CUHK STAT6050: Statistical Learning Theory Spring 2026

Lecture 8: Nonparametric regression in RKHS

March 5, 2026

Lecturer: Ben Dai

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

1 Recall

Based on Lectures 1-7, we are able to compute the convergence rate and establish a probabilistic bound for a general learning method/algorithm. For illustration, we turn to investigate the asymptotics of nonparametric regression in a Reproducing Kernel Hilbert Space (RKHS).

2 RKHS

2.1 Why RKHS?

Motivation. In nonparametric regression, we seek a function space $\mathscr{F}$ that is rich enough to approximate complex functions, yet structured enough to allow for stable estimation.
From Hilbert Spaces to RKHS. A general Hilbert space $\mathscr{H}$ provides a powerful framework for function approximation, but it does not necessarily guarantee that convergence in the norm implies pointwise convergence. In nonparametric regression, we require that if our estimator f_n converges to the true function f in the Hilbert space norm, the predicted values $f_n(x)$ also converge to the true values $f(x)$.

Reproducing Kernel Hilbert Spaces (RKHS) are the specific class of Hilbert spaces that satisfy this requirement. By definition, an RKHS ensures that the evaluation functional $\delta_x(f) = f(x)$ is a continuous linear functional. According to the Riesz representation theorem, this continuity implies the existence of a representer $K_x \in \mathscr{H}$ such that $f(x) = \langle f, K_x \rangle_{\mathscr{H}}$. This reproducing property is the bridge between the abstract Hilbert space structure and the concrete evaluation of functions, enabling the use of kernels to define the space and perform computations.

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

Applying the Riesz representation theorem, if the evaluation functional $\delta_{\mathbf{x}} : f \mapsto f(\mathbf{x})$ is a bounded linear functional on $\mathscr{H}$, there exists a unique representer $K_{\mathbf{x}} \in \mathscr{H}$ such that

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

for all $f \in \mathscr{H}$. This result implies that function evaluation at any point $\mathbf{x}$ can be represented as an inner product with a specific function $K_{\mathbf{x}}$ in the Hilbert space.

Definition 2.2 (Reproducing Kernel Hilbert Space (RKHS))**.** A Hilbert space $\mathscr{H}$ of functions on $\mathscr{X}$ is called a Reproducing Kernel Hilbert Space (RKHS) if the evaluation functional $\delta_{\mathbf{x}} : f \mapsto f(\mathbf{x})$ is a bounded linear functional on $\mathscr{H}$ for every $\mathbf{x} \in \mathscr{X}$.

Theorem 2.3 (Pointwise Convergence in RKHS)**.** *If $\mathscr{H}$ is an RKHS, then convergence in the Hilbert space norm implies pointwise convergence. That is, for any sequence $\{h_n\} \subset \mathscr{H}$ and $h \in \mathscr{H}$,*

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

2.2 From kernel function to RKHS

Motivation. The definition of an RKHS guarantees the existence of a representer K_x for every evaluation functional δ_x. This naturally leads to the question: can we reverse this process? That is, given a series of representers $K_x (x \in \mathscr{X})$, can we construct a Hilbert space $\mathscr{H}$ such that K acts as its "core"? The answer is yes, and this construction is fundamental to the utility of RKHS in machine learning. By starting with a kernel, we can define the inner product structure of the space, which in turn determines the geometry and the properties of the Hilbert space.

By the Riesz representation theorem, the set $\{K_{\mathbf{x}} : \mathbf{x} \in \mathscr{X}\}$ serves as a *natural basis* for the RKHS.

Step 1. Define the inner product between basis elements.

The inner product between these basis elements is determined by the reproducing property:

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

Here, $K(\mathbf{x}, \mathbf{x}') = K_{\mathbf{x}'}(\mathbf{x}) = K_{\mathbf{x}}(\mathbf{x}')$ is a **symmetric bivariate function** known as the kernel function. By defining the kernel, we implicitly define both the basis elements $K_{\mathbf{x}}$ and the inner product between any two basis functions $K(\mathbf{x}, \mathbf{x}')$.

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

To mimic the construction in the finite-dimensional case, we first define a pre-RKHS $\mathscr{H}_0$ as the set of all finite linear combinations of kernel functions:

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

where $n \in \mathbb{N}$, $\alpha_i \in \mathbb{R}$, and $\mathbf{x}_i \in \mathscr{X}$. We equip this space with an inner product defined by the reproducing property:

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

where $f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i K(\mathbf{x}_i, \mathbf{x})$ and $f'(\mathbf{x}) = \sum_{j=1}^{m} \alpha'_j K(\mathbf{x}'_j, \mathbf{x})$.

Step 2. Generate the RKHS by taking the completion of the pre-RKHS.

The pre-RKHS $\mathscr{H}_0$ is generally not complete, meaning it may not contain the limits of all its Cauchy sequences. To ensure that we have a complete Hilbert space, we define $\mathscr{H}$ as the completion of $\mathscr{H}_0$ with respect to the norm induced by the inner product $\langle \cdot, \cdot \rangle_{\mathscr{H}_0}$. This completion includes all limit points of Cauchy sequences in $\mathscr{H}_0$, resulting in a space of functions of the form:

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

where the series converges in the norm of $\mathscr{H}$. The inner product on $\mathscr{H}$ is defined by the continuous extension of the inner product on $\mathscr{H}_0$:

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

where $\{f_n\}$ and $\{f'_n\}$ are Cauchy sequences in $\mathscr{H}_0$ converging to f and f', respectively.

Finally, we verify that $\mathscr{H}$ is indeed an RKHS by confirming that: (i) $\mathscr{H}$ is a Hilbert space; (ii) the evaluation functional δ_x is continuous on $\mathscr{H}$. A formal proof of this construction can be found 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 product**: 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.

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

  which holds for any $f \in \mathscr{H}$ or for any $n \geq 1$, 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 provide the formal definitions and theorems that underpin the construction of an RKHS.

Definition 2.5 (Reproducing kernel)**.** Let $\mathscr{H}$ be a Hilbert space of functions on $\mathscr{X}$. A function $K : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ is called a reproducing kernel of $\mathscr{H}$ if it satisfies the following two conditions:

1. For every $\mathbf{x} \in \mathscr{X}$, the function $K(\cdot, \mathbf{x})$ belongs to $\mathscr{H}$.
2. (Reproducing property). For every $\mathbf{x} \in \mathscr{X}$ and every $h \in \mathscr{H}$, the inner product satisfies $\langle h, K(\cdot, \mathbf{x}) \rangle_{\mathscr{H}} = h(\mathbf{x})$.

Definition 2.6 (Positive semi-definite kernel)**.** A symmetric function $K : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ is called a positive semi-definite kernel if, for any $n \in \mathbb{N}$, any set of points $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \mathscr{X}$, and any set of coefficients $\{\alpha_1, \ldots, \alpha_n\} \subset \mathbb{R}$, the following inequality holds:

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

The Moore-Aronszajn theorem guarantees the validity of the construction presented in Section 2.2.

Theorem 2.7 (Moore-Aronszajn theorem)**.** *Let $K : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ be a symmetric, positive semi-definite kernel. Then there exists a unique Reproducing Kernel Hilbert Space (RKHS) $\mathscr{H}$ such that K is the reproducing kernel of $\mathscr{H}$.*

In summary, the following concepts are equivalent:

$$\text{Reproducing kernel} \iff \text{Positive semi-definite kernel} \iff \text{RKHS}.$$

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 is the difference among different kernels? Theoretically, it affects both estimation and approximation errors; see the discussion in Section 4. Practically, it is highly related to the topic of *multiple kernel learning*; see [Gönen and Alpaydın, 2011] and references therein.

3 Ordinal Regression in RKHS

Let $\mathbf{X} \in \mathbb{R}^d$ be a feature vector and $Y \in \{0, \ldots, K\}$ be an ordinal outcome. We consider a continuous prediction function $f(\mathbf{X}) \in \mathbb{R}$ and use the mean absolute error (MAE) as the loss function:

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

We summarize the quantities of interest as follows:

- **Bayes rule:** $f^*(\mathbf{x}) = \text{Median}(Y|\mathbf{X} = \mathbf{x})$ is the global minimizer of $R(f)$.

- **Excess risk:**
$$\mathscr{E}(f) = R(f) - R(f^*) = \mathbb{E}\Big|Y - f(\mathbf{X})\Big|.$$

- **R-ERM:** Given random samples $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ and an RKHS $\mathscr{H}_K$, the estimator is defined as
$$\widehat{f}_n = \underset{f\in\mathscr{H}_K}{\arg\min} \left(\frac{1}{n}\sum_{i=1}^n \big|Y_i - f(\mathbf{X}_i)\big| + \lambda_n \|f\|_{\mathscr{H}_K}^2 \right).$$

- **Asymptotics:** Finally, we investigate the asymptotic behavior 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 $K(\cdot,\cdot)$ be 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:*
$$\widehat{f}_n = \underset{f\in\mathscr{H}_K}{\arg\min}\ \mathscr{L}_n\big(Y_1,\cdots,Y_n, f(\mathbf{X}_1),\cdots,f(\mathbf{X}_n)\big) + g(\|f\|_{\mathscr{H}_K}^2). \tag{1}$$
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. We refer readers to the summary by Wahba and Wang[1]. As they note:

> "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 for general regularization problems in infinite-dimensional spaces."

According to the Representer Theorem, the R-ERM can be reduced to:
$$\min_{\boldsymbol{\alpha}} \frac{1}{n}\sum_{i=1}^n \Big|Y_i - \sum_{j=1}^n \alpha_j K(\mathbf{X}_j, \mathbf{X}_i)\Big| + \lambda_n \Big\| \sum_{j=1}^n \alpha_j K(\mathbf{X}_j, \cdot)\Big\|_{\mathscr{H}_K}^2 = \min_{\boldsymbol{\alpha}} \frac{1}{n}\sum_{i=1}^n \big|Y_i - \boldsymbol{\alpha}^\intercal \mathbf{K}_i\big| + \lambda_n \boldsymbol{\alpha}^\intercal \mathbf{K}\boldsymbol{\alpha},$$
where $\boldsymbol{\alpha} = (\alpha_1,\cdots,\alpha_n)^\intercal$ and $\mathbf{K} = (K_{ij})_{n\times n}$ is a kernel matrix with $K_{ij} = K(\mathbf{X}_i, \mathbf{X}_j)$.

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

3.1 A first excess risk bound

Next, we turn to bound the excess risk of RKHS regression. For simplicity, we assume $0 \leq \widehat{f}_n(\mathbf{X}_i) \leq K$, otherwise we can consider its truncated estimation. Note that

$$\lambda_n \|\widehat{f}_n\|^2_{\mathscr{H}_K} \leq \widehat{R}_n(\widehat{f}_n) + \lambda_n \|\widehat{f}_n\|^2_{\mathscr{H}_K} \leq \widehat{R}_n(0) \leq K.$$

Thus, $\widehat{f}_n \in \mathscr{H}_n = \{f \in \mathscr{H} : \|f\|_{\mathscr{H}_K} \leq K\lambda_n^{-1/2}\}$. Then, by the standard decomposition of excess risk, we have

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

3.1.1 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

$$\mathbb{E}_\rho \Big(\sup_{f \in \mathscr{H}_n} \Big| \frac{1}{n} \sum_{i=1}^n \rho_i |Y_i - f(\mathbf{X}_i)| \Big| \Big) \leq \mathbb{E}_\rho \|\mathbf{Rad}_n(f)\|_{\mathscr{H}_n}.$$

This bound follows from the Ledoux-Talagrand contraction inequality, as the absolute value function $\phi(z) = |z|$ is 1-Lipschitz. It suffices to bound the Rademacher complexity of $\mathscr{H}_K$, we have the following lemma.

Lemma 3.3. *Suppose K is a uniformly bounded kernel with $\sup_{\mathbf{x} \in \mathscr{X}} \sqrt{K(\mathbf{x},\mathbf{x})} \leq K_0 < \infty$, $\mathscr{H}_K$ is its corresponding RKHS, and $\mathscr{H}_K(r) = \{f \in \mathscr{H}_K : \|f\|_{\mathscr{H}_K} \leq r\}$ is a $\mathscr{H}_K$-ball with a radius r. Then,*

$$\mathbb{E}_\rho \|\mathbf{Rad}_n(f)\|_{\mathscr{H}_K(r)} \leq r K_0 \sqrt{\frac{1}{n}}.$$

Therefore, $\|\mathbf{Rad}_n(l \bullet f)\|_{\mathscr{H}_n} \leq c K_0 (n\lambda_n)^{-1/2}$.

3.1.2 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 3.4 ([Cucker and Zhou, 2007])**.** *Let $\mathscr{X} \subset \mathbb{R}^d$ be a compact domain, and K be a reproducing kernel. Suppose there exists $0 < s \leq 1$, such that $f^* \in Range(L_K^{s/2})$, then*

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

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

3.2 Hyperparameter tuning

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

$$\varepsilon_n \geq A_0 \lambda_n^{s/2} + cK_0(n\lambda_n)^{-1/2} \geq \mathbf{Approx}(\lambda_n) + 8\mathbb{E}\|\mathbf{Rad}_n(l \bullet f)\|_{\mathscr{H}_n},$$

then

$$\mathbb{P}\Big(R(\widehat{f}_n) - R^* \geq \varepsilon_n\Big) \leq \exp\Big(-\frac{n\varepsilon_n^2}{8\big(U^2 + (1/2 + U/6)\varepsilon_n\big)}\Big).$$

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

$$\varepsilon_n^* = \inf_{\lambda_n} A_0 \lambda_n^{s/2} + cK_0(n\lambda_n)^{-1/2} = O(n^{-s/2(1+s)}),$$

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

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

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.