Theoretical Deep Learning

Lecture 4: Kernel methods, representer theorem and RKHSs

July 28, 2021

Lecturer: Lei Wu *Scribe: Lei Wu*

1 Feature-based methods

We start by considering the linear regression, for which the hypothesis class is

$$\mathcal{F} = \{\beta^T x : \beta \in \mathbb{R}^d\},$$

where we omit the bias term for simplicity. The ridge regression penalizes the squared ℓ_2 norm of β:

$$\hat{\beta}_n = \operatorname*{argmin}_{\beta\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n (\beta^T x_i - y_i)^2 + \lambda\|\beta\|_2^2.$$

The minimizer has a closed-form solution:

$$\hat{\beta}_n = \left(\frac{1}{n}X^T X + I\right)^{-1}\frac{1}{n}Xy,$$

where $X = (x_1, \ldots, x_n)^T \in \mathbb{R}^{n\times d}$, $y = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$. Another population one is LASSO, which penalizes the ℓ_1 norm of parameters:

$$\min_{\beta\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n (\beta^T x_i - y_i)^2 + \lambda\|\beta\|_1.$$

To consider nonlinear functions, we can consider the model:

$$f(x;\beta) = \sum_{j=1}^m \beta_j \varphi_j(x).$$

Here, $\varphi_1, \ldots, \varphi_n$ are a set of (nonlinear) basis functions, which are often referred to as *features* in machine learning. Accordingly, the feature map is defined as $\Phi : \mathcal{X} \mapsto \mathbb{R}^m$ with $\Phi(x) = (\varphi_1(x), \ldots, \varphi_n(x))^T \in \mathbb{R}^m$. Typical examples includes

- Spectral methods: $\{\varphi_j\}$ are either Fourier basis or orthogonal polynomials.
- Splines: $\{\varphi_j\}$ are piecewise polynomials.
- Computer vision: Some hand-crafted features.

We can consider two types of feature-based methods.

$$\frac{1}{n}\sum_{i=1}^n (\beta^T \Phi(x_i) - y_i)^2 + \lambda\|\beta\|_2^2$$
$$\frac{1}{n}\sum_{i=1}^n (\beta^T \Phi(x_i) - y_i)^2 + \lambda\|\beta\|_1.$$

1.1 General feature-based methods

ℓ_2 extension. The previous idea can be extended to a general feature-based model:

$$f(x;\beta) = \langle \beta, \Phi(x) \rangle_{\mathcal{H}}, \tag{1.1}$$

where

- $\mathcal{H}$ is the feature space, which can be any Hilbert space;
- $\Phi : \mathcal{X} \mapsto \mathcal{H}$ is the feature map;
- The "coefficients" are $\beta \in \mathcal{H}$.

If taking $\mathcal{H} = \mathbb{R}^m$ and $\Phi(x) = (\phi(x_1), \dots, \phi(x_n))^T \in \mathbb{R}^m$, we recover the classical ones. However, the advantage of the formulation (1.1) is that it includes the case where $m = \infty$. Below is an example:

Random feature models (RFMs). Consider

$$f(x;\beta) = \int \beta(w)\varphi(x;w)\,\mathrm{d}\pi(w) = \langle \beta, \varphi(x;\cdot) \rangle_{L^2(\pi)}, \tag{1.2}$$

where π is a fixed distribution. In this case, the feature map is given by

$$\Phi : \mathcal{X} \mapsto L^2(\pi), \qquad \Phi(x) = \varphi(x;\cdot),$$

and the parameter is $\beta \in L^2(\pi)$. The model (1.2) can be viewed as the continuum limit of the following random feature model

$$f(x;\beta) = \frac{1}{m}\sum_{j=1}^{m} \beta_j \varphi(x;w_j),$$

where $w_1, \dots, w_m$ are independently sampled from π and fixed.

Now, the model (1.1) is well-defined. The objective function of the corresponding ridge regression can be written as

$$\hat{\mathcal{R}}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} (\langle \beta, \Phi(x_i) \rangle_{\mathcal{H}} - y_i)^2 + \lambda \|\beta\|_{\mathcal{H}}^2. \tag{1.3}$$

How can we optimize (1.3), which is an infinitely dimensional problem?

1.2 ℓ_1 extension.

Consider the random feature methods:

$$\frac{1}{n}\sum_{i=1}^{n} \left(\sum_{j=1}^{m} \beta_j \varphi(x_i;w_j) - y_i \right)^2 + \frac{\lambda}{m}\sum_{j=1}^{m} |\beta_j|.$$

Assume for any $x \in \mathcal{X}$, $\operatorname{ess\,sup}_w |\varphi(x;w)| < \infty$. Then, the continuum limit of the above method is given by

$$\min_{\beta \in L^1(\pi)} \frac{1}{n}\sum_{i=1}^{n} \left(\int \beta(w)\varphi(x_i;w)\,\mathrm{d}\pi(w) - y_i \right)^2 + \lambda \int |\beta(w)|\,\mathrm{d}\pi(w).$$

This method can not be analyzed using the kernel theory.

2 Representer theorem and kernel methods

When it is clear from the context, we will drop the subscripts in $\langle , \rangle_{\mathcal{H}}$ and $\| \cdot \|_{\mathcal{H}}$ for simplicity. Let us consider a general problem:

$$\hat{R}_n(\beta) = \frac{1}{2n} \sum_{i=1}^{n} \ell(f(x_i; \beta), y_i) + \lambda r(\|\beta\|), \tag{2.1}$$

- $f(x; \beta) = \langle \beta, \Phi(x) \rangle$
- ℓ is a general loss function.
- $r : [0, \infty) \mapsto [0, \infty)$ is a strictly increasing function.

Theorem 2.1 (Representer theorem)**.** *Let $\hat{\beta}$ the a minimizer of (2.1). Then, there must exist $a_1, \ldots, a_n \in \mathbb{R}$ such that $\hat{\beta} = \sum_{i=1}^{n} a_i \Phi(x_i)$ and*

$$f(x; \hat{\beta}) = \langle \hat{\beta}, \Phi(x) \rangle = \sum_{i=1}^{n} a_i k(x_i, x), \tag{2.2}$$

where $k(x, x') := \langle \Phi(x), \Phi(x') \rangle$.

Proof. Let $V_n = \text{span}\{\Phi(x_1), \ldots, \Phi(x_n)\} \subset \mathcal{H}$. For any $\beta \in \mathcal{H}$, we can decompose it as follows

$$\beta = \beta_{\|} + \beta_{\perp},$$

where $\beta_{\|} \in V_n, \beta_{\perp} \in V_n^{\perp}$. Hence, $\|\beta\|^2 = \|\beta_{\|}\|^2 + \|\beta_{\perp}\|^2$. Since $r(\cdot)$ is non-decreasing,

$$r(\|\beta\|) \geq r(\|\beta_{\|}\|). \tag{2.3}$$

On the other hand, for any x_i,

$$f(x_i; \beta) = \langle \beta, \Phi(x_i) \rangle = \langle \beta_{\|}, \Phi(x_i) \rangle + \langle \beta_{\perp}, \Phi(x_i) \rangle = \langle \beta_{\|}, \Phi(x_i) \rangle, \tag{2.4}$$

where the last equality is due to $\beta_{\perp} \in V_n^{\perp}$. Combining (2.3) and (2.4), we have for any $\beta \in \mathcal{H}$,

$$\hat{\mathcal{R}}_n(\beta) \geq \hat{\mathcal{R}}_n(\beta_{\|}).$$

Therefore, we can take $\hat{\beta}_{\|} = \sum_{i=1}^{n} a_i \Phi(x_i)$. Then, the function represented can be written as

$$f(x; \beta) = \langle \hat{\beta}_{\|}, \Phi(x) \rangle = \sum_{i=1}^{n} a_i \langle \Phi(x_i), \Phi(x) \rangle = \sum_{i=1}^{n} a_i k(x_i, x).$$

□

This theorem allows transforming the infinite-dimensional optimization problem (2.1) into a finite dimensional problem. Moreover, we only need to access the kernel $k(\cdot, \cdot)$ without needing to evaluate the feature maps.

The reduced model. Representer theorem implies that we only need to choose

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

Moreover,

$$\|\beta\|^2 = \langle \sum_{j=1}^{n} a_j \Phi(x_j), \sum_{j=1}^{n} a_j \Phi(x_j) \rangle = \sum_{i,j=1}^{n} k(x_i, x_j) a_i a_j = a^T K a,$$

where $a = (a_1, \ldots, a_n)^T \in \mathbb{R}^n$ and $K = (k(x_i, x_j)) \in \mathbb{R}^{n \times n}$ is the *kernel matrix*.

The kernel ridge regression (KRR) corresponds to the case where $\ell(y, y') = (y - y')^2$ and $r(t) = t^2$, i.e., the problem (1.3). Then, the problem can be reduced to the following n-dimensional problem

$$\hat{\mathcal{R}}_n(a) = \frac{1}{n} \|Ka - y\|_2^2 + \lambda a^T K a, \tag{2.5}$$

whose solution is given by

$$a = (\frac{1}{n} K + I)^{-1} y.$$

In general, kernel methods refer to methods whose hypothesis class is given by

$$\mathcal{F} = \left\{ \sum_{j=1}^{n} a_j k(x_j, \cdot) : a \in \mathbb{R}^n \right\}.$$

Mathematically, the kernel is defined as follows.

Definition 2.2 (kernel)**.** $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is said to be a kernel if there exists a feature map $\Phi : \mathcal{X} \mapsto \mathcal{H}$ such that

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle.$$

Below is a list of popular kernels.

Polynomial kernel: $k(x, x') = (1 + x^T x')^s$ is a kernel for any $s \in \mathbb{N}_+$.

- Linear ($s = 1$). We have $k(x, x') = \langle \Phi(x), \Phi(x') \rangle$ with

$$\Phi(x) = (1, x_1, \ldots, x_d).$$

- Quadratic ($s = 2$): The feature map is given by

$$\Phi(x) = (\underbrace{x_d^2, \ldots, x_1^2}_{\text{quadratic}}, \underbrace{\sqrt{2}x_d x_{d-1}, \ldots, \sqrt{2}x_d x_1, \sqrt{2}x_{d-1}x_{d-2}, \ldots, \sqrt{2}x_2 x_1}_{\text{cross terms}}, \underbrace{\sqrt{2}x_d, \ldots, \sqrt{2}x_1}_{\text{linear terms}}, \underbrace{1}_{\text{constant}}).$$

$$\begin{aligned} \langle \Phi(x), \Phi(x') \rangle &= \sum_{i=1}^{d} (x_i)^2 (x_i')^2 + 2 \sum_{i \neq j} x_i x_j x_i' x_j' + 2 \sum_i x_i x_i' + 1 \\ &= (\sum_{i=1}^{d} x_i x_i')^2 + 2 \sum_i x_i x_i' + 1 \\ &= (x^T x' + 1)^2 \end{aligned} \tag{2.6}$$

Gaussian kernel: $k(x, x') = e^{-\frac{\|x-x'\|_2^2}{2}}$. Considering $d = 1$, we have

$$k(x, x') = e^{-\frac{x^2}{2}-\frac{x'^2}{2}} e^{xx'} = e^{-\frac{x^2}{2}-\frac{x'^2}{2}} \sum_n \frac{1}{n!}(x)^n (x')^n$$
$$= \langle \Phi(x), \Phi(x) \rangle,$$

where $\Phi(x) = e^{-\frac{x^2}{2}}(1, x, \frac{1}{\sqrt{2}}x^2, \ldots, \frac{1}{\sqrt{n!}}x^n, \ldots)$.

Laplace kernel:

$$k(x, x') = e^{-\frac{\|x-x'\|_2}{\sigma}}.$$

This kernel is less smooth than the Gaussian kernel. Recently, it has been shown that the Laplace kernel is intimately related to neural network models in the kernel regime.

For a specific problem, choosing appropriate kernels is highly non-trivial. One may need to incorporate the domain knowledge into the kernel design.

3 Reproducing kernel Hilbert spaces

In this section, we ask the question:

What kind of functions can be "efficiently" learned by kernel methods?

By representer theorem, consider

$$\mathcal{F} = \cup_{n=1}^{\infty} \mathcal{F}_n,$$

where

$$\mathcal{F}_n = \left\{ \sum_{j=1}^{n} a_j k(\cdot, x_j) \,:\, x_j \in \mathcal{X}, a_j \in \mathbb{R}, j \in [n] \right\}.$$

This intuition tells us that what kind of functions can be "approximated" by kernel methods. We are interested in functions $f \in \bar{\mathcal{F}}$. However, the problem is how to take the closure and measure the complexity of $f \in \bar{\mathcal{F}}$? Without imposing constraints on the norm of coefficients $\{a_j\}$ in taking the closure, this space can be extremely large. For example, if the corresponding features are polynomials, then $\mathcal{F}$ contains all the continuous functions because of the Stone-Weierstrass theorem. However, $C(\mathcal{X})$ is too large since the Rademacher complexity is $O(1)$. We hope that the Rademacher complexity is on the order of $O(1/\sqrt{n})$.

We need to define an "appropriate" norm for $f \in \mathcal{F}$.

Let us take a step back to 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 $\|\beta\|^2$ of the hypothesis. This means that $\|\beta\|^2$ should be a good norm of the represented function $f(x; \beta)$, i.e,

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

This intuition can be made rigorous by the following theorem.

Theorem 3.1 (Moore-Aronsajn theorem)**.** *Let $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ be any kernel. Let $\mathcal{H}^0 = span(\{k(\cdot, x) : x \in \mathcal{X}\})$ and endow it with the inner product:*

$$\langle f, g\rangle_{\mathcal{H}^0} = \sum_{i=1}^{n}\sum_{j=1}^{m} \alpha_i\beta_j k(x_i, x'_j), \tag{3.1}$$

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, i.e, the pointwise closure $\mathcal{H}_k = \overline{\mathcal{H}^0}$ is a Hilbert space.

Proof. We show that (3.1) 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. We refer to Link for a complete proof. □

Lemma 3.2. *The Hilbert space 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,

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

□

The reproducing property is the most important property of this Hilbert space.

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 evaluation functional $L_x(f) = f(x)$ is continuous.*

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

$$\sup_{\|f\|_{\mathcal{H}} \leq 1} |L_x(f)| = \sup_{\|f\|_{\mathcal{H}} \leq 1} |\langle f, k(\cdot, x)\rangle_{\mathcal{H}}| \leq \|k(\cdot, x)\|_{\mathcal{H}} < \infty.$$

□

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

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

Lemma 3.5. *For a RKHS, the reproducing kernel k is unique.*

Proof. For any two kernels k_1, k_2,

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

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 3.6. *For any kernel k, there is a unique RKHS, for which k is the reproducing kernel.*

Proof. First, by Moore-Aronsajn theorem, 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 $f = \sum_{j=1}^m a_j k(\cdot, x_j)$,

$$\begin{aligned}\|f\|_{\mathcal{H}_1}^2 &= \langle \sum_i^n a_i k(\cdot, x_i), \sum_{j=1}^m a_j k(\cdot, x_j)\rangle_{\mathcal{H}_1} = \sum_{i,j=1}^n a_i a_j \langle k(\cdot, x_i), k(\cdot, x_j)\rangle_{\mathcal{H}_1} \\ &\overset{(i)}{=} \sum_{i,j=1}^n a_i a_j k(x_i, x_j) = \|f\|_{\mathcal{H}^0}^2.\end{aligned}$$

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

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

□

Theorem 3.7. *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 $\mathcal{H}$. □

4 A generalization analysis of kernel ridge regression

We first provide the upper bound of the Rademacher complexity.

Proposition 4.1. *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*

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

Proof.

$$\begin{aligned}
n\widehat{\mathrm{Rad}}_n(\mathcal{H}_k^Q) &= \mathbb{E}_\xi[\sup_{\|f\|_{\mathcal{H}_k} \leq Q} \sum_{i=1}^n \xi_i f(x_i)] = \mathbb{E}_\xi[\sup_{\|f\|_{\mathcal{H}_k} \leq Q} \sum_{i=1}^n \xi_i \langle f, k(\cdot, x_i) \rangle_{\mathcal{H}_k}] \text{(reproducing property)} \\
&= \mathbb{E}_\xi[\sup_{\|f\|_{\mathcal{H}_k} \leq Q} \langle f, \sum_{i=1}^n \xi_i k(\cdot, x_i) \rangle_{\mathcal{H}}] \leq Q\, \mathbb{E}_\xi[\| \sum_{i=1}^n \xi_i k(\cdot, x_i) \|_{\mathcal{H}_k}] \\
&= Q\, \mathbb{E}_\xi \sqrt{\sum_{i,j=1}^n \xi_i \xi_j k(x_i, x_j)} \leq Q \sqrt{\mathbb{E}_\xi[\sum_{i,j=1}^n \xi_i \xi_j k(x_i, x_j)]} \qquad \text{(Jensen inequality)} \\
&= Q \sqrt{\sum_{i=1}^n k(x_i, x_i)} \qquad (\mathbb{E}[\xi_i \xi_j] = 0, \forall\, i \neq j).
\end{aligned}$$

□

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

$$\hat{f}_n = \operatorname*{argmin}_{f \in \mathcal{H}_k} \hat{\mathcal{R}}_n(f) + \lambda \|f\|_{\mathcal{H}_k}. \tag{4.1}$$

Theorem 4.2 (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 $\hat{f}_n$,

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

which yields

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

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

$$\hat{\mathcal{R}}_n(\mathcal{F}_Q) \le L\hat{\mathcal{R}}_n(\mathcal{H}_k^Q).$$

Using the Rademacher complexity-based generalization bound, we have

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

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

□

The preceding estimate is a priori, since 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 the KRR.

Note that Theorem 4.2 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 level of noise as shown in the following theorem.

Consider the estimator

$$\hat{f}_n = \operatorname*{argmin}_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}(T \circ f(x_i) - y_i)^2 + \lambda\,\|f\|_{\mathcal{H}}, \tag{4.2}$$

where $T(t) = \min(\max(t, -1), 1)$. We make the following non-essential technical assumptions.

- $\sup_x |f^*(x)| \le 1$ and $\sup_x k(x,x) \le 1$.
- $y_i = f^*(x_i) + \xi_i$. $\{\xi\}_i$ are i.i.d. random noise with $|\xi_i| \le \sigma$. Assume that $\sigma \le 1$.

Theorem 4.3. *Under the preceding assumptions and taking $\lambda = \frac{\sigma}{\sqrt{n}}$, for any $\delta \in (0,1)$, we have*

$$\|\hat{f}_n - f^*\|^2_{L^2(\mathbb{P}_x)} \lesssim \frac{\|f^*\|_{\mathcal{H}_k} + \sqrt{\log(2/\delta)}}{\sqrt{n}}.$$

Proof. Let $Q = \|f^*\|_{\mathcal{H}_k}$. By definition,

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

which yields

$$\hat{\mathcal{R}}_n(\hat{f}_n) \le \hat{\mathcal{R}}_n(f^*) + \lambda Q$$
$$\|\hat{f}_n\|_{\mathcal{H}_k} \le Q + \frac{\hat{\mathcal{R}}_n(f^*)}{\lambda}.$$

Notice that $|y_i| \le 1+\sigma$, hence $\phi_i(t) := (t-y_i)^2$ is $2(2+\sigma)$-Lipschitz continuous. Using the contraction lemma and Rademacher complexity-based bound, for any $\delta_1 \in (0,1)$, we have with probability $1-\delta_1$,

$$\begin{aligned}
\mathcal{R}(\hat{f}_n) &\le \hat{\mathcal{R}}_n(\hat{f}_n) + \sup_{\|f\|_{\mathcal{H}_k} \le Q + \frac{\hat{\mathcal{R}}_n(f^*)}{\lambda}} |\mathcal{R}(f) - \hat{\mathcal{R}}_n(f)| \\
&\lesssim \hat{\mathcal{R}}_n(f^*) + \lambda Q + 2(2+\sigma)\,\mathrm{Rad}_n(\mathcal{F}_{Q+\frac{\hat{\mathcal{R}}_n(f^*)}{\lambda}}) + (1+\sigma)\sqrt{\frac{\log(2/\delta_1)}{n}} \\
&\lesssim \hat{\mathcal{R}}_n(f^*) + \lambda Q + \frac{Q + \frac{\hat{\mathcal{R}}_n(f^*)}{\lambda}}{\sqrt{n}} + \sqrt{\frac{\log(2/\delta_1)}{n}}. \qquad (4.3)
\end{aligned}$$

Notice that

$$\hat{\mathcal{R}}_n(\hat{f}_n) = \mathbb{E}[(\hat{f}_n(x) - f^*(x) - \xi)] = \|\hat{f}_n - f^*\|^2_{L^2(\mathbb{P}_x)} + \mathbb{E}[\xi^2]$$
$$\hat{\mathcal{R}}_n(f^*) = \frac{1}{n}\sum_{i=1}^n \xi_i^2.$$

By Hoeffding's inequality, for any $\delta_2 \in (0,1)$, with probability at least $1-\delta_2$,

$$\frac{1}{n}\sum_{i=1}^n \xi_i^2 - \mathbb{E}[\xi^2] \le \sigma\sqrt{\frac{\log(1/\delta_2)}{n}}.$$

Plugging it into (4.3),

$$\begin{aligned}
\|\hat{f}_n - f^*\|_{L^2}^2 &\lesssim \frac{1}{n}\sum_{i=1}^n \xi_i^2 - \mathbb{E}[\xi^2] + \lambda Q + \frac{\frac{1}{n}\sum_{i=1}^n \xi_i^2}{\lambda\sqrt{n}} + \frac{Q}{\sqrt{n}} + \sqrt{\frac{\log(1/\delta_1)}{n}} \\
&\lesssim \lambda Q + \frac{\sigma\sqrt{\log(1/\delta_2)}}{\lambda n} + \frac{Q}{\sqrt{n}} + \sqrt{\frac{\log(1/\delta_1)}{n}}.
\end{aligned}$$

Taking $\lambda = \frac{\sigma}{\sqrt{n}}$ and $\delta_2 = \delta_1 = \delta/2$, we complete the proof. □

4.1 Tightness.

Note that the 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 $\mathrm{Lip}(t^2/2) \le 1$. However, at 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 $O(1/n)$ and this approach is called "local Rademacher complexity". Please refer to [Bartlett et al., 2005] for more details.

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.