Overfitting

B1. Suppose we have N labeled samples $S = \{(x_i, y_i)\}_{i=1}^N$ drawn i.i.d. from an underlying distribution $\mathcal{D}$. Suppose we decide to break this set into a set S_{train} of size N_{train} and a set S_{test} of size N_{test} samples for our training and test set, so $N = N_{\text{train}} + N_{\text{test}}$, and $S = S_{\text{train}} \cup S_{\text{test}}$. Recall the definition of the true least squares error of f:

$$\epsilon(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[(f(x) - y)^2],$$

where the subscript $(x, y) \sim \mathcal{D}$ makes clear that our input-output pairs are sampled according to $\mathcal{D}$. Our training and test losses are defined as:

$$\widehat{\epsilon}_{\text{train}}(f) = \frac{1}{N_{\text{train}}} \sum_{(x,y)\in S_{\text{train}}} (f(x) - y)^2$$

$$\widehat{\epsilon}_{\text{test}}(f) = \frac{1}{N_{\text{test}}} \sum_{(x,y)\in S_{\text{test}}} (f(x) - y)^2$$

We then train our algorithm (for example, using linear least squares regression) using the training set to obtain $\widehat{f}$.

a. *[3 points]* (bias: the test error) For all fixed f (before we've seen any data) show that

$$\mathbb{E}_{\text{train}}[\widehat{\epsilon}_{\text{train}}(f)] = \mathbb{E}_{\text{test}}[\widehat{\epsilon}_{\text{test}}(f)] = \epsilon(f).$$

Use a similar line of reasoning to show that the test error is an unbiased estimate of our true error for $\widehat{f}$. Specifically, show that:

$$\mathbb{E}_{\text{test}}[\widehat{\epsilon}_{\text{test}}(\widehat{f})] = \epsilon(\widehat{f})$$

b. *[4 points]* (bias: the train/dev error) Is the above equation true (in general) with regards to the training loss? Specifically, does $\mathbb{E}_{\text{train}}[\widehat{\epsilon}_{\text{train}}(\widehat{f})]$ equal $\epsilon(\widehat{f})$? If so, why? If not, give a clear argument as to where your previous argument breaks down.

c. *[8 points]* Let $\mathcal{F} = (f_1, f_2, \dots)$ be a collection of functions and let $\widehat{f}_{\text{train}}$ minimize the training error such that $\widehat{\epsilon}_{\text{train}}(\widehat{f}_{\text{train}}) \le \widehat{\epsilon}_{\text{train}}(f)$ for all $f \in \mathcal{F}$. Show that

$$\mathbb{E}_{\text{train}}[\widehat{\epsilon}_{\text{train}}(\widehat{f}_{\text{train}})] \le \mathbb{E}_{\text{train,test}}[\widehat{\epsilon}_{\text{test}}(\widehat{f}_{\text{train}})].$$

(Hint: note that

$$\begin{aligned}
\mathbb{E}_{\text{train,test}}[\widehat{\epsilon}_{\text{test}}(\widehat{f}_{\text{train}})] &= \sum_{f\in\mathcal{F}} \mathbb{E}_{\text{train,test}}[\widehat{\epsilon}_{\text{test}}(f)\mathbf{1}\{\widehat{f}_{\text{train}} = f\}] \\
&= \sum_{f\in\mathcal{F}} \mathbb{E}_{\text{test}}[\widehat{\epsilon}_{\text{test}}(f)]\mathbb{E}_{\text{train}}[\mathbf{1}\{\widehat{f}_{\text{train}} = f\}] \\
&= \sum_{f\in\mathcal{F}} \mathbb{E}_{\text{test}}[\widehat{\epsilon}_{\text{test}}(f)]\mathbb{P}_{\text{train}}(\widehat{f}_{\text{train}} = f)
\end{aligned}$$

where the second equality follows from the independence between the train and test set.)

What to Submit:

- **Part a** Proof
- **Part b** Brief Explanation (3-5 sentences)
- **Part c** Proof

Answer:

Part a:

$$\begin{aligned}
\mathbb{E}_{train}[\widehat{\epsilon}_{train}(f)] &= \mathbb{E}_{train}\left[\frac{1}{N_{train}}\sum_{(x,y)\in S_{train}}(f(x)-y)^2\right] \\
&= \frac{1}{N_{train}}\mathbb{E}_{train}\left[\sum_{(x,y)\in S_{train}}(f(x)-y)^2\right] \quad \text{where points in } S_{train} \text{are i.i.d and f is independent of the training set} \\
&= \frac{1}{N_{train}}N_{train}\mathbb{E}_{(x,y)\sim\mathcal{D}}[(f(x)-y)^2] \\
&= \mathbb{E}_{(x,y)\sim\mathcal{D}}[(f(x)-y)^2] = \epsilon(f) \quad \square
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_{test}[\widehat{\epsilon}_{test}(f)] &= \mathbb{E}_{test}\left[\frac{1}{N_{test}}\sum_{(x,y)\in S_{test}}(f(x)-y)^2\right] \\
&= \frac{1}{N_{test}}\mathbb{E}_{test}\left[\sum_{(x,y)\in S_{test}}(f(x)-y)^2\right] \quad \text{where points in } S_{train} \text{are i.i.d and f is independent of the test set} \\
&= \frac{1}{N_{test}}N_{test}\mathbb{E}_{(x,y)\sim\mathcal{D}}[(f(x)-y)^2] \\
&= \mathbb{E}_{(x,y)\sim\mathcal{D}}[(f(x)-y)^2] = \epsilon(f) \quad \square
\end{aligned}$$

Similarly, f turns to $\widehat{f}$, if the i.i.d assumption still holds, the equation would still hold. Since $\widehat{f}$ is found only in the training set, it does not interfere with the test set. The test set samples are thus i.i.d.

$$\begin{aligned}
\mathbb{E}_{test}[\widehat{\epsilon}_{test}(\widehat{f})] &= \mathbb{E}_{test}\left[\frac{1}{N_{test}}\sum_{(x,y)\in S_{test}}(\widehat{f}(x)-y)^2\right] \\
&= \frac{1}{N_{test}}\mathbb{E}_{test}\left[\sum_{(x,y)\in S_{test}}(\widehat{f}(x)-y)^2\right] \\
&= \frac{1}{N_{test}}N_{test}\mathbb{E}_{(x,y)\sim\mathcal{D}}[(\widehat{f}(x)-y)^2] \\
&= \mathbb{E}_{(x,y)\sim\mathcal{D}}[(\widehat{f}(x)-y)^2] = \epsilon(\widehat{f}) \quad \square
\end{aligned}$$

Part b:

Since $\widehat{f}$ is determined by the training set, the i.i.d assumption does not hold anymore. This implies that we no longer assume that $(\widehat{f}(x)-y)^2$ in $\sum_{(x,y)\in S_{train}}(\widehat{f}(x)-y)^2)$ is i.i.d

Part c:

$$\mathbb{E}_{train}[\widehat{\epsilon}_{train}(\widehat{f}_{train})] = \sum_{f\in\mathbb{F}}\mathbb{E}_{train}[\widehat{\epsilon}_{train}(\widehat{f}_{train}|f)]\mathbb{P}_{train}(\widehat{f}_{train}=f) \text{ by law of total expectation}$$

Since $\widehat{\epsilon}_{train}(\widehat{f}_{train}) \leq \widehat{\epsilon}_{train}(f)$,

$$\begin{aligned}
\mathbb{E}_{train}[\widehat{\epsilon}_{train}(\widehat{f}_{train})] &= \sum_{f \in \mathbb{F}} \mathbb{E}_{train}[\widehat{\epsilon}_{train}(\widehat{f}_{train}|f)]\mathbb{P}_{train}(\widehat{f}_{train} = f) \\
&\leq \sum_{f \in \mathbb{F}} \mathbb{E}_{train}[\widehat{\epsilon}_{train}(f|f)]\mathbb{P}_{train}(\widehat{f}_{train} = f) \\
&\leq \sum_{f \in \mathbb{F}} \mathbb{E}_{train}[\widehat{\epsilon}_{train}]\mathbb{P}_{train}(\widehat{f}_{train} = f)
\end{aligned}$$

Since we know that f is fixed from part a $\implies \mathbb{E}_{train}[\widehat{\epsilon}_{train}(\widehat{f}_{train})] = \mathbb{E}_{test}[\widehat{\epsilon}_{test}(\widehat{f}_{train})]$,

$$\begin{aligned}
\therefore \mathbb{E}_{train}[\widehat{\epsilon}_{train}(\widehat{f}_{train})] &\leq \sum_{f \in \mathbb{F}} \mathbb{E}_{train}[\widehat{\epsilon}_{train}]\mathbb{P}_{train}(\widehat{f}_{train} = f) \\
&\leq \sum_{f \in \mathbb{F}} \mathbb{E}_{test}[\widehat{\epsilon}_{test}]\mathbb{P}_{train}(\widehat{f}_{train} = f) \\
&\leq \mathbb{E}_{train,test}[\widehat{\epsilon}_{test}(\widehat{f}_{train})] \quad \square
\end{aligned}$$