CS 189 Introduction to Machine Learning
Spring 2018

DIS3

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.

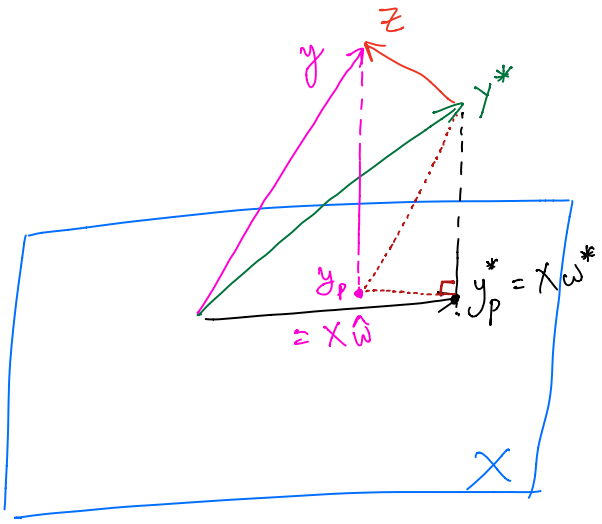
By definition, the projection $P_{c(\mathbf{X})}(\mathbf{y})$ is given by the solution to the following least squares problem:

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

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.**



Pythagoras Thm: $\| y^* - y_p \|^2 = \| y^* - y_p^* \|^2 + \| y_p - y_p^* \|^2$

$\| y^* - y_p \|^2$: Approx. error (bias)

$\| y_p - y_p^* \|^2$: est. error (variance)

$= \| y^* - X w^* \|^2 + \| X \hat{w} - X w^* \|^2$

$\| y^* - X w^* \|^2$: decreases with increase in model complexity

$\| X \hat{w} - X w^* \|^2$: increases with increase in model complexity

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.

Solution: It is easy to derive equation (1) and (2) by following the definition of $P_{\mathbf{X}}(\mathbf{y})$, $\mathbf{w}^*$, and $\hat{\mathbf{w}}$. We now provide two methods for equation (3).

Method 1: Using Pythagoras Theorem: We will show that the minimizer of following two problems would be the same:

$$\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 \quad \text{and} \quad \min_{\mathbf{w}} \|\mathbf{X}\mathbf{w}^* + \mathbf{z} - \mathbf{X}\mathbf{w}\|^2.$$

We start with the first minimization problem and show that finding its argmin is equivalent to finding argmin of the second problem (using parts (b) and (c)).

$$\begin{aligned}
\arg\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 &= \arg\min_{\mathbf{w}} \|(\mathbf{y} - P_{\mathbf{X}}(\mathbf{y})) + P_{\mathbf{X}}(\mathbf{y}) - \mathbf{X}\mathbf{w}\|^2 \\
&\overset{(i)}{=} \arg\min_{\mathbf{w}} \|(\mathbf{y} - P_{\mathbf{X}}(\mathbf{y}))\|^2 + \|P_{\mathbf{X}}(\mathbf{y}) - \mathbf{X}\mathbf{w}\|^2 \\
&\overset{(ii)}{=} \arg\min_{\mathbf{w}} \|P_{\mathbf{X}}(\mathbf{y}) - \mathbf{X}\mathbf{w}\|^2 \\
&\overset{(iii)}{=} \arg\min_{\mathbf{w}} \|P_{\mathbf{X}}(\mathbf{y}^* + \mathbf{z}) - \mathbf{X}\mathbf{w}\|^2
\end{aligned}$$

where step (i) follows by part (b) (Pythagoras theorem), step (ii) from the fact that the first term does not depend on $\mathbf{w}$, and step (iii) from the equality $\mathbf{y} = \mathbf{y}^* + \mathbf{z}$. Now, we have

$$\begin{aligned}
\arg\min_{\mathbf{w}} \|P_{\mathbf{X}}(\mathbf{y}^* + \mathbf{z}) - \mathbf{X}\mathbf{w}\|^2 &\overset{(iv)}{=} \arg\min_{\mathbf{w}} \|P_{\mathbf{X}}(\mathbf{y}^*) + P_{\mathbf{X}}(\mathbf{z}) - \mathbf{X}\mathbf{w}\|^2 \\
&\overset{(v)}{=} \arg\min_{\mathbf{w}} \|P_{\mathbf{X}}(\mathbf{y}^*) + P_{\mathbf{X}}(\mathbf{z}) - \mathbf{X}\mathbf{w}\|^2 + \|\mathbf{z} - P_{\mathbf{X}}(\mathbf{z})\|^2 \\
&\overset{(vi)}{=} \arg\min_{\mathbf{w}} \|\mathbf{z} - P_{\mathbf{X}}(\mathbf{z}) + P_{\mathbf{X}}(\mathbf{y}^*) + P_{\mathbf{X}}(\mathbf{z}) - \mathbf{X}\mathbf{w}\|^2 \\
&= \|\mathbf{X}\mathbf{w}^* + \mathbf{z} - \mathbf{X}\mathbf{w}\|^2
\end{aligned}$$

Here step (iv) follows from the linearity of the projection map, step (v) from the fact that $\|\mathbf{z} - P_{\mathbf{X}}(\mathbf{z})\|^2$ does not depend on $\mathbf{w}$ and step (vi) from Pythagoras' theorem (part (b)).

Methods 2: Using OLS closed form solutions: Using the closed form solutions for OLS, we obtain that

$$\mathbf{w}^* = (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbf{y}^*$$
$$\hat{\mathbf{w}} = (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal (\mathbf{y}^* + \mathbf{z}),$$

which means that

$$\hat{\mathbf{w}} = \mathbf{w}^* + (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbf{z}. \tag{4}$$

This is the solution of (3) if we rewrite (3) into

$$\begin{aligned}\hat{\mathbf{w}} &= \arg\min_{\mathbf{w}} \|\mathbf{X}\mathbf{w}^* + \mathbf{z} - \mathbf{X}\mathbf{w}\|_2^2 \\ &= \arg\min_{\mathbf{w}} \|\mathbf{X}(\mathbf{w} - \mathbf{w}^*) - \mathbf{z}\|_2^2 \\ &= \mathbf{w}^* + \arg\min_{\mathbf{w}'} \|\mathbf{X}\mathbf{w}' - \mathbf{z}\|_2^2.\end{aligned}$$

Methods 3: Using the Linearity of Projection Call $\hat{\mathbf{w}}_1$ the minimizer of $\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$, and $\hat{\mathbf{w}}_2$ the minimizer of $\|\mathbf{X}\mathbf{w}^* + \mathbf{z} - \mathbf{X}\mathbf{w}\|_2^2$. By definition, we have $\mathbf{X}\hat{\mathbf{w}}_1 = P_{\mathbf{X}}(\mathbf{y})$, and $\mathbf{X}\hat{\mathbf{w}}_2 = P_{\mathbf{X}}(\mathbf{X}\mathbf{w}^* + \mathbf{z})$. Since $\mathbf{X}$ is full column rank, $\mathbf{X}$ has trivial nullspace. Thus for any $\mathbf{w}_1$,$\mathbf{w}_2$, to show $\mathbf{w}_1 = \mathbf{w}_2$ it suffices to show $\mathbf{X}\mathbf{w}_1 = \mathbf{X}\mathbf{w}_2$, and $\mathbf{X}(\mathbf{w}_1 - \mathbf{w}_2) = \mathbf{0}$ implies $\mathbf{w}_1 = \mathbf{w}_2$. But this is easy due to properties of the projection. $P_{\mathbf{X}}(\mathbf{y}) = P_{\mathbf{X}}(\mathbf{y}^* + \mathbf{z}) = P_{\mathbf{X}}(\mathbf{y}^*) + P_{\mathbf{X}}(\mathbf{z}) = \mathbf{X}\mathbf{w}^* + P_{\mathbf{X}}(\mathbf{z})$ using linearity, but $\mathbf{X}\mathbf{w}^* = P_{\mathbf{X}}(\mathbf{X}\mathbf{w}^*)$ by idempotency so that $P_{\mathbf{X}}(\mathbf{y}) = P_{\mathbf{X}}(\mathbf{X}\mathbf{w}^* + \mathbf{z})$. Hence $\hat{\mathbf{w}}_1 = \hat{\mathbf{w}}_2$.

(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.

Solution: Recall that the error vector $\mathbf{y}^* - P_{\mathbf{X}}(\mathbf{y}^*)$ is orthogonal to any vector in the column space of $\mathbf{X}$ (review the geometry of least squares if this is not clear to you).

Hence, by the Pythagorean theorem, we have

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

(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 $\mathbf{v} \in \mathcal{C}$?

Solution: Notice that by definition, $P_{\mathcal{C}}(\mathbf{y}) = \mathbf{y}$ when $\mathbf{y} \in \mathcal{C}$. Hence, by the linearity of projections, we may write

$$P_{\mathbf{X}}(\mathbf{y}_P^* + \mathbf{z}) = P_{\mathbf{X}}(\mathbf{y}_P^*) + P_{\mathbf{X}}(\mathbf{z})$$

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

Using the previous part proves the required result.

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

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

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

$$= \underbrace{\left\|\mathbf{y}^* - \mathbb{E}_{\mathbf{z}}\left[\mathbf{X}\hat{\mathbf{w}}\right]\right\|_2^2}_{\text{bias}} + \underbrace{\mathbb{E}_{\mathbf{z}}\left[\left\|\mathbb{E}_{\mathbf{z}}\left[\mathbf{X}\hat{\mathbf{w}}\right] - \mathbf{X}\hat{\mathbf{w}}\right\|_2^2\right]}_{\text{variance}}. \tag{7}$$

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.

Solution: Note that (part (a) equation (1)) $\mathbf{X}\hat{\mathbf{w}} = \mathbf{y}_P = P_{\mathbf{X}}(\mathbf{y}^* + \mathbf{z})$ and hence

$$\begin{aligned}
\|\mathbf{y}^* - \mathbf{X}\hat{\mathbf{w}}\|_2^2 &= \|\mathbf{y}^* - \mathbf{y}_P\|_2^2 \\
&\overset{(i)}{=} \|\mathbf{y}^* - P_{\mathbf{X}}(\mathbf{y}^* + \mathbf{z})\|_2^2 \\
&= \|\mathbf{y}^* - \mathbf{y}_P^*\|_2^2 + \|\mathbf{y}_P^* - \mathbf{y}_P\|_2^2 \quad \text{(using Pythagoras theorem)} \\
&\overset{(ii)}{=} \|\mathbf{y}^* - \mathbf{X}\mathbf{w}^*\|_2^2 + \|\mathbf{X}\mathbf{w}^* - \mathbf{X}\hat{\mathbf{w}}\|_2^2 \quad \text{(using equations (3) and (1)).}
\end{aligned}$$

Taking expectations both sides in step (i) proves the result (5) and in step (ii) proves the result (6).

In the homework, you would derive that $E[\mathbf{X}\hat{\mathbf{w}}] = \mathbf{X}\mathbf{w}^*$ and substituting this fact into the previous result gives the bias-variance decomposition in equation (7), The first term in this expansion is the approximation error term, capturing how far $\mathbf{y}^*$ is from the column space of $\mathbf{X}$. The second term is the estimation term, showing how well we can estimate the true linear model in presence of noise.

(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

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

Solution: We first evaluate the approximation error and estimation error defined in the previous part. In other words:

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

From HW3, we know that the estimation error is given by

$$\frac{1}{n}\mathbb{E}_{\mathbf{z}}\left[\|\mathbf{X}\mathbf{w}^* - \mathbf{X}\hat{\mathbf{w}}\|_2^2\right] = \frac{\sigma^2(D+1)}{n}.$$

For the approximation error, we use the Dth order Taylor series approximation of $\sin x$ about the point 0, which is given (for some $x' \in [0, x]$) by

$$\sin(x) = \underbrace{\sum_{i=1,3,..}^{D} (-1)^{(i-1)/2}\frac{x^i}{i!}}_{\phi_D(x)} + f^{n+1}(\zeta)\frac{x^{D+1}}{(D+1)!}.$$

Now notice that for all $x \in [-3, 3]$ and $|\sin(x)| \leq 1$, we have the relation $|\sin(x) - \phi_D(x)| \leq \frac{3^{D+1}}{(D+1)!}$. Consequently, the approximation error for the sample $\{x_i, y_i^*\}$ is bounded by $\frac{3^{D+1}}{(D+1)!}$. Summing and dividing through by n yields the result.

(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.

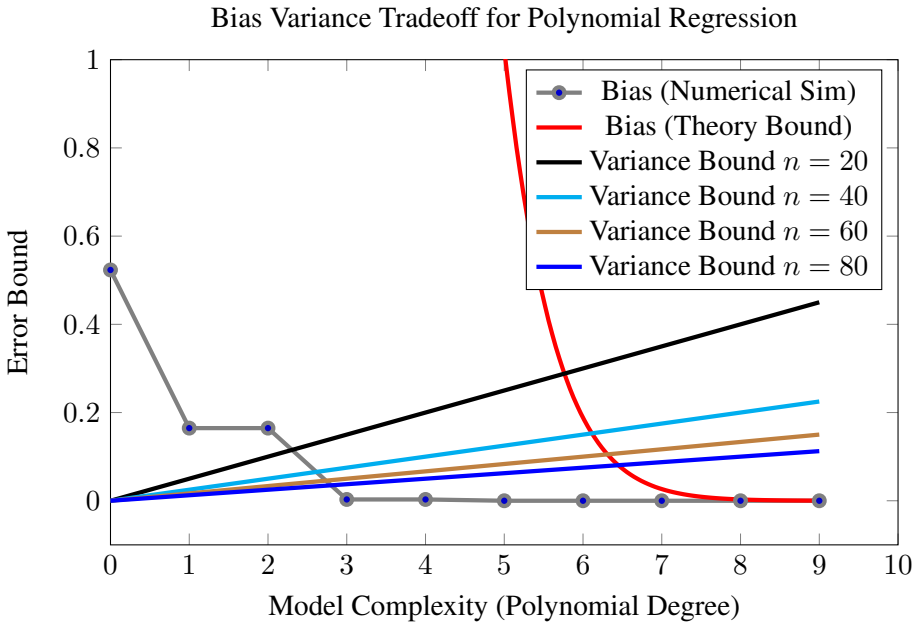


Figure 2: Bias Variance Tradeoff for Polynomial Regression

Solution:

The approximation error measures how well our regression model fits the true data. Since we can better approximate the unknown function with higher order polynomials, this term decreases with D. However, it becomes harder to estimate the correct polynomial, since we now give ourselves extra freedom. This is the bias variance trade-off in action!

Using the intuition from Figure 2, we can equate the two terms in the upper bound of part (e), i.e.,

$$\frac{3^{D+1}}{(D+1)!} \approx \frac{\sigma\sqrt{(D+1)}}{\sqrt{n}}.$$

Applying the Stirling approximation, making a few approximations and substituting $\sigma = 1$ yields the required result.

$$\begin{aligned}
\frac{3^{D+1}}{(D+1)^{D+1}} &\approx \frac{\sqrt{(D+1)}}{\sqrt{n}} \\
\frac{3^D}{D^D} &\approx \frac{\sqrt{D}}{\sqrt{n}} \\
\sqrt{n} &\approx \left(\frac{D}{3}\right)^D \\
\log n &\approx D \log\left(\frac{D}{3}\right).
\end{aligned}$$

Now we verify if the approximation is valid for $D = \log n / \log\log n$:

$$\begin{aligned}
D \log\left(\frac{D}{3}\right) &\approx D \log D \\
&= \frac{\log n}{\log\log n} \cdot \log\left(\frac{\log n}{\log\log n}\right) \\
&= \frac{\log n}{\log\log n} \cdot (\log\log n - \log\log\log n) \\
&\approx \log n.
\end{aligned}$$