1   Projection, Approximation, and Estimation

In this discussion, we will revisit a fundamental issue that ought to have bothered you throughout the class so far. In typical applications, we are dealing with data generated by an *unknown* function (with some noise), and our goal is to estimate this unknown function from samples. So far, we have used linear and polynomial regression as our only methods, but are these good methods when the function is not a polynomial?

We will answer a few aspects of this general question. In particular, we will provide geometric answers to:

- Can we do least squares on an arbitrary problem? What do we end up doing in this case?

- How large does the degree have to be for us to execute reliable polynomial regression?

- Can we formulate the bias-variance trade-off of polynomial regression?

Doing this discussion in sequence will set up all the necessary tools you need to analyze the problem. You are recommended to draw pictures to understand what these projections are doing. That's a great way to develop intuition for what is going on!

Define the projection of a vector $\mathbf{y} \in \mathbb{R}^n$ onto a (closed) set $\mathcal{C}$ as

$$P_{\mathcal{C}}(\mathbf{y}) = \arg \min_{\mathbf{u} \in \mathcal{C}} \|\mathbf{y} - \mathbf{u}\|_2^2.$$

For a matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ having full column rank, let the set $c(\mathbf{X})$ denote its column space.

By definition, the projection $P_{c(\mathbf{X})}(\mathbf{y})$ is given by the solution to the following least squares problem:

$$P_{c(\mathbf{X})}(\mathbf{y}) = \mathbf{X}(\arg \min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2).$$

We will use the notation $P_{\mathbf{X}}(\mathbf{y}) := P_{c(\mathbf{X})}(\mathbf{y})$ in the rest of this discussion section for convenience.

(a) Consider the task of fitting a set of noisy observations using a given set of features.

- Let $\mathbf{y}^* \in \mathbb{R}^n$ denote the true signal, i.e., the set of observations if there was no noise. Note that $\mathbf{y}^*$ **is a given vector.**

- Let $\mathbf{y} = \mathbf{y}^* + \mathbf{z}$ denote the observations that are available to us, where $\mathbf{z}$ denotes the noise that corrupts the true signal. Because $\mathbf{z}$ is a random vector, $\mathbf{y}$ **is a random vector too.**
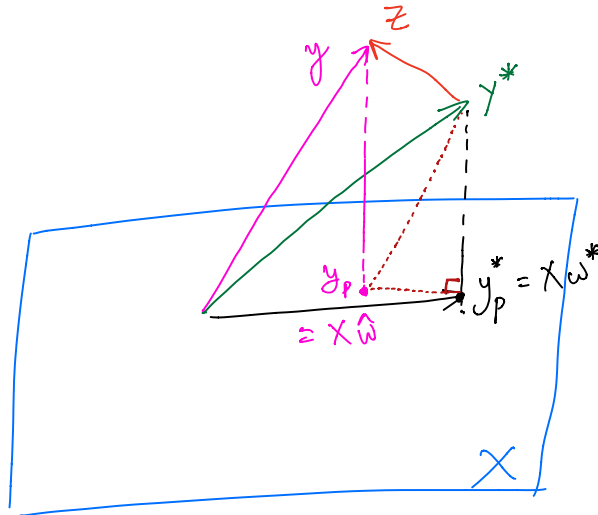
Figure 1: Geometrical representation of the truth, noisy observations, their projections on the feature space and the relations between them.

- Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ denote the given features matrix. Assume that $\mathbf{X}$ has full column rank. Also assume that the matrix $\mathbf{X}$ is fixed.

Given $\mathbf{y}$ and $\mathbf{X}$, our goal is to estimate $\mathbf{y}^*$ as well as possible. Define the vectors $\mathbf{w}^*$ and $\hat{\mathbf{w}}$ as follows:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \|\mathbf{y}^* - \mathbf{X}\mathbf{w}\|_2^2 \qquad \text{and} \qquad \hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2.$$

Note that for a given true signal $\mathbf{y}^*$ the vector $\mathbf{w}^*$ is fixed, but the vector $\hat{\mathbf{w}}$ is a random variable since it is a function of the random noise $\mathbf{z}$. Now, show the following equalities:

$$\mathbf{X}\mathbf{w}^* = P_{\mathbf{X}}(\mathbf{y}^*) = \mathbf{y}_P^*, \qquad \text{and} \tag{1}$$

$$\mathbf{X}\hat{\mathbf{w}} = P_{\mathbf{X}}(\mathbf{y}) =: \mathbf{y}_P, \tag{2}$$

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \|\mathbf{X}\mathbf{w}^* + \mathbf{z} - \mathbf{X}\mathbf{w}\|_2^2. \tag{3}$$

You may use the results from part (b) and (c) to prove these results.

(b) Show that:
$$\|\mathbf{y}^* - P_{\mathbf{X}}(\mathbf{y}^* + \mathbf{z})\|_2^2 = \|\mathbf{y}^* - P_{\mathbf{X}}(\mathbf{y}^*)\|_2^2 + \|P_{\mathbf{X}}(\mathbf{z})\|_2^2.$$

Hint: Recall the Pythagorean theorem.

(c) Let us introduce the shorthand $\mathbf{y}_P^* = P_{\mathbf{X}}(\mathbf{y}^*)$. Use the previous part to show that

$$\|\mathbf{y}^* - P_{\mathbf{X}}(\mathbf{y}^* + \mathbf{z})\|_2^2 = \|\mathbf{y}^* - \mathbf{y}_P^*\|_2^2 + \|\mathbf{y}_P^* - P_{\mathbf{X}}(\mathbf{y}_P^* + \mathbf{z})\|_2^2.$$

Hint: What is the projection $P_{\mathcal{C}}(\mathbf{v})$ when $\mathbf{v} \in \mathcal{C}$?

(d) Use the results obtained in parts (a)-(c) to argue the following equalities:

$$\mathbb{E}_{\mathbf{z}}\left[\|\mathbf{y}^* - \mathbf{X}\hat{\mathbf{w}}\|_2^2\right] = \mathbb{E}_{\mathbf{z}}\left[\|\mathbf{y}^* - P_{\mathbf{X}}(\mathbf{y}^* + \mathbf{z})\|_2^2\right] \tag{4}$$

$$= \|\mathbf{y}^* - \mathbf{X}\mathbf{w}^*\|_2^2 + \mathbb{E}_{\mathbf{z}}\left[\|\mathbf{X}\mathbf{w}^* - \mathbf{X}\hat{\mathbf{w}}\|_2^2\right] \tag{5}$$

$$= \underbrace{\|\mathbf{y}^* - \mathbb{E}_{\mathbf{z}}[\mathbf{X}\hat{\mathbf{w}}]\|_2^2}_{\text{bias}} + \underbrace{\mathbb{E}_{\mathbf{z}}\left[\|\mathbb{E}_{\mathbf{z}}[\mathbf{X}\hat{\mathbf{w}}] - \mathbf{X}\hat{\mathbf{w}}\|_2^2\right]}_{\text{variance}}. \tag{6}$$

Finally, conclude that the error of estimating an arbitrary vector $\mathbf{y}^*$ corrupted by noise via linear regression is bounded by the sum of two terms i) an approximation error, which captures how far $\mathbf{y}^*$ is from the assumed linear model, and ii) an estimation error term, which captures what the error would have been if the true signal indeed were to come from the assumed linear model.

(e) When $\mathbf{X}$ is a full column-rank $n \times d$ matrix and the noise is standard Gaussian, i.e, $\mathbf{z} \sim \mathcal{N}^d(\mathbf{0}, \mathbf{I})$, we derive in HW3 that $\frac{1}{n}\mathbb{E}_{\mathbf{z}}\left[\|\mathbf{X}\mathbf{w}^* - \mathbf{X}\hat{\mathbf{w}}\|_2^2\right] = d/n$.

Let us say that we obtain n samples $\{x_i, y_i\}_{i=1}^n$, where $y_i = \sin(x_i) + \mathbf{z}_i$. Here, each point $x_i \in [-3, 3]$ is distinct, and each $\mathbf{z}_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$ represents independent random noise. Since

the function is unknown to us a-priori, we decide to use a polynomial regression with degree D to estimate the relationship between x_i and y_i. Stack up the noiseless function evaluations into the vector $\mathbf{y}^*$, whose ith coordinate is given by $y_i^* = \sin(x_i)$.

Using Taylor expansion and the above parts, show that if our estimate $\hat{\mathbf{y}}$ is obtained by performing least squares, then we have

$$\frac{1}{n}\mathbb{E}_{\mathbf{z}}\left[\|\mathbf{y}^* - \hat{\mathbf{y}}\|_2^2\right] \leq \left(\frac{3^{D+1}}{(D+1)!}\right)^2 + \frac{D+1}{n}.$$

(f) In the previous part, notice that as D increases, the approximation error decreases but the estimation error increases. Discuss qualitatively why that is true. Given n samples, show that setting $D = O(\log n / \log \log n)$ is an optimal choice for this problem.
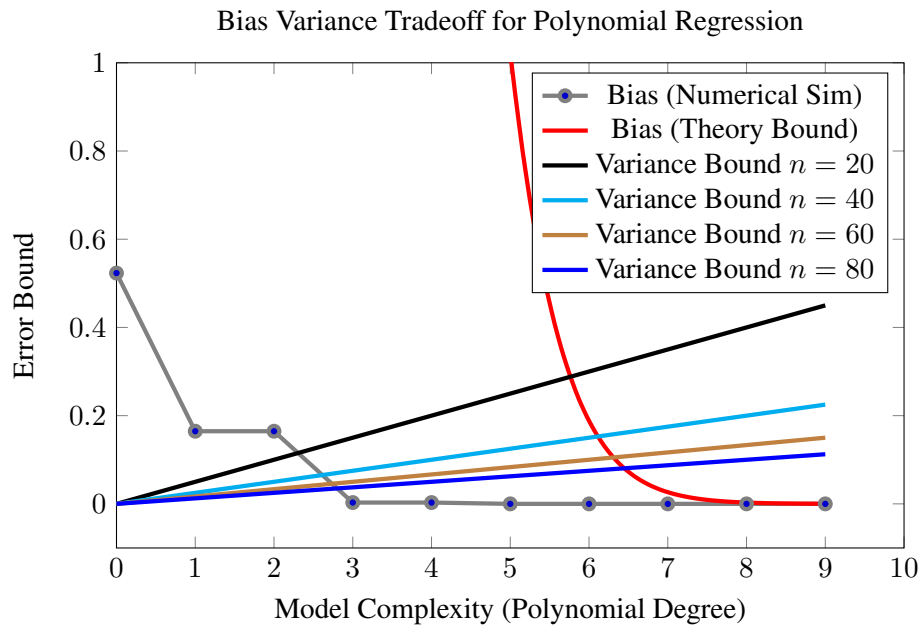

Figure 2: Bias Variance Tradeoff for Polynomial Regression