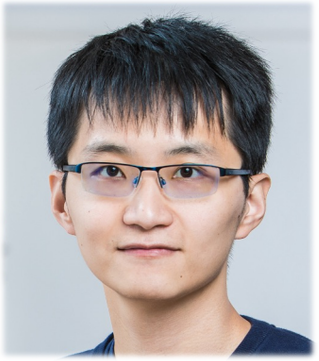
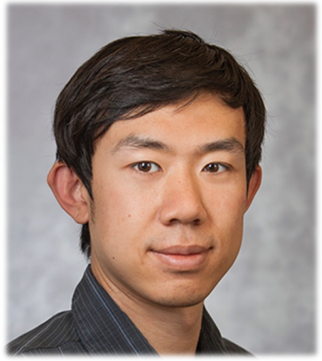
Outlier-Robust Sparse Estimation via Non-Convex Optimization



Yu Cheng
Brown

Ilias Diakonikolas
UW Madison



Rong Ge
Duke

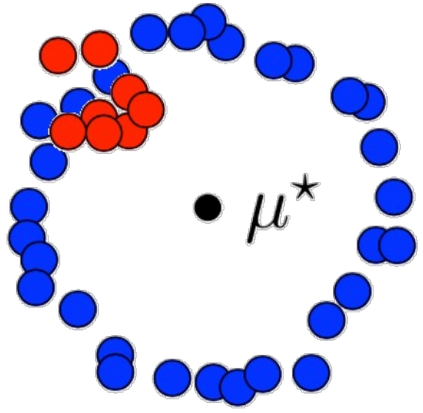
Shivam Gupta
UT Austin

Daniel Kane
UCSD

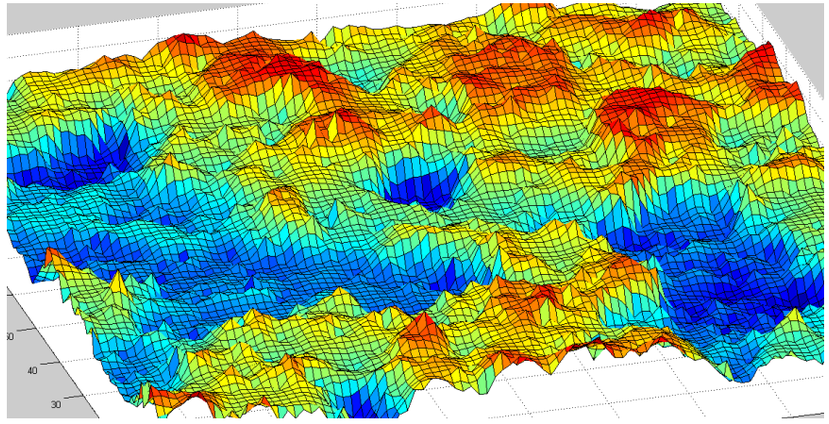
Mahdi Soltanolkotabi
USC

A Tale of Two Research Areas

High-Dimensional Robust Statistics



Non-Convex Optimization



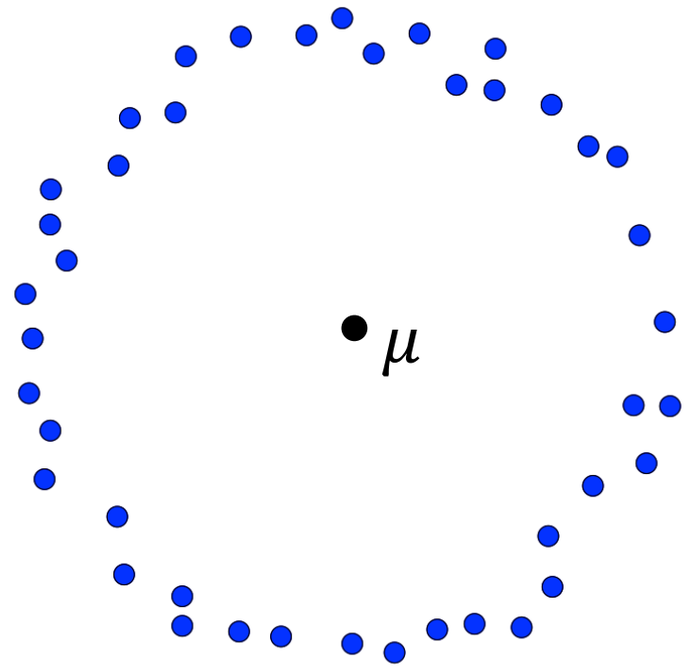
Q: What robust estimation tasks can be solved by gradient descent?

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 Estimation

ϵ-corruption model:

The adversary can inspect
- the true distribution,
- the good samples, and
- the algorithm,

and then replaces ϵ-fraction of the samples with arbitrary points.

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.

Motivation

[BDLS'17][DKKPS'19]: Poly-time algorithms for robust sparse estimation. However, these algorithms are fairly sophisticated (e.g., ellipsoid method) or are not parameter free (e.g., iterative filtering).

[CDGS'20][ZJS'22]: Robust mean estimation via gradient descent.

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

Our Results

- 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.
  - Corollary: gradient descent can solve these problems.
- Our algorithms work for a wider family of distributions.

Our Non-Convex Formulations

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

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 can compute $\nabla f(w)$ using basic matrix-vector operations.

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 two conditions imply the global optimum of f works.

We prove a stronger result: any local optimum of f suffices!

Our Landscape Results

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

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

Theorem (this paper):

Let $\delta = \epsilon\sqrt{\log(1/\epsilon)}$ and let $\gamma = O(n^{1/2}\delta^2\epsilon^{-3/2})$. For any γ-stationary point w of f, we have $\|\mathrm{top_k}(\mu_w) - \mu\|_2 \leq O(\delta)$.

Proof Sketch

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

Let $w^\star$ = putting 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. 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

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

Future Directions

- Other robust estimation tasks solvable by gradient descent? More practical robust estimation algorithms?
- Robust sparse mean estimation in $\tilde{O}(nd)$ time?
  - Input size $= nd$: $n \approx O(k^2 \log d)$ samples of d dimensions.
- 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.