Lecture 19 and 20: MATH 342W: Introduction to Data Science and Machine Learning

Sergio E. Garcia Tapia[*]

April 22nd and 24th, 2025 (last updated May 15, 2025)

R Demo

See `QC_MATH_342W_Spring_2025/practice_lectures/lec19.Rmd`.

Bias-Variance Tradeoff for Regression

Randomness in Ignorance

Let $\mathcal{Y} = \mathbb{R}$ and $y = f(\boldsymbol{x}) + \delta$, where $\boldsymbol{x}$ is a constant. We will make the following two assumptions:

(I) **Zero mean-centered and mean independent**: δ is a realization from Δ, a random variable which is mean-independent and mean zero. That is,

$$\forall \boldsymbol{x} : \mathbb{E}[\Delta \mid \boldsymbol{x}] = 0 \implies \mathbb{E}[\Delta] = 0 \tag{1}$$

where $\mathbb{E}[\cdot]$ denotes expectation.

(II) **Homoskedasticity**: i.e., constant variance:

$$\begin{aligned} \forall \boldsymbol{x} : \sigma^2 &:= \operatorname{Var}[\Delta \mid \boldsymbol{x}] && (2) \\ &= \mathbb{E}[\Delta^2 \mid \boldsymbol{x}] - (\mathbb{E}[\Delta \mid \boldsymbol{x}])^2 \\ &= \mathbb{E}[\Delta^2 \mid \boldsymbol{x}] && \text{(by (I))} \end{aligned}$$

See Figure 1. If δ is random from Δ, then y is also random from Y, so

$$Y = f(\boldsymbol{x}) + \Delta \tag{3}$$

and hence

$$\begin{aligned} \mathbb{E}[Y \mid \boldsymbol{x}] &= \mathbb{E}[f(\boldsymbol{x}) + \Delta \mid \boldsymbol{x}] \\ &= \mathbb{E}[f(\boldsymbol{x}) \mid \boldsymbol{x}] + \mathbb{E}[\Delta \mid \boldsymbol{x}] \\ &= f(\boldsymbol{x}) && \text{(by (I))} \end{aligned}$$

[*] Based on lectures of Dr. Adam Kapelner at Queens College. See also the course GitHub page.

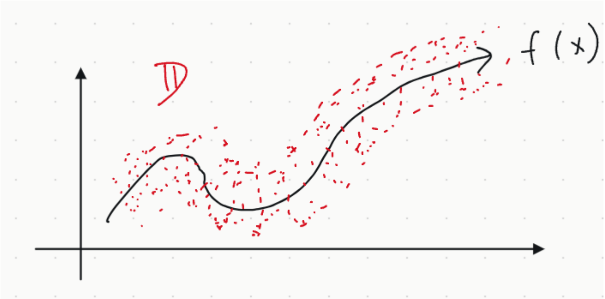

Figure 1: Depiction of zero mean-centered and mean independent Δ, and homoskedasticity.

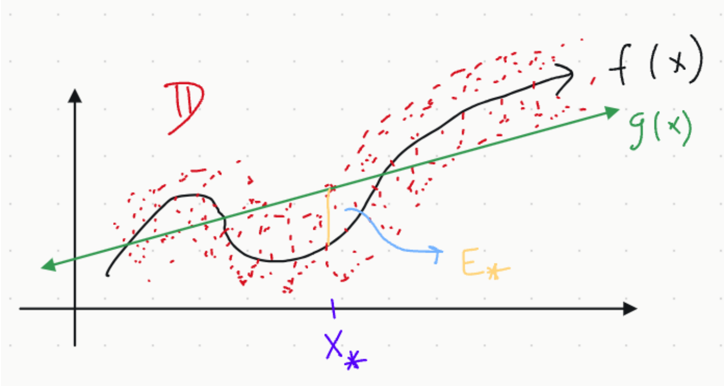

Figure 2: Depiction of random variable E_*

Thus we call f the **conditional expectation function (CEF)**. Let g be a fitted model with $\mathbb{D}$, meaning

$$\begin{aligned} y &= g + e = g + \overbrace{(f-g)+\delta}^{e} \\ e &= f - g + \delta \end{aligned}$$

Here e is the residual, which has both misspecification and estimation. By Equation 1, we can say e is a realization of a random variable E

$$\begin{aligned} Y &= g + \overbrace{(f-g)+\Delta}^{E} \\ E &= (f-g)+\Delta = Y - g \end{aligned}$$

Let's predict on a new observation $\boldsymbol{x}_*$:

$$Y_* = g(\boldsymbol{x}_*) + f(\boldsymbol{x}_*) - g(\boldsymbol{x}_*) + \Delta_*$$

Then the associated error is

$$E_* = f(\boldsymbol{x}_*) - g(\boldsymbol{x}_*) + \Delta_* = Y_* - g(\boldsymbol{x}_*) \tag{4}$$

See Figure 2. Define the **bias** to be

$$\text{Bias}(\boldsymbol{x}_*) := \mathbb{E}[E_* \mid \boldsymbol{x}_*] \tag{5}$$

Then using Equation 4

$$\begin{aligned}
\mathbb{E}[E_* \mid \boldsymbol{x}_*] &= \mathbb{E}[f(\boldsymbol{x}_*) - g(\boldsymbol{x}_*) + \Delta_* \mid \boldsymbol{x}_*] && \text{(by 4)} \\
&= f(\boldsymbol{x}_*) - g(\boldsymbol{x}_*) + \underbrace{\mathbb{E}[\Delta_* \mid \boldsymbol{x}_*]}_{0} && (f \text{ and } g \text{ not random on } \boldsymbol{x}) \\
&= f(\boldsymbol{x}_*) - g(\boldsymbol{x}_*) && \text{(by 1)}
\end{aligned}$$

where $\mathbb{E}$ is expectation on Δ_* since that is what is random. Hence we have

$$\text{Bias}(\boldsymbol{x}_*) = f(\boldsymbol{x}_*) - g(\boldsymbol{x}_*) \tag{6}$$

Define **Mean-Squared Error (MSE)** by

$$MSE(\boldsymbol{x}_*) :=\mathbb{E}[E_*^2 \mid \boldsymbol{x}_*] \tag{7}$$

which we can simplify to

$$\begin{aligned}
\mathbb{E}[E_*^2 \mid \boldsymbol{x}_*] &= \mathbb{E}[(Y_* - g(\boldsymbol{x}_*))^2 \mid \boldsymbol{x}_*] \\
&= \mathbb{E}[Y_*^2 - 2g(\boldsymbol{x}_*)Y_* + g(\boldsymbol{x}_*)^2 \mid \boldsymbol{x}_*] \\
&= \mathbb{E}[Y_*^2 \mid \boldsymbol{x}_*] - 2g(\boldsymbol{x}_*)\mathbb{E}[Y_* \mid \boldsymbol{x}_*] + g(\boldsymbol{x}_*)^2 && (g \text{ is not random in } \boldsymbol{x}_*) \\
&= \mathbb{E}[(f(\boldsymbol{x}_*) + \Delta_*)^2 \mid \boldsymbol{x}_*] - 2g(\boldsymbol{x}_*)\mathbb{E}[f(\boldsymbol{x}_*) + \Delta_* \mid \boldsymbol{x}_*] + g(\boldsymbol{x}_*)^2 \\
& && (\text{since } Y_* = f(\boldsymbol{x}_*) + \Delta_*) \\
&= f(\boldsymbol{x}_*)^2 + 2f(\boldsymbol{x}_*)\underbrace{\mathbb{E}[\Delta_* \mid \boldsymbol{x}]}_{0} + \underbrace{\mathbb{E}[\Delta_*^2 \mid \boldsymbol{x}_*]}_{\sigma^2} - 2g(\boldsymbol{x}_*)f(\boldsymbol{x}_*) + \underbrace{\mathbb{E}(\Delta_* \mid \boldsymbol{x}_*)}_{0} + g(\boldsymbol{x}_*)^2 \\
& && (f \text{ is not random in } \boldsymbol{x}_*) \\
&= \sigma^2 + (f(\boldsymbol{x}_*) - g(\boldsymbol{x}_*))^2 && \text{(by 1 and 2)} \\
&= \sigma^2 + (\text{Bias}[\boldsymbol{x}_*])^2 && \text{(by 6)}
\end{aligned}$$

If $g(\boldsymbol{x}) = f(\boldsymbol{x})$, then the only error is σ^2, called the **irreducible error**. This comes from Δ, which we don't know. Recall bias results from misspecification and estimation.

Allowing Randomness in Responses

Suppose now that $\mathbb{D}$ has randomness in the responses (not in X). Then $y_1, \dots, y_n$ are realized are from $Y_1, \dots, Y_n$ via $\Delta_1, \dots, \Delta_n$, but $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ are constant. We make a third assumption, in addition to 1 and 2:

$$\text{(III) } \Delta_1, \dots, \Delta_n \text{ are independent}$$

What's different? The realizations of Δ are different, which means the models are different. There are uncountably-infinitely many data sets; what changes is the offset from $f(x)$. In the real world, you see one data set $\mathbb{D}$, and you would construct one $g(x)$. Each will be different because g is a function of the algorithm, and since the y's in $\mathbb{D}$ are random, g is in fact a realization of a corresponding random G:

$$G = \mathcal{A}(\mathbb{D}, \mathcal{H}) = \mathcal{A}(\langle X, \vec{Y} \rangle, \mathcal{H})$$

where $\vec{Y}$ is a (vector) random variable. Now consider the Mean-Squared Error again, defined as in 7. In the following, $\mathbb{E}_{\mathbb{D}}$ means expectation involving randomness in the

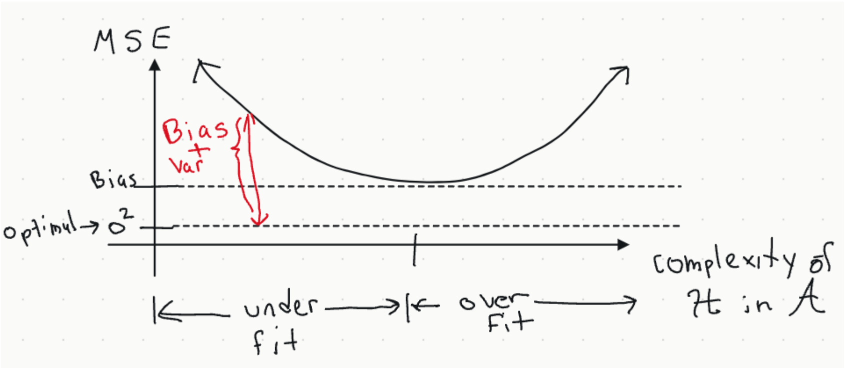

Figure 3: Depiction of bias-variance decomposition.

responses in $\mathbb{D}$ (as a consequence of randomness in $\Delta_1, \ldots, \Delta_n$), while $\mathbb{E}_{\Delta_*}$ means expectation involving randomness in Δ_*:

$$\begin{aligned}
MSE[\boldsymbol{x}_*] :&= \underset{\Delta_*,\mathbb{D}}{\mathbb{E}}[(Y_* - G(\boldsymbol{x}_*))^2 \mid \boldsymbol{x}_*, X] \\
&= \underset{\Delta_*,\mathbb{D}}{\mathbb{E}}[Y_*^2 \mid \boldsymbol{x}_*, X] - 2\underset{\Delta_*,\mathbb{D}}{\mathbb{E}}[Y_* \cdot G(\boldsymbol{x}_*) \mid \boldsymbol{x}_*, X] + \underset{\Delta_*,\mathbb{D}}{\mathbb{E}}[G(\boldsymbol{x}_*)^2 \mid \boldsymbol{x}_*, X] \\
&= \mathbb{E}_{\Delta_*}[Y_*^2 \mid \boldsymbol{x}_*, X] - 2\mathbb{E}_{\Delta_*}[Y_* \mid \boldsymbol{x}_*, X] \cdot \mathbb{E}_{\mathbb{D}}[G(\boldsymbol{x}_*) \mid \boldsymbol{x}_*, X] + \mathbb{E}_{\mathbb{D}}[G(\boldsymbol{x}_*)^2 \mid \boldsymbol{x}_*, X] \\
&\qquad \text{(by independence, and } Y_* \text{ only random in } \Delta_*,\ G \text{ only random in } \mathbb{D}) \\
&= \sigma^2 + f(\boldsymbol{x}_*)^2 - 2f(\boldsymbol{x}_*)\mathbb{E}_{\mathbb{D}}[G(\boldsymbol{x}_*) \mid \boldsymbol{x}_*, X] + \mathbb{E}_{\mathbb{D}}[G(\boldsymbol{x}_*)^2 \mid \boldsymbol{x}_*, X] \quad \text{(by 3, 1, 2)} \\
&= \sigma^2 + f(\boldsymbol{x}_*)^2 - 2f(\boldsymbol{x}_*)\mathbb{E}_{\mathbb{D}}[G(\boldsymbol{x}_*) \mid \boldsymbol{x}_*, X] + (\mathbb{E}_{\mathbb{D}}[G(\boldsymbol{x}_*) \mid \boldsymbol{x}_*, X])^2 + \mathrm{Var}[G(\boldsymbol{x}_*) \mid \boldsymbol{x}_*, X] \\
&\qquad \text{(by } \mathrm{Var}[U] = \mathbb{E}[U^2] - \mathbb{E}[U]^2) \\
&= \sigma^2 + (f(\boldsymbol{x}_*) - \mathbb{E}_{\mathbb{D}}[G(\boldsymbol{x}_*) \mid \boldsymbol{x}_*, X])^2 + \mathrm{Var}[G(\boldsymbol{x}_*) \mid \boldsymbol{x}_*, X] \\
&\qquad \text{(by } (a-b)^2 = a^2 - 2ab + b^2) \\
&= \sigma^2 + (\mathbb{E}_{\mathbb{D}}[f(\boldsymbol{x}_* - G(\boldsymbol{x}_*) \mid \boldsymbol{x}_*, X])^2 + \mathrm{Var}[G(\boldsymbol{x}_*) \mid \boldsymbol{x}_*, X] \\
&\qquad (f \text{ is not random on } \mathbb{D}) \\
&= \sigma^2 + (\mathrm{Bias}[G(\boldsymbol{x}_*)])^2 + \mathrm{Var}[G(\boldsymbol{x}_*)] \qquad \text{(by 5)}
\end{aligned}$$

In summary, we have

$$MSE[\boldsymbol{x}_*] = \sigma^2 + (\mathrm{Bias}[G(\boldsymbol{x}_*)])^2 + \mathrm{Var}[G(\boldsymbol{x}_*)] \tag{8}$$

Equation 8 is called the **Bias-Variance decomposition**, or **Bias-Variance trade-off**. Intrigued by the name, it is natural to ask whether there is in fact a trade-off between bias and variance, and in what sense.

- If you define "trade-off" as a *zero-sum game*, then the answer is *no*, since there are algorithms $\mathcal{A}$ that reduce *both* bias and variance simultaneously.
- If by "trade-off" you instead mean *decomposition*, then yes, there is a "trade-off", as both terms appear in Equation 8.

See Figure 3.

Allowing Randomness in Covariates

We make another generalization: randomness in the $\boldsymbol{x}$'s. Thus, further assume $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n, \boldsymbol{x}_*$ are random, drawn from random variable $\vec{X}$. Before, we calculated the MSE over a fixed

$\boldsymbol{x}$, but now, we do so over *all* observations:

$$\begin{aligned}MSE :&= \mathbb{E}_{\vec{X}}[MSE[\boldsymbol{x}]] \\ &= \mathbb{E}_{\vec{X}}[\sigma^2 + (\text{Bias}[G(\boldsymbol{x}_*)])^2 + \text{Var}[G(\boldsymbol{x}_*)]] && \text{(by 8)} \\ &= \sigma^2 + \mathbb{E}_{\vec{X}}[(\text{Bias}[G(\boldsymbol{x}_*)])^2] + \mathbb{E}_{\vec{X}}[\text{Var}[G(\boldsymbol{x}_*)]] && (9)\end{aligned}$$

Equation 9 is the **general version of the Bias-Variance Decomposition**.

How can we reduce the values of the last two terms in Equation 9? We can make the bias 0 by making $\mathcal{H}$ complex, effectively overfitting. Hence, bias would decrease, but the variance term will in increase.[1] Consider the **model average**:

$$g_{avg} := \frac{\sum_{m=1}^{M} g_m}{M}$$

This can be smart if all M models are known to be good, since this can reduce variation. On the down side, we give up interpretability (suppose, for example, they were linear models). Let's examine the bias-variance decomposition for g_{avg} (henceforth we use G_{avg}, since g_{avg} is realized from it):

$$\begin{aligned}MSE :&= \sigma^2 + \mathbb{E}[\text{Bias}[G_{avg}(\boldsymbol{x}_*)]^2] + \mathbb{E}[\text{Var}[G_{avg}(\boldsymbol{x}_*)]] \\ &= \sigma^2 + \mathbb{E}\left[\left(f - \frac{G_1 + \cdots + G_M}{M}\right)^2\right] + \mathbb{E}\left[\text{Var}\left[\frac{G_1 + \cdots + G_M}{M}\right]\right] && \text{(by 5)} \\ &= \sigma^2 + \mathbb{E}\left[\left(\frac{(f - G_1) + \cdots + (f - G_M)}{M}\right)^2\right] + \frac{1}{M^2} \cdot \mathbb{E}\left[\text{Var}\left[G_1 + \cdots + G_M\right]\right] \\ & && (10)\end{aligned}$$

In our first attempt to simplify this expression, we make the following assumptions:

(i) The bias of each model is the same: $\text{Bias}[G_1] = \cdots = \text{Bias}[G_M]$.

(ii) The models $G_1, \ldots, G_M$ are independent.

(iii) The variance of each model is the same: $\text{Var}[G_1] = \cdots = \text{Var}[G_M]$.

Then continuing from Equation 10:

$$\begin{aligned}MSE &= \sigma^2 + \mathbb{E}[(\text{Bias}[G_1])^2] + \mathbb{E}\left[\frac{1}{M} \cdot \text{Var}[G_1]\right] && \text{(by (i) and (iii))} \\ &\to \sigma^2 && \text{(let } M \to \infty \text{ and use } \mathcal{H} \text{ to overfit)}\end{aligned}$$

What do we make of this? Assumption (ii) said that we have independent models, which is fake! Moreover, in the real world, we have one data set $\mathbb{D}$, and one g, so the g's cannot be independent. However the fact that we eliminated bias by overfitting is reasonable (we can use trees with small N_0). We need to come up with a reasonable way to reduce the variance term that does not use assumption (ii).

[1] Recall setting (I) of our model selection procedure, where we have M pre-specified modeling procedures and we choose the optimal m_*? This may be a good place to apply it.

Review of Correlation and Covariance

Recall from introductory probability that the **covariance** of a pair of random variables X_i and X_j is given by

$$\sigma_{ij} := \mathrm{Cov}[X_i, X_j] := \mathbb{E}[X_i \cdot X_j] - \mu_i \mu_j$$

where $\mu_i = \mathbb{E}[X_j]$. Meanwhile, the **correlation** of random variables X_i and X_j is given by

$$\rho_{ij} := \mathrm{Corr}[X_i, X_j] := \frac{\mathrm{Cov}[X_i, X_j]}{\mathrm{SD}[X_i] \cdot \mathrm{SD}[X_j]}$$

where SD stands for standard deviation. The covariance is a measure of how much X_i and X_j change together, yielding a positive value if both tend to increase together, and negative if both tend to decrease together. The correlation ρ_{ij} is a normalized version of covariance, with $\rho \in [-1, 1]$.

Let $X_1, \ldots, X_n$ be random variables. Assume σ_{ij} is the same for all ordered pairs $i \neq j$ (of which there are $n^2 - n$), and that $\sigma^2 = \sigma_1^2 = \cdots = \sigma_n^2$. Then

$$\rho = \frac{\sigma_{ij}}{\sigma^2} \tag{11}$$

Define the mean of the given random variables

$$\overline{X} := \frac{\sum_{i=1}^n X_i}{n}$$

Then

$$\begin{aligned}
\mathrm{Var}[\overline{X}] &= \frac{1}{n^2}\mathrm{Var}\left[\sum_{i=1}^n X_i\right] \\
&= \frac{1}{n^2}\left(\sum_{i=1}^n \mathrm{Var}[X_i] + \sum_{i \neq j} \mathrm{Cov}[X_i, X_j]\right) && \text{(derived in intro probability)} \\
&= \frac{1}{n^2}\left(n\sigma^2 + (n^2 - n)\sigma_{ij}\right) && \text{(assuming equal } \sigma_{ij}\text{)} \\
&= \frac{1}{n}\left(\sigma^2 + (n-1)\sigma^2\rho\right) && \text{(by 11)} \\
&= \sigma^2\rho + \frac{\sigma^2(1-\rho)}{n}
\end{aligned}$$

In summary, if σ_{ij} is the same for all X_i, X_j, and their variances are all equal, we have

$$\mathrm{Var}[\overline{X}] = \sigma^2\rho + \frac{\sigma^2(1-\rho)}{n} \tag{12}$$

Dependent Models

Continuing to assume equal bias for models $G_1, \ldots, G_M$, that the variance of each model is the same, and replacing assumption (ii) with the assumption that the covariance of each pair of models is the same:

$$MSE := \sigma^2 + \mathbb{E}[\mathrm{Bias}[G_{avg}(\boldsymbol{x}_*)]^2] + \mathbb{E}[\mathrm{Var}[G_{avg}(\boldsymbol{x}_*)]]$$

$$\begin{aligned}
&= \sigma^2 + \mathbb{E}[\mathrm{Bias}[G_1(\boldsymbol{x}_*)]^2] + \mathbb{E}\left[\mathrm{Var}[G_1(\boldsymbol{x}_*)] \cdot \rho + \frac{\mathrm{Var}[G_1] \cdot (1-\rho)}{M}\right] && \text{(by 12)} \\
&\to \sigma^2 + \mathbb{E}[\mathrm{Bias}[G_1(\boldsymbol{x}_*)]^2] + \mathbb{E}[\rho \cdot \mathrm{Var}[G_1]] && \text{(let } M \to \infty\text{)} \\
&\to \sigma^2 + \rho \cdot \mathbb{E}[\mathrm{Var}[G_1]] && \text{(let } \mathcal{A} \text{ overfit, bias becomes 0)} \\
&< \sigma^2 + \mathbb{E}[\mathrm{Var}[G_1]] && \text{(since } \rho \leq 1\text{, plus assume } \rho > 0\text{)}
\end{aligned}$$

The last equation is the MSE for *one* overfit model, whereas the equation that precedes it is the MSE for M overfit models. We cannot make the models independent, but can we make them dependent such that their dependence is nontrivially less than 1? This would help reduce the error in the variance term.

Bootstrap Aggregation

In 1994, **Bootstrap Aggregation (Bagging)** was invented by Breiman. How do we make ρ smaller than, say, 0.9? Even better, smaller than 0.2, which would be close to the theoretical minimum for the MSE (which is σ^2)?

Let's sample the results of $\mathbb{D}$ a total of n times with replacement where n is the number of data points in $\mathbb{D}$. The probability that a given row is chosen on any given trial is $\frac{1}{n}$, and hence the probability that it is omitted is $\frac{n-1}{n} = 1 - \frac{1}{n}$. Thus, the probability that a given row is omitted in all n tries is

$$\left(1 - \frac{1}{n}\right)^n \approx \frac{1}{e} \approx \frac{1}{3}$$

This is called a **bootstrap sample**. Let's take M bootstrap samples. With this we can compute M overfit models

$$g_1 = \mathcal{A}(\mathbb{D}_1, \mathcal{H}), \ldots, g_M = \mathcal{A}(\mathbb{D}_M, \mathcal{H})$$

Then we use the average of them, g_{avg}. Since all g's are built from the same $\mathbb{D}$, they will be a little correlated, and yet they will all be a little different because they are built from a different $\mathbb{D}_m$. Thus ρ will not be zero and will be a little less than 1. Typically, we bag trees with small N_0 (for example $N_0 = 1$), and we let M be as large as practically allowable.