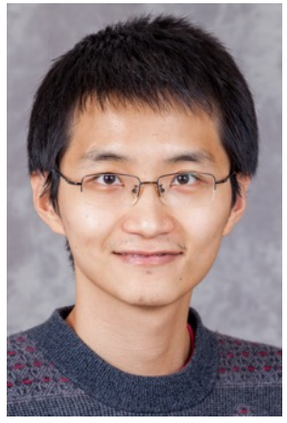
High-Dimensional Robust Mean Estimation via Gradient Descent

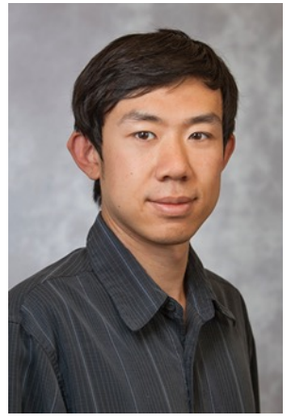


Yu Cheng

UIC

Ilias Diakonikolas

UW Madison



Rong Ge

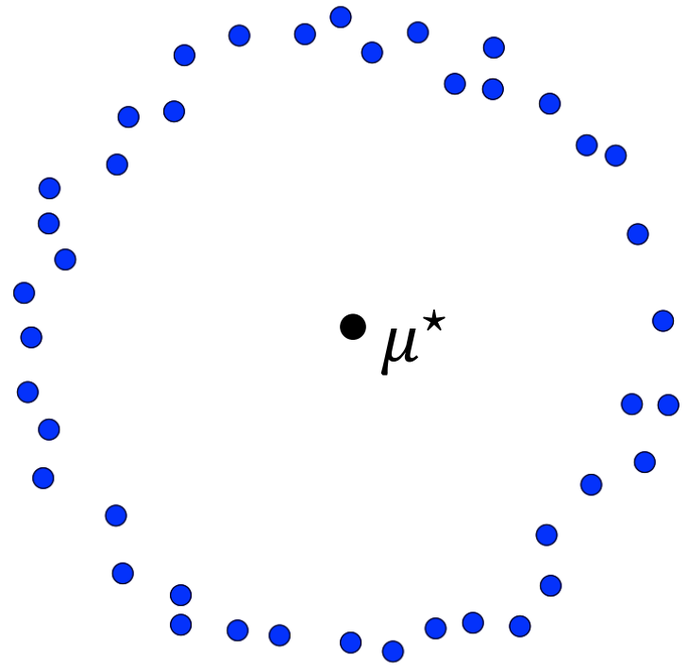
Duke

Mahdi Soltanolkotabi

USC

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

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

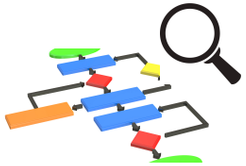
Empirical mean $\widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i$ works:

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

Robust Mean Estimation

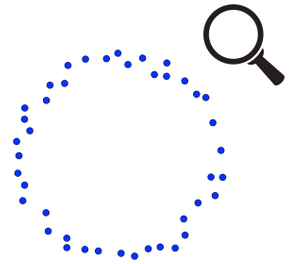
Robust Mean Estimation

ϵ-Corruption:

 specifies N.

 draws N samples from $\mathcal{N}(\mu^\star, I)$.

 replaces ϵN samples with arbitrary points.



Goal: Learn $\mu^\star$ given an ϵ-corrupted set of N samples.

Previous Work

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

Previous Work

Algorithm	Error Guarantee	Runtime
[Lai+ '16]	$O(\epsilon\sqrt{\log d})$	Polynomial
[Diakonikolas+ '16]	$O(\epsilon\sqrt{\log(1/\epsilon)})$	

Previous Work

Algorithm	Error Guarantee	Runtime
[Lai+ '16]	$O(\epsilon\sqrt{\log d})$	Polynomial
[Diakonikolas+ '16]	$O(\epsilon\sqrt{\log(1/\epsilon)})$	
[**C** Diakonikolas Ge '19]		$\tilde{O}(Nd)/\mathrm{poly}(\epsilon)$
[Dong Hopkins Li '19]		$\tilde{O}(Nd)$

These algorithms have near-optimal sample complexity.

Motivation

Existing algorithms are fairly sophisticated (e.g., ellipsoid method, iterative spectral methods, matrix multiplicative weight update) and they are not parameter free.

Is it possible to solve robust estimation tasks by standard first-order methods?

Our Results

- A natural non-convex formulation of robust mean estimation.
- Any approximate stationary point of this non-convex objective gives a good solution for mean estimation.
- Gradient descent converges to an approximate stationary point in a polynomial number of iterations.

Non-Convex Formulation

$$\mu_w = \sum_i w_i X_i \quad \text{and} \quad \Sigma_w = \sum_i w_i (X_i - \mu_w)(X_i - \mu_w)^\top$$

[Diakonikolas+ '16]:

If Σ_w has small spectral norm, then μ_w is close to the true mean.

$$\text{minimize} \quad \|\Sigma_w\|_2$$

$$\text{subject to} \quad w \in \Delta_{N,2\epsilon}$$

$$\Delta_{N,2\epsilon} = \{w \in \mathbb{R}^N : \|w\|_1 = 1 \text{ and } 0 \le w_i \le \tfrac{1}{(1-2\epsilon)N}\}$$

Our Results

$$\begin{aligned} &\text{minimize} && \|\Sigma_w\|_2 \\ &\text{subject to} && w \in \Delta_{N,2\epsilon} \end{aligned}$$

Despite its non-convexity, we can show that any (approximate) stationary point w defines a μ_w that is $O(\epsilon\sqrt{\log(1/\epsilon)})$-close to $\mu^\star$.

Our Results

$\min \|\Sigma_w\|_2 \quad \text{s.t. } w \in \Delta_{N,2\epsilon}$

$\|\Sigma_w\|_2$ may not be differentiable w.r.t. w.

- Sub-gradient: use $\frac{\partial(v^\top \Sigma_w v)}{\partial w}$ where v is any top eigenvector of Σ_w.
- Softmax: minimize $\frac{1}{\rho}\operatorname{tr}\exp(\rho\Sigma_w)$, which is differentiable.

We prove structural and algorithmic results for both approaches.

Our Algorithms

Sub-gradient

Start with any $w_0 \in \mathcal{K} = \Delta_{N,2\epsilon}$.

For $t = 0 \ldots T-1$

Let $v \in \operatorname{argmax}_{\|v\|_2=1} \; v^\top \Sigma_w v$.

$$w_{t+1} \leftarrow \mathcal{P}_{\mathcal{K}}\left(w_t - \eta \frac{\partial (v^\top \Sigma_w v)}{\partial w}\right).$$

end for

# of iterations: $T = \tilde{O}(N^2 d^4)$

Softmax

…

For …

$$w_{t+1} \leftarrow \mathcal{P}_{\mathcal{K}}\left(w_t - \eta \frac{\partial smax(\Sigma_w)}{\partial w}\right).$$

end for

$T = \tilde{O}(Nd^3/\epsilon)$

MATLAB Implementation

Projected Sub-gradient Descent

```
for itr = 1:numItr
    Sigma_w_fun = @(v) X' * (w .* (X * v)) - (X' * w)^2 * v;
    [u, lambda] = eigs(Sigma_w_fun, d, 1);
    nabla_f_w = (X * u) .* (X * u) - 2 * (w' * (X * u)) * (X * u);
    w = w - stepSize * nabla_f_w / norm(nabla_f_w);
    w = project_onto_capped_simplex(w, 1 / (N - epsN));
end
```

Our Results

- A natural non-convex formulation of robust mean estimation.
- Any approximate stationary point of our non-convex objective gives a near-optimal solution for mean estimation.
- Gradient descent converges to an approximate stationary point in a polynomial number of iterations.

Any Stationary Point Suffices

w is a bad solution.

$\Rightarrow v^\top \Sigma_w v$ is much larger than it should be.

$\Rightarrow$ We can find i and j such that

- it is feasible to increase w_i and decrease w_j.
- $v^\top \Sigma_w v$ becomes smaller after the change.

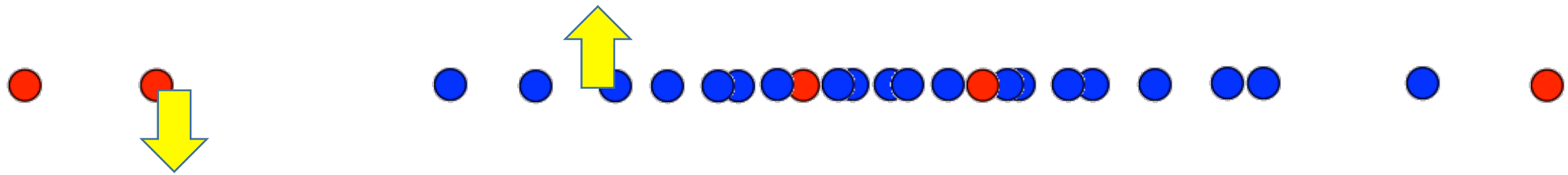
$\Rightarrow w$ is not a first-order stationary point.

Any Stationary Point Suffices

$v^\top \Sigma_w v$ = variance in the direction v.

$\frac{\partial (v^\top \Sigma_w v)}{\partial w}$ = the gradient of w for the

one-dimensional problem with input $\left(X_i^\top v\right)_{i=1}^N$.



Any Stationary Point Suffices

Simple case: $\mu_w = 0$ and Σ_w has a unique largest eigenvector v.

We have $\Sigma_w = \sum_i w_i X_i X_i^\top$ and $v^\top \Sigma_w v = \sum_i w_i y_i^2$ where $y_i = X_i^\top v$.

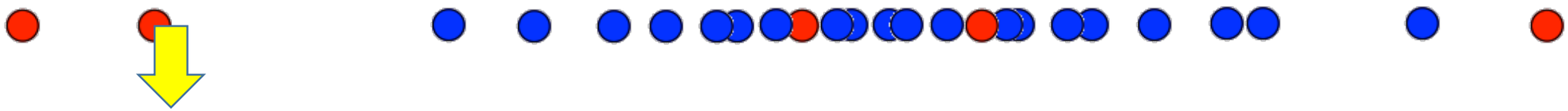
$$\frac{\partial (v^\top \Sigma_w v)}{\partial w_i} = y_i^2$$

Any Stationary Point Suffices

Simple case: $\mu_w = 0$ and Σ_w has a unique largest eigenvector v.

We have $\Sigma_w = \sum_i w_i X_i X_i^\top$ and $v^\top \Sigma_w v = \sum_i w_i y_i^2$ where $y_i = X_i^\top v$.

$$\frac{\partial (v^\top \Sigma_w v)}{\partial w_i} = y_i^2$$



$\sum_{i \in \text{bad}} w_i y_i^2$ is very large $\implies$ $\exists i$ s.t. $w_i > 0$ and y_i^2 is large.

Algorithmic Result

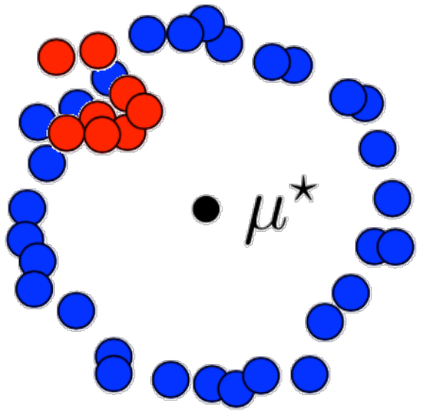
Challenges in bounding # of iterations:

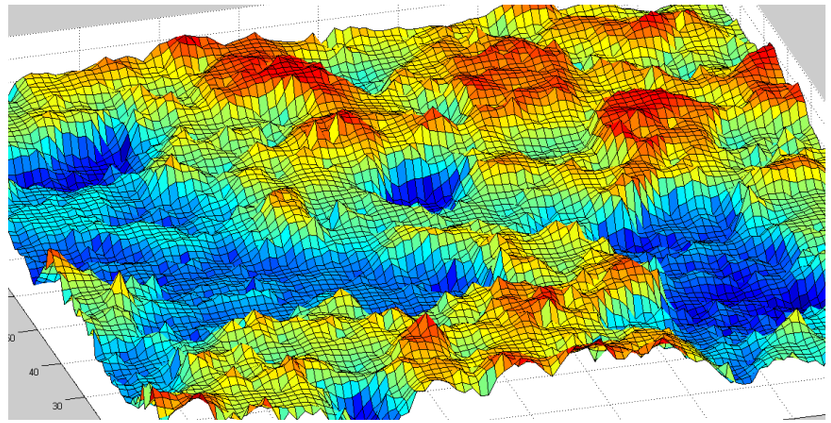
- $w_{t+1} \leftarrow \mathcal{P}_{\mathcal{K}}\left(w_t - \eta \frac{\partial(v^\top \Sigma_w v)}{\partial w}\right)$ can make $\|\Sigma_w\|_2$ larger.

- Non-convex + Non-smooth + Constraints

Our Contributions

High-Dimensional Robust Statistics + **Non-Convex Optimization**





Open Problems

- Faster convergence rate.

- More general robust estimation tasks:
  - Covariance estimation.
  - Sparse PCA.
  - Robust regression.