A. Preliminaries

In this section, we provide some notation and preliminary results that will be used throughout the appendix. Henceforth, we denote the Euclidean norm of a vector $a \in \mathbb{R}^n$ with $\|a\|_2$ and the operator norm of a matrix $A \in \mathbb{R}^{n_1 \times n_2}$ with $\|A\|_2$. Furthermore, we denote with $\|A\|_F$ the Frobenious norm of a matrix or operator A. Let $\mathcal{H}$ be a Hilbert space with $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ as its inner product and $\| \cdot \|_{\mathcal{H}}$ as its norm. We use $\mathrm{Tr}(\cdot)$ to denote the trace of an operator or a matrix. Given a measure $d\rho$, we use $L_2(d\rho)$ to denote the space of square-integrable functions with respect to $d\rho$.

Lemma 1. *(Bernstein inequality, Tropp, 2015, Corollary 7.3.3) Let $\mathbf{R}$ be a fixed $d_1 \times d_2$ matrix over the set of complex/real numbers. Suppose that $\{\mathbf{R}_1, \cdots, \mathbf{R}_n\}$ is an independent and identically distributed sample of $d_1 \times d_2$ matrices such that*

$$\mathbb{E}[\mathbf{R}_i] = \mathbf{R} \qquad \text{and} \qquad \|\mathbf{R}_i\|_2 \leq L,$$

where $L > 0$ is a constant independent of the sample. Furthermore, let $\mathbf{M}_1, \mathbf{M}_2$ be semidefinite upper bounds for the matrix-valued variances

$$\mathrm{Var}_1[\mathbf{R}_i] \preceq \mathbb{E}[\mathbf{R}_i \mathbf{R}_i^T] \preceq \mathbf{M}_1$$
$$\mathrm{Var}_2[\mathbf{R}_i] \preceq \mathbb{E}[\mathbf{R}_i^T \mathbf{R}_i] \preceq \mathbf{M}_2.$$

Let $m = \max(\|\mathbf{M}_1\|_2, \|\mathbf{M}_2\|_2)$ and $d = {}^{Tr(\mathbf{M}_1)+Tr(\mathbf{M}_2)}/_{m}$. Then, for $\epsilon \geq \sqrt{m/n} + 2L/3n$, we can bound

$$\bar{\mathbf{R}}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{R}_i$$

around its mean using the concentration inequality

$$P(\|\bar{\mathbf{R}}_n - \mathbf{R}\|_2 \geq \epsilon) \leq 4d \exp\left(\frac{-n\epsilon^2/2}{m + 2L\epsilon/3} \right).$$

To characterize the stability of a learning algorithm, we need to take into account the complexity of the space of functions. Below, we introduce a particular measure of the complexity over function spaces known as *Rademacher averages*.

Definition 1. *Let P_x be a probability distribution on a set $\mathcal{X}$ and suppose that $\{x_1 \cdots, x_n\}$ are independent samples selected according to P_x. Let $\mathcal{H}$ be a class of functions mapping $\mathcal{X}$ to $\mathbb{R}$. Then, the random variable known as the empirical Rademacher average is defined as*

$$\hat{R}_n(\mathcal{H}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{H}} \left| \frac{2}{n} \sum_{i=1}^{n} \sigma_i f(x_i) \right| \mid x_1, \cdots, x_n \right]$$

where $\sigma_1, \cdots, \sigma_n$ are independent uniform $\{\pm 1\}$-valued random variables. The corresponding Rademacher average is then defined as the expectation of the empirical Rademacher average, i.e.,

$$R_n(\mathcal{H}) = \mathbb{E}\Big[\hat{R}_n(\mathcal{H}) \Big].$$

Lemma 2. *(Bartlett & Mendelson, 2002) Let $\mathcal{H}$ be a reproducing kernel Hilbert space of functions mapping from $\mathcal{X}$ to $\mathbb{R}$ that corresponds to a positive definite kernel k. Let $\mathcal{H}_0$ be the unit ball of $\mathcal{H}$, centered at the origin. Then, we have that $R_n(\mathcal{H}_0) \leq (1/n)\mathbb{E}_X \sqrt{Tr(\mathbf{K})}$, where $\mathbf{K}$ is the Gram matrix for kernel k over an independent and identically distributed sample $X = \{x_1, \cdots, x_n\}$.*

The next lemma states that the expected risk convergence rate of a particular estimator in $\mathcal{H}$ not only depends on the number of data points, but also on the complexity of $\mathcal{H}$.

Lemma 3. *(Bartlett & Mendelson, 2002, Theorem 8) Let $\{x_i, y_i\}_{i=1}^n$ be an independent and identically distributed sample from a probability measure P defined on $\mathcal{X} \times \mathcal{Y}$ and let $\mathcal{H}$ be the space of functions mapping from $\mathcal{X}$ to $\mathcal{A}$. Denote a loss function with $\mathbf{L} : \mathcal{Y} \times \mathcal{A} \to [0, 1]$ and define the expected risk function for all $f \in \mathcal{H}$ to be $\mathcal{E}(f) = \mathbb{E}_P(\mathbf{L}(y, f(x)))$, together with the corresponding empirical risk function $\hat{\mathcal{E}}(f) = (1/n) \sum_{i=1}^n \mathbf{L}(y_i, f(x_i))$. Then, for a sample size n, for all $f \in \mathcal{H}$ and $\delta \in (0, 1)$, with probability $1 - \delta$, we have that*

$$\mathcal{E}(f) \leq \hat{\mathcal{E}}(f) + R_n(\tilde{\mathbf{L}} \circ \mathcal{H}) + \sqrt{\frac{8 \log(2/\delta)}{n}}$$

where $\tilde{\mathbf{L}} \circ \mathcal{H} = \{(x, y) \to \mathbf{L}(y, f(x)) - \mathbf{L}(y, 0) \mid f \in \mathcal{H}\}$.

Similar to Rudi & Rosasco (2017) and Caponnetto & De Vito (2007), we have assumed the existence of $f_{\mathcal{H}} \in \mathcal{H}$ such that $f_{\mathcal{H}} = \inf_{f \in \mathcal{H}} \mathcal{E}(f)$. The assumption implies that there exists some ball of radius $R > 0$ containing $f_{\mathcal{H}}$ in its interior. Our theoretical results do not require prior knowledge of this constant and hold uniformly over all finite radii. To simplify our derivations and constant terms in our bounds, we have (without loss of generality) assumed that $R = 1$.

B. Upper bound on the approximation function norm

Proposition 1. *Assume that the reproducing kernel Hilbert space $\mathcal{H}$ with kernel k admits a decomposition as in Eq. (2) and let $\tilde{\mathcal{H}} \coloneqq \{\tilde{f} \mid \tilde{f} = \sum_{i=1}^s \alpha_i z(v_i, \cdot), \forall \alpha_i \in \mathbb{R}\}$. Then, for all $\tilde{f} \in \tilde{\mathcal{H}}$ it holds that $\|\tilde{f}\|_{\tilde{\mathcal{H}}}^2 \leq s\|\alpha\|_2^2$, where $\tilde{\mathcal{H}}$ is the reproducing kernel Hilbert space with kernel $\tilde{k}$ (see Eq. 3).*

Proof. Let us define a space of functions as

$$\mathcal{H}_1 \coloneqq \{f \mid f(x) = \alpha z(v, x), \alpha \in \mathbb{R}\}.$$

We now show that $\mathcal{H}_1$ is a reproducing kernel Hilbert space with kernel defined as $k_1(x, y) = (1/s) z(v, x) z(v, y)$, where s is a constant.

Define a map $M : \mathbb{R} \to \mathcal{H}_1$ such that $M\alpha = \alpha z(v, \cdot), \forall \alpha \in \mathbb{R}$. The map M is a bijection, i.e. for any $f \in \mathcal{H}_1$ there

exists a unique $\alpha_f \in \mathbb{R}$ such that $M^{-1}f = \alpha_f$. Now, we define an inner product on $\mathcal{H}_1$ as

$$\langle f, g\rangle_{\mathcal{H}_1} = \langle \sqrt{s}M^{-1}f, \sqrt{s}M^{-1}g\rangle_{\mathbb{R}} = s\alpha_f\alpha_g.$$

It is easy to show that this is a well defined inner product and, thus, $\mathcal{H}_1$ is a Hilbert space.

For any instance y, $k_1(\cdot, y) = (1/s)z(v, \cdot)z(v, y) \in \mathcal{H}_1$, since $(1/s)z(v, x) \in \mathbb{R}$ by definition. Take any $f \in \mathcal{H}_1$ and observe that

$$\begin{aligned}\langle f, k_1(\cdot, y)\rangle_{\mathcal{H}_1} &= \langle \sqrt{s}M^{-1}f, \sqrt{s}M^{-1}k_1(\cdot, y)\rangle_{\mathbb{R}} \\ &= s\langle \alpha_f, 1/s z(v, y)\rangle_{\mathbb{R}} \\ &= \alpha_f z(v, y) = f(y).\end{aligned}$$

Hence, we have demonstrated the reproducing property for $\mathcal{H}_1$ and $\|f\|_{\mathcal{H}_1} = s\alpha_f^2$.

Now, suppose we have a sample of features $\{v_i\}_{i=1}^s$. For each v_i, we define the reproducing kernel Hilbert space

$$\mathcal{H}_i := \{f \mid f(x) = \alpha z(v_i, x), \alpha \in \mathbb{R}\}$$

with the kernel $k_i(x, y) = (1/s)z(v_i, x)z(v_i, y)$.

Denoting with

$$\tilde{\mathcal{H}} = \oplus_{i=1}^s \mathcal{H}_i = \{\tilde{f} : \tilde{f} = \sum_{i=1}^s f_i, f_i \in \mathcal{H}_i\}$$

and using the fact that the direct sum of reproducing kernel Hilbert spaces is another reproducing kernel Hilbert space (Berlinet & Thomas-Agnan, 2011), we have that $\tilde{k}(x, y) = \sum_{i=1}^s k_i(x, y) = (1/s)\sum_{i=1}^s z(v_i, x)z(v_i, y)$ is the kernel of $\tilde{\mathcal{H}}$ and that the norm of $\tilde{f} \in \tilde{\mathcal{H}}$ is defined as

$$\min_{f_i \in \mathcal{H}_i \mid f = \sum_{i=1}^s f_i} \sum_{i=1}^s \|f_i\|_{\mathcal{H}_i} =$$

$$\min_{\alpha_i \in \mathbb{R} \mid f_i = \alpha_i z(v_i, \cdot)} \sum_{i=1}^s s\alpha_i^2 = \min_{\alpha_i \in \mathbb{R} \mid f_i = \alpha_i z(v_i, \cdot)} s\|\alpha\|_2^2.$$

Hence, we have that $\|\tilde{f}\|_{\tilde{\mathcal{H}}} \leq s\|\alpha\|_2^2$. □

C. Proofs of Theorems 1 and 3

Before we prove Theorems 1 and 3, we give a general result that provides an upper bound on the approximation error between any function $f \in \mathcal{H}$ and its estimator based on random Fourier features.

C.1. Auxiliary Results

As discussed in Section 2, we would like to approximate a function $f \in \mathcal{H}$ at observation points with preferably as small function norm as possible. The estimation of $\mathbf{f}_x$ can be formulated as the following optimization problem:

$$\min_{\beta \in \mathbb{R}^s} \frac{1}{n}\|\mathbf{f}_x - \mathbf{Z}_q\beta\|_2^2 + \lambda s\|\beta\|_2^2.$$

The following theorem provides the desired upper bound on the approximation error of the estimator based on random Fourier features.

Theorem 5. *Let $\lambda_1 \geq \cdots \geq \lambda_n$ be the eigenvalues of the kernel matrix $\mathbf{K}$ and assume that the regularization parameter satisfies $0 \leq n\lambda \leq \lambda_1$. Let $\tilde{l} : \mathcal{V} \to \mathbb{R}$ be a measurable function such that $\tilde{l}(v) \geq l_\lambda(v)$ $(\forall v \in \mathcal{V})$ and*

$$d_{\tilde{l}} = \int_{\mathcal{V}} \tilde{l}(v)dv < \infty.$$

Suppose $\{v_i\}_{i=1}^s$ are sampled independently according to probability density function $q(v) = \tilde{l}(v)/d_{\tilde{l}}$. If

$$s \geq 5d_{\tilde{l}} \log \frac{16d_{\mathbf{K}}^\lambda}{\delta},$$

then for all $\delta \in (0, 1)$ and $\|f\|_{\mathcal{H}} \leq 1$, with probability greater than $1 - \delta$, we have that it holds

$$\sup_{\|f\|_{\mathcal{H}} \leq 1} \inf_{\sqrt{s}\|\beta\|_2 \leq \sqrt{2}} \frac{1}{n}\|\mathbf{f}_x - \mathbf{Z}_q\beta\|_2^2 \leq 2\lambda. \qquad (12)$$

The following two lemmas are required for our proof of Theorem 5, presented subsequently.

Lemma 4. *Suppose that the assumptions from Theorem 5 hold and let $\epsilon \geq \sqrt{m/s} + 2L/3s$ with constants m and L (see the proof for explicit definition). If the number of features*

$$s \geq d_{\tilde{l}}\left(\frac{1}{\epsilon^2} + \frac{2}{3\epsilon}\right) \log \frac{16d_{\mathbf{K}}^\lambda}{\delta},$$

then for all $\delta \in (0, 1)$, with probability greater than $1 - \delta$,

$$-\epsilon\mathbf{I} \preceq (\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}(\tilde{\mathbf{K}} - \mathbf{K})(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}} \preceq \epsilon\mathbf{I}.$$

Proof. Following the derivations in Avron et al. (2017), we utilize the matrix Bernstein concentration inequality to prove the result. More specifically, we observe that

$$(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}\tilde{\mathbf{K}}(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}} =$$

$$\frac{1}{s}\sum_{i=1}^s (\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}\mathbf{z}_{q,v_i}(\mathbf{x})\mathbf{z}_{q,v_i}(\mathbf{x})^T(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}} =$$

$$\frac{1}{s}\sum_{i=1}^s \mathbf{R}_i =: \bar{\mathbf{R}}_s,$$

with

$$\mathbf{R}_i = (\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}\mathbf{z}_{q,v_i}(\mathbf{x})\mathbf{z}_{q,v_i}(\mathbf{x})^T(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}.$$

Now, observe that

$$\mathbf{R} = \mathbb{E}[\mathbf{R}_i] = (\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}\mathbf{K}(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}.$$

The operator norm of $\mathbf{R}_i$ is equal to

$$\|(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}\mathbf{z}_{q,v_i}(\mathbf{x})\mathbf{z}_{q,v_i}(\mathbf{x})^T(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}\|_2.$$

As $\mathbf{z}_{q,v_i}(\mathbf{x})\mathbf{z}_{q,v_i}(\mathbf{x})^T$ is a rank one matrix, we have that the operator norm of this matrix is equal to its trace, i.e.,

$$\begin{aligned}
&\|\mathbf{R}_i\|_2 = \\
&\mathrm{Tr}((\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}\mathbf{z}_{q,v_i}(\mathbf{x})\mathbf{z}_{q,v_i}(\mathbf{x})^T(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}) = \\
&\frac{p(v_i)}{q(v_i)}\mathrm{Tr}((\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}\mathbf{z}_{v_i}(\mathbf{x})\mathbf{z}_{v_i}(\mathbf{x})^T(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}) = \\
&\frac{p(v_i)}{q(v_i)}\mathrm{Tr}(\mathbf{z}_{v_i}(\mathbf{x})^T(\mathbf{K} + n\lambda\mathbf{I})^{-1}\mathbf{z}_{v_i}(\mathbf{x})) = \\
&\frac{l_\lambda(v_i)}{q(v_i)} =: L_i \quad \text{and} \quad L \coloneqq \sup_i L_i.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\mathbf{R}_i\mathbf{R}_i^T = \\
&(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}\mathbf{z}_{q,v_i}(\mathbf{x})\mathbf{z}_{q,v_i}(\mathbf{x})^T(\mathbf{K} + n\lambda\mathbf{I})^{-1}\mathbf{z}_{q,v_i}(\mathbf{x}) \\
&\cdot \mathbf{z}_{q,v_i}(\mathbf{x})^T(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}} = \\
&\frac{p(v_i)l_\lambda(v_i)}{q^2(v_i)}(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}\mathbf{z}_{v_i}(\mathbf{x})\mathbf{z}_{v_i}(\mathbf{x})^T(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}} \preceq \\
&\frac{\tilde{l}(v_i)}{q(v_i)}\frac{p(v_i)}{q(v_i)}(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}\mathbf{z}_{v_i}(\mathbf{x})\mathbf{z}_{v_i}(\mathbf{x})^T(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}} = \\
&d_{\tilde{l}}\frac{p(v_i)}{q(v_i)}(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}\mathbf{z}_{v_i}(\mathbf{x})\mathbf{z}_{v_i}(\mathbf{x})^T(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}.
\end{aligned}$$

From the latter inequality, we obtain that

$$\mathbb{E}[\mathbf{R}_i\mathbf{R}_i^T] \preceq d_{\tilde{l}}(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}\mathbf{K}(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}} =: \mathbf{M}_1.$$

We also have the following two inequalities

$$\begin{aligned}
m &= \|\mathbf{M}_1\|_2 = d_{\tilde{l}}\frac{\lambda_1}{\lambda_1 + n\lambda} =: d_{\tilde{l}}d_1 \\
d &= \frac{2\,\mathrm{Tr}(\mathbf{M}_1)}{m} = 2\frac{\lambda_1 + n\lambda}{\lambda_1}d_{\mathbf{K}}^\lambda = 2d_1^{-1}d_{\mathbf{K}}^\lambda.
\end{aligned}$$

We are now ready to apply the matrix Bernstein concentration inequality. More specifically, for $\epsilon \geq \sqrt{m/s} + 2L/3s$ and for all $\delta \in (0,1)$, with probability $1-\delta$, we have that

$$\begin{aligned}
\mathrm{P}(\|\bar{\mathbf{R}}_s - \mathbf{R}\|_2 \geq \epsilon) &\leq 4d\exp\left(\frac{-s\epsilon^2/2}{m + 2L\epsilon/3}\right) \\
&= 8d_1^{-1}d_{\mathbf{K}}^\lambda\exp\left(\frac{-s\epsilon^2/2}{d_{\tilde{l}}d_1 + d_{\tilde{l}}2\epsilon/3}\right) \\
&\leq 16d_{\mathbf{K}}^\lambda\exp\left(\frac{-s\epsilon^2}{d_{\tilde{l}}(1 + 2\epsilon/3)}\right) \leq \delta.
\end{aligned}$$

In the third line, we have used the assumption that $n\lambda \leq \lambda_1$ and, consequently, $d_1 \in [1/2, 1)$. □

Remark: We note here that the two considered sampling strategies lead to two different results. In particular, if we let $\tilde{l}(v) = l_\lambda(v)$ then $q(v) = l_\lambda(v)/d_{\mathbf{K}}^\lambda$, i.e., we are sampling proportional to the ridge leverage scores. Thus, the leverage weighted random Fourier features sampler requires

$$s \geq d_{\mathbf{K}}^\lambda\left(\frac{1}{\epsilon^2} + \frac{2}{3\epsilon}\right)\log\frac{16d_{\mathbf{K}}^\lambda}{\delta}. \tag{13}$$

Alternatively, we can opt for the plain random Fourier feature sampling strategy by taking $\tilde{l}(v) = z_0^2 p(v)/\lambda$, with $l_\lambda(v) \leq z_0^2 p(v)/\lambda$. Then, the plain random Fourier features sampling scheme requires

$$s \geq \frac{z_0^2}{\lambda}\left(\frac{1}{\epsilon^2} + \frac{2}{3\epsilon}\right)\log\frac{16d_{\mathbf{K}}^\lambda}{\delta}. \tag{14}$$

Thus, the leverage weighted random Fourier features sampling scheme can dramatically change the required number of features, required to achieve a predefined matrix approximation error in the operator norm.

Lemma 5. *Let $f \in \mathcal{H}$, where $\mathcal{H}$ is the RKHS associated with a kernel k. Let $x_1, \cdots, x_n \in \mathcal{X}$ be a set of instances with $x_i \neq x_j$ for all $i \neq j$. Denote with $\mathbf{f}_x = [f(x_1), \cdots, f(x_n)]^T$ and let $\mathbf{K}$ be the Gram-matrix of the kernel k given by the provided set of instances. Then,*

$$\mathbf{f}_x^T\mathbf{K}^{-1}\mathbf{f}_x \leq 1.$$

Proof. For a vector $\mathbf{a} \in \mathbb{R}^n$ we have that

$$\begin{aligned}
\mathbf{a}^T\mathbf{f}_x\mathbf{f}_x^T\mathbf{a} = \left(\mathbf{f}_x^T\mathbf{a}\right)^2 &= \Big(\sum_{i=1}^n a_i f(x_i)\Big)^2 \\
&= \Big(\sum_{i=1}^n a_i\int_{\mathcal{V}} g(v)z(v,x_i)d\tau(v)\Big)^2 \\
&= \Big(\int_{\mathcal{V}} g(v)\mathbf{z}_v(\mathbf{x})^T\mathbf{a}\,d\tau(v)\Big)^2 \\
&\leq \int_{\mathcal{V}} g(v)^2 d\tau(v)\int_{\mathcal{V}}(\mathbf{z}_v(\mathbf{x})^T\mathbf{a})^2\,d\tau(v) \\
&= \int_{\mathcal{V}}\mathbf{a}^T\mathbf{z}_v(\mathbf{x})\mathbf{z}_v(\mathbf{x})^T\mathbf{a}\,d\tau(v) \\
&= \mathbf{a}^T\int_{\mathcal{V}}\mathbf{z}_v(\mathbf{x})\mathbf{z}_v(\mathbf{x})^T\,d\tau(v)\,\mathbf{a} \\
&= \mathbf{a}^T\mathbf{K}\mathbf{a}.
\end{aligned}$$

The third equality is due to the fact that, for all $f \in \mathcal{H}$, we have that $f(x) = \int_{\mathcal{V}} g(v)z(v,x)p(v)dv$ ($\forall x \in \mathcal{X}$) and

$$\|f\|_{\mathcal{H}} = \min_{\left\{g \mid f(x) = \int_{\mathcal{V}} g(v)z(v,x)p(v)dv\right\}} \|g\|_{L_2(d\tau)}.$$

The first inequality, on the other hand, follows from the Cauchy-Schwarz inequality. The bound implies that $\mathbf{f}_x\mathbf{f}_x^T \preceq \mathbf{K}$ and, consequently, we derive $\mathbf{f}_x^T\mathbf{K}^{-1}\mathbf{f}_x \leq 1$. □

C.1.1. Proof of Theorem 5

Proof. Our goal is to minimize the following objective:

$$\frac{1}{n}\|\mathbf{f}_x - \mathbf{Z}_q\beta\|_2^2 + s\lambda\|\beta\|_2^2. \tag{15}$$

To find the minimizer, we can directly take the derivative with respect to β and, thus, obtain

$$\begin{aligned}\beta &= \frac{1}{s}(\frac{1}{s}\mathbf{Z}_q^T\mathbf{Z}_q + n\lambda\mathbf{I})^{-1}\mathbf{Z}_q^T\mathbf{f}_x \\ &= \frac{1}{s}\mathbf{Z}_q^T(\frac{1}{s}\mathbf{Z}_q\mathbf{Z}_q^T + n\lambda\mathbf{I})^{-1}\mathbf{f}_x \\ &= \frac{1}{s}\mathbf{Z}_q^T(\tilde{\mathbf{K}} + n\lambda\mathbf{I})^{-1}\mathbf{f}_x,\end{aligned}$$

where the second equality follows from the Woodbury inversion lemma.

Substituting β into Eq. (15), we transform the first part as

$$\begin{aligned}\frac{1}{n}\|\mathbf{f}_x - \mathbf{Z}_q\beta\|_2^2 &= \frac{1}{n}\|\mathbf{f}_x - \frac{1}{s}\mathbf{Z}_q\mathbf{Z}_q^T(\tilde{\mathbf{K}} + n\lambda\mathbf{I})^{-1}\mathbf{f}_x\|_2^2 \\ &= \frac{1}{n}\|\mathbf{f}_x - \tilde{\mathbf{K}}(\tilde{\mathbf{K}} + n\lambda\mathbf{I})^{-1}\mathbf{f}_x\|_2^2 \\ &= \frac{1}{n}\|n\lambda(\tilde{\mathbf{K}} + n\lambda\mathbf{I})^{-1}\mathbf{f}_x\|_2^2 \\ &= n\lambda^2\mathbf{f}_x^T(\tilde{\mathbf{K}} + n\lambda\mathbf{I})^{-2}\mathbf{f}_x.\end{aligned}$$

On the other hand, the second part can be transformed as

$$\begin{aligned}s\lambda\|\beta\|_2^2 &= s\lambda\frac{1}{s^2}\mathbf{f}_x^T(\tilde{\mathbf{K}} + n\lambda\mathbf{I})^{-1}\mathbf{Z}_q\mathbf{Z}_q^T(\tilde{\mathbf{K}} + n\lambda\mathbf{I})^{-1}\mathbf{f}_x \\ &= \lambda\mathbf{f}_x^T(\tilde{\mathbf{K}} + n\lambda\mathbf{I})^{-1}\tilde{\mathbf{K}}(\tilde{\mathbf{K}} + n\lambda\mathbf{I})^{-1}\mathbf{f}_x \\ &= \lambda\mathbf{f}_x^T(\tilde{\mathbf{K}} + n\lambda\mathbf{I})^{-1}(\tilde{\mathbf{K}} + n\lambda\mathbf{I})(\tilde{\mathbf{K}} + n\lambda\mathbf{I})^{-1}\mathbf{f}_x \\ &\quad -n\lambda^2\mathbf{f}_x^T(\tilde{\mathbf{K}} + n\lambda\mathbf{I})^{-2}\mathbf{f}_x \\ &= \lambda\mathbf{f}_x^T(\tilde{\mathbf{K}} + n\lambda\mathbf{I})^{-1}\mathbf{f}_x - n\lambda^2\mathbf{f}_x^T(\tilde{\mathbf{K}} + n\lambda\mathbf{I})^{-2}\mathbf{f}_x.\end{aligned}$$

Now, summing up the first and the second part, we deduce

$$\begin{aligned}&\frac{1}{n}\|\mathbf{f}_x - \mathbf{Z}_q\beta\|_2^2 + s\lambda\|\beta\|_2^2 = \\ &\lambda\mathbf{f}_x^T(\tilde{\mathbf{K}} + n\lambda\mathbf{I})^{-1}\mathbf{f}_x = \\ &\lambda\mathbf{f}_x^T(\mathbf{K} + n\lambda\mathbf{I} + \tilde{\mathbf{K}} - \mathbf{K})^{-1}\mathbf{f}_x = \\ &\lambda\mathbf{f}_x^T(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}(\mathbf{I} + (\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}(\tilde{\mathbf{K}} - \mathbf{K}) \\ &\cdot(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}})^{-1}(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}\mathbf{f}_x.\end{aligned}$$

From Lemma 4, it follows that when

$$s \geq d_l(\frac{1}{\epsilon^2} + \frac{2}{3\epsilon})\log\frac{16d_{\mathbf{K}}^{\lambda}}{\delta}$$

then $(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}(\tilde{\mathbf{K}} - \mathbf{K})(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}} \succeq -\epsilon\mathbf{I}$.

We can now upper bound the error as (with $\epsilon = 1/2$):

$$\begin{aligned}&\lambda\mathbf{f}_x^T(\tilde{\mathbf{K}} + n\lambda\mathbf{I})^{-1}\mathbf{f}_x \leq \\ &\lambda\mathbf{f}_x^T(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}(1-\epsilon)^{-1}(\mathbf{K} + n\lambda\mathbf{I})^{-\frac{1}{2}}\mathbf{f}_x = \\ &(1-\epsilon)^{-1}\lambda\mathbf{f}_x^T(\mathbf{K} + n\lambda\mathbf{I})^{-1}\mathbf{f}_x \leq \\ &(1-\epsilon)^{-1}\lambda\mathbf{f}_x^T\mathbf{K}^{-1}\mathbf{f}_x \leq 2\lambda,\end{aligned}$$

where in the last inequality we have used Lemma 5. Moreover, we have that

$$\begin{aligned}&s\|\beta\|_2^2 = \\ &\mathbf{f}_x^T(\tilde{\mathbf{K}} + n\lambda\mathbf{I})^{-1}\mathbf{f}_x - n\lambda\mathbf{f}_x^T(\tilde{\mathbf{K}} + n\lambda\mathbf{I})^{-2}\mathbf{f}_x \leq \\ &\mathbf{f}_x^T(\tilde{\mathbf{K}} + n\lambda\mathbf{I})^{-1}\mathbf{f}_x \leq (1-\epsilon)^{-1}\mathbf{f}_x^T\mathbf{K}^{-1}\mathbf{f}_x \leq 2.\end{aligned}$$

Hence, the squared norm of our approximated function is bounded by $\|\tilde{f}\|_{\tilde{\mathcal{H}}}^2 \leq s\|\beta\|_2^2 \leq 2$. As such, problem (15) can now be written as $\min_\beta(1/n)\|\mathbf{f}_x - \tilde{\mathbf{f}}_\beta\|_2^2$ subject to $\|\tilde{f}\|_{\tilde{\mathcal{H}}}^2 \leq s\|\beta\|_2^2 \leq 2$, which is equivalent to

$$\sup_{\|f\|_{\mathcal{H}}\leq 1}\inf_{\sqrt{s}\|\beta\|_2\leq\sqrt{2}}\frac{1}{n}\|\mathbf{f}_x - \mathbf{Z}\beta\|_2^2,$$

and we have shown that this can be upper bounded by 2λ. □

Before we move to Theorem 1, following Rudi & Rosasco (2017), we prove Lemma 6 which is important in demonstrating the risk convergence rate.

Lemma 6. *Assuming that the conditions of Theorem 1 hold, let $\hat{f}^\lambda$ and f_β^λ be the empirical estimators from problems (6) and (7), respectively. In addition, suppose that $\{v_i\}_{i=1}^s$ are independent samples selected according to a probability measure τ_q with probability density function $q(v)$ such that $p(v)/q(v) > 0$ almost surely. Then, we have*

$$\langle Y - \hat{f}^\lambda, f_\beta^\lambda - \hat{f}^\lambda\rangle = 0.$$

Proof. The solution of problem (7) can be derived as

$$f_\beta^\lambda = \tilde{\mathbf{K}}(\tilde{\mathbf{K}} + n\lambda I)^{-1}Y = \frac{1}{s}\mathbf{Z}_q\mathbf{Z}_q^T(\frac{1}{s}\mathbf{Z}_q\mathbf{Z}_q^T + n\lambda I)^{-1}Y.$$

For all $f \in \mathcal{H}$, let $\mathbf{f} = [f(x_1), \cdots, f(x_n)]^T$. Define $\mathcal{H}_x := \{\mathbf{f} \mid f \in \mathcal{H}\}$. Then we can see that $\mathcal{H}_x$ is a subspace of $\mathbb{R}^n$. Since $Y \in \mathbb{R}^n$, we know there exists an orthogonal projection operator P such that for any vector $Z \in \mathbb{R}^n$, PZ is the projection of Z into $\mathcal{H}_x$. In particular, we have $\hat{f}^\lambda = PY$. In addition, let $\alpha \in \mathbb{R}^n$ and observe that $P\mathbf{K}\alpha = \mathbf{K}\alpha$, as $\mathbf{K}\alpha \in \mathcal{H}_x$. As such, we have that $(I-P)\mathbf{K}\alpha = 0$ for all $\alpha \in \mathbb{R}^n$, implying that $(I-P)\mathbf{K} = 0$. Hence, we have

$$\begin{aligned}&\langle Y - \hat{f}^\lambda, f_\beta^\lambda - \hat{f}^\lambda\rangle = \\ &\langle Y - PY, \tilde{\mathbf{K}}(\tilde{\mathbf{K}} + n\lambda I)^{-1}Y - PY\rangle = \\ &\langle (I-P)Y, \tilde{\mathbf{K}}(\tilde{\mathbf{K}} + n\lambda I)^{-1}Y\rangle - \langle (I-P)Y, PY\rangle = \\ &Y^T(I-P)\tilde{\mathbf{K}}(\tilde{\mathbf{K}} + n\lambda I)^{-1}Y - Y^T(I-P)PY = \\ &\frac{1}{s}Y^T(I-P)\mathbf{Z}_q\mathbf{Z}_q^T(\mathbf{Z}_q\mathbf{Z}_q^T + n\lambda I)^{-1}Y.\end{aligned} \tag{16}$$

The last equality follows from $(I-P)P = P - P^2 = 0$.

We know that the kernel function admits a decomposition as in Eq. (2). Hence, we can express $\mathbf{K}$ as

$$\mathbf{K} = \int_{\mathcal{V}} \mathbf{z}_v(\mathbf{x})\mathbf{z}_v(\mathbf{x})^T d\tau(v),$$

where $\mathbf{z}_v(\mathbf{x}) = [z(v, x_1), \cdots, z(v, x_n)]^T$.

Note that we have $(I-P)\mathbf{K} = 0$, which further implies that $(I-P)\mathbf{K}(I-P) = 0$. As a result, we have the following:

$$\begin{aligned}
0 &= \text{Tr}[(I-P)\mathbf{K}(I-P)] \\
&= \text{Tr}\left[(I-P)\int_{\mathcal{V}} \mathbf{z}_v(\mathbf{x})\mathbf{z}_v(\mathbf{x})^T d\tau(v)(I-P)\right] \\
&= \text{Tr}\left[\int_{\mathcal{V}} (I-P)\mathbf{z}_v(\mathbf{x})\mathbf{z}_v(\mathbf{x})^T(I-P)d\tau(v)\right] \\
&= \int_{\mathcal{V}} \text{Tr}[(I-P)\mathbf{z}_v(\mathbf{x})\mathbf{z}_v(\mathbf{x})^T(I-P)]d\tau(v) \\
&= \int_{\mathcal{V}} \|(I-P)\mathbf{z}_v(\mathbf{x})\|_2^2 d\tau(v) \\
&= \int_{\mathcal{V}} \|(I-P)\mathbf{z}_{q,v}(\mathbf{x})\|_2^2 \frac{p(v)}{q(v)} d\tau_q(v), \qquad (17)
\end{aligned}$$

where $\mathbf{z}_{q,v}(\mathbf{x}) = \sqrt{p(v)/q(v)}\mathbf{z}_v(\mathbf{x})$. Hence, we have $\|(I-P)\mathbf{z}_{q,v}(\mathbf{x})\|_2^2 = 0$ almost surely (a.s.) with respect to measure $d\tau_q$, which further shows that $(I-P)\mathbf{z}_{q,v}(\mathbf{x}) = \mathbf{0}$ a.s. For any $\alpha \in \mathbb{R}^s$, we have:

$$\alpha^T Y^T (I-P)\mathbf{Z}_q = \sum_{i=1} \alpha_i Y^T (I-P)\mathbf{z}_{q,v_i}(\mathbf{x}) = 0.$$

We now let $\alpha = Y^T(I-P)\mathbf{Z}_q$ and obtain

$$\|Y^T(I-P)\mathbf{Z}_q\|_2^2 = 0.$$

Returning back to Eq. (16), we have that

$$\begin{aligned}
&\langle Y - \hat{f}^\lambda, f_\beta^\lambda - \hat{f}^\lambda \rangle = \\
&\frac{1}{s} Y^T(I-P)\mathbf{Z}_q\mathbf{Z}_q^T(\mathbf{Z}_q\mathbf{Z}_q^T + n\lambda I)^{-1}Y.
\end{aligned}$$

Now, observe that

$$\begin{aligned}
&|Y^T(I-P)\mathbf{Z}_q\mathbf{Z}_q^T(\mathbf{Z}_q\mathbf{Z}_q^T + n\lambda I)^{-1}Y| \leq \\
&\|Y^T(I-P)\mathbf{Z}_q\|_2^2 \|\mathbf{Z}_q^T(\mathbf{Z}_q\mathbf{Z}_q^T + n\lambda I)^{-1}Y\|_2^2 = 0.
\end{aligned}$$

Hence, we conclude that $\langle Y - \hat{f}^\lambda, f_\beta^\lambda - \hat{f}^\lambda \rangle = 0$. □

C.2. Proof of Theorem 1

Proof. The proof relies on the decomposition of the expected risk of $\mathcal{E}(f_\beta^\lambda)$ as follows

$$\begin{aligned}
\mathcal{E}(f_\beta^\lambda) = \mathcal{E}(f_\beta^\lambda) - \hat{\mathcal{E}}(f_\beta^\lambda) \qquad (18) \\
+\hat{\mathcal{E}}(f_\beta^\lambda) - \hat{\mathcal{E}}(\hat{f}^\lambda) \qquad (19) \\
+\hat{\mathcal{E}}(\hat{f}^\lambda) - \mathcal{E}(\hat{f}^\lambda) \qquad (20) \\
+\mathcal{E}(\hat{f}^\lambda) - \mathcal{E}(f_{\mathcal{H}}) \qquad (21) \\
+\mathcal{E}(f_{\mathcal{H}}).
\end{aligned}$$

For (18), the bound is based on the Rademacher complexity of the reproducing kernel Hilbert space $\tilde{\mathcal{H}}$, where $\tilde{\mathcal{H}}$ corresponds to the approximated kernel $\tilde{k}$. We can upper bound the Rademacher complexity of this hypothesis space with Lemma 2. As $\mathbf{L}(y, f(x))$ is the squared error loss function with y and $f(x)$ bounded, we have that $\mathbf{L}$ is a Lipschitz continuous function with some constant $L > 0$. Hence,

$$\begin{aligned}
(18) &\leq R_n(\tilde{\mathbf{L}} \circ \tilde{\mathcal{H}}) + \sqrt{\frac{8\log(2/\delta)}{n}} \\
&\leq \sqrt{2}L\frac{1}{n}\mathbb{E}_X\sqrt{\text{Tr}(\tilde{\mathbf{K}})} + \sqrt{\frac{8\log(2/\delta)}{n}} \\
&\leq \sqrt{2}L\frac{1}{n}\sqrt{\mathbb{E}_X\text{Tr}(\tilde{\mathbf{K}})} + \sqrt{\frac{8\log(2/\delta)}{n}} \\
&\leq \sqrt{2}L\frac{1}{n}\sqrt{nz_0^2} + \sqrt{\frac{8\log(2/\delta)}{n}} \\
&\leq \frac{\sqrt{2}Lz_0}{\sqrt{n}} + \sqrt{\frac{8\log(2/\delta)}{n}} \in O(\frac{1}{\sqrt{n}}), \qquad (22)
\end{aligned}$$

where in the last inequality we applied Lemma 3 to $\tilde{\mathcal{H}}$, which is a reproducing kernel Hilbert space with radius $\sqrt{2}$. For (20), a similar reasoning can be applied to the unit ball in the reproducing kernel Hilbert space $\mathcal{H}$.

For (19), we observe that

$$\begin{aligned}
\hat{\mathcal{E}}(f_\beta^\lambda) - \hat{\mathcal{E}}(\hat{f}^\lambda) &= \frac{1}{n}\|Y - f_\beta^\lambda\|_2^2 - \frac{1}{n}\|Y - \hat{f}^\lambda\|_2^2 \\
&= \frac{1}{n}\inf_{\|f_\beta\|}\|Y - f_\beta\|_2^2 - \frac{1}{n}\|Y - \hat{f}^\lambda\|_2^2 \\
&= \frac{1}{n}\inf_{\|f_\beta\|}\Big(\|Y - \hat{f}^\lambda\|_2^2 + \|\hat{f}^\lambda - f_\beta\|_2^2 \\
&\quad +2\langle Y - \hat{f}^\lambda, \hat{f}^\lambda - f_\beta\rangle\Big) - \frac{1}{n}\|Y - \hat{f}^\lambda\|_2^2 \\
&\leq \frac{1}{n}\inf_{\|f_\beta\|}\|\hat{f}^\lambda - f_\beta\|_2^2 \\
&\quad +\frac{2}{n}\inf_{\|f_\beta\|}\langle Y - \hat{f}^\lambda, \hat{f}^\lambda - f_\beta\rangle \\
&\leq \frac{1}{n}\inf_{\|f_\beta\|}\|\hat{f}^\lambda - f_\beta\|_2^2 + \frac{2}{n}\langle Y - \hat{f}^\lambda, \hat{f}^\lambda - f_\beta^\lambda\rangle \\
&= \frac{1}{n}\inf_{\|f_\beta\|}\|\hat{f}^\lambda - f_\beta\|_2^2 \\
&\leq \sup_{\|f\|}\inf_{\|f_\beta\|}\frac{1}{n}\|f - f_\beta\|_2^2 \\
&\leq 2\lambda,
\end{aligned}$$

where in the last step we employ Theorem 5. Combining the three results, we derive

$$\mathcal{E}(f_\beta^\lambda) - \mathcal{E}(f_{\mathcal{H}}) \leq 2\lambda + O(\frac{1}{\sqrt{n}}) + \mathcal{E}(\hat{f}^\lambda) - \mathcal{E}(f_{\mathcal{H}}). \qquad (23)$$

□

C.3. Proofs of Corollaries 1 and 2

Proof. For Corollary 1, we set $\tilde{l}(v) = l_\lambda(v)$ and deduce

$$d_{\tilde{l}} = \int_{\mathcal{V}} l_\lambda(v) dv = d^\lambda_{\mathbf{K}}.$$

For Corollary 2, we set $\tilde{l}(v) = p(v)\frac{z_0^2}{\lambda}$ and derive

$$d_{\tilde{l}} = \int_{\mathcal{V}} p(v)\frac{z_0^2}{\lambda} dv = \frac{z_0^2}{\lambda}.$$

□

C.4. Proof of Theorem 3

Proof. The proof is similar to Theorem 1. In particular, we decompose the expected learning risk as

$$\begin{aligned}\mathcal{E}(g^\lambda_\beta) = \mathcal{E}(g^\lambda_\beta) - \hat{\mathcal{E}}(g^\lambda_\beta) &\quad (24)\\ +\hat{\mathcal{E}}(g^\lambda_\beta) - \hat{\mathcal{E}}(g_\mathcal{H}) &\quad (25)\\ +\hat{\mathcal{E}}(g_\mathcal{H}) - \mathcal{E}(g_\mathcal{H}) &\quad (26)\\ +\mathcal{E}(g_\mathcal{H}).\end{aligned}$$

Now, (24) and (26) can be upper bounded similar to Theorem 1, through the Rademacher complexity bound from Lemma 3. For (25), we have

$$\begin{aligned}&\hat{\mathcal{E}}(g^\lambda_\beta) - \mathcal{E}(g_\mathcal{H}) = \\ &\frac{1}{n}\sum_{i=1}^n \mathbf{L}(y_i, g^\lambda_\beta(x_i)) - \frac{1}{n}\sum_{i=1}^n \mathbf{L}(y_i, g_\mathcal{H}(x_i)) = \\ &\frac{1}{n}\inf_{\|g_\beta\|}\sum_{i=1}^n \mathbf{L}(y_i, g_\beta(x_i)) - \frac{1}{n}\sum_{i=1}^n \mathbf{L}(y_i, g_\mathcal{H}(x_i)) \\ &\leq \inf_{\|g_\beta\|}\frac{1}{n}\sum_{i=1}^n |g_\beta(x_i) - g_\mathcal{H}(x_i)| \\ &\leq \inf_{\|g_\beta\|}\sqrt{\frac{1}{n}\sum_{i=1}^n |g_\beta(x_i) - g_\mathcal{H}(x_i)|^2} \\ &\leq \sup_{\|g\|}\inf_{\|g_\beta\|}\sqrt{\frac{1}{n}\|g - g_\beta\|_2^2} \\ &\leq \sqrt{2\lambda}.\end{aligned}$$

□

C.5. Proofs of Corollaries 3 and 4

The proofs are similar to the proofs of Corollaries 1 and 2.

D. Proof of Theorem 2

In this proof, we rely on the notion of local Rademacher complexity and adjust our notation so that it is easier to cross-reference relevant auxiliary claims from Bartlett et al. (2005). Suppose P is a probability measure on $\mathcal{X} \times \mathcal{Y}$ and let $\{x_i, y_i\}_{i=1}^n$ be an independent sample from P. For any reproducing kernel Hilbert space $\mathcal{H}$ and a loss function l, we define the transformed function class as $l_\mathcal{H} := \{l(f(x), y) \mid f \in \mathcal{H}\}$. We also abbreviate the notation and denote with $l_f = l(f(x), y)$, $Pf = \int f(x)dP(x)$ and $P_n f = 1/n \sum_{i=1}^n f(x_i)$. For the reproducing kernel Hilbert space $\mathcal{H}$, we denote the solution of the kernel ridge regression problem by $\hat{f}$.

For our proof of Theorem 2, we need the following two results from Bartlett et al. (2005).

Theorem 6. *(Bartlett et al., 2005, Theorem 4.1) Let $\mathcal{H}$ be a class of functions with ranges in $[-1, 1]$ and assume that there is some constant B_0 such that for all $f \in \mathcal{H}$, $Pf^2 \leq B_0 Pf$. Let $\hat{\psi}_n$ be a sub-root function and let $\hat{r}^*$ be the fixed point of $\hat{\psi}_n$, i.e., $\hat{\psi}_n(\hat{r}^*) = \hat{r}^*$. Fix any $\delta > 0$, and assume that for any $r \geq \hat{r}^*$,*

$$\hat{\psi}_n(r) \geq e_1 \hat{R}_n\{f \in \text{star}(\mathcal{H}, 0) \mid P_n f^2 \leq r\} + \frac{e_2\delta}{n}$$

where e_1 and e_2 are constants, and

$$\text{star}(\mathcal{H}, f_0) = \{f_0 + \alpha(f - f_0) \mid f \in \mathcal{H} \wedge \alpha \in [0, 1]\}.$$

Then, for all $f \in \mathcal{H}$ and $D > 1$, with probability greater than $1 - 3e^{-\delta}$,

$$Pf \leq \frac{D}{D-1}P_n f + \frac{6D}{B}\hat{r}^* + \frac{e_3\delta}{n}$$

where e_3 is a constant.

Lemma 7. *(Bartlett et al., 2005, Lemma 6.6) Let k be a positive definite kernel function with reproducing kernel Hilbert space $\mathcal{H}$ and let $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_n$ be the eigenvalues of the normalized Gram-matrix $(1/n)\mathbf{K}$. Then, for all $r > 0$*

$$\hat{R}_n\{f \in \mathcal{H} \mid P_n f^2 \leq r\} \leq \Big(\frac{2}{n}\sum_{i=1}^n \min\{r, \hat{\lambda}_i\}\Big)^{1/2}$$

Theorem 6 is crucial for proving sharp convergence rates because it relies on the local Rademacher complexity technique. In order to apply the theorem, we need to find a proper sub-root function $\hat{\psi}_n$ and for that a special property characteristic to the squared error loss is required.

Lemma 8. *(Bartlett et al., 2005, Section 5.2) Let l be the squared error loss function and $\mathcal{H}$ a convex and uniformly bounded hypothesis space. Assume that for every probability distribution P in a class of data-generating distributions, there is an $f^* \in \mathcal{H}$ such that $Pl_{f^*} = \inf_{f\in\mathcal{H}} Pl_f$. Then, there exists a constant $B \geq 1$ such that for all $f \in \mathcal{H}$ and for every probability distribution P*

$$P(f - f^*)^2 \leq BP(l_f - l_{f^*}) \quad (27)$$

Let l be the squared error loss function and observe that for all $f \in \mathcal{H}$ it holds that

$$\begin{aligned} P_n l_f^2 &\geq (P_n l_f)^2 \quad (x^2 \text{ is convex}) \\ &\geq (P_n l_f)^2 - (P_n l_{\hat{f}})^2 \\ &= (P_n l_f + P_n l_{\hat{f}})(P_n l_f - P_n l_{\hat{f}}) \\ &\geq 2 P_n l_{\hat{f}} \, P_n(l_f - l_{\hat{f}}) \\ &\geq \frac{2}{B} P_n l_{\hat{f}} \, P_n(f - \hat{f})^2. \end{aligned} \tag{28}$$

The third inequality holds because $\hat{f}$ achieves the minimal empirical risk. The last inequality is a consequence of Lemma 8 applied to the empirical probability distribution P_n. Hence, to obtain a lower bound on $P_n l_f^2$ expressed solely in terms of $P_n(f - \hat{f})^2$, we need to find a lower bound of $P_n l_{\hat{f}}$. First, observe that it holds

$$P_n l_{\hat{f}} = \frac{1}{n} \|Y - \mathbf{K}(\mathbf{K} + n\lambda I)^{-1} Y\|^2.$$

Then, using this expression we derive

$$\begin{aligned} P_n l_{\hat{f}} &= \frac{1}{n} \|Y - \mathbf{K}(\mathbf{K} + n\lambda I)^{-1} Y\|^2 \\ &= n\lambda^2 Y^T (\mathbf{K} + n\lambda I)^{-2} Y \\ &\geq \frac{n\lambda^2}{(\lambda_1 + n\lambda)^2} Y^T Y \\ &= \left(\frac{n\lambda}{\lambda_1 + n\lambda} \right)^2 \frac{1}{n} \sum_{i=1}^n y_i^2 \\ &\geq \left(\frac{n\lambda}{\lambda_1 + n\lambda} \right)^2 \sigma_y^2 \quad (\text{with } \frac{1}{n} \sum_{i=1}^n y_i^2 \geq \sigma_y^2) \\ &= \sigma_y^2 \left(\frac{1}{1 + \frac{\lambda_1}{n\lambda}} \right)^2 \\ &\geq \sigma_y^2 \left(\frac{1}{\frac{\lambda_1}{n\lambda} + \frac{\lambda_1}{n\lambda}} \right)^2 \\ &= \frac{\sigma_y^2}{4} \left(\frac{n\lambda}{\lambda_1} \right)^2 = c(n\lambda)^2, \end{aligned} \tag{29}$$

where $c = (\sigma_y / 2\lambda_1)^2$ is a constant.

The last equality follows because λ_1 is independent of n and λ, as well as bounded. Hence, Eq.(28) becomes

$$P_n l_f^2 \geq \frac{2c(n\lambda)^2}{B} P_n(f - \hat{f})^2 =: c_1 (n\lambda)^2 P_n(f - \hat{f})^2.$$

As a result of this, we have the following inequality for the two function classes

$$\{l_f \in l_{\mathcal{H}} \mid P_n l_f^2 \leq r\} \subseteq \{l_f \in l_{\mathcal{H}} \mid P_n(f - \hat{f})^2 \leq \frac{r}{c_1 (n\lambda)^2}\}.$$

Recall that for a function class $\mathcal{H}$, we denote its empirical Rademacher complexity by $\hat{R}_n(\mathcal{H})$. Then, we have the following inequality

$$\begin{aligned} &\hat{R}_n\{l_f \in l_{\mathcal{H}} \mid P_n l_f^2 \leq r\} \leq \\ &\hat{R}_n\{l_f \in l_{\mathcal{H}} \mid P_n(f - \hat{f})^2 \leq \frac{r}{c_1 n^2 \lambda^2}\} = \\ &\hat{R}_n\{l_f - l_{\hat{f}} \mid P_n(f - \hat{f})^2 \leq \frac{r}{c_1 n^2 \lambda^2} \ \wedge \ l_f \in l_{\mathcal{H}}\} \leq \\ &L \hat{R}_n\{f - \hat{f} \mid P_n(f - \hat{f})^2 \leq \frac{r}{c_1 n^2 \lambda^2} \ \wedge \ f \in \mathcal{H}\} \leq \\ &L \hat{R}_n\{f - g \mid P_n(f - g)^2 \leq \frac{r}{c_1 n^2 \lambda^2} \ \wedge \ f, g \in \mathcal{H}\} \leq \\ &2L \hat{R}_n\{f \in \mathcal{H} \mid P_n f^2 \leq \frac{1}{4c_1} \frac{r}{n^2 \lambda^2}\} = \\ &2L \hat{R}_n\{f \in \mathcal{H} \mid P_n f^2 \leq \frac{c_2 r}{n^2 \lambda^2}\}, \end{aligned} \tag{30}$$

where the last inequality is due to Bartlett et al. (2005, Corollary 6.7). Now, combining Lemma 7 and Eq. (30) gives us a hint on how to find the sub-root function $\hat{\psi}_n$.

Theorem 7. *Assume $\{x_i, y_i\}_{i=1}^n$ is an independent sample from a probability measure P defined on $\mathcal{X} \times \mathcal{Y}$, with $\mathcal{Y} \in [-1, 1]$. Let k be a positive definite kernel with the reproducing kernel Hilbert space $\mathcal{H}$ and let $\hat{\lambda}_1 \geq \cdots, \geq \hat{\lambda}_n$ be the eigenvalues of the normalized kernel Gram-matrix. Denote the squared error loss function by $l(f(x), y) = (f(x) - y)^2$ and fix $\delta > 0$. If*

$$\hat{\psi}_n(r) = 2Le_1 \left(\frac{2}{n} \sum_{i=1}^n \min\{r, \hat{\lambda}_i\} \right)^{1/2} + \frac{e_3 \delta}{n},$$

then for all $l_f \in l_{\mathcal{H}}$ and $D > 1$, with probability $1 - 3e^{-\delta}$,

$$P l_f \leq \frac{D}{D-1} P_n l_f + \frac{6D}{B} \hat{r}^* + \frac{e_3 \delta}{n}.$$

Moreover, the fixed point $\hat{r}^*$ defined with $\hat{r}^* = \hat{\psi}_n(\hat{r})^*$ can be upper bounded by

$$\hat{r}^* \leq \min_{0 \leq h \leq n} \Big(\frac{h}{n} * \frac{e_4}{n^2 \lambda^2} + \sqrt{\frac{1}{n} \sum_{i > h} \hat{\lambda}_i} \Big),$$

where e_3 and e_4 are constants, and λ is the regularization parameter used in kernel ridge regression.

Proof. As $f(x), y \in [-1, 1]$, we have that $l_f \in [0, 1]$ and $P l_f^2 \leq P l_f$. Hence, we can apply Theorem 6 to function class $l_{\mathcal{H}}$ and obtain that for all $l_f \in l_{\mathcal{H}}$

$$P l_f \leq \frac{D}{D-1} P_n l_f + \frac{6D}{B} \hat{r}^* + \frac{e_3 \delta}{n},$$

as long as there is a sub-root function $\hat{\psi}_n(r)$ such that

$$\hat{\psi}_n(r) \geq e_1 \hat{R}_n\{f \in star(\mathcal{H}, 0) \mid P_n f^2 \leq r\} + \frac{e_2\delta}{n}. \tag{31}$$

We have previously demonstrated that

$$\begin{aligned} & e_1 \hat{R}_n\{f \in star(\mathcal{H}, 0) \mid P_n f^2 \leq r\} + \frac{e_2\delta}{n} \\ & \leq 2e_1 L \hat{R}_n \left\{ f \in \mathcal{H} \mid P_n f^2 \leq \frac{c_2 r}{n^2\lambda^2} \right\} + \frac{e_2\delta}{n} \\ & \leq 2e_1 L \left(\frac{2}{n} \sum_{i=1}^{n} \min\left\{ \frac{c_2 r}{n^2\lambda^2}, \hat{\lambda}_i \right\} \right)^{1/2} + \frac{e_2\delta}{n} \\ & \qquad \text{(by Lemma 7).} \end{aligned} \tag{32}$$

Hence, if choose $\hat{\psi}_n(r)$ to be equal to the right hand side of Eq.(32), then $\hat{\psi}_n(r)$ is a sub-root function that satisfies Eq.(31). Now, the upper bound on the fixed point $\hat{r}^*$ follows from Corollary 6.7 in Bartlett et al. (2005). □

We now deliver the proof of Theorem 2.

Proof. We decompose $\mathcal{E}(f_\beta^\lambda)$ with $D > 1$ as follows

$$\begin{aligned} \mathcal{E}(f_\beta^\lambda) = & \mathcal{E}(f_\beta^\lambda) - \frac{D}{D-1}\hat{\mathcal{E}}(f_\beta^\lambda) \\ & + \frac{D}{D-1}\hat{\mathcal{E}}(f_\beta^\lambda) - \frac{D}{D-1}\hat{\mathcal{E}}(\hat{f}^\lambda) \\ & + \frac{D}{D-1}\hat{\mathcal{E}}(\hat{f}^\lambda) - \mathcal{E}(\hat{f}^\lambda) \\ & + \mathcal{E}(\hat{f}^\lambda) - \mathcal{E}(f_{\mathcal{H}}) \\ & + \mathcal{E}(f_{\mathcal{H}}). \end{aligned}$$

Hence,

$$\mathcal{E}(f_\beta^\lambda) - \mathcal{E}(f_{\mathcal{H}}) \leq \left| \mathcal{E}(f_\beta^\lambda) - \frac{D}{D-1}\hat{\mathcal{E}}(f_\beta^\lambda) \right| \tag{33}$$

$$+ \frac{D}{D-1}(\hat{\mathcal{E}}(f_\beta^\lambda) - \hat{\mathcal{E}}(\hat{f}^\lambda)) \tag{34}$$

$$+ \left| \frac{D}{D-1}\hat{\mathcal{E}}(\hat{f}^\lambda) - \mathcal{E}(\hat{f}^\lambda) \right| \tag{35}$$

$$+ \mathcal{E}(\hat{f}^\lambda) - \mathcal{E}(f_{\mathcal{H}}). \tag{36}$$

We have already demonstrated that

$$\text{Eq. } (34) \leq 2\frac{D}{D-1}\lambda.$$

For Eqs. (33) and (35) we apply Theorem 7. However, note that f_β^λ and $\hat{f}^\lambda$ belong to different reproducing kernel Hilbert spaces. As a result, we have

$$\begin{aligned} \text{Eq. } (33) & \leq \hat{r}^*_{\tilde{\mathcal{H}}} + O(1/n) \\ \text{Eq. } (35) & \leq \hat{r}^*_{\mathcal{H}} + O(1/n) \end{aligned}$$

Now, combining these inequalities together we deduce

$$\begin{aligned} \mathcal{E}(f_\beta^\lambda) - \mathcal{E}(f_{\mathcal{H}}) & \leq \hat{r}^*_{\tilde{\mathcal{H}}} + \hat{r}^*_{\mathcal{H}} + 2\frac{D}{D-1}\lambda + O(1/n) \\ & \quad + \mathcal{E}(\hat{f}^\lambda) - \mathcal{E}(f_{\mathcal{H}}) \\ & \leq 2\hat{r}^*_{\mathcal{H}} + 2\frac{D}{D-1}\lambda + O(1/n) \\ & \quad + \mathcal{E}(\hat{f}^\lambda) - \mathcal{E}(f_{\mathcal{H}}). \end{aligned}$$

The last inequality holds because the eigenvalues of the Gram-matrix for the reproduing kernel Hilbert space $\tilde{\mathcal{H}}$ decay faster than the eigenvalues of $\mathcal{H}$. As a result of this, we have that $\hat{r}^*_{\tilde{\mathcal{H}}} \leq \hat{r}^*_{\mathcal{H}}$.

Now, Theorem 7 implies that

$$\hat{r}^*_{\mathcal{H}} \leq \min_{0 \leq h \leq n} \left(\frac{h}{n} * \frac{e_4}{n^2\lambda^2} + \sqrt{\frac{1}{n}\sum_{i>h} \hat{\lambda}_i} \right). \tag{37}$$

There are two cases worth discussing here. On the one hand, if the eigenvalues of $\mathbf{K}$ decay exponentially, we have

$$\hat{r}^*_{\mathcal{H}} \leq O\left(\frac{\log n}{n}\right)$$

by substituting $h = \lceil \log n \rceil$. Now, according to Caponnetto & De Vito (2007)

$$\mathcal{E}(\hat{f}^\lambda) - \mathcal{E}(f_{\mathcal{H}}) \in O\left(\frac{\log n}{n}\right),$$

and, thus, if we set $\lambda \propto {}^{\log n}/_{n}$ then the expected risk rate can be upper bounded by

$$\mathcal{E}(f_\beta^\lambda) - \mathcal{E}(f_{\mathcal{H}}) \in O\left(\frac{\log n}{n}\right).$$

On the other hand, if $\mathbf{K}$ has finitely many non-zero eigenvalues (t), we then have that

$$\hat{r}^*_{\mathcal{H}} \in O\left(\frac{1}{n}\right),$$

by substituting $h \geq t$. Moreover, in this case, $\mathcal{E}(\hat{f}^\lambda) - \mathcal{E}(f_{\mathcal{H}}) \in O({}^1/_n)$ and setting $\lambda \propto {}^1/_n$, we deduce that

$$\mathcal{E}(f_\beta^\lambda) - \mathcal{E}(f_{\mathcal{H}}) \leq O\left(\frac{1}{n}\right).$$

□

E. Proof of Theorem 4

Proof. Suppose the examples $\{x_i, y_i\}_{i=1}^n$ are independent and identically distributed and that the kernel k can be decomposed as in Eq. (2). Let $\{v_i\}_{i=1}^s$ be an independent

sample selected according to $p(v)$. Then, using these s features we can approximate the kernel as

$$\tilde{k}(x,y) = \frac{1}{s}\sum_{i=1}^{s} z(v_i,x)z(v_i,y) = \int_V z(v,x)z(v,y)d\hat{P}(v), \quad (38)$$

where $\hat{P}$ is the empirical measure on $\{v_i\}_{i=1}^s$. Denote the reproducing kernel Hilbert space associated with kernel $\tilde{k}$ by $\tilde{\mathcal{H}}$ and suppose that kernel ridge regression was performed with the approximate kernel $\tilde{k}$. From Theorem 1 and Corollary 2, it follows that if

$$s \geq \frac{7z_0^2}{\lambda}\log\frac{16d_{\mathbf{K}}^{\lambda}}{\delta},$$

then for all $\delta \in (0,1)$, with probability $1-\delta$, the risk convergence rate of the kernel ridge regression estimator based on random Fourier features can be upper bounded by

$$\mathcal{E}(f_{\alpha}^{\lambda}) \leq 2\lambda + O\left(\frac{1}{\sqrt{n}}\right) + \mathcal{E}(f_{\mathcal{H}}). \quad (39)$$

Let $f_{\tilde{\mathcal{H}}}$ be the function in the reproducing kernel Hilbert space $\tilde{\mathcal{H}}$ achieving the minimal risk, i.e., $\mathcal{E}(f_{\tilde{\mathcal{H}}}) = \inf_{f\in\tilde{\mathcal{H}}}\mathcal{E}(f)$. We now treat $\tilde{k}$ as the actual kernel that can be decomposed via the expectation with respect to the empirical measure in Eq. (38) and re-sample features from the set $\{v_i\}_{i=1}^s$, but this time the sampling is performed using the optimal ridge leverage scores. As $\tilde{k}$ is the actual kernel, it follows from Eq. (5) that the leverage function in this case can be defined by

$$l_{\lambda}(v) = p(v)\mathbf{z}_v(\mathbf{x})^T(\tilde{\mathbf{K}} + n\lambda I)^{-1}\mathbf{z}_v(\mathbf{x}).$$

Now, observe that

$$l_{\lambda}(v_i) = p(v_i)[\mathbf{Z}_s^T(\tilde{\mathbf{K}} + n\lambda I)^{-1}\mathbf{Z}_s]_{ii}$$

where $[A]_{ii}$ denotes the ith diagonal element of matrix A. As $\tilde{\mathbf{K}} = (1/s)\mathbf{Z}_s\mathbf{Z}_s^T$, then the Woodbury inversion lemma implies that

$$l_{\lambda}(v_i) = p(v_i)[\mathbf{Z}_s^T\mathbf{Z}_s(\frac{1}{s}\mathbf{Z}_s^T\mathbf{Z}_s + n\lambda I)^{-1}]_{ii}.$$

If we let $l_{\lambda}(v_i) = p_i$, then the optimal distribution for $\{v_i\}_{i=1}^s$ is multinomial with individual probabilities $q(v_i) = p_i/(\sum_{j=1}^s p_j)$. Hence, we can re-sample l features according to $q(v)$ and perform linear ridge regression using the sampled leverage weighted features. Denoting this estimator with $\tilde{f}_l^{\lambda^*}$ and the corresponding number of degrees of freedom with $d_{\tilde{\mathbf{K}}}^{\lambda} = \mathrm{Tr}\tilde{\mathbf{K}}(\tilde{\mathbf{K}} + n\lambda)^{-1}$, we deduce (using Theorem 1 and Corollary 1)

$$\mathcal{E}(\tilde{f}_l^{\lambda^*}) \leq 2\lambda^* + O\left(\frac{1}{\sqrt{n}}\right) + \mathcal{E}(f_{\tilde{\mathcal{H}}}), \quad (40)$$

with the number of features $l \propto d_{\tilde{\mathbf{K}}}^{\lambda}$.

As $f_{\tilde{\mathcal{H}}}$ is the function achieving the minimal risk over $\tilde{\mathcal{H}}$, we can conclude that $\mathcal{E}(f_{\tilde{\mathcal{H}}}) \leq \mathcal{E}(f_{\alpha}^{\lambda})$. Now, combining Eq. (39) and (40), we obtain the final bound on $\mathcal{E}(\tilde{f}_l^{\lambda^*})$. □

F. Code of Algorithm 1

```
def feat_gen(x,n_feat,lns):
    """
    #function to generate the features for gaussian kernel
    :param x: the data
    :param n_feat: number of features we need
    :param lns: the inverse landscale of the gaussian kernel
    :return: a sequence of features ready for KRR
    """
    n,d = np.shape(x)

    w = np.random.multivariate_normal(np.zeros(d),lns*np.eye(d),n_feat)
    return w


def feat_matrix(x,w):
    """
    #funtion to generate the feature matrix Z
    :param x: the data
    :param w: the features
    :return: the feature matrix of size len(n)*len(s)
    """
    s, dim = w.shape
    # perform the product of x and w transpose
    prot_mat = np.matmul(x,w.T)

    feat1 = np.cos(prot_mat)
    feat2 = np.sin(prot_mat)

    feat_final = np.sqrt(1.0/s)*np.concatenate((feat1,feat2),axis = 1)

    return feat_final

def opm_feat(x,w,lmba):
    """
    #function to select the optimum features
    :param x: the independent variable
    :param w: the first layer features generated according to spectral density
    :return: optimum features with importance weight
    """
    n_num, dim = x.shape
    s, dim = w.shape

    prot_mat = np.matmul(x,w.T)

    feat1 = np.sqrt(1.0/s)*np.cos(prot_mat)
    feat2 = np.sqrt(1.0/s)*np.sin(prot_mat)

    Z_s = feat1 +feat2

    ZTZ = np.matmul(Z_s.T,Z_s)

    ZTZ_inv = np.linalg.inv(ZTZ +n_num*lmba*np.eye(s))
```

```
M = np.matmul(ZTZ,ZTZ_inv)
#M = np.matmul(M0,ZTZ)

l = np.trace(M)
#print l
#n_feat_draw = min(s,max(50,l))
#print n_feat_draw
#n_feat_draw = int(round(n_feat_draw))
n_feat_draw = s
#print n_feat_draw

pi_s = np.diag(M)
#print pi_s
qi_s = pi_s/l
is_wgt = np.sqrt(1/qi_s)
#print is_wgt

wgt_order = np.argsort(is_wgt)

w_order = wgt_order[(s-n_feat_draw):]
#print len(w_order)

#w_order = np.random.choice(s, n_feat_draw, replace=False, p=qi_s)
#print w_order
w_opm = np.zeros((n_feat_draw,dim))

wgh_opm = np.zeros(n_feat_draw)

for ii in np.arange(n_feat_draw):
    order = w_order[ii]
    w_opm[ii,:] = w[order,:]
    wgh_opm[ii] = is_wgt[order]


return w_opm,wgh_opm
```