Beyond Least-Squares: Fast Rates for Regularized Empirical Risk Minimization through Self-Concordance

Ulysse Marteau-Ferey

June 27, 2019

Joint work with Dmitrii Ostrovskii, Francis Bach and Alessandro Rudi

INRIA Paris – ÉNS Paris, CS department – PSL Research University

Presentation of the problem

Learning Problem

Setting: *input* X, *output* $Y \in \mathcal{Y}$

Linear Predictor: $f(x) = \theta \cdot \Phi(x)$, $\Phi(x) \in \mathcal{H}$ *feature map*, $\mathcal{H}$ **infinite dimensional**

Problem: Find

$$\theta^\star \in \arg\min_{\theta \in \mathcal{H}} L(\theta), \qquad L(\theta) = \mathbb{E}\left[\ell(Y, \theta \cdot \Phi(X))\right]$$

$\ell(\cdot,\cdot)$ loss function, (X, Y) unknown, n i.i.d. samples $(x_i, y_i)_{1 \leqslant i \leqslant n}$.

Basic assumption: $\mathcal{H}$ Hilbert space, $Y, \Phi(X)$ bounded.

Regularized Empirical Risk minimization

Problem

$$\theta^\star \in \arg\min_{\theta\in\mathcal{H}} L(\theta), \qquad L(\theta) = \mathbb{E}\left[\ell(Y, \theta \cdot \Phi(X))\right]$$

Classical Estimator : Regularized Empirical Risk Minimizer

$$\widehat{\theta}_\lambda = \arg\min_{\theta\in\mathcal{H}} \widehat{L}(\theta) + \frac{\lambda}{2}\|\theta\|^2, \qquad \widehat{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \theta \cdot \Phi(x_i))$$

λ: regularization parameter $\rightarrow$ *controls overfitting*

Question : Statistical performance of $\widehat{\theta}_\lambda$

$$L(\widehat{\theta}_\lambda) - L(\theta^\star) \leqslant C(n, \lambda)$$

Existing results

A first general result : slow rates

Assumption: $\ell(y, \cdot)$, $y \in \mathcal{Y}$ Lipschitz   Lipschitz constant: R.

Slow rates in $O(1/\sqrt{n})$ (Sridharan et al., 2009)

Bias-variance decomposition

$$L(\widehat{\theta}_\lambda) - L(\theta^\star) \leqslant \|\theta^\star\|^2 \; \lambda + \frac{\mathrm{R}^2\|\Phi\|_\infty^2}{\lambda n}$$

$$L(\widehat{\theta}_\lambda) - L(\theta^\star) \leqslant C \; \frac{1}{\sqrt{n}}, \qquad \lambda = c \; \frac{1}{\sqrt{n}}$$

$C = \mathrm{R}\|\Phi\|_\infty \|\theta^\star\|$ and $c = \mathrm{R}\|\Phi\|_\infty / \|\theta^\star\|$

Fast rates for Least-Squares

Assumption: square loss $\ell(y, y') = \frac{1}{2}(y - y')^2$.
Covariance operator: $\mathbf{\Sigma} = \mathbb{E}\left[\Phi(X) \otimes \Phi(X)\right]$, $\mathbf{\Sigma}_\lambda = \mathbf{\Sigma} + \lambda \mathbf{I}$

Two main quantities

Fast rates for Least-Squares

Assumption: square loss $\ell(y, y') = \frac{1}{2}(y - y')^2$.
Covariance operator: $\boldsymbol{\Sigma} = \mathbb{E}\left[\Phi(X) \otimes \Phi(X)\right]$, $\boldsymbol{\Sigma}_\lambda = \boldsymbol{\Sigma} + \lambda \mathbf{I}$

Two main quantities

- $\mathsf{b}_\lambda = \lambda^2 \|\boldsymbol{\Sigma}_\lambda^{-1/2} \theta^\star\|^2 \leqslant \lambda \|\theta^\star\|^2 \quad \rightarrow$ **bias** regularity of $\theta^\star$

Fast rates for Least-Squares

Assumption: square loss $\ell(y, y') = \frac{1}{2}(y - y')^2$.
Covariance operator: $\boldsymbol{\Sigma} = \mathbb{E}\left[\Phi(X) \otimes \Phi(X)\right]$, $\boldsymbol{\Sigma}_\lambda = \boldsymbol{\Sigma} + \lambda \mathbf{I}$

Two main quantities

- $\mathrm{b}_\lambda = \lambda^2 \|\boldsymbol{\Sigma}_\lambda^{-1/2} \theta^\star\|^2 \leqslant \lambda \|\theta^\star\|^2$ $\rightarrow$ **bias** regularity of $\theta^\star$
- $\mathrm{df}_\lambda = \mathrm{Tr}(\boldsymbol{\Sigma}_\lambda^{-1} \boldsymbol{\Sigma}) \leqslant \|\Phi\|_\infty^2 / \lambda$ $\rightarrow$ **variance** effective dimension

Fast rates for Least-Squares

Assumption: square loss $\ell(y, y') = \frac{1}{2}(y - y')^2$.
Covariance operator: $\boldsymbol{\Sigma} = \mathbb{E}\left[\Phi(X) \otimes \Phi(X)\right]$, $\boldsymbol{\Sigma}_\lambda = \boldsymbol{\Sigma} + \lambda \mathbf{I}$

Two main quantities

- $\mathrm{b}_\lambda = \lambda^2 \|\boldsymbol{\Sigma}_\lambda^{-1/2} \theta^\star\|^2 \leqslant \lambda \|\theta^\star\|^2 \quad \to$ **bias** — regularity of $\theta^\star$
- $\mathrm{df}_\lambda = \mathrm{Tr}(\boldsymbol{\Sigma}_\lambda^{-1} \boldsymbol{\Sigma}) \leqslant \|\Phi\|_\infty^2 / \lambda \quad \to$ **variance** — effective dimension

Fast rates up to $O(1/n)$ (Caponnetto and De Vito, 2007)

Bias-variance decomposition

$$L(\widehat{\theta}_\lambda) - L(\theta^\star) \leqslant \mathrm{b}_\lambda + \sigma^2 \frac{\mathrm{df}_\lambda}{n}, \qquad \sigma^2 \leqslant \|\theta^\star\|^2 \|\Phi\|_\infty^2 \|Y\|_\infty^2$$

Fast rates for Least-Squares

Assumption: square loss $\ell(y, y') = \frac{1}{2}(y - y')^2$.

Covariance operator: $\boldsymbol{\Sigma} = \mathbb{E}[\Phi(X) \otimes \Phi(X)]$, $\boldsymbol{\Sigma}_\lambda = \boldsymbol{\Sigma} + \lambda \mathbf{I}$

Two main quantities

- $\mathrm{b}_\lambda = \lambda^2 \|\boldsymbol{\Sigma}_\lambda^{-1/2} \theta^\star\|^2 \leqslant \lambda \|\theta^\star\|^2 \quad \rightarrow$ **bias** — regularity of $\theta^\star$
- $\mathrm{df}_\lambda = \mathrm{Tr}(\boldsymbol{\Sigma}_\lambda^{-1} \boldsymbol{\Sigma}) \leqslant \|\Phi\|_\infty^2 / \lambda \quad \rightarrow$ **variance** — effective dimension

Fast rates up to $O(1/n)$ (Caponnetto and De Vito, 2007)

Bias-variance decomposition

$$L(\widehat{\theta}_\lambda) - L(\theta^\star) \leqslant \mathrm{b}_\lambda + \sigma^2 \frac{\mathrm{df}_\lambda}{n}, \qquad \sigma^2 \leqslant \|\theta^\star\|^2 \|\Phi\|_\infty^2 \|Y\|_\infty^2$$

Example : for $\mathrm{df}_\lambda \approx d$

$$L(\widehat{\theta}_\lambda) - L(\theta^\star) \leqslant 2 \frac{\sigma^2 d}{n}, \qquad \lambda = \frac{\sigma^2 d}{\|\theta^\star\|^2} \frac{1}{n}$$

Interpretation of the key quantities

Eigen-decomposition: $\boldsymbol{\Sigma} = \sum_{i=0}^{+\infty} \sigma_i \; \psi_i \otimes \psi_i$ $\qquad \sigma_i \searrow 0$

$\theta^\star = \sum_{i=0}^{+\infty} \langle \theta^\star, \psi_i \rangle \; \psi_i$

Interpretation of the key quantities

Eigen-decomposition: $\boldsymbol{\Sigma} = \sum_{i=0}^{+\infty} \sigma_i \; \psi_i \otimes \psi_i$ $\qquad \sigma_i \searrow 0$

$\theta^\star = \sum_{i=0}^{+\infty} \langle \theta^\star, \psi_i \rangle \; \psi_i$

$\mathrm{b}_\lambda \rightarrow$ **bias: regularity of** $\theta^\star$ **w.r.t.** $\boldsymbol{\Sigma}$

$$\mathrm{b}_\lambda \leqslant \mathrm{L}^2 \lambda^{1+2r} \qquad \leftrightarrow \qquad \sum_{i=0}^{+\infty} \frac{\langle \theta^\star, \psi_i \rangle^2}{\sigma_i^{2r}} < \infty$$

Interpretation of the key quantities

Eigen-decomposition: $\boldsymbol{\Sigma} = \sum_{i=0}^{+\infty} \sigma_i \; \psi_i \otimes \psi_i$ $\qquad \sigma_i \searrow 0$

$\theta^\star = \sum_{i=0}^{+\infty} \langle \theta^\star, \psi_i \rangle \; \psi_i$

$\mathrm{b}_\lambda \rightarrow$ **bias: regularity of $\theta^\star$ w.r.t. $\boldsymbol{\Sigma}$**

$$\mathrm{b}_\lambda \leqslant \mathrm{L}^2 \lambda^{1+2r} \qquad \leftrightarrow \qquad \sum_{i=0}^{+\infty} \frac{\langle \theta^\star, \psi_i \rangle^2}{\sigma_i^{2r}} < \infty$$

$\mathrm{df}_\lambda \rightarrow$ **variance: eigenvalue decay of $\boldsymbol{\Sigma}$**

$$\mathrm{df}_\lambda \leqslant \mathrm{Q}^2 \lambda^{-1/\alpha} \qquad \leftrightarrow \qquad \sigma_i = O(i^{-\alpha})$$

Interpretation of the key quantities

Eigen-decomposition: $\boldsymbol{\Sigma} = \sum_{i=0}^{+\infty} \sigma_i \; \psi_i \otimes \psi_i$ $\qquad \sigma_i \searrow 0$

$\theta^\star = \sum_{i=0}^{+\infty} \langle \theta^\star, \psi_i \rangle \; \psi_i$

$\mathrm{b}_\lambda \to$ **bias: regularity of $\theta^\star$ w.r.t. $\boldsymbol{\Sigma}$**

$$\mathrm{b}_\lambda \leqslant \mathrm{L}^2 \lambda^{1+2r} \qquad \leftrightarrow \qquad \sum_{i=0}^{+\infty} \frac{\langle \theta^\star, \psi_i \rangle^2}{\sigma_i^{2r}} < \infty$$

$\mathrm{df}_\lambda \to$ **variance: eigenvalue decay of $\boldsymbol{\Sigma}$**

$$\mathrm{df}_\lambda \leqslant \mathrm{Q}^2 \lambda^{-1/\alpha} \qquad \leftrightarrow \qquad \sigma_i = O(i^{-\alpha})$$

Fast rates (Caponnetto and De Vito, 2007)

$$L(\widehat{\theta}_\lambda) - L(\theta^\star) \leqslant \mathrm{b}_\lambda + \sigma^2 \frac{\mathrm{df}_\lambda}{n}$$

Interpretation of the key quantities

Eigen-decomposition: $\boldsymbol{\Sigma} = \sum_{i=0}^{+\infty} \sigma_i \; \psi_i \otimes \psi_i$ $\qquad \sigma_i \searrow 0$

$\theta^\star = \sum_{i=0}^{+\infty} \langle \theta^\star, \psi_i \rangle \; \psi_i$

$\mathrm{b}_\lambda \rightarrow$ **bias: regularity of $\theta^\star$ w.r.t. $\boldsymbol{\Sigma}$**

$$\mathrm{b}_\lambda \leqslant \mathrm{L}^2 \lambda^{1+2r} \quad \leftrightarrow \quad \sum_{i=0}^{+\infty} \frac{\langle \theta^\star, \psi_i \rangle^2}{\sigma_i^{2r}} < \infty$$

$\mathrm{df}_\lambda \rightarrow$ **variance: eigenvalue decay of $\boldsymbol{\Sigma}$**

$$\mathrm{df}_\lambda \leqslant \mathrm{Q}^2 \lambda^{-1/\alpha} \quad \leftrightarrow \quad \sigma_i = O(i^{-\alpha})$$

Fast rates (Caponnetto and De Vito, 2007)

$$L(\widehat{\theta}_\lambda) - L(\theta^\star) \leqslant \mathrm{L}^2 \lambda^{1+2r} + \sigma^2 \frac{\mathrm{Q}^2 \lambda^{-1/\alpha}}{n}$$

Interpretation of the key quantities

Eigen-decomposition: $\boldsymbol{\Sigma} = \sum_{i=0}^{+\infty} \sigma_i \; \psi_i \otimes \psi_i$ $\quad\quad \sigma_i \searrow 0$

$\theta^\star = \sum_{i=0}^{+\infty} \langle \theta^\star, \psi_i \rangle \; \psi_i$

$\mathrm{b}_\lambda \rightarrow$ **bias: regularity of $\theta^\star$ w.r.t. $\boldsymbol{\Sigma}$**

$$\mathrm{b}_\lambda \leqslant \mathrm{L}^2 \lambda^{1+2r} \quad \leftrightarrow \quad \sum_{i=0}^{+\infty} \frac{\langle \theta^\star, \psi_i \rangle^2}{\sigma_i^{2r}} < \infty$$

$\mathrm{df}_\lambda \rightarrow$ **variance: eigenvalue decay of $\boldsymbol{\Sigma}$**

$$\mathrm{df}_\lambda \leqslant \mathrm{Q}^2 \lambda^{-1/\alpha} \quad \leftrightarrow \quad \sigma_i = O(i^{-\alpha})$$

Fast rates (Caponnetto and De Vito, 2007)

$$L(\widehat{\theta}_\lambda) - L(\theta^\star) \leqslant C \; n^{-\gamma}, \qquad \lambda = c \; n^{-\beta}, \qquad \gamma \in [1/2, 1].$$

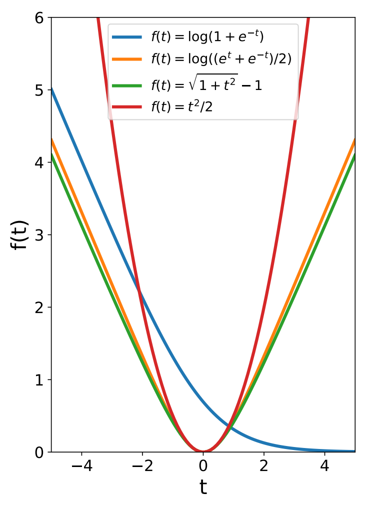
$\gamma = \frac{\alpha(1+2r)}{\alpha(1+2r)+1}$, $\beta = \alpha/(\alpha(1+2r)+1)$, $c = (\sigma \mathrm{Q}/\mathrm{L})^{2\beta}$ and $C = (\sigma^\gamma \mathrm{Q}^\gamma \mathrm{L}^{1-\gamma})^2$

Our contribution

Generalized Self Concordant functions

Regression: $\ell(y, y') = \psi(y - y')$

- Square loss: $\psi(t) = \frac{1}{2}t^2$
- Huber loss 1: $\psi(t) = \sqrt{1 + t^2} - 1$
- Huber loss 2: $\psi(t) = \log \frac{e^t + e^{-t}}{2}$

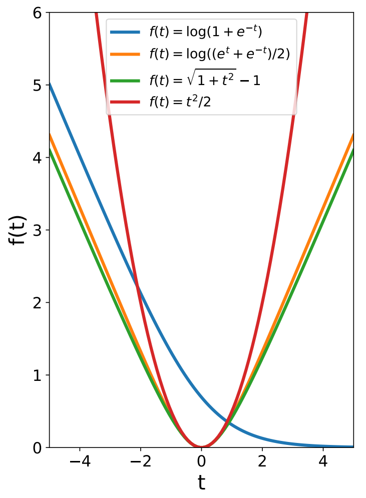


Regression: $\ell(y, y') = \psi(y - y')$

- Square loss: $\psi(t) = \frac{1}{2}t^2$
- Huber loss 1: $\psi(t) = \sqrt{1 + t^2} - 1$
- Huber loss 2: $\psi(t) = \log \frac{e^t + e^{-t}}{2}$

Classification:

- Logistic loss: $\ell(y, y') = \log(1 + e^{-yy'})$
- GLMs: $\ell(y, y') = -y' \cdot y + \log \int_{\mathcal{Y}} \exp(y' \cdot \tilde{y})\, d\mu(\tilde{y})$



Generalized Self Concordant functions

Regression: $\ell(y, y') = \psi(y - y')$

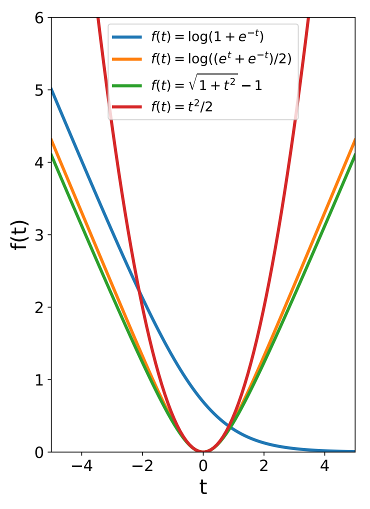
- Square loss: $\psi(t) = \frac{1}{2}t^2$
- Huber loss 1: $\psi(t) = \sqrt{1 + t^2} - 1$
- Huber loss 2: $\psi(t) = \log \frac{e^t + e^{-t}}{2}$

Classification:

- Logistic loss: $\ell(y, y') = \log(1 + e^{-yy'})$
- GLMs: $\ell(y, y') = -y' \cdot y + \log \int_{\mathcal{Y}} \exp(y' \cdot \tilde{y})\, d\mu(\tilde{y})$

Defintion : GSC functions (Bach, 2010)

$$\forall y \in \mathcal{Y}, \;\; \ell^{(3)}(y, \cdot) \leqslant R\ell''(y, \cdot)$$



Fast rates for GSC functions

Assumption: ℓ is GSC

Hessian at optimum: $\mathbf{H} = \mathbb{E}\left[\ell''(Y, \theta^\star \cdot \Phi(X))\ \Phi(X) \otimes \Phi(X)\right]$, $\mathbf{H}_\lambda = \mathbf{H} + \lambda \mathbf{I}$

Fisher information $\mathbf{G} = \mathbb{E}\left[\ell'(Y, \theta^\star \cdot \Phi(X))^2\ \Phi(X) \otimes \Phi(X)\right]$

Fast rates for GSC functions

Assumption: ℓ is GSC

Hessian at optimum: $\mathbf{H} = \mathbb{E}\left[\ell''(Y, \theta^\star \cdot \Phi(X))\ \Phi(X) \otimes \Phi(X)\right]$, $\mathbf{H}_\lambda = \mathbf{H} + \lambda \mathbf{I}$

Fisher information $\mathbf{G} = \mathbb{E}\left[\ell'(Y, \theta^\star \cdot \Phi(X))^2\ \Phi(X) \otimes \Phi(X)\right]$

Two main quantities

- $\mathrm{b}_\lambda = \lambda^2 \|\mathbf{H}_\lambda^{-1/2}\theta^\star\|^2 \leqslant \lambda\|\theta^\star\|^2 \to$ **bias** — regularity of $\theta^\star$
- $\mathrm{df}_\lambda = \mathrm{Tr}(\mathbf{H}_\lambda^{-1/2}\mathbf{G}\mathbf{H}_\lambda^{-1/2}) \leqslant C/\lambda \to$ **variance** — effective dimension

Fast rates for GSC functions

Assumption: ℓ is GSC

Hessian at optimum: $\mathbf{H} = \mathbb{E}\left[\ell''(Y, \theta^\star \cdot \Phi(X))\, \Phi(X) \otimes \Phi(X)\right]$, $\mathbf{H}_\lambda = \mathbf{H} + \lambda \mathbf{I}$

Fisher information $\mathbf{G} = \mathbb{E}\left[\ell'(Y, \theta^\star \cdot \Phi(X))^2\, \Phi(X) \otimes \Phi(X)\right]$

Two main quantities

- $\mathrm{b}_\lambda = \lambda^2 \|\mathbf{H}_\lambda^{-1/2}\theta^\star\|^2 \leqslant \lambda\|\theta^\star\|^2 \to$ **bias** — regularity of $\theta^\star$
- $\mathrm{df}_\lambda = \mathrm{Tr}(\mathbf{H}_\lambda^{-1/2}\mathbf{G}\mathbf{H}_\lambda^{-1/2}) \leqslant C/\lambda \to$ **variance** — effective dimension

Fast rates up to $O(1/n)$ (Marteau-Ferey et al., 2019)

Bias-variance decomposition

$$L(\widehat{\theta}_\lambda) - L(\theta^\star) \leqslant \mathrm{b}_\lambda + \frac{\mathrm{df}_\lambda}{n}$$

Fast rates for GSC functions

Assumption: ℓ is GSC

Hessian at optimum: $\mathbf{H} = \mathbb{E}\left[\ell''(Y, \theta^\star \cdot \Phi(X))\, \Phi(X) \otimes \Phi(X)\right]$, $\mathbf{H}_\lambda = \mathbf{H} + \lambda \mathbf{I}$

Fisher information $\mathbf{G} = \mathbb{E}\left[\ell'(Y, \theta^\star \cdot \Phi(X))^2\, \Phi(X) \otimes \Phi(X)\right]$

Two main quantities

- $\mathrm{b}_\lambda = \lambda^2 \|\mathbf{H}_\lambda^{-1/2} \theta^\star\|^2 \leqslant \lambda \|\theta^\star\|^2 \to$ **bias** — regularity of $\theta^\star$
- $\mathrm{df}_\lambda = \mathrm{Tr}(\mathbf{H}_\lambda^{-1/2} \mathbf{G} \mathbf{H}_\lambda^{-1/2}) \leqslant C/\lambda \to$ **variance** — effective dimension

Fast rates up to $O(1/n)$ (Marteau-Ferey et al., 2019)

Bias-variance decomposition

$$L(\widehat{\theta}_\lambda) - L(\theta^\star) \leqslant \mathrm{b}_\lambda + \frac{\mathrm{df}_\lambda}{n}$$

Example : for $\mathrm{df}_\lambda \approx \sigma^2\ d$

$$L(\widehat{\theta}_\lambda) - L(\theta^\star) \leqslant 2\frac{\sigma^2 d}{n}, \qquad \lambda = \frac{\sigma^2 d}{\|\theta^\star\|^2\ n}$$

Link with least-squares

Assumption: ℓ is GSC

Hessian at optimum: $\mathbf{H} = \mathbb{E}\left[\ell''(Y, \theta^\star \cdot \Phi(X))\ \Phi(X) \otimes \Phi(X)\right]$

Fisher information $\mathbf{G} = \mathbb{E}\left[\ell'(Y, \theta^\star \cdot \Phi(X))^2\ \Phi(X) \otimes \Phi(X)\right]$

Link with least-squares

Assumption: $\ell(y, y') = \frac{1}{2}(y - y')^2$ $\quad\quad$ $\mathbf{\Sigma} = \mathbb{E}\left[\Phi(X) \otimes \Phi(X)\right]$

Hessian at optimum: $\mathbf{H} = \mathbb{E}\left[1\ \Phi(X) \otimes \Phi(X)\right] = \mathbf{\Sigma}$

Fisher information $\quad \mathbf{G} = \mathbb{E}\left[(Y\Phi(X) \cdot \theta^\star)^2\ \Phi(X) \otimes \Phi(X)\right] \preceq \sigma^2 \mathbf{\Sigma}$

Link with least-squares

Assumption: $\ell(y, y') = \frac{1}{2}(y - y')^2$ $\quad\quad\quad\quad$ $\mathbf{\Sigma} = \mathbb{E}\left[\Phi(X) \otimes \Phi(X)\right]$

Hessian at optimum: $\mathbf{H} = \mathbf{\Sigma}$

Fisher information $\mathbf{G} \preceq \sigma^2 \mathbf{\Sigma}$ $\quad\quad\quad\quad$ $\sigma^2 = \|\theta^\star\|^2 \|\Phi\|_\infty^2 \|Y\|_\infty^2$

Link with least-squares

Assumption: $\ell(y, y') = \frac{1}{2}(y - y')^2$ $\qquad \boldsymbol{\Sigma} = \mathbb{E}\left[\Phi(X) \otimes \Phi(X)\right]$

Hessian at optimum: $\mathbf{H} = \boldsymbol{\Sigma}$

Fisher information $\mathbf{G} \preceq \sigma^2 \boldsymbol{\Sigma}$ $\qquad \sigma^2 = \|\theta^\star\|^2 \|\Phi\|_\infty^2 \|Y\|_\infty^2$

Two main quantities

- $\mathrm{b}_\lambda = \lambda^2 \|\mathbf{H}_\lambda^{-1/2} \theta^\star\|^2$ — regularity of $\theta^\star$
- $\mathrm{df}_\lambda = \mathrm{Tr}(\mathbf{H}_\lambda^{-1/2} \mathbf{G} \mathbf{H}_\lambda^{-1/2})$ — effective dimension

Link with least-squares

Assumption: $\ell(y, y') = \frac{1}{2}(y - y')^2$ $\qquad$ $\boldsymbol{\Sigma} = \mathbb{E}\left[\Phi(X) \otimes \Phi(X)\right]$

Hessian at optimum: $\mathbf{H} = \boldsymbol{\Sigma}$

Fisher information $\mathbf{G} \preceq \sigma^2 \boldsymbol{\Sigma}$ $\qquad$ $\sigma^2 = \|\theta^\star\|^2 \|\Phi\|_\infty^2 \|Y\|_\infty^2$

Two main quantities

- $\mathrm{b}_\lambda = \lambda^2 \|\boldsymbol{\Sigma}_\lambda^{-1/2} \theta^\star\|^2$ — regularity of $\theta^\star$
- $\mathrm{df}_\lambda \leqslant \sigma^2 \operatorname{Tr}(\boldsymbol{\Sigma}_\lambda^{-1/2} \boldsymbol{\Sigma} \boldsymbol{\Sigma}_\lambda^{-1/2})$ — effective dimension

Link with least-squares

Assumption: $\ell(y, y') = \frac{1}{2}(y - y')^2$ $\quad\quad \boldsymbol{\Sigma} = \mathbb{E}\left[\Phi(X) \otimes \Phi(X)\right]$

Hessian at optimum: $\mathbf{H} = \boldsymbol{\Sigma}$

Fisher information $\mathbf{G} \preceq \sigma^2 \boldsymbol{\Sigma}$ $\quad\quad \sigma^2 = \|\theta^\star\|^2 \|\Phi\|_\infty^2 \|Y\|_\infty^2$

Two main quantities

- $\mathrm{b}_\lambda = \lambda^2 \|\boldsymbol{\Sigma}_\lambda^{-1/2} \theta^\star\|^2 = \mathrm{b}_\lambda^{\mathtt{ls}}$ — regularity of $\theta^\star$
- $\mathrm{df}_\lambda \leqslant \sigma^2 \operatorname{Tr}(\boldsymbol{\Sigma}_\lambda^{-1/2} \boldsymbol{\Sigma} \boldsymbol{\Sigma}_\lambda^{-1/2}) = \sigma^2 \mathrm{df}_\lambda^{\mathtt{ls}}$ — effective dimension

Fast rates up to $O(1/n)$ (Marteau-Ferey et al., 2019)

Bias-variance decomposition

$$L(\widehat{\theta}_\lambda) - L(\theta^\star) \leqslant \mathrm{b}_\lambda + \frac{\mathrm{df}_\lambda}{n}$$

Link with least-squares

Assumption: $\ell(y, y') = \frac{1}{2}(y - y')^2$ $\quad\quad \mathbf{\Sigma} = \mathbb{E}\left[\Phi(X) \otimes \Phi(X)\right]$

Hessian at optimum: $\mathbf{H} = \mathbf{\Sigma}$

Fisher information $\mathbf{G} \preceq \sigma^2 \mathbf{\Sigma}$ $\quad\quad \sigma^2 = \|\theta^\star\|^2 \|\Phi\|_\infty^2 \|Y\|_\infty^2$

Two main quantities

- $\mathrm{b}_\lambda = \lambda^2 \|\mathbf{\Sigma}_\lambda^{-1/2} \theta^\star\|^2 = \mathrm{b}_\lambda^{\mathtt{ls}}$ — regularity of $\theta^\star$
- $\mathrm{df}_\lambda \leqslant \sigma^2 \operatorname{Tr}(\mathbf{\Sigma}_\lambda^{-1/2} \mathbf{\Sigma} \mathbf{\Sigma}_\lambda^{-1/2}) = \sigma^2 \mathrm{df}_\lambda^{\mathtt{ls}}$ — effective dimension

Fast rates up to $O(1/n)$ (Marteau-Ferey et al., 2019)

Bias-variance decomposition

$$L(\widehat{\theta}_\lambda) - L(\theta^\star) \leqslant \mathrm{b}_\lambda^{\mathtt{ls}} + \frac{\sigma^2 \ \mathrm{df}_\lambda^{\mathtt{ls}}}{n}$$

Conclusion

Thank you for your attention !

Poster 175