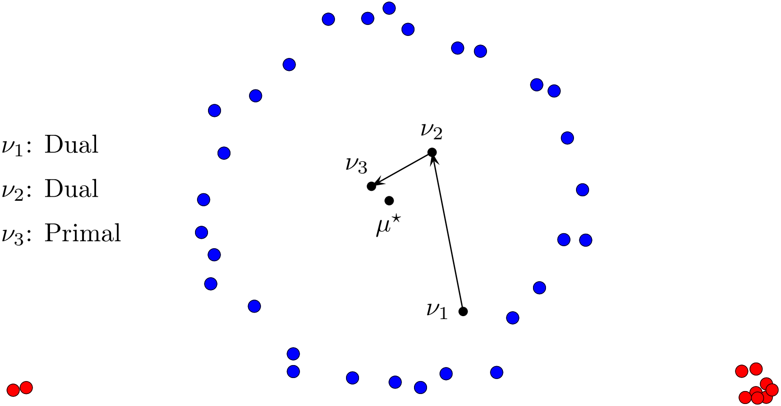
Our Approach

Iteratively move ν closer to $\mu^\star$ using the dual SDP,
until primal SDP has a good solution and we can output $\widehat{\mu}_w$.

ν_1: Dual

ν_2: Dual

ν_3: Primal



Dual SDP

Primal SDP (with parameter ν)

$$\begin{aligned} \text{minimize} \quad & \lambda_{\max}\left(\textstyle\sum_{i=1}^{N} w_i (X_i - \nu)(X_i - \nu)^\top\right) \\ \text{subject to} \quad & w \in \Delta_{N,\epsilon} \end{aligned}$$

Dual SDP

Primal SDP (with parameter ν)

$$\begin{array}{ll} \text{minimize} & \lambda_{\max}\left(\sum_{i=1}^{N} w_i (X_i - \nu)(X_i - \nu)^\top\right) \\ \text{subject to} & w \in \Delta_{N,\epsilon} \end{array}$$

Dual SDP (with parameter ν)

$$\begin{array}{ll} \text{maximize} & \text{Mean of the smallest } (1-\epsilon)\text{-fraction of } \left((X_i - \nu)^\top M (X_i - \nu)\right)_{i=1}^{N} \\ \text{subject to} & M \succeq 0, \operatorname{tr}(M) \le 1 \end{array}$$

Dual SDP

Primal SDP (with parameter ν)

$$\begin{array}{ll} \text{minimize} & \lambda_{\max}\left(\sum_{i=1}^{N} w_i (X_i - \nu)(X_i - \nu)^\top\right) \\ \text{subject to} & w \in \Delta_{N,\epsilon} \end{array}$$

Dual SDP (with parameter ν)

$$\begin{array}{ll} \text{maximize} & \text{Mean of the smallest } (1-\epsilon)\text{-fraction of } \left((X_i - \nu)^\top M (X_i - \nu)\right)_{i=1}^{N} \\ \text{subject to} & M \succeq 0, \operatorname{tr}(M) \le 1 \end{array}$$

- The dual SDP certifies that there are no good weights that can make the spectral norm small.

Dual SDP

Primal SDP (with parameter ν)

$$\begin{aligned} \text{minimize} \quad & \lambda_{\max}\left(\sum_{i=1}^{N} w_i (X_i - \nu)(X_i - \nu)^\top\right) \\ \text{subject to} \quad & w \in \Delta_{N,\epsilon} \end{aligned}$$

Dual SDP (with parameter ν)

$$\begin{aligned} \text{maximize} \quad & \text{Mean of the smallest } (1-\epsilon)\text{-fraction of } \left((X_i - \nu)^\top M (X_i - \nu)\right)_{i=1}^{N} \\ \text{subject to} \quad & M \succeq 0, \operatorname{tr}(M) \le 1 \end{aligned}$$

- The dual SDP certifies that there are no good weights that can make the spectral norm small.
- If the solution is rank-one: $M = yy^\top$, then in the direction of y, the variance is large no matter how we reweight the samples.

Dual SDP

Primal SDP (with parameter ν)

$$\begin{aligned} \text{minimize} \quad & \lambda_{\max}\left(\sum_{i=1}^N w_i (X_i - \nu)(X_i - \nu)^\top\right) \\ \text{subject to} \quad & w \in \Delta_{N,\epsilon} \end{aligned}$$

Dual SDP (with parameter ν)

$$\begin{aligned} \text{maximize} \quad & \text{Mean of the smallest } (1-\epsilon)\text{-fraction of } \left((X_i - \nu)^\top M (X_i - \nu)\right)_{i=1}^N \\ \text{subject to} \quad & M \succeq 0, \operatorname{tr}(M) \le 1 \end{aligned}$$

- The dual SDP certifies that there are no good weights that can make the spectral norm small.
- If the solution is rank-one: $M = yy^\top$, then in the direction of y, the variance is large no matter how we reweight the samples.
- Intuition: When ν is far from $\mu^\star$, y should align with $(\nu - \mu^\star)$.

Dual SDP

Why would the dual SDP pick the direction $(\nu - \mu^\star)$?

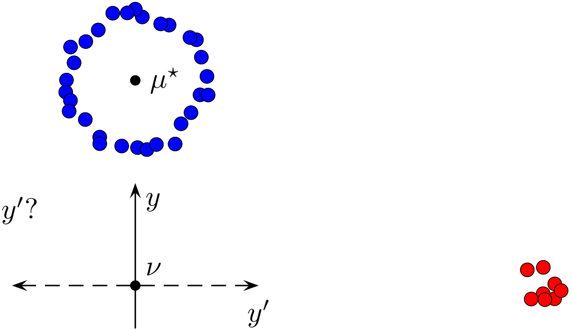
Dual SDP

Why would the dual SDP pick the direction $(\nu - \mu^\star)$?

$M = yy^\top$

$y \approx (\nu - \mu^\star)$

Why is y better than y'?



Conditions on the Good Samples

We require the following deterministic conditions on the good samples:

Concentration Bounds (for Sub-Gaussian Distributions)

For all $w \in \Delta_{N,\epsilon}$ (G is the set of good samples):

$$\left\| \sum_{i \in G} w_i (X_i - \mu^\star) \right\|_2 \le O(\epsilon \sqrt{\log(1/\epsilon)}) =: \delta_1,$$

$$\left\| \sum_{i \in G} w_i (X_i - \mu^\star)(X_i - \mu^\star)^\top - I \right\|_2 \le O(\epsilon \log(1/\epsilon)) =: \delta_2.$$

Conditions on the Good Samples

We require the following deterministic conditions on the good samples:

Concentration Bounds (for Sub-Gaussian Distributions)

For all $w \in \Delta_{N,\epsilon}$ (G is the set of good samples):

$$\left\| \sum_{i \in G} w_i (X_i - \mu^\star) \right\|_2 \le O(\epsilon \sqrt{\log(1/\epsilon)}) =: \delta_1,$$

$$\left\| \sum_{i \in G} w_i (X_i - \mu^\star)(X_i - \mu^\star)^\top - I \right\|_2 \le O(\epsilon \log(1/\epsilon)) =: \delta_2.$$

Removing ϵN samples does not affect the first/second moments too much.

Conditions on the Good Samples

We require the following deterministic conditions on the good samples:

Concentration Bounds (for Sub-Gaussian Distributions)

For all $w \in \Delta_{N,\epsilon}$ (G is the set of good samples):

$$\left\| \sum_{i \in G} w_i (X_i - \mu^\star) \right\|_2 \leq O(\epsilon \sqrt{\log(1/\epsilon)}) =: \delta_1,$$

$$\left\| \sum_{i \in G} w_i (X_i - \mu^\star)(X_i - \mu^\star)^\top - I \right\|_2 \leq O(\epsilon \log(1/\epsilon)) =: \delta_2.$$

Removing ϵN samples does not affect the first/second moments too much.

This is the only place we use the sub-Gaussian assumption.

Optimal Value of the SDPs

Lemma

When $\|\nu - \mu^\star\|_2 \geq \Omega(\beta)$,

$$1 + 0.99 \|\nu - \mu^\star\|_2^2 \leq OPT_\nu \leq 1 + 1.01 \|\nu - \mu^\star\|_2^2 .$$

Optimal Value of the SDPs

Lemma

When $\|\nu - \mu^\star\|_2 \geq \Omega(\beta)$,

$$1 + 0.99\,\|\nu - \mu^\star\|_2^2 \leq OPT_\nu \leq 1 + 1.01\,\|\nu - \mu^\star\|_2^2 .$$

Proof

One feasible primal solution is to set $w_i = \frac{1}{|G|}$ for all $i \in G$.

Optimal Value of the SDPs

Lemma

When $\|\nu - \mu^\star\|_2 \geq \Omega(\beta)$,

$$1 + 0.99 \|\nu - \mu^\star\|_2^2 \leq OPT_\nu \leq 1 + 1.01 \|\nu - \mu^\star\|_2^2 .$$

Proof

One feasible primal solution is to set $w_i = \frac{1}{|G|}$ for all $i \in G$.

$$\mathrm{OPT}_\nu \leq \lambda_{\max}\left(\sum_{i=1}^{N} w_i (X_i - \nu)(X_i - \nu)^\top\right) = \max_{\|y\|_2 = 1} \sum_{i \in G} w_i \langle X_i - \nu, y\rangle^2$$

Optimal Value of the SDPs

Lemma

When $\|\nu - \mu^\star\|_2 \geq \Omega(\beta)$,

$$1 + 0.99 \|\nu - \mu^\star\|_2^2 \leq OPT_\nu \leq 1 + 1.01 \|\nu - \mu^\star\|_2^2 .$$

Proof

One feasible primal solution is to set $w_i = \frac{1}{|G|}$ for all $i \in G$.

$$\mathrm{OPT}_\nu \leq \lambda_{\max}\left(\sum_{i=1}^{N} w_i (X_i - \nu)(X_i - \nu)^\top\right) = \max_{\|y\|_2=1} \sum_{i \in G} w_i \langle X_i - \nu, y\rangle^2$$

$$= \max_{\|y\|_2=1}\left(\sum_{i \in G} w_i \langle X_i - \mu^\star, y\rangle^2 + \langle \mu^\star - \nu, y\rangle^2 + 2\langle \sum_{i \in G} w_i (X_i - \mu^\star), y\rangle \langle \mu^\star - \nu, y\rangle\right)$$

Optimal Value of the SDPs

Lemma

When $\|\nu - \mu^\star\|_2 \geq \Omega(\beta)$,

$$1 + 0.99\,\|\nu - \mu^\star\|_2^2 \leq OPT_\nu \leq 1 + 1.01\,\|\nu - \mu^\star\|_2^2 .$$

Proof

One feasible primal solution is to set $w_i = \frac{1}{|G|}$ for all $i \in G$.

$$\mathrm{OPT}_\nu \leq \lambda_{\max}\left(\sum_{i=1}^{N} w_i (X_i - \nu)(X_i - \nu)^\top\right) = \max_{\|y\|_2=1} \sum_{i \in G} w_i \langle X_i - \nu, y\rangle^2$$

$$= \max_{\|y\|_2=1}\left(\sum_{i \in G} w_i \langle X_i - \mu^\star, y\rangle^2 + \langle \mu^\star - \nu, y\rangle^2 + 2\langle \sum_{i \in G} w_i (X_i - \mu^\star), y\rangle \langle \mu^\star - \nu, y\rangle\right)$$

$$\leq \max_{\|y\|_2=1}\left((1 + \delta_2) + \langle \mu^\star - \nu, y\rangle^2 + 2\delta_1 \langle \mu^\star - \nu, y\rangle\right)$$

Optimal Value of the SDPs

Lemma

When $\|\nu - \mu^\star\|_2 \geq \Omega(\beta)$,

$$1 + 0.99 \|\nu - \mu^\star\|_2^2 \leq OPT_\nu \leq 1 + 1.01 \|\nu - \mu^\star\|_2^2 .$$

Proof

One feasible primal solution is to set $w_i = \frac{1}{|G|}$ for all $i \in G$.

$$\begin{aligned}
\mathrm{OPT}_\nu &\leq \lambda_{\max}\left(\sum_{i=1}^{N} w_i (X_i - \nu)(X_i - \nu)^\top\right) = \max_{\|y\|_2=1} \sum_{i \in G} w_i \langle X_i - \nu, y\rangle^2 \\
&= \max_{\|y\|_2=1}\left(\sum_{i \in G} w_i \langle X_i - \mu^\star, y\rangle^2 + \langle \mu^\star - \nu, y\rangle^2 + 2\langle \sum_{i \in G} w_i (X_i - \mu^\star), y\rangle \langle \mu^\star - \nu, y\rangle\right) \\
&\leq \max_{\|y\|_2=1}\left((1 + \delta_2) + \langle \mu^\star - \nu, y\rangle^2 + 2\delta_1 \langle \mu^\star - \nu, y\rangle\right) \\
&= (1 + \delta_2) + \|\mu^\star - \nu\|_2^2 + 2\delta_1 \|\mu^\star - \nu\|_2
\end{aligned}$$

Optimal Value of the SDPs

Lemma

When $\|\nu - \mu^\star\|_2 \geq \Omega(\beta)$,

$$1 + 0.99\,\|\nu - \mu^\star\|_2^2 \leq OPT_\nu \leq 1 + 1.01\,\|\nu - \mu^\star\|_2^2 .$$

Proof

One feasible primal solution is to set $w_i = \frac{1}{|G|}$ for all $i \in G$.

$$\begin{aligned}
\mathrm{OPT}_\nu &\leq \lambda_{\max}\left(\sum_{i=1}^{N} w_i (X_i - \nu)(X_i - \nu)^\top\right) = \max_{\|y\|_2=1} \sum_{i \in G} w_i \langle X_i - \nu, y\rangle^2 \\
&= \max_{\|y\|_2=1}\left(\sum_{i\in G} w_i \langle X_i - \mu^\star, y\rangle^2 + \langle \mu^\star - \nu, y\rangle^2 + 2\langle \sum_{i \in G} w_i (X_i - \mu^\star), y\rangle \langle \mu^\star - \nu, y\rangle\right) \\
&\leq \max_{\|y\|_2=1}\left((1+\delta_2) + \langle \mu^\star - \nu, y\rangle^2 + 2\delta_1 \langle \mu^\star - \nu, y\rangle\right) \\
&= (1+\delta_2) + \|\mu^\star - \nu\|_2^2 + 2\delta_1 \|\mu^\star - \nu\|_2 \qquad \text{(so } \beta = \sqrt{\delta_2} = \sqrt{\epsilon \ln(1/\epsilon)}.\text{)}
\end{aligned}$$

When Primal SDP Has Good Solutions

Lemma (Good Primal Solutions ⇒ Correct Mean)

For all $w \in \Delta_{N,2\epsilon}$, *if* $\|\widehat{\mu}_w - \mu^\star\|_2 \geq \Omega(\delta)$, *then for all* $\nu \in \mathbb{R}^d$,

$$\lambda_{\max}\left(\sum_{i=1}^{N} w_i (X_i - \nu)(X_i - \nu)^\top\right) \geq 1 + \Omega(\delta^2/\epsilon) = 1 + \Omega(\beta^2).$$

When Primal SDP Has Good Solutions

Lemma (Good Primal Solutions ⇒ Correct Mean)

For all $w \in \Delta_{N,2\epsilon}$, if $\|\widehat{\mu}_w - \mu^\star\|_2 \geq \Omega(\delta)$, then for all $\nu \in \mathbb{R}^d$,

$$\lambda_{\max}\left(\sum_{i=1}^{N} w_i (X_i - \nu)(X_i - \nu)^\top\right) \geq 1 + \Omega(\delta^2/\epsilon) = 1 + \Omega(\beta^2).$$

Implication: if objective value of w is small with any ν, then $\widehat{\mu}_w$ is close to $\mu^\star$.

When Primal SDP Has Good Solutions

Lemma (Good Primal Solutions $\Rightarrow$ Correct Mean)

For all $w \in \Delta_{N,2\epsilon}$, if $\|\widehat{\mu}_w - \mu^\star\|_2 \geq \Omega(\delta)$, then for all $\nu \in \mathbb{R}^d$,

$$\lambda_{\max}\left(\sum_{i=1}^{N} w_i (X_i - \nu)(X_i - \nu)^\top\right) \geq 1 + \Omega(\delta^2/\epsilon) = 1 + \Omega(\beta^2).$$

Implication: if objective value of w is small with any ν, then $\widehat{\mu}_w$ is close to $\mu^\star$.

Proof sketch:

- ν must be close to $\mu^\star$, otherwise $\mathrm{OPT}_\nu \approx 1 + \|\nu - \mu^\star\|_2^2$ is already large.

When Primal SDP Has Good Solutions

Lemma (Good Primal Solutions ⇒ Correct Mean)

For all $w \in \Delta_{N,2\epsilon}$, *if* $\|\widehat{\mu}_w - \mu^\star\|_2 \geq \Omega(\delta)$, *then for all* $\nu \in \mathbb{R}^d$,

$$\lambda_{\max}\left(\sum_{i=1}^{N} w_i (X_i - \nu)(X_i - \nu)^\top\right) \geq 1 + \Omega(\delta^2/\epsilon) = 1 + \Omega(\beta^2).$$

Implication: if objective value of w is small with any ν, then $\widehat{\mu}_w$ is close to $\mu^\star$.

Proof sketch:

- ν must be close to $\mu^\star$, otherwise $\mathrm{OPT}_\nu \approx 1 + \|\nu - \mu^\star\|_2^2$ is already large.
- When ν is close to $\mu^\star$, $(X_i - \nu)(X_i - \nu)^\top$ is close to $(X_i - \mu^\star)(X_i - \mu^\star)^\top$.

When Dual SDP Has Good Solutions

Lemma (Good Dual Solutions $\Rightarrow$ Better ν)

Fix an approximately optimal solution M to the dual SDP with parameter ν. If the objective value of M is at least $1 + \Omega(\beta^2)$, then we can find $\nu' \in \mathbb{R}^d$ such that $\|\nu' - \mu^\star\|_2 \le \frac{9}{10} \|\nu - \mu^\star\|_2$.

When Dual SDP Has Good Solutions

Lemma (Good Dual Solutions ⇒ Better ν)

Fix an approximately optimal solution M to the dual SDP with parameter ν. If the objective value of M is at least $1 + \Omega(\beta^2)$, then we can find $\nu' \in \mathbb{R}^d$ such that $\|\nu' - \mu^\star\|_2 \leq \frac{9}{10}\|\nu - \mu^\star\|_2$.

Intuitively, if the dual SDP throws away all the bad samples,
$1 + \|\nu - \mu^\star\|_2^2 \approx \mathrm{OPT}$

When Dual SDP Has Good Solutions

Lemma (Good Dual Solutions ⇒ Better ν)

Fix an approximately optimal solution M to the dual SDP with parameter ν. If the objective value of M is at least $1 + \Omega(\beta^2)$, then we can find $\nu' \in \mathbb{R}^d$ such that $\|\nu' - \mu^\star\|_2 \le \frac{9}{10} \|\nu - \mu^\star\|_2$.

Intuitively, if the dual SDP throws away all the bad samples,

$1 + \|\nu - \mu^\star\|_2^2 \approx \mathrm{OPT} \approx \mathbb{E}_{X \in G}[(X - \nu)^\top M (X - \nu)]$

When Dual SDP Has Good Solutions

Lemma (Good Dual Solutions $\Rightarrow$ Better ν)

Fix an approximately optimal solution M to the dual SDP with parameter ν. If the objective value of M is at least $1 + \Omega(\beta^2)$, then we can find $\nu' \in \mathbb{R}^d$ such that $\|\nu' - \mu^\star\|_2 \le \frac{9}{10}\|\nu - \mu^\star\|_2$.

Intuitively, if the dual SDP throws away all the bad samples,
$1 + \|\nu - \mu^\star\|_2^2 \approx \mathrm{OPT} \approx \mathbb{E}_{X \in G}[(X-\nu)^\top M (X-\nu)] = \langle M, I + (\nu-\mu^\star)(\nu-\mu^\star)^\top \rangle$.

When Dual SDP Has Good Solutions

Lemma (Good Dual Solutions ⇒ Better ν)

Fix an approximately optimal solution M to the dual SDP with parameter ν. If the objective value of M is at least $1 + \Omega(\beta^2)$, then we can find $\nu' \in \mathbb{R}^d$ such that $\|\nu' - \mu^\star\|_2 \le \frac{9}{10} \|\nu - \mu^\star\|_2$.

Intuitively, if the dual SDP throws away all the bad samples,
$1 + \|\nu - \mu^\star\|_2^2 \approx \mathrm{OPT} \approx \mathbb{E}_{X \in G}\left[(X-\nu)^\top M (X-\nu)\right] = \langle M, I + (\nu-\mu^\star)(\nu-\mu^\star)^\top \rangle$.
Because $\mathrm{tr}(M) = 1$, the top eigenvector of M aligns approx. with $(\nu - \mu^\star)$.

When Dual SDP Has Good Solutions

Lemma (Good Dual Solutions $\Rightarrow$ Better ν)

Fix an approximately optimal solution M to the dual SDP with parameter ν. If the objective value of M is at least $1 + \Omega(\beta^2)$, then we can find $\nu' \in \mathbb{R}^d$ such that $\|\nu' - \mu^\star\|_2 \le \frac{9}{10} \|\nu - \mu^\star\|_2$.

Intuitively, if the dual SDP throws away all the bad samples,
$1 + \|\nu - \mu^\star\|_2^2 \approx \mathrm{OPT} \approx \mathbb{E}_{X\in G}\left[(X-\nu)^\top M (X-\nu)\right] = \langle M, I + (\nu-\mu^\star)(\nu-\mu^\star)^\top \rangle$.
Because $\mathrm{tr}(M) = 1$, the top eigenvector of M aligns approx. with $(\nu - \mu^\star)$.

We will move ν closer to ν':

When Dual SDP Has Good Solutions

Lemma (Good Dual Solutions $\Rightarrow$ Better ν)

Fix an approximately optimal solution M to the dual SDP with parameter ν. If the objective value of M is at least $1 + \Omega(\beta^2)$, then we can find $\nu' \in \mathbb{R}^d$ such that $\|\nu' - \mu^\star\|_2 \le \frac{9}{10}\|\nu - \mu^\star\|_2$.

Intuitively, if the dual SDP throws away all the bad samples,
$1 + \|\nu - \mu^\star\|_2^2 \approx \mathrm{OPT} \approx \mathbb{E}_{X \in G}\left[(X-\nu)^\top M (X-\nu)\right] = \langle M, I + (\nu-\mu^\star)(\nu-\mu^\star)^\top \rangle$.
Because $\mathrm{tr}(M) = 1$, the top eigenvector of M aligns approx. with $(\nu - \mu^\star)$.

We will move ν closer to ν':

- The top eigenvector of M tells us which direction ν should move.

When Dual SDP Has Good Solutions

Lemma (Good Dual Solutions $\Rightarrow$ Better ν)

Fix an approximately optimal solution M to the dual SDP with parameter ν. If the objective value of M is at least $1 + \Omega(\beta^2)$, then we can find $\nu' \in \mathbb{R}^d$ such that $\|\nu' - \mu^\star\|_2 \le \frac{9}{10}\|\nu - \mu^\star\|_2$.

Intuitively, if the dual SDP throws away all the bad samples,
$1 + \|\nu - \mu^\star\|_2^2 \approx \mathrm{OPT} \approx \mathbb{E}_{X\in G}[(X-\nu)^\top M(X-\nu)] = \langle M, I + (\nu-\mu^\star)(\nu-\mu^\star)^\top\rangle.$
Because $\mathrm{tr}(M) = 1$, the top eigenvector of M aligns approx. with $(\nu - \mu^\star)$.

We will move ν closer to ν':

- The top eigenvector of M tells us which direction ν should move.
- The objective value OPT_ν tells us how far ν should move.

When Dual SDP Has Good Solutions

The lemma shows that despite the error from

When Dual SDP Has Good Solutions

The lemma shows that despite the error from

- the errors in the concentration bounds, and
- we are only solving the SDP approximately,

When Dual SDP Has Good Solutions

The lemma shows that despite the error from

- the errors in the concentration bounds, and
- we are only solving the SDP approximately,

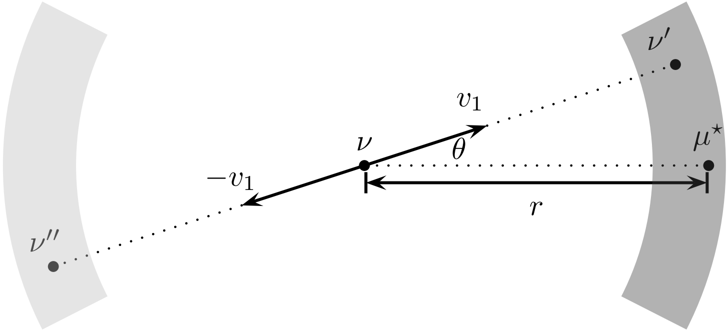
the top eigenvector of M still aligns approximately with $(\nu - \mu^\star)$.

When Dual SDP Has Good Solutions

The lemma shows that despite the error from

- the errors in the concentration bounds, and
- we are only solving the SDP approximately,

the top eigenvector of M still aligns approximately with $(\nu - \mu^\star)$.



Solving the SDPs Approximately

Primal SDP (with parameter ν)

$$\text{minimize} \;\; \lambda_{\max}\left(\sum_{i=1}^{N} w_i (X_i - \nu)(X_i - \nu)^\top\right) \qquad \text{subject to} \;\; w \in \Delta_{N,\epsilon}$$

Solving the SDPs Approximately

Primal SDP (with parameter ν)

$$\text{minimize} \;\; \lambda_{\max}\left(\textstyle\sum_{i=1}^{N} w_i(X_i - \nu)(X_i - \nu)^\top\right) \qquad \text{subject to} \;\; w \in \Delta_{N,\epsilon}$$

Packing/covering SDPs can be solved in nearly-linear time [JY'11, ALO'16, PTZ'16].

Solving the SDPs Approximately

Primal SDP (with parameter ν)

$$\text{minimize} \;\; \lambda_{\max}\left(\textstyle\sum_{i=1}^{N} w_i (X_i - \nu)(X_i - \nu)^\top\right) \qquad \text{subject to} \;\; w \in \Delta_{N,\epsilon}$$

Packing/covering SDPs can be solved in nearly-linear time [JY'11, ALO'16, PTZ'16].

This is not a packing SDP, but we can flip the objective/constraints.

Solving the SDPs Approximately

Primal SDP (with parameter ν)

$$\text{minimize} \quad \lambda_{\max}\left(\sum_{i=1}^{N} w_i (X_i - \nu)(X_i - \nu)^\top\right) \qquad \text{subject to} \quad w \in \Delta_{N,\epsilon}$$

Packing/covering SDPs can be solved in nearly-linear time [JY'11, ALO'16, PTZ'16].

This is not a packing SDP, but we can flip the objective/constraints.

Packing SDP with parameters (ν, ρ)

$$\text{maximize} \quad \|w\|_1 \qquad \text{subject to} \quad 0 \le w_i \le \frac{1}{(1-\epsilon)N}, \sum_i w_i (\rho X_i X_i^\top) \preceq I$$

Solving the SDPs Approximately

Primal SDP (with parameter ν)

$$\text{minimize} \quad \lambda_{\max}\left(\sum_{i=1}^{N} w_i (X_i - \nu)(X_i - \nu)^\top\right) \qquad \text{subject to} \quad w \in \Delta_{N,\epsilon}$$

Packing/covering SDPs can be solved in nearly-linear time [JY'11, ALO'16, PTZ'16].

This is not a packing SDP, but we can flip the objective/constraints.

Packing SDP with parameters (ν, ρ)

$$\text{maximize} \quad \|w\|_1 \qquad \text{subject to} \quad 0 \le w_i \le \frac{1}{(1-\epsilon)N}, \sum_i w_i (\rho X_i X_i^\top) \preceq I$$

Binary search for ρ and check if $\max \|w\|_1 \ge 1$ ($\rho^\star = \frac{1}{\mathrm{OPT}_\nu}$).
Need to handle bi-criteria approximations.

Full Algorithm

Algorithm 1: Robust Mean Estimation for Known Covariance Sub-Gaussian

Let $\nu = \frac{1}{N}\sum_{i=1}^{N} X_i$ be the empirical mean;

for $i = 1$ **to** $O(\log(d \log N/\epsilon))$ **do**

  Compute either

  (i) a good solution $w \in \mathbb{R}^N$ for the primal SDP with parameters $(\nu, 2\epsilon)$; or

  (ii) a good solution $M \in \mathbb{R}^{d \times d}$ for the dual SDP with parameters (ν, ϵ);

  if *the objective value of* w *in primal SDP* $\leq 1 + 400\epsilon \ln(1/\epsilon)$ **then**

    return *the weighted empirical mean* $\widehat{\mu}_w = \sum_{i=1}^{N} w_i X_i$;

  else

    Move ν closer to $\mu^\star$ using the top eigenvector of M.

Full Algorithm

Algorithm 2: Robust Mean Estimation for Known Covariance Sub-Gaussian

Let $\nu = \frac{1}{N}\sum_{i=1}^{N} X_i$ be the empirical mean;

for $i = 1$ **to** $O(\log(d \log N/\epsilon))$ **do**

Compute either

(i) a good solution $w \in \mathbb{R}^N$ for the primal SDP with parameters $(\nu, 2\epsilon)$; or

(ii) a good solution $M \in \mathbb{R}^{d \times d}$ for the dual SDP with parameters (ν, ϵ);

if *the objective value of w in primal SDP* $\leq 1 + 400\epsilon \ln(1/\epsilon)$ **then**

return *the weighted empirical mean* $\widehat{\mu}_w = \sum_{i=1}^{N} w_i X_i$;

else

Move ν closer to $\mu^\star$ using the top eigenvector of M.

Summary

Distribution	Error (δ)	# of Samples (N)	Runtime
Sub-Gaussian	$O(\epsilon\sqrt{\log(1/\epsilon)})$	$O(d/\delta^2)$	$\widetilde{O}(Nd/\epsilon^6)$
Bounded Covariance	$O(\sqrt{\epsilon})$	$\widetilde{O}(d/\delta^2)$	

Summary

Distribution	Error (δ)	# of Samples (N)	Runtime
Sub-Gaussian	$O(\epsilon\sqrt{\log(1/\epsilon)})$	$O(d/\delta^2)$	$\widetilde{O}(Nd/\epsilon^6)$
Bounded Covariance	$O(\sqrt{\epsilon})$	$\widetilde{O}(d/\delta^2)$	

We hope our work will serve as a starting point for the design of faster algorithms for high-dimensional robust estimation.

Open Problems

- Faster algorithms for other high-dimensional robust learning problems (e.g., sparse mean estimation / sparse PCA)?

Open Problems

- Faster algorithms for other high-dimensional robust learning problems (e.g., sparse mean estimation / sparse PCA)?
- Can we avoid the $\text{poly}(1/\epsilon)$ in the runtime?

Open Problems

- Faster algorithms for other high-dimensional robust learning problems (e.g., sparse mean estimation / sparse PCA)?
- Can we avoid the $\mathrm{poly}(1/\epsilon)$ in the runtime?
- Robust covariance estimation in time $\widetilde{O}(Nd)/\epsilon^{O(1)})$?