Homework 1

Abhimanyu Talwar

October 6, 2018

ESL Problem 2.2
I assume that we are given the 10 means for each of the classes - sampled from $\mathcal{N}((1,0)^T, \mathbf{I})$ for BLUE (call these $m_1^B, \cdots, m_{10}^B$), and from $\mathcal{N}((0,1)^T, \mathbf{I})$ for ORANGE (call these $m_1^O, \cdots, m_{10}^O$). The population data points are sampled *given* these 10 means for each class. I also assume that an equal number of data points is sampled from each class, and so $\mathbb{P}(Y = B) = \mathbb{P}(Y = O)$, where Y is a random variable which denotes the class label (B for Blue and O for Orange) of a data point.
Let X be a two dimensional random vector denoting a data point. Then at the Bayes' boundary, we have:

$$\mathbb{P}(Y = B|X = x) = \mathbb{P}(Y = O|X = x) \tag{1}$$

Let $f_{X|Y}(x)$ denote the conditional probability density of the data point X given the color label Y. Then using Bayes' Theorem (and our assumption that $\mathbb{P}(Y = B) = \mathbb{P}(Y = O)$), we can write:

$$\frac{f_{X|Y=B}(x)\mathbb{P}(Y = B)}{f_X(x)} = \frac{f_{X|Y=O}(x)\mathbb{P}(Y = O)}{f_X(x)} \tag{2}$$

$$\implies f_{X|Y=B}(x) = f_{X|Y=O}(x) \tag{3}$$

$$\implies \frac{1}{10}\sum_{i=1}^{10} f_{m_i^B}(x) = \frac{1}{10}\sum_{i=1}^{10} f_{m_i^O}(x) \tag{4}$$

The solutions to Eq. 4 define the optimal Bayes' decision boundary. Here $f_{m_i^O}(x)$ and $f_{m_i^B}(x)$ are probability density functions of the Gaussian distributions $\mathcal{N}(m_i^O, \mathbf{I}/5)$ and $\mathcal{N}(m_i^B, \mathbf{I}/5)$ respectively.

ESL Problem 2.3
Let the N p-dimensional points be represented by the random variables $X_1, \cdots, X_n$, where each random variable X_i is drawn independently from a p-dimensional uniform distribution.

Let $\|X_i\|$ denote the Euclidean distance of a point X_i from the origin. I define the random variable representing distance of the closest point (among the N points) to the origin as:

$$D_{closest} = min\left(\|X_1\|, \|X_2\|, \cdots, \|X_{N-1}\|, \|X_N\|\right) \tag{5}$$

Let r be the median distance of the closest point from the origin. Based on the definition of 'median', we can say:

$$P\left(D_{closest} \geq r\right) = \frac{1}{2} \tag{6}$$

In the situation where the minimum of the N random variables $\|X_1\|, \cdots, \|X_N\|$ is to exceed a number r, each of the N random variables would individually exceed r. This implies that:

$$P\left(D_{closest} \geq r\right) = P\left(\|X_1\| \geq r \cap \|X_2\| \geq r \cap \cdots \|X_N\| \geq r\right) \tag{7}$$

And since the N points are independently drawn, we can say:

$$P\left(D_{closest} \geq r\right) = P\left(\|X_1\| \geq r\right) \times P\left(\|X_2\| \geq r\right) \times \cdots \times P\left(\|X_N\| \geq r\right) \tag{8}$$

Now let $Vol(B_r^p(0))$ denote the volume of a ball of radius r in p-dimensional space. As discussed in the lecture on 17 September, this volume is r^p times some function of p. Then for random variable X_i, the probability that this variable lies outside the ball of radius r centered at origin is given by:

$$P\left(\|X_i\| \geq r\right) = 1 - \frac{Vol(B_r^p(0))}{Vol(B_1^p(0))} \tag{9}$$

$$= 1 - r^p \tag{10}$$

Plugging Eq. 10 in Eq. 8, we get:

$$P\left(D_{closest} \geq r\right) = (1 - r^p)^N \tag{11}$$

And plugging Eq. 11 in Eq. 6, we have:

$$(1 - r^p)^N = \frac{1}{2} \tag{12}$$

$$\implies r = \left(1 - \left(\frac{1}{2}\right)^{\frac{1}{N}}\right)^{\frac{1}{p}} \tag{13}$$

ESL Problem 2.4

Let $(a_1, a_2, \cdots, a_p)$ denote the p components of the unit vector a. Since a is a unit vector, we have:

$$\sum_{j=1}^{p} a_j^2 = 1 \tag{14}$$

As given in the problem statement, let x be a vector drawn from the Spherical Multinormal Distribution $\mathcal{N}(0, \mathbb{I}_p)$, then we define z as:

$$z = a^T x \tag{15}$$

Now my approach for this problem is to show that z has the same Characteristic Function as that of a standard Gaussian random variable $\mathcal{N}(0,1)$. Once I have proved that, we can conclude that z also has a standard Gaussian distribution.

The Characteristic Function for a random variable X having the distribution $\mathcal{N}(0,1)$ can be written as:

$$\varphi_X(t) = e^{-\frac{t^2}{2}} \tag{16}$$

Now I will find the Characteristic Function for the random variable z:

$$\varphi_z(t) = \mathbb{E}\left[e^{itz}\right] \tag{17}$$

$$= \mathbb{E}\left[e^{it\sum_{j=1}^{p} a_j x_j}\right] \tag{18}$$

$$= \mathbb{E}\left[\prod_{j=1}^{p} e^{ita_j x_j}\right] \tag{19}$$

Since x has been drawn from a Spherical Multinormal Distribution, each of its p components, $(x_1, x_2, \cdots, x_p)$, are independent and identically distributed with the standard Gaussian distribution $\mathcal{N}(0,1)$. Therefore, the functions of these p components, $(e^{ita_1x_1}, e^{ita_2x_2}, \cdots, e^{ita_px_p})$ are also independent, for a given a. Then, in Eq. 19, we can write the expectation-of-product as a product-of-expectations. We get:

$$\varphi_z(t) = \prod_{j=1}^{p} \mathbb{E}\left[e^{ita_j x_j}\right] \tag{20}$$

Since each $x_j \sim \mathcal{N}(0,1)$, the j^{th} term inside the product on the right hand side of Eq. 20, represents the Characteristic Function of a standard Gaussian (stated in Eq.16), evaluated at the point ta_j. So we can rewrite Eq. 20 as:

$$\varphi_z(t) = \prod_{j=1}^{p} \varphi_{x_i}(ta_j) \tag{21}$$

$$= \prod_{j=1}^{p} e^{-\frac{t^2 a_j^2}{2}} \tag{22}$$

$$= e^{-\frac{t^2}{2}\sum_{j=1}^{p} a_j^2} \tag{23}$$

$$= e^{-\frac{t^2}{2}} \tag{24}$$

In Eq. 24 above I have used the result from Eq. 14. We can observe from Eq. 24 that z has the same Characteristic Function as that of a standard Gaussian random variable. Hence we conclude that $z \sim \mathcal{N}(0,1)$.

The squared distance for the projection of x onto unit vector a is simply given by z^2. Using the definition of Variance, we have:

$$\begin{aligned} Var(z) &= \mathbb{E}\left[z^2\right] - (\mathbb{E}\left[z\right])^2 && (25) \\ \implies 1 &= \mathbb{E}\left[z^2\right] - 0 && (26) \\ \implies \mathbb{E}\left[z^2\right] &= 1 && (27) \end{aligned}$$

Eq. 27 proves that the expected squared distance of the project of x onto the unit vector a is 1.

ESL Problem 2.5

1. **ESL 2.5 (a): Prove Equation 2.27 from the book** In this problem, we assume that the relationship between X and Y is linear and that it is determined by the following equation, where $\epsilon \sim \mathbb{N}(0, \sigma^2)$.

$$Y = X^T\beta + \epsilon \tag{28}$$

Say we have some training data $\mathbb{T} = (\mathbf{X}, \mathbf{Y})$, and we fit our linear model by least squares to this training data. Let x_0 denote a test datapoint, and let y_0 denote the response variable (given an x_0, the variable y_0 is random due to the presence of noise ϵ). Then the Expected Prediction Error for a prediction made at point x_0 is given by:

$$\begin{aligned} EPE(x_0) &= \mathbb{E}_{y_0|x_0}\mathbb{E}_{\mathbb{T}}\left[(y_0 - \hat{y}_0)^2\right] && (29) \\ &= \mathbb{E}_{y_0|x_0}\mathbb{E}_{\mathbb{T}}\left[(y_0 - x_0^T\beta + x_0^T\beta - \hat{y}_0)^2\right] && (30) \\ &= \mathbb{E}_{y_0|x_0}\mathbb{E}_{\mathbb{T}}\left[(y_0 - x_0^T\beta)^2\right] + \mathbb{E}_{y_0|x_0}\mathbb{E}_{\mathbb{T}}\left[(x_0^T\beta - \hat{y}_0)^2\right] + 2\mathbb{E}_{y_0|x_0}\mathbb{E}_{\mathbb{T}}\left[(y_0 - x_0^T\beta)(x_0^T\beta - \hat{y}_0)\right] && (31) \end{aligned}$$

The right hand side of Eq. 31 consists of summation of three sub-expressions. Below, I will derive them one by one.

For the first sub-expression, using Eq. 28, we have:

$$\begin{aligned} \mathbb{E}_{y_0|x_0}\mathbb{E}_{\mathbb{T}}\left[(y_0 - x_0^T\beta)^2\right] &= \mathbb{E}_{y_0|x_0}\mathbb{E}_{\mathbb{T}}\left[\epsilon^2\right] && (32) \\ &= \sigma^2 && (33) \end{aligned}$$

For the third sub-expression, using Eq. 28, I will replace $(y_0 - x_0^T\beta)$ with ϵ. We then get:

$$\mathbb{E}_{y_0|x_0}\mathbb{E}_{\mathbb{T}}\left[(y_0 - x_0^T\beta)(x_0^T\beta - \hat{y}_0)\right] = \mathbb{E}_{y_0|x_0}\mathbb{E}_{\mathbb{T}}\left[\epsilon(x_0^T\beta - \hat{y}_0)\right] \tag{34}$$

Since ϵ is independent of $(x_0^T\beta - \hat{y}_0)$, we can write the expectation-of-product as a product-of-expectations.

$$\begin{aligned} \mathbb{E}_{y_0|x_0}\mathbb{E}_{\mathbb{T}}\left[(y_0 - x_0^T\beta)(x_0^T\beta - \hat{y}_0)\right] &= \mathbb{E}_{y_0|x_0}\mathbb{E}_{\mathbb{T}}\left[\epsilon\right] \times \mathbb{E}_{y_0|x_0}\mathbb{E}_{\mathbb{T}}\left[(x_0^T\beta - \hat{y}_0)\right] && (35) \\ &= 0 && (36) \end{aligned}$$

The last line follows because $\mathbb{E}\left[\epsilon\right] = 0$.

Finally, for the second sub-expression in Eq. 31, since this sub-expression is not influenced by y_0, we can get rid of the expectation with respect to $y_0|x_0$. We then have:

$$\begin{aligned} \mathbb{E}_{y_0|x_0}\mathbb{E}_{\mathbb{T}}\left[(x_0^T\beta - \hat{y}_0)^2\right] &= \mathbb{E}_{\mathbb{T}}\left[(x_0^T\beta - \hat{y}_0)^2\right] && (37) \\ &= \mathbb{E}_{\mathbb{T}}\left[(x_0^T\beta - \mathbb{E}_{\mathbb{T}}(\hat{y}_0) + \mathbb{E}_{\mathbb{T}}(\hat{y}_0) - \hat{y}_0)^2\right] && (38) \\ &= \mathbb{E}_{\mathbb{T}}\left[(x_0^T\beta - \mathbb{E}_{\mathbb{T}}(\hat{y}_0))^2\right] + \mathbb{E}_{\mathbb{T}}\left[(\mathbb{E}_{\mathbb{T}}(\hat{y}_0) - \hat{y}_0)^2\right] + 2\mathbb{E}_{\mathbb{T}}\left[(x_0^T\beta - \mathbb{E}_{\mathbb{T}}(\hat{y}_0))(\mathbb{E}_{\mathbb{T}}(\hat{y}_0) - \hat{y}_0)\right] && (39) \end{aligned}$$

Now I again have three sub-expressions on the right hand side of Eq. 39. Below, I derive each of them

(a) **Eq. 39 Sub-expression 1**
Since each of $x_0^T\beta$ and $\mathbb{E}_{\mathbb{T}}\left[\hat{y_0}\right]$ are non-random quantities, we can get rid of the expectation with respect to the training set. We then use the fact that for a linear model fitted via the least squares method, the expectation of prediction ($\mathbb{E}_{\mathbb{T}}\left[\hat{y_0}\right]$) is equal to the true mean of the response variable ($x_0^T\beta$). Or in other words, the Bias equals 0. So we get:

$$\begin{aligned} \mathbb{E}_{\mathbb{T}}\left[(x_0^T\beta - \mathbb{E}_{\mathbb{T}}(\hat{y}_0))^2\right] &= x_0^T\beta - \mathbb{E}_{\mathbb{T}}(\hat{y}_0))^2 && (40) \\ &= 0 && (41) \end{aligned}$$

(b) **Eq. 39 Sub-expression 2**
This sub-expression denotes the variance of $\hat{y_0}$. Using the facts that (1) $\mathbb{E}_{\mathbb{T}}[\hat{y_0}] = x_0^T\beta$, and (2) the prediction at x_0, that is $\hat{y_0}$, equals $x_0^T\hat{\beta}$ we get:

$$\mathbb{E}_{\mathbb{T}}\left[(\mathbb{E}_{\mathbb{T}}(\hat{y_0}) - \hat{y_0})^2\right] = Var(\hat{y_0}) \tag{42}$$
$$= Var(x_0^T\hat{\beta}) \tag{43}$$
$$= x_0^T Var(\hat{\beta})x_0^T \tag{44}$$

Now Eq. 3.8 of The Elements of Statistical Learning provides an expression for $Var(\hat{\beta})$ however it assumes that the $\mathbf{X}$ matrix (which consists of observed feature vectors), is non-random. On relaxing that conditioning on $\mathbf{X}$, we should get:

$$Var(\hat{\beta}) = \mathbb{E}_{\mathbb{T}}\left[(\mathbf{X}^T\mathbf{X})^{-1}\right]\sigma^2 \tag{45}$$

Using this result, we get:

$$\mathbb{E}_{\mathbb{T}}\left[(\mathbb{E}_{\mathbb{T}}(\hat{y_0}) - \hat{y_0})^2\right] = x_0^T\mathbb{E}_{\mathbb{T}}\left[(\mathbf{X}^T\mathbf{X})^{-1}\right]x_0\sigma^2 \tag{46}$$

(c) **Eq. 39 Sub-expression 3**
Using the fact that the prediction at x_0, that is $\hat{y_0}$, equals $x_0^T\hat{\beta}$, this sub-expression evaluates to 0.

Substituting the value of these three sub-expressions in Eq. 39, we get:

$$\mathbb{E}_{y_0|x_0}\mathbb{E}_{\mathbb{T}}\left[(x_0^T\beta - \hat{y_0})^2\right] = x_0^T\mathbb{E}_{\mathbb{T}}\left[(\mathbf{X}^T\mathbf{X})^{-1}\right]x_0\sigma^2 \tag{47}$$

Finally using Eq. 33 and Eq. 47 in Eq. 31, we get:

$$EPE(x_0) = \sigma^2 + x_0^T\mathbb{E}_{\mathbb{T}}\left[(\mathbf{X}^T\mathbf{X})^{-1}\right]x_0\sigma^2 \tag{48}$$

2. **ESL 2.5 (b): Prove Equation 2.28 from the book** First I'll argue that assuming $\mathbb{E}[X] = 0$, then $\mathbf{X}^T\mathbf{X} \to NCov(X)$ if N is large.
Consider the $(i,j)^{th}$ element of $\mathbf{X}^T\mathbf{X}$.

$$\left(\mathbf{X}^T\mathbf{X}\right)_{i,j} = \sum_{k=1}^{N}\mathbf{X}_{i,k}^T\mathbf{X}_{k,j} \tag{49}$$
$$= N\left(\frac{1}{N}\sum_{k=1}^{N}\mathbf{X}_{k,i}\mathbf{X}_{k,j}\right) \tag{50}$$
$$\to N\mathbb{E}\left[X_iX_j\right] \tag{51}$$
$$= N\sigma_{i,j} \tag{52}$$

The second last step follows from the Law of Large Numbers. The last step follows from the assumption that $\mathbb{E}[X_i] = \mathbb{E}[X_j] = 0$ and the Covariance formula, $Cov(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.

Using the result from Eq. 52, if N is large, then $\mathbf{X}^T\mathbf{X} \to NCov(X)$. Substituting this expression in Eq. 2.27 from the ESL book.

$$EPE(x_0) = \sigma^2 + x_0^TCov(X)^{-1}x_0\sigma^2/N \tag{53}$$
$$\implies \mathbb{E}_{x_0}EPE(x_0) = \sigma^2 + \mathbb{E}_{x_0}x_0^TCov(X)^{-1}x_0\sigma^2/N \tag{54}$$
$$\tag{55}$$

Now I will use a result from Linear Algebra for the expectation of a quadratic expression. Let B be a random vector of p dimensions, and let A be a $p \times p$ matrix which is non-random, then we have:

$$\mathbb{E}\left[B^TAB\right] = tr(ACov(B)) + \mathbb{E}[B]^T A\mathbb{E}[B] \tag{56}$$

Using this result, and the fact that $\mathbb{E}[x_0] = 0$, we get:

$$\mathbb{E}_{x_0}EPE(x_0) = \sigma^2 + tr\left(Cov(X)^{-1}Cov(x_0)\right)\sigma^2/N \tag{57}$$
$$= \sigma^2 + tr\left(\mathbf{I}\right)\sigma^2/N \tag{58}$$
$$= \sigma^2 + \sigma^2\left(\frac{p}{N}\right) \tag{59}$$

This proves Eq. 2.28 from ESL. The second last line follows from the fact that x_0 and X are essentially random vectors from the same distribution and therefore $Cov(X)$ is the same matrix as $Cov(x_0)$, and their product $Cov(X)^{-1}Cov(x_0)$ is the Identity matrix. The last line follows from the fact that the Trace of a $p \times p$ Identity matrix is simply p (because all diagonal elements are equal to 1).

ESL Problem 2.7

1. **ESL 2.7(a)**

   (a) **k-NN**
   Let $N_k(x_0)$ represent the k nearest neighbors of x_0. Then we define:

   $$l_i(x_0; \mathcal{X}) = \begin{cases} \frac{1}{k}, & \text{if } x_i \in N_k(x_0) \\ 0, & \text{otherwise} \end{cases} \tag{60}$$

   This then gives our estimator for f at x_0 as:

   $$\hat{f}(x_0) = \sum_{i=1}^{N} l_i(x_0; \mathcal{X}) y_i \tag{61}$$

   $$= \frac{1}{k} \sum_{x_i \in N_k(x_0)} y_i \tag{62}$$

   (b) **Linear Regression**
   In the case of Linear Regression, our estimate of f at x_0 can be written as:

   $$\hat{f}(x_0) = x_0^T \hat{\beta} \tag{63}$$

   The parameter estimate $\hat{\beta}$ is given by (here $\mathbf{X}$ represents the $N \times p$ matrix denoting the N observations of x, and p is the number of dimensions of x):

   $$\hat{\beta} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T y \tag{64}$$

   This gives the estimate of f at x_0 as:

   $$\hat{f}(x_0) = x_0^T \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T y \tag{65}$$

   Let w^T denote the $1 \times N$ vector $x_0^T \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T$ and let w_i^T denote the i^{th} element of the vector w^T. Then we define weights as:

   $$l_i(x_0; \mathcal{X}) = w_i^T \tag{66}$$

   Using these weights from Eq. 66, we can now write Eq. 65 in the desired form.

2. **ESL 2.7(b)**
   Adding and subtracting $\mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\hat{f}(x_0)\right]$ within the expression for conditional mean-squared error, we get:

$$\begin{aligned}
\mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[(f(x_0) - \hat{f}(x_0))^2\right] &= \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\left(f(x_0) - \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\hat{f}(x_0)\right] + \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\hat{f}(x_0)\right] - \hat{f}(x_0)\right)^2\right] \\
&= \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\left(f(x_0) - \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\hat{f}(x_0)\right]\right)^2\right] \\
&+ \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\left(\mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\hat{f}(x_0)\right] - \hat{f}(x_0)\right)^2\right] \\
&+ 2\mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\left(f(x_0) - \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\hat{f}(x_0)\right]\right)\left(\mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\hat{f}(x_0)\right] - \hat{f}(x_0)\right)\right]
\end{aligned} \tag{67}$$

   The right hand side of Eq. 67 has three sub-expressions, which I will discuss below. The first two expressions involve the term $\mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\hat{f}(x_0)\right]$, which I will derive first below:

$$\mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\hat{f}(x_0)\right] = \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\sum_{i=1}^{N} l_i(x_0,\mathcal{X})y_i\right] \tag{68}$$

$$= \sum_{i=1}^{N} l_i(x_0,\mathcal{X})\mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[y_i\right] \tag{69}$$

$$= \sum_{i=1}^{N} l_i(x_0,\mathcal{X})f(x_i) \tag{70}$$

$$\tag{71}$$

(a) **Eq. 67 First Sub-expression**
This expression represents the expected squared-distance between (1) the true value of f at x_0, and (2) the expected value of our estimator $\hat{f}$ at x_0. This corresponds to the definition of $Bias^2$ (conditioned on $\mathcal{X}$). We can get rid of the outer expectation since the terms inside are already non-random (conditioned on $\mathcal{X}$).

$$Bias^2(\hat{f}(x_0)|\mathcal{X}) = \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\left(f(x_0) - \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\hat{f}(x_0)\right]\right)^2\right] \tag{72}$$

$$= \left(f(x_0) - \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\hat{f}(x_0)\right]\right)^2 \tag{73}$$

(b) **Eq. 67 Second Sub-expression**
This expression represents the expected squared distance between (1) our estimator $\hat{f}$ evaluated at x_0, and (2) the expected value of the estimator $\hat{f}$ evaluated at x_0. This corresponds to the definition of Variance (conditioned on $\mathcal{X}$).

$$Var(\hat{f}(x_0)|\mathcal{X}) = \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\left(\mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\hat{f}(x_0)\right] - \hat{f}(x_0)\right)^2\right] \tag{74}$$

(c) **Eq. 67 Third Sub-expression**
I will prove below that this sub-expression evaluates to 0. This expression is an expectation of product of two terms. However the first of the two terms is $\left(f(x_0) - \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\hat{f}(x_0)\right]\right)$, which is not dependent on $\mathcal{Y}$ and is non-random conditioned on $\mathcal{X}$. So we will take it out of the expectation to get:

$$\begin{aligned} 2\mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\left(f(x_0) - \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\hat{f}(x_0)\right]\right)\left(\mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\hat{f}(x_0)\right] - \hat{f}(x_0)\right)\right] &= 2\left(f(x_0) - \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\hat{f}(x_0)\right]\right) \\ &\times \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\left(\mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\hat{f}(x_0)\right] - \hat{f}(x_0)\right)\right] \end{aligned} \tag{75}$$

The conditional expectation on the right hand side of Eq. 75 can simply be written as:

$$\mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\left(\mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\hat{f}(x_0)\right] - \hat{f}(x_0)\right)\right] = \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\hat{f}(x_0)\right] - \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\hat{f}(x_0)\right] \tag{76}$$

$$= 0 \tag{77}$$

This proves that the third sub-expression will evaluate to 0.

3. **ESL 2.7(c)**
Using the Tower Rule, we can write:

$$\mathbb{E}_{\mathcal{X},\mathcal{Y}}\left[(f(x_0) - \hat{f}(x_0))^2\right] = \mathbb{E}_{\mathcal{X}}\left[\mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[(f(x_0) - \hat{f}(x_0))^2\right]\right] \tag{78}$$

Using our decomposition of $\mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[(f(x_0) - \hat{f}(x_0))^2\right]$ from ESL 2.7(b) above, we can write:

$$\begin{aligned} \mathbb{E}_{\mathcal{X},\mathcal{Y}}\left[(f(x_0) - \hat{f}(x_0))^2\right] &= \mathbb{E}_{\mathcal{X}}\left[Bias^2(\hat{f}(x_0)|\mathcal{X}) + Var(\hat{f}(x_0)|\mathcal{X})\right] \\ &= \int Bias^2(\hat{f}(x_0)|\mathcal{X})\left(\prod_{i=1}^{N} h(x_i)\right) dx_1 \cdots dx_N \\ &+ \int Var(\hat{f}(x_0)|\mathcal{X})\left(\prod_{i=1}^{N} h(x_i)\right) dx_1 \cdots dx_N \end{aligned} \tag{79}$$

Here, $\left(\prod_{i=1}^{N} h(x_i)\right)$ equals the probability density for a particular instance of training set $\mathcal{X}$.

4. **ESL 2.7(d)**
   Eq. 79 relates the unconditional expected squared error to conditional (on $\mathcal{X}$) Bias and Variance.

ESL Problem 2.9
Note: For this problem I have used the sketch provided in CMU's problem set 2 for their Advanced Methods for Data Analysis course. The calculations are my own. (Click this link for the problem statement.)
I will first prove that the expected test error, $\mathbb{E}\left[R_{te}(\hat{\beta})\right]$ is the same irrespective of our choice of M (the number of test points). Here $\hat{\beta}$ denotes the estimate of our Linear Regression parameter fitted with lease squares on the *training* set. Let $\mathbb{E}[\bullet]$ denote expectation with respect to everything that is random. Then we have:

$$\begin{aligned}\mathbb{E}\left[R_{te}(\hat{\beta})\right] &= \mathbb{E}\left[\frac{1}{M}\sum_{i=1}^{M}(\tilde{y}_i - \hat{\beta}^T\tilde{x}_i)^2\right] \\ &= \frac{1}{M}\sum_{i=1}^{M}\mathbb{E}\left[(\tilde{y}_i - \hat{\beta}^T\tilde{x}_i)^2\right] \\ &= \frac{1}{M}\sum_{i=1}^{M}\left(\mathbb{E}\left[\tilde{y_i}^2\right] + \mathbb{E}\left[(\hat{\beta}^T\tilde{x}_i)^2\right] - 2\mathbb{E}\left[\tilde{y}_i\tilde{x_i}^T\hat{\beta}\right]\right)\end{aligned} \tag{80}$$

Since all datapoints of the Test set, $(\tilde{x}_i, \tilde{y}_i)$ are drawn from the same underlying population distribution, we can replace $\mathbb{E}\left[\tilde{y_i}^2\right]$ for all values of $i = 1, \cdots, M$ with simply $\mathbb{E}[y]^2$ (where y is drawn from the underlying population distribution of responses). Similarly, I will also replace $\mathbb{E}\left[\tilde{x_i}^2\right]$ and $\mathbb{E}\left[\tilde{y}_i\tilde{x_i}^T\right]$ with $\mathbb{E}\left[x^2\right]$ and $\mathbb{E}\left[yx^T\right]$, respectively. Further, since the Test and Training sets have been drawn independently at random, I will use the Tower Rule to decompose expectation of product (of terms involving the Training and Test sets) into product of expectations. I explain this below:

$$\mathbb{E}\left[(\hat{\beta}^T\tilde{x}_i)^2\right] = \mathbb{E}_{Train}\mathbb{E}_{Test|Train}(\hat{\beta}^T\tilde{x}_i)^2 \tag{81}$$

$$= \mathbb{E}_{Train}\hat{\beta}^T\mathbb{E}_{Test}\left[x^2\right] \tag{82}$$

$$= \mathbb{E}\left[\hat{\beta}\right]^T\mathbb{E}\left[x^2\right] \tag{83}$$

Similarly, we also have:

$$\mathbb{E}\left[\tilde{y}_i\tilde{x_i}^T\hat{\beta}\right] = \mathbb{E}[xy]^T\,\mathbb{E}\left[\hat{\beta}\right] \tag{84}$$

Using this results, we can simplify Eq. 80 to say:

$$\mathbb{E}\left[R_{te}(\hat{\beta})\right] = \mathbb{E}\left[y^2\right] + \mathbb{E}\left[\hat{\beta}\right]^T\mathbb{E}\left[x^2\right] + \mathbb{E}[xy]^T\,\mathbb{E}\left[\hat{\beta}\right] \tag{85}$$

As shown above, the expected test error $\mathbb{E}\left[R_{te}(\hat{\beta})\right]$ would be the same expression as in Eq. 85, irrespective of the value of M. This implies that we can equivalently define $\mathbb{E}\left[R_{te}(\hat{\beta})\right]$ in terms of N test points (same number of points as in the training set).

$$\mathbb{E}\left[R_{te}(\hat{\beta})\right] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(\tilde{y}_i - \hat{\beta}^T\tilde{x}_i)^2\right] \tag{86}$$

Now we know that the estimate of Linear Regression coefficient that would minimize the sum of squared residuals for the test set is not $\hat{\beta}$. Instead, it would be some estimate $\hat{\beta}_{Test}$ which would be derived by fitting using least squares on the *test* set instead of on the *training* set. It follows that:

$$\frac{1}{N}\sum_{i=1}^{N}(\tilde{y}_i - \hat{\beta}^T\tilde{x}_i)^2 \geq \frac{1}{N}\sum_{i=1}^{N}(\tilde{y}_i - \hat{\beta}_{Test}^T\tilde{x}_i)^2 \tag{87}$$

$$\implies \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(\tilde{y}_i - \hat{\beta}^T\tilde{x}_i)^2\right] \geq \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(\tilde{y}_i - \hat{\beta}_{Test}^T\tilde{x}_i)^2\right] \tag{88}$$

$$\implies \mathbb{E}\left[R_{te}(\hat{\beta})\right] \geq \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(\tilde{y}_i - \hat{\beta}_{Test}^T\tilde{x}_i)^2\right] \tag{89}$$

Now $(\tilde{x_1}, \tilde{y_1}), \cdots, (\tilde{x_N}, \tilde{y_N})$, is just an arbitrarily chosen set of data points from the underlying population, and $\hat{\beta}_{Test}$ is a Linear Regression parameter estimated by fitted using least squares to this arbitrarily drawn set of N data points, we can say that the two random variables below have the same distribution (and therefore the same expectation):

$$\frac{1}{N}\sum_{i=1}^{N}(\tilde{y_i} - \hat{\beta}_{Test}^T\tilde{x_i})^2 \sim \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\beta}^T x_i)^2 \tag{90}$$

$$\implies \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(\tilde{y_i} - \hat{\beta}_{Test}^T\tilde{x_i})^2\right] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\beta}^T x_i)^2\right] \tag{91}$$

$$\implies \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(\tilde{y_i} - \hat{\beta}_{Test}^T\tilde{x_i})^2\right] = \mathbb{E}\left[R_{tr}(\hat{\beta})\right] \tag{92}$$

Combining Eq. 89 and Eq. 92, we get the desired result:

$$\mathbb{E}\left[R_{tr}(\hat{\beta})\right] \leq \mathbb{E}\left[R_{te}(\hat{\beta})\right] \tag{93}$$