Mathematical Introduction to Machine Learning

Lecture 15: Theoretical foundation of kernel methods

December 12, 2023

Lecturer: Lei Wu *Scribe: Lei Wu*

Reading Guideline. When reading this lecture note (except for Section 5 and 6), you should focus on understanding concepts instead of being obsessed with the technical details. To help you understand concepts, ask yourself the following questions:

- What is the motivation to define RKHS?
- What is RKHS? What is the reproducing property?
- What is the Moore-Aronsajn's explicit construction of inner-product and the connection with KRR?
- What is the Mercer decomposition? What is the spectral/Mercer representation of RKHS?

You also need to master the proofs for the generalization analysis of KRR.

1 Functional Analysis Background

We will make use of a few concepts from functional analysis and here we review what we need.

Definition 1.1 (Function Space)**.** Let $\mathcal{X}$ be the input domain. A function space $\mathcal{F}$ is a space whose elements are functions, e.g. $f : \mathcal{X} \to \mathbb{R}$. We will focus on linear spaces of functions in the sense that if $f, g \in \mathcal{F}$, then $r_1 f + r_2 g \in \mathcal{F}$ for any $r_1, r_2 \in \mathbb{R}$.

Definition 1.2. An inner product is a function $\langle \cdot, \cdot \rangle : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ that satisfies the following properties for every $f, g \in \mathcal{F}$:

1. Symmetric: $\langle f, g \rangle = \langle g, f \rangle$.
2. Linear: $\langle r_1 f_1 + r_2 f_2, g \rangle = r_1 \langle f_1, g \rangle + r_2 \langle f_2, g \rangle$ for any $r_1, r_2 \in \mathbb{R}$.
3. Positive-definite: $\langle f, f \rangle \geq 0$ for all $f \in \mathcal{F}$ and $\langle f, f \rangle = 0$ iff $f = 0$.

Definition 1.3. A norm is a nonnegative function $\| \cdot \| : \mathcal{F} \to \mathbb{R}$ such that for all $f, g \in \mathcal{F}$ and $\alpha \in \mathbb{R}$

- Positivity: $\|f\| \geq 0$ and $\|f\| = 0$ iff $f = 0$;
- Positive homogeneity: $\|\alpha f\| = |\alpha| \|f\|$.
- Triangular inequality: $\|f + g\| \leq \|f\| + \|g\|$;

Lemma 1.4. *Let $(\mathcal{F}, \langle \cdot, \cdot \rangle)$ be an inner product space. Let $\|f\| = \sqrt{\langle f, f \rangle}$. Then, $\| \cdot \|$ is a norm.*

Proof. It is trivial to verify the positivity and positive homogeneity. What we need is to verify the triangular inequality. Noting

$$\|f+g\|^2 = \langle f+g, f+g\rangle = \|f\|^2 + \|g\|^2 + 2\langle f, g\rangle$$
$$(\|f\| + \|g\|)^2 = \|f\|^2 + \|g\|^2 + 2\|f\|\|g\|,$$

we only need to verify the Cauchy-Schwartz inequality

$$\langle f, g\rangle \leq \|f\|\|g\|.$$

To this end, consider

$$\begin{aligned}\|f+\lambda g\|^2 &= \|f\|^2 + 2\lambda\langle f, g\rangle + \lambda^2\|g\|^2 \\ &= \left(\|f\| + \lambda\frac{\langle f, g\rangle}{\|f\|}\right)^2 + \lambda^2\left(\|g\|^2 - \frac{\langle f, g\rangle}{\|f\|}\right)\end{aligned}$$

As the above quantity is non-negative for any $\lambda \in \mathbb{R}$. We must have

$$\|g\|^2 - \frac{\langle f, g\rangle}{\|f\|} \geq 0,$$

which establishes the Cauchy-Schwartz inequality, thereby the triangular inequality.

□

Note that while the dot product in $\mathbb{R}^d$ is an excellent example, an inner product is more general than this, and requires only those properties given above.

Definition 1.5. A Hilbert space is a **complete**, (possibly) infinite-dimensional linear space endowed with an inner product. Let $\mathcal{H}$ be a Hilbert space. Denote by $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ and $\|\cdot\|_{\mathcal{H}}$ the associated inner product and norm.

The most popular finite-dimensional Hilbert space is the Euclidean space $\mathbb{R}^d$ equiped with the $\langle x, y\rangle = \sum_{i=1}^d x_i y_i$. Another popular Hilbert space is $L^2_\mu(\mathcal{X})$ induced by the inner product

$$\langle f, g\rangle_{L^2_\mu(\mathcal{X})} = \int f(x)g(x)\,\mathrm{d}\mu(x) = \mathbb{E}_{x\sim\mu}[f(x)g(x)],$$

where μ is a probability distribution over $\mathcal{X}$.

While this tells us what a Hilbert space is, it is not intuitively clear why we need this mechanism, or what we gain by using it. Essentially, a Hilbert space lets us apply concepts from finite-dimensional linear algebra to infinite-dimensional spaces of functions. In particular, the fact that a Hilbert space is complete will guarantee the convergence of certain algorithms. More importantly, the presence of an inner product allows us to make use of orthogonality and projections, which will later become important.

Definition 1.6 (Linear functional)**.** Let $\mathcal{F}$ be a linear function space. $A : \mathcal{F} \mapsto \mathbb{R}$ is said to be a linear functional if for any $\alpha, \beta \in \mathbb{R}$ and $f, g \in \mathcal{F}$, we have

$$A(\alpha f + \beta g) = \alpha A(f) + \beta A(g)$$

Definition 1.7. Let $(\mathcal{F}, \|\cdot\|_{\mathcal{F}})$ be a normed function space. A linear functional $A : \mathcal{F} \mapsto \mathbb{R}$ is said to be continuous if there exist a constant $C > 0$ such that for any $f, g \in \mathcal{F}$, we have

$$|A(f-g)| \leq C\|f-g\|_{\mathcal{F}}.$$

The norm of A is defined by

$$\|A\| = \sup_{\|f\|_{\mathcal{F}} \leq 1} |A(f)|.$$

Obviously,

$$|A(f-g)| \leq \|A\|\|f-g\|_{\mathcal{F}}.$$

Theorem 1.8 (Riesz representation theorem). *Suppose $\mathcal{H}$ to be a Hilbert space. For any continuous linear functional A, there exist a unique vector $f_A \in \mathcal{H}$, called the Riesz representation of A, such that*

$$A(g) = \langle f_A, g\rangle_{\mathcal{H}} \qquad \forall g \in \mathcal{H}.$$

2 Recap of Kernel Ridge Regression

Kernel Ridge Regression (KRR) can be intepreted as a ridge regression on the feature space. Let $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ be a kernel. Assume that $\phi : \mathcal{X} \mapsto \mathcal{H}$ is the associated feature map with the Hilbert space $\mathcal{H}$ to be feature space. Then,

$$k(x, x') = \langle \phi(x), \phi(x')\rangle_{\mathcal{H}}.$$

Let $f_\beta(x) = \langle \beta, \phi(x)\rangle$ be our model. The feature space ridge regression is given by

$$\hat{\beta} = \operatorname*{argmin}_{\beta \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}(f_\beta(x_i) - y_i)^2 + \lambda\|\beta\|_{\mathcal{H}}^2.$$

By the representer theorem, there exist $\alpha \in \mathbb{R}^n$ such that

$$\hat{\beta} = \sum_{i=1}^{n} \alpha_i \varphi(x_i), \qquad f_{\hat{\beta}}(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x).$$

Hence, in KRR, we only need to consider the model

$$h_\alpha(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x)$$

without need to access the features $\{\varphi(x_i)\}_i$.

In general, the hypothesis class of kernel method is given by

$$\mathcal{H}^0 = \left\{\sum_{i=1}^{n} \alpha_i k(x_i, \cdot) : n \in \mathbb{N}, \alpha \in \mathbb{R}^n \text{ and } x_i \in \mathcal{X} \text{ for } i = 1, \ldots, n\right\} \tag{1}$$

3 Reproducing kernel Hilbert spaces

3.1 Motivation

Our question is: what kind of functions can be learned efficiently by KRR. We anticipate functions in $\mathcal{H}^0$ should be easy to learn? But what norm of a target function f determines the complexity of learning it?

Consider the feature-based representation:

$$\beta = \sum_{j=1}^{n} a_j \Phi(x_j), \qquad f(x;\beta) = \sum_{j=1}^{n} a_j k(x_j, \cdot).$$

In KRR, we penalize the $\|\beta\|^2$ norm. This means that $\|\beta\|^2$ should be a good norm of the represented function $f(x;\beta)$, i.e,

$$\|f(\cdot;\beta)\|^2 = \|\beta\|^2 = \left\langle \sum_{i=1}^{n} a_i \Phi(x_i), \sum_{j=1}^{n} a_j \Phi(x_j) \right\rangle = \sum_{i,j=1}^{n} a_i a_j k(x_i, x_j).$$

This intuition can be made rigorous by the following theorem.

Theorem 3.1 (Moore-Aronsajn theorem, the Moore-Aronsajn definition of a RKHS)**.** *Let $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ be any SPD kernel. Let $\mathcal{H}^0$ be defined in* (1) *and endow it with the inner product:*

$$\langle f, g \rangle_{\mathcal{H}^0} = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(x_i, x'_j), \tag{2}$$

where $f = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i), g = \sum_{j=1}^{m} \beta_j k(\cdot, x'_j)$. Then, $\mathcal{H}^0$ is a valid pre-Hilbert space. This means that the pointwise closure $\mathcal{H}_k = \overline{\mathcal{H}^0}$ is a Hilbert space.

The above theorem construct an inner product in $\mathcal{H}^0$ and we will show later that the induced norm can control the learning complexity. Moreover, this theorem show that we can consider the closure of $\mathcal{H}^0$, i.e., $\mathcal{H}_k$, which contains more functions than $\mathcal{H}^0$.

Proof. We show that (2) indeed defines a valid inner product. First,

$$\langle f, g \rangle_{\mathcal{H}^0} = \sum_{i=1}^{n} \alpha_i g(x_i) = \sum_{j=1}^{n} \beta_j f(x'_j).$$

It is implied that that the **inner product is independent of the specific representation of** f **and** g. The triangular inequality is easy to verify. Next, we show that $\|f\|_{\mathcal{H}^0} = 0$ if and only if $f = 0$. If there exist $x_0 \in \mathcal{X}$ such that $f(x_0) \neq 0$. Assume $f(x) = \sum_{j=1}^{m} a_j k(x_j, \cdot)$ and consider

$$0 \leq \|\lambda f + f(x_0) k(\cdot, x_0)\|^2_{\mathcal{H}^0} = \lambda^2 \|f\|^2_{\mathcal{H}^0} + 2\lambda f^2(x_0) + f^2(x_0) k(x_0, x_0).$$

Taking $\lambda \to -\infty$, the RHS will be negative and this causes contradictory.

What remains is to show that the convergence of Cauchy sequence [1]. We refer to Link for a complete proof. □

[1] You can skip the verification of completeness.

Lemma 3.2 (Reproducing property)**.** *The Hilbert space $\mathcal{H}_k$ defined in Theorem 3.1 satisfies the reproducing property:*

$$\langle f, k(\cdot, x)\rangle_{\mathcal{H}_k} = f(x).$$

Proof. For $f \in \mathcal{H}^0$, we can write $f(x) = \sum_{j=1}^m a_j k(\cdot, x_j)$. By definition,

$$\langle f, k(\cdot, x)\rangle_{\mathcal{H}_k} = \sum_{j=1}^m a_j k(x, x_j) = f(x).$$

For any $f \in \mathcal{H}_k$, let $\lim_{n\to\infty} f_n(x) = f(x)$. Then,

$$\langle f, k(\cdot, x)\rangle_{\mathcal{H}_k} = \lim_{n\to\infty} \langle f_n, k(\cdot, x)\rangle_{\mathcal{H}_k} = \lim_{n\to\infty} f_n(x) = f(x).$$

□

3.2 Definition of RKHS

The reproducing property in Lemma 3.2 plays a fundamental role in the Hilbert space $\mathcal{H}_k$. In fact, the reproducing kernel Hilbert space (RKHS) can be completely determined by this property.

Definition 3.3 (RKHS)**.** Let $\mathcal{X}$ be an arbitrary set and $\mathcal{H}$ a Hilbert space of real-valued functions on $\mathcal{X}$. We say $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) if there is a kernel $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ such that

- $\forall x \in \mathcal{X}$, $k(\cdot, x) \in \mathcal{H}$.
- *Reproducing property:* $\forall x \in X, f \in \mathcal{H}$, $\langle f, k(\cdot, x)\rangle_{\mathcal{H}} = f(x)$

Lemma 3.4. *For a RKHS, the reproducing kernel k is unique.*

Proof. For any two reproducing kernels k_1, k_2, we have

$$\langle f, k_1(\cdot, x) - k_2(\cdot, x)\rangle_{\mathcal{H}} = f(x) - f(x) = 0, \forall x \in X, \forall f \in \mathcal{H}.$$

Taking $f = k_1(\cdot, x) - k_2(\cdot, x)$, we have $\|k_1(\cdot, x) - k_2(\cdot, x)\|_{\mathcal{H}}^2 = 0, \forall x \in X$. Hence, $k_1 = k_2$. □

Theorem 3.5. *For any kernel k, there is a unique RKHS, for which k is the corresponding reproducing kernel. Moreover, this unique RKHS is $\mathcal{H}_k$, i.e., the one constructed in Moore-Aronsajn theorem.*

You only need to know this theorem and can ignore the proof below.

Proof. First, by Moore-Aronsajn theorem and Lemma 3.2, there exists a RKHS with k being the reproducing kernel. Assume $\mathcal{H}_1$ and $\mathcal{H}_2$ be two RKHSs with k being the reproducing kernel. First, by definition, $k(\cdot, x) \in \mathcal{H}_1$ for any $x \in \mathcal{X}$. Hence, $\mathcal{H}^0 \subset \mathcal{H}_1$. Moreover, $\mathcal{H}^0$ is dense in $\mathcal{H}_1$ since if there exists $f \in \mathcal{H}$ such that $f \perp \mathcal{H}^0$, we must have

$$\langle f, k(\cdot, x)\rangle_{\mathcal{H}_1} = f(x) = 0 \qquad \forall x \in \mathcal{X}.$$

For $f = \sum_{j=1}^m a_j k(\cdot, x_j)$,

$$\|f\|_{\mathcal{H}_1}^2 = \left\langle \sum_{i}^n a_i k(\cdot, x_i), \sum_{j=1}^m a_j k(\cdot, x_j) \right\rangle_{\mathcal{H}_1} = \sum_{i,j=1}^n a_i a_j \langle k(\cdot, x_i), k(\cdot, x_j)\rangle_{\mathcal{H}_1}$$

$$\overset{(i)}{=} \sum_{i,j=1}^{n} a_i a_j k(x_i, x_j) = \|f\|_{\mathcal{H}^0}^2.$$

where (i) follows from the reproducing property. Hence, $\|f\|_{\mathcal{H}_1} = \|f\|_{\mathcal{H}^0}$ for $f \in \mathcal{H}_0$. By the same argument, the same results hold for $\mathcal{H}_2$. For any $f \in \mathcal{H}_1$, there must exits $(f_n) \subset \mathcal{H}^0$ such that $f(x) = \lim_{n\to\infty} f_n(x)$. This implies that $f \in \mathcal{H}_2$. Similarly, $\mathcal{H}_1$ and $\mathcal{H}_2$ contains the same functions. What remains is to check that the two norms coincide, which results from

$$\|f\|_{\mathcal{H}_1} = \lim_{n\to\infty} \|f_n\|_{\mathcal{H}_1} = \lim_{n\to\infty} \|f_n\|_{\mathcal{H}^0} = \lim_{n\to\infty} \|f_n\|_{\mathcal{H}_2} = \|f\|_{\mathcal{H}_2}.$$

□

4 The perspective of evaluation functional

You can skip this section when reading this lecture note.

Definition 4.1. For any $x \in \mathcal{X}$, the evaluation functional $L_x : \mathcal{F} \mapsto \mathbb{R}$ is defined by

$$L_x(f) = f(x).$$

Lemma 4.2. *For a RKHS $\mathcal{H}$, the evaluation functional $L_x : \mathcal{H} \mapsto \mathbb{R}$ is continuous.*

Proof. For any $x \in X$ and $f, g \in \mathcal{H}$,

$$|L_x(f) - L_x(g)| = f(x) - g(x) = \langle f - g, k(x, \cdot)\rangle_{\mathcal{H}_k} \leq \|f - g\|_{\mathcal{H}_k} \|k(x, \cdot)\|_{\mathcal{H}_k},$$

where the last step follows from the Cauchy-Schwartz inequality. □

An important implication is that the convergence in norm implies the pointwise convergence. If $\lim_{n\to\infty} \|f_n - f\|_{\mathcal{H}} = 0$, then

$$|f_n(x) - f(x)| \leq \|L_x\| \|f_n - f\|_{\mathcal{H}} \to 0 \qquad \text{as } n \to \infty.$$

This is a major difference between a RKHS and a general Hilbert space. For instance, for the $L^2_\mu(\mathcal{X})$ space, the norm convergence does not imply the pointwise convergence.

This continuity of the evaluation functional is sometimes used as an equivalent definition of RKHS.

Theorem 4.3. *A Hilbert space of functions $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ is a RKHS if and only if the evaluation functional is continuous.*

Proof. If L_x is continuous, by Riesz representation theorem, there exist $K_x \in \mathcal{H}$ such that

$$L_x(f) = \langle K_x, f\rangle_{\mathcal{H}}.$$

Define the kernel:

$$k(x, x') = \langle K_x, K_{x'}\rangle_{\mathcal{H}} = K_{x'}(x) = K_x(x'),$$

for which

$$\langle f, k(\cdot, x)\rangle_{\mathcal{H}} = \langle f, K_x\rangle = f(x), \quad \forall f \in \mathcal{H}.$$

This means $k(\cdot, \cdot)$ is a reproducing kernel of $\mathcal{H}$. □

5 A spectral perspective of RKHS

For a kernel $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$, we define an integral operator $\mathcal{T}_k : L^2_\mu(\mathcal{X}) \mapsto L^2_\mu(\mathcal{X})$ as follows

$$\mathcal{T}_k f(x) = \mathbb{E}_{x' \sim \mu}[k(x, x') f(x')]$$

Theorem 5.1 (Mercer's theorem)**.** *Let k be a continuous kernel on a* ***compact*** *set $\mathcal{X}$. There exist an orthonormal basis $\{e_j\}_{j=1}^\infty$ of $L^2_\mu(\mathcal{X})$ such that $\forall x, x' \in \mathcal{X}$,*

$$k(x, x') = \sum_{j=1}^{\infty} \lambda_j e_j(x) e_j(x').$$

The convergence is uniform on $\mathcal{X} \times \mathcal{X}$ and absolute for each $(x, x') \in \mathcal{X} \times \mathcal{X}$.

This theorem ensures the existence of eigen decomposition of a kernel k, i.e., the corresponding integral operator $\mathcal{T}_k$. Note that $(\lambda_j)_{j \geq 1}$ and $(e_j)_{j \geq 1}$ are the eigenvalues and eigenfunctions of the integral operator $\mathcal{T}_k$ in the sense that

- $\mathcal{T}_k e_j = \lambda_j e_j$, i.e., $\mathbb{E}_{x' \sim \mu}[k(x, x') e_j(x')] = \lambda_j e_j(x)$.
- $\langle e_i, e_j \rangle_{L^2_\mu(\mathcal{X})} = \mathbb{E}_{x \sim \mu}[e_j(x) e_i(x)] = \delta_{i,j}$.

Feature map. Mercer's theorem gives a feature map for the kernel k. Let

$$\Phi : \mathcal{X} \mapsto \ell^2, \qquad \Phi(x) = \left(\sqrt{\lambda_1} e_1(x), \sqrt{\lambda_2} e_2(x), \ldots, \sqrt{\lambda_j} e_j(x), \ldots\right)^T.$$

Then,

$$k(x, x') = \sum_{j=1}^{\infty} \sqrt{\lambda_j} e_j(x) \sqrt{\lambda_j} e_j(x') = \langle \Phi(x), \Phi(x') \rangle_{\ell^2}.$$

Theorem 5.2 (Spectral representation of RKHS)**.** *Let k be a continuous kernel on a compact set $\mathcal{X}$, and $\{e_j\}$ be the orthonormal basis given in Mercer's theorem. Define*

$$\mathcal{H} = \left\{ f = \sum_j a_j e_j : \sum_j \frac{a_j^2}{\lambda_j} < \infty \right\},$$

with the inner product

$$\left\langle \sum_j a_j e_j, \sum_j b_j e_j \right\rangle_{\mathcal{H}} = \sum_j \frac{a_j b_j}{\lambda_j}.$$

Then, $\mathcal{H}$ is the RKHS $\mathcal{H}_k$.

Proof. By Mercer's theorem, $k(\cdot, x) = \sum_j (\lambda_j e_j(x)) e_j$. Hence,

$$\|k(\cdot, x)\|^2_{\mathcal{H}} = \sum_j \frac{\lambda_j^2 e_j(x)^2}{\lambda_j} = \sum_j \lambda_j e_j(x) e_j(x) = k(x, x) < \infty.$$

So, $k(\cdot, x) \in \mathcal{H}$ for any $x \in X$.

Reproducing property: Let $f = \sum_j a_j e_j \in \mathcal{H}$. Then,

$$\langle f, k(\cdot, x)\rangle_{\mathcal{H}} = \sum_j \frac{a_j \lambda_j e_j(x)}{\lambda_j} = f(x). \tag{3}$$

Hence, $\mathcal{H}$ is a RKHS with the reproducing kernel k. By the uniqueness of RKHS, $\mathcal{H} = \mathcal{H}_k$. □

Remark 5.3. Note that the integral operator $\mathcal{T}_k$ and the associated eigenfunctions $\{e_j\}$ depend on the underlying distribution μ. However, $\mathcal{H}$ coincides with the RKHS $\mathcal{H}_k$. This means that $\mathcal{H}$ actually does not depend on the choice of μ at all.

Weighted L^2 space. In this way, RKHS can be viewed as a L^2 space weighted by the eigenvalues.

- L^2 space: $\|f\|^2_{L^2_\mu(\mathcal{X})} = \sum_{j=1}^\infty a_j^2$.
- RKHS/Weighted L^2 space: $\|f\|^2_{\mathcal{H}_k} = \sum_{j=1}^\infty \frac{a_j^2}{\lambda_j}$.

Hence, the faster the eigenvalue decay is, the smaller the RKHS is. Consider $\lambda_j = \frac{1}{j^s}$. Then,

$$\|f\|^2_{\mathcal{H}} = \sum_{j=1}^\infty j^s a_j^2 < \infty \quad \overset{roughly}{\Longrightarrow} \quad a_j^2 = O(\frac{1}{j^{s+1+\alpha}}) \quad \text{for some } \alpha > 0,$$

A larger s leads to a faster the decay of the coefficients.

6 A generalization analysis of kernel ridge regression

We first provide the upper bound of the Rademacher complexity.

Proposition 6.1. *For any kernel k, let $\mathcal{H}_k$ the corresponding RKHS. Let $\mathcal{H}_k^Q = \{f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} \leq Q\}$. Then, we have*

$$\widehat{\mathrm{Rad}}_n(\mathcal{H}_k^Q) \leq \frac{Q}{\sqrt{n}} \sqrt{\frac{1}{n}\sum_{i=1}^n k(x_i, x_i)}.$$

Proof.

$$\begin{aligned} n\widehat{\mathrm{Rad}}_n(\mathcal{H}_k^Q) &= \mathbb{E}_\xi[\sup_{\|f\|_{\mathcal{H}_k}\leq Q} \sum_{i=1}^n \xi_i f(x_i)] = \mathbb{E}_\xi[\sup_{\|f\|_{\mathcal{H}_k}\leq Q} \sum_{i=1}^n \xi_i \langle f, k(\cdot, x_i)\rangle_{\mathcal{H}_k}] \text{(reproducing property)} \\ &= \mathbb{E}_\xi[\sup_{\|f\|_{\mathcal{H}_k}\leq Q} \langle f, \sum_{i=1}^n \xi_i k(\cdot, x_i)\rangle_{\mathcal{H}}] \leq Q\,\mathbb{E}_\xi[\|\sum_{i=1}^n \xi_i k(\cdot, x_i)\|_{\mathcal{H}_k}] \\ &= Q\,\mathbb{E}_\xi\sqrt{\sum_{i,j=1}^n \xi_i\xi_j k(x_i, x_j)} \leq Q\sqrt{\mathbb{E}_\xi[\sum_{i,j=1}^n \xi_i\xi_j k(x_i, x_j)]} \qquad \text{(Jensen inequality)} \end{aligned}$$

$$= Q\sqrt{\sum_{i=1}^{n} k(x_i, x_i)} \qquad (\mathbb{E}[\xi_i \xi_j] = 0, \forall\, i \neq j).$$

□

Given data $\{(x_i, f^*(x_i))\}_{i=1}^n$, consider the kernel ridge regression (KRR) estimator

$$\hat{f}_n = \operatorname*{argmin}_{f \in \mathcal{H}_k} \hat{\mathcal{R}}(f) + \lambda \|f\|_{\mathcal{H}_k}. \tag{4}$$

Theorem 6.2 (A priori estimate)**.** *Assume that $\ell(\cdot, y)$ is L-Lipschitz and bounded by B, and $\sup_{x\in\mathcal{X}} k(x,x) \leq 1$. Then, for any $\delta \in (0,1)$, with probability $1-\delta$ over the choice of training set, we have*

$$\mathcal{R}(\hat{f}_n) \lesssim \lambda \|f^*\|_{\mathcal{H}_k} + \frac{L\|f^*\|_{\mathcal{H}_k}}{\sqrt{n}} + B\sqrt{\frac{\log(1/\delta)}{n}}.$$

Proof. (1) Let $Q = \|f^*\|_{\mathcal{H}_k}$. By the definition of $\hat{f}_n$,

$$\hat{\mathcal{R}}(\hat{f}_n) + \lambda\|\hat{f}_n\|_{\mathcal{H}_k} \leq \hat{\mathcal{R}}(f^*) + \lambda\|f^*\|_{\mathcal{H}_k} = \lambda\|f^*\|_{\mathcal{H}_k} = \lambda Q,$$

which yields

$$\|\hat{f}_n\|_{\mathcal{H}_k} \leq Q, \qquad \hat{\mathcal{R}}(\hat{f}_n) \leq \lambda Q.$$

(2) Let $\mathcal{F}_Q = \{\ell(h(x), h^*(x)) : h \in \mathcal{H}_k^Q\}$. By the contraction property of Rademacher complexity, we have

$$\widehat{\mathrm{Rad}}_n(\mathcal{F}_Q) \leq L\widehat{\mathrm{Rad}}_n(\mathcal{H}_k^Q).$$

Using the Rademacher complexity-based generalization bound, we have

$$\begin{aligned}
|\hat{\mathcal{R}}(\hat{f}_n) - \mathcal{R}(\hat{f}_n)| &\leq \sup_{\|f\|_{\mathcal{H}} \leq Q} |\hat{\mathcal{R}}(f) - \mathcal{R}(f)| \lesssim \widehat{\mathrm{Rad}}_n(\mathcal{F}_Q) + B\sqrt{\frac{\log(4/\delta)}{n}} \\
&\lesssim L\widehat{\mathrm{Rad}}_n(\mathcal{H}_k^Q) + B\sqrt{\frac{\log(4/\delta)}{n}} \leq \frac{LQ}{\sqrt{n}} + B\sqrt{\frac{\log(1/\delta)}{n}} \quad (\text{use} \sup_{x\in X} k(x,x) \leq 1).
\end{aligned}$$

(3) $\mathcal{R}(\hat{f}_n) \leq \hat{\mathcal{R}}(\hat{f}_n) + |\hat{\mathcal{R}}(\hat{f}_n) - \mathcal{R}(\hat{f}_n)| \leq \lambda Q + (LQ + B\sqrt{\log(4/\delta)})/\sqrt{n}$.

□

The preceding estimate is a priori in the sense that it depends on the norm of f^* instead of that of $\hat{f}_n$. Taking $\lambda = O(1/\sqrt{n})$, we have that $\mathcal{R}(\hat{f}_n) = O(1/\sqrt{n})$, which does not suffer from the curse of dimensionality. This means that the functions in the RKHS can be efficiently learned by KRR.

- Similar results hold for any regularizations of the form $r(\|f\|_{\mathcal{H}_k})$, where $r : [0,\infty) \to [0,\infty)$ is strictly increasing.
- Note that Theorem 6.2 holds as long as $\lambda > 0$ and one can even take $\lambda \to 0^+$, which may seem strange at the first glance. This is due to that there is no label noise. In fact, the optimal λ depends on the noise level but we will not discuss the influence of noise in this lecture.

Tightness. It is worth noting that preceding bounds are not tight for the square loss: $\ell(y_1, y_2) = (y_1 - y_2)^2$. When applying the contraction lemma, we use the worst-case Lipschitz norm $\mathrm{Lip}(t^2/2) \leq 1$ for $t \in [0, 1]$. However, around the estimator, we should have $\varepsilon(x) = \hat{f}(x) - f^*(x) \ll 1$. Therefore, we should use the "local" Lipschitz norm to bound the Rademacher complexity. This will in turn gives rise to a fast rate. Usually, the fast rate is close to $O(1/n)$ and this approach is called "local Rademacher complexity". We refer interested readers to [Bartlett et al., 2005, Srebro et al., 2010] for more details.

7 Final Remark

- In this lecture note, we only present the most important aspects of RKHS and illustrate how to analyze KRR with RKHS. Note that RKHS theory plays a critical role in the mathematical analysis of kernel methods (not limited to KRR). If you would like to become an expert in RKHS, we refer to [Wainwright, 2019, Section 12].
- However, it should be stressed that RKHS can be learned efficiently does not imply that there are no other larger function space that can be also learned efficiently. However, RKHS is the one with Hilbert structure and the reproducing property, which makes the mathematical analysis simple and elegant.

References

[Bartlett et al., 2005] Bartlett, P. L., Bousquet, O., Mendelson, S., et al. (2005). Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537.

[Srebro et al., 2010] Srebro, N., Sridharan, K., and Tewari, A. (2010). Smoothness, low noise and fast rates. *Advances in neural information processing systems*, 23.

[Wainwright, 2019] Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.