Recent Advances in High-Dimensional Robust Statistics

Ilias Diakonikolas (UW Madison)

ICML 2020 Tutorial

July 2020

Can we develop learning algorithms that are ***robust*** to a ***constant*** fraction of ***corruptions*** in the data?

PART I: INTRODUCTION

MOTIVATION

- **Model Misspecification/Robust Statistics**
  [Fisher 1920s, Tukey 1960s, Huber 1960s]

- **Outlier Detection/Removal**

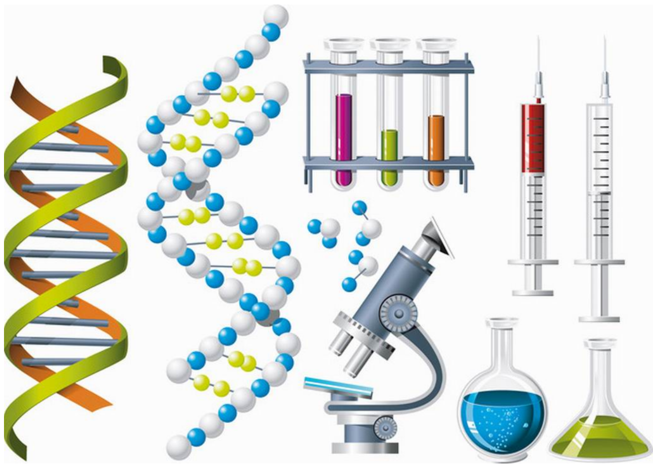
- **Adversarial/Secure ML**

DETECTING OUTLIERS IN REAL DATASETS

- High-dimensional datasets tend to be inherently noisy.

Biological Datasets: POPRES project,
HGDP datasets

[November *et al.*, Nature'08];
[Rosenberg *et al.*, Science'02];
[Li *et al.*, Science'08];
[Paschou *et al*., Medical Genetics'10]

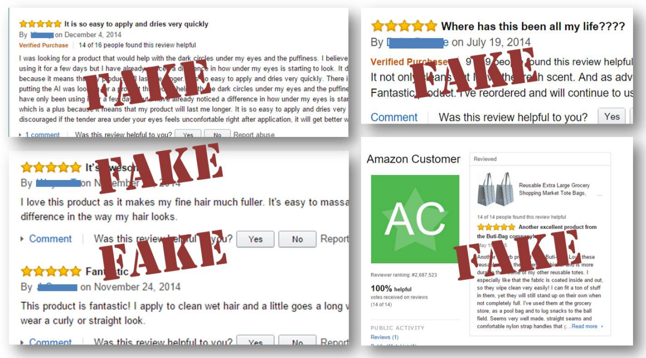


- Outliers: either interesting or can contaminate statistical analysis

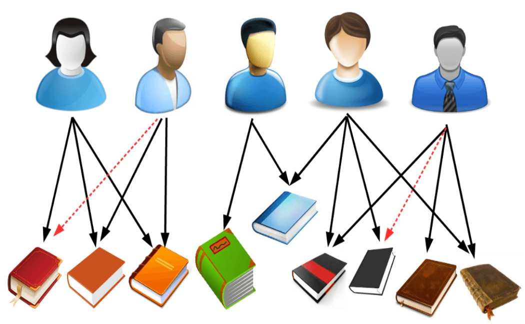
DATA POISONING

Fake Reviews [Mayzlin et al. ‘14]

So Many Misleading, “Fake” Reviews



Recommender Systems



[Li et al. ‘16]

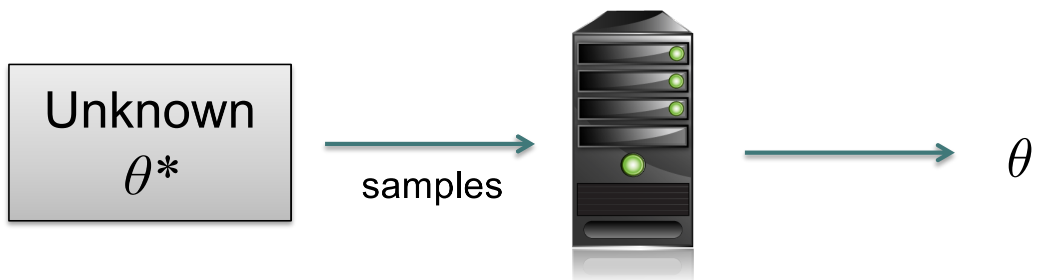
Crowdsourcing

[Wang et al. ‘14]

Malware/spam

[Nelson et al. ‘08]

The Statistical Learning Problem



- *Input*: sample generated by a **statistical model** with unknown θ^*
- *Goal*: estimate parameters θ so that $\theta \approx \theta^*$

Question 1: Is there an *efficient* learning algorithm?

Main performance criteria:

- Sample size
- Running time
- **Robustness**

Question 2: Are there *tradeoffs* between these criteria?

(Outlier-) Robustness in a Generative Model

Strong Contamination Model:

Let $\mathcal{F}$ be a family of statistical models.
We say that a set of N samples is ϵ -corrupted from $\mathcal{F}$ if it is generated as follows:

- N samples are drawn from an unknown $F \in \mathcal{F}$
- An omniscient adversary inspects these samples and changes arbitrarily an ϵ-fraction of them.

cf. Huber's contamination model [1964]

Several Models of Robustness

- Oblivious/Adaptive Adversary
- Additive/Subtractive/ Additive + Subtractive Adversary

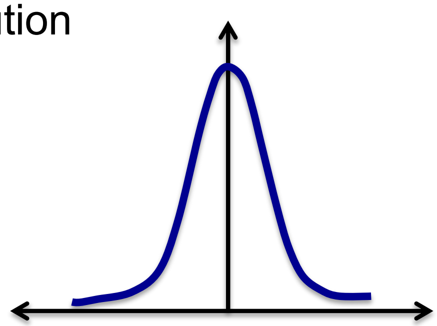
Power of Adversary	Oblivious	Adaptive
Additive Errors	Huber's Contamination Model $D = (1-\epsilon)F + \epsilon B$	Additive Contamination ("Data Poisoning")
Subtractive Errors	$F = (1-\epsilon)D + \epsilon L$	Subtractive Contamination
Additive and Subtractive	Hampel's Contamination $\mathrm{d}_{\mathrm{TV}}(D, F) \le \epsilon$	Strong Contamination

Example: Parameter Estimation

Given i.i.d. samples from an unknown distribution

e.g., a 1-D Gaussian

$$\mathcal{N}(\mu, \sigma^2)$$



how do we accurately estimate its parameters?

empirical mean:

$$\frac{1}{N}\sum_{i=1}^{N} X_i \to \mu$$

empirical variance:

$$\frac{1}{N}\sum_{i=1}^{N} (X_i - \overline{X})^2 \to \sigma^2$$

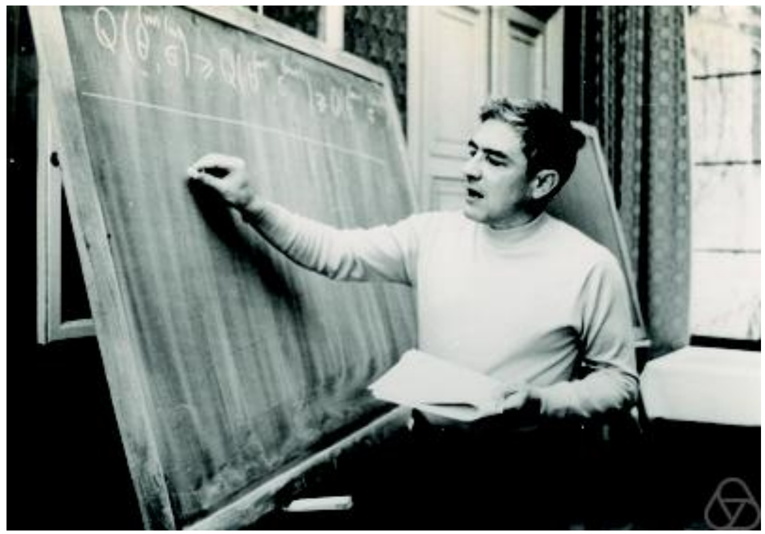
R. A. Fisher

Maximum Likelihood
(1920s)

J. W. Tukey

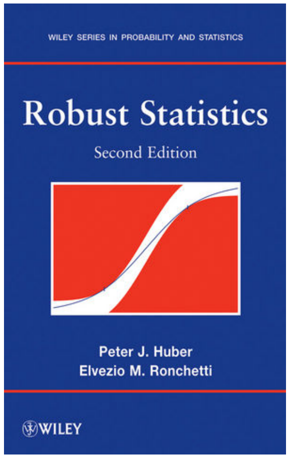
Model Misspecification ?
(1960s)

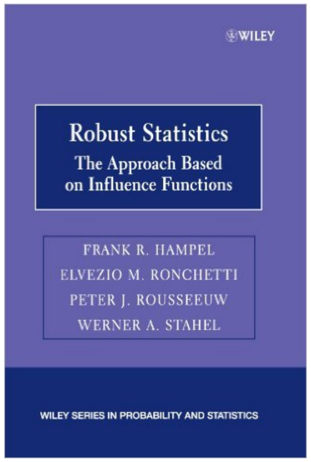
Peter J. Huber

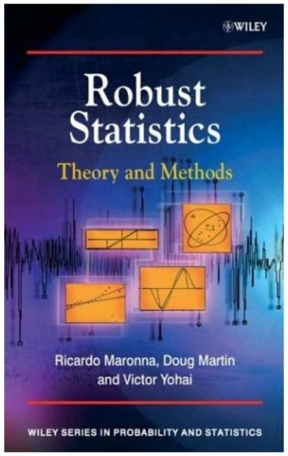


“Robust Estimation of a Location Parameter”
Annals of Mathematical Statistics, 1964.

Robust Statistics







What estimators behave well in the presence of outliers?

Robust Estimation: One Dimension

Given **corrupted** samples from a *one-dimensional* Gaussian, can we accurately estimate its parameters?

- A single corrupted sample can arbitrarily corrupt the empirical mean and variance
- But the **median** and **interquartile range** work

Fact [Folklore]: Given a set S of N ϵ-corrupted samples from a one-dimensional Gaussian

$$\mathcal{N}(\mu, \sigma^2)$$

with high constant probability we have that:

$$|\widehat{\mu} - \mu| \leq O\left(\epsilon + \sqrt{1/N}\right) \cdot \sigma$$

where $\widehat{\mu} = \mathrm{median}(S)$.

What about robust estimation in *high-dimensions*?

HIGH-DIMENSIONAL ROBUST MEAN ESTIMATION

Robust Mean Estimation: Given an ϵ - corrupted set of samples from an **unknown mean**, identity covariance Gaussian $\mathcal{N}(\mu, I)$ in d dimensions, recover $\widehat{\mu}$ with

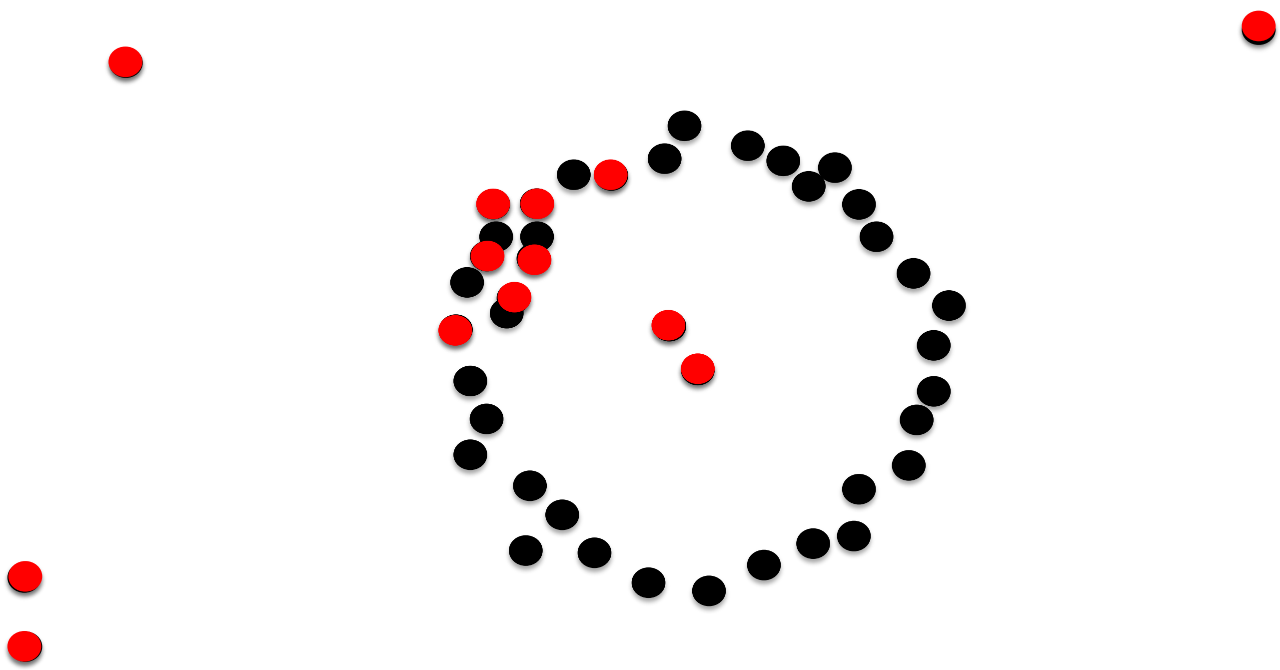
$$\|\widehat{\mu} - \mu\|_2 = O(\epsilon) + O(\sqrt{d/N})$$

Remark: Above convergence rate is optimal [Tukey'75, Donoho'82]

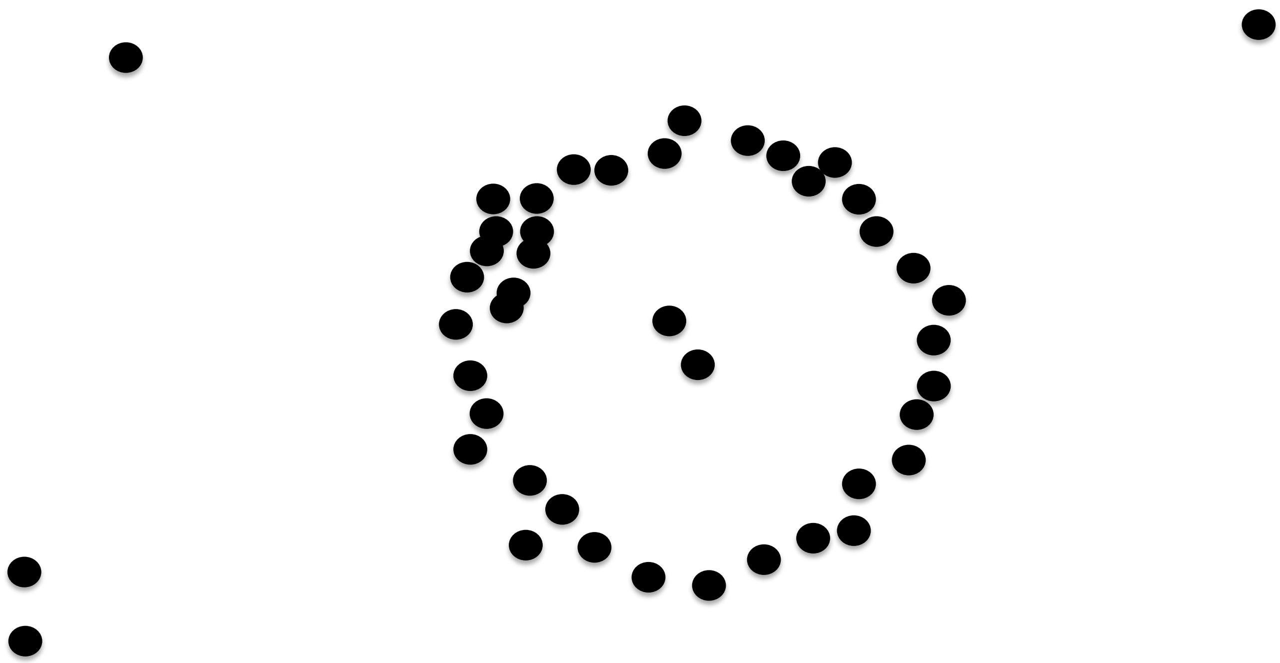
Previous Approaches: Robust Mean Estimation

Estimator	Error Rate	Running Time
Distance-Based Pruning	$\Theta(\epsilon\sqrt{d})$ ✗	$O(dN)$ ✓
Coordinate-wise Median	$\Theta(\epsilon\sqrt{d})$ ✗	$O(dN)$ ✓
Geometric Median	$\Theta(\epsilon\sqrt{d})$ ✗	$\mathrm{poly}(d, N)$ ✓
Tukey Median	$\Theta(\epsilon)$ ✓	NP-Hard ✗
Tournament	$\Theta(\epsilon)$ ✓	$N^{O(d)}$ ✗

DISTANCE-BASED PRUNING



DISTANCE-BASED PRUNING = NAÏVE OUTLIER REMOVAL

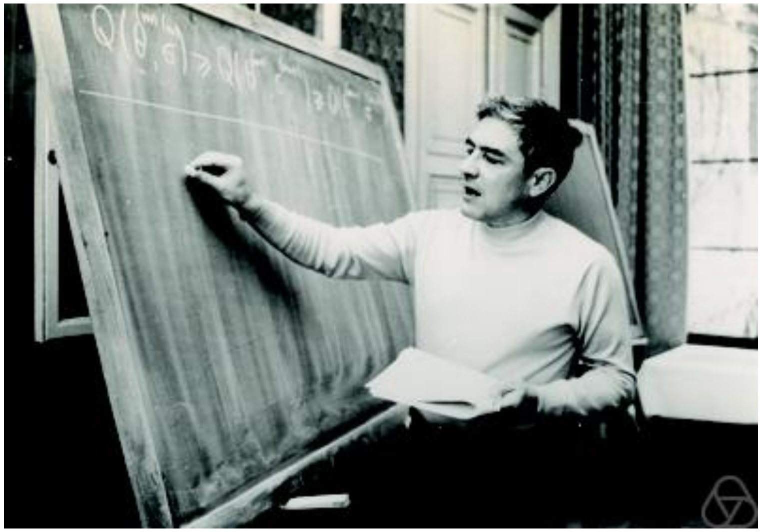


High-Dimensional Robust Statistics: 1960-2016

All known estimators are either **require exponential time to compute** or can tolerate a **negligible fraction of outliers**.

Is robust estimation *algorithmically* possible in high-dimensions?

Peter J. Huber, 1975



"[…] Only simple algorithms (i.e., with **a low degree of computational complexity**) will survive the onslaught of huge data sets. This runs counter to recent developments in computational robust statistics**. It appears to me that none of the above problems will be amenable to a treatment through theorems and proofs.** They will have to be attacked by heuristics and judgment, and by alternative "what if" analyses.[…]"

Robust Statistical Procedures, 1996*, Second Edition.*

Robust estimation in high-dimensions is algorithmically possible!

- Computationally efficient robust estimators that can tolerate a ***constant*** fraction of corruptions.
- Methodology to detect outliers in high dimensions.

Meta-Theorem (Informal): Can obtain *dimension-independent* error guarantees, if distribution on inliers has nice concentration.

First Algorithmic Progress In Unsupervised Setting

[D-Kamath-Kane-Li-Moitra-Stewart, FOCS'16/SICOMP19/CACM'20]

Can tolerate ***constant*** fraction of corruptions.

- Mean and Covariance Estimation
- Mixtures of Spherical Gaussians, Mixtures of Balanced Product Distributions

[**Lai-Rao-Vempala, FOCS'16]**

Can tolerate ***inverse logarithmic*** fraction of corruptions.

- Mean and Covariance Estimation
- Independent Component Analysis, SVD

SUBSEQUENT RELATED WORKS

- Sparse Models [Balakrishan-Du-Li-Singh’17, D-Karmalkar-Kane-Price-Stewart’19, Liu-Shen-Li-Caramanis’19,…]
- Graphical Models [Cheng-D-Kane-Stewart’18, D-Kane-Stewart-Sun’20]
- Robust Regression/Classification [D-Kane-Stewart’18, Klivans-Kothari-Meka’18, D-Kong-Stewart’19, …]
- Robust Stochastic Optimization [Prasad-Suggala-Balakrishnan-Ravikumar’18, D-Kamath-Kane-Li-Steinhard-Stewart’18, …]
- Robust Estimation via SoS [Hopkins-Li’18, Kothari-Steinhardt-Steurer’18, Karmalkar-Klivans-Kothari’19, Raghavendra-Yau’19, Bakshi-Kothari’20, D-Hopkins-Kane-Karmalkar’20, …]
- Near-Linear Time Algorithms [Chen-D-Ge’18, Cheng-D-Ge-Woodruff’19, Depersin-Lecue’19, Dong-Hopkins-Li’19, Li-Ye’20, Cherapanamjeri-Mohanty-Yau’20, …]
- Computational-Statistical Tradeoffs [D-Kane-Stewart’17, D-Kong-Stewart’19, Hopkins-Li’19, …]
- Connections to Non-Convex Optimization [Chen-D-Ge-Soltanolkotabi’20, Zhu-Jiao-Steinhardt’20]
- List-Decodable Learning [Charikar-Steinhardt-Valiant ’17, D-Kane-Stewart’18, Meister-Valiant’18, Karmalkar-Klivans-Kothari’19, Raghavendra-Yau’19, D-Kane-Koongsgard’20, …]
- Applications in Data Analysis [D-Kamath-Kane-Li-Moitra-Stewart’17, D-Kamath-Kane-Li-Steinhard-Stewart’18, …]

High-Dimensional Robust Mean Estimation

Robust Mean Estimation: Gaussian Case

Problem: Given an ϵ-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 mean μ of D.

Theorem 1: Let $\epsilon < 1/2$. If D is a spherical Gaussian, there is an efficient algorithm that outputs an estimate $\widehat{\mu}$ that with high probability satisfies

$$\|\widehat{\mu} - \mu\|_2 = O(\epsilon) + O(\sqrt{d/N})$$

in the **additive contamination** model.

First-term of RHS Independent of d !

[D-Kamath-Kane-Li-Moitra-Stewart, SODA'18]

Robust Mean Estimation: *Sub*-Gaussian Case

Problem: Given an ϵ-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 mean μ of D.

Theorem 2: Let $\epsilon < 1/2$. If D is a spherical *sub-Gaussian*, there is an efficient algorithm that outputs an estimate $\widehat{\mu}$ that with high probability satisfies

$$\|\widehat{\mu} - \mu\|_2 = O(\epsilon\sqrt{\log(1/\epsilon)}) + O(\sqrt{d/N}) \ .$$

in the **strong contamination** model.

Information-theoretically optimal error.

[D-Kamath-Kane-Li-Moitra-Stewart, FOCS'16, ICML'17]

Robust Mean Estimation: Bounded Covariance Case

Problem: Given an ϵ-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 mean μ of D.

Theorem 3: Let $\epsilon < 1/2$. If D has covariance $\Sigma \preceq \sigma^2 \cdot I$, there is an efficient algorithm that outputs an estimate $\widehat{\mu}$ that with high probability satisfies

$$\|\widehat{\mu} - \mu\|_2 = O(\sigma\sqrt{\epsilon}) + O(\sqrt{d/N}) \; .$$

in the **strong contamination** model.

Information-theoretically optimal error.

[D-Kamath-Kane-Li-Moitra-Stewart, ICML'17; Steinhardt, Charikar, Valiant, ITCS'18]

Robust Mean Estimation: Summary

Assumptions on Inliers	Information-Theoretic Bound	Computationally Efficient Estimators	Reference
Gaussian with $\Sigma = I$	$\Theta(\epsilon)$	$O(\epsilon)$	Additive Contamination* [DKKLMS, SODA'18]
Subgaussian with $\Sigma = I$	$\Theta(\epsilon\sqrt{\log(1/\epsilon)})$	$O(\epsilon\sqrt{\log(1/\epsilon)})$	[DKKLMS, FOCS'16]
Bounded t-th Moments $\Sigma = I$	$\Theta(\epsilon^{1-1/t})$	$O(\epsilon^{1-1/t})$	Folklore (see, e.g., survey [DK19])
Unknown Covariance $\Sigma \preceq I$	$\Theta(\sqrt{\epsilon})$	$O(\sqrt{\epsilon})$	[DKKLMS, ICML'17; SCV, ITCS'18]
Bounded t-th Moments	$\Theta(\epsilon^{1-1/t})$	$O(\epsilon^{1-1/t})$	"Niceness" Assumption* [HL, STOC'18; KS, STOC'18]