Chapter 17
Bias-variance tradeoff I

17.1 Hypothesis space $\mathscr{H}$

In this book we choose the hypothesis space $\mathscr{H}$ as a compact subset of the space of continuous functions $\mathbb{C}(X)$ equipped with the standard norm

$$\|f\|_{\mathbb{C}(X)} := \|f\|_\infty := \sup_X |f(\mathbf{x})|, \quad \forall f \in \mathbb{C}(X).$$

While this choice does not cover all the practical problems of interest, it provides rich mathematical structures to study machine learning methods.

17.2 Empirical target function

Recall from Chapter 16 that the best target function is the regression function h^* which is, unfortunately, not computable. In general, h^* does not reside in $\mathscr{H}$, and thus the best that we hope for is to find a *target function* $\hat{h}$ in $\mathscr{H}$ that is closest, e.g. in the least squares sense, to h^*:

$$\hat{h} := \arg\min_{\mathscr{H}} \int_X (f(\mathbf{x}) - h^*(\mathbf{x}))^2 \, d\pi(\mathbf{x}) \tag{17.1}$$

Exercise 17.1. Show that $\hat{h}$ is also closest to $\mathbf{y}$, that is, $\hat{h}$ is a minimizer of the following least squares problem

$$\min_{\mathscr{H}} \int_Z (f(\mathbf{x}) - y)^2 \, d\pi(\mathbf{x}, \mathbf{y})$$

•

Exercise 17.1 implies

$$\hat{h} := \arg\min_{\mathscr{H}} \mathscr{R}(f). \tag{17.2}$$

Compared to (17.1), for which we have no access to h^*, we partially know y in (17.2) through the training set $S := \left\{\left(\mathbf{x}^i, y^i\right)\right\}_{i=1}^N$. Therefore, instead of seeking $\hat{h}$, which is not possible, we resort to a minimizer, the *empirical target function*, $\hat{h}_N$ of the empirical risk minimization problem defined as

$$\boxed{\hat{h}_N := \arg\min_{\mathscr{H}} \mathscr{R}_N(f) = \frac{1}{N}\sum_i^N \left(f\left(\mathbf{x}^i\right) - y^i\right)^2.} \tag{17.3}$$

The question that needs to be addressed is if the target and the empirical target functions exist. To that end, by Theorem 17.1, we need to show that the risk function and the empircal risk function are continuous.

Assumption 17.1 (M-Boundedness of the misfit on $\mathscr{H}$). For any $h \in \mathscr{H}$ and almost everywhere (a.e. or aka a.s.) in X, there holds

$$|h(\mathbf{x}) - y| \leq M. \tag{17.4}$$

Proposition 17.1. *Suppose $\mathscr{H}$ is bounded in the sense of Assumption 17.1. Then $\mathscr{R}, \mathscr{R}_N : \mathscr{H} \to \mathbb{R}$ are Lipschitz continuous.*

Proof. The proof is straightforward using (17.4). Indeed, we have

$$|\mathscr{R}(h_1) - \mathscr{R}(h_2)| = \left|\int_Z (h_1 + h_2 - 2y)(h_1 - h_2)\, d\pi(\mathbf{x}, \mathbf{y})\right| \leq 2M \|h_1 - h2\|_\infty.$$

Similarly,

$$\begin{aligned}|\mathscr{R}_N(h_1) - \mathscr{R}_N(h_2)| &= \left|\frac{1}{N}\sum_i^N \left(h_1\left(\mathbf{x}^i\right) + h_2\left(\mathbf{x}^i\right) - 2y^i\right)\left(h_1\left(\mathbf{x}^i\right) - h_2\left(\mathbf{x}^i\right)\right)\right| \\ &\leq \frac{1}{N}\sum_i^N \left|h_1\left(\mathbf{x}^i\right) + h_2\left(\mathbf{x}^i\right) - 2y^i\right| \|h_1 - h_2\|_\infty \leq 2M \|h_1 - h_2\|_\infty.\end{aligned}$$

□

Let us define the *sampling error* in computing the risk for any h as

$$\mathscr{E}(h) := \mathscr{R}(h) - \mathscr{R}_N(h). \tag{17.5}$$

As will be shown in the next section this sampling error plays the key role in estimating the sample/estimation error.

Exercise 17.2 (Continuity of Sampling Error). Show that $\mathscr{E} : \mathscr{H} \to \mathbb{R}$ is Lipschitz continuous. •

Corollary 17.1 (Existence of $\hat{h}$ and $\hat{h}_N$). *Suppose $\mathscr{H}$ is bounded in the sense of Assumption 17.1. Then $\hat{h}$ and $\hat{h}_N$ exist.*

Proof. The assertion is obvious due to Proposition 17.1 and Theorem 17.1. □

17.3 Bias-Variance Tradeoff

Suppose the empircal target $\hat{h}_N$ is already computed/estimated (to be discussed in details later), the question of interest is how to estimate the actual risk $\mathscr{R}\left(\hat{h}_N\right)$. This section decomposes this actual risk into two parts: the *sample or estimation* error and the approximation error. The compromise between these two errors is known as the *bias-variance tradeoff*. We begin with the following decomposition

$$\mathscr{R}\left(\hat{h}_N\right) = \underbrace{\mathscr{R}\left(\hat{h}_N\right) - \mathscr{R}\left(\hat{h}\right)}_{\mathscr{S}\left(\hat{h}_N\right):=} + \underbrace{\mathscr{R}\left(\hat{h}\right)}_{\mathscr{B}\left(\hat{h}\right):=}.$$

Let us first consider $\mathscr{S}\left(\hat{h}_N\right)$. We have

$$\begin{aligned}\mathscr{S}\left(\hat{h}_N\right) &= \underbrace{\mathscr{R}_N\left(\hat{h}_N\right) - \mathscr{R}_N\left(\hat{h}\right)}_{\leq 0} + \mathscr{R}\left(\hat{h}_N\right) - \mathscr{R}_N\left(\hat{h}_N\right) + \mathscr{R}_N\left(\hat{h}\right) - \mathscr{R}\left(\hat{h}\right) \\ &\leq \left|\mathscr{R}\left(\hat{h}_N\right) - \mathscr{R}_N\left(\hat{h}_N\right)\right| + \left|\mathscr{R}_N\left(\hat{h}\right) - \mathscr{R}\left(\hat{h}\right)\right|, \end{aligned} \tag{17.6}$$

where the negativeness of the first difference is due to the fact that $\hat{h}_N$ is a minimizer of $\mathscr{R}_N$. Note that the last two terms are the sampling errors incurred by approximating $\mathscr{R}\left(\hat{h}_N\right)$ and $\mathscr{R}\left(\hat{h}\right)$ using the training set S. Since $\mathscr{S}\left(\hat{h}_N\right)$ is bounded by the sampling errors, it is conventionally named as the *sample/estimation* error or the *variance*.

Now due to Exercise 16.1 we can rewrite $\mathscr{B}\left(\hat{h}\right)$ as

$$\mathscr{B}\left(\hat{h}\right) = \int_X \left(\hat{h}\left(\mathbf{x}\right) - h^*\left(\mathbf{x}\right)\right)^2 d\pi\left(\mathbf{x}\right) + \sigma_\pi^2, \tag{17.7}$$

which is independent of the training set. It actually depends only on the distance between $\hat{h}$ and the regression function h^*. That is, it is completely determined by the approximation capability of the hypothesis space $\mathscr{H}$. Thus, $\mathscr{B}\left(\hat{h}\right)$ is often called the *bias*.

We are in the position to discuss *bias-variance tradeoff*. For a fixed $\mathscr{H}$, the sample error clearly decreases as the sample size N increases. Now if we fix the sample size N, enlarging the hypothesis space $\mathscr{H}$ clearly reduces the bias but increase the sample error in general (will be shown in the next chapters). A popular tradeoff is to increase the dimension of the hypothesis space $\mathscr{H}$ as the sample size N increases. The question is how fast we should enlarge $\mathscr{H}$. To answer this question in a sensible and rigorous way, we need to first estimate both sample error and the bias, and then balance out these errors. This is the task of the following chapters.

17.4 Appendix

Theorem 17.1. *Let K be a compact set and $L : K \to \mathbb{R}$ be a continuous function. Then L is bounded. In particular, there exist $f, g \in K$ at which L attains its minimum and maximum.*