CUHK STAT5040: Statistical Learning Theory Spring 2023

Lecture 2: Approximation error and estimation error

Lecturer: Ben Dai

"*There is Nothing More Practical Than A Good Theory.*" — Kurt Lewin

1 A fundamental decomposition

Recall the example in Lecture 1.

Example 1.1 (Toy example)**.** *Some examples are illustrated as follows.*

Data. *Suppose* $(Y_1, \cdots, Y_n)$ *is a sequence of i.i.d. random samples with* $\mathbb{E}(Y_i) = \mu = 0$ *and* $\mathbf{Var}(Y_i) = \sigma = 1$.

Risk. $R(\theta) = \mathbb{E}l(Y, \theta) = \mathbb{E}\big((Y-\theta)^2\big)$.

Then,

Bayes decision function: $\theta^* = \mathbb{E}(Y) = \mu$.

Empirical estimator: $\widehat{\theta} = \bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ *is a function of* $(Y_1, \cdots, Y_n)$

Excess risk:

$$\mathscr{E}(\widehat{\theta}) = R(\widehat{\theta}) - R^* = \mathbb{E}\big((Y-\widehat{\theta})^2\big) - \mathbb{E}\big((Y-\mu)^2\big) = \mathbb{E}(\widehat{\theta}^2) = \widehat{\theta}^2.$$

Note that the expectation is taken w.r.t. Y*, which is independent with* $(Y_1, \cdots, Y_n)$.

- *Probabilistic bound. For any* $\delta > 0$*,*

$$\mathbb{P}\big(\mathscr{E}(\widehat{\theta}) \geq \delta^2\big) = \mathbb{P}\big(\widehat{\theta}^2 \geq \delta^2\big) = \mathbb{P}\big(|\widehat{\theta}| \geq \delta\big) \leq \frac{1}{\sqrt{n}\delta},$$

  where the last inequality follows from the Chebyshev's inequality. Alternatively, we can say, for any $\varepsilon > 0$*,*

$$\mathbb{P}\big(\mathscr{E}(\widehat{\theta}) \geq \frac{1}{\varepsilon^2 n}\big) \leq \varepsilon.$$

- *Convergence rate and excess risk consistency.*

$$\mathscr{E}(\widehat{\theta}) = O_P(1/n).$$

The provided toy example in Lecture 1 is a very special case.

- **A1.** The empirical minimizer $\widehat{f}_n$ and Bayes decision function f^* share the *same functional space*. Specifically, $\widehat{f}_n = \widehat{\theta} \in \mathbb{R}$ and $f^* = \mu \in \mathbb{R}$.
- **A2.** The ERM minimizer $\widehat{f}_n = \widehat{\theta} \in \mathbb{R}$ of has an analytic expression, that is, $\widehat{\theta} = \bar{Y}$.

In practice, both **A1** and **A2** are invalid.

- **D1.** Misspecified model. $\widehat{f}_n \in \mathscr{F}$ but $f^* \notin \mathscr{F}$.
- **D2.** $\widehat{f}_n$ can be obtained by some numerical algorithms, but no analytical solution.

The question is: can we provide a general theoretical framework to compute the bound?
A triangle inequality is widely used to address **D1**. Consider the following decomposition:

$$R(\widehat{f}_n) - R^* = \underbrace{R(\widehat{f}_n) - \inf_{f\in\mathscr{F}} R(f)}_{\text{Estimatin Error}} + \underbrace{\inf_{f\in\mathscr{F}} R(f) - R(f^*)}_{\text{Approximation Error}}. \tag{1}$$

To simplicity, let assume that the infimum is achievable: there exists a function $\bar{f} \in \mathscr{F}$, such that $R(\bar{f}) = \inf_{f\in\mathscr{F}} R(f)$. If $f^* \in \mathscr{F}$, then we just set $\bar{f} = f^*$, then the second term (approximation error) is zero. If $f^* \notin \mathscr{F}$, the idea for the decomposition is to "project" (in terms of the risk function) f^* to the given functional space $\mathscr{F}$, which can be interpreted as the best we can do under $\mathscr{F}$. Then, the estimation error and approximation error are treated separately.

Remark 1.2 (Estimation-approximation trade-off). From (1), when we enlarge the candidate class $\mathscr{F}$, then the estimation error will increase, since we have more "parameters" to estimate; yet the approximation error will decrease, since we have more candidates to approximate f^*. The idea is similar to bias-variance trade-off and under-/over-fitting. As a by-product, we would consider a data-dependent candidate class $\mathscr{F} = \mathscr{F}_n$, yielding **the method of sieves** and **the method of penalization**.

1.1 Approximation error

Recall the definition of approximation error:

$$\inf_{f\in\mathscr{F}} R(f) - R(f^*),$$

which only related to f^* and $\mathscr{F}$, either of them is random. Hence, the approximation error is more likely a math problem, which is highly related to approximation theory and functional analysis. The results are usually provided as "complexity of $\mathscr{F}$" and the "regularity of f^*" (such as smoothness).

Example 1.3 (RKHS). *Let $\mathscr{H}_K$ is a reproducing kernel Hilbert space (RKHS) with a Gaussian kernel*

$$K(\mathbf{x},\mathbf{x}') = \exp\big(-\frac{\|\mathbf{x}-\mathbf{x}'\|_2^2}{\sigma^2}\big), \quad \mathscr{X} = [0,1]^d.$$

Then, there exist positive constants c_0 and c_1, such that, for any $f^* \in W^{s,2}(\mathscr{X})$ and $U \geq c_0\|f^*\|_{L^2}$, we have

$$\inf_{\|f\|_{\mathscr{H}_K}\leq U} \|f - f^*\|_{L^2} \leq c_1 \log(U)^{-s/4}, \tag{2}$$

where $W^{s,2}$ is the Sobolev space as a subset of $L^2(\mathscr{X})$ with s-order weak derivatives.

Denote $\mathscr{F} = \{f \in \mathscr{H}_K : \|f\|_{\mathscr{H}_K} \leq U\}$, according to the result in Example 1.3, for most losses, there exist $c_3 > 0$ and $\alpha > 0$, such that

$$\inf_{f\in\mathscr{F}} R(f) - R(f^*) \leq c_3 \inf_{f\in\mathscr{F}} \|f - f^*\|_{L^2}^{\alpha} \leq c_1^{\alpha} c_3 \log(U)^{-s\alpha/4}, \tag{3}$$

which yields an upper bound for the approximation error. A non-technical interpretation is that f^* belongs to a larger space (Sobolev space), and we tend to approximate it with a smaller functional space (RKHS), the approximation error describes how well the ideal decision function f^* is approximated. Recall s is a parameter qualifying the regularity of f^*. Intuitively, a "smoother" function is easier to be approximated. U is a tuning parameter to control the volume of the candidate RKHS: *when U becomes large, then $\mathscr{F}$ will be enlarged, the approximation error becomes smaller,* which echos the estimation-approximation error trade-off in Remark 1.2.

More results on different functional spaces, including RKHS, B-spline, and deep neural networks [Bauer and Kohler, 2019, Yarotsky, 2017], are extensively studied in the literature.

1.2 Estimation error

Next, we focus on the estimation error. Recall "what we have":

- $\widehat{f}_n$ is the minimizer of ERM $\widehat{R}_n$.
- $\widehat{f}_n$ and $\bar{f}$ both belong to the candidate space $\mathscr{F}$.

Consider the following decomposition:

$$\begin{aligned} R(\widehat{f}_n) - \inf_{f\in\mathscr{F}} R(f) &= R(\widehat{f}_n) - R(\bar{f}) = R(\widehat{f}_n) - \widehat{R}_n(\widehat{f}_n) + \widehat{R}_n(\widehat{f}_n) - R(\bar{f}) \\ &= \underbrace{R(\widehat{f}_n) - \widehat{R}_n(\widehat{f}_n)}_{T_1} + \underbrace{\widehat{R}_n(\widehat{f}_n) - \widehat{R}_n(\bar{f})}_{\leq 0:\ \widehat{f}_n \text{ is a minimizer.}} + \underbrace{\widehat{R}_n(\bar{f}) - R(\bar{f})}_{T_2} \leq T_1 + T_2. \end{aligned} \tag{4}$$

It suffices to look at T_1 and T_2. Since $\bar{f}$ is a deterministic (non-random) function, T_2 can be treated by a concentration inequality. However, concentration does not work for T_1, since $\widehat{f}_n$ is a random estimator depend on the training samples $\mathscr{D}_n = (\mathbf{X}_i, \mathbf{Y}_i)_{i=1,\cdots,n}$. Consequently, T_1 yields two-level

nested randomness: (i) $\widehat{R}_n \to R$; and (ii) $\mathscr{D}_n \to \widehat{f}_n$. One solution is to consider uniform concentration:

$$T_1 \leq \sup_{f\in\mathscr{F}} \big(R(f) - \widehat{R}_n(f)\big).$$

The most important benefit of the provided upper bound is that we decouple the nested randomness, yet the price is to consider the concentration uniformly over a set of candidate functions. Note that $\big\{R(f) - \widehat{R}_n(f); f \in \mathscr{F}\big\}$ is so-called an (non-scaled) empirical process indexed by $\mathscr{F}$. Since $\bar{f} \in \mathscr{F}$, it suffices to consider a two-side empirical process to control the estimation error, that is,

$$R(\widehat{f}_n) - \inf_{f\in\mathscr{F}} R(f) \leq T_1 + T_2 \leq 2 \sup_{f\in\mathscr{F}} \big|R(f) - \widehat{R}_n(f)\big|. \tag{5}$$

The center of this course is to investigate the asymptotics of the empirical process

$$\mathbb{G}_n(f) = \sqrt{n}\big|R(f) - \widehat{R}_n(f)\big|,\ f \in \mathscr{F}.$$

Here, we use the scaled empirical process for consistency of the definition of the literature. Before digging deeply into the technical details, let's develop an overall insight of this measure. Again, when we enlarge the candidate space $\mathscr{F}$, it is clear that $\mathbb{G}_n$ becomes large. Back to the estimation error, the interpretation is that it is more difficult to search a good estimator in a more complicated candidate space.

1.3 Excess risk bounds

Combine the results of estimation error and approximation error, the regret is bounded by

$$\begin{aligned} R(\widehat{f}_n) - R^* &= R(\widehat{f}_n) - \inf_{f\in\mathscr{F}} R(f) + \inf_{f\in\mathscr{F}} R(f) - R(f^*) \\ &\leq \sup_{f\in\mathscr{F}} |\widehat{R}_n(f) - R(f)| + \inf_{f\in\mathscr{F}} R(f) - R(f^*). \end{aligned} \tag{6}$$

It is clear that when U increases, the approximation error will decrease, and the estimation error will increase. Therefore, after bound the estimation and approximation errors, we tend to (theoretically) find optimal tuning parameters to improve the convergence rate of the regret bound.

For example, in Example 1.3, according to (3),

$$R(\widehat{f}_n) - R^* \leq \sup_{f\in\mathscr{F}} |\widehat{R}_n(f) - R(f)| + c_1^{\alpha} c_3 \log(U)^{-s\alpha/4}.$$

Then, the probabilistic bound is given as

$$\mathbb{P}\big(R(\widehat{f}_n) - R^* \geq \varepsilon\big) \leq \mathbb{P}\big(\sup_{f\in\mathscr{F}} |\widehat{R}_n(f) - R(f)| \geq \varepsilon - c_1^{\alpha} c_3 \log(U)^{-s\alpha/4}\big). \tag{7}$$

Therefore, it suffices to investigate the asymptotics of $\sup_{f\in\mathscr{F}} |\widehat{R}_n(f) - R(f)|$.

The approximation error is an important topic in learning theory, yet its tools and results probably are "complementary" to asymptotic analysis in statistics. Partly to this reason, many works would either omit the approximation error, nor assume that $f^* \in \mathscr{F}$, to highlight the statistical properties of methods. In the sequel of this course, we will mainly focus on the estimation error.

References

[Bauer and Kohler, 2019] Bauer, B. and Kohler, M. (2019). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics*, 47(4):2261–2285.

[Yarotsky, 2017] Yarotsky, D. (2017). Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103–114.