MATH888: High-dimensional probability and statistics Fall 2021

Lecture 16 — October 13, 2021

Sebastien Roch, UW-Madison *Scribe: Yuchen Zeng, Sebastien Roch*

1 Overview

In the last lecture we (a) studied least square estimator of linear regression model, and (b) proved the upper bound of mean squared error of estimated predicted variables.

In this lecture we (a) derive the concentration bound of the least square estimator of linear regression model, and (b) investigate a more general case: non-linear models.

2 Reveiw of eigenvalues

We begin by reviewing a few basic factors about matrices.

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

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

Theorem 2 (Spectral Theorem)**.** *Let $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ be a symmetric matrix, that is, $\boldsymbol{A}^\top = \boldsymbol{A}$. Then $\boldsymbol{A}$ has d orthonormal eigenvectors $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_d$ 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*

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

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

Definition 3 (Rayleigh Quotient)**.** *Let $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ be a symmetric matrix. The Rayleigh quotient is defined as*

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

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

Theorem 4 (Courant-Fischer)**.** *Let $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ be a symmetric matrix with spectral decomposition $\boldsymbol{A} = \sum_{i=1}^d \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^\top$ where $\lambda_1 \geq \cdots \geq \lambda_d$. For each $k = 1, \ldots, d$, define the subspace*

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

Then, for all $k = 1, \ldots, d$,

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

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

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

Lemma 5 (Weyl's Inequality)**.** *Let* $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ *and* $\boldsymbol{B} \in \mathbb{R}^{d\times d}$ *be symmetric matrices. Then, for all* $j = 1, \ldots, d$,

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

where $\|\boldsymbol{C}\|_2$ *is the induced 2-norm of* $\boldsymbol{C}$.

Theorem 6 (Davis-Kahan)**.** *Let* $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ *and* $\boldsymbol{B} \in \mathbb{R}^{d\times d}$ *be symmetric matrices. For an* $i \in \{1, \ldots, d\}$*, assume that*

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

then

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

3 Controlling the coefficients

Assume we have n observation pairs $\{(\mathbf{x}_i, \mathrm{y}_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^p$, $\mathrm{y}_i = \mathbf{x}_i^\top \boldsymbol{\theta}^\star + \varepsilon_i$, $\boldsymbol{\theta}^\star \in \mathbb{R}^p$ is unknown, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n) \in \mathbb{R}^n$ is of mean zero. Denote the $\mathbf{y} = (\mathrm{y}_1, \ldots, \mathrm{y}_n)$, $\mathbf{X}$ as the matrix with $\mathbf{x}_i^\top$ on row i. Therefore, the linear regression model can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon}.$$

The least square estimator of $\boldsymbol{\theta}^\star$ is defined as $\widehat{\boldsymbol{\theta}}^{\mathrm{LS}} \in \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2$. Then $\widehat{\boldsymbol{\theta}}^{\mathrm{LS}} = \mathbf{X}\mathbf{y}$ is the minimum norm solution in the set of least square estimators, which is generally unique.

In last lecture, we proved the following theorem, which provides a upper bound for the mean square error of the estimations.

Theorem 7. *Suppose* $\boldsymbol{\varepsilon}$ *is subgaussian,* i.e, $\|\boldsymbol{\varepsilon}\|_{\psi_2} \le K$*. Then* w.p.$1 - \delta$*, we have*

$$\mathrm{MSE}(\mathbf{X}\widehat{\boldsymbol{\theta}}^{\mathrm{LS}}) \lesssim \frac{K^2}{n}(\mathrm{rank}(\mathbf{X}) + \log(\frac{1}{\delta})),$$

where $\mathrm{MSE}(\mathbf{X}\widehat{\boldsymbol{\theta}}^{\mathrm{LS}}) = \frac{1}{n}\left\|\mathbf{X}\widehat{\boldsymbol{\theta}}^{\mathrm{LS}} - \mathbf{X}\boldsymbol{\theta}^\star\right\|_2^2$.

We derive a corollary which provides an concentration bound for $\widehat{\boldsymbol{\theta}}^{\mathrm{LS}}$, based on the theorem above.

Corollary 8. *Suppose* $p \le n$ *and* $\mathbf{B} = \frac{\mathbf{X}^\top\mathbf{X}}{n}$ *has rank* p*. Then* w.p.$1 - \delta$,

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

Proof. By Courant-Fischer characterization, we have

$$\lambda_{\min}(\mathbf{B}) = \min_{\boldsymbol{u}\in\mathbb{R}^p} \frac{\langle \boldsymbol{u}, \mathbf{B}\boldsymbol{u}\rangle}{\langle \boldsymbol{u}, \boldsymbol{u}\rangle}. \tag{1}$$

Note that

$$\frac{1}{n}\left\|\mathbf{X}\widehat{\boldsymbol{\theta}}^{\mathrm{LS}} - \mathbf{X}\boldsymbol{\theta}^\star\right\|_2^2 = (\widehat{\boldsymbol{\theta}}^{\mathrm{LS}} - \boldsymbol{\theta}^\star)^\top \frac{\mathbf{X}^\top\mathbf{X}}{n}(\widehat{\boldsymbol{\theta}}^{\mathrm{LS}} - \boldsymbol{\theta}^\star) = (\widehat{\boldsymbol{\theta}}^{\mathrm{LS}} - \boldsymbol{\theta}^\star)^\top \mathbf{B}(\widehat{\boldsymbol{\theta}}^{\mathrm{LS}} - \boldsymbol{\theta}^\star). \tag{2}$$

Combining (1) and (2) yields

$$\left\|\widehat{\boldsymbol{\theta}}^{\mathrm{LS}} - \boldsymbol{\theta}^\star\right\|_2^2 \le \frac{1}{\lambda_{\min}(\mathbf{B})}\,\mathrm{MSE}(\mathbf{X}\widehat{\boldsymbol{\theta}}^{\mathrm{LS}}).$$

Plugging the upper bound of $\mathrm{MSE}(\mathbf{X}\widehat{\boldsymbol{\theta}}^{\mathrm{LS}})$ in Theorem 7 completes the proof. □

4 Modeling non-linearities and an oracle bound

In this section, we extend the results to some nonlinear models. Consider a dictionary $\mathcal{H} = \{\phi_1, \ldots, \phi_M\}$, where $\phi_i : \mathbb{R}^p \to \mathbb{R}$ is a map, for $i = 1, \ldots, M$. Then we can consider functions of $\mathbf{x}$ which can be written as $\sum_{i=1}^M \theta_i\phi(\mathbf{x})$. Given n observation pairs $\{(\mathbf{x}_i, \mathrm{y}_i)\}_{i=1}^M$, we minimize the mean square error and obtain the least square estimator

$$\widehat{\boldsymbol{\theta}}^{\mathrm{LS}} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^M} \|\mathbf{y} - \Phi\boldsymbol{\theta}\|_2^2,$$

where the ith row of Φ is $\Phi_i^\top = (\phi_1(\mathbf{x}_i), \ldots, \phi_M(\mathbf{x}_i))$.

Next, we provide a quick example to motivate the analysis on nonlinear model.

Example 1. *Suppose we observe a quadratic pattern between $\boldsymbol{x}$ and* y*. We select the features to be $\phi_1(x) = 1, \phi_2(x) = x, \phi_3(x) = x^2$. Then we have $\widehat{\boldsymbol{\theta}}^{\mathrm{LS}} \in \arg\min_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^n \left(\mathrm{y}_i - (\theta_1 + \theta_2\boldsymbol{x}_i + \theta_3\boldsymbol{x}_i^2)\right)^2$.*

Lastly, we present a theorem, which provides an upper bound for mean square error of the estimation under non-linear models' cases—without assuming that the data is generated by the class of models used to fit it. This type of bound is called an *oracle bound*, and we will have more to say about them later in the course.

Theorem 9 (See Theorem 3.3 in [1])**.** *Assume the data $\mathrm{y}_i = f(\mathbf{x}_i) + \varepsilon_i$ with $\|\boldsymbol{\varepsilon}\|_{\psi_2} \le K$, for all $i = 1, \ldots, n$. Fix the dictionary $\mathcal{H}$. Then* w.p. $1 - \delta$*,*

$$\mathrm{MSE}(\Phi\widehat{\boldsymbol{\theta}}^{\mathrm{LS}}) \le \inf_{\boldsymbol{\theta}\in\mathbb{R}^p} \mathrm{MSE}(\Phi\boldsymbol{\theta}) + \frac{cK^2}{n}(M + \log(\frac{1}{\delta})),$$

where $\mathrm{MSE}(\Phi\boldsymbol{\theta}) = \frac{1}{n}\|\Phi\boldsymbol{\theta} - \boldsymbol{f}\|_2^2$, and $\boldsymbol{f} = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))$.

Proof. A solution to $\inf_{\boldsymbol{\theta}\in\mathbb{R}^p} \mathrm{MSE}(\Phi\boldsymbol{\theta})$ is given by $\bar{\boldsymbol{\theta}} = \Phi^+\boldsymbol{f}$, which corresponds to projecting $\boldsymbol{f}$ onto the column space of Φ. By the properties of the orthogonal projection, for any other vector $\Phi\boldsymbol{\theta}$ in the span of the columns of Φ, we have the orthogonality

$$\langle \Phi\boldsymbol{\theta} - \Phi\bar{\boldsymbol{\theta}}, \boldsymbol{f} - \Phi\bar{\boldsymbol{\theta}}\rangle = 0.$$

That applies in particular to the choice $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{\mathrm{LS}}$. Hence, by Pythagoras,

$$\begin{aligned}\frac{1}{n}\left\|\Phi\widehat{\boldsymbol{\theta}}^{\mathrm{LS}} - \boldsymbol{f}\right\|_2^2 &= \frac{1}{n}\left\|\Phi\widehat{\boldsymbol{\theta}}^{\mathrm{LS}} - \Phi\bar{\boldsymbol{\theta}} + \Phi\bar{\boldsymbol{\theta}} - \boldsymbol{f}\right\|_2^2 \\ &= \frac{1}{n}\left\|\Phi\bar{\boldsymbol{\theta}} - \boldsymbol{f}\right\|_2^2 + \frac{1}{n}\left\|\Phi\widehat{\boldsymbol{\theta}}^{\mathrm{LS}} - \Phi\bar{\boldsymbol{\theta}}\right\|_2^2.\end{aligned}$$

The claim then follows from the definition of $\bar{\boldsymbol{\theta}}$ and our previous bound on the MSE of linear regression. □

References

[1] Philippe Rigollet and Jan-Christian Hutter, *High Dimensional Statistics*, Lecture Notes, 2019. `http://www-math.mit.edu/~rigollet/PDFs/RigNotes17.pdf`