Topics in Deep Learning Theory (Spring 2025)

Lecture 2: Linear Method for Regression

Instructor: Lei Wu Date: April 20, 2025

Abstract

In Lecture 1, we introduced the general framework for analyzing generalization via uniform concentration. While this framework is broadly applicable, working with analytically tractable models is often valuable, as they provide explicit derivations that offer deeper insights beyond the general theory. In this lecture, we focus on the linear regression problem, following [Bach, 2024, Section 3].

We shall use big-O notations like $O(\cdot), o(\cdot), \Omega(\cdot)$ and $\Theta(\cdot)$ to hide constants and $\tilde{O}(\cdot)$ to further hide logarithmic factors. We also use $a \lesssim b$ to mean $a = O(b)$ and $a \gtrsim b$ is defined analogously. We use $a \asymp b$ if $a = \Theta(b)$. We also use C to represent an abstract positive constant, whose value may change from line to line.

1 Motivation

In this lecture, we focus on the model given below

$$f(x;\theta) = \sum_{j=1}^{m} \phi_j(x)\theta_j = \phi(x)^\top \theta.,$$

where $x \in \mathcal{X}$ and $\{\phi_j\}_{j=1}^m$ are m basis/feature functions and $\phi(x) = (\phi_1(x), \cdots, \phi_m(x))^\top$. Obviously, this model is linear in the parameter θ, yet it can represent nonlinear functions. When $\phi_j(x) = x_j$ with $m = d$, it reduces to standard linear regression.

Remark 1.1. *We can also consider a model that is linear in x but nonlinear in its parameters. An example is the so-called diagonal linear network*

$$f(x;\alpha,\beta) = \sum_{j=1}^{m} x_j \alpha_j \beta_j.$$

where $\alpha = (\alpha_1, \cdots, \alpha_d) \in \mathbb{R}^d$ and $\beta = (\beta_1, \cdots, \beta_d) \in \mathbb{R}^d$ are parameter vectors. Being linear in x versus being linear in θ is quite different! This stylized model has been widely used in deep learning theory to study how SGD facilitates feature learning and how this contributes to improved sample complexity. For further discussion, we refer to [Woodworth et al., 2020] and its follow-up works, including [Pesme et al., 2021, Andriushchenko et al., 2023, Wu and Su, 2023].

While linear models differ substantially from popular deep learning architectures like MLPs, CNNs, and Transformers, they still play a pivotal role in understanding modern machine learning. Here are three compelling reasons why:

[0]Special thanks to Yuhao Liu and Zilin Wang for scribing the notes.

- **Analytical clarity and tractability:** Linear models are much simpler than their deep counterparts, making them invaluable for deriving clear theoretical insights and foundational results.
- **Sharper theoretical bounds:** The analytical tractability allows us to prove tighter theoretical guarantees, such as sample complexity bounds, that are usually out of reach for more complex models.
- **Beyond traditional statistical learning:** The analytical tractability of linear models makes them an ideal testbed for studying fundamental phenomena in modern machine learning — such as double descent, transfer learning, benign overfitting, and out-of-distribution generalization. These insights extend beyond the in-distribution generalization that traditional statistical learning theory primarily focuses on.

2 Setup

Let $S = \{(x_i, y_i)\}_{i=1}^n$ be our training data and assume that the labels are generated by

$$y_i = f^*(x_i) + \xi_i,$$

where the noise terms $\{\xi_i\}$ are independent random variables with zero mean expectation $\mathbb{E}[\xi_i] = 0$ and finite variance $\mathbb{E}[\xi_i^2] = \sigma^2$. Recall that our model takes the following linear form:

$$f(x; \theta) = \phi(x)^\top \theta.$$

We may omit the dependence on θ and write the model simply as f whenever there is no ambiguity. In this lecture, we make the following assumption on the target function

$$f^*(x) = \phi(x)^\top \theta_*.$$

This is often referred to as the *well-specified* setting, as f^* lies within the chosen model class.

Remark 2.1. *In most practical scenarios, the target function typically falls outside the model class, making the model mis-specified. Let $\mathcal{F}$ denote the class of functions our model can represent. A common approach to handling the mis-specified case is to decompose the target function as*

$$f^* = \bar{f}^* + \epsilon^*,$$

where $\bar{f}^* \in \mathcal{F}$ is the best approximation within the model class, and $\epsilon^*(x) := f^*(x) - \bar{f}^*(x)$ represents the residual and the term ϵ^* quantifies the approximation error. We will discuss this issue in future lectures.

Depending on how the inputs are sampled, there are two classical scenarios:

- *Fixed design.* In this setting, the inputs $\{x_1, \cdots, x_n\}$ are predetermined and non-stochastic. This setting is predominantly employed in computational mathematics (such as grid design for numerical solutions of partial differential equations) and classical statistical applications, such as experimental design.
- *Random design.* In this setting, $\{x_i\}_{i=1}^n$ are i.i.d. samples drawn from an underlying distribution ρ. This is the common setup in statistical learning.

Excess risk. Let $(x, y) \sim \mathcal{D}$ and $\mathcal{F}$ be the hypothesis space of our model can represent. Given a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}_{\geq 0}$, let

$$\mathcal{R}(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f(x), y)] \quad \text{and} \quad \mathcal{R}^* = \min_{f\in\mathcal{F}} \mathcal{R}(f).$$

Then we can measure the model's performance using the *excess risk*:

$$\mathcal{E}(f) = \mathcal{R}(f) - \mathcal{R}^*. \tag{1}$$

Note that the above definition is quite general without making any assumption on the joint distribution $\mathcal{D}$ and the loss function.

In our setting, where the loss function is squared and the data distribution follows $x \sim \pi \in \mathcal{P}(\mathcal{X})$ with $y = f^*(x) + \xi$, where ξ is independent of x and satisfies $\mathbb{E}[\xi] = 0, \mathbb{E}[\xi^2] = \sigma^2$, the risk function can be expressed as

$$\begin{aligned}\mathcal{R}(f) &= \mathbb{E}_{x\sim\pi}(f(x) - f^*(x) - \xi)^2 \\ &= \mathbb{E}_{x\sim\pi}(f(x) - f^*(x))^2 + \mathbb{E}[\xi^2] \\ &= \mathbb{E}_{x\sim\pi}(f(x) - f^*(x))^2 + \sigma^2.\end{aligned}$$

Since $f^* \in \mathcal{F}$, it follows that the optimal risk is

$$\mathcal{R}^* = \sigma^2.$$

Thus, the *excess risk* becomes the more intuitive squared fitting error:

$$\mathcal{E}_\pi(f) = \mathbb{E}_{x\sim\pi}(f(x) - f^*(x))^2 = \|f - f^*\|_\pi^2, \tag{2}$$

where $\|\cdot\|_\pi$ denotes the $L^2(\pi)$ norm. Note that in general, π is not restricted to ρ, the distribution from which our training data is drawn. However, we will focus on two important special cases:

- **Training error:** When π is the empirical distribution of the training data, given by $\pi = \hat{\rho} := \frac{1}{n}\sum_{i=1}^n \delta(\cdot - x_i)$, this quantity corresponds to the training error, which reflects the model's ability to denoise the training data.
- **Test error:** When $\pi = \rho$, the distribution of the underlying data-generating process, this corresponds to the test error, which measures the model's generalization performance.

In our well-specified setting for linear model, the *excess risk* under a given distribution π can be expressed in the following form

$$\begin{aligned}\mathcal{E}_\pi(f) &= \mathbb{E}_{x\sim\pi}(f(x) - f^*(x))^2 \\ &= \mathbb{E}_{x\sim\pi}(\phi(x)^\top(\theta - \theta^*))^2 \\ &= (\theta - \theta^*)^\top\Sigma_\pi(\theta - \theta^*) \\ &= \|\theta - \theta^*\|_{\Sigma_\pi}^2,\end{aligned} \tag{3}$$

where Σ_π denotes the covariance matrix of the feature map under the distribution π:

$$\Sigma_\pi = \mathbb{E}_\pi[\phi(x)\phi(x)^\top],$$

and $\|u\|_A^2 := u^\top A u$ for a psd matrix A.

To evaluate training and test error, we first define the empirical and population feature covariance matrices as follows:

$$\hat{\Sigma} = \mathbb{E}_{\hat{\rho}}[\phi(x_i)\phi(x_i)^\top] = \frac{1}{n}\sum_{i=1}^{n}\phi(x_i)\phi(x_i)^\top, \quad \Sigma = \mathbb{E}_{x\sim\rho}[\phi(x)\phi(x)^\top]. \tag{4}$$

In the random design setting, it is evident that $\hat{\Sigma}$ converges to Σ in a suitable sense as $n \to \infty$.

Bias-variance decomposition. Our analysis will frequently rely on the celebrated bias-variance decomposition. Let $\hat{f}$ denote the model trained on the dataset S. The bias-variance decomposition for the learned model is given by

$$\hat{f} = \underbrace{\mathbb{E}[\hat{f}]}_{\text{bias}} + \underbrace{\hat{f} - \mathbb{E}[\hat{f}]}_{\text{variance}},$$

where the expectation is taken over the randomness in S. For instance, in the fixed-design setting (or analyzing the training error), the expectation is taken with respect to the noise $\{\xi_i\}_{i=1}^n$.

Applying this decomposition to the excess risk (2), we obtain

$$\begin{aligned}\mathbb{E}[\mathcal{E}(\hat{f})] &= \mathbb{E}\|\hat{f} - f^*\|_\pi^2 = \mathbb{E}\|\mathbb{E}[\hat{f}] - f^* + \hat{f} - \mathbb{E}[\hat{f}]\|_\pi^2 \\ &= \underbrace{\|\mathbb{E}[\hat{f}] - f^*\|_\pi^2}_{\text{bias}} + \underbrace{\mathbb{E}\|\hat{f} - \mathbb{E}[\hat{f}]\|_\pi^2}_{\text{variance}}.\end{aligned} \tag{5}$$

3 OLS estimator

In this section, we analyze the performance of the ordinary least squares (OLS) estimator, which is obtained by minimizing the empirical risk:

$$\hat{\theta} = \operatorname*{argmin}_\theta \frac{1}{n}\sum_{i=1}^{n}(\phi(x_i)^\top\theta - y_i)^2 = \operatorname*{argmin}_\theta \frac{1}{n}\|\Phi\theta - y\|^2$$

with $\Phi = (\phi(x_1), \cdots, \phi(x_n))^\top \in \mathbb{R}^{n\times m}$. The minimizer is given by

$$\hat{\theta} = \left(\frac{1}{n}\Phi^\top\Phi\right)^{-1}\frac{1}{n}\Phi^\top y = \left(\frac{1}{n}\Phi^\top\Phi\right)^{-1}\frac{1}{n}\Phi^\top(\Phi\theta_* + \xi) = \theta_* + \widehat{\Sigma}^{-1}\frac{1}{n}\Phi^\top\xi, \tag{6}$$

where we use the fact that $\frac{1}{n}\Phi^\top\Phi \in \mathbb{R}^{m\times m} = \widehat{\Sigma}$ is the empirical covariance matrix (see Eq. (4)), assumed to be invertible.

Eq. (6) provides a closed-form expression for the OLS estimator. Notably, since $\mathbb{E}\,\hat{\theta} = \theta_*$, the OLS estimator is unbiased. To assess its performance, we first analyze the training error. The following result characterizes the estimator $\hat{f}_{\text{ols}} = f(\cdot;\hat{\theta})$.

Proposition 3.1. *The expected training error of OLS estimator is given by*

$$\widehat{\mathcal{E}}_n := \mathbb{E}_\xi[\mathcal{E}_{\hat{\rho}}(\hat{f}_{\text{ols}})] = \frac{\sigma^2 m}{n}. \tag{7}$$

Proof. According to (5) and (6), we have

$$\begin{aligned}
\mathbb{E}_\xi[\mathcal{E}_{\hat{\rho}}(\hat{f})] &= \mathbb{E}_\xi \|\hat{\theta} - \mathbb{E}_\xi[\hat{\theta}]\|_{\widehat{\Sigma}}^2 + \|\mathbb{E}_\xi[\hat{\theta}] - \theta_*\|_{\widehat{\Sigma}}^2 \\
&= \frac{1}{n^2}\mathbb{E}_\xi[\xi^\top \Phi \widehat{\Sigma}^{-1}\widehat{\Sigma}\widehat{\Sigma}^{-1}\Phi^\top \xi] \\
&= \frac{1}{n^2}\operatorname{Tr}\left(\widehat{\Sigma}^{-1}\Phi^\top \mathbb{E}_\xi[\xi\xi^\top]\Phi\right) \\
&= \frac{\sigma^2}{n}\operatorname{Tr}\left(\widehat{\Sigma}^{-1}\widehat{\Sigma}\right) \\
&= \frac{\sigma^2 m}{n}.
\end{aligned}$$

□

Remark 3.2. *To achieve a small training error, we require at least $\Omega(m)$ samples. This can be costly when the feature space is high-dimensional.*

Next, we analyze the generalization error. We can express the risk of the as follows.

Proposition 3.3. *The excess risk of the OLS estimator under ρ is given by*

$$\mathcal{E}_n := \mathbb{E}_\xi[\mathcal{E}_\rho(\hat{f}_{\text{ols}})] = \frac{\sigma^2}{n}\operatorname{Tr}(\Sigma\widehat{\Sigma}^{-1}). \tag{8}$$

Proof. Recall the expression of the OLS estimator in (6). From (5) we have

$$\begin{aligned}
\mathbb{E}_\xi[\mathcal{E}_\rho(\hat{f})] &= \mathbb{E}_\xi \|\hat{\theta} - \mathbb{E}_\xi[\hat{\theta}]\|_{\Sigma}^2 + \|\mathbb{E}_\xi[\hat{\theta}] - \theta_*\|_{\Sigma}^2 \\
&= \frac{1}{n^2}\mathbb{E}_\xi[\xi^\top \Phi \widehat{\Sigma}^{-1}\Sigma\widehat{\Sigma}^{-1}\Phi^\top \xi] \\
&= \frac{1}{n^2}\operatorname{Tr}(\Sigma\widehat{\Sigma}^{-1}\Phi^\top \mathbb{E}_\xi[\xi\xi^\top]\Phi\widehat{\Sigma}^{-1}) \\
&= \frac{\sigma^2}{n}\operatorname{Tr}(\Sigma\widehat{\Sigma}^{-1}).
\end{aligned}$$

□

Intuitively, when $n \gg 1$, we expect $\widehat{\Sigma} \approx \Sigma$. As a result, it follows that $\widehat{\Sigma}^{-1}\Sigma \approx I_m$, leading to the approximation $\mathcal{E}_n \approx \frac{\sigma^2 m}{n}$.

To rigorously estimate the generalization error, we introduce a method based on concentration inequalities for matrices. Our analysis is motivated by considering the trace of the inverse term $\operatorname{Tr}(\Sigma^{-1}\widehat{\Sigma})$ and rewriting it as

$$\begin{aligned}
\operatorname{Tr}(\Sigma^{-1}\widehat{\Sigma}) &= \operatorname{Tr}(\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2}) \\
&= \frac{1}{n}\operatorname{Tr}\left(\Sigma^{-1/2}\left(\sum_{i=1}^n \phi(x_i)\phi(x_i)^\top\right)\Sigma^{-1/2}\right) \\
&= \operatorname{Tr}\left(\frac{1}{n}\sum_{i=1}^n z_i z_i^\top\right)
\end{aligned}$$

where we define $z_i = \Sigma^{-1/2}\phi(x_i)$. The following lemma provides a bound on the rate at which $\frac{1}{n}\sum_{i=1}^n z_i z_i^\top$ concentrates around I_m.

Lemma 3.4 (Matrix Bernstein bound)**.** *Given n independent symmetric matrices $M_i \in \mathbb{R}^{d\times d}$, such that for all $i \in \{1, \cdots, n\}$, $\mathbb{E}[M_i] = 0$, $\|M_i\|_2 \le b$ almost surely and $\|\frac{1}{n}\sum_{i=1}^n \mathbb{E}[M_i^2]\|_2 \le \sigma^2$. Then for all $t \ge 0$, we have*

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n M_i\right\|_2 \ge t\right) \le d \cdot \exp\left(-\frac{nt^2/2}{\sigma^2 + bt/3}\right).$$

We are now ready to prove Proposition 3.5, which provides a high-probability bound on the excess risk.

Proposition 3.5. *Let $z = \Sigma^{-1/2}\phi(x)$ satisfy $\|z\|_2 \le Cm$ almost surely. Then, for any $\delta \in (0,1)$, if the number of samples satisfies $n \ge Cm\log(m/\delta)$, it follows that with probability at least $1-\delta$ over the sampling of $x_1, \ldots, x_n$, the test error is bounded as*

$$\mathcal{E}_n \le \frac{2\sigma^2 m}{n}.$$

Proof. Let $M_i = z_i z_i^\top - I_m$. We can check that it holds for each $i \in [n]$ that:

$$\mathbb{E}[M_i] = 0,$$

and

$$\|M_i\|_2 \le \max\left\{\|z_i\|^2, 1\right\} \le Cm,$$

as well as

$$\|\,\mathbb{E}[M_i^2]\|_2 = \|\,\mathbb{E}[\|z_i\|^2 z_i z_i^\top] - 2z_i z_i^\top + I_d\|_2 \le Cm.$$

Let $\Delta_n = \frac{1}{n}\sum_{i=1}^n z_i z_i^\top - I_m$. By applying Lemma 3.4 with $b = \sigma^2 = Cm$, we obtain

$$\mathbb{P}\left\{\|\Delta_n\|_2 \ge t\right\} \le m \cdot \exp\left(-\frac{nt^2/2}{Cm(1+t/3)}\right).$$

For any $\delta \in (0,1)$, setting $t = 1/2$ and $n \ge \frac{C}{d}\log(d/\delta)$, we conclude that with probability at least $1-\delta$,

$$\left\|\frac{1}{n}\sum_{i=1}^n z_i z_i^\top - I_m\right\|_2 \le \frac{1}{2}.$$

This implies the following spectral bound:

$$\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2} \succeq \frac{1}{2}I_m.$$

Equivalently, taking inverses,

$$\Sigma\widehat{\Sigma}^{-1} \preceq 2I_m.$$

Combining this with equation (8) completes the proof. □

We have discussed a method based on matrix concentration inequalities to obtain an non-asymptotic bound of the test error. However, there exist alternative approaches that can achieve similar results. Below, we briefly introduce two such methods:

- **Gaussian design.** Suppose that the feature mapping $\phi(x)$ follows a Gaussian distribution with mean zero and covariance matrix Σ. In this case, $z = \Sigma^{-1/2}\phi(x)$ follows a standard Gaussian distribution, leading to the formation of the matrix $Z = (z_1, \cdots, z_n)^\top \in \mathbb{R}^{n\times m}$. Under this setting, the inverse of the Gram matrix, $(Z^\top Z)^{-1}$, follows an *inverse Wishart distribution*. It is known that $\mathbb{E}[(Z^\top Z)^{-1}] = \frac{1}{n-m-1} I_m$. Furthermore, the expectation of the trace term satisfies $\mathbb{E}[\mathrm{Tr}(\Sigma\widehat{\Sigma}^{-1})] = n\,\mathbb{E}[\mathrm{Tr}(Z^\top Z)^{-1}]$. This allows us to derive an explicit bound on the expected excess risk.

- **Random matrix theory.**

4 Ridge Regression

From the discussion in Section 3, the required sample complexity is $n = \Omega(m)$. In high-dimensional settings, where $m \gg 1$, this dependency leads to an impractically large number of required samples. This raises a natural question: Can we improve our results to achieve lower training and test errors, ideally making them independent of m? Two principal strategies have emerged in machine learning to tackle this challenge:

- **Dimension reduction.** This approach assumes that the feature vector $\Phi(x)$ resides within a low-dimensional manifold. Dimension reduction techniques aim to replace the original feature vector $\Phi(x)$ with a lower-dimensional representation obtained through classical methods, such as Principal Component Analysis (PCA) or random projections.

- **Regularization.** This approach leverages structural assumptions on the parameter θ_*, such as sparsity or bounded norm constraints. To enforce sparsity, an ℓ_1 penalty (known as the Least Absolute Shrinkage and Selection Operator, or LASSO) is applied to θ. Alternatively, to impose norm constraints, an ℓ_2 penalty is utilized, yielding **ridge regression**.

Remark 4.1. *In the limiting scenario where the number of features $m \to \infty$, regularization methods exhibit interesting connections to other learning models: the LASSO regularization corresponds to a two-layer neural network, whereas ridge regression corresponds to kernel-based methods.*

Here, our primary focus is on ridge regression.

4.1 Ridge Regression

We begin by making a bounded-norm assumption on the target function.

Assumption 4.2. $\|\theta_*\|_2 \le 1$.

The ridge regression (Ridge) estimator is given by

$$\hat{\theta}_\lambda = \operatorname*{argmin}_\theta \frac{1}{n}\|\Phi\theta - y\|_2^2 + \lambda\|\theta\|_2^2,$$

where $\lambda > 0$ is the regularization parameter. This estimator admits the following closed-form solution:

$$\hat{\theta}_\lambda = \frac{1}{n}(\widehat{\Sigma} + \lambda I)^{-1}\Phi^\top y.$$

Due to the regularization term λI, invertibility of $\Phi^\top \Phi$ is no longer required, distinguishing Ridge from OLS.

Denote $\hat{f}_\lambda = f(x; \hat{\theta}_\lambda)$. We then have the following proposition regarding the training error of $\hat{f}_\lambda$ by setting $\pi = \hat{\rho}$:

Proposition 4.3. *The expected excess risk of the prediction function $\hat{f}_\lambda$ is given by*

$$\mathbb{E}_\xi[\mathcal{E}_{\hat{\rho}}(\hat{f}_\lambda)] = \lambda^2 \theta_*^\top (\widehat{\Sigma} + \lambda I)^{-2} \widehat{\Sigma} \theta_* + \frac{\sigma^2}{n} \operatorname{Tr}\left[\widehat{\Sigma}^2 (\widehat{\Sigma} + \lambda I)^{-2}\right],$$

where the first term represents the bias, and the second term corresponds to the variance.

Proof. We use the risk decomposition in (5) into a bias term $B(\lambda)$ and a variance term $V(\lambda)$. Since we have

$$\begin{aligned}
\mathbb{E}_\xi\left[\hat{\theta}_\lambda\right] &= \mathbb{E}_\xi\left[\frac{1}{n}(\widehat{\Sigma} + \lambda I)^{-1} \Phi^\top (\Phi \theta_* + \xi)\right] \\
&= \frac{1}{n}(\widehat{\Sigma} + \lambda I)^{-1} \Phi^\top \Phi \theta_* = (\widehat{\Sigma} + \lambda I)^{-1} \widehat{\Sigma} \theta_* \\
&= \theta_* - \lambda (\widehat{\Sigma} + \lambda I)^{-1} \theta_*,
\end{aligned}$$

it follows that

$$B(\lambda) = \| \mathbb{E}_\xi[\theta] - \theta^* \|_{\widehat{\Sigma}}^2 = \lambda^2 \theta_*^\top (\widehat{\Sigma} + \lambda I)^{-2} \widehat{\Sigma} \theta_*,$$

with the fact that $\widehat{\Sigma}$ and $(\widehat{\Sigma} + \lambda I)^{-1}$ commute. For the variance term $V(\lambda)$, since $\mathbb{E}[\xi \xi^\top] = \sigma^2 I$ and $\operatorname{Tr}(AB) = \operatorname{Tr}(BA)$, we have

$$\begin{aligned}
V(\lambda) &= \mathbb{E}_\xi\left[\left\| \hat{\theta}_\lambda - \mathbb{E}_\xi\left[\hat{\theta}_\lambda\right] \right\|_{\widehat{\Sigma}}^2\right] \\
&= \mathbb{E}_\xi\left[\left\| \frac{1}{n}(\widehat{\Sigma} + \lambda I)^{-1} \Phi^\top \xi \right\|_{\widehat{\Sigma}}^2\right] \\
&= \mathbb{E}_\xi\left[\frac{1}{n^2} \operatorname{Tr}\left(\xi^\top \Phi (\widehat{\Sigma} + \lambda I)^{-1} \widehat{\Sigma} (\widehat{\Sigma} + \lambda I)^{-1} \Phi^\top \xi\right)\right] \\
&= \mathbb{E}_\xi\left[\frac{1}{n^2} \operatorname{Tr}\left(\Phi^\top \xi \xi^\top \Phi (\widehat{\Sigma} + \lambda I)^{-1} \widehat{\Sigma} (\widehat{\Sigma} + \lambda I)^{-1}\right)\right] \\
&= \frac{\sigma^2}{n} \operatorname{Tr}\left(\widehat{\Sigma} (\widehat{\Sigma} + \lambda I)^{-1} \widehat{\Sigma} (\widehat{\Sigma} + \lambda I)^{-1}\right).
\end{aligned}$$

Summing the bias term $B(\lambda)$ and variance term $B(\lambda)$ completes the proof. □

We highlight the following key observations:

- **Bias term.** This term captures the approximation error introduced by the ridge penalty, which restricts the model's expressivity.
- **Variance Term.** In contrast to Proposition 3.1, this term no longer depends on the number of parameters m. Instead, it depends on the *degrees of freedom* (DoF):

$$\operatorname{DoF}(\lambda) \;=\; \operatorname{Tr}\big(\widehat{\Sigma}^2 (\widehat{\Sigma} + \lambda I)^{-2}\big) \;=\; \sum_{j=1}^{m} \frac{\mu_j^2}{(\mu_j + \lambda)^2}, \tag{9}$$

where $\{\mu_j\}_{j=1}^m$ are the eigenvalues of $\widehat{\Sigma}$ in decreasing order. Intuitively, $\mathrm{DoF}(\lambda)$ measures how many eigenvalues exceed the threshold λ, as discussed in Lemma 4.4. When $\lambda = 0$, $\mathrm{DoF}(0) = m$, representing the total degrees of freedom in the unregularized model.

- **Bias-variance trade-off.** Because the bias $B(\lambda)$ and variance $V(\lambda)$ vary oppositely as functions of λ, a natural trade-off emerges between these two terms.

Let $m(\lambda) = \#\{j : \mu_j \geq \lambda\}$ be the number of eigenvalues that is greater than λ. Then, the following lemma characterizes how the DoF depends on spectrum of $\widehat{\Sigma}$:

Lemma 4.4. *When $\mu_j \asymp j^{-\beta}$, then we have $\mathrm{DoF}(\lambda) \asymp m(\lambda) = \lambda^{-1/\beta}$.*

Proof. First, under the eigenvalue decay assumption, we have $m(\lambda) \asymp \lambda^{-1/\beta}$. Moreover, noting

$$\mathrm{DoF}(\lambda) \geq \sum_{\mu_j \geq \lambda} \left(\frac{\mu_j}{\mu_j + \lambda}\right)^2 \geq \frac{1}{4} m(\lambda)$$

and

$$\begin{aligned}
\mathrm{DoF}(\lambda) &= \sum_j \frac{\mu_j^2}{(\mu_j + \lambda)^2} = \sum_{\mu_j \geq \lambda} \frac{\mu_j^2}{(\mu_j + \lambda)^2} + \sum_{\mu_j \leq \lambda} \frac{\mu_j^2}{(\mu_j + \lambda)^2} \\
&\leq \sum_{\mu_j \geq \lambda} 1 + \sum_{\mu_j \leq \lambda} \frac{\mu_j^2}{(\mu_j + \lambda)^2} \\
&\leq m(\lambda) + \frac{1}{\lambda^2} \int_{\lambda^{-1/\beta}}^{\infty} x^{-2\beta} \mathrm{d}x \leq C_1 \lambda^{-\frac{1}{\beta}} \leq C \lambda^{-1/\beta},
\end{aligned}$$

we can conclude that $\mathrm{DoF}(\lambda) \asymp m(\lambda)$. □

Next, we present an upper bound for $\mathbb{E}_\xi[\mathcal{E}_{\hat{\rho}}(\hat{f}_\lambda)]$.

Proposition 4.5. *The expected training error satisfies the following upper bound:*

$$\mathbb{E}_\xi[\mathcal{E}_{\hat{\rho}}(\hat{f}_\lambda)] \leq \frac{\lambda}{2} + \frac{\sigma^2 \operatorname{Tr}(\widehat{\Sigma})}{2\lambda n}.$$

Proof. Note that the eigenvalues of the matrix $(\widehat{\Sigma} + \lambda I)^{-2} \lambda \widehat{\Sigma}$ are bounded above by $1/2$, which is a direct consequence of the inequality $(\mu + \lambda)^2 \geq 2\mu\lambda$. Thus, the bias term can be bounded as follows:

$$\begin{aligned}
B(\lambda) &= \lambda^2 \theta_*^\top (\widehat{\Sigma} + \lambda I)^{-2} \widehat{\Sigma} \theta_* \\
&= \lambda \theta_*^T (\widehat{\Sigma} + \lambda I)^{-2} \lambda \widehat{\Sigma} \theta_* \\
&\leq \frac{\lambda}{2} \|\theta_*\|^2 \leq \frac{\lambda}{2}
\end{aligned}$$

where the last inequality follows directly from Assumption 4.2. Similarly, the variance term $V(\lambda)$ can be bounded by:

$$\begin{aligned}
V(\lambda) &= \frac{\sigma^2}{n} \operatorname{Tr}\left[\widehat{\Sigma}^2 (\widehat{\Sigma} + \lambda I)^{-2}\right] \\
&= \frac{\sigma^2}{\lambda n} \operatorname{Tr}\left[\widehat{\Sigma} \lambda \widehat{\Sigma} (\widehat{\Sigma} + \lambda I)^{-2}\right] \\
&\leqslant \frac{\sigma^2 \operatorname{Tr} \widehat{\Sigma}}{2\lambda n}.
\end{aligned}$$

Summing the bias term $B(\lambda)$ and variance term $V(\lambda)$ completes the proof. □

The upper bound derived above takes the form $a\lambda + b/\lambda$ for positive constant $a, b > 0$. This expression attains its minimum at $\lambda = \sqrt{b/a}$, yielding an optimal value of $2\sqrt{ab}$. Taking $a = \frac{1}{2}$ and $b = \frac{\sigma^2 \operatorname{Tr}(\widehat{\Sigma})}{2n}$,we obtain the optimal regularization parameter: $\lambda^* = \frac{\sigma \operatorname{Tr}(\widehat{\Sigma})^{1/2}}{\sqrt{n}}$ and consequently, the optimal upper bound:

$$\mathbb{E}_\xi[\mathcal{E}_{\hat{\rho}}(\hat{f}_\lambda)] \leq \frac{\sigma \operatorname{tr}(\widehat{\Sigma})^{1/2}}{\sqrt{n}}$$

The expected excess risk decays as $n^{-1/2}$, independent of the dimension m, assuming $\operatorname{Tr}(\widehat{\Sigma}) \leq C$. However, the amplification of the mean value inequality in the aforementioned proof is excessively relaxed, resulting in the loss of certain spectral information of $\widehat{\Sigma}$. To derive more precise results, additional assumptions about the eigenvalue distribution are beneficial. Specifically, we introduce the following assumption regarding the polynomial decay of eigenvalues:

Assumption 4.6. *(Polynomial decay) The eigenvalues μ_j of $\widehat{\Sigma}$ exhibit polynomial decay $\mu_j \asymp j^{-\beta}$ with $\beta > 1$.*

Then, combining Lemma 4.4 into Proposition 4.5 gives

$$\mathbb{E}_\xi[\mathcal{E}_{\hat{\rho}}(\hat{f}_\lambda)] \leq C\left(\lambda + \frac{\sigma^2}{n}\lambda^{-\frac{1}{\beta}}\right).$$

By choosing the optimal $\lambda^* = (\sigma^2/n)^{\frac{\beta}{1+\beta}}$, we obtain the following bound:

$$\mathbb{E}_\xi[\mathcal{E}_{\hat{\rho}}(\hat{f}_\lambda)] \leq C\left(\frac{\sigma^2}{n}\right)^{\frac{\beta}{1+\beta}}. \tag{10}$$

We summarize the key insights as follows:

- As $\beta \to 1$, $\mathbb{E}_\xi[\mathcal{E}_{\hat{\rho}}(\hat{f}_\lambda)]$ decays at a rate of $\frac{1}{\sqrt{n}}$, recovering the rate established earlier derived by Proposition 4.5.
- As β increases, the eigenvalues of $\widehat{\Sigma}$ decay more rapidly, resulting in a faster decay of the expected excess risk $\mathbb{E}_\xi[\mathcal{E}_{\hat{\rho}}(\hat{f}_\lambda)]$ with respect to n. When $\beta \to \infty$, we approach the *fast rate*: $1/n$.

We note that by following a similar strategy as in Proposition 3.5, one can derive an upper bound for the test error analogous to (10). We defer this derivation to the analysis of kernel methods, which can be viewed as ridge-regularized linear models in the limit $m \to \infty$.

5 Feature Extraction and Reduced Models

Beyond directly penalizing the model's expressivity, as in ridge regression, another popular approach to improving generalization is through feature extraction and constructing certain reduced models.

In our case, we aim to construct a set of basis functions $\{h_1(x), \cdots, h_k(x)\}$ with $k \ll m$ to approximate the original raw features through linear combinations:

$$h_t(x) = \sum_{j=1}^{m} w_{jt}\phi_j(x), \quad t = 1, \cdots, k$$

where $W = (w_{jt}) \in \mathbb{R}^{m\times k}$ is the projection matrix. This yields a compressed representation

$$f_\alpha(x) = \sum_{t=1}^{k} \alpha_t h_t(x) = \phi(x)^\top W\alpha$$

where $\alpha = (\alpha_1, \cdots, \alpha_k) \in \mathbb{R}^k$. This naturally leads to a two-stage learning framework:

- **Stage I: (Unsupervised) feature extraction.** Learn the combination matrix W using our data, such as the spectral property, which are independent of the labels $\{y_i\}$.
- **Stage II: Supervised learning in the reduced space.** Learn coefficients $\alpha \in \mathbb{R}^k$ via some regression methods.

Remark 5.1. *In practice, we are usually given a labeled dataset $\{(x_i, y_i)\}_{i=1}^n$ along with a large amount of unlabeled data $\{x_i\}_{i=n+1}^N$. In Stage I, we can utilize those enormous unlabeled data to learn a good feature.*

Remark 5.2. *If both α and W are learnable in simultaneously, this framework can be viewed as a simplified two-layer neural network:*

$$f_{W,\alpha}(x) = \phi(x)^\top W\alpha \quad \xrightarrow{\textit{with activation}} \quad \sigma(\phi(x)^\top W)\alpha$$

5.1 Spectral truncation

Assume the empirical covariance matrix $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n \phi(x_i)\phi(x_i)^\top$ admits eigen-decomposition as follows:

$$\widehat{\Sigma} = \sum_{j=1}^{m} \mu_j u_j u_j^\top \quad \text{with} \quad \mu_1 \geq \mu_2 \geq \cdots \geq \mu_m$$

For some estimator $\hat{\theta}$, let $\delta = \hat{\theta} - \theta_*$. The training error is

$$\hat{\mathcal{E}}_n = \|\hat{\theta} - \theta_*\|_{\widehat{\Sigma}}^2 = \sum_{j=1}^{m} \mu_j (\delta_j^\top u_j)^2.$$

Now, consider a reduced model where $\theta = W\alpha$ lies within a k-dimensional subspace. Intuitively, this constraint implies that at most k coordinates of δ can be optimized. Given the decay in the eigenvalues of the covariance matrix, it is natural to construct W using the leading eigenvectors. Thus, we define

$$W = (u_1, \cdots, u_k) \in \mathbb{R}^{m\times k}$$

Let $\Phi = (\phi(x_1), \cdots, \phi(x_n))^\top \in \mathbb{R}^{n\times m}$. The estimator is obtained by solving the following regression problem:

$$\begin{aligned}
\hat{\alpha} &= \operatorname*{argmin}_\alpha \frac{1}{n}\|\Phi W\alpha - y\|^2 \\
&= \left(\frac{1}{n}W^\top\Phi^\top\Phi W\right)^{-1}\frac{1}{n}W^\top\Phi^\top y \\
&= \left(\frac{1}{n}W^\top\Phi^\top\Phi W\right)^{-1}\frac{1}{n}W^\top\Phi^\top(\Phi\theta_* + \xi).
\end{aligned} \tag{11}$$

We then proceed to analyze the error of this estimator. We have the following result.

Proposition 5.3. *Assuming that the eigenvalues of the covariance matrix satisfy $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_m$. the empirical excess risk of the estimator obtained via the spectral truncation approach, as given in (11), is bounded by*

$$\mathcal{E}_n \leq \mu_{k+1} + \frac{\sigma^2 k}{n}. \tag{12}$$

Proof. Recall the bias-variance decomposition we introduced in (5). For the bias term we have

$$\begin{aligned} B(k) &= \| \mathbb{E}(W\hat{\alpha}) - \theta_* \|_{\widehat{\Sigma}}^2 \\ &= \left\| W(W^\top \widehat{\Sigma} W)^{-1} W^\top \widehat{\Sigma} \theta_* - \theta_* \right\|_{\widehat{\Sigma}}^2 . \end{aligned}$$

Let $\theta_* = W\alpha_* + W_\perp \beta_*$ be the decomposition with $\|\alpha_*\|^2 + \|\beta_*\|^2 = \|\theta_*\|^2$, where $W_\perp$ is the orthogonal complement of W. Note that $W^\top \widehat{\Sigma} W_\perp = 0$. Then

$$\begin{aligned} B(k) &= \left\| W(W^\top \widehat{\Sigma} W)^{-1} W^\top \widehat{\Sigma}(W\alpha_* + W_\perp \beta_*) - \theta_* \right\|_{\widehat{\Sigma}}^2 \\ &= \|W\alpha_* - \theta_*\|_{\widehat{\Sigma}}^2 \\ &= \|W_\perp \beta_*\|_{\widehat{\Sigma}}^2 \\ &\leq \|W_\perp^\top \widehat{\Sigma} W_\perp\|_2 \\ &\leq \mu_{k+1}. \end{aligned}$$

For the variance term,

$$\begin{aligned} V(k) &= \mathbb{E}_\xi \left[\left\| W(W^\top \widehat{\Sigma} W)^{-1} \frac{1}{n} W^\top \Phi^\top \xi \right\|_{\widehat{\Sigma}}^2 \right] \\ &= \frac{1}{n^2} \mathbb{E}_\xi \left[\xi^\top \Phi W (W^\top \widehat{\Sigma} W)^{-1} (W^\top \widehat{\Sigma} W)(W^\top \widehat{\Sigma} W)^{-1} W^\top \Phi^\top \xi \right] \\ &= \frac{1}{n^2} \operatorname{Tr}\left((W^\top \widehat{\Sigma} W)^{-1} W^\top \Phi^\top \mathbb{E}[\xi\xi^\top] \Phi W \right) \\ &= \frac{\sigma^2}{n} \operatorname{Tr}(I_k) \\ &= \frac{\sigma^2 k}{n}. \end{aligned}$$

Combining the two terms finishes the proof. □

If we further assume a power-law decay of the eigenvalues, we can determine an optimal dimension for the reduced model and establish the corresponding convergence rate.

Proposition 5.4. *Let $\mu_j \asymp j^{-\beta}$ for some $\beta > 0$. Then the optimal dimension is given by*

$$k_{\text{op}} = \left(\frac{n\beta}{\sigma^2} \right)^{\frac{1}{\beta+1}},$$

and we have

$$\mathcal{E}_n \leq \left(\frac{\sigma^2}{\beta n} \right)^{\frac{\beta}{\beta+1}}$$

The proof is straightforward. Comparing with (10), we see that the reduced model achieves the same rate as ridge regression.

Phase Transitions The error rate exhibits distinct phases depending on eigenvalue decay:

$$\mathcal{E}_n \sim n^{-\frac{\beta}{\beta+1}} \approx \begin{cases} n^{-1} & \beta \to \infty \text{ (Parametric/fast rate)} \\ n^{-1/2} & \beta \to 1 \text{ (Infinite-dimensional)} \\ \text{arbitrarily slow} & \beta \to 0 \text{ (Ill-posed problem)} \end{cases}$$

- When $\beta \to \infty$, the function being learned resides in a space of fixed, finite dimension. In this scenario, the convergence rate of n^{-1} is known as the parametric/fast rate in statistics.
- As $\beta \to 1$ and $m \to \infty$, the function space transitions to an infinite-dimensional setting and $\widehat{\Sigma}$ takes the form of an operator. For such an operator to possess an eigendecomposition, it must be compact. A sufficient condition for $\widehat{\Sigma}$ to be compact is that it belongs to the trace class, which holds if the following series converges:

$$\sum_k \mu_k = \sum_k k^{-\beta} < \infty.$$

  This convergence occurs when $\beta > 1$.

6 Summary

In this lecture, we have examined linear regression in detail, focusing on fundamental principles that can generalize to broader settings. In particular, we introduced a general approach to generalization analysis, applicable to both ridge regression and reduced models, summarized as follows.

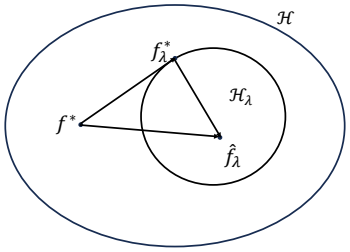

Figure 1: Error decomposition

Consider the task of learning f^* using a model class $\mathcal{H}$. Since $\mathcal{H}$ may be highly expressive, we often impose constraints to improve generalization. Let $\mathcal{H}_\lambda$ be a constrained model class parameterized by λ, which controls the size of the hypothesis space. Denote by $\hat{f}_\lambda \in \mathcal{H}_\lambda$ the model learned from n samples, and let $f_\lambda^* \in \mathcal{H}_\lambda$ a good approximation of f^* within $\mathcal{H}_\lambda$, often chosen (though not necessary) as

$$f_\lambda^* = \inf_{f \in \mathcal{H}_\lambda} \|f - f^*\|.$$

Then, the error decomposition follows as:

$$\begin{aligned}\|\hat{f}_\lambda - f^*\| &\leq \|\hat{f}_\lambda - f^*_\lambda\| + \|f^*_\lambda - f^*\| \\ &\leq \frac{C(\lambda)}{\sqrt{n}} + A(\lambda) \\ &\leq \inf_\lambda \left(\frac{C(\lambda)}{\sqrt{n}} + A(\lambda) \right),\end{aligned}$$

as illustrated in Fig. 1. Note that $\|\hat{f}_\lambda - f^*_\lambda\|$ goes in the Monte-Carlo rate $O(n^{-1/2})$ as we impose constraint on $\mathcal{H}_\lambda$, which should behave well.

When the approximation error $A(\lambda)$ (bias term) is small, the convergence rate approaches the Monte-Carlo rate $O(n^{-1/2})$ (aka fast rate and parametric rate). However, if $A(\lambda)$ is large, the bias-variance trade-off may lead to a significantly slower rate of n^{-r} for some $0 < r < \frac{1}{2}$. Thus, the harder our model approximates the target f^*, the slower the learning error converges as increasing number of samples.

References

[Andriushchenko et al., 2023] Andriushchenko, M., Varre, A. V., Pillaud-Vivien, L., and Flammarion, N. (2023). SGD with large step sizes learns sparse features. In *International Conference on Machine Learning*, pages 903–925. PMLR.

[Bach, 2024] Bach, F. (2024). *Learning theory from first principles*. MIT press.

[Pesme et al., 2021] Pesme, S., Pillaud-Vivien, L., and Flammarion, N. (2021). Implicit bias of SGD for diagonal linear networks: a provable benefit of stochasticity. *Advances in Neural Information Processing Systems*, 34:29218–29230.

[Woodworth et al., 2020] Woodworth, B., Gunasekar, S., Lee, J. D., Moroshko, E., Savarese, P., Golan, I., Soudry, D., and Srebro, N. (2020). Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pages 3635–3673. PMLR.

[Wu and Su, 2023] Wu, L. and Su, W. J. (2023). The implicit regularization of dynamical stability in stochastic gradient descent. In *International Conference on Machine Learning*, pages 37656–37684. PMLR.