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 $O(1/\sqrt{n})$).