MATH524 – Spring 2025

Problem Set: Week 10

1. (**Conditional expectation**) Show that $m(x) = \mathbb{E}[Y|X = x]$ minimizes

$$\mathbb{E}[|f(X) - Y|^2]$$

over all measurable functions $f : \mathbb{R}^d \to \mathbb{R}$. That is, the regression function minimizes the expected squared-loss.

Solution: For any $f : \mathbb{R}^d \to \mathbb{R}$,

$$\mathbb{E}[|f(X) - Y|^2] = \mathbb{E}[|f(X) - m(X) + m(X) - Y|^2] = \mathbb{E}[|f(X) - m(X)|^2] + \mathbb{E}[|m(X) - Y|^2] \tag{1}$$

since

$$\begin{aligned}
\mathbb{E}[(f(X) - m(X))(m(X) - Y)] &= \mathbb{E}[\mathbb{E}[(f(X) - m(X))(m(X) - Y)|X]] \\
&= \mathbb{E}[(f(X) - m(X))\mathbb{E}[(m(X) - Y)|X]] \\
&= \mathbb{E}[(f(X) - m(X))(m(X) - m(X))] \\
&= 0.
\end{aligned}$$

Thus, for minimizing the squared loss, the first term on the RHS of (1) is always nonnegative and zero if and only if $f(X) = m(X)$. And so $m = \mathbb{E}[Y|X]$ minimizes the L_2 loss.

2. (**Error decomposition**) Let $\mathcal{F}_n$ be some class of functions $f : \mathbb{R}^d \to \mathbb{R}$ that depend in some way on the data $\{(X_i, Y_i)\}_{i=1}^n$ that is generated by a standard regression model:

$$Y_i = m(X_i) + \varepsilon_i.$$

Consider the regression estimator m_n that satisfies

$$m_n = \arg\min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^n |f(X_i) - Y_i|^2.$$

Show the following

$$\int |m_n(x) - m(x)|^2 \mu(\,\mathrm{d}x) \le 2 \sup_{f \in \mathcal{F}_n} \left| \frac{1}{n} \sum_{i=1}^n |f(X_i) - Y_i|^2 - \mathbb{E}[(f(X) - Y)^2] \right| + \inf_{f \in \mathcal{F}_n} \int |f(x) - m(x)|^2 \mu(\,\mathrm{d}x).$$

Solution: Let $D_n = \{(X_i, Y_i)\}_{i=1}^n$. We will start by decomposing the LHS by using the decomposition from Question 1:

$$\begin{aligned}
&\int |m_n(x) - m(x)|^2 \mu(\,\mathrm{d}x) \\
&= \mathbb{E}[|m_n(X) - Y|^2 | D_n] - \mathbb{E}[|m(X) - Y|^2] \\
&= (\mathbb{E}[|m_n(X) - Y|^2 | D_n] - \inf_{f \in \mathcal{F}_n} \mathbb{E}[|f(X) - Y|^2]) + (\inf_{f \in \mathcal{F}_n} \mathbb{E}[|f(X) - Y|^2] - \mathbb{E}[|m(X) - Y|^2]).
\end{aligned} \tag{2}$$

Again, by the observation in Question 1, the second term in (2) is equal to

$$\inf_{f \in \mathcal{F}_n} \mathbb{E}[|f(X) - Y|^2] - \mathbb{E}[|m(X) - Y|^2] = \inf_{f \in \mathcal{F}_n} \int |f(x) - m(x)| \mu(\,\mathrm{d}x).$$

For the first term, by the definition of m_n being a minimizing function,

$$\begin{aligned}
&\mathbb{E}[|m_n(X) - Y|^2 | D_n] - \inf_{f \in \mathcal{F}_n} \mathbb{E}[|f(X) - Y|^2] \\
&= \sup_{f \in \mathcal{F}_n} \Bigg(\mathbb{E}[|m_n(X) - Y|^2 | D_n] - \frac{1}{n} \sum_{i=1}^{n} |m_n(X_i) - Y_i|^2 + \frac{1}{n} \sum_{i=1}^{n} |m_n(X_i) - Y_i|^2 \\
&\qquad - \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2 + \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2 - \mathbb{E}[|f(X) - Y|^2] \Bigg) \\
&\leq \sup_{f \in \mathcal{F}_n} \Bigg(\mathbb{E}[|m_n(X) - Y|^2 | D_n] - \frac{1}{n} \sum_{i=1}^{n} |m_n(X_i) - Y_i|^2 + \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2 - \mathbb{E}[|f(X) - Y|^2] \Bigg) \\
&\leq \sup_{f \in \mathcal{F}_n} \left| \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2 - \mathbb{E}[|f(X) - Y|^2] \right|.
\end{aligned}$$

Typically the first term in (2) is referred to as the *estimation error* and the second term as the *approximation error*. This reason for this is the first term is the loss due to approximation to a function in the class $\mathcal{F}_n$ while the second term is the loss due to the true regression function not being captured by $\mathcal{F}_n$.