CSE 546: Machine Learning **Lecture 3**

Bias-Variance Tradeoff and Dimension-Free Regression

Instructor: Sham Kakade

1 Risk, in the well specified case

Suppose now that the linear model is correct. In particular, assume that:

$$Y = w_*^\top X + \eta$$

where $\eta \sim \mathcal{N}(0, \sigma^2)$ and where X is a vector (we use bold face to denote matrices). Here, η is referred to as the *noise*.

Again, suppose we observe data:

$$\mathcal{T} = (x_1, y_1), \dots (x_n, y_n)$$

where:

$$y_i = w_*^\top x_i + \eta_i$$

so η_i is the noise in the i-th observation.

Define

$$\mathbf{\Sigma} := \frac{1}{n}\mathbf{X}^\top\mathbf{X}$$

Let $\hat{w}_{\mathcal{T}}$ be any estimation procedure using the training set $\mathcal{T}$. We can define the risk of this procedure as:

$$R(\hat{w}_{\mathcal{T}}) = E_{\mathbf{Y}}(\hat{w}_{\mathcal{T}} - w_*)^\top \mathbf{\Sigma}(\hat{w}_{\mathcal{T}} - w_*) := E_{\mathbf{Y}}\|\hat{w}_{\mathcal{T}} - w_*\|^2_{\mathbf{\Sigma}}$$

where the expectation is over the $\mathbf{Y} = [Y_1, \dots Y_n]$. We condition on $X_1, \dots X_n$. Intuitively, this risk is a measure of the error in the parameters.

It is straightforward to see that risk is equivalent the following:

$$R(\hat{w}) = \frac{1}{n}E_{\mathbf{Y}}\|\mathbf{X}\hat{w}_{\mathbf{Y}} - \mathbf{X}w_*\|^2 = \frac{1}{n}\sum_i E_{\mathbf{Y}}(\hat{w}_{\mathbf{Y}}^\top X_i - E[Y|X_i])^2$$

Sometimes this is referred to as *de-noising* (or fixed design) regression, as we are looking at the error on the training set.

1.1 Risk Bounds for Least Squares

Recall that:

$$\hat{w}_{\text{least squares}} = \frac{1}{n}\Sigma^{-1}\mathbf{X}^\top\mathbf{Y}$$

Lemma 1.1. *(Risk Bound) We have that:*

$$R(\hat{w}_{\textit{least squares}}) = \frac{d}{n}\sigma^2$$

Proof. Define η as the noise vector $[\eta_1, \eta_2, \dots \eta_n]$. So we have:

$$\mathbf{Y} = \mathbf{X}w_* + \eta$$

and so:

$$\hat{w} - w_* = \frac{1}{n}\Sigma^{-1}\mathbf{X}^\top(\mathbf{X}w_* + \eta) - w_* = \frac{1}{n}\Sigma^{-1}\mathbf{X}^\top\eta$$

using the definition of $\hat{w}$.

The risk is then:

$$R(\hat{w}) = \frac{1}{n^2}E_\eta\eta^\top\mathbf{X}\Sigma^{-1}\mathbf{X}^\top\eta = \frac{1}{n^2}E_\eta\eta^\top UU^\top\eta = \frac{1}{n}E_\eta\eta^\top UU^\top\eta = \frac{d}{n}\sigma^2$$

where U is the left orthogonal matrix of the thin SVD of $\mathbf{X}$. Here, U is an $n \times d$ orthogonal matrix so $UU^\top$, so $U^\top\eta$ is a d-dimensional Gaussian vector whose distribution is $N(0, \mathrm{I}_d)$ where I_d is the $d \times d$ identity matrix. □

2 What about if $d > n$?

If $d > n$, the risk of the least squares estimator is not useful. There are two common approaches we seek to understand in detail:

- Regularization. The idea is to “shrink” w in a certain manner to reduce variance (and increase bias).
- Feature Selection. The idea is to fit w only in certain directions (and exclude other irrelevant directions).

3 Bias-Variance Tradeoff (in the well specified case)

Lemma 3.1. *(bias-variance for risk) Define $\overline{w} = \mathbb{E}[\hat{w}]$ We can decompose the expected risk as:*

$$\begin{aligned} R(\hat{w}) &= \mathbb{E}_{\mathbf{Y}}\|\hat{w} - \overline{w}\|^2_{\mathbf{\Sigma}} + \|\overline{w} - w_*\|^2_{\mathbf{\Sigma}} \\ &= \frac{1}{n}\mathbb{E}_{\mathbf{Y}}\|\mathbf{X}\hat{w} - \mathbf{X}\overline{w}\|^2 + \frac{1}{n}\|\mathbf{X}w_* - \mathbf{X}\overline{w}\|^2 \end{aligned}$$

where we have that:

$$\textit{variance} = \mathbb{E}_{\mathbf{Y}}\|\hat{w} - \overline{w}\|^2_{\mathbf{\Sigma}} = \frac{1}{n}\mathbb{E}_{\mathbf{Y}}\|\mathbf{X}\hat{w} - \mathbf{X}\overline{w}\|^2$$

and

$$\textit{prediction bias vector} = \mathbf{X}w_* - \mathbf{X}\overline{w}$$

4 Ridge Regression

The ridge regression estimator using an outcome $\mathbf{Y}$ is just:

$$\hat{w}_\lambda = \arg\min_w \frac{1}{n}\|\mathbf{Y} - \mathbf{X}w\|^2 + \lambda\|w\|^2$$

The estimator is then:

$$\hat{w}_\lambda = (\mathbf{\Sigma} + \lambda I)^{-1}(\frac{1}{n}\mathbf{X}^\top\mathbf{Y}) = (\mathbf{\Sigma} + \lambda I)^{-1}(\frac{1}{n}\sum Y_i X_i)$$

4.1 Risk of Ridge Regression

There following bound characterizes the risk of the ridge regression estimator, for a particular choice of λ.

Theorem 4.1. *Assume the linear model is correct: Define $\overline{d}$ as:*

$$\overline{d} = \frac{1}{n} \sum_i \|X_i\|^2$$

For $\lambda = \frac{\sqrt{\overline{d}}}{\|w_*\|\sqrt{n}}$, then:

$$R(\hat{w}_\lambda) \leq \frac{\|w_*\|\sqrt{\overline{d}}}{\sqrt{n}} \leq \frac{\|w_*\|X_+}{\sqrt{n}}$$

where X_+ is a bound on the norm of $\|X\|_i$.

5 What about prediction error and model mis-specification?

We have worked under a somewhat unrealistic setting in that:

- We have assumed the model is correct.
- We have assumed the noise is Gaussian
- Our notion of Risk is measured under the observed points $X_1, \ldots X_n$, while we often care about our prediction on new points?

Roughly speaking, most of these results transfer over when all of these assumptions are relaxed. We will see one such example later, where we look at stochastic gradient descent.

6 Coordinate Ascent

How should we fit a 'big' model?

Suppose we want to optimize a function $L(w_1, w_2, \ldots w_d)$ where d is large. A simple iterative method is to start with some vector w. Then one can randomly pick a subset of coordinates of coordinates and improve the value of f where we only change the values of w on this subset.