Exercise 1.1

We want to show that the empirical risk $R_{\text{emp}}(f|\mathcal{D})$ converges to the expected risk $R(f)$ as the number of training samples $N \to \infty$.

Let $Z_i = \mathcal{L}(y_i, f(x_i))$, where (x_i, y_i) are i.i.d. samples and $\mathcal{L}$ - loss function. Then each Z_i is a bounded random variable, and all Z_i are i.i.d.

By the Law of Large Numbers, the sample average converges to the expected value:

$$\frac{1}{N}\sum_{i=1}^{N} Z_i \to \mathbb{E}[Z_i] \quad \text{as } N \to \infty.$$

This implies:

$$R_{\text{emp}}(f|\mathcal{D}) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}(y_i, f(x_i)) \to \mathbb{E}_{p(x,y)}[\mathcal{L}(y, f(x))] = R(f).$$

So the empirical risk converges to the expected risk, meaning R_{emp} is a consistent estimator of R.

Exercise 1.2

(a)

Model $f_1 \in \mathcal{H}_1$ has higher bias but lower variance. Since $\mathcal{H}_1$ is a simple linear functions), the model cannot capture complex relationships in the data, which results in high bias. However, due to simplicity, it also means that it behaves more consistently across different training sets, hence the low variance.

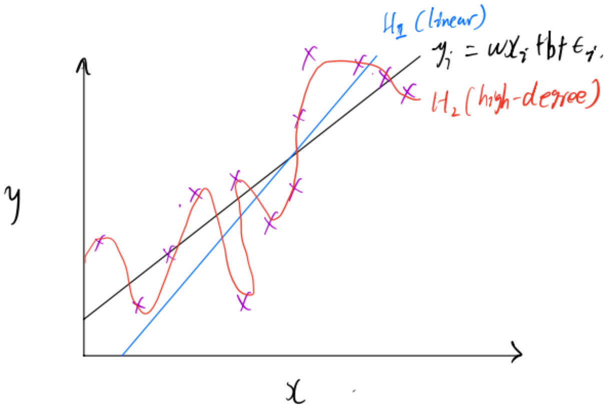
Model $f_2 \in \mathcal{H}_2$, a complex hypothesis space with high-degree polynomials has lower bias because it can fit a wider range of functions, including the training data very closely. But this increased flexibility leads to higher variance—small changes in the training data can lead to large changes in the fitted model.

This tradeoff explains how we can observe:

$$R_{\text{emp}}(f_2|\mathcal{D}) < R_{\text{emp}}(f_1|\mathcal{D}) \quad \text{but} \quad R(f_2) > R(f_1).$$

Model f_2 overfits the training data: it achieves a low empirical risk by closely fitting even the noise in the training set, but generalizes poorly, resulting in a higher expected risk.

Ex 1.2 (b).



It can be easily seen that H_2 is obviously overfitting.

Figure 1: (b) sketch

(c)

k-fold cross validation helps us compare models by checking how well they perform on data they have not seen. By spliting the data into k parts, train on some, and test on the rest, then repeat this k times, it gives a better idea of how the model will work on new data. It's useful because it gives a more accurate estimate of the model's true error $R(f)$, and helps avoid choosing a model that overfits the training data.

Exercise 1.3

(a)

The misclassification loss can be reformulated as:

$$R(f) = \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{y|\boldsymbol{x}}\left[\mathbb{1}_{\{f(\boldsymbol{x}) \neq y\}}\right]\right].$$

Note that the inner expectation is:

$$\mathbb{E}_{y|\boldsymbol{x}}\left[\mathbb{1}_{\{f(\boldsymbol{x}) \neq y\}}\right] = 1 - P(y = f(\boldsymbol{x})|\boldsymbol{x}).$$

Minimizing the misclassification loss is equivalent to maximizing $P(y|\boldsymbol{x})$ for given $\boldsymbol{x}$ w.r.t. y (MAP of y):

$$\hat{y}_{\text{MAP}} = \arg \max_{y \in \{0,1\}} = P(y|\boldsymbol{x}).$$

(b)

We have $R(\boldsymbol{f}) = \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}[\|\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x})\|_2^2]\right]$ for the Linear Regression problem. Our task is to find the $\boldsymbol{f}(\boldsymbol{x})$ that minimizes $R(\boldsymbol{f})$.

Let $\frac{\partial}{\partial \boldsymbol{f}}\mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}[\|\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x})\|_2^2] = 0$, we can solve for that optimal regression function:

$$\boldsymbol{f}^*(\boldsymbol{x}) = \mathbb{E}[\boldsymbol{y}|\boldsymbol{x}].$$

Exercise 1.4

(a)

Substitute X_i by $\mathcal{L}(\boldsymbol{y}_i, \boldsymbol{f}(\boldsymbol{x}_i))$, note that $R_{\text{emp}}(\boldsymbol{f}|\ \mathcal{D}) = \frac{1}{N}\sum_{i=1}^{N} X_i$ and $R(\boldsymbol{f}) = \mathbb{E}[X_i]$.

According to Hoeffding's inequality:

$$\Pr\left(\left|\frac{1}{N}\sum_{i=1}^{N} X_i - \mathbb{E}[X_i]\right| \geq \varepsilon\right) \leq 2\exp\left(-\frac{2N\varepsilon^2}{M^2}\right).$$

Thus,

$$\Pr\left(\left|\frac{1}{N}\sum_{i=1}^{N} X_i - \mathbb{E}[X_i]\right| \leq \varepsilon\right) \geq 1 - 2\exp\left(-\frac{2N\varepsilon^2}{M^2}\right).$$

Let $\varepsilon = \sqrt{\frac{M^2 \ln(2/\delta)}{2N}}$, we can show that:

$$\Pr\left(\left|\frac{1}{N}\sum_{i=1}^{N} X_i - \mathbb{E}[X_i]\right| \leq \sqrt{\frac{M^2 \ln(2/\delta)}{2N}}\right) \geq 1 - \delta.$$

(b)

For finite number of bounded training samples, it is always possible for ERM to approximate the correct expectation value, if the size of the training set is big enough. However, the decreasing speed of the absolute value maybe relatively slow (of order $O(1/\sqrt{n})$).