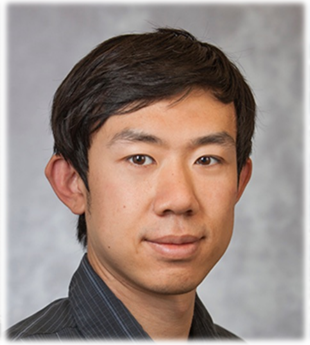
Bridging High-Dimensional Robust Statistics and Non-Convex Optimization

Yu Cheng (Brown University)

@ TTIC, August 2024

Based on joint works with:

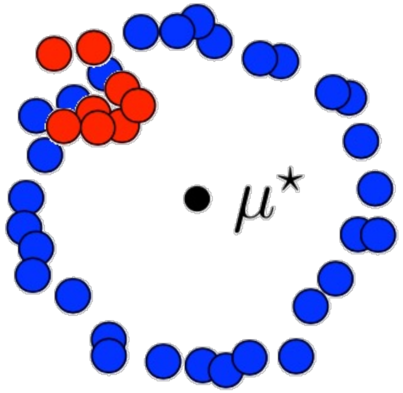




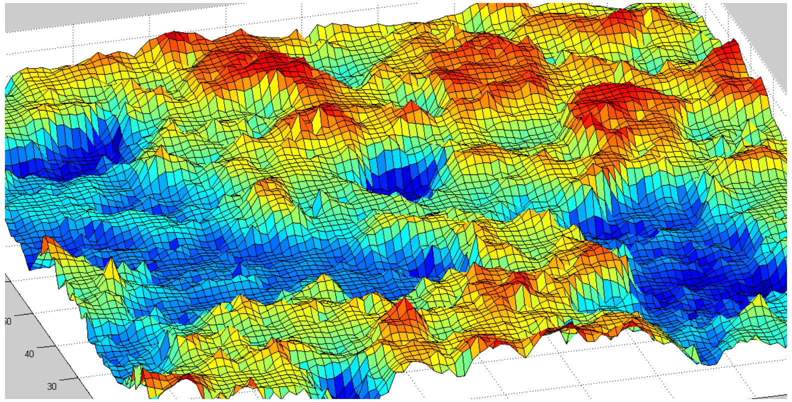



A Tale of Two Research Areas

High-Dimensional
Robust Statistics

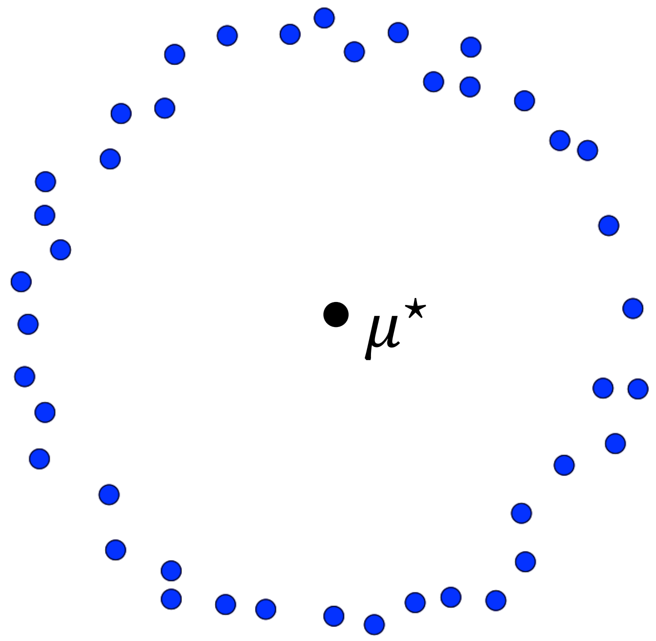


Non-Convex
Optimization

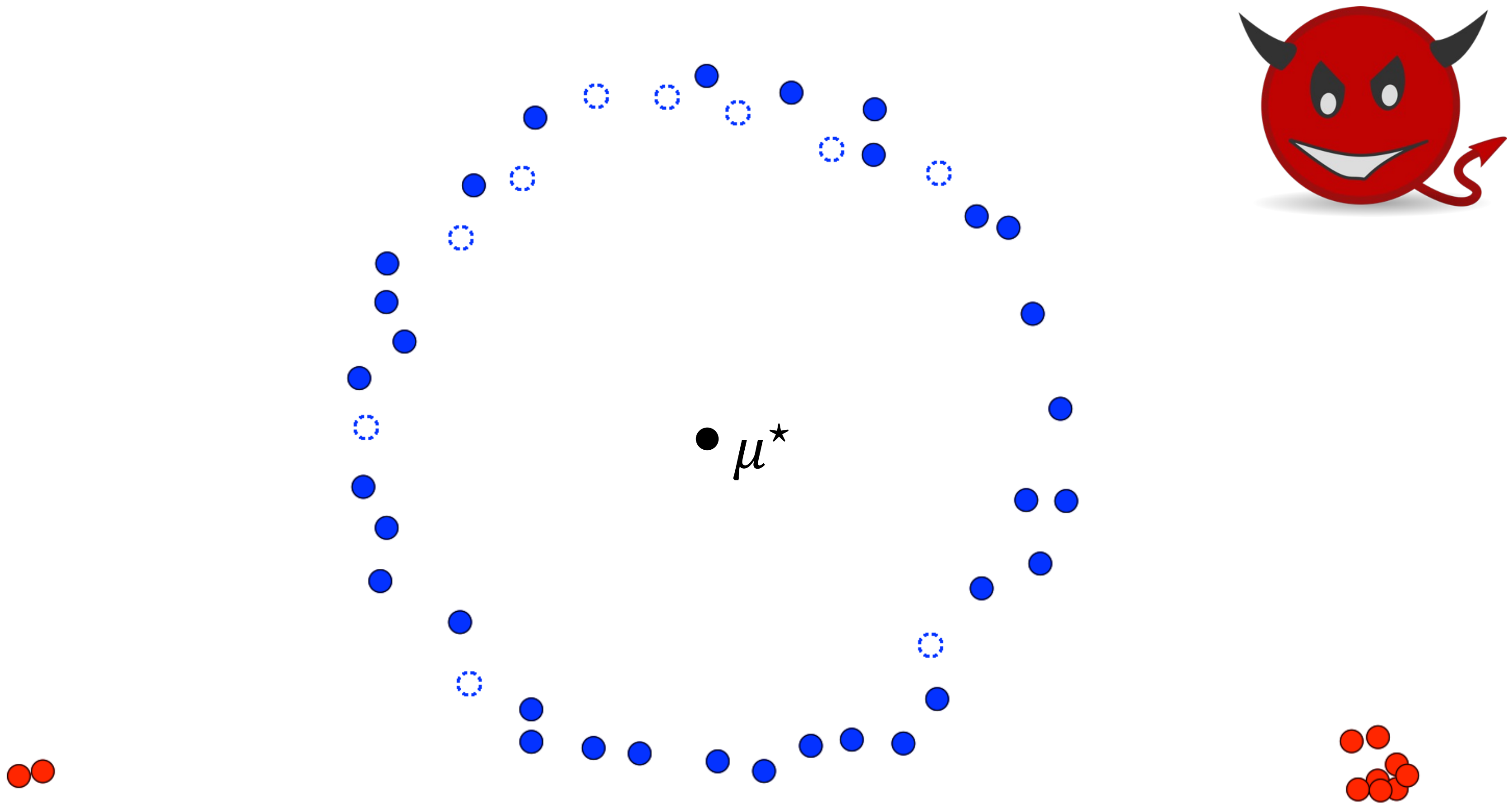


Mean Estimation

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

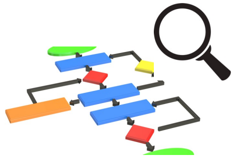


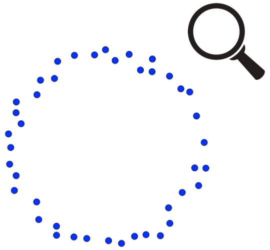
Robust Mean Estimation



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.

Robust Mean Estimation: Prior 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

Robust Mean Estimation: Prior Work

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

These algorithms have near-optimal sample complexity.

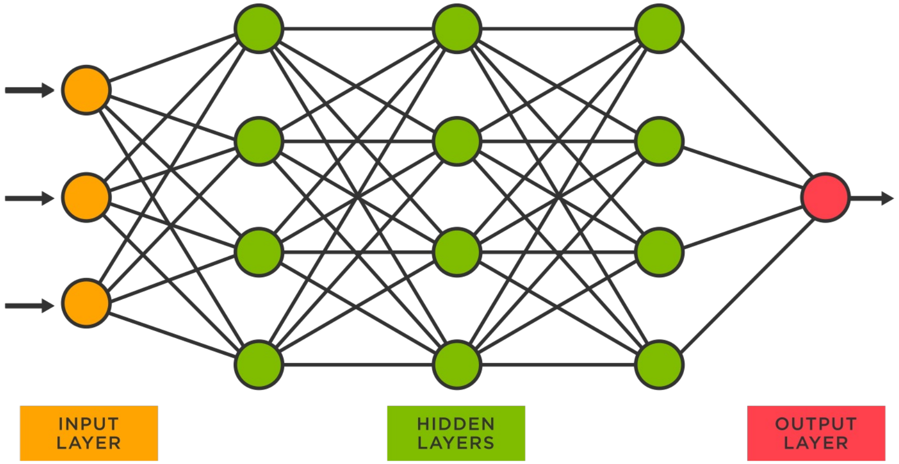
Motivation #1

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?

Non-Convex Optimization

Extremely successful in practice.



Non-Convex Optimization

Extremely successful in practice.

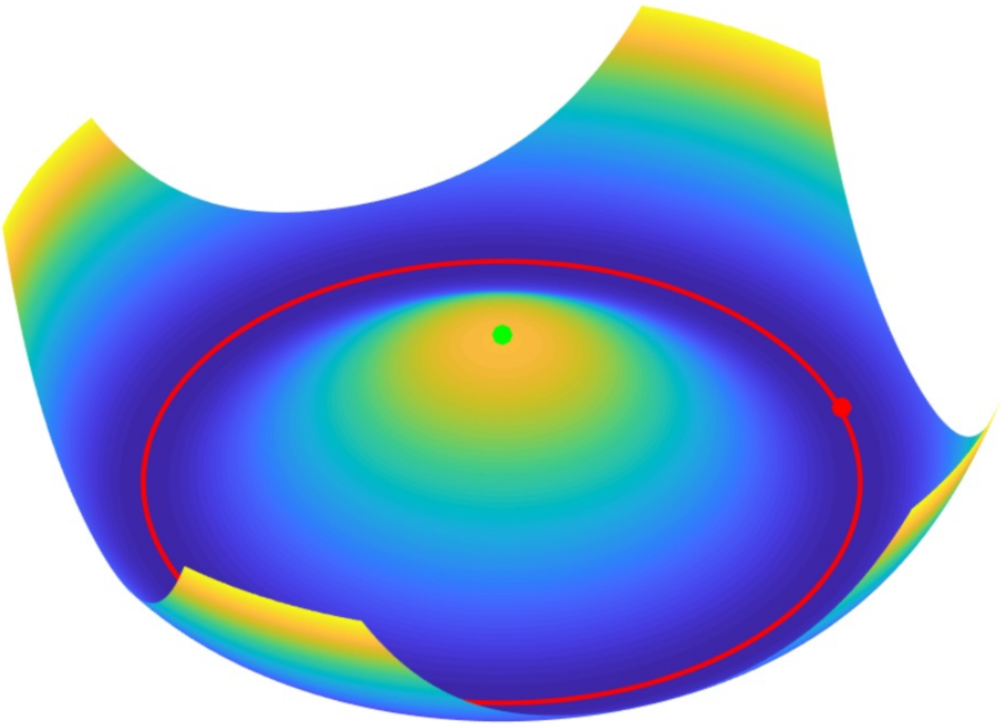
- In theory: NP-Hard.
- In practice: can be solved via (stochastic) gradient descent.

Why does non-convex optimization work?

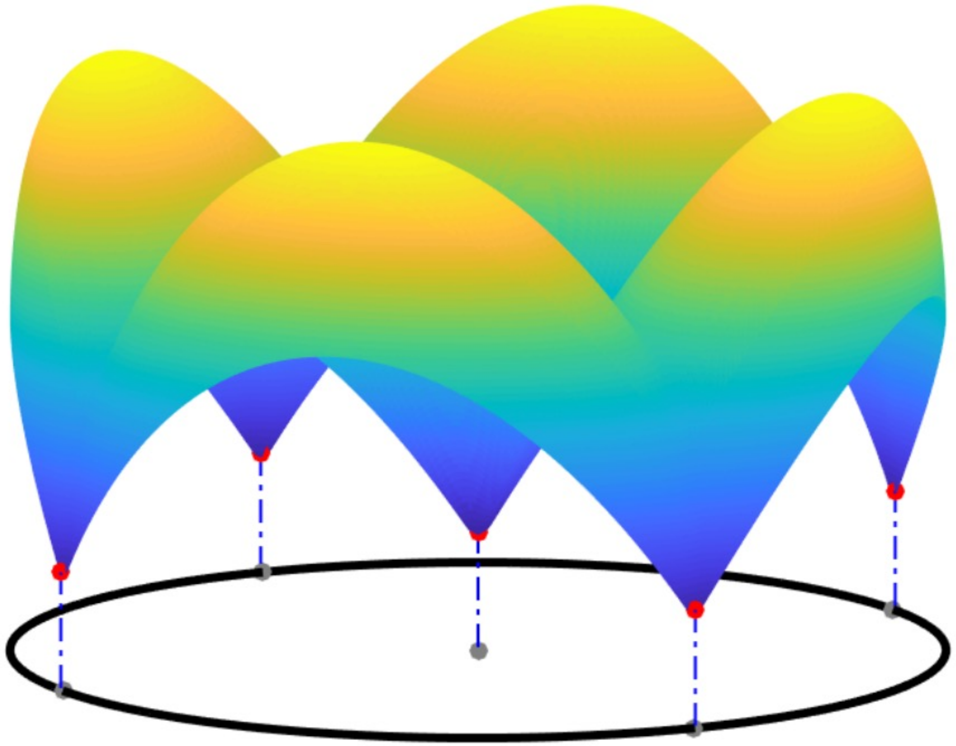
- One possible explanation:

All local optima are globally optimal!

Non-Convex Optimization



Rotational symmetry



Discrete symmetry

Non-Convex Optimization

All local optima are globally optimal!

- Matrix factorization / Matrix completion.
- Matrix sensing / Phase retrieval.
- Eigenvector computation.
- Tensor decomposition.
- Dictionary learning.
- Training neural networks.
- …

Motivation #1, Revisited

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

Are all local optima globally optimal for natural non-convex formulations of robust estimation tasks?

Robust Gradient Descent

Robust meta-algorithms for stochastic optimization [Diakonikolas+ '19][Prasad+ '20].

- Unknown true distribution $\mathcal{D}$ of labelled data (X, Y).
- Input: $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ where ϵ-fraction is arbitrarily corrupted.
- Goal: $\min \quad \bar{L}(\theta) \coloneqq \mathbb{E}_{(X,Y)\sim\mathcal{D}}[L(\theta, X, Y)]$.

Example: Robust linear regression, $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$

$$\min \quad \Sigma_{i=1}^{n} L_i(\theta) = \Sigma_{i=1}^{n} (\theta^\top X_i - Y_i)^2 \quad \text{under } \epsilon\text{-corruption.}$$

Robust Gradient Descent

Robust meta-algorithms for stochastic optimization [Diakonikolas+ '19][Prasad+ '20].

Goal: $\min \ \bar{L}(\theta) \coloneqq \mathbb{E}_{(X,Y)\sim\mathcal{D}}[L(\theta, X, Y)]$.

Input: $\min \ \Sigma_{i=1}^{n} L_i(\theta)$, ϵ-fraction of the L_i is corrupted.

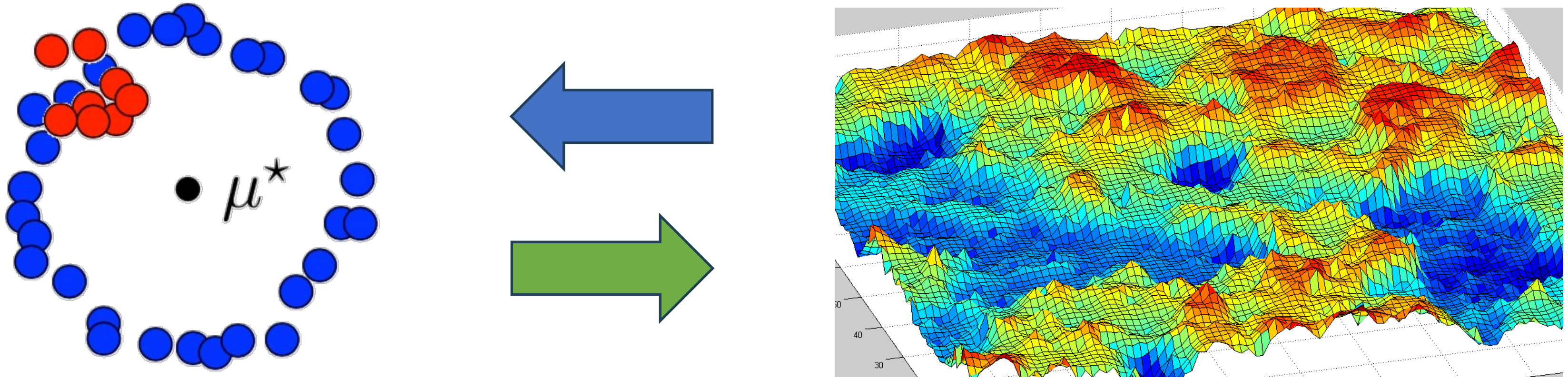
Key idea:

- The gradients $\left(\nabla L_i(\theta)\right)_i$ is an ϵ-corrupted set of vectors with true mean $\nabla\bar{L}(\theta)$.
- Can robustly estimate the true gradient $\nabla\bar{L}(\theta)$.
- Can converge to a (local) optima of $\bar{L}(\theta)$ despite ϵ-corruption.

Motivation #2

Can we design provably robust algorithms for tractable non-convex problems?

A Tale of Two Research Areas



- New algorithms for robust statistics via optimization
- New robust algorithms for tractable non-convex problems.

Outline

- Robust Mean Estimation via Gradient Descent
- Robust Sparse Estimation via Gradient Descent
- Robust Second-Order Nonconvex Optimization

Motivating Question

Can we solve **robust mean estimation** using standard **first-order methods?**

Our Results [CDGS '20]

- A natural non-convex formulation of robust mean estimation.
- Any approximate stationary point of this 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.

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.

$$\min \quad \|\Sigma_w - I\|_2 \qquad \text{s.t.} \quad w \in \Delta_{n,\epsilon}$$

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

Our Results [CDGS '20]

- A natural non-convex formulation of robust mean estimation.
- Any approximate stationary point of this 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.

Our Results [CDGS '20]

$$\min \quad \|\Sigma_w - I\|_2 \qquad \text{s.t.} \quad w \in \Delta_{n,\epsilon}$$

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

No Bad Local Optima [CDGGKS'22]

$$\min \quad f(w) = \|\Sigma_w - I\|_2$$

Let $w^\star$ = uniform weight on the remaining good samples.

We prove that for any w with $f(w) \gg \epsilon$, moving w toward $w^\star$ decreases the value of f.

No Bad Local Optima [CDGGKS'22]

$$\min \quad f(w) = \|\Sigma_w - I\|_2$$

Formally, for any $0 < \eta < 1$,

$$\Sigma_{(1-\eta)w+\eta w^\star} = (1-\eta)\Sigma_w + \eta\Sigma_{w^\star} + \eta(1-\eta)(\mu_w - \mu_{w^\star})(\mu_w - \mu_{w^\star})^\top$$

We show that the third term can essentially be ignored, so

$$f((1-\eta)w + \eta w^\star) \lessapprox (1-\eta)f(w) + \eta f(w^\star) < f(w)$$

No Bad Local Optima [CDGGKS'22]

$$\Sigma_{(1-\eta)w+\eta w^\star} = (1-\eta)\Sigma_w + \eta\Sigma_{w^\star} + \eta(1-\eta)(\mu_w - \mu_{w^\star})(\mu_w - \mu_{w^\star})^\top$$

- Upper bounding the third term:

$$\textit{for any } w \in \Delta_{n,\epsilon}, \textit{ we have}$$

$$\|\mu_w - \mu_{w_\star}\|_2^2 \leq 4\epsilon\left(\|\Sigma_w - I\|_2 + O(\frac{\delta^2}{\epsilon})\right)$$

- Proof similar to structural lemma for robust mean estimation.

Another Proof [CDGS ’20]

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

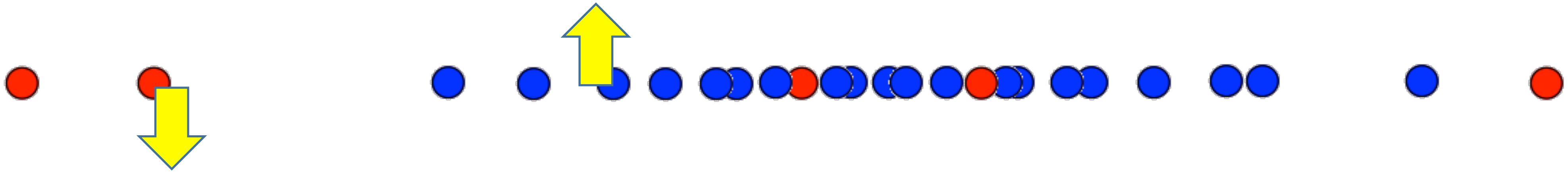
- it is feasible to increase w_i and decrease w_j.
- $v^\top \Sigma_w v$ becomes smaller after the change.

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

Another Proof [CDGS '20]

$$v^\top \Sigma_w v \;=\; \text{variance in the direction } v.$$

$$\frac{\partial (v^\top \Sigma_w v)}{\partial w} \;=\; \text{the gradient of } w \text{ for the } \textcolor{red}{\text{1-D problem}} \text{ with input } \left(X_i^\top v\right)_{i=1}^{n}.$$

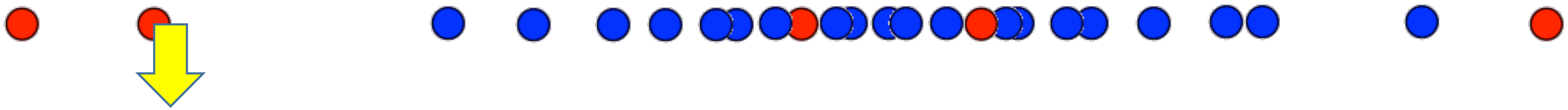


Another Proof [CDGS '20]

Simple case: $\mu_w = 0$ and Σ_w has a unique top 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.

Our Results [CDGS ’20]

- A natural non-convex formulation of robust mean estimation.
- Any approximate stationary point of this 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.

Algorithmic Results

$\min \|\Sigma_w\|_2$ s.t. $w \in \Delta_{n,\epsilon}$

$\|\Sigma_w\|_2$ 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.

Algorithmic Results

Sub-gradient

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

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

Softmax

…

For …

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

end for

Implementation

Projected Sub-gradient Descent

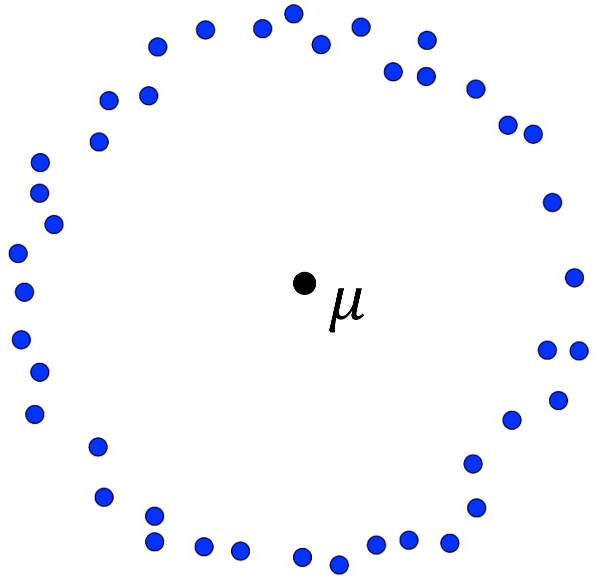
```
for itr = 1:numItr
    Sigma_w_fun = @(v) X' * (w .* (X * v)) - (X' * w)^2 * v;
    [u, lambda] = eigs(Sigma_w_fun, d, 1);
    nabla_f_w = (X * u) .* (X * u) - 2 * (w' * (X * u)) * (X * u);
    w = w - stepSize * nabla_f_w / norm(nabla_f_w);
    w = project_onto_capped_simplex(w, 1 / (N - epsN));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

Outline

- Robust Mean Estimation via Gradient Descent
- Robust Sparse Estimation via Gradient Descent
- Robust Second-Order Nonconvex Optimization

Sparse Mean Estimation

- Input: n samples $\{X_1, \ldots, X_n\}$ drawn from $\mathcal{N}(\mu, I)$ where $\mu \in \mathbb{R}^d$ is unknown and k-sparse.
- Goal: Learn μ.



Without sparsity: $n \approx O(d)$.

With sparsity: $n \approx O(k^2 \log d)$.

Robust Sparse Mean and Sparse PCA

Robust sparse mean estimation:

- Input: An ϵ-corrupted set of n samples drawn from $\mathcal{N}(\mu, I)$
  where $\mu \in \mathbb{R}^d$ is unknown and k-sparse.
- Goal: Learn μ.

Robust sparse PCA (with spiked covariance):

- Input: An ϵ-corrupted set of n samples drawn from $\mathcal{N}(0, I + vv^\top)$
  where $v \in \mathbb{R}^d$ is unknown and k-sparse.
- Goal: Learn v.

Motivating Question

Can we solve **robust sparse estimation** tasks using standard **first-order methods?**

Our Results [CDGGKS'22]

- We design new optimization formulations for robust sparse mean estimation and robust sparse PCA.
- We show that any (approximate first-order) stationary point provides a good solution for robust sparse estimation.
- Our algorithms work for a wider family of distributions.

Our Non-Convex Formulations [CDGGKS'22]

$$\min \quad f(w) = \|\Sigma_w - I\|_{F,k,k}$$

μ_w and Σ_w are the weighted empirical mean and covariance matrix.

$\|A\|_{F,k,k}$ is the maximum Frobenius norm of any k^2 entries of A, where these entries are chosen from k rows with k entries in each row.

We prove that f has no bad first-order stationary points!

Intuition for Choosing $f(w) = \|\Sigma_w - I\|_{F,k,k}$

Structural result from [BDLS'17]: If the variance in all **sparse** directions is close to 1, then the empirical mean is close to the true mean.

Our choice of f satisfies:

- $f(w) \geq v^\top(\Sigma_w - I)v$ for all k-sparse unit vector v.
  - $v^\top \Sigma_w v$ is the sample variance in direction v (weighted by w).
- We show that $f(w) \leq \tilde{O}(\epsilon)$ if w puts weight only on good samples.

These conditions imply the global optimum of f works.

We prove any local optimum of f suffices!

No Bad Local Optima (w/o Sparsity)

$$\min \quad f(w) = \|\Sigma_w - I\|_2$$

Formally, for any $0 < \eta < 1$,

$$\Sigma_{(1-\eta)w+\eta w^\star} = (1-\eta)\Sigma_w + \eta\Sigma_{w^\star} + \eta(1-\eta)(\mu_w - \mu_{w^\star})(\mu_w - \mu_{w^\star})^\top$$

The third term can essentially be ignored: $\|\mu_w - \mu_{w_\star}\|_2^2 \le 4\epsilon\left(\|\Sigma_w - I\|_2 + O(\frac{\delta^2}{\epsilon})\right)$

so

$$f((1-\eta)w + \eta w^\star) \lessapprox (1-\eta)f(w) + \eta f(w^\star) < f(w)$$

No Bad Local Optima (w/ Sparsity)

$$\min \quad f(w) = \|\Sigma_w - I\|_{F,k,k}$$

Formally, for any $0 < \eta < 1$,

$$\Sigma_{(1-\eta)w+\eta w^\star} = (1-\eta)\Sigma_w + \eta\Sigma_{w^\star} + \eta(1-\eta)(\mu_w - \mu_{w^\star})(\mu_w - \mu_{w^\star})^\top$$

The third term can essentially be ignored: $\left\|(\mu_w - \mu_{w_\star})(\mu_w - \mu_{w_\star})^\top\right\|_{F,k,k} \leq 4\epsilon\left(\|\Sigma_w - I\|_{F,k,k} + O(\delta^2/\epsilon)\right)$

so

$$f((1-\eta)w + \eta w^\star) \lessapprox (1-\eta)f(w) + \eta f(w^\star) < f(w)$$

Outline

- Robust Mean Estimation via Gradient Descent
- Robust Sparse Estimation via Gradient Descent
- Robust Second-Order Nonconvex Optimization

Previous Work

Robust meta-algorithms for stochastic optimization [Diakonikolas+ '19][Prasad+ '20].

Goal: $\min \; \bar{L}(\theta) \coloneqq \mathbb{E}_{(X,Y)\sim\mathcal{D}}[L(\theta, X, Y)]$.

Input: $\min \; \Sigma_{i=1}^{n} L_i(\theta)$, ϵ-fraction of the L_i is corrupted.

Key idea:

- The gradients $\left(\nabla L_i(\theta)\right)_i$ is an ϵ-corrupted set of vectors with true mean $\nabla \bar{L}(\theta)$.
- Can robustly estimate the true gradient $\nabla \bar{L}(\theta)$.
- Can converge to a (local) optima of $\bar{L}(\theta)$ despite ϵ-corruption.

Motivating Question

- Prior works can robustly find First-Order Stationary Points (FOSP).

- In many tractable non-convex problems, FOSPs may be bad solutions, but Second-Order Stationary Points (SOSPs) are guaranteed to be globally optimal.

Motivating Question

Can we develop a general framework for finding **second-order stationary points** in robust stochastic optimization?

Our Results [LCDDGW'23]

- We can robustly find SOSPs despite ϵ-corruption.
  - Robustly estimate the Hessian matrix
  - Require $\tilde{O}(d^2)$ samples.

- As an application, we apply our framework to low-rank matrix sensing, developing provably robust algorithms that can tolerate corruptions in both the sensing matrices and the measurements.

Our Results [LCDDGW'23]

$g_k = \text{RobustMeanEstimation}(\{\nabla f_i(x_k)\})$ such that $\|g_k - \nabla \bar{f}(x_k)\| \le \epsilon_g/3$
$H_k = \text{RobustMeanEstimation}(\{\nabla^2 f_i(x_k)\})$ such that $\left\|H_k - \nabla^2 \bar{f}(x_k)\right\|_{\text{op}} \le \epsilon_H/9$

Algorithm 1: [LW23]

Input: $\epsilon_g = O(\sigma_g\sqrt{\epsilon})$, $\epsilon_H = O(\sigma_H\sqrt{\epsilon})$, Initialization x_0, Lipschitzness constants
Output: $(2\epsilon_g, 2\epsilon_H)$-approximate SOSP
Runtime: $O(1/\epsilon_g^2, 1/\epsilon_H^3)$ iterations in expectation
for $k = 1, 2, \ldots$ **do**
  if $\|g_k\| > \epsilon_g$ **then**
    $x_{k+1} = x_k - \frac{1}{L_g} g_k$; // gradient step
  else if $\hat{\lambda}_k := \lambda_{\min}(H_k) < -\epsilon_H$ **then**
    $\hat{p}_k \leftarrow$ unit minimum eigenvector of H_k
    Draw $\sigma_k \leftarrow \pm 1$ with probability $\frac{1}{2}$
    $x_{k+1} = x_k + \frac{2\epsilon_H}{L_H}\sigma_k \hat{p}_k$; // negative curvature step
  else
    return x_k

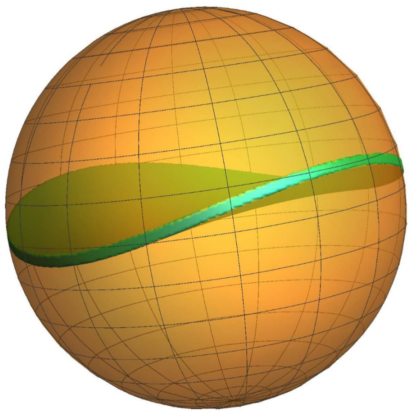
Open Problems

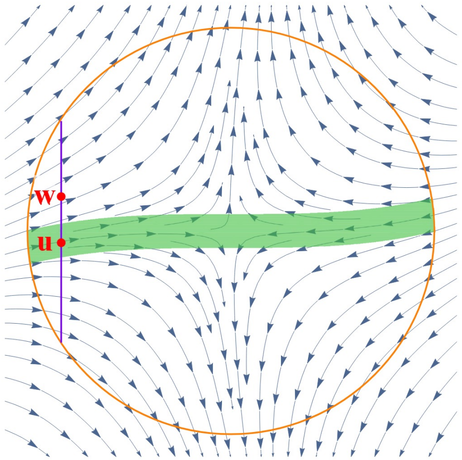
- Other robust estimation tasks via optimization
  - Covariance estimation.
  - …
- Robust mean estimation via first-order optimization in nearly-linear time?
- Can we compute the gradient of (a smoothed version of) $f(w) = \|\Sigma_w - I\|_{F,k,k}$ without writing down Σ_w explicitly?
  - Writing down Σ_w takes $d^2 \gg nd$ time.

Open Problems

- Provably robust algorithms for other tractable non-convex problems using tools in robust statistics.

- Robustly finding SOSP without robust Hessian estimation?





Thank You!
Q&A