High Dimensional Robust M-Estimation: Arbitrary Corruption and Heavy Tails

Liu Liu

The University of Texas at Austin

July, 2021

Table of Contents

1. Introduction and Motivation
   - *M*-estimation in high dimensions
   - Robust statistics models
2. High dimensional robust *M*-estimation
   - Problem formulation
   - Robust Descent Condition
   - Main results
3. Low rank matrix regression under heavy tails
   - Problem formulation
   - Main results

Table of Contents

1. Introduction and Motivation
   - *M*-estimation in high dimensions
   - Robust statistics models
2. High dimensional robust *M*-estimation
   - Problem formulation
   - Robust Descent Condition
   - Main results
3. Low rank matrix regression under heavy tails
   - Problem formulation
   - Main results

Introduction: background of the dissertation

- Large-scale statistical problems: both the dimension d and the sample size n may be large (possibly $n \ll d$).
- Low dimensional structures in the high dimensional setting.

Introduction: background of the dissertation

- Large-scale statistical problems: both the dimension d and the sample size n may be large (possibly $n \ll d$).
- Low dimensional structures in the high dimensional setting.
- Many examples of this:
  - Sparse regression.
  - Compressed Sensing of low rank matrices.
  - Low rank matrix completion.
  - Low rank + sparse matrix decomposition.
  - etc...

M-estimation in high dimensions

Suppose we observe n i.i.d. samples: $\{\boldsymbol{z}_i\}_{i=1}^n$.

M-estimation with constraint

$$\widehat{\boldsymbol{\beta}} = \arg\min \underbrace{\sum_{i=1}^{n} \ell_i(\boldsymbol{\beta}; \boldsymbol{z}_i)}_{\text{empirical risk}}, \quad \text{subject to} \quad \underbrace{\boldsymbol{\beta} \in \mathcal{C}.}_{\text{low dimensional structure}}$$

M-estimation in high dimensions

Suppose we observe n i.i.d. samples: $\{\boldsymbol{z}_i\}_{i=1}^n$.

M-estimation with constraint

$$\widehat{\beta} = \arg\min \underbrace{\sum_{i=1}^{n} \ell_i(\beta; \boldsymbol{z}_i)}_{\text{empirical risk}}, \quad \text{subject to} \quad \underbrace{\beta \in \mathcal{C}.}_{\substack{\text{low dimensional}\\ \text{structure}}}$$

In regression, $\boldsymbol{z}_i = (y_i, \boldsymbol{x}_i) \in \mathbb{R} \times \mathbb{R}^d$,

Lasso as an example

$$\widehat{\beta} = \arg\min \underbrace{\sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^\top \beta)^2}_{\text{empirical risk}}, \quad \text{subject to} \quad \underbrace{\|\beta\|_1 \leq R.}_{\substack{\ell_1 \text{ norm}\\ \text{enforces sparsity}}}$$

Sufficient conditions for sparse regression

ℓ_1 relaxation

- Computationally tractable compared to ℓ_0 optimization.
- Minimax optimal under restrictive conditions.

- Computationally tractable approaches (e.g., ℓ_1 minimization, Iterative Hard Thresholding) rely on restrictive conditions:
  - Restricted isometry (Candes & Tao '05).
  - Restricted eigenvalue (Bickel, Ritov & Tsybakov '08).
  - Restricted strong convexity (Negahban et al. '12).

Sufficient conditions for sparse regression

ℓ_1 relaxation

- Computationally tractable compared to ℓ_0 optimization.
- Minimax optimal under restrictive conditions.

- Computationally tractable approaches (e.g., ℓ_1 minimization, Iterative Hard Thresholding) rely on restrictive conditions:
  - Restricted isometry (Candes & Tao '05).
  - Restricted eigenvalue (Bickel, Ritov & Tsybakov '08).
  - Restricted strong convexity (Negahban et al. '12).

- Certifying these conditions is NP-hard.

- Instead, we impose strong assumptions on the probabilistic models of the data, such as sub-Gaussianity.

Table of Contents

1. Introduction and Motivation
   - *M*-estimation in high dimensions
   - Robust statistics models
2. High dimensional robust *M*-estimation
   - Problem formulation
   - Robust Descent Condition
   - Main results
3. Low rank matrix regression under heavy tails
   - Problem formulation
   - Main results

Contamination model

[G. Box] "All models are wrong, but some are useful."

What if the real data violate the assumptions required: Huber's contamination model (Huber '64):

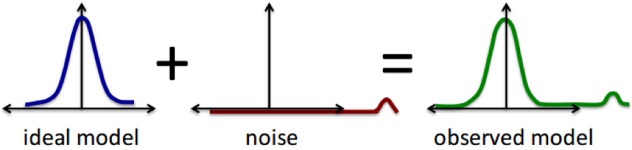


Figure: ϵ-fraction are arbitrary corruptions.

Contamination model

[G. Box] "All models are wrong, but some are useful."

What if the real data violate the assumptions required: Huber's contamination model (Huber '64):

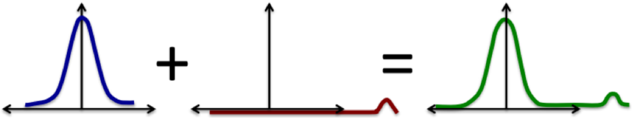


Figure: ϵ-fraction are arbitrary corruptions.

- A single corrupted sample can arbitrarily corrupt the original *M*-estimation (e.g., maximum likelihood estimation).
- In $\mathbb{R}^1$ case, trimmed mean has optimal guarantee $|\hat{\mu} - \mu| \leq O(\epsilon)$.

Heavy tailed model

Another way to model outliers is via heavy-tailed distributions.

A random variable X has heavy-tailed distribution if $\mathbb{E}|X|^k = \infty$ for some $k > 0$. For bounded second moment P, we have

$$\mathbb{E}_P(X) = \mu, \quad \mathrm{Var}_P(X) \le \sigma^2.$$

Heavy tailed model

Another way to model outliers is via heavy-tailed distributions.

A random variable X has heavy-tailed distribution if $\mathbb{E}|X|^k = \infty$ for some $k > 0$. For bounded second moment P, we have

$$\mathbb{E}_P(X) = \mu, \quad \mathrm{Var}_P(X) \leq \sigma^2.$$

The guarantees for empirical mean estimator are not satisfactory

$$\Pr\left(|\widehat{\mu} - \mu| \geq \sigma\sqrt{\frac{1/\alpha}{N}}\right) \leq \alpha.$$

Mean estimation in $\mathbb{R}^1$ under heavy tails

Median-of-means (MOM) estimator (Nemirovski & Yudin 1983):
Split samples into $k = \lceil \log(1/\alpha) \rceil$ groups $G_1, \cdots, G_k$ of size N/k:

$$\underbrace{\underbrace{\overbrace{X_1, \dots, X_{|G_1|}}^{G_1}}_{\bar{\mu}_1 := \frac{1}{|G_1|} \sum_{X_i \in G_1} X_i} \;\cdots\cdots\; \underbrace{\overbrace{X_{N-|G_k|+1}, \dots, X_N}^{G_k}}_{\bar{\mu}_k := \frac{1}{|G_k|} \sum_{X_i \in G_k} X_i}}_{\widehat{\mu}^{(k)} := \mathrm{median}(\bar{\mu}_1, \dots, \bar{\mu}_k)}$$

We recover the sub-Gaussian concentration

$$\Pr\left(\left| \widehat{\mu}^{(k)} - \mu \right| \geq 6.4\sigma \sqrt{\frac{\log(1/\alpha)}{N}} \right) \leq \alpha.$$

Robust statistics review: somewhat recent history

Arbitrary corruption

- Robust mean estimation (Diakonikolas et al., Lai, Rao & Vempala '16).
- Robust sparse mean estimation (Balakrishnan et al '17, **Liu et al** '18).
- Robust regression using robust gradient descent (Chen, Su & Xu '17, Prasad et al '18).
- Least Trimmed Squares type (Alfons et al. '13, Yang, Lozano & Aravkin '18, Shen & Sanghavi '19).

Heavy tailed distribution

- Catoni's mean estimator using Huber loss (Catoni '12).
- Covariance estimation with heavy-tailed entries (Minsker '18).
- MOM tournaments for ERM (Lugosi & Mendelson '16, Lecué & Lerasle '17, Jalal et al '20).

Summary

1. Restrictive conditions (RIP/RE/RSC) $\rightarrow$ optimal estimation in high dimensions.

Summary

1. Restrictive conditions (RIP/RE/RSC) $\rightarrow$ optimal estimation in high dimensions.

2. Many existing algorithms are efficient to deal with low dimensional structure in high dimensions.

Summary

1. Restrictive conditions (RIP/RE/RSC) $\rightarrow$ optimal estimation in high dimensions.
2. Many existing algorithms are efficient to deal with low dimensional structure in high dimensions.

Question

1. Under heavy tails or arbitrary corruption, what assumptions are sufficient to enable efficient and robust algorithms for high dimensional M-estimation?

Summary

1. Restrictive conditions (RIP/RE/RSC) $\rightarrow$ optimal estimation in high dimensions.
2. Many existing algorithms are efficient to deal with low dimensional structure in high dimensions.

Question

1. Under heavy tails or arbitrary corruption, what assumptions are sufficient to enable efficient and robust algorithms for high dimensional M-estimation?
2. Can we obtain robust algorithms without losing any computational efficiency?

Table of Contents

1. Introduction and Motivation
   - *M*-estimation in high dimensions
   - Robust statistics models
2. High dimensional robust *M*-estimation
   - Problem formulation
   - Robust Descent Condition
   - Main results
3. Low rank matrix regression under heavy tails
   - Problem formulation
   - Main results

Problem setup: heavy tailed distribution in $\mathbb{R}^d$

For a distribution P of $\boldsymbol{x} \in \mathbb{R}^d$ with mean $\mathbb{E}(\boldsymbol{x})$ and covariance $\boldsymbol{\Sigma}$,

Bounded $2k$-th moment

We say that P has bounded $2k$-th moment, if there is a universal constant C_{2k} such that, for a unit vector $\boldsymbol{v} \in \mathbb{R}^d$, we have

$$\mathbb{E}_P \left|\langle \boldsymbol{v}, \boldsymbol{x} - \mathbb{E}(\boldsymbol{x})\rangle\right|^{2k} \leq C_{2k}\, \mathbb{E}_P(\left|\langle \boldsymbol{v}, \boldsymbol{x} - \mathbb{E}(\boldsymbol{x})\rangle\right|^2)^k.$$

Problem setup: heavy tailed distribution in $\mathbb{R}^d$

For a distribution P of $\boldsymbol{x} \in \mathbb{R}^d$ with mean $\mathbb{E}(\boldsymbol{x})$ and covariance $\boldsymbol{\Sigma}$,

Bounded $2k$-th moment

We say that P has bounded $2k$-th moment, if there is a universal constant C_{2k} such that, for a unit vector $\boldsymbol{v} \in \mathbb{R}^d$, we have

$$\mathbb{E}_P \left|\langle \boldsymbol{v}, \boldsymbol{x} - \mathbb{E}(\boldsymbol{x})\rangle\right|^{2k} \leq C_{2k}\, \mathbb{E}_P(|\langle \boldsymbol{v}, \boldsymbol{x} - \mathbb{E}(\boldsymbol{x})\rangle|^2)^k.$$

For example, we will study sparse linear regression with bounded 4-th moments for $\boldsymbol{x}$ and bounded variance for y and noise.

Problem setup: ϵ-corrupted samples

Sparse regression model:

- $y_i = \boldsymbol{x}_i{}^T \beta^* + \xi_i$.
- sub-Gaussian covariates: $\mathrm{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}$.
- sub-Gaussian noise: $\mathrm{Var}(\xi) \leq \sigma^2$.

Contamination model:

- First, $\{z_i\} \sim P$.
- We observe $\{z_i, i \in \mathcal{S}\}$.
- P: sparse regression model.
- $\mathcal{S}$: Samples with corruption.
- ϵ: fraction of outliers.

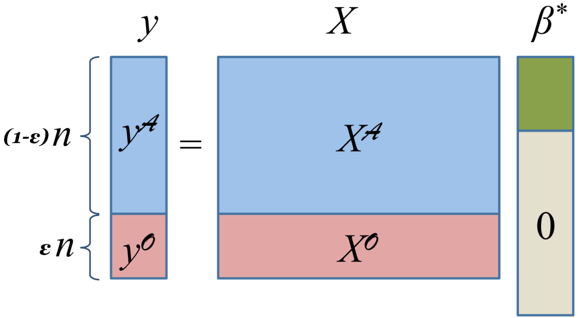
Problem setup: ϵ-corrupted samples

Sparse regression model:

- $y_i = \boldsymbol{x}_i{}^T \beta^* + \xi_i$.
- sub-Gaussian covariates: $\mathrm{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}$.
- sub-Gaussian noise: $\mathrm{Var}(\xi) \leq \sigma^2$.

Contamination model:

- First, $\{z_i\} \sim P$.
- We observe $\{z_i, i \in \mathcal{S}\}$.
- P: sparse regression model.
- $\mathcal{S}$: Samples with corruption.
- ϵ: fraction of outliers.



Related work for robust sparse regression

Arbitrary corruption

- Wright & Ma '10, Li '12, Bhatia, Jain & Kar '15, Karmalkar & Price '19: Robust regression resilient to a constant fraction of corruptions only in y.
- Chen, Caramanis & Mannor '13: Robust sparse regression resilient to corruptions in $\boldsymbol{x}$ and y.
- Balakrishnan et al '17, **Liu et al '18**, Diakonikolas et al '19: Robust sparse regression resilient to a constant fraction of corruptions in $\boldsymbol{x}$ and y. They only deal with identity/sparse covariance.

Heavy tailed distribution

- Hsu & Sabato '16, Loh '17: heavy tailed distribution only in y.
- Fan, Wang & Zhu '16: heavy tailed distribution in $\boldsymbol{x}$ and y.
- Lugosi & Mendelson '16: MOM tournaments, but not computationally tractable.

Dealing with corruption/heavy tails in $(\boldsymbol{x}, y)$

Chen, Caramanis & Mannor '13 and Fan, Wang & Zhu '16:

1. Pre-process $(\boldsymbol{x}, y)$ by trimming or shrinking.
2. The impacts of corruption/heavy tails are controlled.

Dealing with corruption/heavy tails in $(\boldsymbol{x}, y)$

Chen, Caramanis & Mannor ’13 and Fan, Wang & Zhu ’16:

1. Pre-process $(\boldsymbol{x}, y)$ by trimming or shrinking.
2. The impacts of corruption/heavy tails are controlled.
3. Restricted Eigenvalue condition holds on the processed data.
4. Common ℓ_1 strategy works on the processed data.

Dealing with corruption/heavy tails in $(\boldsymbol{x}, y)$

Chen, Caramanis & Mannor '13 and Fan, Wang & Zhu '16:

1. Pre-process $(\boldsymbol{x}, y)$ by trimming or shrinking.
2. The impacts of corruption/heavy tails are controlled.
3. Restricted Eigenvalue condition holds on the processed data.
4. Common ℓ_1 strategy works on the processed data.

However, this leads to sub-optimal recovery guarantees.

Dealing with corruption/heavy tails in $(\boldsymbol{x}, y)$

Chen, Caramanis & Mannor '13 and Fan, Wang & Zhu '16:

1. Pre-process $(\boldsymbol{x}, y)$ by trimming or shrinking.
2. The impacts of corruption/heavy tails are controlled.
3. Restricted Eigenvalue condition holds on the processed data.
4. Common ℓ_1 strategy works on the processed data.

However, this leads to sub-optimal recovery guarantees.

A simple example: sparse linear equations with outliers.

- A simple exhaustive search algorithm guarantees exact recovery.
- If the pre-processing does not remove all the outliers, exact recovery is impossible.
- Hence the pre-processing idea is not optimal.

Thought experiment

For the population risk $f(\boldsymbol{\beta}) = \mathbb{E}_{\boldsymbol{z}_i \sim P}\, \ell_i(\boldsymbol{\beta}; \boldsymbol{z}_i)$, suppose we had access to the population gradient $\boldsymbol{G}(\boldsymbol{\beta}) = \mathbb{E}_{\boldsymbol{z}_i \sim P}\, \nabla \ell_i(\boldsymbol{\beta}; \boldsymbol{z}_i)$.

Thought experiment

For the population risk $f(\boldsymbol{\beta}) = \mathbb{E}_{\boldsymbol{z}_i \sim P}\, \ell_i(\boldsymbol{\beta}; \boldsymbol{z}_i)$, suppose we had access to the population gradient $\boldsymbol{G}(\boldsymbol{\beta}) = \mathbb{E}_{\boldsymbol{z}_i \sim P}\, \nabla \ell_i(\boldsymbol{\beta}; \boldsymbol{z}_i)$.

We use Population Hard Thresholding

1. At current $\boldsymbol{\beta}^t$, we obtain $\boldsymbol{G}^t$.
2. Update the parameter[a]: $\boldsymbol{\beta}^{t+1} = \mathrm{P}_{k'}\left(\boldsymbol{\beta}^t - \eta \boldsymbol{G}^t\right)$.

[a] The hard thresholding operator keeps the largest (in magnitude) k' elements of a vector, and k' is proportional to k.

Thought experiment

For the population risk $f(\boldsymbol{\beta}) = \mathbb{E}_{\boldsymbol{z}_i \sim P}\, \ell_i(\boldsymbol{\beta}; \boldsymbol{z}_i)$, suppose we had access to the population gradient $\boldsymbol{G}(\boldsymbol{\beta}) = \mathbb{E}_{\boldsymbol{z}_i \sim P}\, \nabla \ell_i(\boldsymbol{\beta}; \boldsymbol{z}_i)$.

We use Population Hard Thresholding

1. At current $\boldsymbol{\beta}^t$, we obtain $\boldsymbol{G}^t$.
2. Update the parameter[a]: $\boldsymbol{\beta}^{t+1} = \mathrm{P}_{k'}\left(\boldsymbol{\beta}^t - \eta \boldsymbol{G}^t\right)$.

[a] The hard thresholding operator keeps the largest (in magnitude) k' elements of a vector, and k' is proportional to k.

If the population risk f satisfies μ_α-strong convexity & μ_β-smoothness:

$$\frac{\mu_\alpha}{2}\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2 \le f(\boldsymbol{\beta}_1) - f(\boldsymbol{\beta}_2) - |\langle \nabla f(\boldsymbol{\beta}_2), \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 \rangle| \le \frac{\mu_\beta}{2}\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2,$$

then Population Hard Thresholding with $\eta = \frac{1}{\mu_\beta}$ has linear convergence

$$\left\|\boldsymbol{\beta}^{t+1} - \boldsymbol{\beta}^*\right\|_2 \le \left(1 - \frac{\mu_\alpha}{\mu_\beta}\right)\left\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^*\right\|_2.$$

Finite-sample analysis and robustness

- In practice: no access to population gradient $\boldsymbol{G}(\boldsymbol{\beta})$.

- For authentic sub-Gaussian samples, empirical gradient $\widehat{\boldsymbol{G}}(\boldsymbol{\beta})$ should have well-controlled stochastic fluctuation.

- For ϵ-corrupted samples, empirical average $\widehat{\boldsymbol{G}}(\boldsymbol{\beta})$ can be arbitrarily bad.

Finite-sample analysis and robustness

- In practice: no access to population gradient $\boldsymbol{G}(\beta)$.
- For authentic sub-Gaussian samples, empirical gradient $\widehat{\boldsymbol{G}}(\beta)$ should have well-controlled stochastic fluctuation.
- For ϵ-corrupted samples, empirical average $\widehat{\boldsymbol{G}}(\beta)$ can be arbitrarily bad.
- We use a robust gradient estimator $\widehat{\boldsymbol{G}}_{\mathrm{rob}}(\beta)$, as a robust counterpart of the population version $\boldsymbol{G}(\beta)$.

Finite-sample analysis and robustness

- In practice: no access to population gradient $\boldsymbol{G}(\beta)$.

- For authentic sub-Gaussian samples, empirical gradient $\widehat{\boldsymbol{G}}(\beta)$ should have well-controlled stochastic fluctuation.

- For ϵ-corrupted samples, empirical average $\widehat{\boldsymbol{G}}(\beta)$ can be arbitrarily bad.

- We use a robust gradient estimator $\widehat{\boldsymbol{G}}_{\mathrm{rob}}(\beta)$, as a robust counterpart of the population version $\boldsymbol{G}(\beta)$.

- Question: a way to measure how close the robust version is to the population version in high dimensions?

Table of Contents

1. Introduction and Motivation
   - *M*-estimation in high dimensions
   - Robust statistics models
2. High dimensional robust *M*-estimation
   - Problem formulation
   - Robust Descent Condition
   - Main results
3. Low rank matrix regression under heavy tails
   - Problem formulation
   - Main results

$\widehat{\boldsymbol{G}}_{\mathrm{rob}}(\beta)$ vs. $\boldsymbol{G}(\beta)$ – how close?

- Past results for robust gradient descent in low dimensions (Chen, Su & Xu '17, Prasad et al '18) establish bounds on

$$\left\| \widehat{\boldsymbol{G}}_{\mathrm{rob}}(\beta) - \boldsymbol{G}(\beta) \right\|_2 .$$

$\widehat{\boldsymbol{G}}_{\mathrm{rob}}(\beta)$ vs. $\boldsymbol{G}(\beta)$ – how close?

- Past results for robust gradient descent in low dimensions (Chen, Su & Xu '17, Prasad et al '18) establish bounds on

$$\left\| \widehat{\boldsymbol{G}}_{\mathrm{rob}}(\beta) - \boldsymbol{G}(\beta) \right\|_2 .$$

- **Liu et al** '18 proposed Robust Sparse Gradient Estimator (RSGE) to bound $\| \widehat{\boldsymbol{G}}_{\mathrm{rob}}(\beta) - \boldsymbol{G}(\beta) \|_2$ in high dimensions.

- Stability of IHT + RSGE lead to optimal recovery (**Liu et al** '18).

$\widehat{\boldsymbol{G}}_{\mathrm{rob}}(\beta)$ vs. $\boldsymbol{G}(\beta)$ – how close?

- Past results for robust gradient descent in low dimensions (Chen, Su & Xu '17, Prasad et al '18) establish bounds on

$$\left\| \widehat{\boldsymbol{G}}_{\mathrm{rob}}(\beta) - \boldsymbol{G}(\beta) \right\|_2 .$$

- **Liu et al** '18 proposed Robust Sparse Gradient Estimator (RSGE) to bound $\| \widehat{\boldsymbol{G}}_{\mathrm{rob}}(\beta) - \boldsymbol{G}(\beta) \|_2$ in high dimensions.

- Stability of IHT + RSGE lead to optimal recovery (**Liu et al** '18).

- However, ℓ_2 norm bound may be too much to ask.
  - For general (non-sparse, non-identity) covariance?
  - Sparse logistic regression?

Robust Descent Condition

- RSGE $\|\widehat{\boldsymbol{G}}_{\mathrm{rob}}(\beta) - \boldsymbol{G}(\beta)\|_2$ requires bounds in all directions in high dimensions $\mathbb{R}^d$.

Robust Descent Condition

- RSGE $\|\widehat{\boldsymbol{G}}_{\mathrm{rob}}(\boldsymbol{\beta}) - \boldsymbol{G}(\boldsymbol{\beta})\|_2$ requires bounds in all directions in high dimensions $\mathbb{R}^d$.

- Intuition: IHT guarantees that the trajectory goes through sparse vectors, we only need to bound a small number of directions for robust gradients in $\mathbb{R}^d$.

Robust Descent Condition

- RSGE $\|\widehat{\boldsymbol{G}}_{\text{rob}}(\boldsymbol{\beta}) - \boldsymbol{G}(\boldsymbol{\beta})\|_2$ requires bounds in all directions in high dimensions $\mathbb{R}^d$.
- Intuition: IHT guarantees that the trajectory goes through sparse vectors, we only need to bound a small number of directions for robust gradients in $\mathbb{R}^d$.
- We propose a Robust Descent Condition (RDC).

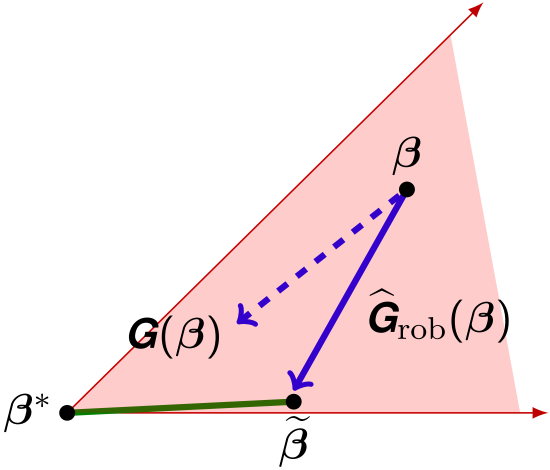
$$\left|\langle \widehat{\boldsymbol{G}}_{\text{rob}}(\boldsymbol{\beta}) - \boldsymbol{G}(\boldsymbol{\beta}), \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle\right| \leq \left(\alpha \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 + \psi\right) \left\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_2$$

  - $\boldsymbol{\beta}$ and $\widetilde{\boldsymbol{\beta}}$ are the subsequent iterates of the algorithm.
  - ψ is the accuracy of the robust gradient estimator.

- We show a Meta Theorem (Stability of Robust Hard Thresholding)
  - If we have a (α, ψ)-RDC, it guarantees $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O(\psi)$.

RDC: a geometric illustration

$$\left|\langle \widehat{\boldsymbol{G}}_{\mathrm{rob}}(\boldsymbol{\beta}) - \boldsymbol{G}(\boldsymbol{\beta}), \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle\right| \leq \left(\alpha \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 + \psi\right) \left\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_2$$



The stability property for Robust Hard Thresholding

Theorem 1 (Meta-Theorem)

Suppose we observe samples from a statistical model with population risk f satisfying μ_α-strong convexity and μ_β-smoothness.

If a robust gradient estimator satisfies (α, ψ)-Robust Descent Condition where $\alpha \leq \frac{1}{32}\mu_\alpha$, then Robust Hard Thresholding with $\eta = 1/\mu_\beta$ outputs $\widehat{\boldsymbol{\beta}}$ such that

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O(\psi/\mu_\alpha),$$

by setting $T = O\left(\log\left(\mu_\alpha \|\boldsymbol{\beta}^*\|_2/\psi\right)\right)$.

The stability property for Robust Hard Thresholding

Theorem 1 (Meta-Theorem)

Suppose we observe samples from a statistical model with population risk f satisfying μ_α-strong convexity and μ_β-smoothness.

If a robust gradient estimator satisfies (α, ψ)-Robust Descent Condition where $\alpha \leq \frac{1}{32}\mu_\alpha$, then Robust Hard Thresholding with $\eta = 1/\mu_\beta$ outputs $\widehat{\boldsymbol{\beta}}$ such that

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O(\psi/\mu_\alpha),$$

by setting $T = O\left(\log\left(\mu_\alpha \|\boldsymbol{\beta}^*\|_2/\psi\right)\right)$.

- We prefer a sufficiently small ψ.

The stability property for Robust Hard Thresholding

Theorem 1 (Meta-Theorem)

Suppose we observe samples from a statistical model with population risk f satisfying μ_α-strong convexity and μ_β-smoothness.

If a robust gradient estimator satisfies (α, ψ)-Robust Descent Condition where $\alpha \le \frac{1}{32}\mu_\alpha$, then Robust Hard Thresholding with $\eta = 1/\mu_\beta$ outputs $\widehat{\boldsymbol{\beta}}$ such that

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O(\psi/\mu_\alpha),$$

by setting $T = O\left(\log\left(\mu_\alpha\|\boldsymbol{\beta}^*\|_2/\psi\right)\right)$.

- We prefer a sufficiently small ψ.
- This Meta-Theorem is flexible enough to recover existing results.

Using RDC to recover existing results: I

We can use the RDC and the Meta-Theorem to recover existing results in the literature. Some immediate examples are as follows.

Using RDC to recover existing results: I

We can use the RDC and the Meta-Theorem to recover existing results in the literature. Some immediate examples are as follows.

When we have uncorrupted sub-Gaussian samples.

Suppose the samples follow from sparse linear regression with sub-Gaussian covariates and noise $\mathcal{N}(0, \sigma^2)$.

- The empirical average of gradients $\widehat{\boldsymbol{G}}$ satisfies the RDC with $\psi = O(\sigma\sqrt{\frac{k \log(d)}{n}})$.

- Plugging in this ψ to the Meta-Theorem recovers the well-known minimax rate for sparse linear regression.

Using RDC to recover existing results: II

When we have a constant fraction of arbitrary corruption.

When $\mathbf{\Sigma} = \boldsymbol{I}_d$ or is sparse, [BDLS17, LSLC18, DKK$^+$19] provide RSGE which upper bounds $\|\widehat{\boldsymbol{G}}_{\mathrm{rob}}(\boldsymbol{\beta}) - \boldsymbol{G}(\boldsymbol{\beta})\|_2 \leq \alpha\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 + \psi$, for a constant fraction ϵ of corrupted samples.

- Since $|\langle \widehat{\boldsymbol{G}}_{\mathrm{rob}}(\boldsymbol{\beta}) - \boldsymbol{G}(\boldsymbol{\beta}), \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle| \leq \|\widehat{\boldsymbol{G}}_{\mathrm{rob}}(\boldsymbol{\beta}) - \boldsymbol{G}(\boldsymbol{\beta})\|_2 \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$, we observe that RSGE implies RDC.

Using RDC to recover existing results: II

When we have a constant fraction of arbitrary corruption.

When $\boldsymbol{\Sigma} = \boldsymbol{I}_d$ or is sparse, [BDLS17, LSLC18, DKK+19] provide RSGE which upper bounds $\|\widehat{\boldsymbol{G}}_{\mathrm{rob}}(\boldsymbol{\beta}) - \boldsymbol{G}(\boldsymbol{\beta})\|_2 \le \alpha\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 + \psi$, for a constant fraction ϵ of corrupted samples.

- Since $|\langle \widehat{\boldsymbol{G}}_{\mathrm{rob}}(\boldsymbol{\beta}) - \boldsymbol{G}(\boldsymbol{\beta}), \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle| \le \|\widehat{\boldsymbol{G}}_{\mathrm{rob}}(\boldsymbol{\beta}) - \boldsymbol{G}(\boldsymbol{\beta})\|_2 \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2$, we observe that RSGE implies RDC.

- Hence any RSGE can be used.
  - For $\boldsymbol{\Sigma} = I$, [BDLS17, DKK+19] guarantees an RDC with $\psi = O(\sigma\epsilon)$ when $n = \Omega(k^2 \log d/\epsilon^2)$;
  - For unknown sparse $\boldsymbol{\Sigma}$, [LSLC18] guarantees $\psi = O(\sigma\sqrt{\epsilon})$ when $n = \Omega(k^2 \log d/\epsilon)$.

- Plugging in this ψ to the Meta-Theorem recovers the State-of-the-Art results for robust sparse regression.

Table of Contents

1. Introduction and Motivation
   - *M*-estimation in high dimensions
   - Robust statistics models
2. High dimensional robust *M*-estimation
   - Problem formulation
   - Robust Descent Condition
   - Main results
3. Low rank matrix regression under heavy tails
   - Problem formulation
   - Main results

Our robust algorithms based on RDC

Robust Descent Condition

$$\left|\langle \widehat{\boldsymbol{G}}_{\mathrm{rob}}(\boldsymbol{\beta}) - \boldsymbol{G}(\boldsymbol{\beta}), \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle\right| \leq \left(\alpha \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 + \psi\right) \left\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_2.$$

- When $\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ only takes a small number of directions, then it is a much easier condition to satisfy than the ℓ_2 norm.

Our robust algorithms based on RDC

Robust Descent Condition

$$\left|\langle \widehat{\boldsymbol{G}}_{\mathrm{rob}}(\boldsymbol{\beta}) - \boldsymbol{G}(\boldsymbol{\beta}), \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle\right| \leq \left(\alpha \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 + \psi\right) \left\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_2.$$

- When $\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ only takes a small number of directions, then it is a much easier condition to satisfy than the ℓ_2 norm.

- If $\boldsymbol{\beta}^*$ is sparse, and the algorithm guarantees that the trajectory goes through sparse vectors, then $\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ will always be sparse.

Our robust algorithms based on RDC

Robust Descent Condition

$$\left|\langle \widehat{\boldsymbol{G}}_{\mathrm{rob}}(\boldsymbol{\beta}) - \boldsymbol{G}(\boldsymbol{\beta}), \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle\right| \leq \left(\alpha \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 + \psi\right) \left\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_2.$$

- When $\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ only takes a small number of directions, then it is a much easier condition to satisfy than the ℓ_2 norm.
- If $\boldsymbol{\beta}^*$ is sparse, and the algorithm guarantees that the trajectory goes through sparse vectors, then $\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ will always be sparse.
- We only need to guarantee $\|\widehat{\boldsymbol{G}}_{\mathrm{rob}}(\boldsymbol{\beta}) - \boldsymbol{G}(\boldsymbol{\beta})\|_\infty$, and coordinate-wise technique suffices to obtain minimax result.

Our robust algorithms based on RDC

Robust Descent Condition

$$\left|\langle \widehat{\boldsymbol{G}}_{\mathrm{rob}}(\boldsymbol{\beta}) - \boldsymbol{G}(\boldsymbol{\beta}), \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle\right| \le \left(\alpha \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 + \psi\right) \left\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_2.$$

- When $\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ only takes a small number of directions, then it is a much easier condition to satisfy than the ℓ_2 norm.
- If $\boldsymbol{\beta}^*$ is sparse, and the algorithm guarantees that the trajectory goes through sparse vectors, then $\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ will always be sparse.
- We only need to guarantee $\|\widehat{\boldsymbol{G}}_{\mathrm{rob}}(\boldsymbol{\beta}) - \boldsymbol{G}(\boldsymbol{\beta})\|_\infty$, and coordinate-wise technique suffices to obtain minimax result.
- For $\mathbb{R}^1$ mean estimation, we can use trimmed mean for corrupted samples and median-of-means for heavy tails.

Our robust algorithms based on RDC

- We only need to guarantee $\|\widehat{\boldsymbol{G}}_{\text{rob}}(\boldsymbol{\beta}) - \boldsymbol{G}(\boldsymbol{\beta})\|_\infty$, and coordinate-wise technique suffices to obtain minimax result.

- For $\mathbb{R}^1$ mean estimation, we can use trimmed mean for corrupted samples and median-of-means for heavy tails.

Our robust algorithms based on RDC

- We only need to guarantee $\|\widehat{\boldsymbol{G}}_{\mathrm{rob}}(\boldsymbol{\beta}) - \boldsymbol{G}(\boldsymbol{\beta})\|_\infty$, and coordinate-wise technique suffices to obtain minimax result.

- For $\mathbb{R}^1$ mean estimation, we can use trimmed mean for corrupted samples and median-of-means for heavy tails.

Robust Hard Thresholding

1. At current $\boldsymbol{\beta}^t$, calculate all gradients: $\boldsymbol{g}_i^t = \nabla \ell_i(\boldsymbol{\beta}^t)$, $i \in [n]$.
2. For $\{\boldsymbol{g}_i^t\}_{i=1}^n$, we obtain $\widehat{\boldsymbol{G}}_{\mathrm{rob}}^t$ satisfying the RDC by using two options:

   (♠) trimmed gradient estimator for arbitrary corruption.
   (♣) MOM gradient estimator for heavy tailed distribution.
3. Update the parameter: $\boldsymbol{\beta}^{t+1} = \mathrm{P}_{k'}\left(\boldsymbol{\beta}^t - \eta \widehat{\boldsymbol{G}}_{\mathrm{rob}}^t\right)$.

Main results

Simple coordinate-wise technique gives sharp results

Corollary for arbitrary corruptions

- Resilient to a $(1/\sqrt{k})$-fraction of arbitrary outliers.
- When $\epsilon \to 0$, we have minimax rate.
- When $\sigma^2 \to 0$, we have exact recovery.

Main results

Simple coordinate-wise technique gives sharp results

Corollary for arbitrary corruptions

- Resilient to a $(1/\sqrt{k})$-fraction of arbitrary outliers.
- When $\epsilon \to 0$, we have minimax rate.
- When $\sigma^2 \to 0$, we have exact recovery.

Corollary for heavy tailed distribution

- Can deal with bounded 4-th moment covariates.
- The same minimax rate as the sub-Gaussian case.
- When $\sigma^2 \to 0$, we have exact recovery.

Main results

Simple coordinate-wise technique gives sharp results

Corollary for arbitrary corruptions

- Resilient to a $(1/\sqrt{k})$-fraction of arbitrary outliers.
- When $\epsilon \to 0$, we have minimax rate.
- When $\sigma^2 \to 0$, we have exact recovery.

Corollary for heavy tailed distribution

- Can deal with bounded 4-th moment covariates.
- The same minimax rate as the sub-Gaussian case.
- When $\sigma^2 \to 0$, we have exact recovery.

Computational complexity: both of them are nearly linear time.

Simulation study: arbitrary corruption

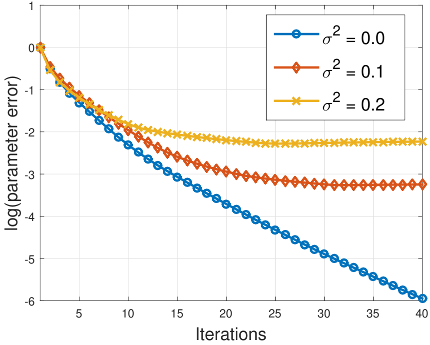


Figure: The corruption level ϵ is fixed and we use trimmed gradient for different noise level σ^2. We plot $\log(\left\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^*\right\|_2)$ vs. iterates.

Simulation study: heavy tailed distribution

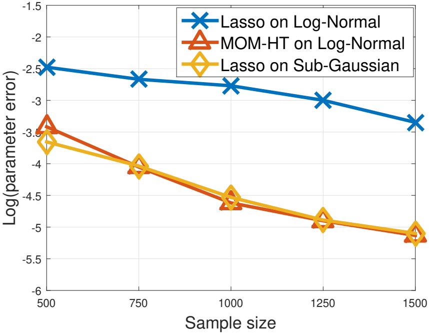


Figure: We consider log-normal samples, and we use MOM gradient for different sample size to compare with baselines (Lasso on heavy tailed data, and Lasso on sub-Gaussian data). We plot $\log(\|\beta^t - \beta^*\|_2)$ vs. sample size.

Summary

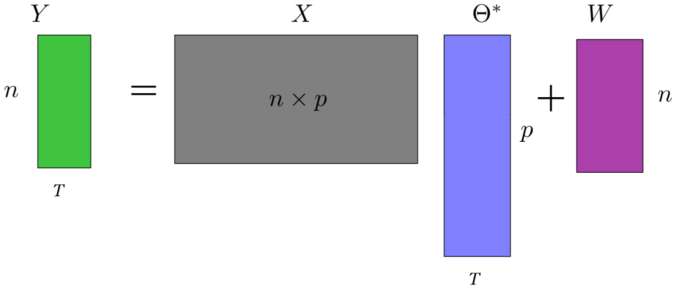
- Important distinction in high dimensional statistics: corruption/heavy tails both in $(\boldsymbol{x}, y)$ vs. only in y.
- A natural condition we call the Robust Descent Condition.
- RDC + Robust Hard Thresholding: fast linear convergence to minimax rate.
- Sharpest available error bound for corruption/heavy tails models.

Table of Contents

1. Introduction and Motivation
   - *M*-estimation in high dimensions
   - Robust statistics models
2. High dimensional robust *M*-estimation
   - Problem formulation
   - Robust Descent Condition
   - Main results
3. Low rank matrix regression under heavy tails
   - Problem formulation
   - Main results

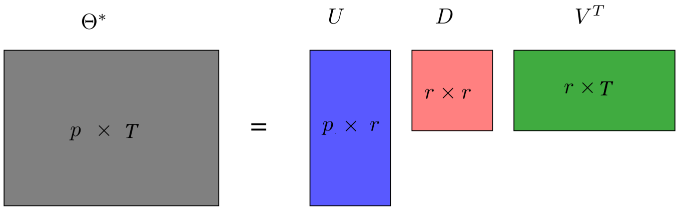
Low rank matrix regression

Matrix regression (multivariate regression) has n samples which considers prediction with T tasks by mapping $\boldsymbol{x} \in \mathbb{R}^p$ to $\boldsymbol{y} \in \mathbb{R}^T$.



Low rank matrix regression

We are interested in the low rank structure of $\boldsymbol{\Theta} \in \mathbb{R}^{p\times T}$.



- For sub-Gaussian data X and W, rank-r assumption for $\boldsymbol{\Theta}^*$ guarantees the estimation error $\sqrt{\frac{r(p+T)}{n}}$, instead of $\sqrt{\frac{pT}{n}}$.
- Nuclear norm regularization* (similar to ℓ_1 regularization) or Singular Value Projection† (SVP, similar to IHT).

*The nuclear norm is the summation of the singular values.

†The SVP iteratively makes an orthogonal projection onto a set of low-rank matrices.

Table of Contents

1. Introduction and Motivation
   - *M*-estimation in high dimensions
   - Robust statistics models
2. High dimensional robust *M*-estimation
   - Problem formulation
   - Robust Descent Condition
   - Main results
3. Low rank matrix regression under heavy tails
   - Problem formulation
   - Main results

RDC in matrix space

What if the explanatory variable $\boldsymbol{x}$ and the stochastic noise $\boldsymbol{w}$ follow heavy tailed distribution (bounded 4-th moment)?

RDC in matrix space

What if the explanatory variable $\boldsymbol{x}$ and the stochastic noise $\boldsymbol{w}$ follow heavy tailed distribution (bounded 4-th moment)?

- Recall that IHT + RDC $\rightarrow$ a robust estimator for heavy tailed sparse regression.

RDC in matrix space

What if the explanatory variable $\boldsymbol{x}$ and the stochastic noise $\boldsymbol{w}$ follow heavy tailed distribution (bounded 4-th moment)?

- Recall that IHT + RDC $\rightarrow$ a robust estimator for heavy tailed sparse regression.
- We can use Singular Value Projection + matrix version of RDC.

RDC in matrix space

What if the explanatory variable $\boldsymbol{x}$ and the stochastic noise $\boldsymbol{w}$ follow heavy tailed distribution (bounded 4-th moment)?

- Recall that IHT + RDC $\rightarrow$ a robust estimator for heavy tailed sparse regression.
- We can use Singular Value Projection + matrix version of RDC.

RDC in vector space

$$\left|\langle \widehat{\boldsymbol{G}}_{\mathrm{rob}}(\boldsymbol{\beta}) - \boldsymbol{G}(\boldsymbol{\beta}), \widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\rangle\right| \leq \left(\alpha\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 + \psi\right)\left\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_2.$$

RDC in matrix space

$$\left|\langle \widehat{\boldsymbol{G}}_{\mathrm{rob}}(\boldsymbol{\Theta}) - \boldsymbol{G}(\boldsymbol{\Theta}), \widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\rangle\right| \leq \left(\alpha\left\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\right\|_{\mathrm{F}} + \psi\right)\left\|\widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\right\|_{\mathrm{F}}.$$

Robust gradient in matrix space

RDC in matrix space

$$\left|\langle \widehat{\boldsymbol{G}}_{\mathrm{rob}}(\boldsymbol{\Theta}) - \boldsymbol{G}(\boldsymbol{\Theta}), \widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^* \rangle\right| \lesssim (\alpha \left\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\right\|_{\mathrm{F}} + \psi) \left\|\widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\right\|_{\mathrm{F}}.$$

- The trajectory $\widetilde{\boldsymbol{\Theta}}$ is guaranteed to be low rank by SVP.
- We only need to guarantee $\|\widehat{\boldsymbol{G}}_{\mathrm{rob}}(\boldsymbol{\Theta}) - \boldsymbol{G}(\boldsymbol{\Theta})\|_{\mathrm{op}}$.

Robust gradient in matrix space

RDC in matrix space

$$\left|\langle \widehat{\boldsymbol{G}}_{\mathrm{rob}}(\boldsymbol{\Theta}) - \boldsymbol{G}(\boldsymbol{\Theta}), \widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^* \rangle\right| \lesssim \left(\alpha \left\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\right\|_{\mathrm{F}} + \psi\right) \left\|\widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\right\|_{\mathrm{F}}.$$

- The trajectory $\widetilde{\boldsymbol{\Theta}}$ is guaranteed to be low rank by SVP.
- We only need to guarantee $\|\widehat{\boldsymbol{G}}_{\mathrm{rob}}(\boldsymbol{\Theta}) - \boldsymbol{G}(\boldsymbol{\Theta})\|_{\mathrm{op}}$.
- We leverage a robust matrix estimator from (Minsker '18):
  - trim the spectrum of each sample, and the remaining average will have sub-Gaussian concentration bound.

Robust gradient in matrix space

RDC in matrix space

$$\left|\langle \widehat{\boldsymbol{G}}_{\mathrm{rob}}(\boldsymbol{\Theta}) - \boldsymbol{G}(\boldsymbol{\Theta}), \widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^* \rangle\right| \lesssim (\alpha \left\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\right\|_{\mathrm{F}} + \psi) \left\|\widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\right\|_{\mathrm{F}}.$$

- The trajectory $\widetilde{\boldsymbol{\Theta}}$ is guaranteed to be low rank by SVP.
- We only need to guarantee $\|\widehat{\boldsymbol{G}}_{\mathrm{rob}}(\boldsymbol{\Theta}) - \boldsymbol{G}(\boldsymbol{\Theta})\|_{\mathrm{op}}$.
- We leverage a robust matrix estimator from (Minsker '18):
  - trim the spectrum of each sample, and the remaining average will have sub-Gaussian concentration bound.
- This robust gradient estimator satisfies matrix version of RDC, and the Robust SVP converges linearly to the error rate $\sqrt{\frac{r(p+T)}{n}}$.

Robust gradient in matrix space

RDC in matrix space

$$\left|\left\langle \widehat{\boldsymbol{G}}_{\mathrm{rob}}(\boldsymbol{\Theta}) - \boldsymbol{G}(\boldsymbol{\Theta}), \widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^* \right\rangle\right| \lesssim \left(\alpha \left\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\right\|_{\mathrm{F}} + \psi\right) \left\|\widetilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\right\|_{\mathrm{F}}.$$

- The trajectory $\widetilde{\boldsymbol{\Theta}}$ is guaranteed to be low rank by SVP.
- We only need to guarantee $\|\widehat{\boldsymbol{G}}_{\mathrm{rob}}(\boldsymbol{\Theta}) - \boldsymbol{G}(\boldsymbol{\Theta})\|_{\mathrm{op}}$.
- We leverage a robust matrix estimator from (Minsker '18):
    - trim the spectrum of each sample, and the remaining average will have sub-Gaussian concentration bound.
- This robust gradient estimator satisfies matrix version of RDC, and the Robust SVP converges linearly to the error rate $\sqrt{\frac{r(p+T)}{n}}$.
- The Robust SVP takes $O(npT)$-time complexity per iteration.

Robust factorized gradient descent

Speed up by Burer-Monteiro formulation $\boldsymbol{\Theta} = \boldsymbol{U}\boldsymbol{V}^\top$, where $\boldsymbol{U} \in \mathbb{R}^{p\times r}$, and $\boldsymbol{V} \in \mathbb{R}^{T\times r}$.

Robust factorized gradient descent

$\widehat{\boldsymbol{G}}_{\boldsymbol{U}}$ and $\widehat{\boldsymbol{G}}_{\boldsymbol{V}}$ are robust versions of gradients on $\boldsymbol{U}$ and $\boldsymbol{V}$,

$$\boldsymbol{U}^{t+1} = \boldsymbol{U}^{t} - \eta\widehat{\boldsymbol{G}}_{\boldsymbol{U}},$$
$$\boldsymbol{V}^{t+1} = \boldsymbol{V}^{t} - \eta\widehat{\boldsymbol{G}}_{\boldsymbol{V}}.$$

Robust factorized gradient descent

Speed up by Burer-Monteiro formulation $\boldsymbol{\Theta} = \boldsymbol{U}\boldsymbol{V}^\top$, where $\boldsymbol{U} \in \mathbb{R}^{p \times r}$, and $\boldsymbol{V} \in \mathbb{R}^{T \times r}$.

Robust factorized gradient descent

$\widehat{\boldsymbol{G}}_{\boldsymbol{U}}$ and $\widehat{\boldsymbol{G}}_{\boldsymbol{V}}$ are robust versions of gradients on $\boldsymbol{U}$ and $\boldsymbol{V}$,

$$\boldsymbol{U}^{t+1} = \boldsymbol{U}^t - \eta \widehat{\boldsymbol{G}}_{\boldsymbol{U}},$$
$$\boldsymbol{V}^{t+1} = \boldsymbol{V}^t - \eta \widehat{\boldsymbol{G}}_{\boldsymbol{V}}.$$

- An element-wise MOM gradient estimator for $\boldsymbol{U}$ and $\boldsymbol{V}$.

Robust factorized gradient descent

Speed up by Burer-Monteiro formulation $\boldsymbol{\Theta} = \boldsymbol{U}\boldsymbol{V}^\top$, where $\boldsymbol{U} \in \mathbb{R}^{p\times r}$, and $\boldsymbol{V} \in \mathbb{R}^{T\times r}$.

Robust factorized gradient descent

$\widehat{\boldsymbol{G}}_{\boldsymbol{U}}$ and $\widehat{\boldsymbol{G}}_{\boldsymbol{V}}$ are robust versions of gradients on $\boldsymbol{U}$ and $\boldsymbol{V}$,

$$\boldsymbol{U}^{t+1} = \boldsymbol{U}^t - \eta\widehat{\boldsymbol{G}}_{\boldsymbol{U}},$$
$$\boldsymbol{V}^{t+1} = \boldsymbol{V}^t - \eta\widehat{\boldsymbol{G}}_{\boldsymbol{V}}.$$

- An element-wise MOM gradient estimator for $\boldsymbol{U}$ and $\boldsymbol{V}$.
- Nearly the same statistical results as the Robust SVP.
- Time complexity $O(nr(p+T))$ per iteration.

Robust factorized gradient descent

Speed up by Burer-Monteiro formulation $\boldsymbol{\Theta} = \boldsymbol{U}\boldsymbol{V}^\top$, where $\boldsymbol{U} \in \mathbb{R}^{p \times r}$, and $\boldsymbol{V} \in \mathbb{R}^{T \times r}$.

Robust factorized gradient descent

$\widehat{\boldsymbol{G}}_{\boldsymbol{U}}$ and $\widehat{\boldsymbol{G}}_{\boldsymbol{V}}$ are robust versions of gradients on $\boldsymbol{U}$ and $\boldsymbol{V}$,

$$\boldsymbol{U}^{t+1} = \boldsymbol{U}^t - \eta \widehat{\boldsymbol{G}}_{\boldsymbol{U}},$$
$$\boldsymbol{V}^{t+1} = \boldsymbol{V}^t - \eta \widehat{\boldsymbol{G}}_{\boldsymbol{V}}.$$

- An element-wise MOM gradient estimator for $\boldsymbol{U}$ and $\boldsymbol{V}$.
- Nearly the same statistical results as the Robust SVP.
- Time complexity $O(nr(p+T))$ per iteration.
- Local linear convergence guarantee.

Summary

- A natural extension of the RDC to the low-rank setting.
- For covariates $\boldsymbol{x}$ with 4-th moment bound, we show that a gradient estimator adapted from (Minsker '18) satisfies the RDC.
- Our algorithm, Robust SVP, obtains the sub-Gaussian rate, with time complexity $O(npT)$ per iteration.
- Factorized robust gradient descent uses element-wise MOM.
  - Local linear convergence to the sub-Gaussian rate.
  - The time complexity is reduced to $O(nr(p+T))$ per iteration.

Publications during PhD

- Zhuo, J., **Liu, L.**, & Caramanis, C. (2020). Robust Structured Statistical Estimation via Conditional Gradient Type Methods. arXiv preprint arXiv:2007.03572.
- Jalal, A., **Liu, L.**, Dimakis, A. G., & Caramanis, C. (2020). Robust compressed sensing of generative models. In NeurIPS 2020.
- **Liu, L.**, Li, T., & Caramanis, C. (2019). Low Rank Matrix Regression under Heavy Tailed Distribution. Submitted.
- **Liu, L.**, Li, T., & Caramanis, C. (2019). High Dimensional Robust M-Estimation: Arbitrary Corruption and Heavy Tails. arXiv preprint arXiv:1901.08237.
- **Liu, L.**, Shen, Y., Li, T., & Caramanis, C. (2020). High dimensional robust sparse regression. In AISTATS 2020.
- Li, T., Kyrillidis, A., **Liu, L.**, & Caramanis, C. (2018). Approximate Newton-based statistical inference using only stochastic gradients. arXiv preprint arXiv:1805.08920.
- Li, T., **Liu, L.**, Kyrillidis, A., & Caramanis, C. (2018). Statistical Inference Using SGD. In AAAI 2018.

Thank you

Many thanks to all of my collaborators:

Constantine Caramanis, Alex Dimakis, Ajil Jalal, Anastasios Kyrillidis, Tianyang Li, Yanyao Shen, and Jiacheng Zhuo.

References I

[BDLS17] Sivaraman Balakrishnan, Simon S. Du, Jerry Li, and Aarti Singh. Computationally efficient robust sparse estimation in high dimensions. In *Proceedings of the 2017 Conference on Learning Theory*, 2017.

[DKK+19] Ilias Diakonikolas, Daniel Kane, Sushrut Karmalkar, Eric Price, and Alistair Stewart. Outlier-robust high-dimensional sparse estimation via iterative filtering. *Advances in Neural Information Processing Systems*, 32:10689–10700, 2019.

[LSLC18] Liu Liu, Yanyao Shen, Tianyang Li, and Constantine Caramanis. High dimensional robust sparse regression. *arXiv preprint arXiv:1805.11643*, 2018.