University of Waterloo **Scribed By: Gustavo Sutter**
STAT 946 - Deep Learning Theory **Lecture #10**
Instructor: Mufan Li **October 8, 2025**

High Dimensional Linear Regression

Recap: Setup and Underparametrized Regime

From Bach 2024 we have the following model

$$y_i = x_i^\top \theta_* + \epsilon_i \tag{1}$$

where $x_i \sim \mathcal{N}(0,1)$ and $\epsilon_i \sim \mathcal{N}(0,\sigma^2)$.

In vector form Equation 1 can be rewritten

$$\underbrace{\boldsymbol{y}}_{n\times 1} = \underbrace{X}_{n\times d}\underbrace{\boldsymbol{\theta}_*}_{d\times 1} + \underbrace{\boldsymbol{\epsilon}}_{n\times 1} \tag{2}$$

We are doing regression using random projections, that is

$$\begin{aligned} \underbrace{\hat{\boldsymbol{y}}}_{n\times 1} &= \underbrace{X}_{n\times d}\underbrace{S}_{d\times m}\underbrace{\hat{\boldsymbol{\eta}}}_{m\times 1}, \qquad S_{ij} \sim \mathcal{N}(0,1) \\ \hat{\boldsymbol{\theta}} &= S\hat{\boldsymbol{\eta}} \\ \hat{\boldsymbol{\eta}} &= \lim_{\lambda\to 0} \arg\min_{\boldsymbol{\eta}} \{\|\boldsymbol{y} - XS\boldsymbol{\eta}\|^2 + \lambda\|\boldsymbol{\eta}\|^2\} \end{aligned} \tag{3}$$

The risk $R(\hat{\boldsymbol{\theta}}) = \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*\|$ can be decomposed into bias and variance components by taking the expectation over $\boldsymbol{\epsilon}$ as follows:

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\epsilon}}[R(\hat{\boldsymbol{\theta}})] &= \mathbb{E}_{\boldsymbol{\epsilon}}[\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*\|^2] \\ &= \mathbb{E}_{\boldsymbol{\epsilon}}[\|\underbrace{S(S^\top XX^\top S + n\lambda I)^{-1}S^\top X^\top}_{M}\boldsymbol{y} - \boldsymbol{\theta}_*\|^2] \\ &= \mathbb{E}_{\boldsymbol{\epsilon}}[\|M\boldsymbol{y} - \boldsymbol{\theta}_*\|^2] \\ &= \mathbb{E}_{\boldsymbol{\epsilon}}[\|M(X\boldsymbol{\theta}_* + \boldsymbol{\epsilon}) - \boldsymbol{\theta}_*\|^2] \\ &= \mathbb{E}_{\boldsymbol{\epsilon}}[\|(MX - I)\boldsymbol{\theta}_* + M\boldsymbol{\epsilon}\|^2] \\ &= \|(MX - I)\boldsymbol{\theta}_*\|^2 + \mathbb{E}_{\boldsymbol{\epsilon}}[\|M\boldsymbol{\epsilon}\|^2] \\ &= R^{(\text{bias})}(\hat{\boldsymbol{\theta}}) + \mathbb{E}_{\boldsymbol{\epsilon}}[R^{(\text{var})}(\hat{\boldsymbol{\theta}})] \end{aligned} \tag{4}$$

By taking the limit $d, n, m \to \infty$ such that $\frac{d}{n} \to \gamma$ and $\frac{m}{n} \to \delta$ we showed that in the under-parametrized regime ($\delta < 1, \gamma < 1$) we get

$$\mathbb{E}_{\boldsymbol{\epsilon}}[R^{(\mathrm{var})}(\hat{\boldsymbol{\theta}})] \sim \frac{\sigma^2 \delta}{1-\delta} \tag{5}$$

In addition one can show that

$$R^{(\mathrm{bias})}(\hat{\boldsymbol{\theta}}) \sim \frac{\gamma - \delta}{1-\delta} \frac{\|\boldsymbol{\theta}_*\|^2}{\gamma} \tag{6}$$

On this lecture we will see the derivation of the limit in the overparametrized case.

Overparametrized Regime

Recall that $\hat{\Sigma} = \frac{1}{n} X X^\top$ and $\mathrm{Tr}(\hat{\Sigma}(\hat{\Sigma} + \lambda I)^{-1}) \sim \mathrm{Tr}(\Sigma(\Sigma + \kappa(\lambda) I)^{-1})$, where $\kappa(\lambda) - \frac{1}{\varphi(-\lambda)}$. When taking $\lambda \to 0$ we get

$$\kappa(\lambda) = \begin{cases} 0, \gamma \leq 1 \quad \text{(underparametrized)} \\ \gamma - 1, \gamma > 1 \quad \text{(overparametrized)} \end{cases} \tag{7}$$

This means that in the overparametrized regime, even when we have no regularization ($\lambda \to 0$) we have asymptotic regularization ($\kappa(\lambda) = \gamma - 1 > 0$).

Now let us derive the limit of the bias component of the risk in the overparametrized regime. Taking the expectation of the term defined in Equation 4 we get

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\epsilon}}[\|M\boldsymbol{\epsilon}\|^2] &= \sigma^2 \,\mathrm{Tr}(M^\top M) \\ &= \frac{\sigma^2}{n} \mathrm{Tr}[S^\top S (S^\top \hat{\Sigma} S + \lambda I)^{-1} S \hat{\Sigma} S^\top (S^\top \hat{\Sigma} S + \lambda I)^{-1}] \end{aligned} \tag{8}$$

The trace expression surely looks unwieldy, but we can make use of the following result:

Result (Proposition 2, Bach 2024).

$$\begin{aligned} \mathrm{Tr}\big[A Z^\top (Z\Sigma Z^\top - nzI)^{-1} Z B Z^\top (Z \Sigma Z^\top - nzI)^{-1} Z\big] &\sim \mathrm{Tr}\big[A(\Sigma + \tfrac{1}{\varphi(z)} I)^{-1} B (\Sigma + \tfrac{1}{\varphi(z)} I)^{-1}\big] \\ &+ \frac{1}{\varphi(z)^2} \mathrm{Tr}\big[A(\Sigma + \tfrac{1}{\varphi(z)} I)^{-2}\big] \,\mathrm{Tr}\big[B(\Sigma + \tfrac{1}{\varphi(z)} I)^{-2}\big] \cdot \frac{1}{n - \mathrm{df}_2(1/\varphi(z))} \end{aligned} \tag{9}$$

Using this result (taking $Z \to S^\top, \Sigma \to \hat{\Sigma}, n \to m, z \to -\lambda, A \to I, B \to \hat{\Sigma}, \kappa(\lambda) \to \tilde{\kappa}(\lambda), \mathrm{df}_2 \to \tilde{\mathrm{df}}_2$ wrt $\hat{\Sigma}$), we have the following asymptotic equivalence:

$$\mathbb{E}_{\boldsymbol{\epsilon}}[R^{(\mathrm{var})}(\hat{\boldsymbol{\theta}})] \sim \frac{\sigma^2}{n} \operatorname{Tr}[\hat{\Sigma}(\hat{\Sigma} + \tilde{\kappa}(\lambda) I)^{-2}] + \frac{\sigma^2 \tilde{\kappa}(\lambda)^2}{n(m - \tilde{\mathrm{df}}_2(\tilde{\kappa}(\lambda))} \operatorname{Tr}[(\hat{\Sigma} + \tilde{\kappa}(\lambda) I)^{-2}] \operatorname{Tr}[\hat{\Sigma}(\hat{\Sigma} + \tilde{\kappa}(\lambda) I)^{-2}] \tag{10}$$

To simplify this formula we will use the following results:

- $\lambda \to 0 \quad \tilde{\kappa}(\lambda) = \frac{1}{\varphi(-\lambda)} \to 0$
- $\tilde{\mathrm{df}}_2(\tilde{\kappa}(\lambda)) = \operatorname{Tr}[\hat{\Sigma}^2 (\hat{\Sigma}^+)^2] = n$
- Using the push-through identity[1] we have $\hat{\Sigma}(\hat{\Sigma}+\tilde{\kappa} I)^{-2} = nX^\top (XX^\top + n\tilde{\kappa} I)^{-2} X \to nX^\top (XX^\top)^{-2} X$
- Using the Woodbury matrix identity [2] we have

$$\begin{aligned}
\tilde{\kappa}(\lambda)^2 (\hat{\Sigma} + \tilde{\kappa}(\lambda) I)^{-2} &= (I - X^\top (XX^\top + n\tilde{\kappa}(\lambda) I)^{-1}) X)^2 \\
&\to (I - \underbrace{X^\top (XX^\top)^{-1} X}_{\text{projection matrix } P_X})^2 \\
&= (I - P_X)^2 \\
&= I - P_X
\end{aligned} \tag{11}$$

Putting it all together we have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\epsilon}}[R^{(\mathrm{var})}(\hat{\boldsymbol{\theta}})] &\sim \sigma^2 \operatorname{Tr}[X^\top (XX^\top)^{-2} X] + \frac{\sigma^2}{m-n} \operatorname{Tr}[I - P_X] \operatorname{Tr}[(XX^\top)^{-1}] \\
&= \sigma^2 \operatorname{Tr}[(XX^\top)^{-1}] + \frac{\sigma^2}{m-n} \operatorname{Tr}[I - P_X] \operatorname{Tr}[(XX^\top)^{-1}]
\end{aligned} \tag{12}$$

Note that $\operatorname{Tr}[(XX^\top - nzI)^{-1}] = \hat{\varphi}(z) \xrightarrow{n\to\infty} \varphi(z) \xrightarrow{z \to 0} \frac{1}{\kappa(0)} = \frac{1}{\gamma - 1}$. Also, $\operatorname{Tr}[I - P_X] = d - n$, because $\operatorname{rank}(P_X) = n$. Thus

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\epsilon}}[R^{(\mathrm{var})}(\hat{\boldsymbol{\theta}})] &\sim \sigma^2 \operatorname{Tr}[(XX^\top)^{-1}] + \frac{\sigma^2}{m-n} \operatorname{Tr}[I - P_X] \operatorname{Tr}[(XX^\top)^{-1}] \\
&\sim \sigma^2 \frac{1}{\gamma - 1} + \frac{\sigma^2}{m-n} (d-n) \frac{1}{\gamma - 1} \\
&= \sigma^2 \frac{1}{\gamma - 1} \left(1 + \frac{d-n}{m-n} \right) \\
&= \sigma^2 \left(\frac{1}{\gamma - 1} + \frac{1}{\delta - 1} \right)
\end{aligned} \tag{13}$$

[1] Push-through identity: $X(X^\top X + \kappa I)^{-1} = (XX^\top + \kappa I)^{-1} X$

[2] Woodbury matrix identity: $\kappa(X^\top X + \kappa I)^{-1} = I - X^\top (XX^\top + \kappa I)^{-1} X$

Therefore we can see that $\mathbb{E}_{\boldsymbol{\epsilon}}[R^{(\text{var})}(\hat{\boldsymbol{\theta}})]$ decreases as δ or γ increases.

The result for the bias component of the risk is the following (we will not prove it):

$$R^{(\text{bias})} \sim \left(1 - \frac{1}{\gamma}\right) \frac{\delta}{1-\delta} \|\boldsymbol{\theta}_*\|^2 \tag{14}$$

Thus $R^{(\text{bias})}$ increases with γ and decreases with δ.

Double Descent

Combining the results from the previous lecture (underparametrized regime) and today's results (overparametrized regime), we illustrate the double descent phenomenon in Figure 1.

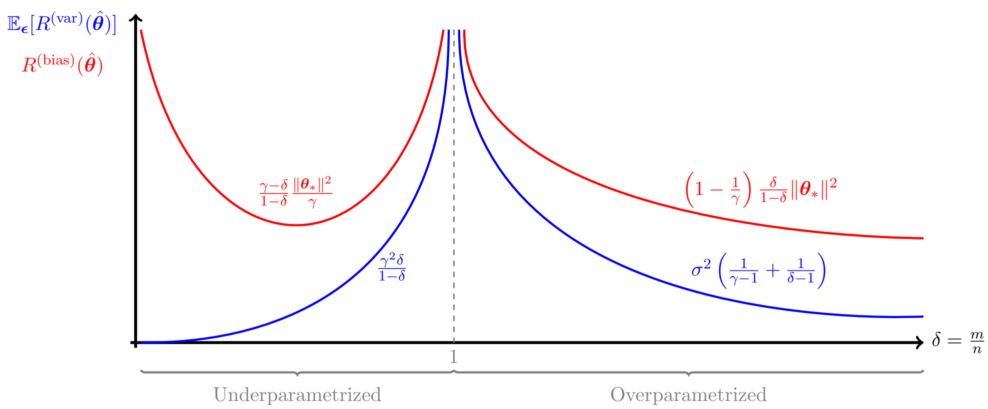

Figure 1: Illustration of the double descent phenomenon.

References

Bach, Francis (2024). “High-Dimensional Analysis of Double Descent for Linear Regression with Random Projections”. In: *SIAM Journal on Mathematics of Data Science* 6.1, pp. 26–50.