Supplementary Material
Nyström Method vs Random Fourier Features: A Theoretical and Empirical Comparison

Anonymous Author(s)
Affiliation
Address
email

Appendix A: Proof of Proposition 1

For any $f \in \mathcal{H}_a^f$, we can write

$$f(\cdot) = \sum_{k=1}^m w_k^s s(\cdot) + w_k^c c(\cdot),$$

where $\mathbf{w} = (w_1^s, w_1^c, \cdots, w_m^s, w_m^c)^\top$ are the coefficients. Then we have

$$f(\mathbf{x}) = \mathbf{w}^\top \mathbf{z}_f(\mathbf{x}), \quad \|f\|_{\mathcal{H}_\kappa}^2 \propto \int d\mathbf{u} \|\widehat{f}(\mathbf{u})\|^2 \exp(\sigma^2 \|\mathbf{u}\|^2/2) = \mathbf{w}^\top (\mathbf{w} \circ \gamma),$$

where the functional norm follows [2], $\circ$ denotes element-wise dot product, and

$$\gamma = (\gamma_1^s, \gamma_1^c, \cdots, \gamma_m^s, \gamma_m^c)^\top, \quad \gamma_i^s = \gamma_i^c = \exp(\sigma^2 \|\mathbf{u}_i\|^2/2),$$

weights the Fourier component $\mathbf{u}_i$. We then finish the proof by replacing the linear form of $f(\mathbf{x})$ and the functional norm of f into equation (5) to obtain equation (6) in the paper.

Appendix B: Proof of Proposition 2

By noting that for any $f(\cdot) = \sum_{j=1}^r w_j \widehat{\varphi}_j(\cdot) \in \mathcal{H}_a^n$, we have

$$f(\mathbf{x}_i) = \sum_{j=1}^r w_j \widehat{\varphi}_j(\mathbf{x}_i) = \mathbf{w}^\top \mathbf{z}_n(\mathbf{x}_i), \quad \|f\|_{\mathcal{H}_\kappa}^2 = \|\mathbf{w}\|_2^2,$$

where $\widehat{\varphi}_j(\mathbf{x}_i) = \mathbf{z}_n(\mathbf{x}_i)$ follows the equality in (2) in the paper, and the functional norm follows the fact that $\widehat{\varphi}_j, j = 1, \cdots, r$ are normalized orthonomal eigenfunctions. By replacing these into equation (10) in the paper, we obtain an equivalent formulation in equation (9) in the paper.

Appendix C: Proof of Theorem 2

To prove Theorem 2. We first prove the following lemma.

Lemma 1. *Define $\mathcal{L}(f)$ as*

$$\mathcal{L}(f) = \frac{\lambda}{2} \|f\|_{\mathcal{H}_\kappa}^2 + \frac{1}{N} \sum_{i=1}^N \ell(f(\mathbf{x}_i), y_i).$$

We have

$$0 \leq \mathcal{L}(f_m^*) - \mathcal{L}(f_N^*) \leq \frac{C}{2N\lambda} \|K - \widehat{K}_r\|_2,$$

where C is the bound on the gradient of the loss function, i.e., $|\nabla_z \ell(z,y)| \leq C$, $\|\cdot\|_2$ stands for the spectral norm of a matrix.

Proof. Let $\ell^*(\alpha)$ denote the Fenchel conjugate of $\ell(z,y)$ in terms of z, i.e.,

$$\ell^*(\alpha) = \sup_{\alpha\in\Omega} (\alpha z - \ell(z,y))$$

where Ω is the range of the mapping $\nabla_z \ell(z,y) : \mathbb{R} \to \mathbb{R}$. Using the conjugate of $\ell(z,y)$, we can write

$$\mathcal{L}(f_N^*) = \min_{f\in\mathcal{H}_\mathcal{D}} \frac{\lambda}{2}\|f\|_{\mathcal{H}_\kappa}^2 + \frac{1}{N}\sum_{i=1}^N \ell(f(\mathbf{x}_i), y_i),$$

into an equivalent form

$$\mathcal{L}(f_N^*) \quad = \quad \max_{\{\alpha_i\in\Omega\}_{i=1}^N} -\frac{1}{N}\sum_{i=1}^N \ell^*(\alpha_i) - \frac{1}{2\lambda N^2}(\boldsymbol{\alpha}\circ\mathbf{y})^\top K(\boldsymbol{\alpha}\circ\mathbf{y})$$

where $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_N)^\top$. Similarly, we can cast

$$\mathcal{L}(f_m^*) = \min_{f\in\mathcal{H}_n^a} \frac{\lambda}{2}\|f\|_{\mathcal{H}_\kappa}^2 + \frac{1}{N}\sum_{i=1}^N \ell(f(\mathbf{x}_i), y_i),$$

into

$$\mathcal{L}(f_m^*) \quad = \quad \max_{\{\alpha_i\in\Omega\}_{i=1}^N} -\frac{1}{N}\sum_{i=1}^N \ell^*(\alpha_i) - \frac{1}{2\lambda N^2}(\alpha\circ\mathbf{y})^\top K_b \widehat{K}^\dagger K_b^\top (\alpha\circ\mathbf{y}).$$

Then

$$\begin{aligned}
\mathcal{L}(f_m^*) = &\max_{\{\alpha_i\in\Omega\}_{i=1}^N} -\frac{1}{N}\sum_{i=1}^N \ell^*(\alpha_i) - \frac{1}{2\lambda N^2}(\boldsymbol{\alpha}\circ\mathbf{y})^\top K(\boldsymbol{\alpha}\circ\mathbf{y}) \\
&+ \frac{1}{2\lambda N^2}(\boldsymbol{\alpha}\circ\mathbf{y})^\top (K - K_b\widehat{K}^\dagger K_b^\top)(\boldsymbol{\alpha}\circ\mathbf{y}) \\
\leq &\max_{\{\alpha_i\in\Omega\}_{i=1}^N} -\frac{1}{N}\sum_{i=1}^N \ell^*(\alpha_i) - \frac{1}{2\lambda N^2}(\boldsymbol{\alpha}\circ\mathbf{y})^\top K(\boldsymbol{\alpha}\circ\mathbf{y}) \\
&+ \max_{\{\alpha_i\in\Omega\}_{i=1}^N} \frac{1}{2\lambda N^2}(\boldsymbol{\alpha}\circ\mathbf{y})^\top (K - K_b\widehat{K}^\dagger K_b^\top)(\boldsymbol{\alpha}\circ\mathbf{y}) \\
\leq &\,\mathcal{L}(f_N^*) + \frac{1}{2\lambda N^2}\|\boldsymbol{\alpha}\|_2^2 \|K - K_b\widehat{K}^\dagger K_b^\top\|_2 \leq \mathcal{L}_N(f_N^*) + \frac{C}{2\lambda N}\|K - \widehat{K}_r\|_2.
\end{aligned}$$

where the last inequality follows $|\alpha| \leq C, \forall \alpha\in\Omega$ due to $|\nabla_z \ell(z,y)| \leq C$, and the definition of $\widehat{K}_r$ in equation (7) in the paper. □

Proof of Theorem 2

Define loss function $\bar{\ell}(f(\mathbf{x}), y) = \frac{\lambda}{2}\|f\|_{\mathcal{H}_\kappa}^2 + \ell(f(\mathbf{x}), y)$. To simplify our notation, we define P_N and P as

$$P_N(\bar{\ell}\circ f) = \frac{1}{N}\sum_{i=1}^N \bar{\ell}(f(\mathbf{x}_i), y_i) = \mathcal{L}(f),\ P(\bar{\ell}\circ f) = \mathrm{E}\left[\bar{\ell}(f(\mathbf{x}), y)\right] = F(f).$$

Using this notation, we have

$$\begin{aligned}
&\Lambda(f_m^*) - \Lambda(f_N^*) = P(\bar{\ell}\circ f_m^* - \bar{\ell}\circ f_N^*) = P_N(\bar{\ell}\circ f_m^* - \bar{\ell}\circ f_N^*) + (P - P_N)(\bar{\ell}\circ f_m^* - \bar{\ell}\circ f_N^*) \\
&\leq P_N(\bar{\ell}\circ f_m^* - \bar{\ell}\circ f_N^*) + \max_{(f,f')\in\mathcal{G}} (P - P_N)(\bar{\ell}\circ f - \bar{\ell}\circ f'),
\end{aligned}$$

where $\mathcal{G}$ is defined as

$$\mathcal{G} = \Big\{(f, f') \in \mathcal{H}_\kappa \times \mathcal{H}_\kappa : \|f - f'\|_{\ell_2} \leq r, \|f - f'\|_{\mathcal{H}_\kappa} \leq R\Big\},$$

where $\|f\|_{\ell_2} = \sqrt{\mathrm{E}[f^2(\mathbf{x})]} \leq \|f\|_{\mathcal{H}_\kappa}$, and r, R are given by $r = R = \|f_m^* - f_N^*\|_{\mathcal{H}_\kappa} \leq 4/\sqrt{\lambda}$. Using Lemma 9 from [1], we have with a probability $1 - 2N^{-3}$, for any $\epsilon r \leq e^N$, $\epsilon^2 R \leq e^N$,

$$\sup_{(f,f')\in\mathcal{G}} (P - P_N)(\bar{\ell} \circ f' - \bar{\ell} \circ f') \leq C_0 L(r\epsilon + R\epsilon^2 + e^{-N}) \leq C_1 L(r\epsilon + e^{-N}),$$

where C_0, C_1 are numerical constants, and L is the upper bound of the gradient of $\bar{\ell}$ for functions in $\mathcal{G}$ and is given by $L \leq \sqrt{\lambda} + C$. Since $\max\left(\|f_N^*\|_{\mathcal{H}_\kappa}, \|f_m^*\|_{\mathcal{H}_\kappa}\right) \leq 2/\sqrt{\lambda}$ and $16\epsilon^2 e^{-2N} \leq \lambda$, we have the condition $\epsilon r \leq e^N$ satisfied, and therefore, with a probability $1 - 2N^{-3}$, we have

$$\begin{aligned} P(\bar{\ell} \circ f_m^* - \bar{\ell} \circ f_N^*) &\leq P_N(\bar{\ell} \circ f_m^* - \bar{\ell} \circ f_N^*) + C_1(\sqrt{\lambda} + C)(r\epsilon + e^{-N}) \\ &\leq \frac{C\|K - \widehat{K}_r\|_2}{2N\lambda} + C_1(\sqrt{\lambda} + C)(r\epsilon + e^{-N}), \end{aligned}$$

where we use the result in Lemma 1. Using the definition of r, and for any scaler s, we have

$$\begin{aligned} 2rs &\leq \frac{\lambda\|f_m^* - f_N^*\|_{\mathcal{H}_\kappa}^2}{8} + \frac{8s^2}{\lambda} \leq \frac{\lambda}{4}\|f_m^* - f^*\|_{\mathcal{H}_\kappa}^2 + \frac{\lambda}{4}\|f^* - f_N^*\|_{\mathcal{H}_\kappa}^2 + \frac{8s^2}{\lambda} \\ &\leq \frac{1}{2}P(\bar{\ell} \circ f_m^* - \bar{\ell} \circ f^*) + \frac{1}{2}P(\bar{\ell} \circ f_N^* - \bar{\ell} \circ f^*) + \frac{8s^2}{\lambda}, \end{aligned}$$

where the last step explores the strong convexity of $\bar{\ell}$. Using the bound for $2r[C_1(\sqrt{\lambda} + C)\epsilon/2]$, we have

$$\begin{aligned} &P(\bar{\ell} \circ f_m^* - \bar{\ell} \circ f_N^*) - C_1\left(\sqrt{\lambda} + C\right)e^{-N} \\ \leq \quad &\frac{C\|K - \widehat{K}_r\|_2}{2N\lambda} + \frac{1}{2}P(\bar{\ell} \circ f_m^* - \bar{\ell} \circ f^*) + \frac{1}{2}P(\bar{\ell} \circ f_N^* - \bar{\ell} \circ f^*) + \frac{2C_1^2(\sqrt{\lambda} + C)^2}{\lambda}\epsilon^2. \end{aligned}$$

Using the fact $\Lambda(f) = P(\bar{\ell} \circ f - \bar{\ell} \circ f^*)$, we have

$$\frac{1}{2}\Lambda(f_m^*) \leq \frac{3}{2}\Lambda(f_N^*) + C_1\left(\sqrt{\lambda} + C\right)e^{-N} + \frac{C\|K - \widehat{K}_r\|_2}{2N\lambda} + \frac{2C_1^2(\sqrt{\lambda} + C)^2}{\lambda}\epsilon^2.$$

We complete the proof by absorbing the constant terms into a numerical constant C_2.

Appendix D: Proof of Theorem 3

To prove Theorem 3, we fist prove the following lemma that relates H_r and $\widehat{H}_r$ to matrices K_r and $\widehat{K}_r$, respectively.

Lemma 2. *Assume $\widehat{\lambda}_r > 0$ and $\lambda_r > 0$. We have*

$$\left[\widehat{K}_r\right]_{i,j} = \langle \kappa(\mathbf{x}_i, \cdot), \widehat{H}_r \kappa(\mathbf{x}_j, \cdot)\rangle_{\mathcal{H}_\kappa}, \quad [K_r]_{i,j} = \langle \kappa(\mathbf{x}_i, \cdot), H_r \kappa(\mathbf{x}_j, \cdot)\rangle_{\mathcal{H}_\kappa},\ i, j = 1, \ldots, N.$$

Proof. By the definition of $\widehat{H}_r$ and expression of $\widehat{\varphi}_j$, we have

$$\begin{aligned} &\langle \kappa(\mathbf{x}_i, \cdot), \widehat{H}_r \kappa(\mathbf{x}_j, \cdot)\rangle = \sum_{k=1}^r \frac{1}{\widehat{\lambda}_k} \langle \kappa(\mathbf{x}_i, \cdot), \widehat{\varphi}_k\rangle \langle \kappa(\mathbf{x}_j, \cdot), \widehat{\varphi}_k\rangle \\ &= \sum_{s,t=1}^m \sum_{k=1}^r \frac{1}{\widehat{\lambda}_k} \widehat{V}_{s,k} \widehat{V}_{t,k} \langle \kappa(\mathbf{x}_i, \cdot), \kappa(\widehat{\mathbf{x}}_s, \cdot)\rangle \langle \kappa(\mathbf{x}_j, \cdot), \kappa(\widehat{\mathbf{x}}_t, \cdot)\rangle \\ &= \sum_{s,t=1}^m \sum_{k=1}^r \frac{\widehat{V}_{s,k}\widehat{V}_{t,k}}{\widehat{\lambda}_k} [K_b]_{i,s}[K_b]_{j,t} = \sum_{s,t=1}^m [K_b]_{i,s}\widehat{K}^\dagger_{s,t}[K_b]_{j,t} = [K_b\widehat{K}^\dagger K_b^\top]_{i,j} = [\widehat{K}_r]_{i,j} \end{aligned}$$

where in the second equality we use equation (2) in the paper for $\widehat{\varphi}_j$, and last line uses $\widehat{K} = \sum_{k=1}^r \frac{1}{\widehat{\lambda}_k}\widehat{\mathbf{v}}_k\widehat{\mathbf{v}}_k^\top$. Similarly, we have

$$\begin{aligned}
\langle \kappa(\mathbf{x}_i,\cdot), H_r\kappa(\mathbf{x}_j,\cdot)\rangle &= \sum_{k=1}^r \frac{1}{\lambda_k}\langle \kappa(\mathbf{x}_i,\cdot),\varphi_k\rangle\langle\kappa(\mathbf{x}_j,\cdot),\varphi_k\rangle \\
&= \sum_{s,t=1}^N\sum_{k=1}^r \frac{1}{\lambda_k}V_{s,k}V_{t,k}\langle\kappa(\mathbf{x}_i,\cdot),\kappa(\mathbf{x}_s,\cdot)\rangle\langle\kappa(\mathbf{x}_j,\cdot),\kappa(\mathbf{x}_t,\cdot)\rangle \\
&= \sum_{s,t=1}^N\sum_{k=1}^r \frac{V_{s,k}V_{t,k}}{\lambda_k}[K]_{i,s}[K]_{j,t} = \sum_{s,t=1}^N [K]_{i,s}[K_r^{-1}]_{s,t}[K]_{j,t} = [KK_r^{-1}K^\top]_{i,j} = [K_r]_{i,j}
\end{aligned}$$

$\square$

Proof of Theorem 3

Define $\bar{\kappa}(\mathbf{x}_a,\mathbf{x}_i) = \langle\kappa(\mathbf{x}_a,\cdot),\Delta H\kappa(\mathbf{x}_i,\cdot)\rangle_{\mathcal{H}_\kappa}$. By Lemma 2, we have

$$[K_r - \widehat{K}_r]_{ia} = \bar{\kappa}(\mathbf{x}_a,\mathbf{x}_i).$$

We have

$$\begin{aligned}
\|K_r - \widehat{K}_r\|_2^2 &= \max_{\|\mathbf{u}\|_2\le 1}\sum_{i,j=1}^N\sum_{a=1}^N u_iu_j\bar{\kappa}(\mathbf{x}_a,\mathbf{x}_i)\bar{\kappa}(\mathbf{x}_a,\mathbf{x}_j) \\
&= \max_{\|\mathbf{u}\|_2\le 1}\sum_{i,j=1}^N\sum_{a=1}^N u_iu_j\langle\kappa(\mathbf{x}_a,\cdot),\Delta H\kappa(\mathbf{x}_i,\cdot)\rangle_{\mathcal{H}_\kappa}\langle\kappa(\mathbf{x}_a,\cdot),\Delta H\kappa(\mathbf{x}_j,\cdot)\rangle_{\mathcal{H}_\kappa} \qquad (1)
\end{aligned}$$

Using the definition of L_N in equaiton (1) in the paper and the reproducing property, for any $f, g\in\mathcal{H}_\kappa$, we can have

$$\sum_{i=1}^N\langle\kappa(\mathbf{x}_i,\cdot),f\rangle\langle\kappa(\mathbf{x}_i,\cdot),g\rangle = \sum_{i=1}^N\langle\kappa(\mathbf{x}_i,\cdot),g\rangle f(\mathbf{x}_i) = N\langle L_Nf,g\rangle$$

Applying the above equality to $f = \Delta H\kappa(\mathbf{x}_i,\cdot), g = \Delta H\kappa(\mathbf{x}_j,\cdot)$, we have

$$\begin{aligned}
\sum_{a=1}^N\langle\kappa(\mathbf{x}_a,\cdot),\Delta H\kappa(\mathbf{x}_i,\cdot)\rangle\langle\kappa(\mathbf{x}_a,\cdot),\Delta H\kappa(\mathbf{x}_j,\cdot)\rangle &= N\langle L_N\Delta H\kappa(\mathbf{x}_i,\cdot),\Delta H\kappa(\mathbf{x}_j,\cdot)\rangle \\
&= N\langle\kappa(\mathbf{x}_i,\cdot),\Delta HL_N\Delta H\kappa(\mathbf{x}_j,\cdot)\rangle \qquad (2)
\end{aligned}$$

Using equality (2) in (1), we have

$$\|K_r - \widehat{K}_r\|_2^2 = \max_{\|\mathbf{u}\|_2\le 1} N\sum_{i,j=1}^N u_iu_j\langle\kappa(\mathbf{x}_i,\cdot),\Delta HL_N\Delta H\kappa(\mathbf{x}_j,\cdot)\rangle_{\mathcal{H}_\kappa}.$$

Define $f(\cdot) = \sum_{i=1}^N u_i\kappa(\mathbf{x}_i,\cdot)$. We note that $\kappa(\mathbf{x}_i,\cdot)$ can be expressed as an eigen-expansion form, i.e.,

$$\kappa(\mathbf{x}_i,\cdot) = \sum_{j=1}^N\sqrt{\lambda_j}V_{i,j}\varphi_j(\cdot), \qquad (3)$$

Using the relationship between $\kappa(\mathbf{x}_i,\cdot)$ and eigenfunctions $\varphi_j(\cdot)$ in equation (3), we have

$$f(\cdot) = \sum_{i=1}^N\sum_{j=1}^N u_i\sqrt{\lambda_i}V_{i,j}\varphi_j(\cdot) = \sqrt{N}L_N^{1/2}[g](\cdot),$$

where $g(\cdot) = \sum_{i,j=1}^N u_iV_{i,j}\varphi_j(\cdot)$. Since $\|\mathbf{u}\|_2\le 1$, we have $\|g\|_{\mathcal{H}_\kappa}^2 = \mathbf{u}^\top VV^\top\mathbf{u}\le 1$, where $V = (\mathbf{v}_1,\ldots,\mathbf{v}_N)$. We thus have

$$\begin{aligned}
\|K_r - \widehat{K}_r\|_2^2 &= N^2\max_{\|g\|_{\mathcal{H}_\kappa}\le 1}\langle L_N^{1/2}g,\Delta HL_N\Delta HL_N^{1/2}g\rangle_{\mathcal{H}_\kappa} = N^2\max_{\|g\|_{\mathcal{H}_\kappa}\le 1}\langle g, L_N^{1/2}\Delta HL_N\Delta HL_N^{1/2}g\rangle_{\mathcal{H}_\kappa} \\
&= N^2\|L_N^{1/2}\Delta HL_N\Delta HL_N^{1/2}\|_2 = N^2\|L_N^{1/2}\Delta HL_N^{1/2}\|_2^2.
\end{aligned}$$

Appendix E: Proof of Theorem 4

To prove Theorem 4, we first prove the following lemma.

Lemma 3. *Assume* $\Delta = (\lambda_r - \lambda_{r+1})/N > 3\|L_N - L_m\|_{HS}$, *where* $\|\cdot\|_{HS}$ *stands for the Hilbert Schmidt norm of an integral operator. Let* $\Theta = (\widehat{\varphi}_1, \ldots, \widehat{\varphi}_r), \Phi = (\varphi_1, \ldots, \varphi_r), \overline{\Phi} = (\varphi_{r+1}, \ldots, \varphi_N)$ *. Then, there exists a matrix* $P \in \mathbb{R}^{(N-r)\times r}$ *satisfying*

$$\|P\|_F \leq \frac{2\|L_N - L_m\|_{HS}}{\Delta - \|L_N - L_m\|_{HS}}.$$

such that $\Theta = (\Phi + \overline{\Phi}P)(I + P^\top P)^{-1/2}$.

Proof of Lemma 3. we need the following perturbation result [3] [1].

Lemma 4. *(Theorem 2.7 of Chapter 6 [3]) Let* $(\lambda_i, \mathbf{v}_i), i \in [n]$ *be the eigenvalues and eigenvectors of a symmetric matrix* $A \in \mathbb{R}^{n\times n}$ *ranked in the descending order of eigenvalues. Set* $X = (\mathbf{v}_1, \ldots, \mathbf{v}_r)$ *and* $Y = (\mathbf{v}_{r+1}, \ldots, \mathbf{v}_n)$. *Given a symmetric perturbation matrix* E, *let*

$$\widehat{E} = (X, Y)^\top E(X, Y) = \begin{pmatrix} \widehat{E}_{11} & \widehat{E}_{12} \\ \widehat{E}_{21} & \widehat{E}_{22} \end{pmatrix}.$$

Let $\|\cdot\|$ *represent a consistent family of norms and set*

$$\gamma = \|\widehat{E}_{21}\|, \delta = \lambda_r - \lambda_{r+1} - \|\widehat{E}_{11}\| - \|\widehat{E}_{22}\|.$$

If $\delta > 0$ *and* $2\gamma < \delta$, *then there exists a unique matrix* $P \in \mathbb{R}^{(n-r)\times r}$ *satisfying* $\|P\| < \frac{2\gamma}{\delta}$ *such that*

$$\begin{aligned} X' &= (X + YP)(I + P^\top P)^{-1/2}, \\ Y' &= (Y - XP^\top)(I + PP^\top)^{-1/2}, \end{aligned}$$

are the eigenvectors of $A + E$.

Define matrix B as

$$B_{i,j} = \frac{1}{m}\sum_{k=1}^{m} \widehat{\lambda}_k \langle \widehat{\varphi}_k, \varphi_i\rangle \langle \widehat{\varphi}_k, \varphi_j\rangle.$$

Let $\mathbf{z}_i$ be the eigenvector of B corresponding to eigenvalue $\widehat{\lambda}_i/m$. It is straightforward to show that

$$\mathbf{z}_i = (\langle \varphi_1, \widehat{\varphi}_i\rangle_{\mathcal{H}_\kappa}, \ldots, \langle \varphi_N, \widehat{\varphi}_i\rangle_{\mathcal{H}_\kappa})^\top, i \in [m],$$

and therefore we have

$$\widehat{\varphi}_i = \sum_{k=1}^{N} z_{i,k}\varphi_k, i \in [m], \text{ or } \Theta = (\Phi, \overline{\Phi})Z,$$

where $Z = (\mathbf{z}_1, \cdots, \mathbf{z}_r)$. To decide the relationship between $\{\widehat{\varphi}_i\}_{i=1}^r$ and $\{\varphi_i\}_{i=1}^N$, we need to determine matrix Z. We define matrix $D = \text{diag}(\lambda_1/N, \ldots, \lambda_N/N)$ and matrix $E = B - D$, i.e.

$$E_{i,j} = B_{i,j} - \lambda_i \delta_{i,j}/N = \langle \varphi_i, (L_m - L_N)\varphi_j\rangle_{\mathcal{H}_\kappa}.$$

Following the notation of Lemma 4, we define $X = (e_1, \ldots, e_r)$ and $Y = (e_{r+1}, \ldots, e_N)$, where $e_1, \ldots, e_N$ are the canonical bases of $\mathbb{R}^N$, which are also eigenvectors of D (which corresponds to matrix A in Lemma 4). Define δ and γ as follows

$$\begin{aligned} \gamma &= \sqrt{\sum_{i=1}^{r}\sum_{j=r+1}^{N} \langle \varphi_i, (L_N - L_m)\varphi_j\rangle_{\mathcal{H}_\kappa}^2}, \\ \delta &= \Delta - \sqrt{\sum_{i,j=1}^{r} \langle \varphi_i, (L_N - L_m)\varphi_j\rangle_{\mathcal{H}_\kappa}^2} - \sqrt{\sum_{i,j=r+1}^{N} \langle \varphi_i, (L_N - L_m)\varphi_j\rangle_{\mathcal{H}_\kappa}^2}. \end{aligned}$$

[1] We simplify the statement to make it better fit with our objective

It is easy to verify that γ, δ are defined with respect to the Frobenius norm of $\widehat{E}$ in Lemma 4. In order to apply the result in Lemma 4, we need to show $\delta > 0$ and $\gamma < \delta/2$. To this end, we need to provide the lower and upper bounds for γ and δ, respectively. We first bound δ as

$$\delta - \Delta \quad \geq \quad -\sqrt{\sum_{i,j=1}^{N} \langle \varphi_i, (L_N - L_m)\varphi_j \rangle_{\mathcal{H}_\kappa}^2} = -\|L_N - L_m\|_{HS}.$$

We then bound γ as

$$\begin{aligned} \gamma &= \sqrt{\sum_{i=1}^{r} \sum_{j=r+1}^{N} \langle \varphi_i, (L_N - L_m)\varphi_j \rangle_{\mathcal{H}_\kappa}^2} \\ &\leq \sqrt{\sum_{i=1}^{N} \sum_{j=1}^{N} \langle \varphi_i, (L_N - L_m)\varphi_j \rangle_{\mathcal{H}_\kappa}^2} = \|L_N - L_m\|_{HS}. \end{aligned}$$

Hence, when $\Delta > 3\|L_N - L_m\|_{HS}$, we have $\delta > 2\gamma > 0$, which satisfies the condition specified in Lemma 4. Thus, according to Lemma 4, there exists a $P \in \mathbb{R}^{(N-r)\times r}$ satisfying $\|P\| < 2\gamma/\delta$, such that

$$Z = (\mathbf{z}_1, \ldots, \mathbf{z}_r) = (X + YP)(I + P^\top P)^{-1/2} = (X, Y) \begin{pmatrix} I_{r\times r} \\ P \end{pmatrix} (I + P^\top P)^{-1/2}.$$

implying

$$\Theta = (\Phi, \overline{\Phi}) Z = (\Phi, \overline{\Phi})(X, Y) \begin{pmatrix} I_{r\times r} \\ P \end{pmatrix} (I + P^\top P)^{-1/2} = (\Phi + \overline{\Phi} P)(I + P^\top P)^{-1/2},$$

where we use $(X, Y) = I$. $\square$

Proof of Theorem 4

Since

$$\|L_N^{1/2} \Delta H L_N^{1/2}\|_2 = \max_{\|f\|_{\mathcal{H}_\kappa} \leq 1} \langle f, L_N^{1/2} \Delta H L_N^{1/2} f \rangle_{\mathcal{H}_\kappa},$$

we need to bound $\langle f, L_N^{1/2} \Delta H L_N^{1/2} f \rangle_{\mathcal{H}_\kappa}$. To this end, we write $f(\cdot) = \sum_{i=1}^N u_i \varphi_i(\cdot)$, where $\sum_{i=1}^N u_i^2 = 1$, and have

$$\begin{aligned} \langle f, L_N^{1/2} \widehat{H}_r L_N^{1/2} f \rangle_{\mathcal{H}_\kappa} &= \sum_{i,j=1}^{N} u_i u_j \langle \varphi_i, L_N^{1/2} \widehat{H}_r L_N^{1/2} \varphi_j \rangle_{\mathcal{H}_\kappa} \\ &= \frac{1}{N} \sum_{i,j=1}^{N} u_i u_j \sqrt{\lambda_i \lambda_j} \langle \varphi_i, \widehat{H}_r \varphi_j \rangle_{\mathcal{H}_\kappa} = \mathbf{u}^\top D A^2 D \mathbf{u}, \end{aligned}$$

where $\mathbf{u} = (u_1, \ldots, u_N)^\top$, $D = \text{diag}(\sqrt{\lambda_1/N}, \ldots, \sqrt{\lambda_N/N})$ and $A = [\langle \varphi_i, \widehat{\varphi}_j \rangle_{\mathcal{H}_\kappa}]_{N\times r} = (\Phi, \overline{\Phi})^\top \Theta$. Since $\Delta > 3\|L_N - L_m\|_{HS}$, according to Lemma 3, there exists a matrix $P \in \mathbb{R}^{(N-r)\times r}$ satisfying

$$\|P\|_F \leq \frac{2\|L_N - L_m\|_{HS}}{\Delta - \|L_N - L_m\|_{HS}},$$

such that $\Theta = \left(\Phi + \overline{\Phi} P\right)(I + P^\top P)^{-1/2}$. Using the expression of Θ, we compute A as

$$\begin{aligned} A &= (\Phi, \overline{\Phi})^\top \Theta = (\Phi, \overline{\Phi})^\top \left(\Phi + \overline{\Phi} P\right)(I + P^\top P)^{-1/2} \\ &= \begin{pmatrix} I \\ P \end{pmatrix} (I + P^\top P)^{-1/2}. \end{aligned}$$

Thus, we have

$$\begin{aligned} \langle f, L_N^{1/2} \Delta H L_N^{1/2} f \rangle_{\mathcal{H}_\kappa} &= \mathbf{u}^\top D \left(\begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} - A A^\top \right) D \mathbf{u} \\ &= \mathbf{u}^\top D B D \mathbf{u}, \end{aligned}$$

where B is given by

$$
\begin{aligned}
B &= \begin{pmatrix} I-(I+P^\top P)^{-1} & (I+P^\top P)^{-1}P^\top \\ P(I+P^\top P)^{-1} & P(I+P^\top P)^{-1}P^\top \end{pmatrix} \\
&= \begin{pmatrix} P(I+P^\top P)^{-1}P^\top & (I+P^\top P)^{-1}P^\top \\ P(I+P^\top P)^{-1} & P(I+P^\top P)^{-1}P^\top \end{pmatrix}.
\end{aligned}
$$

Rewrite $\mathbf{u} = (\mathbf{u}_a, \mathbf{u}_b)$ where $\mathbf{u}_a \in \mathbb{R}^r$ includes the first r entries in $\mathbf{u}$ and $\mathbf{u}_b$ includes the rest of the entries in $\mathbf{u}$. Define $D_a = \mathrm{diag}(\sqrt{\lambda_1/N}, \ldots, \sqrt{\lambda_r/N})$ and $D_b = \mathrm{diag}(\sqrt{\lambda_{r+1}/N}, \ldots, \sqrt{\lambda_N/N})$. Let $Q = (I + P^\top P)^{-1}$. We have

$$
\begin{aligned}
&\mathbf{u}^\top DBD\mathbf{u} \le \|\mathbf{u}_a\|^2 \|D_a PQP^\top D_a\|_2 + \|\mathbf{u}_b\|^2 \|D_b PQP^\top D_b\|_2 + 2\|\mathbf{u}_a\|\|\mathbf{u}_b\|\|D_a QP^\top D_b\|_2 \\
&\le (\|\mathbf{u}_a\| + \|\mathbf{u}_b\|)^2 \max\left(\frac{\lambda_1}{N}\|P\|_2^2, \frac{\sqrt{\lambda_1\lambda_{r+1}}}{N}\|P\|_2\right) \le 2\max\left(\|P\|_2^2, \sqrt{\lambda_{r+1}/N}\|P\|_2\right),
\end{aligned}
$$

where we use the following facts $\left\|PQP^\top\right\|_2 \le \|P\|_2^2$, $\|QP^\top\|_2 = \sqrt{\|PQ^2P^\top\|_2} \le \|P\|_2$, and $\lambda_1 \le N$. We complete the proof by using the fact

$$
\|L_N^{1/2}\Delta H L_N^{1/2}\|_2 = \max_{\|f\|_{\mathcal{H}_\kappa}\le 1} \langle f, L_N^{1/2}\Delta H L_N^{1/2} f\rangle_{\mathcal{H}_\kappa} = \max_{\|\mathbf{u}\|_2 \le 1} \mathbf{u}^\top DBD\mathbf{u}.
$$

and the bound on $\|P\|_2$ in Lemma 3.

References

[1] V. Koltchinskii and M. Yuan. Sparsity in multiple kernel learning. *Annuals of Statistics*, 38:3660–3694, 2010.

[2] H. Q. Minh. Some properties of gaussian reproducing kernel hilbert spaces and their implications for function approximation and learning theory. *Constructive Approximation*, 32(2):307–338, 2010.

[3] G. W. Stewart and J. Sun. *Matrix Perturbation Theory*. Academic Press, 1990.