Loss minimization and parameter estimation with heavy tails

Daniel Hsu[⋆] Sivan Sabato[#†]

[⋆]Department of Computer Science, Columbia University

[#]Microsoft Research New England

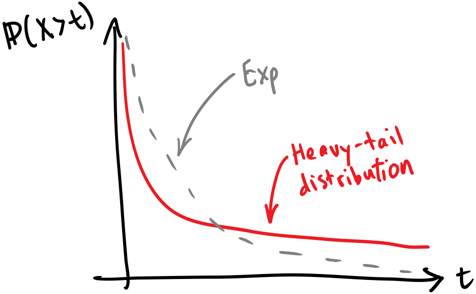
[†]On the job market—don't miss this amazing hiring opportunity!

Outline

1. Introduction

Heavy-tail distributions

Distribution with "tail" that is "heavier" than that of Exponential.



For random vectors, consider the distribution of $\|\boldsymbol{X}\|$.

Multivariate heavy-tail distributions

Heavy-tail distributions for random vectors $\boldsymbol{X} \in \mathbb{R}^d$:

- Marginal distributions of X_i have heavy tails, or
- Strong dependencies between the X_i.

Multivariate heavy-tail distributions

Heavy-tail distributions for random vectors $\boldsymbol{X} \in \mathbb{R}^d$:

- Marginal distributions of X_i have heavy tails, or
- Strong dependencies between the X_i.

Can we use the same procedures originally designed for distributions without heavy tails?

Or do we need new procedures?

Minimax optimal but not deviation optimal

Empirical mean achieves minimax rate for estimating $\mathbb{E}(X)$, but suboptimal when deviations are concerned:

Squared error of empirical mean is

$$\Omega\left(\frac{\sigma^2}{n\delta}\right)$$

with probability $\geq 2\delta$ *for some distribution.*
($n =$ *sample size,* $\sigma^2 = \mathrm{var}(X) < \infty$*.)*

Minimax optimal but not deviation optimal

Empirical mean achieves minimax rate for estimating $\mathbb{E}(X)$, but suboptimal when deviations are concerned:

Squared error of empirical mean is

$$\Omega\left(\frac{\sigma^2}{n\delta}\right)$$

with probability $\geq 2\delta$ for some distribution.
($n =$ sample size, $\sigma^2 = \mathrm{var}(X) < \infty$.)

Note: If data were Gaussian, squared error would be

$$O\left(\frac{\sigma^2 \log(1/\delta)}{n}\right).$$

Main result

New computationally efficient estimator for least squares linear regression when distributions of $\boldsymbol{X} \in \mathbb{R}^d$ and $Y \in \mathbb{R}$ may have heavy tails.

Main result

New computationally efficient estimator for least squares linear regression when distributions of $\boldsymbol{X} \in \mathbb{R}^d$ and $Y \in \mathbb{R}$ may have heavy tails.

Assuming bounded $(4 + \epsilon)$-order moments and regularity conditions, convergence rate is

$$O\left(\frac{\sigma^2 d \log(1/\delta)}{n}\right)$$

with probability $\geq 1 - \delta$ as soon as $n \geq \tilde{O}(d \log(1/\delta) + \log^2(1/\delta))$.

(n = sample size, σ^2 = optimal squared error.)

Main result

New computationally efficient estimator for least squares linear regression when distributions of $\boldsymbol{X} \in \mathbb{R}^d$ and $Y \in \mathbb{R}$ may have heavy tails.

Assuming bounded $(4+\epsilon)$-order moments and regularity conditions, convergence rate is

$$O\left(\frac{\sigma^2 d \log(1/\delta)}{n}\right)$$

with probability $\geq 1-\delta$ as soon as $n \geq \tilde{O}(d \log(1/\delta) + \log^2(1/\delta))$.

(n = sample size, σ^2 = optimal squared error.)

Previous state-of-the-art: [Audibert and Catoni, AoS 2011], essentially same conditions and rate, but computationally inefficient.

General technique with many other applications: ridge, Lasso, matrix approximation, *etc.*

2. Warm-up: estimating a scalar mean

Warm-up: estimating a scalar mean

Forget $\boldsymbol{X}$; how do we estimate $\mathbb{E}(Y)$?

(Set $\mu := \mathbb{E}(Y)$ and $\sigma^2 := \mathrm{var}(Y)$; assume $\sigma^2 < \infty$.)

Empirical mean

Let $Y_1, Y_2, \ldots, Y_n$ be iid copies of Y, and set

$$\widehat{\mu} \;:=\; \frac{1}{n}\sum_{i=1}^{n} Y_i$$

(empirical mean).

Empirical mean

Let $Y_1, Y_2, \ldots, Y_n$ be iid copies of Y, and set

$$\widehat{\mu} \;:=\; \frac{1}{n}\sum_{i=1}^{n} Y_i$$

(empirical mean).

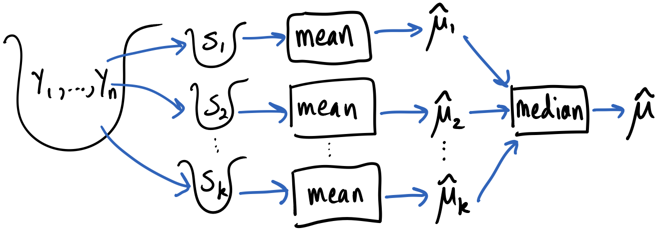
There exists distributions for Y with $\sigma^2 < \infty$ s.t.

$$\mathbb{P}\left((\widehat{\mu} - \mu)^2 \geq \frac{\sigma^2}{2n\delta}\,(1 - 2e\delta/n)^{n-1} \right) \;\geq\; 2\delta.$$

(Catoni, 2012)

Median-of-means

[Nemirovsky and Yudin, 1983; Alon, Matias, and Szegedy, JCSS 1999]



Median-of-means

[Nemirovsky and Yudin, 1983; Alon, Matias, and Szegedy, JCSS 1999]

1. Split the sample $\{Y_1, \ldots, Y_n\}$ into k parts $S_1, S_2, \ldots, S_k$ of equal size (say, randomly).
2. For each $i = 1, 2, \ldots, k$: set $\widehat{\mu}_i := \text{mean}(S_i)$.
3. Return $\widehat{\mu} := \text{median}(\{\widehat{\mu}_1, \widehat{\mu}_2, \ldots, \widehat{\mu}_k\})$.

Median-of-means

[Nemirovsky and Yudin, 1983; Alon, Matias, and Szegedy, JCSS 1999]

1. Split the sample $\{Y_1, \dots, Y_n\}$ into k parts $S_1, S_2, \dots, S_k$ of equal size (say, randomly).
2. For each $i = 1, 2, \dots, k$: set $\widehat{\mu}_i := \text{mean}(S_i)$.
3. Return $\widehat{\mu} := \text{median}(\{\widehat{\mu}_1, \widehat{\mu}_2, \dots, \widehat{\mu}_k\})$.

Theorem (Folklore)

Set $k := 4.5 \ln(1/\delta)$. *With probability at least* $1 - \delta$*,*

$$(\widehat{\mu} - \mu)^2 \;\leq\; O\left(\frac{\sigma^2 \log(1/\delta)}{n}\right).$$

Analysis of median-of-means

1. Assume $|S_i| = k/n$ for simplicity. By Chebyshev's inequality, for each $i = 1, 2, \ldots, k$:

$$\Pr\left(|\widehat{\mu}_i - \mu| \leq \sqrt{\frac{6\sigma^2 k}{n}} \right) \;\geq\; 5/6.$$

Analysis of median-of-means

1. Assume $|S_i| = k/n$ for simplicity. By Chebyshev's inequality, for each $i = 1, 2, \ldots, k$:

$$\Pr\left(|\widehat{\mu}_i - \mu| \leq \sqrt{\frac{6\sigma^2 k}{n}}\right) \;\geq\; 5/6.$$

2. Let $b_i := \mathbb{1}\{|\widehat{\mu}_i - \mu| \leq \sqrt{6\sigma^2 k/n}\}$. By Hoeffding's inequality,

$$\Pr\left(\sum_{i=1}^{k} b_i > k/2\right) \;\geq\; 1 - \exp(-k/4.5).$$

Analysis of median-of-means

1. Assume $|S_i| = k/n$ for simplicity. By Chebyshev's inequality, for each $i = 1, 2, \ldots, k$:

$$\Pr\left(|\widehat{\mu}_i - \mu| \leq \sqrt{\frac{6\sigma^2 k}{n}}\right) \;\geq\; 5/6.$$

2. Let $b_i := \mathbb{1}\{|\widehat{\mu}_i - \mu| \leq \sqrt{6\sigma^2 k/n}\}$. By Hoeffding's inequality,

$$\Pr\left(\sum_{i=1}^{k} b_i > k/2\right) \;\geq\; 1 - \exp(-k/4.5).$$

3. In the event that more than half of the $\widehat{\mu}_i$ are within $\sqrt{6\sigma^2 k/n}$ of μ, the median $\widehat{\mu}$ is as well.

Alternative: minimize a robust loss function

Alternative is to minimize a "robust" loss function [Catoni, 2012]:

$$\widehat{\mu} \;:=\; \arg\min_{\mu \in \mathbb{R}} \sum_{i=1}^{n} \ell\left(\frac{\mu - Y_i}{\sigma}\right).$$

Example: $\ell(z) := \log\cosh(z)$. Optimal rate and constants.

Catch: need to know σ^2.

3. Linear regression with heavy-tail distributions

Linear regression (for out-of-sample prediction)

1. **Response variable**: random variable $Y \in \mathbb{R}$.
2. **Covariates**: random vector $\boldsymbol{X} \in \mathbb{R}^d$.
   (Assume $\boldsymbol{\Sigma} := \mathbb{E}\boldsymbol{X}\boldsymbol{X}^\top \succ 0$.)
3. **Given**: Sample S of n iid copies of $(\boldsymbol{X}, Y)$.
4. **Goal**: find $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(S) \in \mathbb{R}^d$ to minimize population loss

$$L(\boldsymbol{\beta}) \ := \ \mathbb{E}(\boldsymbol{Y} - \boldsymbol{\beta}^\top \boldsymbol{X})^2.$$

Linear regression (for out-of-sample prediction)

1. **Response variable**: random variable $Y \in \mathbb{R}$.
2. **Covariates**: random vector $\boldsymbol{X} \in \mathbb{R}^d$.
   (Assume $\boldsymbol{\Sigma} := \mathbb{E}\boldsymbol{X}\boldsymbol{X}^\top \succ 0$.)
3. **Given**: Sample S of n iid copies of $(\boldsymbol{X}, Y)$.
4. **Goal**: find $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(S) \in \mathbb{R}^d$ to minimize population loss

$$L(\boldsymbol{\beta}) \ := \ \mathbb{E}(\boldsymbol{Y} - \boldsymbol{\beta}^\top \boldsymbol{X})^2.$$

Recall: Let $\boldsymbol{\beta}_\star := \arg\min_{\boldsymbol{\beta}' \in \mathbb{R}^d} L(\boldsymbol{\beta}')$. For any $\boldsymbol{\beta} \in \mathbb{R}^d$,

$$L(\boldsymbol{\beta}) - L(\boldsymbol{\beta}_\star) = \left\| \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}_\star) \right\|^2 =: \|\boldsymbol{\beta} - \boldsymbol{\beta}_\star\|_{\boldsymbol{\Sigma}}^2 .$$

Generalization of median-of-means

1. Split the sample S into k parts $S_1, S_2, \ldots, S_k$ of equal size (say, randomly).
2. For each $i = 1, 2, \ldots, k$: set $\widehat{\boldsymbol{\beta}}_i :=$ ordinary least squares(S_i).
3. Return $\widehat{\boldsymbol{\beta}} :=$ select good one$\left(\left\{\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \ldots, \widehat{\boldsymbol{\beta}}_k\right\}\right)$.

Generalization of median-of-means

1. Split the sample S into k parts $S_1, S_2, \ldots, S_k$ of equal size (say, randomly).
2. For each $i = 1, 2, \ldots, k$: set $\widehat{\boldsymbol{\beta}}_i :=$ ordinary least squares(S_i).
3. Return $\widehat{\boldsymbol{\beta}} :=$ select good one $\left(\left\{\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \ldots, \widehat{\boldsymbol{\beta}}_k\right\}\right)$.

Questions:

1. Guarantees for $\widehat{\boldsymbol{\beta}}_i = \mathrm{OLS}(S_i)$?
2. How to select a good $\widehat{\boldsymbol{\beta}}_i$?

Ordinary least squares

Under moment conditions*, $\widehat{\boldsymbol{\beta}}_i := \mathrm{OLS}(S_i)$ satisfies

$$\left\|\widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_\star\right\|_{\boldsymbol{\Sigma}} = O\left(\sqrt{\frac{\sigma^2 d}{|S_i|}}\right)$$

with probability at least $5/6$ as soon as $|S_i| \geq O(d \log d)$.**

* Requires Kurtosis condition for this simplified bound.

** Can replace $d \log d$ with d under some regularity conditions [Srivastava and Vershynin, AoP 2013].

Ordinary least squares

Under moment conditions*, $\widehat{\boldsymbol{\beta}}_i := \mathrm{OLS}(S_i)$ satisfies

$$\left\| \widehat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_\star \right\|_{\boldsymbol{\Sigma}} = O\left(\sqrt{\frac{\sigma^2 d}{|S_i|}} \right)$$

with probability at least $5/6$ as soon as $|S_i| \geq O(d \log d)$.**

Upshot: If $k := O(\log(1/\delta))$, then with probability $\geq 1 - \delta$, more than half of the $\widehat{\boldsymbol{\beta}}_i$ will be within $\varepsilon := \sqrt{\sigma^2 d \log(1/\delta)/n}$ of $\boldsymbol{\beta}_\star$.

* Requires Kurtosis condition for this simplified bound.

** Can replace $d \log d$ with d under some regularity conditions [Srivastava and Vershynin, AoP 2013].

Selecting a good $\widehat{\boldsymbol{\beta}}_i$ assuming $\boldsymbol{\Sigma}$ is known

Consider metric $\rho(\boldsymbol{a}, \boldsymbol{b}) := \|\boldsymbol{a} - \boldsymbol{b}\|_{\boldsymbol{\Sigma}}$.

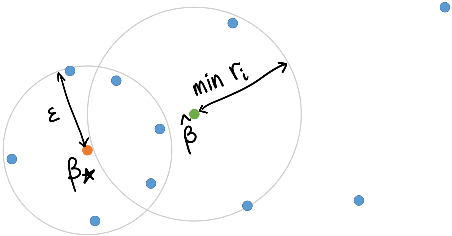
1. For each $i = 1, 2, \ldots, k$:
   Let $r_i := \text{median}\left\{\rho(\widehat{\boldsymbol{\beta}}_i, \widehat{\boldsymbol{\beta}}_j) : j = 1, 2, \ldots, k\right\}$.
2. Let $i_\star := \arg\min r_i$.
3. Return $\widehat{\boldsymbol{\beta}} := \widehat{\boldsymbol{\beta}}_{i_\star}$.

Selecting a good $\widehat{\boldsymbol{\beta}}_i$ assuming $\boldsymbol{\Sigma}$ is known

Consider metric $\rho(\boldsymbol{a}, \boldsymbol{b}) := \|\boldsymbol{a} - \boldsymbol{b}\|_{\boldsymbol{\Sigma}}$.

1. For each $i = 1, 2, \ldots, k$:
   Let $r_i := \text{median}\left\{\rho(\widehat{\boldsymbol{\beta}}_i, \widehat{\boldsymbol{\beta}}_j) : j = 1, 2, \ldots, k\right\}$.
2. Let $i_\star := \arg\min r_i$.
3. Return $\widehat{\boldsymbol{\beta}} := \widehat{\boldsymbol{\beta}}_{i_\star}$.

Claim: If more than half of the $\widehat{\boldsymbol{\beta}}_i$ are within distance ε of $\boldsymbol{\beta}_\star$, then $\widehat{\boldsymbol{\beta}}$ is within distance 3ε of $\boldsymbol{\beta}_\star$.



Selecting a good $\widehat{\boldsymbol{\beta}}_i$ when $\boldsymbol{\Sigma}$ is unknown

General case: $\boldsymbol{\Sigma}$ is unknown; can't compute distances $\|\boldsymbol{a} - \boldsymbol{b}\|_{\boldsymbol{\Sigma}}$.

Selecting a good $\widehat{\boldsymbol{\beta}}_i$ when $\boldsymbol{\Sigma}$ is <u>unknown</u>

General case: $\boldsymbol{\Sigma}$ is unknown; can't compute distances $\|\boldsymbol{a} - \boldsymbol{b}\|_{\boldsymbol{\Sigma}}$.

Solution: Estimate $\binom{k}{2}$ distances using fresh (unlabeled) samples.

Selecting a good $\widehat{\boldsymbol{\beta}}_i$ when $\boldsymbol{\Sigma}$ is <u>unknown</u>

General case: $\boldsymbol{\Sigma}$ is unknown; can't compute distances $\|\boldsymbol{a} - \boldsymbol{b}\|_{\boldsymbol{\Sigma}}$.

Solution: Estimate $\binom{k}{2}$ distances using fresh (unlabeled) samples.

- Only require constant fraction of these estimates to be accurate within constant multiplicative factors.
- Extra $O(k^2) = O(\log^2(1/\delta))$ (unlabeled) samples suffice.

Another interpretation: multiplicative approximation

With probability $\geq 1 - \delta$,

$$L(\widehat{\boldsymbol{\beta}}) \;\leq\; \left(1 + O\left(\frac{d\log(1/\delta)}{n}\right)\right) L(\boldsymbol{\beta}_\star)$$

(as soon as $n \geq \tilde{O}(d\log(1/\delta) + \log^2(1/\delta))$).

For instance, get 2-approximation with

$$n = \tilde{O}\left(d\log(1/\delta) + \log^2(1/\delta)\right)$$

—no dependence on $L(\boldsymbol{\beta}_\star)$.

(cf. [Mahdavi and Jin, COLT 2013].)

4. Concluding remarks

Concluding remarks

1. **This talk**: Linear regression with heavy-tail distributions in finite dimensions.
   Paper: Other applications (*e.g.*, ridge, Lasso, matrix approximation). `http://arxiv.org/abs/1307.1827`

Concluding remarks

1. **This talk**: Linear regression with heavy-tail distributions in finite dimensions.
   Paper: Other applications (*e.g.*, ridge, Lasso, matrix approximation). `http://arxiv.org/abs/1307.1827`

2. Simple algorithms + simple statistics:
   Avoid unnecessary assumptions made in statistical learning theory for classical problems.

Concluding remarks

1. **This talk**: Linear regression with heavy-tail distributions in finite dimensions.
   Paper: Other applications (*e.g.*, ridge, Lasso, matrix approximation). `http://arxiv.org/abs/1307.1827`

2. Simple algorithms + simple statistics:
   Avoid unnecessary assumptions made in statistical learning theory for classical problems.

3. Open questions:
   - Remove extraneous log factors?
   - Validation sets: not just for parameter tuning?

Concluding remarks

1. **This talk**: Linear regression with heavy-tail distributions in finite dimensions.
   Paper: Other applications (*e.g.*, ridge, Lasso, matrix approximation). `http://arxiv.org/abs/1307.1827`

2. Simple algorithms + simple statistics:
   Avoid unnecessary assumptions made in statistical learning theory for classical problems.

3. Open questions:
   - Remove extraneous log factors?
   - Validation sets: not just for parameter tuning?

Thanks!