LAMDA Learning And Mining from DatA

南京大学 NANJING UNIVERSITY

Fast Rates for Empirical Risk Minimization: Beyond the $O(1/n)$ Risk Bound

Lijun Zhang

LAMDA group, Nanjing University, China

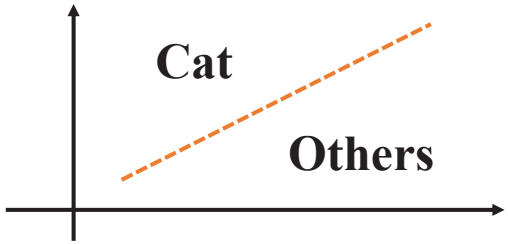
The 16th China Conference on Machine Learning (CCML 2017)

Outline

Outline

Statistical Machine Learning

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



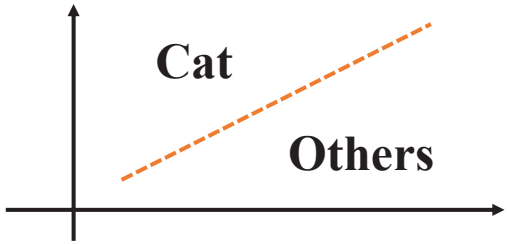
Statistical Machine Learning

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

$$h(\mathbf{x}) \approx y$$

  - where $(\mathbf{x}, y)$ is a testing pair

Cat

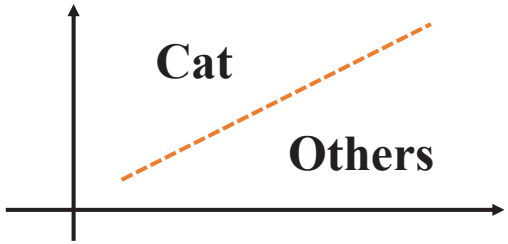


?

Statistical Machine Learning

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

Cat



- Output: $h \in \mathcal{H}$
- Goal—Prediction

$$h(\mathbf{x}) \approx y$$

  - where $(\mathbf{x}, y)$ is a testing pair

?

- Statistical Assumption
  - Training and testing data are sampled independently from $\mathbb{D}$

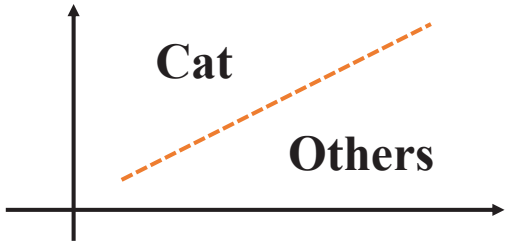
LAMDA Learning And Mining from DatA

Statistical Machine Learning

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

Cat

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



- Goal—Risk Minimization (RM)

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

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

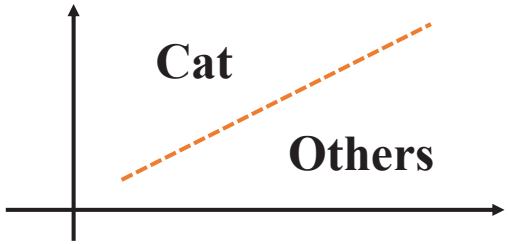
?

- Statistical Assumption
  - Training and testing data are sampled independently from $\mathbb{D}$

LAMDA Learning And Mining from DatA

Statistical Machine Learning

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



- Goal—Risk Minimization (RM)

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

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

?

- Statistical Assumption
  - Training and testing data are sampled independently from $\mathbb{D}$

LAMDA Learning And Mining from DatA

Empirical Risk Minimization (ERM)

- Risk Minimization (RM)

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

  - The distribution $\mathbb{D}$ is unknown

Empirical Risk Minimization (ERM)

- Risk Minimization (RM)

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

  - The distribution $\mathbb{D}$ is unknown

- Empirical Risk Minimization (ERM)

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

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

Empirical Risk Minimization (ERM)

- Risk Minimization (RM)

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

  - The distribution $\mathbb{D}$ is unknown

- Empirical Risk Minimization (ERM)

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

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

- Examples—Least Squares

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

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

LAMDA Learning And Mining from DatA

Empirical Risk Minimization (ERM)

- Risk Minimization (RM)

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

  - The distribution $\mathbb{D}$ is unknown

- Empirical Risk Minimization (ERM)

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

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

- Examples—Support Vector Machine (SVM)

$$\min_{\mathbf{w} \in \mathbb{R}^d} \widehat{F}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{x}_i^\top \mathbf{w}, y_i) + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

  - $\ell(u, v) = \max(0, 1 - uv)$ is the hinge loss

LAMDA Learning And Mining from DatA

Theoretical Guarantees of ERM

Foundational Problem of Statistical Learning Theory

Theoretical Guarantees of ERM

Foundational Problem of Statistical Learning Theory

- Generalization Error

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

  - where $\hat{h} = \mathrm{argmin}_{h\in\mathcal{H}}\,\widehat{F}(h)$ is the empirical minimizer

Theoretical Guarantees of ERM

Foundational Problem of Statistical Learning Theory

- Generalization Error

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

  - where $\hat{h} = \mathrm{argmin}_{h\in\mathcal{H}}\,\widehat{F}(h)$ is the empirical minimizer

- Excess Risk

$$F(\hat{h}) - \min_{\mathbf{w}\in\mathcal{W}} F(h) = F(\hat{h}) - F(h_*)$$

$$= \mathrm{E}_{(\mathbf{x},y)\sim\mathbb{D}}\left[\ell(\hat{h}(\mathbf{x}), y)\right] - \mathrm{E}_{(\mathbf{x},y)\sim\mathbb{D}}\left[\ell(h_*(\mathbf{x}), y)\right]$$

  - where $h_* = \mathrm{argmin}_{h\in\mathcal{H}}\,F(h)$ is the optimal solution

LAMDA Learning And Mining from DatA

Outline

Stochastic Optimization

- Stochastic Optimization

$$\min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w}) = \mathrm{E}_{f\sim\mathbb{P}}\left[f(\mathbf{w})\right]$$

  - $f(\cdot) : \mathcal{W} \mapsto \mathbb{R}$ is a random function sampled from $\mathbb{P}$

Stochastic Optimization

- Stochastic Optimization

$$\min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w}) = \mathrm{E}_{f\sim\mathbb{P}}\left[f(\mathbf{w})\right]$$

  - $f(\cdot) : \mathcal{W} \mapsto \mathbb{R}$ is a random function sampled from $\mathbb{P}$

- Examples—Risk Minimization (RM)

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

  - $h \to \mathbf{w}$, $\mathcal{H} \to \mathcal{W}$, $\ell(h(\mathbf{x}), y) \to f(\mathbf{w})$

Stochastic Optimization

- Stochastic Optimization

$$\min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w}) = \mathrm{E}_{f\sim\mathbb{P}}\left[f(\mathbf{w})\right]$$

  - $f(\cdot) : \mathcal{W} \mapsto \mathbb{R}$ is a random function sampled from $\mathbb{P}$

- Examples—Risk Minimization (RM)

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

  - $h \to \mathbf{w}$, $\mathcal{H} \to \mathcal{W}$, $\ell(h(\mathbf{x}), y) \to f(\mathbf{w})$

- Examples—Newsvendor Problem (报童问题)

$$\max_{x} F(x) = \mathrm{E}_{\xi\sim\mathbb{D}}\left[p\min(x,\xi) - cx\right]$$

  - x is the supply, ξ is the random demand
  - p is the price, c is the cost

LAMDA Learning And Mining from DatA

Sample Average Approximation (SAA)

- Stochastic Optimization

$$\min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w}) = \mathrm{E}_{f\sim\mathbb{P}}\left[f(\mathbf{w})\right]$$

  - $f(\cdot) : \mathcal{W} \mapsto \mathbb{R}$ is a random function sampled from $\mathbb{P}$

- Sample Average Approximation (SAA)

$$\min_{\mathbf{w}\in\mathcal{W}} \widehat{F}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{w})$$

  - $f_1, \ldots, f_n$ are sampled independently from $\mathbb{P}$

Sample Average Approximation (SAA)

- Stochastic Optimization

$$\min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w}) = \mathrm{E}_{f\sim\mathbb{P}}\left[f(\mathbf{w})\right]$$

  - $f(\cdot) : \mathcal{W} \mapsto \mathbb{R}$ is a random function sampled from $\mathbb{P}$

- Sample Average Approximation (SAA)

$$\min_{\mathbf{w}\in\mathcal{W}} \widehat{F}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{w})$$

  - $f_1, \ldots, f_n$ are sampled independently from $\mathbb{P}$

- Examples—Newsvendor Problem (报童问题)

$$\max_{x} \widehat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \left(p\min(x, \xi_i) - cx\right)$$

  - $\xi_1, \ldots, \xi_n$ are i.i.d. samples

LAMDA Learning And Mining from DatA

Sample Average Approximation (SAA)

- Stochastic Optimization

$$\min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w}) = \mathrm{E}_{f\sim\mathbb{P}}\left[f(\mathbf{w})\right]$$

  - $f(\cdot) : \mathcal{W} \mapsto \mathbb{R}$ is a random function sampled from $\mathbb{P}$

- Sample Average Approximation (SAA)

$$\min_{\mathbf{w}\in\mathcal{W}} \widehat{F}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{w})$$

  - $f_1, \ldots, f_n$ are sampled independently from $\mathbb{P}$

- Excess Risk

$$F(\widehat{\mathbf{w}}) - \min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w}) = F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*)$$

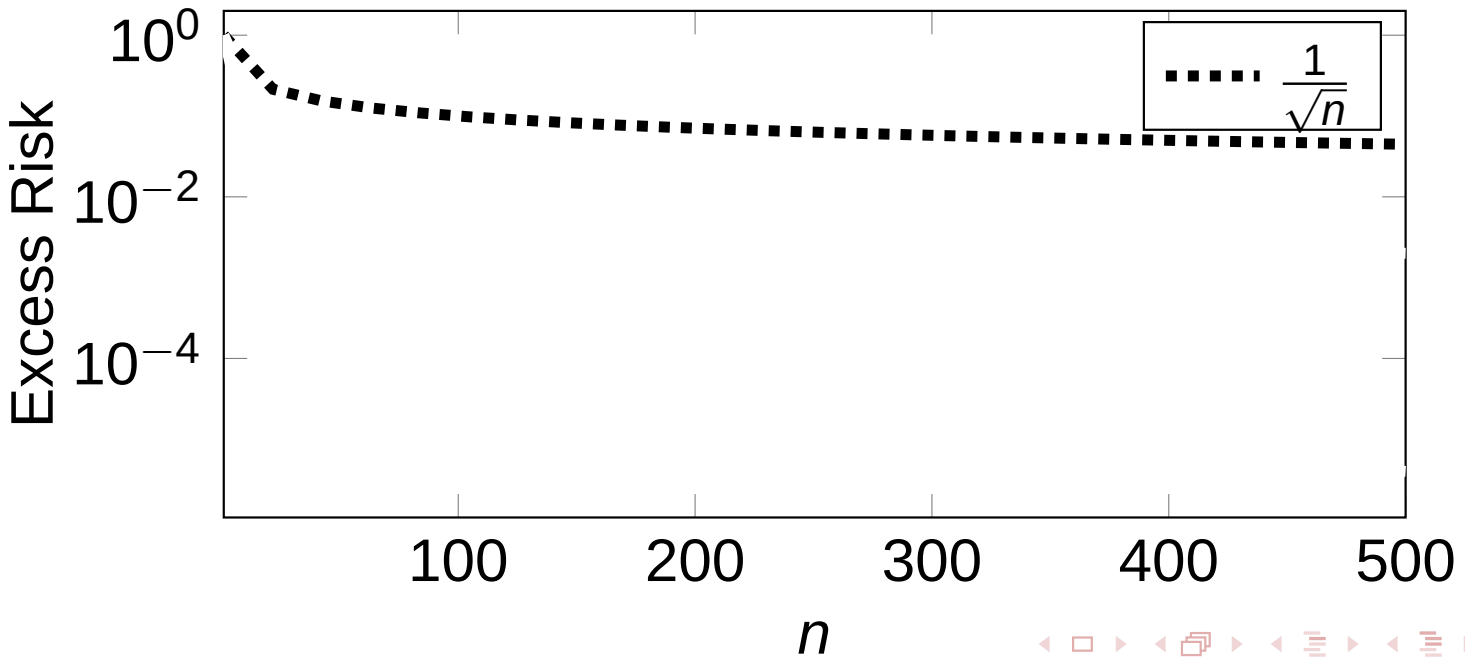
  - $\widehat{\mathbf{w}} = \mathrm{argmin}_{\mathbf{w}\in\mathcal{W}} \widehat{F}(\mathbf{w})$ is the empirical minimizer
  - $\mathbf{w}_* = \mathrm{argmin}_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w})$ is the optimal solution

LAMDA Learning And Mining from DatA

Our Contributions [Zhang et al., 2017]

- ERM/SAA for Statistical Machine Learning

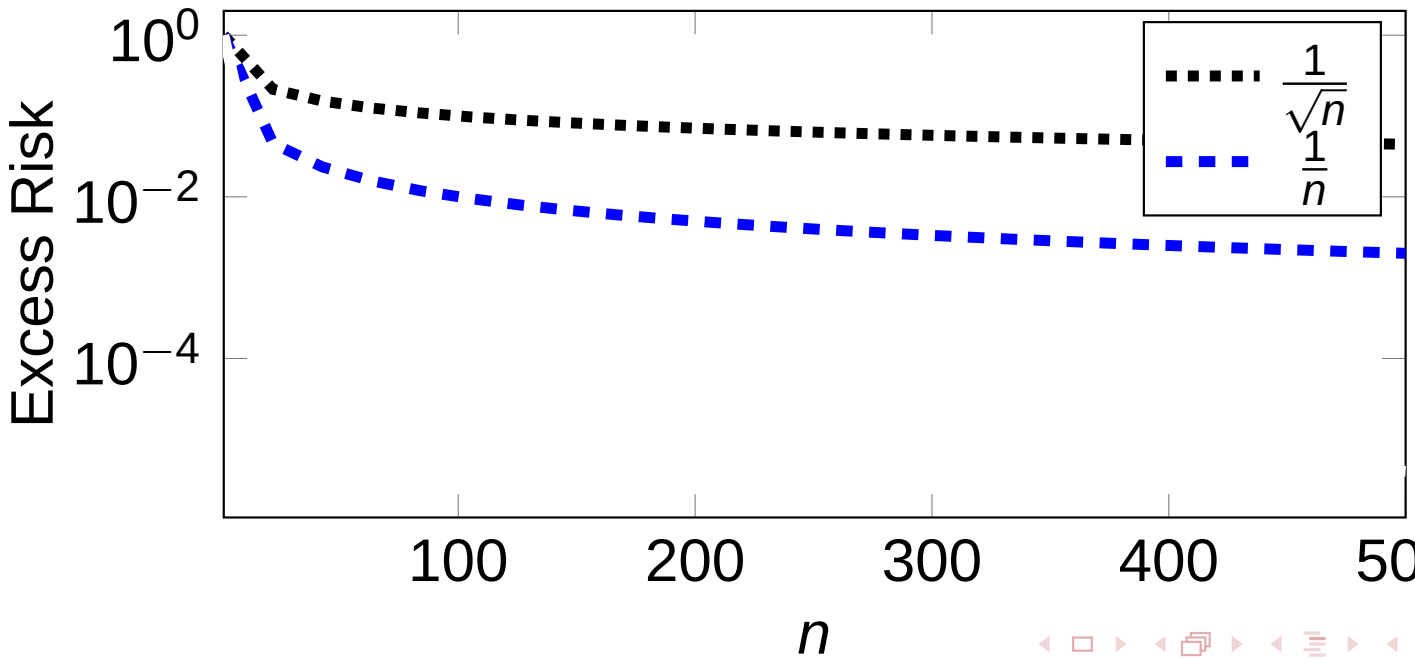
$$F(\hat{h}) - F(h_*) = O\left(\frac{1}{\sqrt{n}}\right)$$



Our Contributions [Zhang et al., 2017]

- ERM/SAA for Statistical Machine Learning

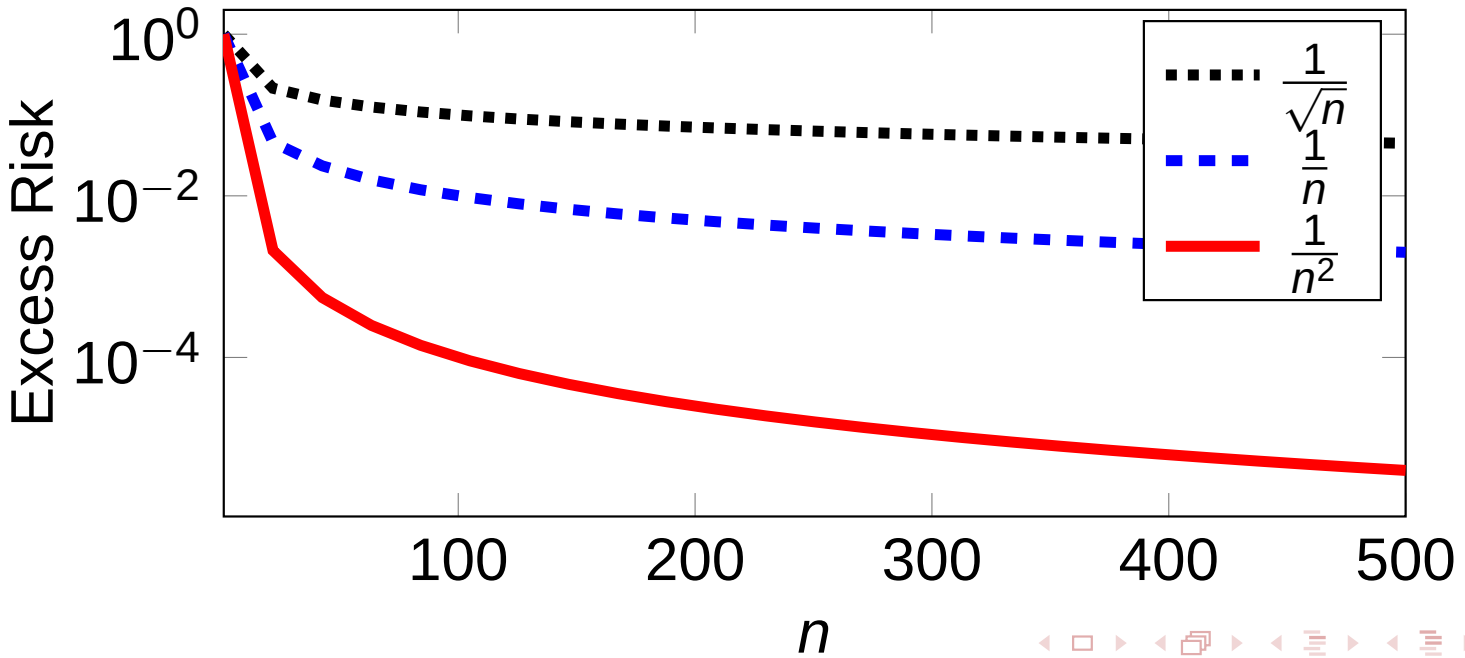
$$F(\hat{h}) - F(h_*) = O\left(\frac{1}{\sqrt{n}}\right) \overset{\text{Smoothness}}{\underset{\text{Strong Convexity}}{\Longrightarrow}} O\left(\frac{1}{n}\right)$$



Our Contributions [Zhang et al., 2017]

- ERM/SAA for Statistical Machine Learning

$$F(\hat{h}) - F(h_*) = O\left(\frac{1}{\sqrt{n}}\right) \xrightarrow[\text{Strong Convexity}]{\text{Smoothness}} O\left(\frac{1}{n}\right) \xrightarrow[\text{Strong Convexity}]{\text{Smoothness \&}} O\left(\frac{1}{n^2}\right)$$



LAMDA Learning And Mining from DatA

Our Contributions [Zhang et al., 2017]

- ERM/SAA for Statistical Machine Learning

$$F(\hat{h}) - F(h_*) = O\left(\frac{1}{\sqrt{n}}\right) \underset{\text{Strong Convexity}}{\overset{\text{Smoothness}}{\Longrightarrow}} O\left(\frac{1}{n}\right) \underset{\text{Strong Convexity}}{\overset{\text{Smoothness \&}}{\Longrightarrow}} O\left(\frac{1}{n^2}\right)$$

- ERM/SAA for Stochastic Optimization

$$F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) = O\left(\frac{1}{\sqrt{n}}\right)$$

LAMDA Learning And Mining from DatA

Our Contributions [Zhang et al., 2017]

- ERM/SAA for Statistical Machine Learning

$$F(\hat{h}) - F(h_*) = O\left(\frac{1}{\sqrt{n}}\right) \underset{\text{Strong Convexity}}{\overset{\text{Smoothness}}{\Longrightarrow}} O\left(\frac{1}{n}\right) \underset{\text{Strong Convexity}}{\overset{\text{Smoothness \&}}{\Longrightarrow}} O\left(\frac{1}{n^2}\right)$$

- ERM/SAA for Stochastic Optimization

$$F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) = O\left(\frac{1}{\sqrt{n}}\right) \underset{\text{Strong Convexity}}{\Longrightarrow} O\left(\frac{1}{n}\right)$$

LAMDA Learning And Mining from DatA

Our Contributions [Zhang et al., 2017]

- ERM/SAA for Statistical Machine Learning

$$F(\hat{h}) - F(h_*) = O\left(\frac{1}{\sqrt{n}}\right) \underset{\text{Strong Convexity}}{\overset{\text{Smoothness}}{\Longrightarrow}} O\left(\frac{1}{n}\right) \overset{\text{Smoothness \&}}{\underset{\text{Strong Convexity}}{\Longrightarrow}} O\left(\frac{1}{n^2}\right)$$

- ERM/SAA for Stochastic Optimization

$$F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) = O\left(\frac{1}{\sqrt{n}}\right) \underset{\text{Strong Convexity}}{\overset{\text{Smoothness}}{\Longrightarrow}} O\left(\frac{1}{n}\right)$$

Our Contributions [Zhang et al., 2017]

- ERM/SAA for Statistical Machine Learning

$$F(\hat{h}) - F(h_*) = O\left(\frac{1}{\sqrt{n}}\right) \underset{\text{Strong Convexity}}{\overset{\text{Smoothness}}{\Longrightarrow}} O\left(\frac{1}{n}\right) \underset{\text{Strong Convexity}}{\overset{\text{Smoothness \&}}{\Longrightarrow}} O\left(\frac{1}{n^2}\right)$$

- ERM/SAA for Stochastic Optimization

$$F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) = O\left(\frac{1}{\sqrt{n}}\right) \underset{\text{Strong Convexity}}{\overset{\text{Smoothness}}{\Longrightarrow}} O\left(\frac{1}{n}\right) \underset{\text{Strong Convexity}}{\overset{\text{Smoothness \&}}{\Longrightarrow}} O\left(\frac{1}{n^2}\right)$$

LAMDA Learning And Mining from DatA

Our Contributions [Zhang et al., 2017]

- ERM/SAA for Statistical Machine Learning

$$F(\hat{h}) - F(h_*) = O\left(\frac{1}{\sqrt{n}}\right) \underset{\text{Strong Convexity}}{\overset{\text{Smoothness}}{\Longrightarrow}} O\left(\frac{1}{n}\right) \xrightarrow[\text{Strong Convexity}]{\text{Smoothness \&}} O\left(\frac{1}{n^2}\right)$$

- ERM/SAA for Stochastic Optimization

$$F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) = O\left(\frac{1}{\sqrt{n}}\right) \underset{\text{Strong Convexity}}{\overset{\text{Smoothness}}{\Longrightarrow}} O\left(\frac{1}{n}\right) \xrightarrow[\text{Strong Convexity}]{\text{Smoothness \&}} O\left(\frac{1}{n^2}\right)$$

- The 30th Annual Conference on Learning Theory (COLT'17)
  - 1 of 18 long talks

LAMDA Learning And Mining from DatA

Outline

LAMDA
Learning And Mining from DatA

Statistical Machine Learning—0−1 Losses

- 0−1 Losses [Vapnik and Chervonenkis, 1971]

$$\text{RM:} \quad \min_{h \in \mathcal{H}} \ F(h) = \mathrm{E}_{(\mathbf{x},y)\sim\mathbb{D}}\left[\mathbb{1}(h(\mathbf{x}) \neq y)\right]$$

$$\text{ERM:} \quad \min_{h \in \mathcal{H}} \ \widehat{F}(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(h(\mathbf{x}_i) \neq y_i)$$

Statistical Machine Learning—0−1 Losses

- 0−1 Losses [Vapnik and Chervonenkis, 1971]

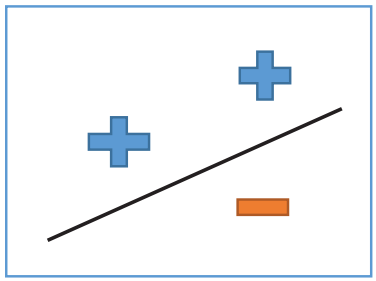
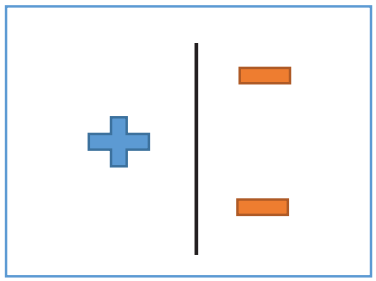
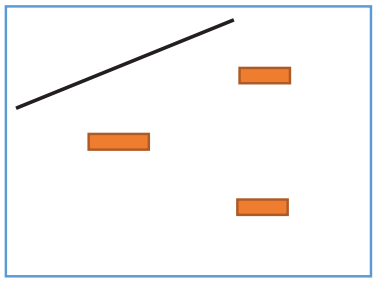
$$\text{RM:} \quad \min_{h \in \mathcal{H}} F(h) = \mathrm{E}_{(\mathbf{x},y)\sim\mathbb{D}}\left[\mathbb{1}(h(\mathbf{x}) \neq y)\right]$$

$$\text{ERM:} \quad \min_{h \in \mathcal{H}} \widehat{F}(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(h(\mathbf{x}_i) \neq y_i)$$

  - Risk Bound

$$F(\hat{h}) - F(h_*) = O\left(\sqrt{\frac{\mathrm{VC}(\mathcal{H})}{n}}\right)$$

$\mathrm{VC}(\mathcal{H})$ is the VC-dimension of $\mathcal{H}$

LAMDA Learning And Mining from DatA

Statistical Machine Learning—0−1 Losses

- 0−1 Losses [Vapnik and Chervonenkis, 1971]

$$\text{RM:} \quad \min_{h \in \mathcal{H}} F(h) = \mathrm{E}_{(\mathbf{x},y)\sim\mathbb{D}}\left[\mathbb{1}(h(\mathbf{x}) \neq y)\right]$$

$$\text{ERM:} \quad \min_{h \in \mathcal{H}} \widehat{F}(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(h(\mathbf{x}_i) \neq y_i)$$

  - Risk Bound

$$F(\hat{h}) - F(h_*) = O\left(\sqrt{\frac{\mathrm{VC}(\mathcal{H})}{n}}\right)$$

    $\mathrm{VC}(\mathcal{H})$ is the VC-dimension of $\mathcal{H}$

  - Limitations:
    1. Minimizing 0−1 losses is intractable
    2. VC-dimension is data-independent
    3. Vacuous when $\mathrm{VC}(\mathcal{H}) = \infty$ (e.g., kernels)

LAMDA Learning And Mining from DatA

Statistical Machine Learning—Lipschitz Losses

- Lipschitz Continuous Losses [Bartlett and Mendelson, 2002]

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

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

  - $\ell(\cdot, y)$ is Lipschitz continuous for any $y \in \mathcal{Y}$

Definition 1

A function $f : \mathcal{W} \in \mathbb{R}$ is G-Lipschitz continuous if

$$|f(x) - f(y)| \leq G|x - y|, \ \forall x, y \in \mathcal{W}$$

LAMDA Learning And Mining from DatA

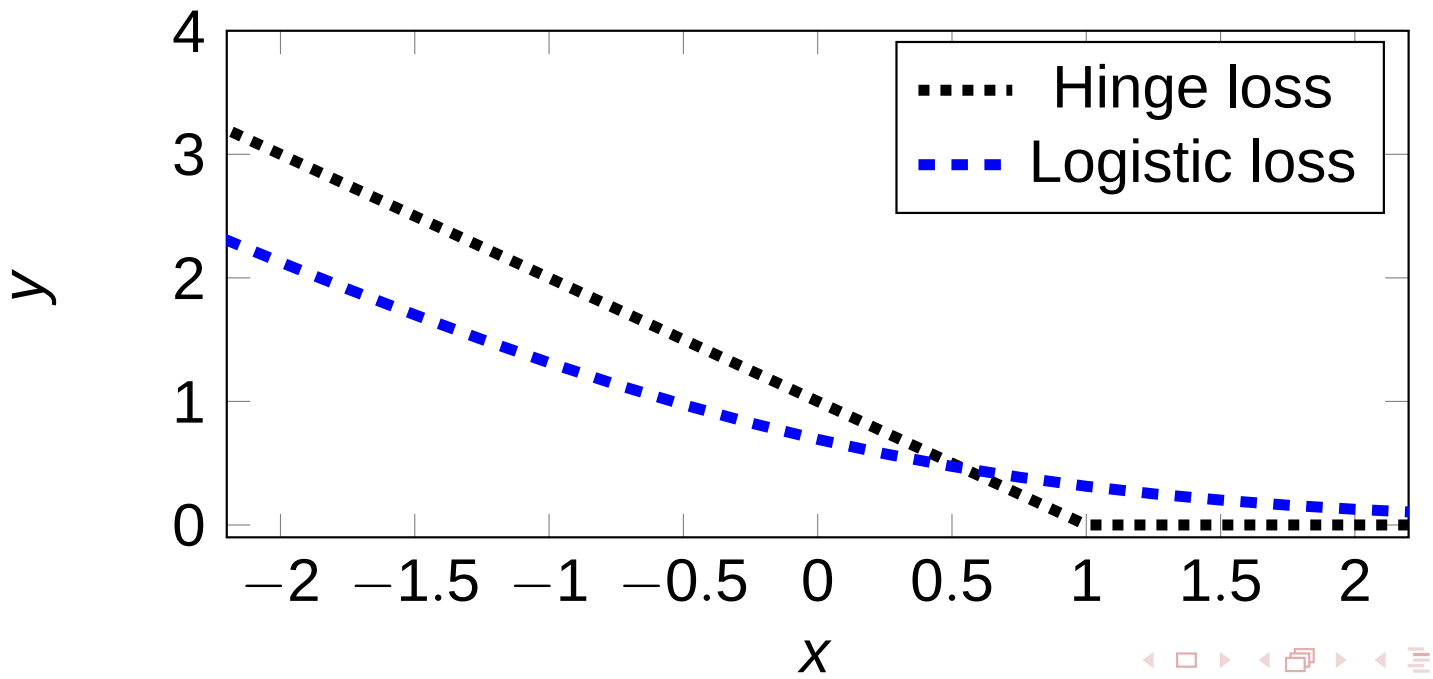
Statistical Machine Learning—Lipschitz Losses

- Lipschitz Continuous Losses [Bartlett and Mendelson, 2002]

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

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

  - $\ell(\cdot, y)$ is Lipschitz continuous for any $y \in \mathcal{Y}$



Statistical Machine Learning—Lipschitz Losses

- Lipschitz Continuous Losses [Bartlett and Mendelson, 2002]

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

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

  - $\ell(\cdot, y)$ is Lipschitz continuous for any $y \in \mathcal{Y}$
  - Risk Bound

$$F(\hat{h}) - F(h_*) = O\left(\frac{1}{\sqrt{n}} + \mathcal{R}_n(\mathcal{H})\right)$$

$\mathcal{R}_n(\mathcal{H})$ is the Rademacher complexity

$$\mathcal{R}_n(\mathcal{H}) = \mathrm{E}\left[\sup_{h \in \mathcal{H}} \left|\frac{1}{n}\sum_{i=1}^{n} \sigma_i h(\mathbf{x}_i)\right|\right]$$

$\sigma_i \in \{\pm 1\}$ is Rademacher random variable

LAMDA Learning And Mining from DatA

Statistical Machine Learning—Lipschitz Losses

- Lipschitz Continuous Losses [Bartlett and Mendelson, 2002]

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

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

  - $\ell(\cdot, y)$ is Lipschitz continuous for any $y \in \mathcal{Y}$
  - Risk Bound

$$F(\hat{h}) - F(h_*) = O\left(\frac{1}{\sqrt{n}} + \mathcal{R}_n(\mathcal{H})\right)$$

  - Suppose

$$\mathcal{H} = \{f_{\mathbf{w}} : \mathbf{x} \mapsto \langle \mathbf{w}, \phi(\mathbf{x})\rangle : \|\mathbf{w}\| \leq R\}$$

  then

$$\mathcal{R}_n(\mathcal{H}) = \mathrm{E}\left[\sup_{h \in \mathcal{H}} \left|\frac{1}{n}\sum_{i=1}^{n} \sigma_i h(\mathbf{x}_i)\right|\right] = O\left(\frac{1}{\sqrt{n}}\right)$$

LAMDA Learning And Mining from DatA

Statistical Machine Learning—Lipschitz Losses

- Lipschitz Continuous Losses [Bartlett and Mendelson, 2002]

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

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

  - $\ell(\cdot, y)$ is Lipschitz continuous for any $y \in \mathcal{Y}$
  - Risk Bound

$$F(\hat{h}) - F(h_*) = O\left(\frac{1}{\sqrt{n}} + \mathcal{R}_n(\mathcal{H})\right) = O\left(\frac{1}{\sqrt{n}}\right)$$

  - Suppose

$$\mathcal{H} = \{f_{\mathbf{w}} : \mathbf{x} \mapsto \langle \mathbf{w}, \phi(\mathbf{x})\rangle : \|\mathbf{w}\| \leq R\}$$

  then

$$\mathcal{R}_n(\mathcal{H}) = \mathrm{E}\left[\sup_{h \in \mathcal{H}} \left|\frac{1}{n}\sum_{i=1}^{n} \sigma_i h(\mathbf{x}_i)\right|\right] = O\left(\frac{1}{\sqrt{n}}\right)$$

LAMDA Learning And Mining from DatA

Statistical Machine Learning—Lipschitz Losses

- Lipschitz Continuous Losses [Bartlett and Mendelson, 2002]

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

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

  - $\ell(\cdot, y)$ is Lipschitz continuous for any $y \in \mathcal{Y}$
  - Risk Bound

$$F(\hat{h}) - F(h_*) = O\left(\frac{1}{\sqrt{n}} + \mathcal{R}_n(\mathcal{H})\right) = O\left(\frac{1}{\sqrt{n}}\right)$$

  - Advantages:
    1. Most convex losses are Lipschitz continues
    2. Rademacher complexity is data-dependent
    3. Rademacher complexity could be applied even when $\mathrm{VC}(\mathcal{H}) = \infty$ (e.g., kernels)

LAMDA Learning And Mining from DatA

Statistical Machine Learning—Smooth Losses

- Smooth Losses [Srebro et al., 2010]

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

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

  - $\ell(\cdot, y)$ is L-smooth for any $y \in \mathcal{Y}$

Definition 2

A function $f : \mathcal{W} \in \mathbb{R}$ is L-smooth if

$$|f'(x) - f'(y)| \leq L|x - y|, \ \forall x, y \in \mathcal{W}$$

LAMDA Learning And Mining from DatA

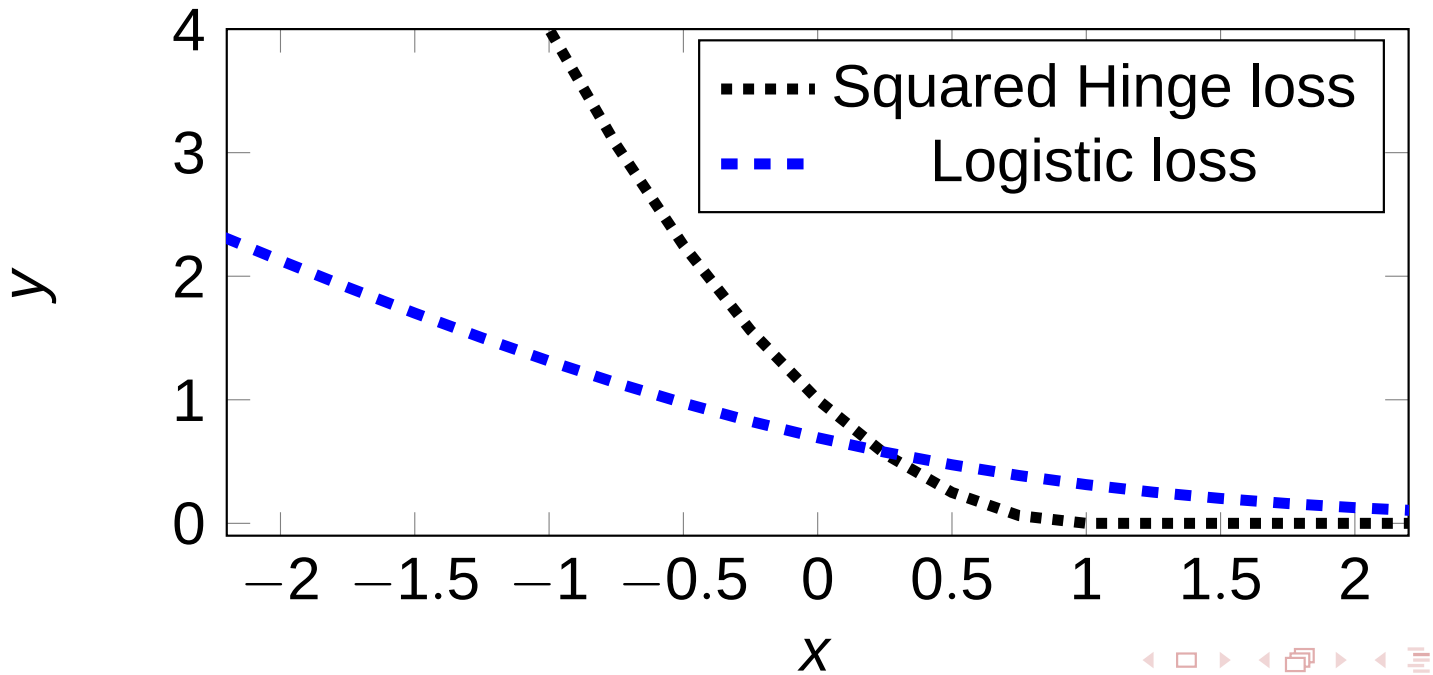
Statistical Machine Learning—Smooth Losses

- Smooth Losses [Srebro et al., 2010]

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

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

  - $\ell(\cdot, y)$ is L-smooth for any $y \in \mathcal{Y}$



LAMDA Learning And Mining from DatA

Statistical Machine Learning—Smooth Losses

- Smooth Losses [Srebro et al., 2010]

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

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

  - $\ell(\cdot, y)$ is L-smooth for any $y \in \mathcal{Y}$
  - Risk Bound

$$F(\hat{h}) - F(h_*) = \widetilde{O}\left(\mathcal{R}_n^2(\mathcal{H}) + \mathcal{R}_n(\mathcal{H})\sqrt{F_*}\right)$$

$\mathcal{R}_n(\mathcal{H})$ is the Rademacher complexity

$$\mathcal{R}_n(\mathcal{H}) = \mathrm{E}\left[\sup_{h \in \mathcal{H}} \left|\frac{1}{n}\sum_{i=1}^{n} \sigma_i h(\mathbf{x}_i)\right|\right]$$

$F_* = F(h_*)$ is the minimal risk

LAMDA Learning And Mining from DatA

Statistical Machine Learning—Smooth Losses

- Smooth Losses [Srebro et al., 2010]

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

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

  - $\ell(\cdot, y)$ is L-smooth for any $y \in \mathcal{Y}$
  - Risk Bound

$$F(\hat{h}) - F(h_*) = \widetilde{O}\left(\mathcal{R}_n^2(\mathcal{H}) + \mathcal{R}_n(\mathcal{H})\sqrt{F_*}\right) = \widetilde{O}\left(\frac{1}{n}\right)$$

  - When

$$F_* = O\left(\frac{1}{n}\right)$$

$$\mathcal{R}_n(\mathcal{H}) = \mathrm{E}\left[\sup_{h \in \mathcal{H}} \left|\frac{1}{n}\sum_{i=1}^{n} \sigma_i h(\mathbf{x}_i)\right|\right] = O\left(\frac{1}{\sqrt{n}}\right)$$

LAMDA Learning And Mining from DatA

Statistical Machine Learning—Strongly Convex

- Strongly Convex Losses [Sridharan et al., 2009]

$$\text{RM:} \quad \min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w}) = \mathrm{E}_{(\mathbf{x},y)\sim\mathbb{D}}\left[\ell(\langle\mathbf{w},\phi(\mathbf{x})\rangle, y) + r(\mathbf{w})\right]$$

$$\text{ERM:} \quad \min_{\mathbf{w}\in\mathcal{W}} \widehat{F}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}\ell(\langle\mathbf{w},\phi(\mathbf{x}_i)\rangle, y_i)) + r(\mathbf{w})$$

  - $F(\cdot)$ is λ-strongly convex over domain $\mathcal{W}$

Definition 3

A function $f : \mathcal{W} \in \mathbb{R}$ is λ-strongly convex if

$$f(\mathbf{x}) + \langle\nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle + \frac{\lambda}{2}\|\mathbf{y}-\mathbf{x}\|^2 \leq f(\mathbf{y}),\ \forall \mathbf{x},\mathbf{y}\in\mathcal{W}.$$

LAMDA Learning And Mining from DatA

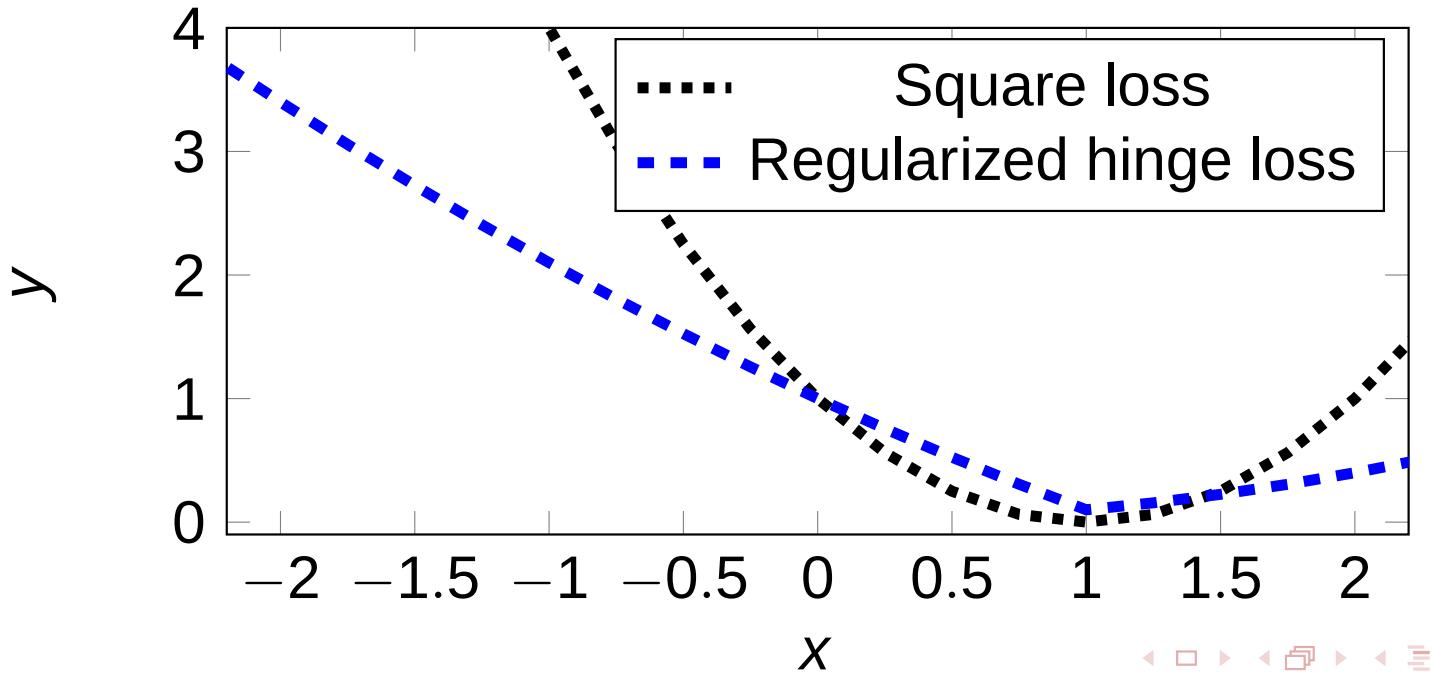
Statistical Machine Learning—Strongly Convex

- Strongly Convex Losses [Sridharan et al., 2009]

$$\text{RM:} \quad \min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w}) = \mathrm{E}_{(\mathbf{x},y)\sim\mathbb{D}}\left[\ell(\langle\mathbf{w},\phi(\mathbf{x})\rangle, y) + r(\mathbf{w})\right]$$

$$\text{ERM:} \quad \min_{\mathbf{w}\in\mathcal{W}} \widehat{F}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\langle\mathbf{w},\phi(\mathbf{x}_i)\rangle, y_i)) + r(\mathbf{w})$$

  - $F(\cdot)$ is λ-strongly convex over domain $\mathcal{W}$



Statistical Machine Learning—Strongly Convex

- Strongly Convex Losses [Sridharan et al., 2009]

$$\text{RM:} \quad \min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w}) = \mathrm{E}_{(\mathbf{x},y)\sim\mathbb{D}}\left[\ell(\langle\mathbf{w},\phi(\mathbf{x})\rangle, y) + r(\mathbf{w})\right]$$

$$\text{ERM:} \quad \min_{\mathbf{w}\in\mathcal{W}} \widehat{F}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\langle\mathbf{w},\phi(\mathbf{x}_i)\rangle, y_i)) + r(\mathbf{w})$$

  - $F(\cdot)$ is λ-strongly convex over domain $\mathcal{W}$
  - $\ell(\cdot, y)$ is Lipschitz continuous for any $y \in \mathcal{Y}$

Statistical Machine Learning—Strongly Convex

- Strongly Convex Losses [Sridharan et al., 2009]

$$\text{RM:} \quad \min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w}) = \mathrm{E}_{(\mathbf{x},y)\sim\mathbb{D}}\left[\ell(\langle\mathbf{w},\phi(\mathbf{x})\rangle, y) + r(\mathbf{w})\right]$$

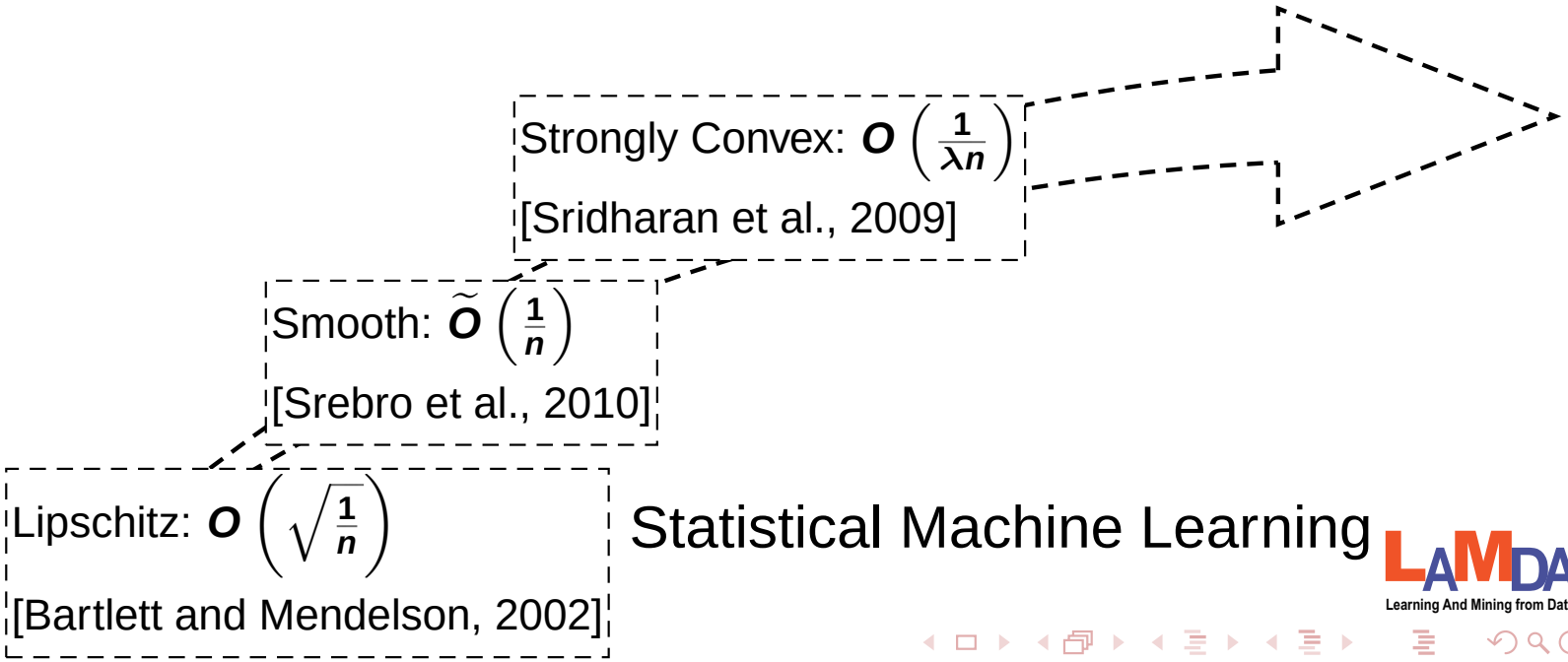
$$\text{ERM:} \quad \min_{\mathbf{w}\in\mathcal{W}} \widehat{F}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}\ell(\langle\mathbf{w},\phi(\mathbf{x}_i)\rangle, y_i)) + r(\mathbf{w})$$

  - $F(\cdot)$ is λ-strongly convex over domain $\mathcal{W}$
  - $\ell(\cdot, y)$ is Lipschitz continuous for any $y \in \mathcal{Y}$
  - Risk Bound

$$F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) = O\left(\frac{1}{\lambda n}\right)$$

    Based on the Rademacher complexity

Risk Bounds of Empirical Risk Minimization



Stochastic Optimization

- Stochastic Optimization and Sample Average Approximation

$$\text{SO:} \quad \min_{\mathbf{w}\in\mathcal{W}} \; F(\mathbf{w}) = \mathrm{E}_{f\sim\mathbb{P}}\left[f(\mathbf{w})\right]$$

$$\text{SAA:} \quad \min_{\mathbf{w}\in\mathcal{W}} \; \widehat{F}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{w})$$

Stochastic Optimization

- Stochastic Optimization and Sample Average Approximation

$$\text{SO:} \quad \min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w}) = \mathrm{E}_{f\sim\mathbb{P}}[f(\mathbf{w})]$$

$$\text{SAA:} \quad \min_{\mathbf{w}\in\mathcal{W}} \widehat{F}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{w})$$

  - Maximum Likelihood Estimate [Wald, 1949, Huber, 1967]
    - Asymptotic analysis

Stochastic Optimization

- Stochastic Optimization and Sample Average Approximation

$$\text{SO:} \quad \min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w}) = \mathrm{E}_{f\sim\mathbb{P}}[f(\mathbf{w})]$$

$$\text{SAA:} \quad \min_{\mathbf{w}\in\mathcal{W}} \widehat{F}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{w})$$

- Maximum Likelihood Estimate [Wald, 1949, Huber, 1967]
  - Asymptotic analysis
- $f(\cdot)$ is Lipschitz continuous [Shalev-Shwartz et al., 2009]

$$F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) = \widetilde{O}\left(\sqrt{\frac{d}{n}}\right)$$

LAMDA Learning And Mining from DatA

Stochastic Optimization

- Stochastic Optimization and Sample Average Approximation

$$\text{SO:} \quad \min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w}) = \mathrm{E}_{f\sim\mathbb{P}}[f(\mathbf{w})]$$

$$\text{SAA:} \quad \min_{\mathbf{w}\in\mathcal{W}} \widehat{F}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{w})$$

- Maximum Likelihood Estimate [Wald, 1949, Huber, 1967]
  - Asymptotic analysis
- $f(\cdot)$ is Lipschitz continuous [Shalev-Shwartz et al., 2009]

$$F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) = \widetilde{O}\left(\sqrt{\frac{d}{n}}\right)$$

- $f(\cdot)$ is λ-strongly convex and Lipschitz continuous [Shalev-Shwartz et al., 2009]

$$\mathrm{E}\left[F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*)\right] = O\left(\frac{1}{\lambda n}\right)$$

LAMDA Learning And Mining from DatA

Risk Bounds of Empirical Risk Minimization

Stochastic Optimization

Strongly Convex: $O\left(\frac{1}{\lambda n}\right)$ in expection

[Shalev-Shwartz et al., 2009]

Strongly Convex: $O\left(\frac{1}{\lambda n}\right)$

[Sridharan et al., 2009]

Lipschitz: $\widetilde{O}\left(\sqrt{\frac{d}{n}}\right)$

[Shalev-Shwartz et al., 2009]

Smooth: $\widetilde{O}\left(\frac{1}{n}\right)$

[Srebro et al., 2010]

Lipschitz: $O\left(\sqrt{\frac{1}{n}}\right)$

[Bartlett and Mendelson, 2002]

Statistical Machine Learning

LAMDA Learning And Mining from DatA

Outline

Statistical Machine Learning

- Risk Minimization (RM)

$$\min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w}) = \mathrm{E}_{(\mathbf{x},y)\sim\mathbb{D}}\left[\ell(\langle\mathbf{w},\phi(\mathbf{x})\rangle, y)\right] + r(\mathbf{w})$$

$$\mathbf{w}_* = \operatorname*{argmin}_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w})$$

- Empirical Risk Minimization (ERM)

$$\min_{\mathbf{w}\in\mathcal{W}} \widehat{F}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\langle\mathbf{w},\phi(\mathbf{x}_i)\rangle, y_i)) + r(\mathbf{w})$$

$$\widehat{\mathbf{w}} = \operatorname*{argmin}_{\mathbf{w}\in\mathcal{W}} \widehat{F}(\mathbf{w})$$

- Excess Risk

$$F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*)$$

LAMDA Learning And Mining from DatA

Risk Bounds of Empirical Risk Minimization

Stochastic Optimization

Strongly Convex: $O\left(\frac{1}{\lambda n}\right)$ in expection

[Shalev-Shwartz et al., 2009]

Strongly Convex: $O\left(\frac{1}{\lambda n}\right)$

[Sridharan et al., 2009]

Lipschitz: $\widetilde{O}\left(\sqrt{\frac{d}{n}}\right)$

[Shalev-Shwartz et al., 2009]

Smooth: $\widetilde{O}\left(\frac{1}{n}\right)$

[Srebro et al., 2010]

Lipschitz: $O\left(\sqrt{\frac{1}{n}}\right)$

[Bartlett and Mendelson, 2002]

Statistical Machine Learning

LAMDA Learning And Mining from DatA

Risk Bounds of Empirical Risk Minimization

Stochastic Optimization

Strongly Convex: $O\left(\frac{1}{\lambda n}\right)$ in expection

[Shalev-Shwartz et al., 2009]

Strongly Convex & Smooth: $O\left(\frac{1}{\lambda n^2}\right)$

[Zhang et al., 2017]

Strongly Convex: $O\left(\frac{1}{\lambda n}\right)$

[Sridharan et al., 2009]

Lipschitz: $\widetilde{O}\left(\sqrt{\frac{d}{n}}\right)$

[Shalev-Shwartz et al., 2009]

Smooth: $\widetilde{O}\left(\frac{1}{n}\right)$

[Srebro et al., 2010]

Lipschitz: $O\left(\sqrt{\frac{1}{n}}\right)$

[Bartlett and Mendelson, 2002]

Statistical Machine Learning

LAMDA Learning And Mining from DatA

Smoothness & Strong Convexity

Theorem 1

Assume

- $\ell(\cdot, y)$ *is nonnegative, and* L*-smooth*
- $r(\cdot)$ *is Lipschitz continuous*
- $F(\cdot)$ *is* λ*-strongly convex*

When $n = \Omega(\kappa^2)$*, with high probability, we have*

$$F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) = O\left(\frac{1}{\lambda n^2} + \frac{\kappa H_*}{n}\right)$$

where $\kappa = L/\lambda$ *and* $H_* = F(\mathbf{w}_*) - r(\mathbf{w}_*)$.

LAMDA Learning And Mining from DatA

Smoothness & Strong Convexity

Theorem 1

Assume

- $\ell(\cdot, y)$ *is nonnegative, and L-smooth*
- $r(\cdot)$ *is Lipschitz continuous*
- $F(\cdot)$ *is* λ*-strongly convex*

When $n = \Omega(\kappa^2)$*, with high probability, we have*

$$F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) = O\left(\frac{1}{\lambda n^2} + \frac{\kappa H_*}{n}\right)$$

where $\kappa = L/\lambda$ *and* $H_* = F(\mathbf{w}_*) - r(\mathbf{w}_*)$.

Corollary 1

When $n = \Omega(\kappa^2)$ *and* $H_* = O(1/n)$,

$$F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) = O\left(\frac{1}{\lambda n^2}\right)$$

LAMDA Learning And Mining from DatA

Examples

- Least Squares

$$\min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w}) = \mathrm{E}_{(\mathbf{x},y)\sim\mathbb{D}}\left[\left(\mathbf{x}^\top\mathbf{w} - y\right)^2\right]$$

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

  - $F(\cdot)$ is strongly convex if $\mathrm{E}[\mathbf{x}\mathbf{x}^\top]$ is full-rank

Examples

- Least Squares

$$\min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w}) = \mathrm{E}_{(\mathbf{x},y)\sim\mathbb{D}}\left[\left(\mathbf{x}^\top\mathbf{w} - y\right)^2\right]$$

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

  - $F(\cdot)$ is strongly convex if $\mathrm{E}[\mathbf{x}\mathbf{x}^\top]$ is full-rank

- Regularized Logistic Regression

$$\min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w}) = \mathrm{E}_{(\mathbf{x},y)\sim\mathbb{D}}\left[\log\left(1 + e^{-y\mathbf{x}^\top\mathbf{w}}\right)\right] + \lambda\|\mathbf{w}\|^2$$

$$\min_{\mathbf{w}\in\mathcal{W}} \widehat{F}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}\log\left(1 + e^{-y_i\mathbf{x}_i^\top\mathbf{w}}\right) + \lambda\|\mathbf{w}\|^2$$

LAMDA Learning And Mining from DatA

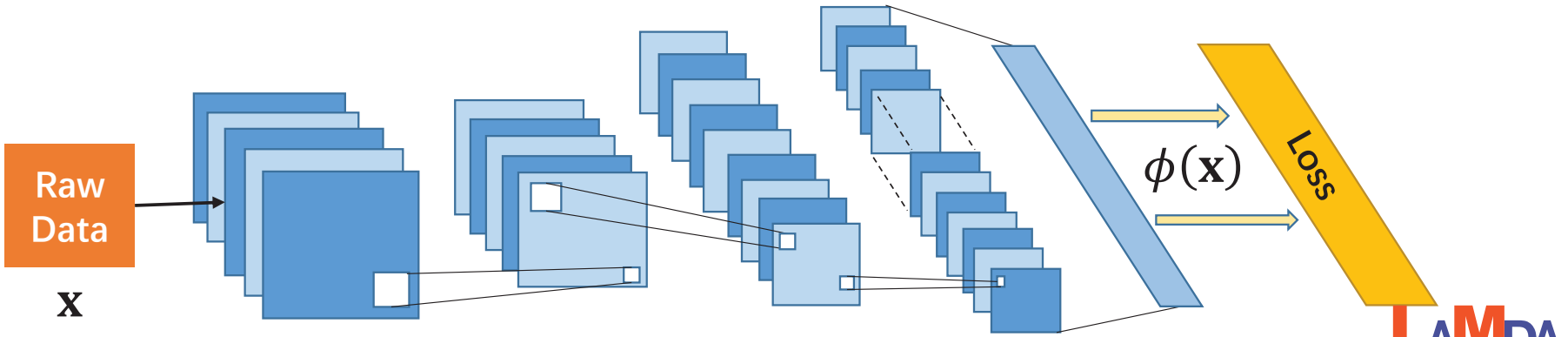
Implications

When n is large and H_* is small, ERM has $O(1/n^2)$ rate.

Implications

When n is large and H_* is small, ERM has $O(1/n^2)$ rate.

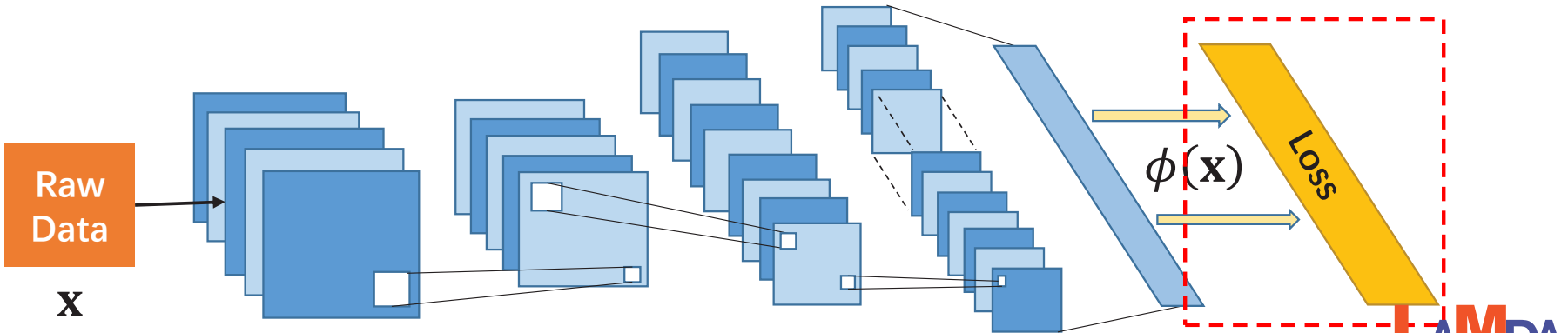
- The number of training data n is large
  - Big data is powerful
- The minimal risk H_* is small
  - Representation learning is necessary



Implications

When n is large and H_* is small, ERM has $O(1/n^2)$ rate.

- The number of training data n is large
  - Big data is powerful
- The minimal risk H_* is small
  - Representation learning is necessary



Outline

Stochastic Optimization

- Stochastic Optimization

$$\min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w}) = \mathrm{E}_{f\sim\mathbb{P}}[f(\mathbf{w})] + r(\mathbf{w})$$

$$\mathbf{w}_* = \operatorname*{argmin}_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w})$$

- Sample Average Approximation (SAA)

$$\min_{\mathbf{w}\in\mathcal{W}} \widehat{F}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{w}) + r(\mathbf{w})$$

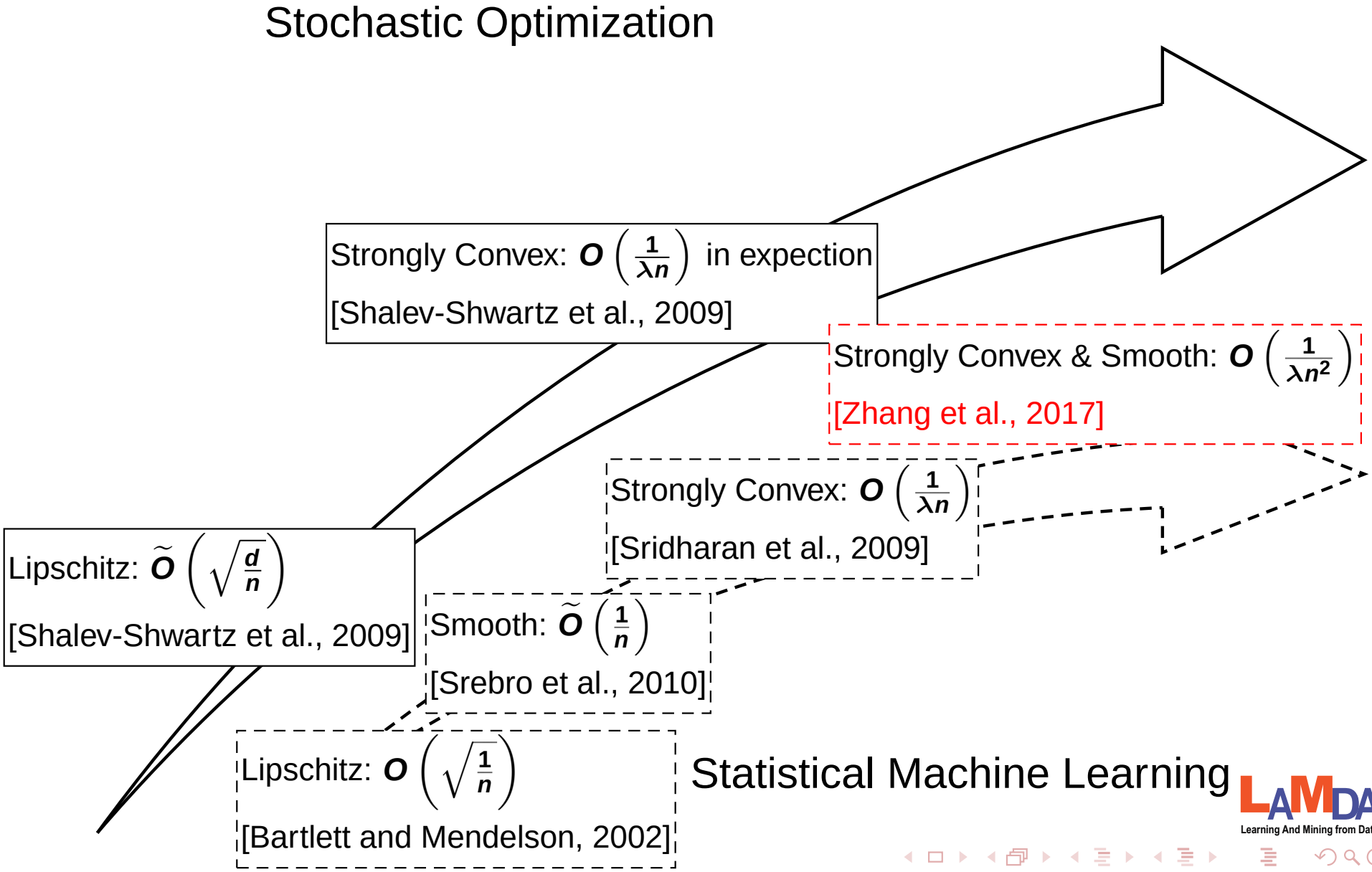
$$\widehat{\mathbf{w}} = \operatorname*{argmin}_{\mathbf{w}\in\mathcal{W}} \widehat{F}(\mathbf{w})$$

- Excess Risk

$$F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*)$$

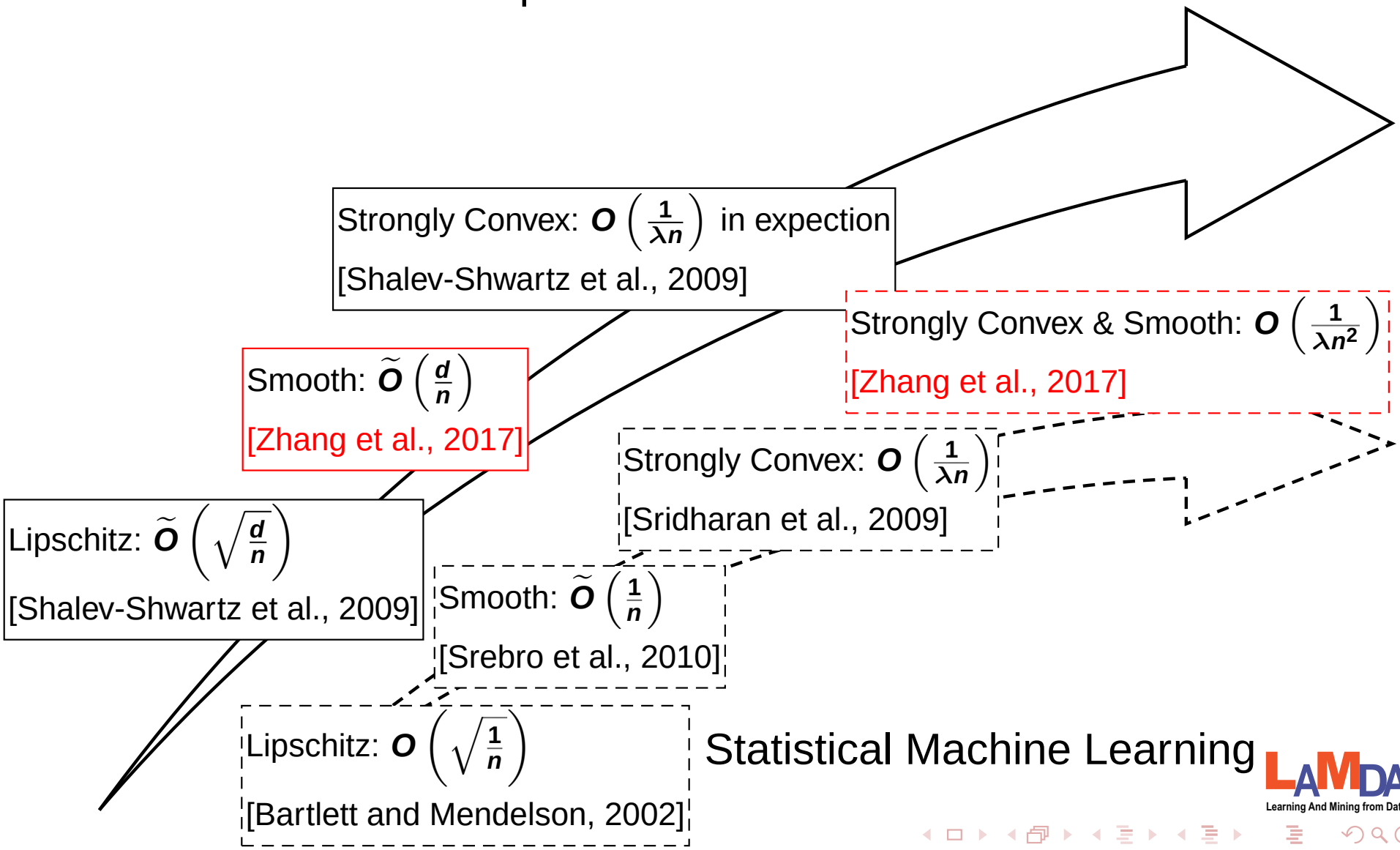
LAMDA Learning And Mining from DatA

Risk Bounds of Empirical Risk Minimization



Risk Bounds of Empirical Risk Minimization

Stochastic Optimization



Statistical Machine Learning

LAMDA Learning And Mining from DatA

Smoothness

Theorem 2

Assume

- $f(\cdot)$ *is nonnegative,* L*-smooth, and convex*
- $F(\cdot)$ *is Lipschitz continuous*

Then, with high probability, we have

$$F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) = O\left(\frac{d \log n}{n} + \sqrt{\frac{H_*}{n}}\right) = \widetilde{O}\left(\frac{d}{n} + \sqrt{\frac{H_*}{n}}\right)$$

where $H_* = F(\mathbf{w}_*) - r(\mathbf{w}_*)$.

Smoothness

Theorem 2

Assume

- $f(\cdot)$ *is nonnegative,* L*-smooth, and convex*
- $F(\cdot)$ *is Lipschitz continuous*

Then, with high probability, we have

$$F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) = O\left(\frac{d \log n}{n} + \sqrt{\frac{H_*}{n}}\right) = \widetilde{O}\left(\frac{d}{n} + \sqrt{\frac{H_*}{n}}\right)$$

where $H_* = F(\mathbf{w}_*) - r(\mathbf{w}_*)$.

Corollary 2

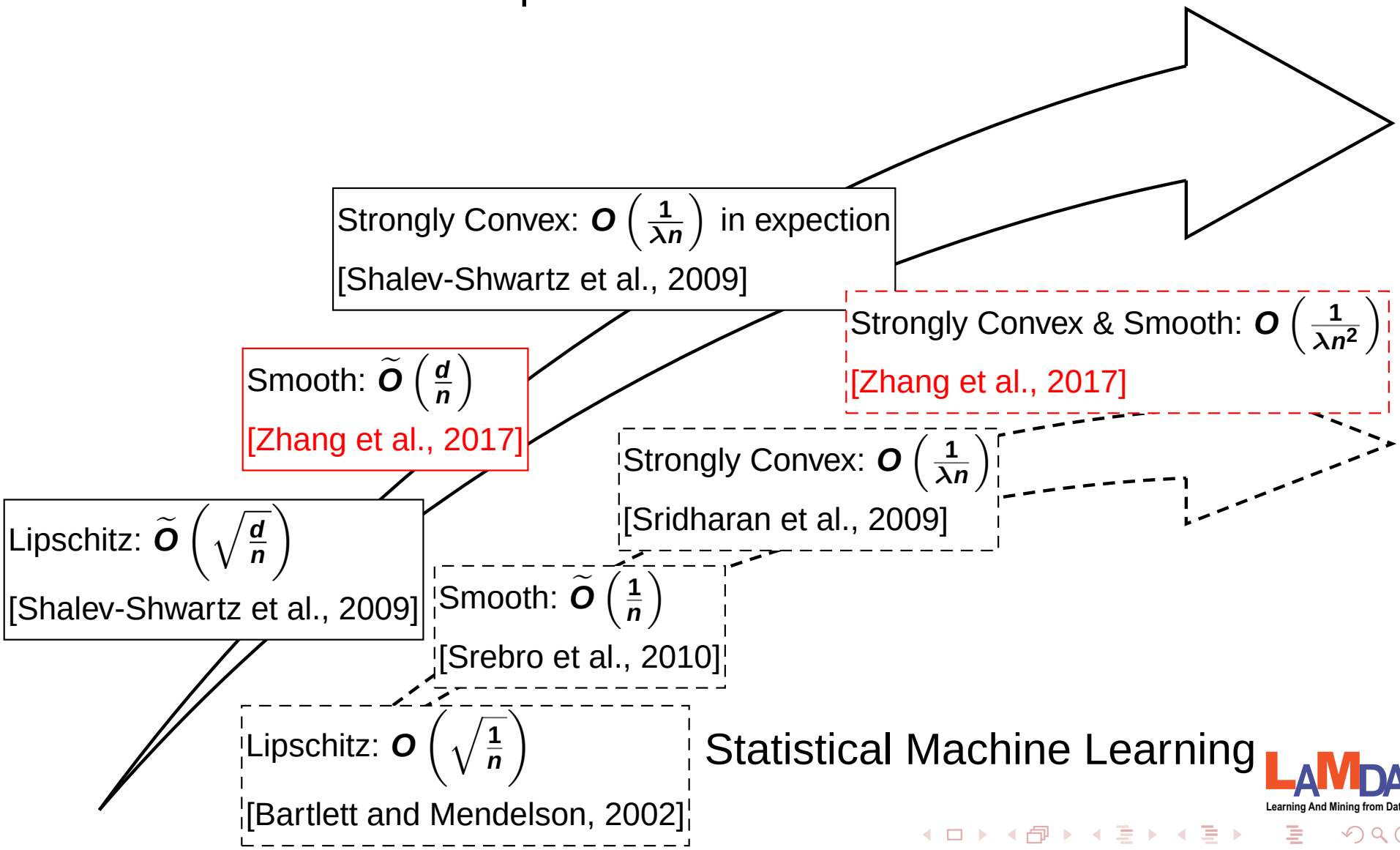
Under the above assumptions, when $H_* = O(d^2/n)$

$$F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) = \widetilde{O}\left(\frac{d}{n}\right)$$

LAMDA Learning And Mining from DatA

Risk Bounds of Empirical Risk Minimization

Stochastic Optimization

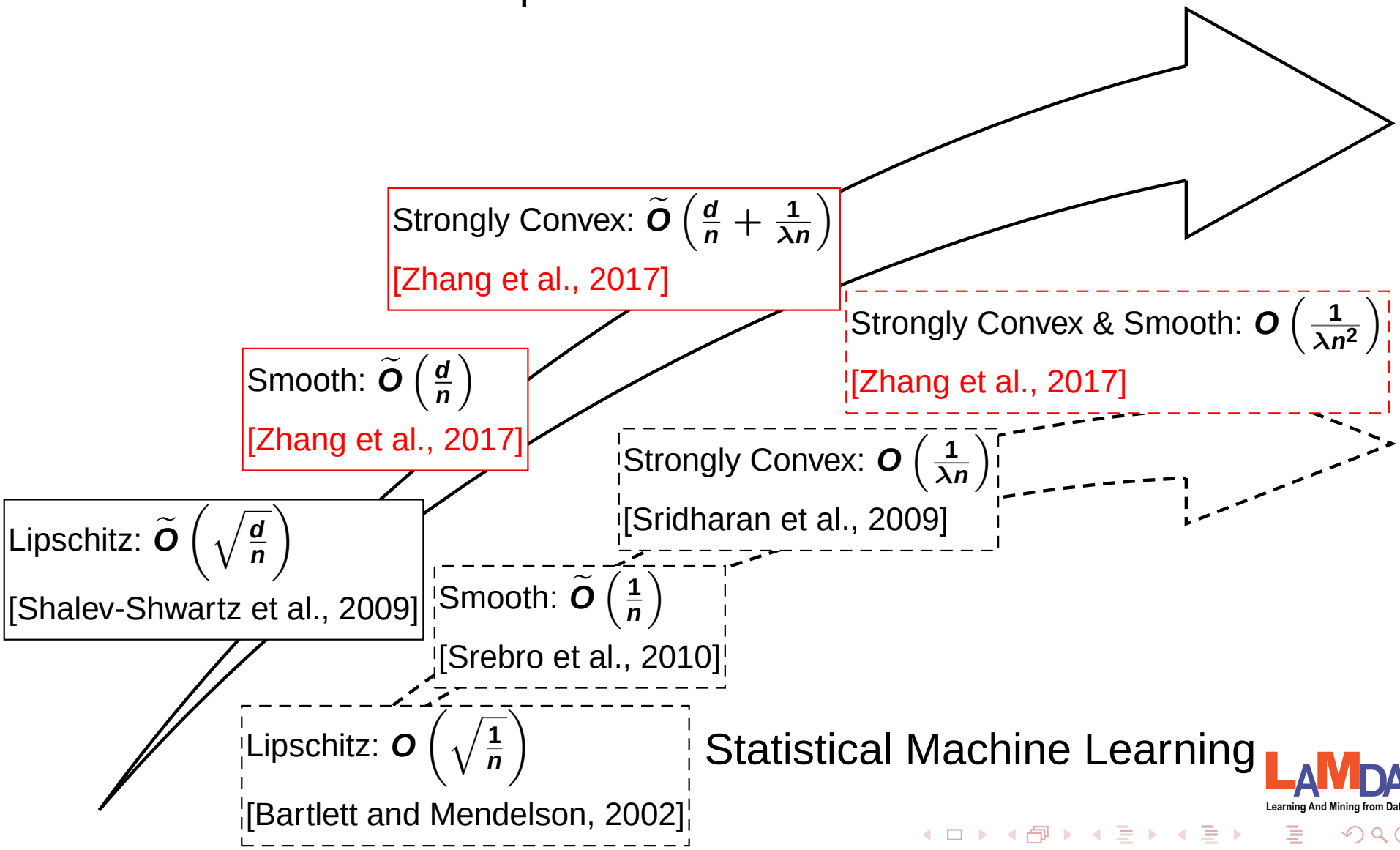


Statistical Machine Learning

LAMDA Learning And Mining from DatA

Risk Bounds of Empirical Risk Minimization

Stochastic Optimization



Statistical Machine Learning

LAMDA Learning And Mining from DatA

Smoothness & Strong Convexity I

Theorem 3

Assume

- $f(\cdot)$ *is nonnegative,* L*-smooth, and convex*
- $F(\cdot)$ *is Lipschitz continuous*
- $F(\cdot)$ *is* λ*-strongly convex*

Then, with high probability, we have

$$F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) = O\left(\frac{d \log n}{n} + \frac{\kappa H_*}{n}\right) = \widetilde{O}\left(\frac{d}{n} + \frac{1}{\lambda n}\right)$$

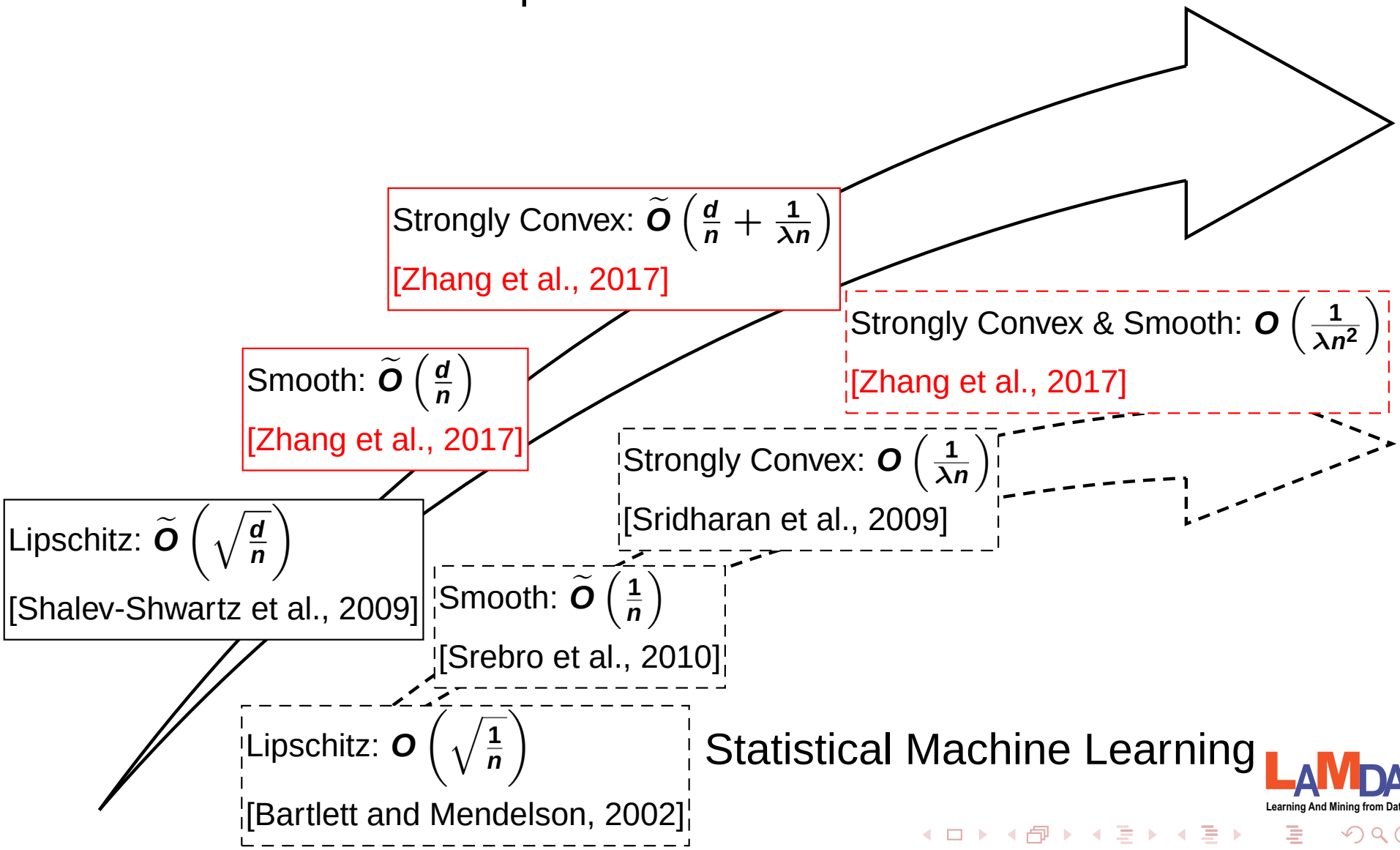
where

$$\kappa = \frac{L}{\lambda} \text{ and } H_* = F(\mathbf{w}_*) - r(\mathbf{w}_*).$$

LAMDA Learning And Mining from DatA

Risk Bounds of Empirical Risk Minimization

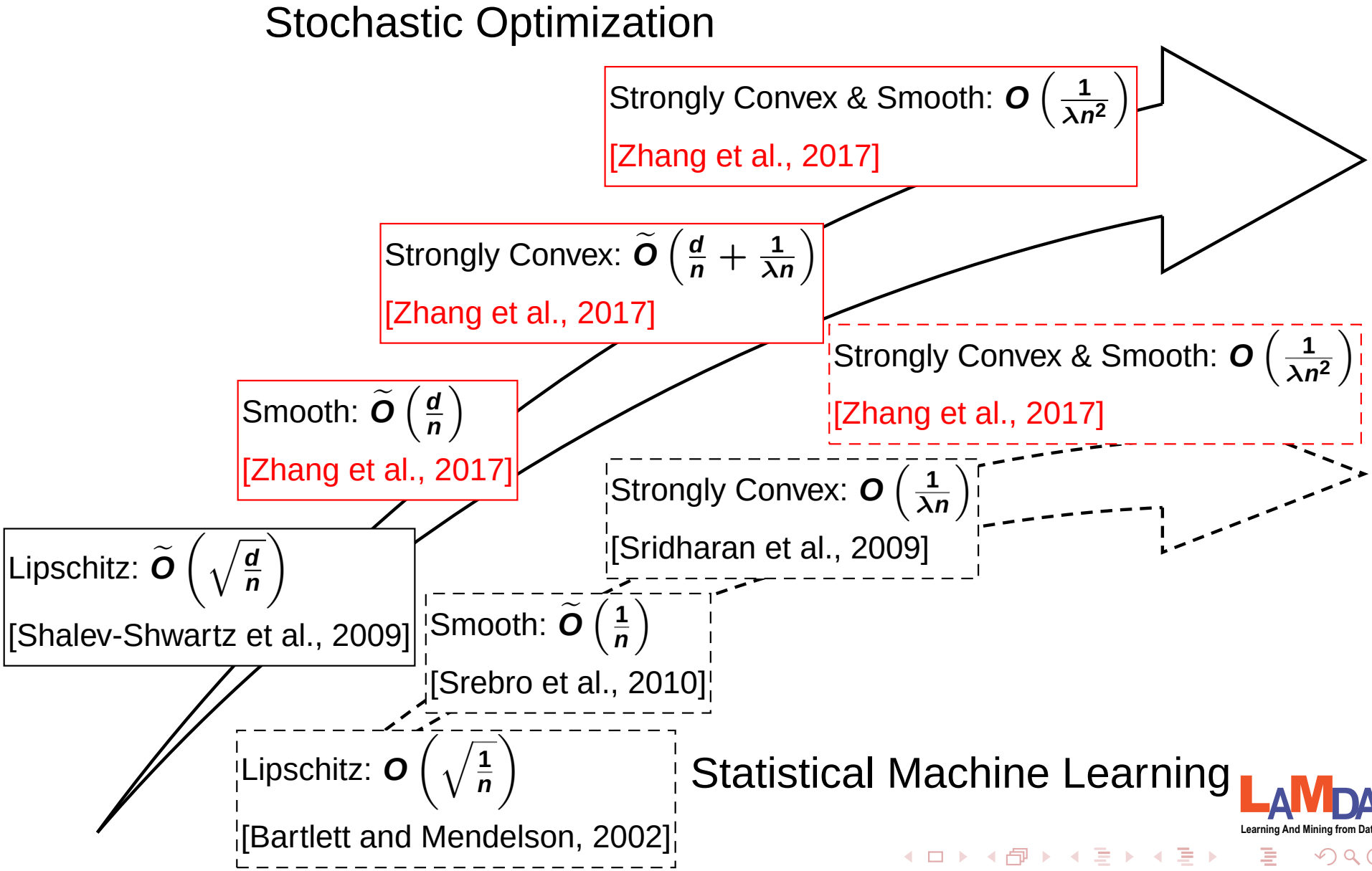
Stochastic Optimization



Statistical Machine Learning

LAMDA Learning And Mining from DatA

Risk Bounds of Empirical Risk Minimization



Smoothness & Strong Convexity II

Theorem 4

Assume

- $f(\cdot)$ *is nonnegative,* L*-smooth, and convex*
- $F(\cdot)$ *is Lipschitz continuous*
- $F(\cdot)$ *is* λ*-strongly convex*

When $n = \Omega(\kappa d \log n) = \widetilde{\Omega}(\kappa d)$*, with high probability, we have*

$$F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) = O\left(\frac{1}{\lambda n^2} + \frac{\kappa H_*}{n}\right)$$

where $\kappa = L/\lambda$ *and* $H_* = F(\mathbf{w}_*) - r(\mathbf{w}_*)$.

Smoothness & Strong Convexity II

Theorem 4

Assume

- $f(\cdot)$ *is nonnegative,* L*-smooth, and convex*
- $F(\cdot)$ *is Lipschitz continuous*
- $F(\cdot)$ *is* λ*-strongly convex*

When $n = \Omega(\kappa d \log n) = \widetilde{\Omega}(\kappa d)$*, with high probability, we have*

$$F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) = O\left(\frac{1}{\lambda n^2} + \frac{\kappa H_*}{n}\right)$$

where $\kappa = L/\lambda$ *and* $H_* = F(\mathbf{w}_*) - r(\mathbf{w}_*)$.

Corollary 3

When $n = \widetilde{\Omega}(\kappa d)$ *and* $H_* = O(1/n)$,

$$F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) = O\left(\frac{1}{\lambda n^2}\right)$$

LAMDA Learning And Mining from DatA

Analysis I

- Uniform Convergence of Functions (Traditional Analysis)

$$\sup_{\mathbf{w}\in\mathcal{W}} \left| F(\mathbf{w}) - \widehat{F}(\mathbf{w}) \right|$$

  - $O(1/n)$ is the best rate

Analysis I

- Uniform Convergence of Functions (Traditional Analysis)

$$\sup_{\mathbf{w}\in\mathcal{W}} \left| F(\mathbf{w}) - \widehat{F}(\mathbf{w}) \right|$$

  - $O(1/n)$ is the best rate

- Uniform Convergence of Gradients (Our Analysis)

$$\sup_{\mathbf{w}\in\mathcal{W}} \left\| \nabla F(\mathbf{w}) - \nabla \widehat{F}(\mathbf{w}) \right\|$$

  - $O(1/n)$ is the best rate

LAMDA Learning And Mining from DatA

Analysis II

- The Basic Idea

$$F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) + \frac{\lambda}{2}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2$$
$$\leq \left\|\nabla F(\widehat{\mathbf{w}}) - \nabla \widehat{F}(\widehat{\mathbf{w}})\right\| \|\widehat{\mathbf{w}} - \mathbf{w}_*\|$$

Analysis II

- The Basic Idea

$$
\begin{aligned}
& F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) + \frac{\lambda}{2}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2 \\
\leq & \left\|\nabla F(\widehat{\mathbf{w}}) - \nabla \widehat{F}(\widehat{\mathbf{w}})\right\| \|\widehat{\mathbf{w}} - \mathbf{w}_*\| \\
\leq & \sup_{\mathbf{w}\in\mathcal{W}} \left\|\nabla F(\mathbf{w}) - \nabla \widehat{F}(\mathbf{w})\right\| \|\widehat{\mathbf{w}} - \mathbf{w}_*\|
\end{aligned}
$$

Analysis II

- The Basic Idea

$$
\begin{aligned}
& F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) + \frac{\lambda}{2}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2 \\
\leq & \left\|\nabla F(\widehat{\mathbf{w}}) - \nabla \widehat{F}(\widehat{\mathbf{w}})\right\| \|\widehat{\mathbf{w}} - \mathbf{w}_*\| \\
\leq & \sup_{\mathbf{w}\in\mathcal{W}} \left\|\nabla F(\mathbf{w}) - \nabla \widehat{F}(\mathbf{w})\right\| \|\widehat{\mathbf{w}} - \mathbf{w}_*\| \\
\leq & \frac{c}{n}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|
\end{aligned}
$$

Analysis II

- The Basic Idea

$$
\begin{aligned}
& F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) + \frac{\lambda}{2}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2 \\
\leq & \left\|\nabla F(\widehat{\mathbf{w}}) - \nabla \widehat{F}(\widehat{\mathbf{w}})\right\| \|\widehat{\mathbf{w}} - \mathbf{w}_*\| \\
\leq & \sup_{\mathbf{w}\in\mathcal{W}} \left\|\nabla F(\mathbf{w}) - \nabla \widehat{F}(\mathbf{w})\right\| \|\widehat{\mathbf{w}} - \mathbf{w}_*\| \\
\leq & \frac{c}{n}\|\widehat{\mathbf{w}} - \mathbf{w}_*\| \overset{2ab\leq a^2+b^2}{\leq} \frac{c^2}{2\lambda n^2} + \frac{\lambda}{2}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2
\end{aligned}
$$

Analysis II

- The Basic Idea

$$
\begin{aligned}
& F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) + \frac{\lambda}{2}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2 \\
\leq & \left\|\nabla F(\widehat{\mathbf{w}}) - \nabla \widehat{F}(\widehat{\mathbf{w}})\right\| \|\widehat{\mathbf{w}} - \mathbf{w}_*\| \\
\leq & \sup_{\mathbf{w}\in\mathcal{W}} \left\|\nabla F(\mathbf{w}) - \nabla \widehat{F}(\mathbf{w})\right\| \|\widehat{\mathbf{w}} - \mathbf{w}_*\| \\
\leq & \frac{c}{n}\|\widehat{\mathbf{w}} - \mathbf{w}_*\| \overset{2ab\leq a^2+b^2}{\leq} \frac{c^2}{2\lambda n^2} + \frac{\lambda}{2}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2
\end{aligned}
$$

implying

$$
F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) \leq \frac{c^2}{2\lambda n^2} = O\left(\frac{1}{\lambda n^2}\right)
$$

LAMDA Learning And Mining from DatA

Analysis II

- The Basic Idea

$$
\begin{aligned}
& F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) + \frac{\lambda}{2}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2 \\
\leq & \left\|\nabla F(\widehat{\mathbf{w}}) - \nabla \widehat{F}(\widehat{\mathbf{w}})\right\| \|\widehat{\mathbf{w}} - \mathbf{w}_*\| \\
\leq & \sup_{\mathbf{w}\in\mathcal{W}} \left\|\nabla F(\mathbf{w}) - \nabla \widehat{F}(\mathbf{w})\right\| \|\widehat{\mathbf{w}} - \mathbf{w}_*\| \\
\leq & \frac{c}{n}\|\widehat{\mathbf{w}} - \mathbf{w}_*\| \overset{2ab\leq a^2+b^2}{\leq} \frac{c^2}{2\lambda n^2} + \frac{\lambda}{2}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2
\end{aligned}
$$

implying

$$
F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) \leq \frac{c^2}{2\lambda n^2} = O\left(\frac{1}{\lambda n^2}\right)
$$

- Rademacher Complexity, Covering Number
- Concentration Inequality of Vectors
- Convexity, Smoothness

LAMDA Learning And Mining from DatA

Outline

LAMDA
Learning And Mining from DatA

Conclusions

- ERM/SAA for Statistical Machine Learning

$f(\cdot)$	$F(\cdot)$	Risk Bounds
Smooth	Strongly convex	$O(\frac{1}{\lambda n^2} + \frac{\kappa H_*}{n})$ when $n = \Omega(\kappa^2)$

- ERM/SAA for Stochastic Optimization

$f(\cdot)$	$F(\cdot)$	Risk Bounds
Smooth	Lipschitz	$\widetilde{O}(\frac{d}{n} + \sqrt{\frac{H_*}{n}})$
Smooth	Lipschitz Strongly Convex	$\widetilde{O}(\frac{d}{n} + \frac{\kappa H_*}{n})$ $O(\frac{1}{\lambda n^2} + \frac{\kappa H_*}{n})$ when $n = \widetilde{\Omega}(\kappa d)$

LAMDA
Learning And Mining from DatA

Future Work

- Optimality of Our Bounds
  - Is the risk bound tight?
  - Is the lower bound on n unavoidable?
  - Can the assumptions (e.g., strong convexity) be relaxed?
- Fast Rates for Stochastic Approximation (SA)
  - Stochastic gradient descent (SGD)
- Fast Rates for Non-convex Losses
  - Deep learning

Reference I

Thanks!


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


Huber, P. J. (1967).
The behavior of maximum likelihood estimates under nonstandard conditions.
In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 221–233.


Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. (2009).
Stochastic convex optimization.
In *Proceedings of the 22nd Annual Conference on Learning Theory*.


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

LAMDA
Learning And Mining from DatA

Reference II

Vapnik, V. N. and Chervonenkis, A. Y. (1971).
On the uniform convergence of relative frequencies of events to their probabilities.

Theory of Probability & Its Applications, 16(2):264–280.

Wald, A. (1949).
Note on the consistency of the maximum likelihood estimate.
The Annals of Mathematical Statistics, 20(4):595–601.

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.