MATH888: High-dimensional probability and statistics Fall 2021

Lecture 16 — Oct 13, 2021

Sebastien Roch, UW-Madison *Scribe: Jiaxin Hu*

1 Overview

In the last lecture we studied the linear regression model with sub-Gaussian noise and proved the upper bound of the mean squared error (MSE) of the predicted values with least squared estimate.

In this lecture we derive the upper bound for the least squared estimate itself and relax the linear assumption on the predictor $\mathbf{X}$ by considering $\varphi(\mathbf{X})$ with non-linear functions φ.

2 Review of Matrix Perturbation theory

We review the matrix perturbation theory before going to the linear regression.

Definition 1 (Induced Norm)**.** *The 2-norm of a matrix* $\mathbf{A} \in \mathbb{R}^{n\times m}$ *is*

$$\|\mathbf{A}\|_2 = \max_{\mathbf{0}\neq\mathbf{x}\in\mathbb{R}^m} \frac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{x}\|} = \max_{\mathbf{x}\in\mathbb{S}^{m-1}} \|\mathbf{A}\mathbf{x}\|.$$

Theorem 1 (Spectral Theorem)**.** *Let* $\mathbf{A} \in \mathbb{R}^{d\times d}$ *be a symmetric matrix, that is,* $\mathbf{A}^T = \mathbf{A}$*. Then* $\mathbf{A}$ *has* d *orthonormal eigenvectors* $\mathbf{q}_1, ..., \mathbf{q}_d$ *with corresponding (not necessarily distinct) real eigenvalues* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$*. In matrix form, this is written as the matrix factorization*

$$\mathbf{A} = \mathbf{Q}\Lambda\mathbf{Q}^T = \sum_{i=1}^{d} \lambda_i \mathbf{q}_i \mathbf{q}_i^T,$$

where $\mathbf{Q}$ *has columns* $\mathbf{q}_1, ..., \mathbf{q}_d$ *and* $\Lambda = diag(\lambda_1, ..., \lambda_d)$*. We refer to this factorization as a spectral decomposition of* $\mathbf{A}$*.*

Definition 2 (Rayleigh Quotient)**.** *Let* $\mathbf{A} \in \mathbb{R}^{d\times d}$ *be a symmetric matrix. The Rayleigh quotient is defined as*

$$\mathcal{R}_{\mathbf{A}}(\mathbf{u}) = \frac{\langle \mathbf{u}, \mathbf{A}\mathbf{u}\rangle}{\langle \mathbf{u}, \mathbf{u}\rangle},$$

which is defined for any $\mathbf{0} \neq \mathbf{u} \in \mathbb{R}^d$*.*

Theorem 2 (Courant-Fischer)**.** *Let* $\mathbf{A} \in \mathbb{R}^{d\times d}$ *be a symmetric matrix with spectral decomposition* $\mathbf{A} = \sum_{i=1}^d \lambda_i \mathbf{v}_i \mathbf{v}_i^T$ *where* $\lambda_1 \geq \cdots \geq \lambda_d$*. For each* $k = 1, ..., d$*, define the subspace*

$$\mathcal{V}_k = span(\mathbf{v}_1, ..., \mathbf{v}_k) \text{ and } \mathcal{W}_{d-k+1} = span(\mathbf{v}_k, ..., \mathbf{v}_d).$$

Then, for all $k = 1, ..., d$*,*

$$\lambda_k = \min_{\mathbf{u}\in\mathcal{V}_k} \mathcal{R}_{\mathbf{A}}(\mathbf{u}) = \max_{\mathbf{u}\in\mathcal{W}_{d-k+1}} \mathcal{R}_{\mathbf{A}}(\mathbf{u}).$$

Furthermore we have the following min-max formulas, which do not depend on the choice of spectral decomposition, for all $k = 1, ..., d$,

$$\lambda_k = \max_{dim(\mathcal{V})=k} \min_{\mathbf{u}\in\mathcal{V}} \mathcal{R}_{\mathbf{A}}(\mathbf{u}) = \min_{dim(\mathcal{W})=d-k+1} \max_{\mathbf{u}\in\mathcal{W}} \mathcal{R}_{\mathbf{A}}(\mathbf{u}).$$

Lemma 3 (Weyl)**.** *Let $\mathbf{A} \in \mathbb{R}^{d\times d}$ and $\mathbf{B} \in \mathbb{R}^{d\times d}$ be symmetric matrices. Then, for all $j = 1, ..., d$*

$$\max_{j\in[d]} |\lambda_j(\mathbf{B}) - \lambda_j(\mathbf{A})| \leq \|\mathbf{B} - \mathbf{A}\|_2 ,$$

where $\|\mathbf{X}\|_2$ is the induced 2-norm of $\mathbf{X}$.

Theorem 4 (Davis-Kahan)**.** *Let $\mathbf{A} \in \mathbb{R}^{d\times d}$ and $\mathbf{B} \in \mathbb{R}^{d\times d}$ be symmetric matrices. For an $i \in [d]$, assume that*

$$\delta := \min_{j\neq i} |\lambda_i(\mathbf{A}) - \lambda_j(A)| > 0.$$

Then,

$$\min_{s\in\{+1,-1\}} \|\mathbf{v}_i(\mathbf{A}) - s\mathbf{v}_i(\mathbf{B})\|^2 \leq \frac{8\,\|\mathbf{A} - \mathbf{B}\|_2^2}{\delta^2}.$$

See the slide on the course website for more details.

3 Error bound for the least square estimator

Consider n observations $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ where $\mathbf{x}_i \in \mathbb{R}^p$ and

$$y_i = \mathbf{x}_i^T \boldsymbol{\theta}^* + \epsilon_i,$$

where $\boldsymbol{\theta}^* \in \mathbb{R}^p$ is the unknown coefficient and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n) \in \mathbb{R}^n$ is a mean-zero random vector. Rewrite the linear regression model in the matrix form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta}^* + \boldsymbol{\epsilon},$$

where $\mathbf{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$ and $\mathbf{X} \in \mathbb{R}^{n\times p}$ is a matrix with $\mathbf{x}_i^T$ on the row i. The least squared estimator of $\boldsymbol{\theta}^*$ is

$$\hat{\boldsymbol{\theta}}^{LS} \in \arg\min_{\hat{\boldsymbol{\theta}}\in\mathbb{R}^p} \frac{1}{n} \left\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}}\right\|_2^2 .$$

We can choose $\hat{\boldsymbol{\theta}}^{LS} = \mathbf{X}^+\mathbf{y}$, where $\mathbf{X}^+$ is the persudo inverse of $\mathbf{X}$. Note that $\hat{\boldsymbol{\theta}}^{LS}$ is the minimal norm solution and is not necessarily unique.

First, we recall the upper bound for the predicted values with least squared estimate.

Theorem 5 (Upper bound of MSE for predicted values)**.** *Suppose $\boldsymbol{\epsilon}$ is a sub-Gaussian vector with $\|\boldsymbol{\epsilon}\|_{\psi_2} \leq K$. Then, with probability $1 - \delta$, we have*

$$MSE(\mathbf{X}\hat{\boldsymbol{\theta}}^{LS}) = \frac{1}{n} \left\|\mathbf{X}\hat{\boldsymbol{\theta}}^{LS} - \mathbf{X}\boldsymbol{\theta}^*\right\|_2^2 \lesssim \frac{K^2}{n} \left(rank(\mathbf{X}) + \log\frac{1}{\delta}\right).$$

Next, we derive a corollary of Theorem 5 to find the upper bound for $\hat{\boldsymbol{\theta}}^{LS}$ itself rather than the predicted values.

Corollary 6 (Upper bound of MSE for least squared estimate)**.** *Suppose $\boldsymbol{\epsilon}$ is a sub-Gaussian vector with $\|\boldsymbol{\epsilon}\|_{\psi_2} \leq K$. Assume $p \leq n$ and $\mathbf{B} = \frac{\mathbf{X}^T\mathbf{X}}{n}$ has rank p. Then, with probability $1-\delta$, we have*

$$\left\|\hat{\boldsymbol{\theta}}^{LS} - \boldsymbol{\theta}^*\right\|_2^2 \lesssim \frac{1}{\lambda_{\min}(\mathbf{B})} \frac{K^2}{n} \left(p + \log\frac{1}{\delta}\right),$$

where $\lambda_{\min}(\mathbf{B})$ is the smallest eigenvalue of $\mathbf{B}$.

Proof of Corollary 6. By Courant-Fischer characterization, we have

$$\lambda_{\min}(\mathbf{B}) = \min_{\mathbf{0}\neq\mathbf{u}\in\mathbb{R}^p} \frac{\langle \mathbf{u}, \mathbf{B}\mathbf{u}\rangle}{\langle \mathbf{u}, \mathbf{u}\rangle}.$$

Note that

$$\frac{1}{n}\left\|\mathbf{X}\hat{\boldsymbol{\theta}}^{LS} - \mathbf{X}\boldsymbol{\theta}^*\right\|_2^2 = \left(\hat{\boldsymbol{\theta}}^{LS} - \boldsymbol{\theta}^*\right)^T \frac{\mathbf{X}^T\mathbf{X}}{n}\left(\hat{\boldsymbol{\theta}}^{LS} - \boldsymbol{\theta}^*\right) = \left(\hat{\boldsymbol{\theta}}^{LS} - \boldsymbol{\theta}^*\right)^T \mathbf{B}\left(\hat{\boldsymbol{\theta}}^{LS} - \boldsymbol{\theta}^*\right),$$

where the second equation follows by the definition of $\mathbf{B}$. Therefore, we have

$$\left\|\hat{\boldsymbol{\theta}}^{LS} - \boldsymbol{\theta}^*\right\|_2^2 \leq \frac{1}{\lambda_{\min}(\mathbf{B})}\mathrm{MSE}(\mathbf{X}\hat{\boldsymbol{\theta}}^{LS}),$$

which completes the proof by Theorem 5. □

4 Non-linear regression and oracle bounds

The linear assumption on $\mathbf{X}$ is to restrictive for practical application. We model the non-linearity in this section.

Consider the dictionary $\mathcal{H} = \{\varphi_1, \ldots, \varphi_M\}$ where $\varphi_i : \mathbb{R}^p \mapsto \mathbb{R}$ for all $i \in [M]$. Assume the response $y(\mathbf{x}) \sim \sum_{i=1}^M \theta_i\varphi_i(\mathbf{x})$. Given n observations $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, we can obtain the least squared estimator via

$$\hat{\boldsymbol{\theta}}^{LS} \in \arg\min_{\hat{\boldsymbol{\theta}}\in\mathbb{R}^M} \left\|\mathbf{y} - \Phi\hat{\boldsymbol{\theta}}\right\|_2^2,$$

where the i-th row of Φ is $\Phi_i^T = (\varphi_1(\mathbf{x}_i), \ldots, \varphi_M(\mathbf{x}_i))$.

Example 1 (Polynomial regression)**.** *Consider the polynomial regression with $\varphi_1(x) = 1, \varphi_2(x) = x, \varphi_3(x) = x^2$. Then, with observations $\{x_i, y_i\}_{i=1}^n$we have least squared estimator*

$$\hat{\boldsymbol{\theta}}^{LS} \in \arg\min_{\hat{\boldsymbol{\theta}}\in\mathbb{R}^3} \frac{1}{n}\sum_{i=1}^n \left[y_i - \left(\theta_1 + \theta_2 x_i + \theta_3 x_i^2\right)\right]^2.$$

Assume the data is generated from the class of function you assumed. We describe such assumption using the word "oracle". Next, we establish the oracle bounds for the predicted values of non-linear regression.

Theorem 7 (Upper bound of MSE for predicted values in non-linear regression)**.** *Suppose the data is of form $y_i = f(\mathbf{x}_i) + \epsilon_i$ for $i \in [n]$ with $\|\boldsymbol{\epsilon}\|_{\psi_2} = K$. Fix the dictionary $\mathcal{H}$. Then, with probability $1 - \delta$, we havev*

$$MSE(\Phi\hat{\boldsymbol{\theta}}^{LS}) \leq \inf_{\hat{\boldsymbol{\theta}}} MSE(\Phi\hat{\boldsymbol{\theta}}) + \frac{CK^2}{n}\left(M + \log\frac{1}{\delta}\right),$$

where $MSE(\Phi\hat{\boldsymbol{\theta}}) = \frac{1}{n}\left\|\Phi\hat{\boldsymbol{\theta}} - \mathbf{f}\right\|_2^2$ and $\mathbf{f} = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))$. Note that the inf *chooses the best $\boldsymbol{\theta}$ in the class. If the data is indeed generated from this class (realizable), then the term $\inf_{\hat{\boldsymbol{\theta}}} MSE(\Phi\hat{\boldsymbol{\theta}})$ degenerates to 0.*