Topics in Deep Learning Theory (Spring 2025)

Lecture 4: Theoretical Analysis of KRR

Instructor: Lei Wu Date: May 1, 2025

Abstract

Kernel Ridge Regression (KRR) serves as a cornerstone kernel method, providing a compelling and approachable scenario for studying and understanding generalization. This lecture note delves into the technical details of analyzing KRR, underscoring its significance in illuminating fundamental concepts within learning theory. We examine two distinct approaches to derive generalization bounds for KRR:

1. **Empirical process**: This leverages uniform concentration inequalities to establish bounds on the generalization error, offering a broad and versatile perspective applicable to broader nonlinear models such as neural networks.
2. **Integral operator**: This approach capitalizes on KRR's explicit closed-form solution, employing integral operator techniques to obtain precise generalization bounds. These techniques are also applicable for analyzing properties of linear methods beyond traditional generalization.

By exploring these complementary methods, we illustrate how KRR bridges general theoretical tools and specialized algorithmic insights, making it an exemplary subject for learning theory analysis.

1 Preliminary

Let $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ be a kernel and $\mathcal{H}_k$ be the corresponding RKHS. The kernel ridge regression (KRR) using the RKHS $\mathcal{H}_k$ as its hypothesis space and fitting data using the following strategy:

$$\hat{f}_\lambda = \operatorname*{argmin}_{f \in \mathcal{H}_k} J_\lambda(f) := \frac{1}{n} \sum_{i=1}^n \left(f(x_i) - y_i\right)^2 + \lambda \|f\|_{\mathcal{H}_k}^q, \tag{1}$$

where $q \geq 1$. We shall focus on the classical case where $q = 2$. However, one may wonder how the choice of q, such as $q = 1$ or other values, affects the conclusions on the optimality of KRR.

We make the following assumptions to simplify our derivation.

Assumption 1.1. *The kernel satisfies* $\sup_{x \in \mathcal{X}} k(x, x) \leq 1$.

Assumption 1.2. $y_i = f^*(x_i) + \xi_i$ *where* $f^* \in \mathcal{H}_k$ *satisfies* $\|f^*\|_{\mathcal{H}_k} \geq 1$. *Suppose* $x_1, \ldots, x_n$ *are iid samples drawn from* $\rho \in \mathcal{P}(\mathcal{X})$ *and the noise* ξ_i *is independent of* x_i *with* $\mathbb{E}[\xi_i] = 0$ *and* $|\xi_i| \leq \sigma \leq 1$.

Question 1. Note that assuming the well-specified case where $f^* \in \mathcal{H}$ is a natural starting point for theoretical analysis. However, in practice, it is quite possible that $f^* \notin \mathcal{H}$. How can we address such a scenario?

2 Derivation via Empirical Process

Theorem 2.1. *Under Assumption 1.1 and 1.2, with probability at least* $1-\delta$*, the estimator obtained from KRR satisfies*

$$\|\hat{f}_\lambda - f^*\|_{L^2(\rho)}^2 \lesssim \frac{\sqrt{\log(1/\delta)}+1}{\sqrt{n}}\|f^*\|_{\mathcal{H}_k}^2,$$

if we set $\lambda = \frac{3\left(1+\sqrt{2\log(2/\delta)}\right)}{\sqrt{n}}\sigma$.

Without making any additional assumption, the above bound is tight. However, as we will show later, this bound becomse loose when extra structures are available:

- **Target function.** f^* may lie in a function space smaller than $\mathcal{H}_k$, i.e., it has extra smoothness beyond $\mathcal{H}_k$.
- **Model.** The eigenvalues of k may decay very fast. Faster decay corresponds to a smaller model, which should have led to a faster rate for the estimation error.

2.1 Proof

We shall state a few lemmas related to RKHSs.

Lemma 2.2. *For any* $f \in \mathcal{H}_k$ *and* $x \in \mathcal{X}$, $|f(x)| \le \|f\|_{\mathcal{H}_k}\sqrt{k(x,x)}$.

Proof. By the reproducing property, we have

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

□

Lemma 2.3. *Let* $\mathcal{F}_Q = \{f : \|f\|_{\mathcal{H}_k} \le Q\}$. *Then, we have* $\widehat{\mathrm{Rad}}_n(\mathcal{F}_Q) \le Q\sqrt{\frac{1}{n^2}\sum_{i=1}^n k(x_i,x_i)}$.

Proof. By the definition of the empirical Rademacher complexity, we have

$$\begin{aligned}
\widehat{\mathrm{Rad}}_n(\mathcal{F}_Q) &= \mathbb{E}_\epsilon\left[\sup_{\|f\|_{\mathcal{H}_k}\le Q}\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)\right] \\
&= \mathbb{E}_\epsilon\left[\sup_{\|f\|_{\mathcal{H}_k}\le Q}\left\langle f, \frac{1}{n}\sum_{i=1}^n \epsilon_i k(\cdot,x_i)\right\rangle_{\mathcal{H}_k}\right] \\
&\le \mathbb{E}_\epsilon\left[\sup_{\|f\|_{\mathcal{H}_k}\le Q}\|f\|_{\mathcal{H}_k}\left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i k(\cdot,x_i)\right\|_{\mathcal{H}_k}\right] \le Q\mathbb{E}_\epsilon\left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i k(\cdot,x_i)\right\|_{\mathcal{H}_k},
\end{aligned}$$

where the second equality follows from the reproducing property of the kernel and the first inequality derives from Cauchy-Schwarz inequality. By Jensen's inequality, we then obtain

$$\begin{aligned}\mathbb{E}_\epsilon \left\| \frac{1}{n}\sum_{i=1}^n \epsilon_i k(\cdot, x_i) \right\|_{\mathcal{H}_k} &\leq \sqrt{\mathbb{E}_\epsilon \left\| \frac{1}{n}\sum_{i=1}^n \epsilon_i k(\cdot, x_i) \right\|_{\mathcal{H}_k}^2} \\ &= \sqrt{\mathbb{E}_\epsilon \left\langle \frac{1}{n}\sum_{i=1}^n \epsilon_i k(\cdot, x_i), \frac{1}{n}\sum_{i=1}^n \epsilon_i k(\cdot, x_i) \right\rangle_{\mathcal{H}_k}} \\ &= \sqrt{\frac{1}{n^2}\mathbb{E}_\epsilon \sum_{i,j=1}^n \epsilon_i \epsilon_j k(x_i, x_j)} = \sqrt{\frac{1}{n^2}\sum_{i=1}^n k(x_i, x_i)}.\end{aligned}$$

Combining the above estimation, we complete the proof. □

2.1.1 Step 1: Comparison inequality

In order to extract some basic properties of $\hat{f}_\lambda$, we can compare it with certain reference solutions, whose properties are well understood. For instance, a typical choice for such a reference solution is the ground truth f^* , though **it is not strictly necessary**.

Since $\hat{f}_\lambda$ minimizes $J_\lambda(\cdot)$ over $\mathcal{H}_k$, and $f^* \in \mathcal{H}_k$, we have

$$J_\lambda(\hat{f}_\lambda) \leq J_\lambda(f^*). \tag{2}$$

Thus, we have

$$\frac{1}{n}\sum_{i=1}^n (\hat{f}_\lambda(x_i) - f^*(x_i) - \xi_i)^2 + \lambda \|\hat{f}_\lambda\|_{\mathcal{H}_k}^2 \leq 0 + \lambda \|f^*\|_{\mathcal{H}_k}^2,$$

which leads to

$$\frac{1}{n}\sum_{i=1}^n (\hat{f}_\lambda(x_i) - f^*(x_i))^2 \leq \frac{2}{n}\sum_{i=1}^n \xi_i(\hat{f}_\lambda(x_i) - f^*(x_i)) + \lambda(\|f^*\|_{\mathcal{H}_k}^2 - \|\hat{f}_\lambda\|_{\mathcal{H}_k}^2). \tag{3}$$

We next show that we can obtain an upper bound of empirical risk and the the norm of $\hat{f}_\lambda$ by utilizing this comparison inequality.

Question. In the case where $f^* \notin \mathcal{H}_k$, how should the reference solution be chosen? Since f^* cannot be selected as the reference solution in this scenario, what alternative approach should be taken?

2.1.2 Step 2: Controlling the noise interaction term

We next to deal with the noise term $\frac{1}{n}\sum_{i=1}^n \xi_i(\hat{f}_\lambda(x_i) - f^*(x_i))$ in (3). It follows from the reproducing property that $\hat{f}_\lambda(x_i) - f^*(x_i) = \langle \hat{f}_\lambda - f^*, k(\cdot, x_i)\rangle_{\mathcal{H}_k}$ and thus, we have

$$\frac{1}{n}\sum_{i=1}^n \xi_i(\hat{f}_\lambda(x_i) - f^*(x_i)) = \left\langle \hat{f}_\lambda - f^*, \frac{1}{n}\sum_{i=1}^n \xi_i k(\cdot, x_i) \right\rangle_{\mathcal{H}_k}.$$

Applying the Cauchy-Schwarz inequality, we get:

$$\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i(\hat{f}_\lambda(x_i)-f^*(x_i))\right| \le \|\hat{f}_\lambda - f^*\|_{\mathcal{H}_k}\left\|\frac{1}{n}\sum_{i=1}^{n}\xi_i k(\cdot,x_i)\right\|_{\mathcal{H}_k}. \tag{4}$$

To bound the second term on the right-hand side, we then need the following lemma.

Lemma 2.4. *Let $k:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ be a PSD kernel satisfying $\sup_{x\in\mathcal{X}}k(x,x)\le 1$. Let $\{\xi_i\}_{i=1}^n$ be i.i.d. random variables with mean zero and $|\xi_i|\le\sigma$. Given fixed points $x_1,\dots,x_n\in\mathcal{X}$, for any $\delta\in(0,1)$, with probability at least $1-\delta$, the following holds:*

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\xi_i k(x_i,\cdot)\right\|_{\mathcal{H}_k} \le \frac{1+\sqrt{2\log(1/\delta)}}{\sqrt{n}}\sigma.$$

It is worth noting that one can apply certain concentration inequalities for Hilbert-valued random vectors or the Hanson–Wright inequality (Theorem 3.11) to obtain a similar upper bound without assuming bounded noise. Nevertheless, we provide an elementary, self-contained proof here that leverages the special structure of RKHS.

Proof. We aim to bound the RKHS norm of the random element $\frac{1}{n}\sum_{i=1}^n\xi_i k(x_i,\cdot)$ in $\mathcal{H}_k$. Using the properties of the RKHS, the square of RKHS norm is:

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\xi_i k(x_i,\cdot)\right\|_{\mathcal{H}_k}^2 = \left\langle\frac{1}{n}\sum_{i=1}^{n}\xi_i k(x_i,\cdot),\frac{1}{n}\sum_{j=1}^{n}\xi_j k(x_j,\cdot)\right\rangle_{\mathcal{H}_k} = \frac{1}{n^2}\sum_{i,j=1}^{n}\xi_i\xi_j k(x_i,x_j),$$

where the last equation is due to the reproducing property. This expresses the square of RKHS norm as a quadratic form in the random variables ξ_i.

Next, we compute the expectation of the squared norm:

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\xi_i k(x_i,\cdot)\right\|_{\mathcal{H}_k}^2\right] = \frac{1}{n^2}\sum_{i,j=1}^{n}\mathbb{E}[\xi_i\xi_j]k(x_i,x_j) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}[\xi_i^2]k(x_i,x_i),$$

since $\{\xi_i\}$ are independent and $\mathbb{E}[\xi_i]=0$. Given $|\xi_i|\le\sigma$, we have $\mathbb{E}[\xi_i^2]\le\sigma^2$. Additionally, since $k(x_i,x_i)\le 1$, it follows that

$$\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}[\xi_i^2]k(x_i,x_i) \le \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2\cdot 1 = \frac{\sigma^2}{n}.$$

Applying Jensen's inequality, we get:

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\xi_i k(x_i,\cdot)\right\|_{\mathcal{H}_k}\right] \le \sqrt{\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\xi_i k(x_i,\cdot)\right\|_{\mathcal{H}_k}^2\right]} \le \frac{\sigma}{\sqrt{n}}.$$

This gives an upper bound on the expected norm.

To obtain a high-probability bound, define

$$O(\xi_1, \ldots, \xi_n) = \left\| \frac{1}{n} \sum_{i=1}^{n} \xi_i k(x_i, \cdot) \right\|_{\mathcal{H}_k}.$$

To apply McDiarmid's inequality, we should bound the difference in f when one variable changes. Let ξ and $\xi^{(i)}$ differ only in the i-th component. Then

$$|O(\xi) - O(\xi^{(i)})| \leq \frac{1}{n} \|(\xi_i - \xi_i')k(x_i, \cdot)\|_{\mathcal{H}_k}.$$

Since $\|k(x_i, \cdot)\|_{\mathcal{H}_k} = \sqrt{k(x_i, x_i)} \leq 1$ and $|\xi_i - \xi_i'| \leq 2\sigma$, we have

$$|O(\xi) - O(\xi^{(i)})| \leq \frac{1}{n} \cdot 2\sigma \cdot 1 = \frac{2\sigma}{n}.$$

The bounded difference condition holds with $c_i = \frac{2\sigma}{n}$. By McDiarmid's inequality,

$$\mathbb{P}\left(O \geq \mathbb{E}\, O + t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n} c_i^2}\right) = \exp\left(-\frac{nt^2}{2\sigma^2}\right).$$

To ensure this probability is at most δ, set

$$\exp\left(-\frac{nt^2}{2\sigma^2}\right) \leq \delta \implies t \geq \frac{\sqrt{2\log(1/\delta)}\sigma}{\sqrt{n}}.$$

Combining with $\mathbb{E}\, O \leq \frac{\sigma}{\sqrt{n}}$, with probability at least $1 - \delta$:

$$O \leq \frac{\sigma}{\sqrt{n}} + \frac{\sqrt{2\log(1/\delta)}\sigma}{\sqrt{n}} = \frac{1 + \sqrt{2\log(1/\delta)}}{\sqrt{n}}\sigma.$$

This completes the proof. □

Applying Lemma 2.4, with probability at least $1 - \delta/2$, we have

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \xi_i k(\cdot, x_i) \right\|_{\mathcal{H}_k} \leq \lambda_n,$$

where $\lambda_n = \frac{1+\sqrt{2\log(2/\delta)}}{\sqrt{n}}\sigma$. Thus, with probability at least $1 - \delta/2$:

$$\frac{1}{n} \sum_{i=1}^{n} \xi_i (\hat{f}_\lambda(x_i) - f^*(x_i)) \leq \lambda_n \|\hat{f}_\lambda - f^*\|_{\mathcal{H}_k}. \tag{5}$$

2.1.3 Step 3: Controlling the training error and the norm of $\hat{f}_\lambda$

Combining (3) and (5), we have

$$\frac{1}{n} \sum_{i=1}^{n} (\hat{f}_\lambda(x_i) - f^*(x_i))^2 \leq 2\lambda_n \left\| f^* - \hat{f}_\lambda \right\|_{\mathcal{H}_k} + \lambda \left(\|f^*\|_{\mathcal{H}_k}^2 - \left\| \hat{f}_\lambda \right\|_{\mathcal{H}_k}^2 \right) \tag{6}$$

Let $\lambda \geq 3\lambda_n$. We know from (6) that the following two are satisfied with probability at least $1 - \delta/2$:

- The norm of the estimator is bounded by

$$\begin{aligned} 0 \leq \frac{2}{3}\lambda\left(\|f^*\|_{\mathcal{H}_k} + \left\|\hat{f}_\lambda\right\|_{\mathcal{H}_k}\right) + \lambda\left(\|f^*\|_{\mathcal{H}_k}^2 - \left\|\hat{f}_\lambda\right\|_{\mathcal{H}_k}^2\right) \\ \implies \left(\left\|\hat{f}_\lambda\right\|_{\mathcal{H}_k} - \frac{1}{3}\right)^2 \leq \|f^*\|_{\mathcal{H}_k}^2 + \frac{2}{3}\|f^*\|_{\mathcal{H}_k} + \frac{1}{9} \\ \implies \|\hat{f}_\lambda\|_{\mathcal{H}_k} \leq 2\|f^*\|_{\mathcal{H}_k}. \end{aligned} \tag{7}$$

  where we have used the assumption $\|f^*\|_{\mathcal{H}_k} \geq 1$.

- The training error is bounded by

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n (\hat{f}_\lambda(x_i) - f^*(x_i))^2 &\leq \frac{2}{3}\lambda\left(\|f^*\|_{\mathcal{H}_k} + \left\|\hat{f}_\lambda\right\|_{\mathcal{H}_k}\right) + \lambda\left(\|f^*\|_{\mathcal{H}_k}^2 - \left\|\hat{f}_\lambda\right\|_{\mathcal{H}_k}^2\right) \\ &\leq 2\lambda\|f^*\|_{\mathcal{H}_k} + \lambda\|f^*\|_{\mathcal{H}_k}^2 \\ &\leq 3\lambda\|f^*\|_{\mathcal{H}_k}^2. \end{aligned} \tag{8}$$

2.1.4 Step 4: Estimating the Rademacher complexity

Define the shift class $\mathcal{G}_Q$ by

$$\mathcal{G}_Q = \{x \mapsto f(x) - f^*(x) : f \in \mathcal{F}_Q\},$$

and let

$$\mathcal{L}_Q = \{x \mapsto \phi(g(x)) : g \in \mathcal{G}_Q\} = \{x \mapsto (f(x) - f^*(x))^2 : f \in \mathcal{F}_Q\}.$$

where $\phi(t) = t^2$. The following lemma provides a bound on $\widehat{\mathrm{Rad}}(\mathcal{L}_Q)$.

Lemma 2.5. *Suppose Assumption 1.1 holds. Then, we have* $\widehat{\mathrm{Rad}}_n(\mathcal{L}_Q) \leq \frac{2Q}{\sqrt{n}}(Q + \|f^*\|_{\mathcal{H}_k})$.

Proof. For any $g \in \mathcal{G}_Q$, we see that

$$\begin{aligned} |g(x)| &= |f(x) - f^*(x)| \\ &\leq |f(x)| + |f^*(x)| \\ &= |\langle f, k(\cdot,x)\rangle_{\mathcal{H}_k}| + |\langle f^*, k(\cdot,x)\rangle_{\mathcal{H}_k}| \\ &\leq \sqrt{k(x,x)}\,[\|f\|_{\mathcal{H}_k} + \|f^*\|_{\mathcal{H}_k}] \leq Q + \|f^*\|_{\mathcal{H}_k}. \end{aligned}$$

Since $\phi(t)$ is 2M-Lipschitz on the interval $[-M, M]$, then by the contraction lemma, we have

$$\begin{aligned} \widehat{\mathrm{Rad}}_n(\mathcal{L}_Q) &= \widehat{\mathrm{Rad}}_n(\phi \circ \mathcal{G}_Q) \\ &\leq 2(Q + \|f^*\|_{\mathcal{H}_k})\widehat{\mathrm{Rad}}_n(\mathcal{G}_Q) \\ &\leq 2(Q + \|f^*\|_{\mathcal{H}_k})\widehat{\mathrm{Rad}}_n(\mathcal{F}_Q) \\ &\leq \frac{2Q}{\sqrt{n}}(Q + \|f^*\|_{\mathcal{H}_k}), \end{aligned}$$

where the equality $\widehat{\mathrm{Rad}}_n(\mathcal{G}_Q) = \widehat{\mathrm{Rad}}_n(\mathcal{F}_Q)$ follows from the shift invariance of the Rademacher complexity, and the last inequality uses Lemma 2.3 in conjunction with Assumption 1.1. This completes the proof. □

2.1.5 Step 5: Putting all together

According to (7), we can set $Q = 2\|f^*\|_{\mathcal{H}_k}$. Conditional on $\|\hat{f}_\lambda\|_{\mathcal{H}_k} \leq Q$, and with probability at least $1 - \delta/2$, we have that

$$\mathbb{E}_x[(\hat{f}_\lambda(x) - f^*(x))^2] \leq \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_\lambda(x_i) - f^*(x_i))^2 + 2\widehat{\mathrm{Rad}}_n(\mathcal{L}_Q) + 2B\sqrt{\frac{2\log(8/\delta)}{n}}.$$

where $B = \sup_{x\in\mathcal{X}}(\hat{f}_\lambda(x) - f^*(x))^2 \leq (Q + \|f^*\|_{\mathcal{H}_k})^2$. Combining this with (8) and Lemma 2.5, and setting

$$\lambda = 3\lambda_n = \frac{3\left(1 + \sqrt{2\log(2/\delta)}\right)}{\sqrt{n}}\sigma,$$

we finally get

$$\begin{aligned}\mathbb{E}_x[(\hat{f}_\lambda(x) - f^*(x))^2] &\lesssim \lambda\|f^*\|^2_{\mathcal{H}_k} + \frac{\|f^*\|^2_{\mathcal{H}_k}}{\sqrt{n}} + \|f^*\|^2_{\mathcal{H}_k}\sqrt{\frac{\log(1/\delta)}{n}} \\ &\asymp \frac{\sqrt{\log(1/\delta)} + 1}{\sqrt{n}}\|f^*\|^2_{\mathcal{H}_k}.\end{aligned}$$

with probability at least $1 - \delta$.

2.2 Remarks

The choice of q. If going through the proof, we can see that the choice of q is not essential for deriving the generalization bound; a similar bound can be obtained for $q = 1$. In practice, the choice $q = 2$ is primarily motivated by the availability of a closed-form solution.

Handle the case where $f^* \notin \mathcal{H}_k$. In certain scenarios—especially within computational mathematics—one may encounter cases where the true target function $f^* \notin \mathcal{H}_k$. For example, if $f^* \in W^{r,2}([0,1]^d)$ and $r < \frac{d}{2}$, then $W^{k,2}$ is not an RKHS, and the standard KRR analysis may not directly apply. To address this, one can select $f^*_\epsilon \in \mathcal{H}_k$ such that $\|f^*_\epsilon - f^*\|_{L^2(\rho)} \leq \epsilon$. This yields

$$\|\hat{f}_\lambda - f^*\|_{L^2(\rho)} \leq \|\hat{f}_\lambda - f^*_\epsilon\|_{L^2(\rho)} + \|f^*_\epsilon - f^*\|_{L^2(\rho)} \leq \|\hat{f}_\lambda - f^*_\epsilon\|_{L^2(\rho)} + \epsilon.$$

The term $\|\hat{f}_\lambda - f_\epsilon\|_{L^2(\rho)}$ can be bounded using an approach similar to above analysis but with f_ϵ as the new target function. However, $q^*(\epsilon) := \|f_\epsilon\|_{\mathcal{H}_k}$ grows with ϵ. Roughly speaking, we may get somethimg like

$$\|\hat{f}_\lambda - f^*\|_{L^2(\rho)} \leq \frac{\|f^*_\epsilon\|}{\sqrt{n}} + \epsilon = \frac{q^*(\epsilon)}{\sqrt{n}} + \epsilon.$$

Then, we can obtain the rate is

$$\|\hat{f}_\lambda - f^*\|_{L^2(\rho)} \leq \inf_\epsilon\left(\frac{q^*(\epsilon)}{\sqrt{n}} + \epsilon\right).$$

Sharper bounds via localization. The reason that above derivation stems from that when applying the uniform bound, we consider something like

$$\|\hat{f}_\lambda - f^*\|_\rho \le \sup_{f \in \mathcal{H}_k^Q} \|f - f^*\|_\rho.$$

However, argubaly, $\hat{f}_\lambda$ is close to f^*, as the final bound tells us $\|\hat{f}_\lambda - f^*\|_\rho = O(n^{-1/2})$. Therefore, we should be above to obtain a tighter bound by using this information. Specifically, define the localized hypothesis:

$$\mathcal{H}_{r,Q} = \{f : \|f\|_{\mathcal{H}_k} \le Q, \|f - f^*\|_\rho \le r\}.$$

- We first apply the above derivation to $\mathcal{H}_{1,Q}$, then obtain

$$\hat{f}_\lambda \in \mathcal{H}_{\epsilon_{n,1},Q}, \text{ with } \epsilon_{n,1} = \frac{1}{\sqrt{n}}.$$

- We then apply uniform bound over a smaller hypoheis $\mathcal{H}_{\epsilon_n^1,Q}$ to obtain an error bound $\epsilon_{n,2}$, with $\epsilon_{n,2} \le \epsilon_{n,1}$.
- Repeat the above process until the estimate does not improve anymore.

The idea of localization to obtaining sharper bounds is very intuitive. However, rigorously formalizing this intuition is nontrivial and beyond our scope. For further details, we refer readers to [Wainwright, 2019, Section 13–14].

3 Derivation via Integral Operator

Express KRR solution using operator. Consider the integral operator $\mathcal{T} : \mathcal{H}_k \mapsto \mathcal{H}_k$ given by

$$\mathcal{T}f = \int_{\mathcal{X}} k(\cdot, x) f(x) \, \mathrm{d}\rho(x) = \int_{\mathcal{X}} k(\cdot, x) \langle k(x, \cdot), f \rangle_{\mathcal{H}_k} \, \mathrm{d}\rho(x) = \mathbb{E}_x[k(x, \cdot) \otimes k(x, \cdot)]f$$

Then, its empirical version is given by

$$\widehat{\mathcal{T}}f = \frac{1}{n} \sum_{i=1}^n k(\cdot, x_i) f(x_i) = \left(\frac{1}{n} \sum_{i=1}^n k(\cdot, x_i) \otimes k(\cdot, x_i) \right) f.$$

Here, the outer product is defined as follows: Given two Hilbert spaces $\mathcal{H}_1, \mathcal{H}_2$ and $v \in \mathcal{H}_1$, $u \in \mathcal{H}_2$, the operator $u \otimes v : \mathcal{H}_1 \to \mathcal{H}_2$ is defined by $(u \otimes v)f = \langle v, f \rangle_{\mathcal{H}_1} u$ for any $f \in \mathcal{H}_1$.

Thus, $\mathcal{T} : \mathcal{H}_k \mapsto \mathcal{H}_k$ and its empirical $\widehat{\mathcal{T}} : \mathcal{H}_k \mapsto \mathcal{H}_k$ are well-defined. Then, using these operators, the objective of KRR can be rewritten as

$$\begin{aligned} J_\lambda(f) &= \frac{1}{n} \sum_{i=1}^n \left(\langle f, k(x_i, \cdot) \rangle_{\mathcal{H}_k} - y_i \right)^2 + \lambda \|f\|_{\mathcal{H}_k}^2 \\ &= \left\langle f, (\widehat{\mathcal{T}} + \lambda) f \right\rangle_{\mathcal{H}_k} - 2 \left\langle f, \frac{1}{n} \sum_{i=1}^n y_i k(x_i, \cdot) \right\rangle_{\mathcal{H}_k} + \frac{1}{n} \sum_{i=1}^n y_i^2. \end{aligned}$$

A simple variation calculus gives

$$\hat{f}_\lambda = (\hat{T} + \lambda)^{-1}\hat{g}, \qquad \hat{g} = \frac{1}{n}\sum_{i=1}^{n} y_i k(x_i, \cdot). \tag{9}$$

Note that

$$\begin{aligned}
\hat{g} &= \frac{1}{n}\sum_{i=1}^{n} f^*(x_i)k(x_i,\cdot) + \frac{1}{n}\sum_{i=1}^{n}\xi_i k(x_i,\cdot) \\
&= \frac{1}{n}\sum_{i=1}^{n} k(x_i,\cdot)\langle k(x_i,\cdot), f^*\rangle_{\mathcal{H}_k} + \frac{1}{n}\sum_{i=1}^{n}\xi_i k(x_i,\cdot) \\
&= \hat{T}f^* + \frac{1}{n}\sum_{i=1}^{n}\xi_i k(x_i,\cdot).
\end{aligned}$$

Plugging it into (9) gives

$$\hat{f}_\lambda = \underbrace{(\hat{T}+\lambda)^{-1}\hat{T}f^*}_{\text{bias}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\xi_i(\hat{T}+\lambda)^{-1}k(x_i,\cdot)}_{\text{variance}}$$

The integral operator approach leverages the above closed-form solution to derive sharp error estimates; for details, we refer interested readers to [Fischer and Steinwart, 2020, Zhang et al., 2023]. However, because this approach relies heavily on operator calculus and remains somewhat abstract, we instead adopt equivalent methods that naturally reduce the problem to finite-dimensional linear regression, offering greater intuition and ease of understanding.

Translating KRR to a (infinitely-dimensional) linear ridge regression. Suppose that Mercer's theorem holds, i.e., the kernel admits the following spectral decomposition

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

where $p = \infty$. One can treat p as a finite integer without losing any intuition (even rigor). Let ℓ_p^2 be $\mathbb{R}^p$ equipped with the ℓ^2 norm. Then, consider the feature map

$$\varphi(x) = (\sqrt{\lambda_1}e_1(x), \ldots, \sqrt{\lambda_p}e_p(x)) \in \ell_p^2, \tag{10}$$

under which

$$\Sigma = \mathbb{E}_x[\varphi(x)\varphi(x)^\top] = \mathrm{diag}(\lambda_1, \ldots, \lambda_p).$$

We shall denote its empirical version as

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\varphi(x_i)\varphi(x_i)^\top.$$

By our previous theory, KRR is equivalent to perform linear regression for $f_\theta = \varphi(x)^\top\theta$.

Assumption 3.1. *Let $f^*(x) = \varphi(x)^\top \theta^*$ with $\|\theta^*\| \le 1$.*

This assumption is equivalent to assume $\|f^*\|_{\mathcal{H}_k} \le 1$. Then, the KRR is equivalent to linear ridge regression

$$\hat{\theta}_\lambda = \operatorname*{argmin}_{\theta \in \ell_p^2} \left(\frac{1}{n} \sum_{i=1}^{n} (\varphi(x_i)^\top \theta - y_i) + \lambda \|\theta\|^2 \right). \tag{11}$$

It is easy to obtain the close-form solution of ridge regression:

$$\hat{\theta}_\lambda = (\widehat{\Sigma} + \lambda)^{-1} \widehat{\Sigma} \theta^* + (\widehat{\Sigma} + \lambda)^{-1} \frac{1}{n} \Phi^\top \xi,$$

Note that for any $\theta \in \ell_p^2$, $\mathcal{E}(f_\theta) = \|f_\theta - f^*\|_{L^2(\mu)} = \|\theta - \theta^*\|_\Sigma$. Thus, by the triangle inequality, we have the following bias-variance decomposition:

$$\mathcal{E}(f_{\hat{\theta}_\lambda}) \le \underbrace{\left\| (\widehat{\Sigma} + \lambda)^{-1} \widehat{\Sigma} \theta^* - \theta^* \right\|_\Sigma}_{B(\lambda)} + \underbrace{\left\| (\widehat{\Sigma} + \lambda)^{-1} \frac{1}{n} \Phi^\top \xi \right\|_\Sigma}_{V(\lambda)},$$

for which we shall estimate the two terms separately.

Remark 3.2. *We remark that all of the following derivations, in principle, hold for any feature map $\varphi : \mathcal{X} \to \mathcal{H}$ satisfying $k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}}$. In what follows, we consider the specific feature map given in (10), as it makes all computations explicit, particularly the following max degree of freedom (DoF).*

To establish a high-probability bound, we first introduce the following degree-of-freedom (DoF):

Definition 3.3 (Max-DoF)**.** Let $F(\lambda) = \sup_{x \in \mathcal{X}} \|(\Sigma + \lambda)^{-1/2} \varphi(x)\|^2$.

Due to the choice of the feature map φ in (10), we have

$$F(\lambda) = \sup_{x \in \mathcal{X}} \sum_{j=1}^{\infty} \frac{\lambda_j}{\lambda_j + \lambda} e_j^2(x).$$

It follows that

$$F(\lambda) = \sup_{x \in \mathcal{X}} \sum_{j=1}^{\infty} \frac{\lambda_j}{\lambda_j + \lambda} e_j^2(x) \ge \int \sum_{j=1}^{\infty} \frac{\lambda_j}{\lambda_j + \lambda} e_j^2(x) \, d\mu(x) = \sum_{j=1}^{\infty} \frac{\lambda_j}{\lambda_j + \lambda} =: N(\lambda),$$

implying that $F(\lambda)$ lower bounds the classical average DoF $N(\lambda) = \operatorname{tr}\big[(\Sigma + \lambda)^{-1} \Sigma\big]$,

Lemma 3.4. *If the eigenfunctions are uniformly bounded, i.e., $\sup_j \|e_j\|_\infty \le C$, then*

$$N(\lambda) \le F(\lambda) \le C \, N(\lambda).$$

The proof is straightforward and implies that the max DoF is equivalent to the average DoF up to a multiplicative constant. For periodic kernels—whose eigenfunctions are Fourier basis functions—the condition holds. However, for dot-product kernels $k(x, x') = \kappa(x^\top x')$ with $x, x' \in \mathbb{S}^{d-1}$, the eigenfunctions are spherical harmonics, which are not uniformly bounded. Essentially, we expect the max DoF to behave similarly to the average DoF when the eigenfunctions are reasonably well-behaved. Uniform boundedness is only one sufficient condition; alternative conditions, such as the embedding property introduced in [Fischer and Steinwart, 2020], can also be used to ensure a similar behavior. Nonetheless, the reader may treat these quantities as essentially equivalent without overly concerning themselves with the precise conditions.

Theorem 3.5. *For any $\delta \in (0, 1)$, let*

$$\lambda \geq \lambda_{n,\delta} := \inf \left\{ \lambda : n \gtrsim F(\lambda) \max(1, \log(F(\lambda)/\delta)) \right\}.$$

Then, it holds w.p. at least $1 - \delta$ that

$$\|\hat{f}_\lambda - f^*\|_{L^2(\rho)}^2 \lesssim \lambda + \frac{C_\delta \sigma^2}{n} F(\lambda),$$

where $C_\delta = 1 + \log(1/\delta)$.

When $\lambda_j \asymp j^{-\beta}$ with $\beta > 1$ and the eigenfunctions are uniformly bounded, we have $F(\lambda) \leq \lambda^{-1/\beta}$. Then, we have

$$\inf_\lambda \left(\lambda + \frac{C_\delta \sigma^2}{n} \lambda^{-1/\beta} \right) = \left(\frac{C_\delta \sigma^2}{\beta n} \right)^{\frac{\beta}{\beta+1}} \asymp n^{-\frac{\beta}{\beta+1}}.$$

Similar to the bound of training error, we observe that the test error also converges to zero at a rate of $O(n^{-\beta/(1+\beta)})$. As β approaches infinity, this rate improves to the *fast rate*. Conversely, as β approaches 1 from above, the rate reduces to $O(n^{-1/2})$, which matches the rate obtained through the above coarse-grained application of the empirical process technique.

Remark 3.6. *Consider the noiseless regime where $\sigma = 0$. The generalization error bound simplifies to*

$$\|\hat{f}_\lambda - f^*\|_{L^2(\rho)}^2 \lesssim \lambda,$$

where we can take $\lambda = \lambda_{n,\delta}$. With $\max(1, \log(F(\lambda)/\delta)) \lesssim \lambda^{-\epsilon}$ for any $\epsilon > 0$ when $0 < \lambda < 1$ and $F(\lambda) \leq \lambda^{-1/\beta}$, we obtain an upper bound for $\lambda_{n,\delta}$ as $\lambda_{n,\delta} \lesssim n^{-\frac{\beta}{1+\epsilon\beta}}$. Hence

$$\|\hat{f}_\lambda - f^*\|_{L^2(\rho)}^2 \lesssim n^{-\frac{\beta}{1+\epsilon\beta}} \to n^{-\beta}.$$

In the noisy regime, the rate cannot improve upon the standard parametric/fast rate $O(n^{-1})$. In contrast, in the noiseless regime, the rate can exceed $O(n^{-1})$ when $\beta > 1$, implying that substantially fewer samples may be required compared to the noisy setting.

Remark 3.7. *Note that the integral operator approach is limited to the squared loss and to penalties with $q = 2$. In contrast, the empirical process approach applies to general Lipschitz loss functions and to penalties with $q \geq 1$.*

3.1 Proof

3.1.1 Step 1: Concentration of the empirical covariance

We need the following dimension-free matrix/operator concentration inequality:

Theorem 3.8 (Theorem 3.1 in [Minsker, 2017]). *Let $X_1, \dots, X_n$ be a sequence of independent self-adjoint random operators on Hilbert space $\mathcal{H}$ such that $\mathbb{E}X_i = 0$ for $i = 1, \dots, n$ and $\left\|\sum_{i=1}^n \mathbb{E}X_i^2\right\| \le \sigma^2$. Assume that $\|X_i\| \le U$ almost surely for all $1 \le i \le n$ and some positive $U \in \mathbb{R}$. Then, for any $t \ge \frac{1}{6}\left(U + \sqrt{U^2 + 36\sigma^2}\right)$,*

$$\mathbb{P}\left(\left\|\sum_{i=1}^n X_i\right\| > t\right) \le 14\frac{\mathrm{Tr}(\sum_{i=1}^n \mathbb{E}X_i^2)}{\sigma^2}\exp\left[-\frac{t^2/2}{\sigma^2 + tU/3}\right]. \tag{12}$$

One major difference compared to the matrix concentration inequality utilized in Lecture 2 is that the multiplicative factor on the right-hand side is now $\frac{\mathrm{Tr}(\sum_{i=1}^n \mathbb{E}X_i^2)}{\sigma^2}$. This factor essentially reflects the stable rank of the operators rather than their dimension, thereby rendering the bound dimension-free.

We remark that the above theorem is slightly tighter than the version stated in [Minsker, 2017, Theorem 3.1], where $\frac{\mathrm{Tr}(\sum_{i=1}^n \mathbb{E}X_i^2)}{\sigma^2}$ in (12) is replaced by $\frac{\mathrm{Tr}(\sum_{i=1}^n \mathbb{E}X_i^2)}{\|\sum_{i=1}^n \mathbb{E}X_i^2\|}$. However, [Minsker, 2017] in fact has proved the tighter presented here; we refer to the paragraphs below Eq. (3.9) in [Minsker, 2017] for more details. It is worth noting that a larger σ imposes a stricter condition on t, which limits the size of deviation for which we can provide concentration guarantees.

Proposition 3.9. *For any $\delta \in (0,1)$, if $\lambda \ge \lambda_{n,\delta}$, it holds w.p. at least $1-\delta$ that*

$$\|(\Sigma + \lambda)^{-1/2}(\Sigma - \widehat{\Sigma})(\Sigma + \lambda)^{-1/2}\| \le 1/4.$$

Proof. The proof follows from the observation that

$$(\Sigma + \lambda)^{-1/2}\widehat{\Sigma}(\Sigma + \lambda)^{-1/2} = \frac{1}{n}\sum_{i=1}^n (\Sigma + \lambda)^{-1/2}\varphi(x_i)\varphi(x_i)^\top(\Sigma + \lambda)^{-1/2} = \frac{1}{n}\sum_{i=1}^n z_i z_i^\top,$$

where $z_i = z(x_i) = (\Sigma + \lambda)^{-1/2}\varphi(x_i)$. Noting that

$$\begin{aligned}\|z(x)z(x)^\top\| = \|z_i\|^2 &\le F(\lambda)\\ \mathbb{E}\left[\sum (z_i z_i^\top)^2\right] = \mathbb{E}\left[\sum \|z_i\|^2 z_i z_i^\top\right] &\le F(\lambda)^2.\end{aligned}$$

Then, applying Theorem 3.8, we complete the proof. □

We shall mostly use the following corollary of Proposition 3.9:

Corollary 3.10. *For any $\delta \in (0,1)$ and $n \in \mathbb{N}$, if $\lambda \ge \lambda_{n,\delta}$, it holds with probability $1-\delta$ that*

$$\|(\widehat{\Sigma} + \lambda)^{-1/2}\Sigma^{1/2}\| \le \|(\widehat{\Sigma} + \lambda)^{-1/2}(\Sigma + \lambda)^{1/2}\| \le 2.$$

Proof. By Proposition 3.9, it holds w.p. at least $1-\delta$ that

$$\frac{1}{4}I - (\Sigma+\lambda)^{-1/2}(\Sigma-\widehat{\Sigma})(\Sigma+\lambda)^{-1/2} \succeq 0.$$

Thus,

$$\widehat{\Sigma}+\lambda \succeq \frac{3}{4}(\Sigma+\lambda) \Longrightarrow (\Sigma+\lambda)^{1/2}(\widehat{\Sigma}+\lambda)^{-1}(\Sigma+\lambda)^{1/2} \preceq \frac{4}{3}I.$$

Lastly,

$$\|(\widehat{\Sigma}+\lambda)^{-1/2}(\Sigma+\lambda)^{1/2}\| = \sqrt{\|(\Sigma+\lambda)^{1/2}(\widehat{\Sigma}+\lambda)^{-1}(\Sigma+\lambda)^{1/2}\|} \leq \sqrt{4/3}.$$

□

3.1.2 Step 2: Control the bias term

Noting

$$(\widehat{\Sigma}+\lambda)^{-1}\widehat{\Sigma}\theta^* - \theta^* = -\lambda(\widehat{\Sigma}+\lambda)^{-1}\theta^*,$$

we thus have

$$\begin{aligned} B(\lambda) &= \lambda\|\Sigma^{1/2}(\widehat{\Sigma}+\lambda)^{-1}\theta^*\| \leq \lambda\|\Sigma^{1/2}(\widehat{\Sigma}+\lambda)^{-1}\| \\ &= \lambda\|\Sigma^{1/2}(\widehat{\Sigma}+\lambda)^{-1/2}\|\|(\widehat{\Sigma}+\lambda)^{-1/2}\| \\ &\leq 2\lambda\|(\widehat{\Sigma}+\lambda)^{-1/2}\| \leq 2\lambda^{1/2}, \end{aligned}$$

where the third step uses Corollary 3.10.

3.1.3 Step 3: Control the variance term

To bound the variance termm we will need the following concentration inequality for quadratic form (see, e.g., [Vershynin, 2018, Theorem 6.2.1]):

Theorem 3.11 (Hanson-Wright inequality)**.** *Let $X = (X_1, \ldots, X_n)$ be a vector of independent, mean zero, sub-Gaussian R.V. such that $\|X_i\|_{\psi_2} \leq K$ for all $i = 1, \ldots, n$. Let A be an $n \times n$ symmetric matrix. Then, for any $t > 0$,*

$$\mathbb{P}\left(\left|X^\top AX - \mathbb{E}[X^\top AX]\right| \geq t\right) \leq 2\exp\left(-c\cdot\min\left(\frac{t^2}{K^4\|A\|_F^2}, \frac{t}{K^2\|A\|_{\text{op}}}\right)\right),$$

where $c > 0$ is a universal constant.

Let us examine the variance term:

$$V(\lambda)^2 = \left\|(\widehat{\Sigma}+\lambda)^{-1}\frac{1}{n}\Phi^\top\xi\right\|_\Sigma^2 = \frac{1}{n^2}\xi^\top\Phi(\widehat{\Sigma}+\lambda)^{-1}\Sigma(\widehat{\Sigma}+\lambda)^{-1}\Phi\xi =: \frac{1}{n}\xi^\top\hat{A}\xi,$$

where

$$\hat{A} = \frac{1}{n}\Phi(\widehat{\Sigma}+\lambda)^{-1}\Sigma(\widehat{\Sigma}+\lambda)^{-1}\Phi.$$

Noting

$$\text{Tr}[\hat{A}] = \text{Tr}[(\widehat{\Sigma}+\lambda)^{-1}\Sigma(\widehat{\Sigma}+\lambda)^{-1}\frac{1}{n}\Phi\Phi^\top] = \text{Tr}[(\widehat{\Sigma}+\lambda)^{-1}\Sigma(\widehat{\Sigma}+\lambda)^{-1}\widehat{\Sigma}]$$

$$= \|\Sigma^{1/2}(\widehat{\Sigma}+\lambda)^{-1}\widehat{\Sigma}^{1/2}\|_F^2 =: Q^2.$$

we have

$$\mathbb{E}_\xi[V(\lambda)^2] = \frac{\sigma^2}{n}\operatorname{Tr}[\hat{A}] = \frac{\sigma^2}{n}Q^2.$$

By the Hanson-Wright inequality, the essential step is to bound $\operatorname{Tr}[\hat{A}] = Q^2$, for which

$$\begin{aligned}
Q &\le \|\Sigma^{1/2}(\Sigma+\lambda)^{-1}\widehat{\Sigma}^{1/2}\|_F + \|\Sigma^{1/2}\left((\Sigma+\lambda)^{-1} - (\widehat{\Sigma}+\lambda)^{-1}\right)\widehat{\Sigma}^{1/2}\|_F \\
&= \|\Sigma^{1/2}(\Sigma+\lambda)^{-1}\widehat{\Sigma}^{1/2}\|_F + \|\Sigma^{1/2}(\Sigma+\lambda)^{-1}\left(\widehat{\Sigma}-\Sigma\right)(\widehat{\Sigma}+\lambda)^{-1}\widehat{\Sigma}^{1/2}\|_F \\
&= Q_1 + Q_2.
\end{aligned}$$

We next shall bound Q_1 and Q_2 separately.

Bound Q_1.

$$\begin{aligned}
Q_1^2 &\le \|(\Sigma+\lambda)^{-1/2}\widehat{\Sigma}^{1/2}\|_F^2 = \operatorname{Tr}[(\Sigma+\lambda)^{-1/2}\widehat{\Sigma}(\Sigma+\lambda)^{-1/2}] = \operatorname{Tr}[(\Sigma+\lambda)^{-1}\widehat{\Sigma}] \\
&= \frac{1}{n}\sum_{i=1}^n \operatorname{Tr}[(\Sigma+\lambda)^{-1/2}\varphi(x_i)\varphi(x_i)^\top(\Sigma+\lambda)^{-1/2}] \\
&= \frac{1}{n}\sum_{i=1}^n \|(\Sigma+\lambda)^{-1/2}\varphi(x_i)\|^2 \\
&\le F(\lambda). \qquad (13)
\end{aligned}$$

Bound Q_2.

$$\begin{aligned}
&\|\Sigma^{1/2}(\Sigma+\lambda)^{-1}\left(\widehat{\Sigma}-\Sigma\right)(\widehat{\Sigma}+\lambda)^{-1}\widehat{\Sigma}^{1/2}\|_F \\
&\quad\le \|(\Sigma+\lambda)^{-1/2}\left(\widehat{\Sigma}-\Sigma\right)(\widehat{\Sigma}+\lambda)^{-1}\widehat{\Sigma}^{1/2}\|_F \\
&\quad\le \|(\Sigma+\lambda)^{-1/2}\left(\widehat{\Sigma}-\Sigma\right)(\Sigma+\lambda)^{-1/2}\|\|(\Sigma+\lambda)^{1/2}(\widehat{\Sigma}+\lambda)^{-1}\widehat{\Sigma}^{1/2}\|_F \\
&\quad\le \frac{1}{4}\|(\Sigma+\lambda)^{1/2}(\widehat{\Sigma}+\lambda)^{-1}\widehat{\Sigma}^{1/2}\|_F \\
&\quad\le \frac{1}{4}\|(\widehat{\Sigma}+\lambda)^{-1/2}\widehat{\Sigma}^{1/2}\|_F \\
&\quad\le \frac{1}{4}\|(\Sigma+\lambda)^{-1/2}\widehat{\Sigma}^{1/2}\|_F \\
&\quad= \frac{1}{4}\sqrt{\operatorname{Tr}[(\Sigma+\lambda)^{-1}\widehat{\Sigma}]} = \frac{1}{4}F(\lambda)^{1/2}. \qquad (14)
\end{aligned}$$

Combining (13) and (14), we obtain

$$\operatorname{Tr}[\hat{A}] \lesssim F(\lambda).$$

Thus, it is easy to obtain

$$\begin{aligned}
\|\hat{A}\|_{\mathrm{op}} &\le \operatorname{Tr}[\hat{A}] \lesssim F(\lambda) \\
\|\hat{A}\|_F^2 &= \|\hat{A}\|_{\mathrm{op}}\operatorname{Tr}[\hat{A}] \lesssim F(\lambda)^2.
\end{aligned}$$

Then, applying the Hanson-Wright inequality, we know w.p. $1-\delta$, we have

$$\xi^\top\hat{A}\xi \le \mathbb{E}[\xi^\top\hat{A}\xi] + F(\lambda)\log(1/\delta) \le F(\lambda)(1+\log(1/\delta)).$$

3.1.4 Step 4: Putting all together

We obtain w.p. at least $1-\delta$ that

$$\|\hat{f}_\lambda - f^*\|_{L^2(\rho)}^2 \lesssim B^2(\lambda) + V^2(\lambda) \lesssim \lambda + \frac{\sigma^2(1+\log(1/\delta))}{n} F(\lambda).$$

3.2 Smooth adaptation

3.3 The source condition

Definition 3.12 (Sobolev-type interpolation spaces)**.** Let $\mathcal{H}^s = \left\{\sum_{j=1}^\infty a_j \lambda_j^{s/2} e_j : \sum_{j=1}^\infty a_j^2 < \infty\right\}$ equipped with inner product

$$\left\langle \sum_{j=1}^\infty a_j \lambda_j^{s/2} e_j, \sum_{j=1}^\infty b_j \lambda_j^{s/2} e_j \right\rangle_{\mathcal{H}^s} = \sum_{j=1}^\infty a_j b_j.$$

Note that when $s=1$, $\mathcal{H}^1 = \mathcal{H}_k$, when $s=0$, $\mathcal{H}^1 = L^2(\rho)$ (assuming the eigenfunctions $\{e_j\}_{j=1}^\infty$ forms a complete basis of $L^2(\rho)$). Thus, $\mathcal{H}^s$ defines function spaces interpolate "between" $L^2(\rho)$ and $\mathcal{H}_k$. It is often referred "Sobolev-type" as it generalizes the classic Sobolev spaces. Consider the periodic kernel $k(x-x') = \kappa(x-x')$, for which the eigenfunctions are $e_j(x) := e^{2\pi \mathrm{i} j x}$ for $j \in \mathbb{Z}$. Suppose that

$$\hat{\kappa}(j) \asymp (1+|j|^2)^{-1}.$$

Then, we have $\mathcal{H}_k = H^1(\mathbb{T})$, the Sobolev space with the first-order weak derivate belonging to $L^2(\mathbb{T})$. Then,

$$\|f\|_{\mathcal{H}^s}^2 = \sum_{j\in\mathbb{Z}} \frac{\langle f, e_j\rangle_{L^2}^2}{\lambda_j^s} = \sum_{j\in\mathbb{Z}} (1+|j|^2)^s \hat{f}(j)^2 = \|f\|_{H^s(\mathbb{T})}^2.$$

We make the following assumption:

Assumption 3.13 (Source condition)**.** $f^* \in \mathcal{H}_k^s$.

Here, s describes the relative smoothness of f^* for our model $\mathcal{H}_k$. We shall show that KRR can adapt to the relative smoothness f^*. The case of $s \geq 1$ is referred to as the well-specified case, while $s < 1$, it means that f^* is less smoothness than the model, a regime referred to the misspecified case as $f^* \notin \mathcal{H}_k$. Notably, when $s > 1$, it means that f^* has extra smoothness.

3.3.1 KRR learns smoother functions more efficiently

Suppose that the target function f^* is in $\mathcal{H}^s$ for some $s \in [1,2)$, and we perform KRR in $\mathcal{H}^1$. Let $\varphi(x) = (\sqrt{\lambda_1}e_1(x), \cdots, \sqrt{\lambda_p}e_p(x))$ be the feature map. The target function can be represented as

$$f^*(x) = \varphi(x)^\top \theta^* = \sum_{j=1}^p \theta_j^* \sqrt{\lambda_j} e_j(x)$$

where

$$\theta_j^* = \lambda_j^{\frac{s-1}{2}} a_j^* \iff \theta^* = \Sigma^{\frac{s-1}{2}} a^*$$

for some $a^* \in \ell_2^p$. We assume $\|a^*\| \le 1$. Then, $\Sigma = \mathrm{diag}(\lambda_1, \cdots, \lambda_p)$ is the covariance matrix.

Recall that in the bias-variance decomposition, only the bias term relies on the target θ^*. We then provide an intuitive derivation by assuming $\widehat{\Sigma} \approx \Sigma$:

$$\begin{aligned} B(\lambda) &= \|\Sigma^{1/2}\lambda(\widehat{\Sigma}+\lambda)^{-1}\theta^*\| \\ &= \|\Sigma^{1/2}\lambda(\widehat{\Sigma}+\lambda)^{-1}\Sigma^{\frac{s-1}{2}}a^*\| \\ &\le \|\Sigma^{1/2}\lambda(\widehat{\Sigma}+\lambda)^{-1}\Sigma^{\frac{s-1}{2}}\| \\ &\approx \|\Sigma^{1/2}\lambda(\Sigma+\lambda)^{-1}\Sigma^{\frac{s-1}{2}}\| \\ &\le \lambda \max_j \frac{\lambda_j^{\frac{s}{2}}}{\lambda_j+\lambda} \\ &\le \lambda \max_{t\in[0,\lambda_1]} \frac{t^{\frac{s}{2}}}{t+\lambda}. \end{aligned} \tag{15}$$

Lemma 3.14. *For any $s \ge 0$, $\max_{t\in[0,\lambda_1]} \frac{t^{\frac{s}{2}}}{t+\lambda} \lesssim \lambda^{\min(s,2)/2-1}$.*

Proof. Let $h(t) = \frac{t^\alpha}{t+\lambda}$. Then,

$$h'(t) = \frac{\alpha t^{\alpha-1}(t+\lambda) - t^\alpha}{(t+\lambda)^2} = \frac{(1-\alpha)t^{\alpha-1}}{(t+\lambda)^2}\left(\frac{\lambda\alpha}{1-\alpha} - t\right).$$

For $\alpha \in (0,1)$, f is increasing in $[0,t^*]$ and decreasing in $[t^*,\infty)$ with $t^* = \lambda\alpha/(1-\alpha) =: c_\alpha\lambda$. Thus,

$$h(t) \le h(t^*) = c_\alpha^\alpha \frac{\lambda^\alpha}{c_\alpha\lambda+\lambda} = \frac{c_\alpha^\alpha}{1+c_\alpha}\lambda^{\alpha-1}.$$

For $\alpha \ge 1$, h is monotonically increasing. Therefore

$$\max_{t\in[0,\lambda_1]} \frac{t^{\frac{s}{2}}}{t+\lambda} \lesssim \lambda^{\min(s,2)/2-1}.$$

□

Using the above lemma, we obtain that when n is sufficiently large,

$$B(\lambda) = O(\lambda^{s/2}) \tag{16}$$

Combining with the variance term (which does not change), we have

$$\|\hat{f}_\lambda - f^*\|_{L^2(\rho)}^2 \lesssim B^2(\lambda) + V^2(\lambda) \lesssim \lambda^s + \frac{\sigma^2}{n}F(\lambda)$$

where we have ignored the dependence on δ. Recall that when $\lambda_j \asymp j^{-\beta}$ with $\beta > 1$ and the eigenfunctions are uniformly bounded, we have $F(\lambda) \le \lambda^{-1/\beta}$. Finally we obtain the optimal λ and the convergence rate as

$$\lambda_{\mathrm{op}} \propto n^{-\frac{\beta}{s\beta+1}}, \qquad \|\hat{f}_\lambda - f^*\|_{L^2(\rho)}^2 \lesssim n^{-\frac{s\beta}{s\beta+1}}.$$

The learning benefits from the smoothness since the rate improves from $n^{-\frac{\beta}{\beta+1}}$ to $n^{-\frac{s\beta}{s\beta+1}}$ for $s \in [1,2]$.

Remark 3.15. *When $s > 2$, the last equation in (15) is increasing in t. The maximum is attained at $t = \lambda_1$ and $B(\lambda) = O(\lambda)$. This is the same rate as in the case $s = 2$. Therefore, the rate cannot be further improved for $s > 2$.*

Remark 3.16. *The convergence rate $n^{-\frac{s\beta}{s\beta+1}}$ reveals that the model class (an RKHS with $s = 1$) can leverage the smoothness of the target function f^* up to order $s = 2$. Why can a model class with first-order smoothness detect up to second-order smoothness in the target function, yet not benefit from higher-order smoothness?*

A concrete illustration is provided by approximating a function $f \in C^\infty[a, b]$ via piecewise linear functions. In this scenario, we have

$$|f(x) - f(a) - f'(a)(x - a)| = \frac{|f''(\xi)|}{2}(x - a)^2$$

where $\xi \in (a, b)$. This bound demonstrates that approximation error using first-order smooth functions is governed by the second-order smoothness of the target function. Consequently, information regarding any higher-order smoothness (e.g., third-order or above) remains unexploited by a model class constrained to first-order smoothness.

3.3.2 Choosing a smoother kernel according to target function

We make an informal argument to show how we should choose the kernel according to the smoothness of the target function. Suppose the target f^* lies in a γ-th order Sobolev space $H^\gamma(\mathbb{T})$. Specifically,

$$f^*(x) = \sum_j a_j (1 + |j|^2)^{-\frac{\gamma}{2}} e_j(x). \tag{17}$$

where $\sum_j a_j^2 < \infty$. For simplicity, let us naively suppose $a_j \sim j^{-1/2}$. Next, we consider using KRR to learn f^*. Let the kernel be

$$k(x, x') = \sum_j \lambda_j e_j(x) e_j(x')$$

where $\lambda_j \sim j^{-\beta}$. In order for KRR to incur no approximation error, f^* must lie in $\mathcal{H}_k$, defined by

$$\mathcal{H}_k^s = \left\{ \sum_{j=1}^\infty b_j \lambda_j^{s/2} e_j : \sum_{j=1}^\infty b_j^2 < \infty \right\}$$

If we again assume $b_j \sim j^{-1/2}$, we obtain

$$f^*(x) = \sum_j b_j \lambda_j^{s/2} e_j \sim \sum_j a_j j^{-\frac{s\beta}{2}} e_j.$$

Comparing this with (17) leads to

$$j^{-\gamma} \sim j^{-\frac{s\beta}{2}} \implies s = \frac{2\gamma}{\beta}.$$

This equation reveals: the smoother the kernel (i.e., the larger β), the smaller the corresponding smoothness parameter s in an RKHS. According to the analysis above, given a kernel with eigenvalue decay with $\lambda_j \sim j^{-\beta}$, we can bound the bias term by $O(\lambda^{\tilde{s}})$ where

$$\tilde{s} = \min\left\{2, \frac{2\gamma}{\beta}\right\}$$

Therefore, we should at least set $\beta = \gamma$ so that KRR can fully adapt to the smoothness of the target function.

Considering the Gaussian kernel as

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2}\right),$$

whose eigenvalues decay on the order of

$$\lambda_j \asymp \exp(-cj \log j) \text{ as } j \to +\infty,$$

which effectively corresponds to $\beta = \infty$. Therefore, the Gaussian kernel can adapt to any level of smoothness in the target function, helping to explain its enduring popularity in KRR and general kernel-based methods.