# Chapter 7

Abhimanyu Talwar

October 10, 2018

**ESL Problem 7.1**
Let $\omega$ denote the expected value of *optimism*, that is $\mathbb{E}_y[op]$. We have:

$$\omega = \mathbb{E}_y[Err_{in}] - \mathbb{E}_y[\overline{err}] \tag{1}$$

$$\begin{aligned} &= \mathbb{E}_y\left[\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{y^o}[L(y_i^o, \hat{y}_i)]\right] \\ &- \mathbb{E}_y\left[\frac{1}{N}\sum_{i=1}^{N}L(y_i, \hat{y}_i)\right] \end{aligned} \tag{2}$$

For squared error, we have:

$$\begin{aligned} \omega &= \mathbb{E}_y\left[\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{y^o}\left[(y_i^o - \hat{y}_i)^2\right]\right] \\ &- \mathbb{E}_y\left[\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2\right] \\ &= \mathbb{E}_y\left[\frac{1}{N}\sum_{i=1}^{N}\left(\mathbb{E}_{y^o}\left[(y_i^o)^2\right] + (\hat{y}_i)^2 - 2\mathbb{E}_{y^o}[y_i^o]\,\hat{y}_i\right)\right] \\ &- \mathbb{E}_y\left[\frac{1}{N}\sum_{i=1}^{N}\left(y_i^2 + \hat{y_i}^2 - 2y_i\hat{y}_i\right)\right] \end{aligned} \tag{3}$$

Notice in Eq. 3 that the two terms $\mathbb{E}_{y^o}\left[(y_i^o)^2\right]$ and $\mathbb{E}_{y^o}[y_i^o]$ are already expectations with respect to $y$ (with the training set of features $\mathbb{X}$ held fixed) and will not change when we once again take the expectation $\mathbb{E}_y[\bullet]$ (because this expectation also assumed $\mathbb{X}$ held fixed). So I will simply replace them with $\mathbb{E}_y\left[(y_i)^2\right]$ and $\mathbb{E}_y[y_i]$ respectively. After making this substitution in Eq. 3 and canceling terms, we have:

$$\begin{aligned} \omega &= \frac{-2}{N}\sum_{i=1}^{N}\mathbb{E}_y[y_i]\,\mathbb{E}_y[\hat{y}_i] + \sum_{i=1}^{N}\frac{2}{N}\mathbb{E}_y[y_i\hat{y}_i] \\ &= \frac{2}{N}\sum_{i=1}^{N}Cov(\hat{y}_i, y_i) \end{aligned} \tag{4}$$

From Eq. 1 and Eq. 4, we have:

$$\mathbb{E}_y[Err_{in}] = \mathbb{E}_y[\overline{err}] + \frac{2}{N}\sum_{i=1}^{N}Cov(\hat{y}_i, y_i) \tag{5}$$

Now we assume that the underlying model has addition noise, i.e. $Y = f(X) + \epsilon$ where $\epsilon \sim \mathbb{N}(0, \sigma^2)$. We further assume that we have fitted a linear prediction function (with $d$ predictor variables) using least squares.
**Claim:** Under the assumptions stated above, we have:

$$\sum_{i=1}^{N}Cov(\hat{y}_i, y_i) = d\sigma^2 \tag{6}$$

**Proof:** We have:

$$\sum_{i=1}^{N} Cov(\hat{y}_i, y_i) = \sum_{i=1}^{N} \left(\mathbb{E}_y\left[y_i\hat{y}_i\right] - \mathbb{E}_y\left[y_i\right]\mathbb{E}_y\left[\hat{y}_i\right]\right) \tag{7}$$

$$= \sum_{i=1}^{N} \left(\mathbb{E}_y\left[(f(x_i)+\epsilon)\hat{y}_i\right] - \mathbb{E}_y\left[(f(x_i)+\epsilon)\right]\mathbb{E}_y\left[\hat{y}_i\right]\right) \tag{8}$$

$$= \sum_{i=1}^{N} \mathbb{E}_y\left[\epsilon\hat{y}_i\right] \tag{9}$$

$$= \sum_{i=1}^{N} x_i^T(X^TX)^{-1}X^T\mathbb{E}_y\left[\epsilon y\right] \tag{10}$$

Now since the truth value $y_i = f(x_i)+\epsilon$, the $N \times 1$ vector $\mathbb{E}_y\left[\epsilon y\right]$ has each of its entry equal to $\sigma^2$. Let me define $\epsilon_N$ as an $N \times 1$ random vector, each of whose entry is the random noise $\epsilon$. Then converting the summation in Eq. 10 into a matrix product, and taking terms related to $X$ inside the expectation (because the expectation is only over $y$), we get:

$$\sum_{i=1}^{N} Cov(\hat{y}_i, y_i) = \mathbb{E}_y\left[\epsilon_N^T X(X^TX)^{-1}X^T\epsilon_N\right] \tag{11}$$

Now I will apply the Linear Algebra identity, $\mathbb{E}\left[B^TAB\right] = trace(ACov(B)) + \mathbb{E}\left[B\right]^T A\mathbb{E}\left[B\right]$, to Eq. 11 to get:

$$\sum_{i=1}^{N} Cov(\hat{y}_i, y_i) = tr((X^TX)^{-1}Cov(X^T\epsilon_N)) + \mathbb{E}_y\left[\epsilon_N^TX\right](X^TX)^{-1}\mathbb{E}_y\left[X^T\epsilon_N\right] \tag{12}$$

$$= trace((X^TX)^{-1}(X^TX)\sigma^2) + 0 \tag{13}$$

$$= d\sigma^2 \tag{14}$$

The last step follows from the fact that the *trace* of an identity matrix of dimension $d \times d$ is simply $d$. Using the result from Eq. 14 in Eq. 5, we get the desired result.

$$\boxed{\mathbb{E}_y\left[Err_{in}\right] = \mathbb{E}_y\left[\overline{err}\right] + 2\frac{d}{N}\sigma^2} \tag{15}$$

**ESL Problem 7.5**

For this problem, I will use similar calculations as I used to prove Eq. 6 above. Assume $y$ arises from the additive-noise model, that is $y = f(x)+\epsilon$ (where $\epsilon \sim N(0,\sigma^2)$), we have:

$$\sum_{i=1}^{N} Cov(\hat{y}_i, y_i) = \sum_{i=1}^{N} \left(\mathbb{E}_y\left[y_i\hat{y}_i\right] - \mathbb{E}_y\left[y_i\right]\mathbb{E}_y\left[\hat{y}_i\right]\right) \tag{16}$$

$$= \sum_{i=1}^{N} \left(\mathbb{E}_y\left[(f(x_i)+\epsilon)\hat{y}_i\right] - \mathbb{E}_y\left[(f(x_i)+\epsilon)\right]\mathbb{E}_y\left[\hat{y}_i\right]\right) \tag{17}$$

$$= \sum_{i=1}^{N} \mathbb{E}_y\left[\epsilon\hat{y}_i\right] \tag{18}$$

$$\tag{19}$$

Let me define $\epsilon_N$ as an $N \times 1$ random vector, each of whose entry is the random noise $\epsilon$.

$$\sum_{i=1}^{N} Cov(\hat{y}_i, y_i) = \mathbb{E}_y\left[\epsilon_N^TSy\right] \tag{20}$$

$$= \mathbb{E}_y\left[\epsilon_N^TS(f(\mathbf{X})+\epsilon_N)\right] \tag{21}$$

$$= \mathbb{E}_y\left[\epsilon_N^TSf(\mathbf{X})\right] + \mathbb{E}_y\left[\epsilon_N^TS\epsilon_N\right] \tag{22}$$

$$= 0 + \mathbb{E}_y\left[\epsilon_N^TS\epsilon_N\right] \tag{23}$$

$$\tag{24}$$

Now I will apply the Linear Algebra identity, $\mathbb{E}\left[B^T AB\right] = trace(ACov(B)) + \mathbb{E}\left[B\right]^T A\mathbb{E}\left[B\right]$. We then have:

$$\begin{aligned}\sum_{i=1}^{N} Cov(\hat{y}_i, y_i) &= trace\left(SCov(\epsilon_N)\right) + \mathbb{E}_y\left[\epsilon_N^T\right] S\mathbb{E}_y\left[\epsilon_N\right] \qquad (25)\\ &= trace(S)\sigma^2 \qquad (26)\end{aligned}$$

**ESL Problem 7.2**

$$\begin{aligned}Err(x_0) &= \mathbb{E}\left[\mathbb{I}(Y \neq \hat{G}(x_0)|X = x_0)\right] \qquad (27)\\ &= P(Y \neq \hat{G}(x_0)|X = x_0) \qquad (28)\end{aligned}$$

Now for this problem, we are *given* $x_0$, and so two cases are possible:

1. **Case 1:** $f(x_0) > 1/2$
   In this case, $G(x_0) = 1$. We have:

$$\begin{aligned}P(Y \neq \hat{G}(x_0)|X = x_0) &= P(Y = 1, \hat{G}(x_0) = 0|X = x_0)\\ &+ P(Y = 0, \hat{G}(x_0) = 1|X = x_0)\\ &= f(x_0)P(\hat{G}(x_0) \neq G(x_0)|X = x_0)\\ &+ P(Y \neq G(x_0)|X = x_0)(1 - P(\hat{G}(x_0) \neq G(x_0)|X = x_0))\\ &= f(x_0)P(\hat{G}(x_0) \neq G(x_0)|X = x_0)\\ &- (1 - f(x_0))P(\hat{G}(x_0) \neq G(x_0)|X = x_0)\\ &+ P(Y \neq G(x_0)|X = x_0)\\ &= (2f(x_0) - 1)P(\hat{G}(x_0) \neq G(x_0)|X = x_0) + P(Y \neq G(x_0)|X = x_0)\end{aligned} \qquad (29)$$

2. **Case 2:** $f(x_0) \leq 1/2$
   In this case, $G(x_0) = 0$. Similar to calculations done for Case 1 above, we can derive:

$$P(Y \neq \hat{G}(x_0)|X = x_0) = (1 - 2f(x_0))P(\hat{G}(x_0) \neq G(x_0)|X = x_0) + P(Y \neq G(x_0)|X = x_0) \qquad (30)$$

   Combining Eq. 29 and Eq. 30, and writing $P(Y \neq G(x_0)|X = x_0)$ as $Err_B(x_0)$, we prove the desired result:

$$\boxed{Err(x_0) = |2f(x_0) - 1|P(\hat{G}(x_0) \neq G(x_0)|X = x_0) + Err_B(x_0)} \qquad (31)$$

   For the second part of the question, we are given $\hat{f}(x_0) \sim \mathbb{N}\left(\mathbb{E}\left[\hat{f}(x_0)\right]), Var(\hat{f}(x_0))\right)$. If we standardize the random variable $\hat{f}(x_0)$, we then have:

$$\frac{\left(\hat{f}(x_0) - \mathbb{E}\left[\hat{f}(x_0)\right]\right)}{\sqrt{Var(\hat{f}(x_0))}} \sim \mathbb{N}(0, 1) \qquad (32)$$

   Again we are *given* $x_0$ and therefore two cases are possible:

   (a) **Case 1:** $f(x_0) > 1/2$

$$\begin{aligned}P(G(x_0) \neq \hat{G}(x_0)|X = x_0) &= P(\hat{f}(x_0) < 1/2) \qquad (33)\\ &= P\left(\frac{\left(\hat{f}(x_0) - \mathbb{E}\left[\hat{f}(x_0)\right]\right)}{\sqrt{Var(\hat{f}(x_0))}} < \frac{\left(1/2 - \mathbb{E}\left[\hat{f}(x_0)\right]\right)}{\sqrt{Var(\hat{f}(x_0))}}\right) \qquad (34)\\ &= \Phi\left(\frac{\left(1/2 - \mathbb{E}\left[\hat{f}(x_0)\right]\right)}{\sqrt{Var(\hat{f}(x_0))}}\right) \qquad (35)\end{aligned}$$

(b) **Case 2:** $f(x_0) \leq 1/2$
Similar to calculations above for Case 1, we can write:

$$P(G(x_0) \neq \hat{G}(x_0)|X = x_0) = \Phi\left(\frac{\left(\mathbb{E}\left[\hat{f}(x_0)\right] - 1/2\right)}{\sqrt{Var(\hat{f}(x_0))}}\right) \tag{36}$$

Combining Eq. 35 and Eq. 36, we can get the desired result:

$$\boxed{P(G(x_0) \neq \hat{G}(x_0)|X = x_0) = \Phi\left(\frac{sign(1/2 - f(x_0))\left(\mathbb{E}\left[\hat{f}(x_0)\right] - 1/2\right)}{\sqrt{Var(\hat{f}(x_0))}}\right)} \tag{37}$$

**ESL Problem 7.4**
I have already proved this result as part of my proof for Problem 7.1 above, and I will restate it here. Expected optimism $\omega$ can be written as:

$$\begin{aligned}\omega &= \mathbb{E}_y\left[\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{y^o}\left[(y_i^o - \hat{y}_i)^2\right]\right] \\ &- \mathbb{E}_y\left[\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2\right] \\ &= \mathbb{E}_y\left[\frac{1}{N}\sum_{i=1}^{N}\left(\mathbb{E}_{y^o}\left[(y_i^o)^2\right] + (\hat{y}_i)^2 - 2\mathbb{E}_{y^o}\left[y_i^o\right]\hat{y}_i\right)\right] \\ &- \mathbb{E}_y\left[\frac{1}{N}\sum_{i=1}^{N}\left(y_i^2 + \hat{y_i}^2 - 2y_i\hat{y}_i\right)\right]\end{aligned} \tag{38}$$

Since each of the expectations, $\mathbb{E}_{y_o}[\bullet]$ and $\mathbb{E}_y[\bullet]$, are taken assuming the training feature set $\mathbb{X}$ is held fixed, we can replace $\mathbb{E}_{y^o}\left[(y_i^o)^2\right]$ and $\mathbb{E}_{y^o}[y_i^o]$ with $\mathbb{E}_y\left[(y_i)^2\right]$ and $\mathbb{E}_y[y_i]$ respectively. After making this substitution and canceling terms in Eq. 38, we get:

$$\begin{aligned}\omega &= \frac{-2}{N}\sum_{i=1}^{N}\mathbb{E}_y[y_i]\,\mathbb{E}_y[\hat{y}_i] + \sum_{i=1}^{N}\frac{2}{N}\mathbb{E}_y[y_i\hat{y}_i] \\ &= \frac{2}{N}\sum_{i=1}^{N}Cov(\hat{y}_i, y_i)\end{aligned} \tag{39}$$

**ESL Problem 7.6**
We can express k-NN regression as a Linear Smoother of the form $\hat{y} = Sy$. Let $(\mathbf{X}, \mathbf{y})$ be our training set and let $(X_i, y_i)$ represent the $i^{th}$ datapoint out of a total $N$ training datapoints. Let $S_{ij}$ represent the element of $S$ in row $i$ and column $j$. We define:

$$S_{ij} = \begin{cases} 1/k, & \text{if } X_j \in N_k(X_i) \\ 0, & \text{otherwise} \end{cases} \tag{40}$$

Here $N_k(X_i)$ represents the set of k Nearest Neighbors of $X_i$. Notice that $S_{ii} = 1/k$ for all $i = 1, \cdots, N$. This is because a datapoint will always lie in the set of its own k Nearest Neighbors. Since the effective degrees of freedom is simply equal to $trace(S)$ (by definition, Eq. 7.32 in *The Elements of Statistical Learning*), we have:

$$df(S) = trace(S) \tag{41}$$

$$= \sum_{i=1}^{N} S_{ii} \tag{42}$$

$$= \sum_{i=1}^{N} 1/k \tag{43}$$

$$= \frac{N}{k} \tag{44}$$

**ESL Problem 7.8**
I will make use of the fact that $sign(sin(\pi x)) = (-1)^{\lfloor x \rfloor}$, where $\lfloor \bullet \rfloor$ represents the Floor operator. For a given $l$ and some configuration of binary labels over those $l$ points $z_1, \cdots, z_l$, let $K_0$ denote the set such that for any $k \in K_0$, where $k \in [1, 2, \cdots, l]$, and the point $z_k = 10^{-k}$ is assigned the label 0 in this configuration. Let me define:

$$\alpha = \pi \sum_{k \in K_0} 10^k \tag{45}$$

**Claim:** The function $\mathbb{I}(sin(\alpha x) > 0))$ will shatter the given configuration of $l$ points.
**Proof:** In the given label configuration, a point $z_p$ can take one of two labels:

1. **Label of $z_p$ is** 0: In this case, $p \in K_0$, and we have:

$$\begin{aligned}
sin(\alpha z_p) &= sin\left(\pi z_p \sum_{k \in K_0} 10^k\right) && (46)\\
&= sin\left(\pi 10^{-p} \sum_{k \in K_0} 10^k\right) && (47)\\
&= sin\left(\pi 10^{-p}\left(10^p + \sum_{k \in K_0, k<p} 10^k + \sum_{k \in K_0, k>p} 10^k\right)\right) && (48)\\
&= sin\left(\pi\left(1 + r + 10m\right)\right) && (49)
\end{aligned}$$

   Where $r < 1$ and $m$ is a positive integer. Now we have:

$$\begin{aligned}
sign(sin(\alpha z_p)) &= (-1)^{\lfloor 1+r+10m \rfloor} && (50)\\
&= (-1)^1 && (51)\\
&= -1 && (52)
\end{aligned}$$

   The label predicted by $\mathbb{I}(sin(\alpha z_p) > 0))$ is 0. Hence, for points which are assigned label 0, our function is able to correctly classify those points.

2. **Label of $z_p$ is** 0: In this case, $p \notin K_0$ and so we have:

$$\begin{aligned}
sin(\alpha z_p) &= sin\left(\pi z_p \sum_{k \in K_0} 10^k\right) && (53)\\
&= sin\left(\pi 10^{-p} \sum_{k \in K_0} 10^k\right) && (54)\\
&= sin\left(\pi 10^{-p}\left(\sum_{k \in K_0, k<p} 10^k + \sum_{k \in K_0, k>p} 10^k\right)\right) && (55)\\
&= sin\left(\pi\left(r + 10m\right)\right) && (56)
\end{aligned}$$

   Where $r < 1$ and $m$ is a positive integer. So we have:

$$\begin{aligned}
sign(sin(\alpha z_p)) &= (-1)^{\lfloor r+10m \rfloor} && (57)\\
&= (-1)^0 && (58)\\
&= +1 && (59)
\end{aligned}$$

   The label predicted by $\mathbb{I}(sin(\alpha z_p) > 0))$ is 1.

**Hence, in both cases our function is correctly able to classify, and hence shatter, these points. As $l$ was chosen arbitrarily, we conclude that the VC Dimension is infinity.**