Lecture 6: Method of regularization

sieve estimator, hyperparameter, and tuning

Lecturer: Ben Dai

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

1 Recall

Based on Lectures 1-5, we are ready to investigate the asymptotics of the excess risk for a given function class $\mathscr{F}$. Yet, for different function class, the convergence rates might be different.

A consequent question is: can we find a "optimal" function class (depending on the sample size) yielding a sharp convergence rate?

Recall the approximation error and estimation error trade-off. When the complexity of $\mathscr{F}$ is increasing, then the estimation error is increasing, and the approximation is decreasing. Therefore, we can find a "optimal" function class (depending on the sample size) via balancing the estimation/approximation errors. Alternatively, a sequence of data-dependent function classes is constructed:

$$\mathscr{F}_1 \subset \mathscr{F}_2 \subset \cdots \mathscr{F}_n \subset \cdots \to \mathscr{F}^*,$$

where $\mathscr{F}^*$ is the target function class such that $f^* \in \mathscr{F}^*$. Specifically, we summarize our goal as the following **Aim**.

Aim. Find a sequence of function classes $(\mathscr{F}_n)_{n=1,2,\cdots}$, to achieve the "optimal" convergence rate of the excess risk:

$$\mathscr{E}(\widehat{f}_n) = R(\widehat{f}_n) - R(f^*), \quad \text{where } \widehat{f}_n = \operatorname*{argmin}_{f \in \mathscr{F}_n} \widehat{R}_n(f).$$

2 Method of regularization

Aim provides a high-level idea of the sieve estimator, more specific question is how can we construct a sequence of function classes. The answer is introducing **hyperparameter** or **tuning parameters**. Some examples are itemized as follows.

- Regularization. $\mathscr{F}_n = \{f \in \mathscr{F} : \|f\|_{\mathscr{F}} \leq \xi_n\}$.
- Feedforward neural networks. $\mathscr{F}_n = \{f \in \mathscr{F}(W_1, \cdots, W_D)\}$, W_j is #node in the j-th layer.
- Decision tree. $\mathscr{F}_n = \{f \in \mathscr{F}(D_n, W_n)\}$, D_n is #layer, and W_n is #leaf node.

We will focus on *regularization*, yet the analytic tools can be extended to a more general setup. To proceed, we give the constrained ERM (C-ERM):

$$\min_{f\in\mathscr{F}} \widehat{R}_n(f), \quad \text{subj to. } \|f\|_{\mathscr{F}}^2 \le \xi_n^2.$$

Ideally, we turn to investigate the asymptotics of $\widehat{f}_n$ from C-ERM, yet it is usually difficult to directly solve a constrained optimization as in C-ERM. Instead, a penalized (regularized) version (R-ERM) is much easier to solve.

$$\min_{f\in\mathscr{F}} \widehat{R}_n(f) + \lambda_n \|f\|_{\mathscr{F}}^2, \tag{1}$$

where $\lambda_n \to 0$ is a hyperparameter replacing ξ_n to control the complexity of $\mathscr{F}_n$. In fact, when $l(\cdot,\cdot)$ and $\|\cdot\|_{\mathscr{F}}$ are convex, then C-ERM and R-ERM are equivalent, yet the correspondence between ξ_n and λ_n is usually unclear. Lemma 2.1 provides a potential finite-sample functional class $\mathscr{F}_n$ including $\widehat{f}_n$.

Lemma 2.1. *Suppose $\widehat{f}_n$ is a minimizer of R-ERM in* (1)*, then $\widehat{f}_n \in \mathscr{F}_n = \big\{f \in \mathscr{F} : \|f\|_{\mathscr{F}} \le c\lambda_n^{-1/2}\big\}$.*

Suppose $\widehat{f}_n$ is a minimizer of R-ERM in (1), we can slightly revise **Aim** as **Aim′**.

Aim′. Find a sequence of λ_n to achieve the "optimal" convergence rate of the excess risk $\mathscr{E}(\widehat{f}_n)$.

Remark 2.2 (ERM and R-ERM). The difference between ERM and R-ERM can be summarized as follows.

- ERM: $\widehat{f}_n$ is a minimizer of $\widehat{R}_n(f)$; R-ERM: $\widehat{f}_n$ is a minimizer of $\widehat{R}_n(f) + \lambda_n\|f\|_{\mathscr{F}}^2$.
- ERM: the approximation function $\bar{f}$ is usually fixed; R-ERM: the approximation function $\bar{f}_n$ is different based on different λ_n

2.1 New decomposition

Again, we decompose the excess risk of a regularized estimator as estimation/approximation errors:

$$\begin{aligned}
\mathscr{E}(\widehat{f}_n) &= R(\widehat{f}_n) - R(\bar{f}_n) + R(\bar{f}_n) - R(f^*) = R(\widehat{f}_n) - \widehat{R}_n(\widehat{f}_n) + \widehat{R}_n(\bar{f}_n) - R(\bar{f}_n) \\
&\quad + R(\widehat{f}_n) - \widehat{R}_n(\bar{f}_n) + R(\bar{f}_n) - R(f^*) \\
&= R(\widehat{f}_n) - \widehat{R}_n(\widehat{f}_n) + \widehat{R}_n(\bar{f}_n) - R(\bar{f}_n) \\
&\quad + \widehat{R}_n(\widehat{f}_n) + \lambda_n\|\widehat{f}_n\|_{\mathscr{F}}^2 - R(\bar{f}_n) - \lambda_n\|\bar{f}_n\|_{\mathscr{F}}^2 \\
&\quad + \lambda_n\|\bar{f}_n\|_{\mathscr{F}}^2 - \lambda_n\|\widehat{f}_n\|_{\mathscr{F}}^2 + R(\bar{f}_n) - R(f^*) \\
&\le 2\sup_{f\in\mathscr{F}_n} \big|\widehat{R}_n(f) - R(f)\big| + \lambda_n\|\bar{f}_n\|_{\mathscr{F}}^2 + R(\bar{f}_n) - R(f^*),
\end{aligned} \tag{2}$$

where the last inequality follows from the fact that both $\widehat{f}_n$ and $\bar{f}_n$ belong to $\mathscr{F}_n$, and $\widehat{f}_n$ is a minimizer of R-EMR.

Note that the decomposition in (2) differs from the previous one with regard to the approximation error. We define the new approximation error:

$$\mathbf{Approx}(\lambda_n) = \inf_{f \in \mathscr{F}} R(f) - R(f^*) + \lambda_n \|f\|_{\mathscr{F}}^2.$$

For simplicity, we assume that the "optimal" approximation function $\bar{f}_n$ is achievable

$$\bar{f}_n = \arg\min_{f} \; R(f) - R(f^*) + \lambda_n \|f\|_{\mathscr{F}}^2.$$

Note that

$$R(f) - R(f^*) + \lambda_n \|f\|_{\mathscr{F}}^2 \leq R(0) - R(f^*),$$

yielding that $\bar{f}_n \in \mathscr{F}_n = \{f \in \mathscr{F} : \|f\|_{\mathscr{F}} \leq c\lambda_n^{-1/2}\}$.

Remark 2.3 (Approximation error bounds). Under some conditions, a polynomial decay bound can be obtained for the approximation error:

$$\mathbf{Approx}(\lambda_n) \leq c\lambda_n^s,$$

where s is related to the regularity/smoothness of f^*. Moreover, there is ample literature devoted to bound the approximation error; see, for instance, Sections 5.4-5.6 in [Steinwart and Christmann, 2008], and Chapters 4 and 6 in [Cucker and Zhou, 2007] and references therein.

3 Probabilistic bound

Now, we are ready to derive the probabilistic bound for R-ERM estimator.

$$\mathbb{P}\big(\mathscr{E}(\widehat{f}_n) \geq \varepsilon_n\big) \leq \mathbb{P}\Big(\sup_{f \in \mathscr{F}_n} \big|\widehat{R}_n(f) - R(f)\big| \geq \frac{1}{2}\big(\varepsilon_n - \mathbf{Approx}(\lambda_n)\big)\Big).$$

According to Corollary 5.1 in Lecture 4, we derive the following corollary.

Corollary 3.1. *Suppose the loss function $l(\cdot,\cdot)$ is uniformly bounded by a constant U, then for any*

$$\varepsilon_n \geq \mathbf{Approx}(\lambda_n) + 8\mathbb{E} \sup_{f \in \mathscr{F}_n} |\mathbf{Rad}_n(l \bullet f)|,$$

with $(l \bullet f)(\mathbf{Z}) = l\big(\mathbf{Y}, f(\mathbf{X})\big)$, we have

$$\mathbb{P}\Big(\sup_{f \in \mathscr{F}_n} |\widehat{R}_n(f) - R(f)| \geq \varepsilon_n\Big) \leq \exp\Big(-\frac{n\varepsilon_n^2}{8\big(\sigma_{\mathscr{F}_n}^2 + (1/2 + U/3)\varepsilon_n\big)}\Big). \tag{3}$$

References

[Cucker and Zhou, 2007] Cucker, F. and Zhou, D. X. (2007). *Learning theory: an approximation theory viewpoint*, volume 24. Cambridge University Press.

[Steinwart and Christmann, 2008] Steinwart, I. and Christmann, A. (2008). *Support vector machines*. Springer Science & Business Media.