PART III: EXPERIMENTS AND EXTENSIONS
OUTLINE

Part III

• General Framework for Robust Mean Estimation

• Experiments

• Robust Stochastic Optimization

• Learning with Majority of Outliers
OUTLINE

Part III

• General Framework for Robust Mean Estimation

• Experiments

• Robust Stochastic Optimization

• Learning with Majority of Outliers
Non-Convex Optimization Formulation (I)

Optimization Formulation:
Assign weights to the samples so that weighted empirical mean works.

Let
\[ \Delta_{N,\epsilon} = \left\{ w \in \mathbb{R}^N : \|w\|_1 = 1 \text{ and } 0 \leq w_i \leq \frac{1}{(1 - \epsilon)N} \right\} \]

\[ \hat{\mu}_w = \sum_{i=1}^{N} w_i X_i \quad \text{and} \quad \hat{\Sigma}_w = \sum_{i=1}^{N} w_i (X_i - \hat{\mu}_w)(X_i - \hat{\mu}_w)^T. \]

Generalization of Key Lemma: For any \( w \in \Delta_{N,2\epsilon} \)

\[ \|\hat{\Sigma}_w\|_2 \leq 1 + O(\epsilon \log(1/\epsilon)) \quad \rightarrow \quad \|\hat{\mu}_w - \mu\|_2 = O(\epsilon \sqrt{\log(1/\epsilon)}) \]
Non-Convex Optimization Formulation (II)

Notation: \( \Delta_{N,\epsilon} = \left\{ w \in \mathbb{R}^N : \|w\|_1 = 1 \text{ and } 0 \leq w_i \leq \frac{1}{(1 - \epsilon)N} \right\} \)

\[
\hat{\mu}_w = \sum_{i=1}^{N} w_i X_i \\
\hat{\Sigma}_w = \sum_{i=1}^{N} w_i (X_i - \hat{\mu}_w)(X_i - \hat{\mu}_w)^T.
\]

Generalization of Key Lemma

\[
\|\hat{\Sigma}_w\|_2 \leq 1 + O(\epsilon \log(1/\epsilon)) \\
\|\hat{\mu}_w - \mu\|_2 = O(\epsilon \sqrt{\log(1/\epsilon)})
\]

Non-Convex Formulation:

\[
\min_w \|\hat{\Sigma}_w\|_2 \text{ subject to } w \in \Delta_{N,2\epsilon}
\]
NON-CONVEX OPTIMIZATION FORMULATION (III)

Notation: \( \Delta_{N,\epsilon} = \left\{ w \in \mathbb{R}^N : \|w\|_1 = 1 \text{ and } 0 \leq w_i \leq \frac{1}{(1-\epsilon)N} \right\} \)

\[
\hat{\mu}_w = \sum_{i=1}^{N} w_i X_i \\
\hat{\Sigma}_w = \sum_{i=1}^{N} w_i (X_i - \hat{\mu}_w)(X_i - \hat{\mu}_w)^T.
\]

Non-Convex Formulation:

\[
\min_w \|\hat{\Sigma}_w\|_2 \text{ subject to } w \in \Delta_{N,2\epsilon}
\]

Algorithmic Approaches:

• This is what filtering does!
• Ellipsoid Method [DKKLMS’16]
• Bi-level optimization [Cheng-D-Ge’18] (near-linear time!)
• Gradient Descent [Cheng-D-Ge-Soltanolkotabi, ICML’20]
**ROBUST MEAN ESTIMATION VIA GRADIENT-DESCENT**

**Notation:** \( \Delta_{N,\epsilon} = \left\{ w \in \mathbb{R}^N : \|w\|_1 = 1 \text{ and } 0 \leq w_i \leq \frac{1}{(1-\epsilon)N} \right\} \)

\[
\hat{\mu}_w = \sum_{i=1}^{N} w_i X_i \\
\hat{\Sigma}_w = \sum_{i=1}^{N} w_i (X_i - \hat{\mu}_w)(X_i - \hat{\mu}_w)^T.
\]

**Non-Convex Formulation:**

\[
\min_w \|\hat{\Sigma}_w\|_2 \text{ subject to } w \in \Delta_{N,2\epsilon}
\]

**Theorem** [Cheng-D-Ge-Soltanolkotabi, ICML’20, Paper Id: #6611]

Any approximate stationary point \( w \) defines \( \hat{\mu}_w \) that is close to \( \mu \).

See also [Zhu et al., Arxiv, May 2020]
Outline

Part III

• General Framework for Robust Mean Estimation

• Experiments

• Robust Stochastic Optimization

• Learning with Majority of Outliers
EXPERIMENTS

Being Robust (in High Dimensions) Can Be Practical
D., Kamath, Kane, Li, Moitra, Stewart, ICML’17
SYNTHEtic EXPERImENTS: UNKnOWn MEAn

Error rates on synthetic data (unknown mean):

\[ \mathcal{N}(\mu, I) + 10\% \text{ noise} \]
SYNTHETIC EXPERIMENTS: UNKNOWN MEAN

Error rates on synthetic data (unknown mean):

- Filtering
- Sample mean w/ noise
- Pruning
- RANSAC
- Geometric Median
ROBUST COVARIANCE ESTIMATION

**Problem**: Given an $\epsilon$-corrupted set of points $x_1, \ldots, x_N \in \mathbb{R}^d$ from an unknown distribution $D$ in a known family $\mathcal{F}$, estimate the covariance of $D$.

**Theorem**: Let $\epsilon < 1/2$. We can efficiently recover $\hat{\sum}$ such that

$$\| \sum^{-1/2}(\hat{\sum} - \sum)\sum^{-1/2} \|_F \leq f(\epsilon) + \tilde{O}(d/\sqrt{N}) ,$$

where $f$ depends on the concentration of $D$.

If $D$ is a Gaussian, then $f(\epsilon) = O(\epsilon \log(1/\epsilon))$.

[D-Kamath-Kane-Li-Moitra-Stewart, FOCS’16]
SYNTHETIC EXPERIMENTS: UNKNOWN COVARIANCE (I)

Error rates on synthetic data (**unknown covariance, isotropic**):

\[ \mathcal{N}(0, \Sigma) + 10\% \text{ noise} \]

\[ \text{close to identity} \]
SYNTHETIC EXPERIMENTS: UNKNOWN COVARIANCE (I)

Error rates on synthetic data (unknown covariance, isotropic):

![Graphs showing error rates with different methods as a function of dimension.](image-url)
SYNTHETIC EXPERIMENTS: UNKNOWN COVARIANCE (II)

Error rates on synthetic data (unknown covariance):

\[ \mathcal{N}(0, \Sigma) + 10\% \text{ noise} \]

far from identity
SYNTHETIC EXPERIMENTS: UNKNOWN COVARIANCE (II)

Error rates on synthetic data (unknown covariance, anisotropic):

- Filtering
- LRV Cov
- Sample covariance w/ noise
- Pruning
- RANSAC
REAL DATA EXPERIMENTS

[Novembre et al. ’08]: Take top two singular vectors of people x SNP matrix (POPRES)

“Genes Mirror Geography in Europe”
EXPERIMENTS: PRUNING PROJECTION

A comparison of error rate on semi-synthetic data:
EXPERIMENTS: RANSAC PROJECTION

A comparison of error rate on semi-synthetic data:
EXPERIMENTS: ROBUST PCA [XCS]

A comparison of error rate on semi-synthetic data:
EXPERIMENTS: FILTER PROJECTION
**Outline**

**Part III**

- General Framework for Robust Mean Estimation
- Experiments
- Robust Stochastic Optimization
- Learning with Majority of Outliers
Sever: A Robust Meta-Algorithm for Stochastic Optimization.
[D-Kamath-Kane-Li-Steinhardt-Stewart, ICML’19]
**ROBUST STOCHASTIC CONVEX OPTIMIZATION**

**Problem**: Given loss function $\ell(X, w)$ and $\epsilon$-corrupted samples from a distribution $\mathcal{D}$ over $X$, minimize $f(w) = \mathbb{E}_{X \sim \mathcal{D}}[\ell(X, w)]$

**Difficulty**: Corrupted data can move the gradients.

**Theorem**: Suppose $\ell$ is convex and $\text{Cov}_{X \sim \mathcal{D}}[\nabla \ell(X, w)] \preceq \sigma^2 \cdot I$. Under mild assumptions on $\mathcal{D}$, can recover a point such that

$$f(\hat{w}) - \min_w f(w) \leq O(\sigma \sqrt{\epsilon}) .$$

**Main Idea**: Filter at minimizer of empirical risk.
**Specific Applications**

**Corollary**: Outlier-robust learning algorithms with dimension-independent error guarantees for:
- SVMs
- Linear Regression
- Logistic Regression
- GLMs
- **Experimental Performance Against Data Poisoning Attacks.**

Concurrent works obtained tighter guarantees in terms of either sample complexity or error, by focusing on specific tasks and distributional assumptions [Klivans-Kothari-Meka’18, Diakonikolas-Kong-Stewart’18, …].
EXPERIMENTS: RIDGE REGRESSION

Regression: Synthetic data

Regression: Drug discovery data

Regression: Drug discovery data, attack targeted against SEVER

Test Error vs. Outlier Fraction $\epsilon$

- *uncorrupted*  
- *l2*  
- *loss*  
- *gradientCentered*  
- *SEVER*
OUTLINE

Part III

• General Framework for Robust Mean Estimation

• Experiments

• Robust Stochastic Optimization

• Learning with Majority of Outliers
Learning with a Majority of Outliers

- So far focused on setting where $\epsilon < 1/2$.

- What can we learn from a dataset in which the majority of points are corrupted?

**Problem:** Given a set of points $x_1, \ldots , x_N \in \mathbb{R}^d$ and $0 < \alpha \leq 1/2$ such that:
- An unknown subset of $\alpha N$ points are drawn from an unknown $D \in \mathcal{F}$, and
- The remaining $(1 - \alpha) N$ points are arbitrary, approximate the mean $\mu$ of $D$.

Which is the “real” $D$?
**LIST-DECODABLE LEARNING**

- Return *several hypotheses* with the guarantee that at least one is close.

**List-Decodable Mean Estimation:**
Given a set of points $x_1, \ldots, x_N \in \mathbb{R}^d$ and $0 < \alpha \leq 1/2$ such that:
- An unknown subset of $\alpha N$ points are drawn from an unknown $D \in \mathcal{F}$, and
- The remaining $(1 - \alpha)N$ points are arbitrary,

output a small list of $s$ hypotheses vectors such that one is close to the mean $\mu$ of $D$.

- Model defined in [Balcan-Blum-Vempala’08]
- First studied for mean estimation [Charikar-Steinhardt-Valiant’17]
- Application: Learning Mixture Models
**List-Decodable Mean Estimation**

**Theorem [Charikar-Steinhardt-Valiant’17]:** Let $0 < \alpha \leq 1/2$. If $D$ has covariance $\Sigma \preceq I$ there is an efficient algorithm that uses $N \geq d/\alpha$ corrupted points, and outputs a list of $s = O(1/\alpha)$ vectors $\hat{\mu}_1, \ldots, \hat{\mu}_s$ such that with high probability

$$\min_i \|\hat{\mu}_i - \mu\|_2 = \tilde{O}(1/\sqrt{\alpha}) .$$

**Theorem [Diakonikolas-Kane-Stewart’18] Any** list-decodable mean estimator for bounded covariance distributions must have error $\Omega(1/\sqrt{\alpha})$ as long as the list size is any function of $\alpha$.

- Initial algorithm [CSV’17] based on ellipsoid method.
- Generalization of filtering (“multi-filtering”) works for list-decodable setting [DKS’18].