CUHK STAT6050: Statistical Learning Theory Spring 2022

Lecture 7: Nonparametric regression in RKHS

kernel method, Representer theorem

Lecturer: Ben Dai

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

1 Recall

Based on Lectures 1-6, we are able to compute the convergence rate and establish a probabilistic bound for a general learning method/algorithm. For illustrate, we turn to investigate the asymptotics of the nonparameteric regression in Reproducing kernel Hilbert space (RKHS).

2 RKHS

2.1 Why RKHS?

Requirements. At least, we require pointwise convergence, that is,

$$\|f_n - f\|_{\mathscr{F}} \to 0 \quad \Longrightarrow \quad f_n(x) \to f(x), \text{ for any } x \in \mathscr{X}.$$

Theorem 2.1 (The Riesz Representation Theorem for Hilbert Spaces)**.** *Let $\mathscr{H}$ be a Hilbert space, and $L : \mathscr{H} \to \mathbb{R}$ is a linear continuous functional on $\mathscr{H}$. Then there exists some $K \in \mathscr{H}$ such that for every $h \in \mathscr{H}$, we have $L(h) = \langle h, K \rangle_{\mathscr{H}}$.*

According to Riesz representation theorem, if $L = \delta_x$ is a linear continuous functional on $\mathscr{H}$, then we have

$$f(\mathbf{x}) = \delta_{\mathbf{x}}(f) = \langle f, K_{\mathbf{x}} \rangle_{\mathscr{H}},$$

which means that we can represent function evaluation as the inner production on $K_{\mathbf{x}}$.

Definition 2.2 (RKHS)**.** A Hilbert space $\mathscr{H}$ is said to be a Reproducing Kernel Hilbert Space (RKHS) if $\delta_{\mathbf{x}}$ is a linear continuous functional on $\mathscr{H}$, for any $\mathbf{x} \in \mathscr{X}$.

Theorem 2.3. *Suppose $\mathscr{H}$ is a RKHS, then*

$$\|h_n - h\|_{\mathscr{H}} \to 0 \quad \Longrightarrow \quad h_n(x) \to h(x), \textit{ for any } x \in \mathscr{X}.$$

2.2 From kernel function to RKHS

The overall idea in this section is to construct a RKHS from a kernel function. Recall the construction in finite-dimension space: (i) basis; (ii) inner production among basis.

From Riesz representation theorem, $\{K_{\mathbf{x}} : \mathbf{x} \in \mathscr{X}\}$ will be a good option as a basis function, and their inner production is given as:

$$\langle K_{\mathbf{x}'}, K_{\mathbf{x}} \rangle_{\mathscr{H}} = \delta_{\mathbf{x}}(K_{\mathbf{x}'}) = K_{\mathbf{x}'}(\mathbf{x}).$$

Note that $K_{\mathbf{x}'}(\mathbf{x}) : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ is a symmetric bivariate function (so-called kernel function). Once we define a $K(\mathbf{x}, \mathbf{x}') = K_{\mathbf{x}'}(\mathbf{x}) = K_{\mathbf{x}}(\mathbf{x}')$, then we define the basis $K_{\mathbf{x}}$ and the inner production between two basis $K(\mathbf{x}, \mathbf{x}')$.

Step 1. Define a pre-RKHS by linear span of kernel functions.

To mimic the construction in finite-dimension case, we first construct a pre-RKHS $\mathscr{H}_0$ as a set of functions:

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i K_{\mathbf{x}_i} = \sum_{i=1}^{n} \alpha_i K(\mathbf{x}_i, \mathbf{x}),$$

equipped with an inner production:

$$\langle f, f' \rangle_{\mathscr{H}_0} = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j)$$

Step 2. Generate a RKHS by taking "closure" of the pre-RKHS.

Then, we create $\mathscr{H}$ as an "closure" of $\mathscr{H}$:

$$\mathscr{H} = \overline{\mathscr{H}_0} = \mathscr{H}_0 \bigcup \{\text{limit points of all } \mathscr{H}_0\text{-Cauchy sequences}\}, \quad f(\mathbf{x}) = \sum_{i=1}^{\infty} \alpha_i K(\mathbf{x}_i, \mathbf{x}) \in \mathscr{H},$$

equipped with the inner production:

$$\langle f, f' \rangle_{\mathscr{H}} = \lim_{n \to \infty} \langle f, f' \rangle_{\mathscr{H}_0}$$

Then, we turn to verify $\mathscr{H}$ is indeed a RKHS, and find requirements for a valid kernel function $K(\cdot, \cdot)$. Recall the definition of RKHS: (i) $\mathscr{H}$ equipped with $\langle f, f' \rangle_{\mathscr{H}}$ is a Hilbert space; (ii) δ_x is a continuous functional on $\mathscr{H}$. One can find the formal proof in [Sejdinovic and Gretton, 2012].

Remark 2.4 (Positive-definite kernel). One quick observation is that a valid kernel function should be symmetric and positive definite.

- From **inner production**: a kernel function should be symmetric.

$$K(\mathbf{x}, \mathbf{x}') = \langle K_{\mathbf{x}}, K_{\mathbf{x}'} \rangle_{\mathscr{H}} = \langle K_{\mathbf{x}'}, K_{\mathbf{x}} \rangle_{\mathscr{H}} = K(\mathbf{x}', \mathbf{x})$$

- From **norm**: a kernel function should be positive definite.

$$0 \le \|f\|_{\mathscr{H}}^2 = \| \sum_{i=1}^{n} \alpha_i K(\mathbf{x}_i, \mathbf{x}) \|_{\mathscr{H}}^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \langle K_{\mathbf{x}_i}, K_{\mathbf{x}_j} \rangle_{\mathscr{H}} = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j),$$

  which holds for any $f \in \mathscr{H}$ or for any $n \ge 1$, any $(\alpha_1, \cdots, \alpha_n) \in \mathbb{R}^n$, any $(\mathbf{x}_1, \cdots, \mathbf{x}_n) \in \mathscr{X}^n$.

2.3 Definitions and theorems

In this section, we give formal definitions and theorems used in Section 2.2.

Definition 2.5 (Reproducing kernel)**.** Let $\mathscr{H}$ be a RKHS. A function $K : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ is called a reproducing kernel of $\mathscr{H}$ if it satisfies

- For any $\mathbf{x} \in \mathscr{X}$, $K_{\mathbf{x}} = K(\cdot, \mathbf{x}) \in \mathscr{H}$.
- (Reproducing property). For any $\mathbf{x} \in \mathscr{X}$, and any $h \in \mathscr{H}$, $\langle h, K_{\mathbf{x}} \rangle_{\mathscr{H}} = h(\mathbf{x})$.

Definition 2.6 (Positive semi-definite (Mercer) Kernel)**.** A symmetric function $K : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ is positive definite kernel, if for any $n \geq 1$, any $(\alpha_1, \cdots, \alpha_n) \in \mathbb{R}^n$, any $(\mathbf{x}_1, \cdots, \mathbf{x}_n) \in \mathscr{X}^n$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \geq 0.$$

Next, we have Moore-Aronszajn theorem to guarantee the legality of the construction in Section 2.2.

Theorem 2.7 (Moore-Aronszajn theorem)**.** *Let $K : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ be positive semi-definite. Then there is unique RKHS $\mathscr{H}$ with reproducing kernel K.*

Finally, we summarize as:

"legal" kernel $\iff$ reproducing kernel $\iff$ positive semi-definite kernel $\iff$ *RKHS*.

2.4 Examples

- Linear kernel.
$$K(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\intercal \mathbf{x}'$$
- Gaussian kernel.
$$K(\mathbf{x}, \mathbf{x}') = \exp\big(-\frac{\|\mathbf{x} - \mathbf{x}'\|_2^2}{\sigma^2}\big)$$
- γ-degree polynomial kernel.
$$K(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^\intercal \mathbf{x}' + c)^\gamma$$

Remark 2.8. What's the difference among different kernels? Theoretically, it effects both estimation/approximation errors, see discussion in Section 4. Practically, it highly related to the topic *multiple kernel learning*, see [Gönen and Alpaydın, 2011] and references therein.

3 Regression in RKHS

We denote vector of features as $\mathbf{X} \in \mathbb{R}^d$, a scalar outcome as $Y \in \mathbb{R}$. Suppose $\mathbf{Z} = (\mathbf{X}, Y)$ satisfy a nonparametric regression model:

$$Y = f^*(\mathbf{X}) + \varepsilon,$$

where ε is a random noise with $\mathbb{E}(\varepsilon) = 0$ and $\varepsilon \perp \mathbf{X}$, and f^* is the true conditional mean function with $\|f^*\|_\infty < \infty$. Our goal is to find a decision function f minimizing the mean squared loss:

$$R(f) = \mathbb{E}\Big(l\big(Y, f(\mathbf{X})\big)\Big) = \mathbb{E}\Big(\big(Y - f(\mathbf{X})\big)^2\Big).$$

Let's summarize the quantities of interests.

- **Bayes rule.** $f^*(\mathbf{x}) = \mathbb{E}(Y|\mathbf{X} = \mathbf{x})$ is the global minimizer of $R(f)$.
- **Excess risk.**

$$\mathscr{E}(f) = R(f) - R(f^*) = \mathbb{E}\Big(\big(f(\mathbf{X}) - f^*(\mathbf{X})\big)^2\Big) = \big\|f - f^*\big\|^2_{L^2(\mathbb{P}_\mathbf{X})}$$

- **R-ERM.** Given random samples $(\mathbf{X}_i, Y_i)_{i=1,\cdots,n}$, and a RKHS $\mathscr{H}_K$ associated with a kernel K,

$$\widehat{f}_n = \underset{f \in \mathscr{H}_K}{\arg\min} \frac{1}{n} \sum_{i=1}^n \big(Y_i - f(\mathbf{X}_i)\big)^2 + \lambda_n \|f\|^2_{\mathscr{H}_K}$$

- **Asymptotics.** Finally, we aim to investigate the asymptotics of $\mathscr{E}(\widehat{f}_n)$.

First, we consider the empirical optimization of **ERM** on RKHS. Indeed, this can be challenging, since $f \in \mathscr{H}_K$, and the RKHS $\mathscr{H}_K$ is an infinity-dimensional function class. Fortunately, we have the Representer Theorem, which implies that ERM reduces to a finite dimensional optimization problem.

Theorem 3.1 (Representer Theorem [Kimeldorf and Wahba, 1970, Wahba, 1990])**.** *Let $K(\cdot,\cdot)$ is a kernel function and $\mathscr{H}_K$ be its associated RKHS. Given a training sample $(\mathbf{X}_i, Y_i)_{i=1,\cdots,n}$, consider the R-ERM:*

$$\widehat{f}_n = \underset{f \in \mathscr{H}_K}{\arg\min}\ \mathscr{L}_n\big(Y_1, \cdots, Y_n, f(\mathbf{X}_1), \cdots, f(\mathbf{X}_n)\big) + g(\|f\|^2_{\mathscr{H}_K}). \tag{1}$$

Suppose $g(\cdot)$ is an increasing function $\mathscr{L}_n$ depends on f only through $f(\mathbf{X}_1), \cdots, f(\mathbf{X}_n)$. Then, every minimizer of (1) *has form:*

$$\widehat{f}_n(\mathbf{x}) = \sum_{i=1}^n \widehat{\alpha}_i K(\mathbf{X}_i, \mathbf{x}),$$

for some $(\widehat{\alpha}_i)_{i=1,\cdots,n} \in \mathbb{R}^n$.

Remark 3.2. I recommend readers to check the summary[1] by Grace Wahba and Yuedong Wang. Some quotas:

- *The significance of the representer theorem is that the solution in an infinite dimensional space falls in a finite dimensional space. This property makes it possible to compute estimates of general regularization problems in infinite dimensional spaces.*

According to the Representer Theorem, the R-ERM can be reduced to:

$$\min_{\boldsymbol{\alpha}} \frac{1}{n}\sum_{i=1}^n \big(Y_i - \sum_{j=1}^n \alpha_j K(\mathbf{X}_j,\mathbf{X}_i)\big)^2 + \lambda_n \|\sum_{j=1}^n \alpha_j K(\mathbf{X}_j,\cdot)\|^2_{\mathscr{H}_K} = \min_{\boldsymbol{\alpha}} \frac{1}{n}\sum_{i=1}^n \big(Y_i - \boldsymbol{\alpha}^\intercal \mathbf{K}_i\big)^2 + \lambda_n \boldsymbol{\alpha}^\intercal \mathbf{K}\boldsymbol{\alpha},$$

where $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)^\intercal$ and $\mathbf{K} = (K_{ij})_{n\times n}$ is a kernel matrix with $K_{ij} = K(\mathbf{X}_i, \mathbf{X}_j)$. Note that the optimization in the right-hand side is reduced to a linear regression with a structured penalization, and the solution is given as:

$$\widehat{\boldsymbol{\alpha}} = \big(\mathbf{K} + \lambda_n \mathbf{I}\big)^{-1} \mathbf{Y}_{1:n}, \quad \widehat{f}_n(\mathbf{x}) = \sum_{j=1}^n \widehat{\alpha}_j K(\mathbf{X}_j, \mathbf{x}).$$

4 A loose excess risk upper bound

Next, we turn to bound the excess risk of RKHS regression. For simplicity, we assume $|\varepsilon_i| \le U$, and we have $\widehat{f}_n(\mathbf{X}_i) \le U$, unless we can truncate it by U to further minimize R-ERM. Note that $\widehat{f}_n \in \mathscr{H}_K$ is a minimizer of:

$$\frac{1}{n}\sum_{i=1}^n \Big(\big(f(\mathbf{X}_i) - f^*(\mathbf{X}_i)\big)^2 + 2\varepsilon_i\big(f(\mathbf{X}_i) - f^*(\mathbf{X}_i)\big)\Big) + \lambda_n \|f\|^2_{\mathscr{H}_K} := \widehat{R}_n(f) + \lambda_n \|f\|^2_{\mathscr{H}_K}.$$

Then, let $\mathscr{H}_n = \{f \in \mathscr{H} : \|f\|_{\mathscr{H}_K} \le c\lambda_n^{-1/2}\}$,

$$\mathscr{E}(\widehat{f}_n) = R(\widehat{f}_n) - R(f^*) \le 2 \sup_{f\in\mathscr{H}_n} |\widehat{R}_n(f) - R(f)| + \mathbf{Approx}(\lambda_n)$$

4.1 Estimation error in RKHS

Based on Corollary 3.1 in Lecture 6, it suffices to bound $\mathbf{Rad}_n(l \bullet f)$ and $\mathbf{Approx}(\lambda_n)$. We treat them separately. Note that

$$\begin{aligned}
\mathbf{Rad}_n(l\bullet f) &= \Big\|\frac{1}{n}\sum_{i=1}^n \rho_i\Big(\big(f(\mathbf{X}_i) - f^*(\mathbf{X}_i)\big)^2 + 2\varepsilon_i\big(f(\mathbf{X}_i) - f^*(\mathbf{X}_i)\big)\Big)\Big\|_{\mathscr{H}_n} \\
&= \Big\|\frac{1}{n}\sum_{i=1}^n \rho_i\big(f(\mathbf{X}_i) - f^*(\mathbf{X}_i)\big)^2\Big\|_{\mathscr{H}_n} + 2\Big\|\frac{1}{n}\sum_{i=1}^n \rho_i\varepsilon_i\big(f(\mathbf{X}_i) - f^*(\mathbf{X}_i)\big)\Big\|_{\mathscr{H}_n} \\
&\le 4U\mathbf{Rad}_n(f).
\end{aligned}$$

It suffices to bound the Rademacher complexity of $\mathscr{H}_K$, we have the following lemma.

[1] http://pages.stat.wisc.edu/~wahba/ftp1/wahba.wang.2019submit.pdf

Lemma 4.1. *Suppose K is a uniformly bounded kernel with $\sup_{\mathbf{x}\in\mathscr{X}} K(\mathbf{x},\mathbf{x}) \le K_0 < \infty$, $\mathscr{H}_K$ is its corresponding RKHS, and $\mathscr{H}_K(r) = \{f \in \mathscr{H}_K : \|f\|_{\mathscr{H}_K} \le r\}$ is a $\mathscr{H}_K$-ball with a radius r. Then,*

$$\mathbb{E}_\rho \|\mathbf{Rad}_n(f)\|_{\mathscr{H}_K(r)} \le rK_0\sqrt{\frac{1}{n}}.$$

Therefore, $\mathbf{Rad}_n(l \bullet f) \le cUK_0(n\lambda_n)^{-1/2}$.

4.2 Approximation error in RKHS

Then, we turn to bound $\mathbf{Approx}(\lambda_n)$. We present Proposition 8.5 in [Cucker and Zhou, 2007] to illustrate the approximation error for RKHS. Recall $\mathbf{Approx}(\lambda_n)$ in RKHS regression:

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

provided that $R(f) = \mathbb{E}(Y - f(\mathbf{X}))^2$.

Theorem 4.2 ([Cucker and Zhou, 2007])**.** *Let $\mathscr{X} \subset \mathbb{R}^d$ be a compact domain, and K be a reproducing kernel. Suppose there exists $0 < s \le 1$, such that $f^* \in Range(L_K^{s/2})$, then*

$$\mathbf{Approx}(\lambda_n) \le A_0\lambda_n^s,$$

where $A_0 = \mathscr{E}\big(L_K^{-s/2} f^*\big)$.

4.3 Hyperparameter tuning

Taken together, if we further assume that $\mathscr{X}$ is a compact domain almost surely, and

$$\varepsilon_n \ge A_0\lambda_n^s + cUK_0(n\lambda_n)^{-1/2} \ge \mathbf{Approx}(\lambda_n) + 8\|\mathbb{E}\mathbf{Rad}_n(l \bullet f)\|_{\mathscr{H}_n},$$

then

$$\mathbb{P}\Big(\|f - f^*\|^2_{L^2(\mathbb{P}_\mathbf{x})} \ge \varepsilon_n\Big) \le \exp\Big(-\frac{n\varepsilon_n^2}{8\big(U^2 + (1/2 + U/3)\varepsilon_n\big)}\Big).$$

Note that the developed inequality is valid for any λ_n, thus we can tune λ_n to improve the convergence rate:

$$\varepsilon_n^* = \inf_{\lambda_n} A_0\lambda_n^s + cUK_0(n\lambda_n)^{-1/2} = O(n^{-s/(1+2s)}),$$

obtained by $\lambda_n = O(n^{-1/(1+2s)})$. Therefore, the convergence rate is given as:

$$\mathscr{E}(\widehat{f}_n) = O_P(\varepsilon_n^*) = O_P(n^{-\frac{s}{1+2s}}).$$

Remark 4.3. Note that the result is a simple but loose bound for the excess risk, we can improve the convergence rate from $n^{-s/(1+2s)}$ to $n^{-2s/(1+2s)}$ in the next few lectures. Yet, we are looking for where there is potential for improvement...

References

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

[Gönen and Alpaydın, 2011] Gönen, M. and Alpaydın, E. (2011). Multiple kernel learning algorithms. *The Journal of Machine Learning Research*, 12:2211–2268.

[Kimeldorf and Wahba, 1970] Kimeldorf, G. S. and Wahba, G. (1970). A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2):495–502.

[Sejdinovic and Gretton, 2012] Sejdinovic, D. and Gretton, A. (2012). What is an rkhs? *Lecture Notes*.

[Wahba, 1990] Wahba, G. (1990). *Spline models for observational data*. SIAM.