MLE Sheet 3

May 2025

Exercise 1: Empirical Risk Minimization

Task 1

Let the training dataset be

$$D = \{(x_i, y_i)\}_{i=1}^N \sim p(x, y) \quad \text{i.i.d.}$$

Let $f \in \mathcal{H}$ be a fixed hypothesis, and let $L(y, f(x))$ be a bounded loss function. That is:

$$\exists M < \infty \quad \text{such that} \quad |L(y, f(x))| \leq M \quad \text{for all } (x, y).$$

The **empirical risk** is defined as:

$$R_{\text{emp}}(f \mid D) = \frac{1}{N} \sum_{i=1}^N L(y_i, f(x_i)),$$

and the **expected risk** as:

$$R(f) = \mathbb{E}_{(x,y)\sim p(x,y)} \left[L(y, f(x)) \right].$$

Proof

The random variables are defined as:

$$Z_i := L(y_i, f(x_i)) \quad \text{for } i = 1, \ldots, N.$$

Since the pairs (x_i, y_i) are i.i.d., and f is fixed, the Z_i are i.i.d. random variables.
Furthermore, since L is bounded, there exists $M < \infty$ such that $|Z_i| \leq M$, implying that $\mathbb{E}[Z_i]$ is finite.

Using **SLLN**:

$$\frac{1}{N} \sum_{i=1}^N Z_i \xrightarrow{\text{a.s.}} \mathbb{E}[Z_1] = R(f),$$

as $N \to \infty$.

Therefore:

$$R_{\text{emp}}(f \mid D) = \frac{1}{N} \sum_{i=1}^N L(y_i, f(x_i)) \xrightarrow{\text{a.s.}} R(f).$$

Conclusion

Under the assumptions that the training data are i.i.d. and the loss function is bounded, the empirical risk converges almost surely to the expected risk. Thus, the empirical risk is a consistent estimator of the expected risk.

Task 2

a

Let $f_1 \in \mathcal{H}_1$ be a linear model and $f_2 \in \mathcal{H}_2$ a more flexible model such as a high-degree polynomial. Suppose:

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

This scenario is a classic example of **overfitting**. Although f_2 achieves lower empirical risk, it generalizes worse than f_1. The reason lies in the **bias-variance trade-off**:

- **Bias**: A model with high bias (such as a linear model) makes strong assumptions about the target function, which may lead to underfitting. It cannot capture complex patterns, resulting in systematic error.
- **Variance**: A model with high variance (such as a high-degree polynomial) is very sensitive to fluctuations in the training data. It can fit noise, resulting in poor generalization on unseen data.

Thus, even though f_2 fits the training data better, its expected risk $R(f_2)$ is higher due to increased variance. f_1, though less flexible, generalizes better by maintaining a better balance between bias and variance.

b

Assume the target function is a linear trend contaminated with Gaussian noise:

$$y_i = wx_i + b + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

We compare fits from:

- $\mathcal{H}_1$: a simple linear model.
- $\mathcal{H}_2$: a complex high-degree polynomial model.

Overfitting scenario:

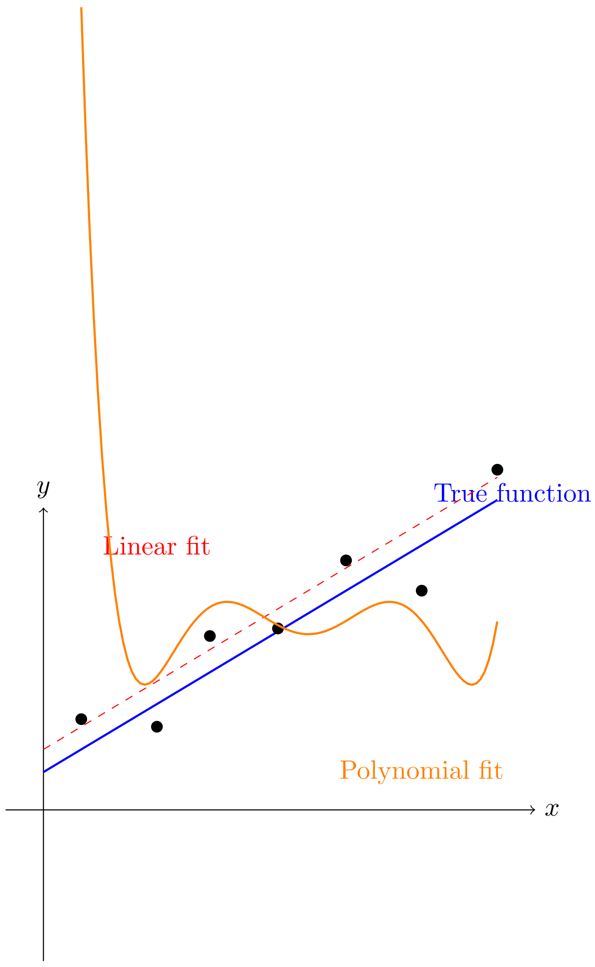
- The linear model from $\mathcal{H}_1$ captures the general trend but cannot fit all data points exactly. It ignores noise and may underfit slightly.
- The polynomial model from $\mathcal{H}_2$ can pass exactly through all training points, including noisy outliers, resulting in an oscillating curve that fits training data perfectly but fails to generalize.

We consider a regression setting where the true relationship is linear and corrupted with Gaussian noise:

$$y_i = wx_i + b + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2).$$

The plot below shows:

- The **true function** in blue (ground truth).
- **Noisy training samples** (black points).
- A **linear fit** (red dashed line), which captures the trend well and generalizes effectively.
- A **polynomial fit** (orange curve), which overfits by following noise in the data.



c

k**-fold cross-validation** partitions the dataset D into k equally-sized folds. The procedure is:

1. For each $j = 1, \dots, k$:
   - Train the model on $D \setminus D_j$, the data excluding fold j.
   - Evaluate the model on the held-out fold D_j.
2. Compute the average validation error across all folds:

$$R_{\text{cv}}(f) = \frac{1}{k} \sum_{j=1}^{k} R_{\text{val}}^{(j)}(f)$$

Advantages of cross-validation as an estimator for $R(f)$:

- **Better generalization estimate**: It gives a more reliable approximation of the true risk $R(f)$ than the empirical risk on the training data.
- **Model selection**: It helps compare models (e.g., from $\mathcal{H}_1$ and $\mathcal{H}_2$) by evaluating how well they generalize, not just how well they fit the training data.
- **Efficient use of data**: Every data point is used for both training and validation, improving the robustness of the evaluation.

Task 3

a

Consider a binary classification problem where $y \in \{0, 1\}$, and the loss function is the **misclassification loss**:

$$L(y, f(x)) = \mathbf{1}\{f(x) \neq y\},$$

where $\mathbf{1}\{A\}$ is the indicator function, equal to 1 if event A is true, and 0 otherwise.

The expected risk (also called the **true risk**) is defined as:

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

Let's minimize the **conditional risk** $R(f|x)$ for each input x:

$$R(f|x) = \mathbb{E}_{p(y|x)}[\mathbf{1}\{f(x) \neq y\}] = p(y = 1|x) \cdot \mathbf{1}\{f(x) \neq 1\} + p(y = 0|x) \cdot \mathbf{1}\{f(x) \neq 0\}.$$

This simplifies to:

$$R(f|x) = \begin{cases} p(y = 1|x) & \text{if } f(x) = 0, \\ p(y = 0|x) & \text{if } f(x) = 1. \end{cases}$$

To minimize $R(f|x)$, we choose the label $f(x) \in \{0,1\}$ that minimizes this conditional risk. Hence, the optimal classifier is:

$$f^*(x) = \arg\min_{c\in\{0,1\}} \mathbb{P}(y \neq c \mid x) = \arg\max_{c\in\{0,1\}} \mathbb{P}(y = c \mid x).$$

This is the **Bayes classifier**, which predicts the class with the highest posterior probability:

$$f^*(x) = \begin{cases} 1 & \text{if } \mathbb{P}(y=1|x) > \mathbb{P}(y=0|x), \\ 0 & \text{otherwise.} \end{cases}$$

Thus, the Bayes decision rule (MAP estimate) minimizes the expected misclassification risk.

b

Consider the regression case, where $y \in \mathbb{R}$, and the loss is the squared error:

$$L(y, f(x)) = \|y - f(x)\|_2^2.$$

Again, we consider the expected risk:

$$R(f) = \mathbb{E}_{p(x,y)}\left[\|y - f(x)\|^2\right] = \mathbb{E}_{p(x)}\left[\mathbb{E}_{p(y|x)}[\|y - f(x)\|^2]\right] = \mathbb{E}_{p(x)}[R(f|x)].$$

We now minimize the conditional risk:

$$R(f|x) = \mathbb{E}_{p(y|x)}\left[\|y - f(x)\|^2\right].$$

This is minimized when $f(x)$ is the conditional expectation of y given x. To show this, we differentiate:

$$\frac{d}{df(x)}\mathbb{E}_{p(y|x)}[(y - f(x))^2] = -2\mathbb{E}_{p(y|x)}[y - f(x)].$$

Setting the derivative to zero:

$$\mathbb{E}_{p(y|x)}[y - f(x)] = 0 \quad \Rightarrow \quad f(x) = \mathbb{E}[y \mid x].$$

Hence, the optimal regression function under squared loss is:

$$f^*(x) = \mathbb{E}[y \mid x],$$

i.e., the **conditional mean** of y given x.