Learning under Heavy-tailed Distributions

Lijun Zhang

Nanjing University, China

The 2nd International Symposium on Image Computing and Digital Medicine (ISICDM 2018)

Outline

Outline

1. Introduction
2. Related Work
3. Our Approach
4. Conclusion

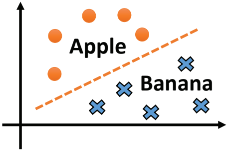
Machine Learning

Supervised Learning

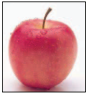

Apple

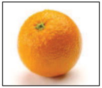

Banana

$$\begin{matrix}(\mathbf{x}_1, y_1) & \\ \cdots & \Longrightarrow y \approx h(\mathbf{x}) \\ (\mathbf{x}_n, y_n) & \end{matrix}$$

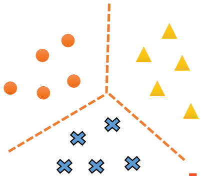


Unsupervised Learning





$$\begin{matrix}\mathbf{x}_1 & \\ \cdots & \Longrightarrow h(\mathbf{x}) \\ \mathbf{x}_n & \end{matrix}$$



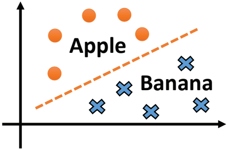
LAMDA Learning And Mining from DatA

Supervised Learning

- Input
  - Training data:
    $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$
  - A hypothesis class:
    $\mathcal{H} = \{h : \mathcal{X} \mapsto \mathbb{R}\}$

Supervised Learning

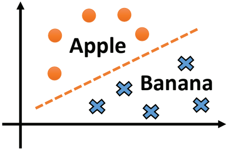
- Input
  - Training data: $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$
  - A hypothesis class: $\mathcal{H} = \{h : \mathcal{X} \mapsto \mathbb{R}\}$
- Output
  - A classifier: $h \in \mathcal{H}$



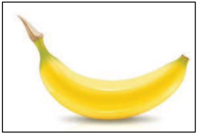
Supervised Learning

- Input
  - Training data: $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$
  - A hypothesis class: $\mathcal{H} = \{h : \mathcal{X} \mapsto \mathbb{R}\}$

- Output
  - A classifier: $h \in \mathcal{H}$



- Goal
  - Predict y by $h(\mathbf{x})$



⇨ ?

LAMDA Learning And Mining from DatA

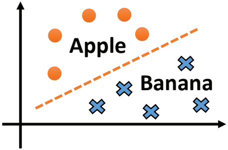
Supervised Learning

Input

- Training data: $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$
- A hypothesis class: $\mathcal{H} = \{h : \mathcal{X} \mapsto \mathbb{R}\}$

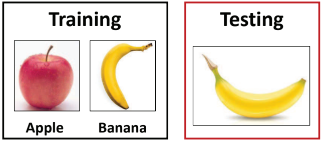
Output

- A classifier: $h \in \mathcal{H}$



Assumption

- I.I.D. sampled

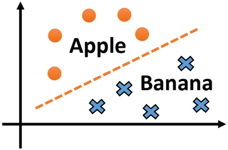


Goal

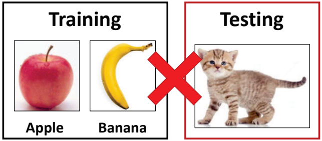
- Predict y by $h(\mathbf{x})$

Supervised Learning

- **Input**
  - Training data: $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$
  - A hypothesis class: $\mathcal{H} = \{h : \mathcal{X} \mapsto \mathbb{R}\}$
- **Output**
  - A classifier: $h \in \mathcal{H}$



- **Assumption**
  - I.I.D. sampled



- **Goal**
  - Predict y by $h(\mathbf{x})$

Mathematical Formulation

- Testing—Risk Minimization

$$\min_{h\in\mathcal{H}} \ell(h(\mathbf{x}), y)$$

- $\ell(\cdot,\cdot) : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ is certain loss
- E.g., $0-1$ loss, hinge loss, squared loss

Mathematical Formulation

- Testing—Risk Minimization

$$\min_{h \in \mathcal{H}} R(h) = \mathrm{E}_{(\mathbf{x},y)\sim\mathbb{D}}\left[\ell(h(\mathbf{x}), y)\right]$$

- $\ell(\cdot,\cdot) : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ is certain loss
- E.g., $0-1$ loss, hinge loss, squared loss

LAMDA Learning And Mining from DatA

Mathematical Formulation

Training—Empirical Risk Minimization (ERM)

$$\min_{h \in \mathcal{H}} \widehat{R}(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h(\mathbf{x}_i), y_i))$$

- $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ are sampled independently from $\mathbb{D}$

Testing—Risk Minimization

$$\min_{h \in \mathcal{H}} R(h) = \mathrm{E}_{(\mathbf{x},y)\sim\mathbb{D}}\left[\ell(h(\mathbf{x}), y)\right]$$

- $\ell(\cdot, \cdot) : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ is certain loss
- E.g., $0-1$ loss, hinge loss, squared loss

Mathematical Formulation

■ Training—Empirical Risk Minimization (ERM)

$$\min_{h \in \mathcal{H}} \widehat{R}(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(\mathbf{x}_i), y_i))$$

- $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ are sampled independently from $\mathbb{D}$

■ Testing—Risk Minimization

$$\min_{h \in \mathcal{H}} R(h) = \mathrm{E}_{(\mathbf{x},y)\sim\mathbb{D}}\left[\ell(h(\mathbf{x}), y)\right]$$

- $\ell(\cdot,\cdot) : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ is certain loss
- E.g., $0-1$ loss, hinge loss, squared loss

■ Examples—Least Squares

$$\min_{\mathbf{w} \in \mathcal{W}} \frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i^\top \mathbf{w} - \mathbf{y}_i)^2$$

- $\mathcal{W} = \{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\| \leq B\}$ is the domain

LAMDA Learning And Mining from DatA

Mathematical Formulation

Training—Empirical Risk Minimization (ERM)

$$\min_{h\in\mathcal{H}} \widehat{R}(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(\mathbf{x}_i), y_i))$$

- $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ are sampled independently from $\mathbb{D}$

Testing—Risk Minimization

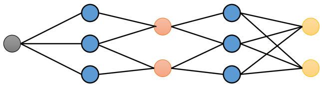
$$\min_{h\in\mathcal{H}} R(h) = \mathrm{E}_{(\mathbf{x},y)\sim\mathbb{D}}\left[\ell(h(\mathbf{x}), y)\right]$$

- $\ell(\cdot,\cdot) : \mathbb{R}\times\mathbb{R} \mapsto \mathbb{R}$ is certain loss
- E.g., $0-1$ loss, hinge loss, squared loss

Examples—Neural Networks

$$\min_{h\in\mathcal{H}} \widehat{R}(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(\mathbf{x}_i), y_i))$$

$\mathcal{H} = \{$



$\}$

Fundamentals of Supervised Learning

- Training—Empirical Risk Minimization (ERM)

$$\min_{h \in \mathcal{H}} \widehat{R}(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h(\mathbf{x}_i), y_i))$$

  - $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ are sampled independently

- Testing—Risk Minimization

$$\min_{h \in \mathcal{H}} R(h) = \mathrm{E}_{(\mathbf{x}, y) \sim \mathbb{D}} \left[\ell(h(\mathbf{x}), y) \right]$$

  - $\ell(\cdot, \cdot) : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ is certain loss

Optimization Theory

- Optimization Error

$$\widehat{R}(\hat{h}) - \min_{h \in \mathcal{H}} \widehat{R}(h)$$

Learning Theory

- Excess Risk

$$R(\hat{h}) - \min_{h \in \mathcal{H}} R(h)$$

Fundamentals of Supervised Learning

- Training—Empirical Risk Minimization (ERM)

$$\min_{h\in\mathcal{H}} \widehat{R}(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(\mathbf{x}_i), y_i))$$

  - $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ are sampled independently

Optimization Theory

- Optimization Error

$$\widehat{R}(\hat{h}) - \min_{h\in\mathcal{H}} \widehat{R}(h)$$

- Testing—Risk Minimization

$$\min_{h\in\mathcal{H}} R(h) = \mathrm{E}_{(\mathbf{x},y)\sim\mathbb{D}}\left[\ell(h(\mathbf{x}), y)\right]$$

  - $\ell(\cdot,\cdot) : \mathbb{R}\times\mathbb{R}\mapsto\mathbb{R}$ is certain loss

Learning Theory

- Excess Risk

$$R(\hat{h}) - \min_{h\in\mathcal{H}} R(h)$$

LAMDA Learning And Mining from DatA

Excess Risk of ERM

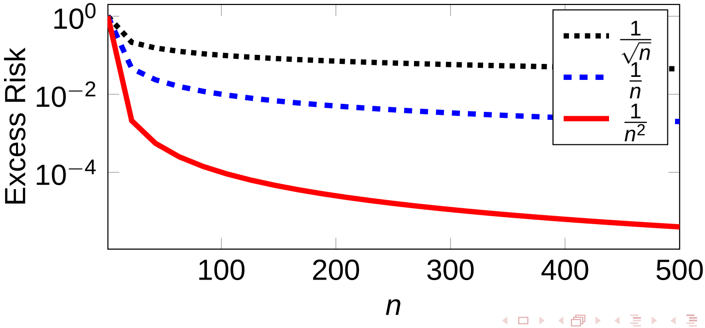
[Srebro et al., 2010]

[Bartlett and Mendelson, 2002] Smooth [Zhang et al., 2017]

$$R(\hat{h}) - R(h_*) \xrightarrow{\text{Lipschitz}} O\left(\frac{1}{\sqrt{n}}\right) \overset{\text{Smooth}}{\underset{\text{Strongly Convex}}{\rightrightarrows}} O\left(\frac{1}{n}\right) \xrightarrow[\text{Strongly Convex}]{\text{Smooth \&}} O\left(\frac{1}{n^2}\right)$$

Strongly Convex

[Sridharan et al., 2009]



Rationale of ERM

Bounded or Sub-Gaussian Distributions

Empirical risk is a good approximation of risk when the distribution is bounded or sub-Gaussian

Rationale of ERM

Bounded or Sub-Gaussian Distributions

Empirical risk is a good approximation of risk when the distribution is bounded or sub-Gaussian, i.e., for any $h \in \mathcal{H}$,

$$\left| \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell(h(\mathbf{x}_i), y_i))}_{\widehat{R}(h)} - \underbrace{\mathrm{E}_{(\mathbf{x},y)\sim\mathbb{D}}\left[\ell(h(\mathbf{x}), y)\right]}_{R(h)} \right| = O\left(\frac{1}{n^{\alpha}}\right)$$

- Losses: $\ell(h(\mathbf{x}_1), y_1)), \ldots, \ell(h(\mathbf{x}_n), y_n))$
- Predictions: $h(\mathbf{x}_1), \ldots, h(\mathbf{x}_n)$

LAMDA Learning And Mining from DatA

Rationale of ERM

Bounded or Sub-Gaussian Distributions

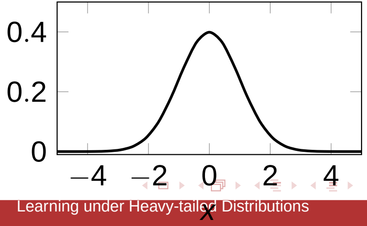
Empirical risk is a good approximation of risk when the distribution is bounded or sub-Gaussian, i.e., for any $h \in \mathcal{H}$,

$$\left| \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell(h(\mathbf{x}_i), y_i))}_{\widehat{R}(h)} - \underbrace{\mathrm{E}_{(\mathbf{x},y)\sim\mathbb{D}}\left[\ell(h(\mathbf{x}), y)\right]}_{R(h)} \right| = O\left(\frac{1}{n^{\alpha}}\right)$$

- Losses: $\ell(h(\mathbf{x}_1), y_1)), \ldots, \ell(h(\mathbf{x}_n), y_n))$
- Predictions: $h(\mathbf{x}_1), \ldots, h(\mathbf{x}_n)$

■ Sub-Gaussian Distributions

$$\mathbb{P}(|X| \geq t) \leq Ce^{-\nu t^2}$$



Heavy-tailed Distributions

- Heavy-tailed Distributions [Foss et al., 2013]

$$\int_{-\infty}^{\infty} e^{tx} dF(x) = \infty, \forall t > 0$$

where $F(x)$ is the distribution function

Learning And Mining from DatA

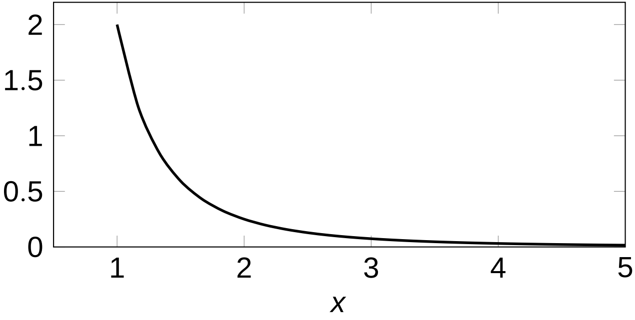
Heavy-tailed Distributions

- Heavy-tailed Distributions [Foss et al., 2013]

$$\int_{-\infty}^{\infty} e^{tx} dF(x) = \infty, \forall t > 0$$

where $F(x)$ is the distribution function

- Pareto distribution



- Long-tailed distribution

Heavy-tailed Distributions

- Heavy-tailed Distributions [Foss et al., 2013]

$$\int_{-\infty}^{\infty} e^{tx} dF(x) = \infty, \forall t > 0$$

where $F(x)$ is the distribution function

- Occur in Physics, Geoscience and Economics

Heavy-tailed Distributions

- Heavy-tailed Distributions [Foss et al., 2013]

$$\int_{-\infty}^{\infty} e^{tx} dF(x) = \infty, \forall t > 0$$

where $F(x)$ is the distribution function

  - Occur in Physics, Geoscience and Economics

- Learning under Heavy-tailed Distributions
  - ERM fails!

$$\left| \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell(h(\mathbf{x}_i), y_i))}_{\widehat{R}(h)} - \underbrace{\mathrm{E}_{(\mathbf{x},y)\sim\mathbb{D}}\left[\ell(h(\mathbf{x}), y)\right]}_{R(h)} \right| = ?$$

  - Truncated Minimization [Zhang and Zhou, 2018]

LAMDA Learning And Mining from DatA

Outline

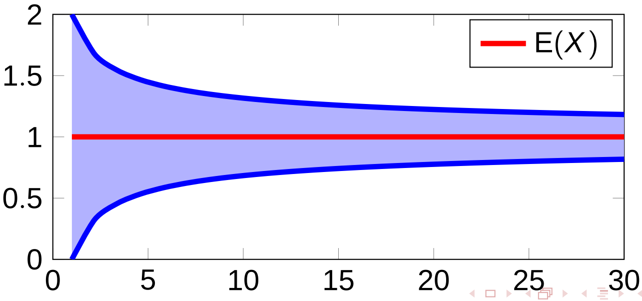
Bounded Distributions

- Estimation of the mean

Hoeffding's inequality [Lugosi, 2009]

Let $X_1, \ldots, X_n$ be independent random variables such that $|X_i| \leq C$. Then, with probability at least $1-\delta$,

$$\left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mathrm{E}[X] \right| \leq C \sqrt{\frac{2}{n} \log \frac{2}{\delta}}$$



Bounded Distributions

- Estimation of the mean

Hoeffding's inequality [Lugosi, 2009]

Let $X_1, \ldots, X_n$ be independent random variables such that $|X_i| \leq C$. Then, with probability at least $1 - \delta$,

$$\left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mathrm{E}[X] \right| \leq C \sqrt{\frac{2}{n} \log \frac{2}{\delta}}$$

- Empirical Risk Minimization (ERM)

$$\left| \underbrace{\frac{1}{n} \sum_{i=1}^{n} \ell(h(\mathbf{x}_i), y_i))}_{\widehat{R}(h)} - \underbrace{\mathrm{E}_{(\mathbf{x},y)\sim\mathbb{D}} \left[\ell(h(\mathbf{x}), y) \right]}_{R(h)} \right| = O\left(\frac{1}{\sqrt{n}} \right)$$

$$\Rightarrow \min_{h \in \mathcal{H}} \widehat{R}(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h(\mathbf{x}_i), y_i))$$

Heavy-tailed Distributions

- Robust estimation of the mean [Catoni, 2012]

$$\sum_{i=1}^{n} (X_i - \widehat{\theta}) = 0$$

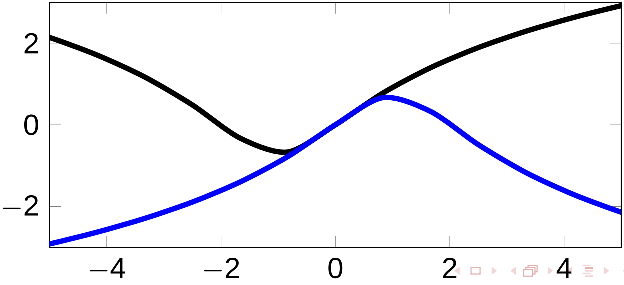
Heavy-tailed Distributions

- Robust estimation of the mean [Catoni, 2012]

$$\sum_{i=1}^{n} \psi\left[\alpha(X_i - \widehat{\theta})\right] = 0$$

$\alpha > 0$, and $\psi(\cdot) : \mathbb{R} \mapsto \mathbb{R}$ is non-decreasing

$$-\log\left(1 - x + \frac{x^2}{2}\right) \leq \psi(x) \leq \log\left(1 + x + \frac{x^2}{2}\right)$$



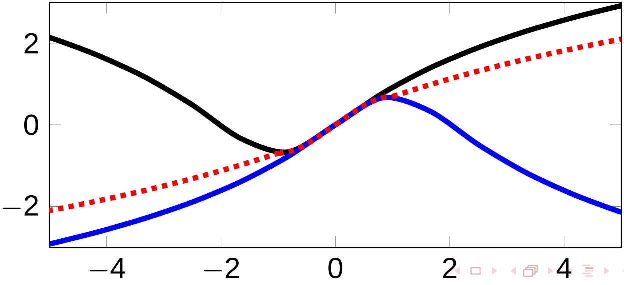
Heavy-tailed Distributions

- Robust estimation of the mean [Catoni, 2012]

$$\sum_{i=1}^{n} \psi\left[\alpha(X_i - \widehat{\theta})\right] = 0$$

$\alpha > 0$, and $\psi(\cdot) : \mathbb{R} \mapsto \mathbb{R}$ is non-decreasing

$$-\log\left(1 - x + \frac{x^2}{2}\right) \leq \psi(x) \leq \log\left(1 + x + \frac{x^2}{2}\right)$$



Heavy-tailed Distributions

- Robust estimation of the mean [Catoni, 2012]

$$\sum_{i=1}^{n} \psi\left[\alpha(X_i - \widehat{\theta})\right] = 0$$

$\alpha > 0$, and $\psi(\cdot) : \mathbb{R} \mapsto \mathbb{R}$ is non-decreasing

$$-\log\left(1 - x + \frac{x^2}{2}\right) \leq \psi(x) \leq \log\left(1 + x + \frac{x^2}{2}\right)$$

- $\widehat{\theta}$ is a good approximation of the mean

$$\left|\widehat{\theta} - \mathrm{E}[X]\right| = O\left(\sqrt{\frac{v}{n}}\right)$$

where $v = \mathrm{Var}(X)$

Robust ℓ_2-regression [Audibert and Catoni, 2011]

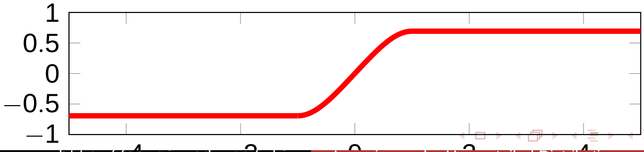
- Training Data
  - $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ where $\mathbf{x}_i \in \mathbb{R}^d$ and $\mathbf{y}_i \in \mathbb{R}$
  - Both $\mathbf{x}$ and y could be heavy-tailed

Robust ℓ_2-regression [Audibert and Catoni, 2011]

- Training Data
  - $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ where $\mathbf{x}_i \in \mathbb{R}^d$ and $\mathbf{y}_i \in \mathbb{R}$
  - Both $\mathbf{x}$ and y could be heavy-tailed
- Min-max Estimator

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{\mathbf{u}\in\mathcal{W}} \lambda \left(\|\mathbf{w}\|^2 - \|\mathbf{u}\|^2\right) + \frac{1}{\alpha n} \sum_{i=1}^{n} \psi\left[\alpha(y_i - \mathbf{w}^\top \mathbf{x}_i)^2 - \alpha(y_i - \mathbf{u}^\top \mathbf{x}_i)^2\right]$$

$$\psi(x) = \begin{cases} -\log\left(1 - x + \frac{x^2}{2}\right), 0 \leq x \leq 1; \\ \log(2), x \geq 1; \\ -\psi(-x), x \leq 0. \end{cases}$$



Robust ℓ_2-regression [Audibert and Catoni, 2011]

Training Data

- $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ where $\mathbf{x}_i \in \mathbb{R}^d$ and $\mathbf{y}_i \in \mathbb{R}$
- Both $\mathbf{x}$ and y could be heavy-tailed

Min-max Estimator

$$\min_{\mathbf{w}\in\mathcal{W}} \max_{\mathbf{u}\in\mathcal{W}} \lambda\left(\|\mathbf{w}\|^2 - \|\mathbf{u}\|^2\right) + \frac{1}{\alpha n}\sum_{i=1}^{n} \psi\left[\alpha(y_i - \mathbf{w}^\top \mathbf{x}_i)^2 - \alpha(y_i - \mathbf{u}^\top \mathbf{x}_i)^2\right]$$

Excess Risk

$$\mathrm{E}\left[(y - \widehat{\mathbf{w}}^\top \mathbf{x})^2\right] + \lambda\|\widehat{\mathbf{w}}\|^2 - \min_{\mathbf{w}\in\mathcal{W}} \left\{\mathrm{E}\left[(y - \mathbf{w}^\top \mathbf{x})^2\right] + \lambda\|\mathbf{w}\|^2\right\}$$

$$= O\left(\frac{d}{n}\right)$$

- Optimization is unclear

Learning with Heavy-tailed Losses [Brownlees et al., 2015]

- Input
  - n random variables $X_1, \ldots, X_n$
  - A functional space $\mathcal{F} = \{f : \mathcal{X} \mapsto \mathbb{R}\}$

Learning with Heavy-tailed Losses [Brownlees et al., 2015]

- Input
  - n random variables $X_1, \ldots, X_n$
  - A functional space $\mathcal{F} = \{f : \mathcal{X} \mapsto \mathbb{R}\}$
- Optimization Problem

$$\sum_{i=1}^{n} \psi\left[\alpha(X_i - \widehat{\theta})\right] = 0 \Rightarrow \begin{array}{ll} \min_{f \in \mathcal{F}} & \widehat{\theta}_f \\ \text{s. t.} & \frac{1}{n\alpha} \sum_{i=1}^{n} \psi\left[\alpha(f(X_i) - \widehat{\theta}_f)\right] = 0 \end{array}$$

Learning with Heavy-tailed Losses [Brownlees et al., 2015]

- Input
  - n random variables $X_1, \ldots, X_n$
  - A functional space $\mathcal{F} = \{f : \mathcal{X} \mapsto \mathbb{R}\}$
- Optimization Problem

$$\sum_{i=1}^{n} \psi\left[\alpha(X_i - \widehat{\theta})\right] = 0 \Rightarrow \begin{array}{ll} \min_{f \in \mathcal{F}} & \widehat{\theta}_f \\ \text{s. t.} & \frac{1}{n\alpha} \sum_{i=1}^{n} \psi\left[\alpha(f(X_i) - \widehat{\theta}_f)\right] = 0 \end{array}$$

- The theoretical guarantee is unsatisfying
  - Their risk bounds also hold for ERM
  - In most cases, they require the bounded assumption
  - Optimization is unclear

Outline

The Big Picture

- Supervised Learning under Heavy-tailed Distributions

Regression ($\mathbf{x}$ and y are heavy-tailed)	ℓ_2-regression $(\mathbf{x}^\top \mathbf{w} - y)^2$	ℓ_1-regression $\lvert \mathbf{x}^\top \mathbf{w} - y \rvert$	$\cdots$
Classification ($\mathbf{x}$ is heavy-tailed)	SVM $\max(0, 1 - y\mathbf{w}^\top \mathbf{x})$	Logistic Regression $\log\left(1 + e^{-y\mathbf{w}^\top \mathbf{x}}\right)$	$\cdots$

The Big Picture

- Supervised Learning under Heavy-tailed Distributions

Regression ($\mathbf{x}$ and y are heavy-tailed)	ℓ_2-regression $(\mathbf{x}^\top \mathbf{w} - y)^2$ [Audibert and Catoni, 2011]	ℓ_1-regression $\lvert\mathbf{x}^\top \mathbf{w} - y\rvert$?	$\cdots$
Classification ($\mathbf{x}$ is heavy-tailed)	SVM $\max(0, 1 - y\mathbf{w}^\top \mathbf{x})$?	Logistic Regression $\log\left(1 + e^{-y\mathbf{w}^\top \mathbf{x}}\right)$?	$\cdots$

LAMDA Learning And Mining from DatA

The Big Picture

- Supervised Learning under Heavy-tailed Distributions

Regression ($\mathbf{x}$ and y are heavy-tailed)	ℓ_2-regression $(\mathbf{x}^\top \mathbf{w} - y)^2$ [Audibert and Catoni, 2011]	ℓ_1-regression $\lvert \mathbf{x}^\top \mathbf{w} - y \rvert$ [Zhang and Zhou, 2018]	$\cdots$
Classification ($\mathbf{x}$ is heavy-tailed)	SVM $\max(0, 1 - y\mathbf{w}^\top \mathbf{x})$ [Zhang and Zhou, 2018]	Logistic Regression $\log\left(1 + e^{-y\mathbf{w}^\top \mathbf{x}}\right)$ [Zhang and Zhou, 2018]	$\cdots$

The Big Picture

- Supervised Learning under Heavy-tailed Distributions

Regression ($\mathbf{x}$ and y are heavy-tailed)	ℓ_2-regression $(\mathbf{x}^\top \mathbf{w} - y)^2$ [Audibert and Catoni, 2011]	ℓ_1-regression $\lvert\mathbf{x}^\top \mathbf{w} - y\rvert$ [Zhang and Zhou, 2018]	$\cdots$
Classification ($\mathbf{x}$ is heavy-tailed)	SVM $\max(0, 1 - y\mathbf{w}^\top \mathbf{x})$ [Zhang and Zhou, 2018]	Logistic Regression $\log\left(1 + e^{-y\mathbf{w}^\top \mathbf{x}}\right)$ [Zhang and Zhou, 2018]	$\cdots$

- Lipschitz Losses

$$|\ell(\mathbf{x}^\top \mathbf{w}, y) - \ell(\mathbf{x}^\top \mathbf{w}', y)| \leq |\mathbf{x}^\top \mathbf{w} - \mathbf{x}^\top \mathbf{w}'|$$

ℓ_1-regression under Heavy-tailed Distributions

- Training Data
  - $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ where $\mathbf{x}_i \in \mathbb{R}^d$ and $\mathbf{y}_i \in \mathbb{R}$
  - Both $\mathbf{x}$ and y could be heavy-tailed

ℓ_1-regression under Heavy-tailed Distributions

- Training Data
  - $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ where $\mathbf{x}_i \in \mathbb{R}^d$ and $\mathbf{y}_i \in \mathbb{R}$
  - Both $\mathbf{x}$ and y could be heavy-tailed

- Traditional ERM

$$\min_{\mathbf{w} \in \mathcal{W}} \frac{1}{n} \sum_{i=1}^{n} |y_i - \mathbf{x}_i^\top \mathbf{w}|$$

ℓ_1-regression under Heavy-tailed Distributions

- Training Data
  - $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ where $\mathbf{x}_i \in \mathbb{R}^d$ and $\mathbf{y}_i \in \mathbb{R}$
  - Both $\mathbf{x}$ and y could be heavy-tailed

- Traditional ERM

$$\min_{\mathbf{w}\in\mathcal{W}} \frac{1}{n}\sum_{i=1}^{n} |y_i - \mathbf{x}_i^\top \mathbf{w}|$$

- Truncated Minimization

$$\min_{\mathbf{w}\in\mathcal{W}} \frac{1}{n\alpha}\sum_{i=1}^{n} \psi\left(\alpha |y_i - \mathbf{x}_i^\top \mathbf{w}|\right)$$

$\alpha > 0$, and $\psi(\cdot) : \mathbb{R} \mapsto \mathbb{R}$ is non-decreasing

$$-\log\left(1 - x + \frac{x^2}{2}\right) \leq \psi(x) \leq \log\left(1 + x + \frac{x^2}{2}\right)$$

LAMDA Learning And Mining from DatA

Theoretical Guarantees

- Truncated Minimization for ℓ_1-regression

$$\min_{\mathbf{w}\in\mathcal{W}} \frac{1}{n\alpha}\sum_{i=1}^{n}\psi\left(\alpha|y_i - \mathbf{x}_i^\top \mathbf{w}|\right)$$

- Excess Risk

$$\mathrm{E}\left[|y-\widehat{\mathbf{w}}^\top\mathbf{x}|\right] - \min_{\mathbf{w}\in\mathcal{W}}\mathrm{E}\left[|y-\mathbf{w}^\top\mathbf{x}|\right] = O\left(\sqrt{\frac{d}{n}}\right)$$

Theoretical Guarantees

- Truncated Minimization for ℓ_1-regression

$$\min_{\mathbf{w}\in\mathcal{W}} \frac{1}{n\alpha}\sum_{i=1}^{n}\psi\left(\alpha|y_i-\mathbf{x}_i^\top\mathbf{w}|\right)$$

- Excess Risk

$$\mathrm{E}\left[|y-\widehat{\mathbf{w}}^\top\mathbf{x}|\right]-\min_{\mathbf{w}\in\mathcal{W}}\mathrm{E}\left[|y-\mathbf{w}^\top\mathbf{x}|\right]=O\left(\sqrt{\frac{d}{n}}\right)$$

- Min-max Estimator for ℓ_2-regression

[Audibert and Catoni, 2011]

$$\min_{\mathbf{w}\in\mathcal{W}}\max_{\mathbf{u}\in\mathcal{W}}\lambda\left(\|\mathbf{w}\|^2-\|\mathbf{u}\|^2\right)+\frac{1}{\alpha n}\sum_{i=1}^{n}\psi\left[\alpha(y_i-\mathbf{w}^\top\mathbf{x}_i)^2-\alpha(y_i-\mathbf{u}^\top\mathbf{x}_i)^2\right]$$

- Excess Risk

$$\mathrm{E}\left[(y-\widehat{\mathbf{w}}^\top\mathbf{x})^2\right]+\lambda\|\widehat{\mathbf{w}}\|^2-\min_{\mathbf{w}\in\mathcal{W}}\left\{\mathrm{E}\left[(y-\mathbf{w}^\top\mathbf{x})^2\right]+\lambda\|\mathbf{w}\|^2\right\}=O\left(\frac{d}{n}\right)$$

ℓ_1-regression with Bounded Features

- Training Data
  - $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ where $\mathbf{x}_i \in \mathbb{R}^d$ and $\mathbf{y}_i \in \mathbb{R}$
  - $\mathbf{x}$ is bounded and y could be heavy-tailed

LAMDA Learning And Mining from DatA

ℓ_1-regression with Bounded Features

- Training Data
  - $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ where $\mathbf{x}_i \in \mathbb{R}^d$ and $\mathbf{y}_i \in \mathbb{R}$
  - $\mathbf{x}$ is bounded and y could be heavy-tailed
- Traditional ERM

$$\min_{\mathbf{w}\in\mathcal{W}} \frac{1}{n}\sum_{i=1}^{n} |y_i - \mathbf{x}_i^\top \mathbf{w}|$$

- Excess Risk

$$\mathrm{E}\left[|y - \widehat{\mathbf{w}}^\top \mathbf{x}|\right] - \min_{\mathbf{w}\in\mathcal{W}} \mathrm{E}\left[|y - \mathbf{w}^\top \mathbf{x}|\right] = O\left(\frac{D}{\sqrt{n}}\right)$$

where $\|\mathbf{x}\|_2 \leq D$

LAMDA Learning And Mining from DatA

Experimental Setting

- Truncated Minimization Problem

$$\min_{\mathbf{w}\in\mathcal{W}} \widehat{R}_\psi(\mathbf{w}) = \frac{1}{n\alpha}\sum_{i=1}^{n}\psi(\alpha|y_i - \mathbf{x}_i^\top \mathbf{w}|)$$

  - Sum of quasiconvex functions

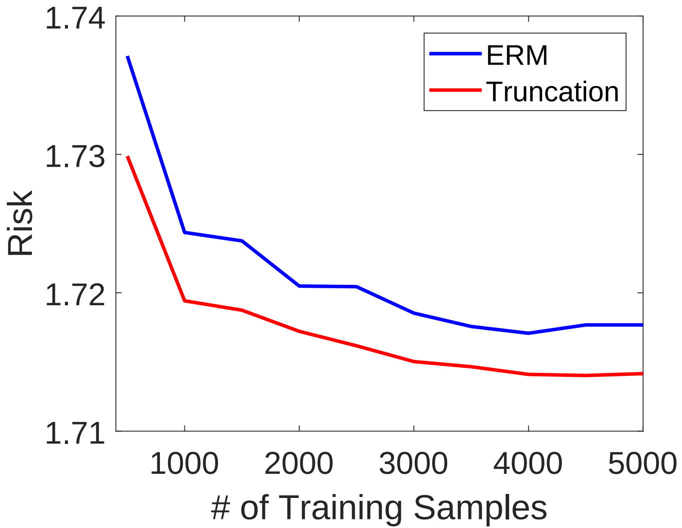
- Normalized Gradient Descent (NGD)

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\frac{\partial\widehat{R}_\psi(\mathbf{w}_t)}{\|\partial\widehat{R}_\psi(\mathbf{w}_t)\|_2}$$

- Data Sets
  - Both feature and label are heavy-tailed
  - Feature is bounded, and label is heavy-tailed
  - Both feature and label are bounded

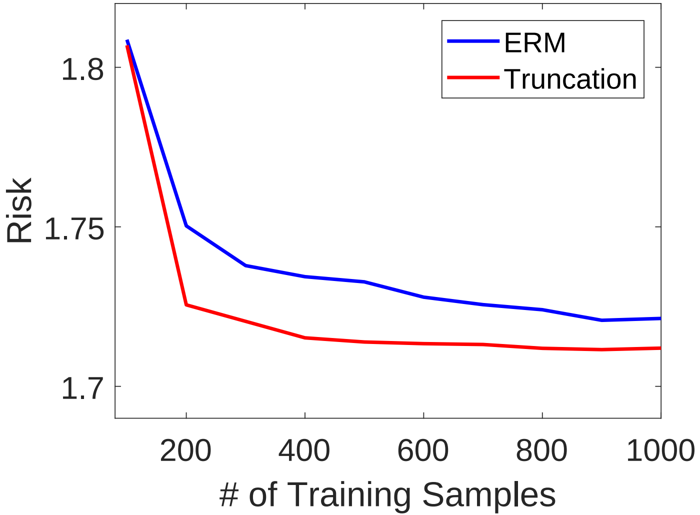
Heavy-tailed Feature and Label

- t-distribution and $W = \mathrm{sign}(V)/|V|^{1/2.01}$, $V \sim \mathcal{N}(0,1)$



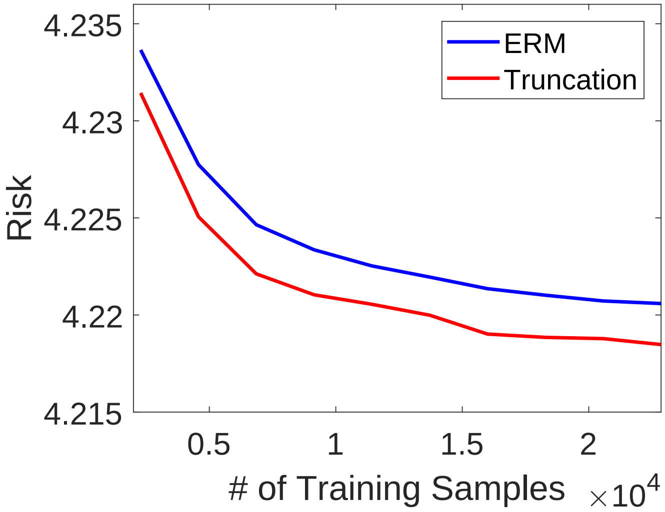
Bounded Feature and Heavy-tailed Label

- $W = \text{sign}(V)/|V|^{1/2.01}$, $V \sim \mathcal{N}(0,1)$



Bounded Feature and Label

- CASP dataset from UCI



Outline

Conclusion and Future Work

Conclusion

- Truncated Minimization for Heavy-tailed Distributions

$$\min_{\mathbf{w}\in\mathcal{W}} \frac{1}{n\alpha}\sum_{i=1}^{n}\psi(\alpha|y_i - \mathbf{x}_i^\top \mathbf{w}|)$$

- Learning Theory (Excess Risk)

$$\mathrm{E}\left[|y - \widehat{\mathbf{w}}^\top \mathbf{x}|\right] - \min_{\mathbf{w}\in\mathcal{W}} \mathrm{E}\left[|y - \mathbf{w}^\top \mathbf{x}|\right] = O\left(\sqrt{\frac{d}{n}}\right)$$

Future Work

- Optimization Theory for the Non-convex Problem
- Median-of-means Approaches [Hsu and Sabato, 2014]

Reference I

Thanks!

Audibert, J.-Y. and Catoni, O. (2011).
Robust linear least squares regression.
The Annals of Statistics, 39(5):2766–2794.

Bartlett, P. L. and Mendelson, S. (2002).
Rademacher and gaussian complexities: risk bounds and structural results.
Journal of Machine Learning Research, 3:463–482.

Brownlees, C., Joly, E., and Lugosi, G. (2015).
Empirical risk minimization for heavy-tailed losses.
The Annals of Statistics, 43(6):2507–2536.

Catoni, O. (2012).
Challenging the empirical mean and empirical variance: A deviation study.
Annales de l'Institut Henri Poincaré, Probabilités et Statistiques, 48(4):1148–1185.

Foss, S., Korshunov, D., and Zachary, S. (2013).
An Introduction to Heavy-Tailed and Subexponential Distributions.
Springer.

Hsu, D. and Sabato, S. (2014).
Heavy-tailed regression with a generalized median-of-means.
In *Proceedings of the 31st International Conference on Machine Learning*, pages 37–45.

Lugosi, G. (2009).
Concentration-of-measure inequalities.
Technical report, Department of Economics, Pompeu Fabra University.

Reference II

Srebro, N., Sridharan, K., and Tewari, A. (2010).
Optimistic rates for learning with a smooth loss.
ArXiv e-prints, arXiv:1009.3896.

Sridharan, K., Shalev-shwartz, S., and Srebro, N. (2009).
Fast rates for regularized objectives.
In *Advances in Neural Information Processing Systems 21*, pages 1545–1552.

Zhang, L., Yang, T., and Jin, R. (2017).
Empirical risk minimization for stochastic convex optimization: $O(1/n)$- and $O(1/n^2)$-type of risk bounds.
In *Proceedings of the 30th Annual Conference on Learning Theory*, pages 1954–1979.

Zhang, L. and Zhou, Z.-H. (2018).
ℓ_1-regression with heavy-tailed distributions.
In *Advances in Neural Information Processing Systems 31*.