Approximate Log-Hilbert-Schmidt Distances between Covariance Operators for Image Classification Supplementary Material

Hà Quang Minh[1] Marco San Biagio[1] Loris Bazzani[2] Vittorio Murino[1]

[1] Pattern Analysis and Computer Vision (PAVIS), Istituto Italiano di Tecnologia (IIT), Italy
[2] Department of Computer Science, Dartmouth College, USA
{minh.haquang,marco.sanbiagio,vittorio.murino}@iit.it, loris.bazzani@gmail.com

Abstract

The Supplementary Material contains the proofs for all mathematical results stated in the main paper. We also describe in more detail the Quasi-random Fourier feature map approach in Section 3.1 of the main paper.

1. Proofs for main mathematical results

For clarity, we restate all the mathematical results that we wish to prove here.

Theorem 1. *Assume that $\gamma \neq \mu$, $\gamma > 0$, $\mu > 0$. Then*

$$\lim_{D\to\infty} \left\|\log(C_{\hat{\Phi}_D(\mathbf{x})} + \gamma I_D) - \log(C_{\hat{\Phi}_D(\mathbf{y})} + \mu I_D)\right\|_F = \infty. \tag{1}$$

Theorem 2. *Assume that $\gamma = \mu > 0$ and that $\lim_{D\to\infty} \hat{K}_D(x,y) = K(x,y)$ for every pair $(x,y) \in \mathcal{X}\times\mathcal{X}$. Then*

$$\lim_{D\to\infty} \left\|\log(C_{\hat{\Phi}_D(\mathbf{x})} + \gamma I_D) - \log(C_{\hat{\Phi}_D(\mathbf{y})} + \gamma I_D)\right\|_F = ||\log(C_{\Phi(\mathbf{x})} + \gamma I_{\mathcal{H}}) - \log(C_{\Phi(\mathbf{y})} + \gamma I_{\mathcal{H}})||_{\mathrm{eHS}}. \tag{2}$$

We need the following preliminary results. Let $\mathcal{H}$ be a separable Hilbert space, equipped with norm $||\ ||$, and $A : \mathcal{H} \to \mathcal{H}$ be a bounded linear operator. We recall that the operator norm of A is defined to be

$$||A|| = \sup_{x\neq 0} \frac{||Ax||}{||x||}. \tag{3}$$

If A is self-adjoint, compact, and positive, then

$$||A|| = \lambda_{\max}(A), \tag{4}$$

where $\lambda_{\max}(A)$ denotes the largest eigenvalue of A. The trace norm of A in this case is given by

$$||A||_{\mathrm{tr}} = \sum_{k=1}^{\infty} \lambda_k(A) = \mathrm{tr}(A). \tag{5}$$

Lemma 1. *Let $\mathcal{H}$ be a separable Hilbert space. Let $r \in \mathbb{N}$ be fixed. Let $A \in \mathcal{L}(\mathcal{H})$ be a self-adjoint, positive operator with finite rank $r < \infty$. Then*

$$||A|| \leq ||A||_{\mathrm{HS}} \leq \sqrt{r}||A||, \tag{6}$$

$$||A|| \leq ||A||_{\mathrm{tr}} \leq r||A||. \tag{7}$$

Thus convergences in the $||\ ||_{\mathrm{HS}}$ norm, the $||\ ||_{\mathrm{tr}}$ norm, and the $||\ ||$ norm are all equivalent to each other.

Proof. By definition of the $|| \ ||$ and $|| \ ||_{\mathrm{HS}}$ norms and the finite rank assumption, we have

$$||A||^2 = \lambda_{\max}^2(A) \leq \sum_{j=1}^{r} \lambda_j^2(A) = ||A||_{\mathrm{HS}}^2 \leq r\lambda_{\max}^2(A),$$

from which the first inequality follows. Similarly, for the second inequality, we have

$$||A|| = \lambda_{\max}(A) \leq \sum_{j=1}^{r} \lambda_k(A) = ||A||_{\mathrm{tr}} \leq r\lambda_{\max}(A) = r||A||.$$

This completes the proof of the lemma. □

Lemma 2. *Let $\mathcal{H}$ be a separable Hilbert space. Let $r \in \mathbb{N}$ be fixed. Let $A, \{A_k\}_{k\in\mathbb{N}}$ be self-adjoint, positive operators of rank at most r, such that $\lim_{k\to\infty} ||A_k - A||_{\mathrm{HS}} = 0$. Then*

$$\lim_{k\to\infty} ||\log(I + A_k) - \log(I + A)||_{\mathrm{HS}} = 0. \tag{8}$$

Proof. By assumption, the operators in the sequence $(A_k - A)_{k\in\mathbb{N}}$ all have rank at most $2r$. Thus from Lemma 1, the convergence $||A_k - A||_{\mathrm{HS}}$ is equivalent to the convergence $||A_k - A||$.

The operators in the sequence $(\log(I + A_k))_{k\in\mathbb{N}}$ are also self-adjoint, positive, and of rank at most r. The operators in the sequence $(\log(I + A_k) - \log(I + A))_{k\in\mathbb{N}}$ have rank at most $2r$ and thus the convergence $||\log(I + A_k) - \log(I + A)||_{\mathrm{HS}}$ is equivalent to the convergence $||\log(I + A_k) - \log(I + A)||$. Thus we have

$$||A_k - A||_{\mathrm{HS}} \to 0 \Longleftrightarrow ||A_k - A|| \to 0 \Longleftrightarrow \lambda_{\max}(A_k) \to \lambda_{\max}(A).$$

It follows that

$$\begin{aligned} \log(1 + \lambda_{\max}(A_k)) \to \log(1 + \lambda_{\max}(A)) &\Longleftrightarrow ||\log(I + A_k) - \log(I + A)|| \to 0 \\ &\Longleftrightarrow ||\log(I + A_k) - \log(I + A)||_{\mathrm{HS}} \to 0. \end{aligned}$$

This completes the proof of the lemma. □

Lemma 3. *Let $\mathcal{H}$ be a separable Hilbert space. Let $r \in \mathbb{N}$ be fixed. Let $A, \{A_k\}_{k\in\mathbb{N}}$ be self-adjoint, positive operators of rank at most r, such that $\lim_{k\to\infty} ||A_k - A||_{\mathrm{HS}} = 0$. Then*

$$\lim_{k\to\infty} \mathrm{tr}[\log(I + A_k) - \log(I + A)] = 0. \tag{9}$$

Proof. By assumption, the operators in the sequence $(\log(I + A_k))_{k\in\mathbb{N}}$ are also self-adjoint, positive, and of rank at most r. The operators in the sequence $(\log(I + A_k) - \log(I + A))_{k\in\mathbb{N}}$ have rank at most $2r$ and by Lemma 2, $\lim_{k\to\infty} ||\log(I + A_k) - \log(I + A)||_{\mathrm{HS}} = 0$. By Lemma 1, this convergence is equivalent to convergence in the $|| \ ||_{\mathrm{tr}}$ norm. Thus we have

$$|\mathrm{tr}[\log(I + A_k) - \log(I + A)]| \leq ||\log(I + A_k) - \log(I + A)||_{\mathrm{tr}} \to 0$$

as $k \to \infty$. This completes the proof of the lemma. □

Lemma 4. *Let $\mathcal{H}$ be a separable Hilbert space. Let $r \in \mathbb{N}$ be fixed. Let $A, \{A_k\}_{k\in\mathbb{N}}$, $B, \{B_k\}_{k\in\mathbb{N}}$ be self-adjoint, positive operators of rank at most r, such that $\lim_{k\to\infty} ||A_k - A||_{\mathrm{HS}} = 0$ and $\lim_{k\to\infty} ||B_k - B||_{\mathrm{HS}} = 0$. Then*

$$\lim_{k\to\infty} \mathrm{tr}[\log(I + A_k)\log(I + B_k)] = \mathrm{tr}[\log(I + A)\log(I + B)]. \tag{10}$$

Proof. From Lemma 2, we have

$$\lim_{k\to\infty} ||\log(I + A_k) - \log(I + A)||_{\mathrm{HS}} = 0,$$
$$\lim_{k\to\infty} ||\log(I + B_k) - \log(I + B)||_{\mathrm{HS}} = 0.$$

Thus, using Cauchy-Schwarz inequality and the definition $\langle A, B\rangle_{\mathrm{HS}} = \mathrm{tr}(A^T B)$, we obtain

$$
\begin{aligned}
&|\mathrm{tr}[\log(I+A_k)\log(I+B_k) - \log(I+A)\log(I+B)]| \\
&= |\mathrm{tr}[\log(I+A_k)\log(I+B_k) - \log(I+A_k)\log(I+B) + \log(I+A_k)\log(I+B) - \log(I+A)\log(I+B)]| \\
&= |\mathrm{tr}[\log(I+A_k)(\log(I+B_k) - \log(I+B)) + (\log(I+A_k) - \log(I+A))\log(I+B)]| \\
&= |\langle \log(I+A_k), \log(I+B_k) - \log(I+B)\rangle_{\mathrm{HS}} + \langle \log(I+A_k) - \log(I+A), \log(I+B)\rangle_{\mathrm{HS}}| \\
&\leq |\langle \log(I+A_k), \log(I+B_k) - \log(I+B)\rangle_{\mathrm{HS}}| + |\langle \log(I+A_k) - \log(I+A), \log(I+B)\rangle_{\mathrm{HS}}| \\
&\leq ||\log(I+A_k)||_{\mathrm{HS}} ||\log(I+B_k) - \log(I+B)||_{\mathrm{HS}} + ||\log(I+A_k) - \log(I+A)||_{\mathrm{HS}} ||\log(I+B)||_{\mathrm{HS}}.
\end{aligned}
$$

Taking limit on both sides as $k \to \infty$, we obtain

$$\lim_{k\to\infty} \mathrm{tr}[\log(I+A_k)\log(I+B_k) - \log(I+A)\log(I+B)] = 0.$$

This completes the proof of the lemma. ☐

Lemma 5. *[2] Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two separable Hilbert spaces. Let $A : \mathcal{H}_1 \to \mathcal{H}_2$ and $B : \mathcal{H}_2 \to \mathcal{H}_1$ be two bounded linear operators. Then the nonzero eigenvalues of $BA : \mathcal{H}_1 \to \mathcal{H}_1$ and $AB : \mathcal{H}_2 \to \mathcal{H}_2$, if they exist, are the same.*

In the following, we identify $\mathcal{H}$ with ℓ^2 and $\mathbb{R}^D$ with a D-dimensional subspace of ℓ^2, that is

$$a = (a_j)_{j=1}^D \in \mathbb{R}^D \Longleftrightarrow a = (a_1, \ldots, a_D, 0, 0, \ldots) \in \ell^2. \tag{11}$$

For the data matrices $\mathbf{x} = [x_1, \ldots, x_m]$, $\mathbf{y} = [y_1, \ldots, y_m]$, let $K[\mathbf{x}], K[\mathbf{y}], \hat{K}_D[\mathbf{x}], \hat{K}_D[\mathbf{y}]$ be the $m \times m$ Gram matrices, defined by

$$(K[\mathbf{x}])_{ij} = K(x_i, x_j), \quad (\hat{K}_D[\mathbf{x}])_{ij} = \hat{K}_D(x_i, x_j), \quad (K[\mathbf{y}])_{ij} = K(y_i, y_j), \quad (\hat{K}_D[\mathbf{y}])_{ij} = \hat{K}_D(y_i, y_j).$$

Lemma 6. *Let $\mathcal{H} = \ell^2$, with $\mathbb{R}^D$ identified with a D-dimensional subspace of $\mathcal{H}$ as in Eq. (11). Assume that $\lim_{D\to\infty} \hat{K}_D(x, y) = K(x, y)$ for all pairs $(x, y) \in \mathcal{X} \times \mathcal{X}$. Then*

$$\lim_{D\to\infty} ||C_{\hat{\Phi}_D(\mathbf{x})} - C_{\Phi(\mathbf{x})}||_{\mathrm{HS}(\mathcal{H})} = 0. \tag{12}$$

Proof. Let $A = \frac{1}{\sqrt{m}}\Phi(\mathbf{x})J_m : \mathbb{R}^m \to \mathcal{H}$, then

$$AA^T = C_{\Phi(\mathbf{x})}, \quad A^T A = \frac{1}{m} J_m K[\mathbf{x}] J_m.$$

By Lemma 5, the nonzero eigenvalues of $C_{\Phi(\mathbf{x})} = AA^T$ are the same as those of $\frac{1}{m}J_m K[\mathbf{x}]J_m = A^T A$. Similarly, the nonzero eigenvalues of $C_{\hat{\Phi}_D(\mathbf{x})}$ are the same as those of $\frac{1}{m}J_m \hat{K}_D[\mathbf{x}]J_m$. This also implies that both $C_{\Phi(\mathbf{x})}$ and $C_{\hat{\Phi}_D(\mathbf{x})}$ have rank at most $m-1$, since $\mathrm{rank}(J_m) = m-1$.

Since $\lim_{D\to\infty} \hat{K}_D(x_i, x_j) = K(x_i, x_j)$ for all pairs (x_i, x_j), $1 \leq i, j \leq m$, we have, as $m \times m$ matrices,

$$\lim_{D\to\infty} ||J_m \hat{K}_D[\mathbf{x}]J_m - J_m K[\mathbf{x}]J_m||_F = 0.$$

Since $J_m \hat{K}_D[\mathbf{x}]J_m$ and $J_m K[\mathbf{x}]J_m$ are finite matrices, convergence in the $||\ ||_F$ norm is equivalent to convergence in the operator $||\ ||$ norm. Thus we have

$$
\begin{aligned}
\lim_{D\to\infty} ||J_m \hat{K}_D[\mathbf{x}]J_m - J_m K[\mathbf{x}]J_m|| = 0 &\Longleftrightarrow \lim_{D\to\infty} \lambda_{\max}(J_m \hat{K}_D[\mathbf{x}]J_m) = \lambda_{\max}(J_m K[\mathbf{x}]J_m) \\
&\Longleftrightarrow \lim_{D\to\infty} \lambda_{\max}(C_{\hat{\Phi}_D(\mathbf{x})}) = \lambda_{\max}(C_{\Phi(\mathbf{x})}) \Longleftrightarrow \lim_{D\to\infty} ||C_{\hat{\Phi}_D(\mathbf{x})} - C_{\Phi(\mathbf{x})}|| = 0 \\
&\Longleftrightarrow \lim_{D\to\infty} ||C_{\hat{\Phi}_D(\mathbf{x})} - C_{\Phi(\mathbf{x})}||_{\mathrm{HS}(\mathcal{H})} = 0,
\end{aligned}
$$

by Lemma 1, since both $C_{\hat{\Phi}_D(\mathbf{x})}, C_{\Phi(\mathbf{x})}$ have rank at most $m-1$. This completes the proof of the lemma. ☐

Proof of Theorem 1. Consider the expansion

$$\left\|\log\left(C_{\hat{\Phi}_D(\mathbf{x})}+\gamma I_D\right)-\log\left(C_{\hat{\Phi}_D(\mathbf{y})}+\mu I_D\right)\right\|_F^2=\left\|\log\left(\frac{C_{\hat{\Phi}_D(\mathbf{x})}}{\gamma}+I_D\right)-\log\left(\frac{C_{\hat{\Phi}_D(\mathbf{y})}}{\mu}+I_D\right)+(\log\gamma-\log\mu)I_D\right\|_F^2$$
$$=\left\|\log\left(\frac{C_{\hat{\Phi}_D(\mathbf{x})}}{\gamma}+I_D\right)-\log\left(\frac{C_{\hat{\Phi}_D(\mathbf{y})}}{\mu}+I_D\right)\right\|_F^2+2(\log\gamma-\log\mu)\mathrm{tr}\left(\log\left(\frac{C_{\hat{\Phi}_D(\mathbf{x})}}{\gamma}+I_D\right)-\log\left(\frac{C_{\hat{\Phi}_D(\mathbf{y})}}{\mu}+I_D\right)\right)$$
$$+(\log\gamma-\log\mu)^2D. \tag{13}$$

With $\mathbb{R}^D$ identified as a subspace of $\mathcal{H}=\ell^2$, we have by Lemma 6 (with the scaling factors γ, μ), that

$$\lim_{D\to\infty}\left\|\frac{C_{\hat{\Phi}_D(\mathbf{x})}}{\gamma}-\frac{C_{\Phi(\mathbf{x})}}{\gamma}\right\|_{\mathrm{HS}(\mathcal{H})}^2=0,\quad \lim_{D\to\infty}\left\|\frac{C_{\hat{\Phi}_D(\mathbf{y})}}{\mu}-\frac{C_{\Phi(\mathbf{y})}}{\mu}\right\|_{\mathrm{HS}(\mathcal{H})}^2=0.$$

By Lemma 3, we have

$$\lim_{D\to\infty}\mathrm{tr}\left(\log\left(\frac{C_{\hat{\Phi}_D(\mathbf{x})}}{\gamma}+I_D\right)\right)=\mathrm{tr}\left(\log\left(\frac{C_{\Phi(\mathbf{x})}}{\gamma}+I_{\mathcal{H}}\right)\right)=\mathrm{tr}\left[\log\left(\frac{1}{\gamma m}J_mK[\mathbf{x}]J_m+I_m\right)\right],$$
$$\lim_{D\to\infty}\mathrm{tr}\left(\log\left(\frac{C_{\hat{\Phi}_D(\mathbf{y})}}{\mu}+I_D\right)\right)=\mathrm{tr}\left(\log\left(\frac{C_{\Phi(\mathbf{y})}}{\mu}+I_{\mathcal{H}}\right)\right)=\mathrm{tr}\left[\log\left(\frac{1}{\mu m}J_mK[\mathbf{y}]J_m+I_m\right)\right].$$

Since these two quantities are both finite, for $\gamma\neq\mu$, as $D\to\infty$, clearly the right hand side of Eq. (13) goes to infinity. This gives us the desired limit. □

Proof of Theorem 2. Without loss of generality, we identify $\mathcal{H}$ with ℓ^2 as above and identify $\mathbb{R}^D$ with a D-dimensional subspace of ℓ^2 as in Eq. (11). When $\gamma=\mu$, we have

$$||\log(C_{\Phi(\mathbf{x})}+\gamma I_{\mathcal{H}})-\log(C_{\Phi(\mathbf{y})}+\gamma I_{\mathcal{H}})||_{\mathrm{eHS}}^2=\left\|\log\left(\frac{C_{\Phi(\mathbf{x})}}{\gamma}+I_{\mathcal{H}}\right)-\log\left(\frac{C_{\Phi(\mathbf{y})}}{\gamma}+I_{\mathcal{H}}\right)\right\|_{\mathrm{HS}}^2$$
$$=\left\|\log\left(\frac{C_{\Phi(\mathbf{x})}}{\gamma}+I_{\mathcal{H}}\right)\right\|_{\mathrm{HS}}^2+\left\|\log\left(\frac{C_{\Phi(\mathbf{y})}}{\gamma}+I_{\mathcal{H}}\right)\right\|_{\mathrm{HS}}^2-2\mathrm{tr}\left[\log\left(\frac{C_{\Phi(\mathbf{x})}}{\gamma}+I_{\mathcal{H}}\right)\log\left(\frac{C_{\Phi(\mathbf{y})}}{\gamma}+I_{\mathcal{H}}\right)\right].$$

It follows from Lemma 5 that the first term is

$$\left\|\log\left(\frac{1}{\gamma}C_{\Phi(\mathbf{x})}+I_{\mathcal{H}}\right)\right\|_{\mathrm{HS}}^2=\left\|\log\left(\frac{1}{\gamma m}\Phi(\mathbf{x})J_m^2\Phi(\mathbf{x})^T+I_{\mathcal{H}}\right)\right\|_{\mathrm{HS}}^2=\left\|\log\left(\frac{1}{\gamma m}J_m\Phi(\mathbf{x})^T\Phi(\mathbf{x})J_m+I_m\right)\right\|_{\mathrm{HS}}^2$$
$$=\left\|\log\left(\frac{1}{\gamma m}J_mK[\mathbf{x}]J_m+I_m\right)\right\|_{\mathrm{HS}}^2=\mathrm{tr}\left[\log\left(\frac{1}{\gamma m}J_mK[\mathbf{x}]J_m+I_m\right)\right]^2.$$

Similarly, the second term is

$$\left\|\log\left(\frac{1}{\gamma}C_{\Phi(\mathbf{y})}+I_{\mathcal{H}}\right)\right\|_{\mathrm{HS}}^2=\mathrm{tr}\left[\log\left(\frac{1}{\gamma m}J_mK[\mathbf{y}]J_m+I_m\right)\right]^2.$$

Thus we have

$$||\log(C_{\Phi(\mathbf{x})}+\gamma I_{\mathcal{H}})-\log(C_{\Phi(\mathbf{y})}+\gamma I_{\mathcal{H}})||_{\mathrm{eHS}}^2=\mathrm{tr}\left[\log\left(\frac{1}{\gamma m}J_mK[\mathbf{x}]J_m+I_m\right)\right]^2+\mathrm{tr}\left[\log\left(\frac{1}{\gamma m}J_mK[\mathbf{y}]J_m+I_m\right)\right]^2$$
$$-2\mathrm{tr}\left[\log\left(\frac{C_{\Phi(\mathbf{x})}}{\gamma}+I_{\mathcal{H}}\right)\log\left(\frac{C_{\Phi(\mathbf{y})}}{\gamma}+I_{\mathcal{H}}\right)\right]. \tag{14}$$

Similarly,

$$||\log(C_{\hat{\Phi}_D(\mathbf{x})}+\gamma I_D)-\log(C_{\hat{\Phi}_D(\mathbf{y})}+\gamma I_D)||_F^2=\mathrm{tr}\left[\log\left(\frac{1}{\gamma m}J_m\hat{K}_D[\mathbf{x}]J_m+I_m\right)\right]^2+\mathrm{tr}\left[\log\left(\frac{1}{\gamma m}J_m\hat{K}_D[\mathbf{y}]J_m+I_m\right)\right]^2$$
$$-2\mathrm{tr}\left[\log\left(\frac{C_{\hat{\Phi}_D(\mathbf{x})}}{\gamma}+I_D\right)\log\left(\frac{C_{\hat{\Phi}_D(\mathbf{y})}}{\gamma}+I_D\right)\right]. \tag{15}$$

With $\mathbb{R}^D$ identified as a subspace of $\mathcal{H} = \ell^2$, we have by Lemma 6 (with the scaling factor γ), that

$$\lim_{D\to\infty} \left\|\frac{C_{\hat{\Phi}_D(\mathbf{x})}}{\gamma} - \frac{C_{\Phi(\mathbf{x})}}{\gamma}\right\|^2_{\mathrm{HS}(\mathcal{H})} = 0, \quad \lim_{D\to\infty} \left\|\frac{C_{\hat{\Phi}_D(\mathbf{y})}}{\gamma} - \frac{C_{\Phi(\mathbf{y})}}{\gamma}\right\|^2_{\mathrm{HS}(\mathcal{H})} = 0,$$

with the operators $C_{\hat{\Phi}_D(\mathbf{x})}, C_{\Phi(\mathbf{x})}, C_{\hat{\Phi}_D(\mathbf{y})}, C_{\Phi(\mathbf{y})}$ all have rank at most $m-1$. It thus follows from Lemma 4 that

$$\lim_{D\to\infty} \mathrm{tr}\left[\log\left(\frac{C_{\hat{\Phi}_D(\mathbf{x})}}{\gamma} + I_D\right)\log\left(\frac{C_{\hat{\Phi}_D(\mathbf{y})}}{\gamma} + I_D\right)\right] = \mathrm{tr}\left[\log\left(\frac{C_{\Phi(\mathbf{x})}}{\gamma} + I_{\mathcal{H}}\right)\log\left(\frac{C_{\Phi(\mathbf{y})}}{\gamma} + I_{\mathcal{H}}\right)\right]. \tag{16}$$

Similarly, since $\lim_{D\to\infty} \hat{K}_D(x_i, x_j) = K(x_i, x_j)$ for all pairs (x_i, x_j) and $\lim_{D\to\infty} \hat{K}_D(y_i, y_j) = K(y_i, y_j)$ for all pairs (y_i, y_j), $1 \leq i, j \leq m$, we have, as $m \times m$ matrices,

$$\lim_{D\to\infty} ||J_m\hat{K}_D[\mathbf{x}]J_m - J_mK[\mathbf{x}]J_m||_F = 0, \quad \lim_{D\to\infty} ||J_m\hat{K}_D[\mathbf{y}]J_m - J_mK[\mathbf{y}]J_m||_F = 0.$$

It also follows from Lemma 4 that

$$\lim_{D\to\infty} \mathrm{tr}\left[\log\left(\frac{1}{\gamma m}J_m\hat{K}_D[\mathbf{x}]J_m + I_m\right)\right]^2 = \mathrm{tr}\left[\log\left(\frac{1}{\gamma m}J_mK[\mathbf{x}]J_m + I_m\right)\right]^2, \tag{17}$$

$$\lim_{D\to\infty} \mathrm{tr}\left[\log\left(\frac{1}{\gamma m}J_m\hat{K}_D[\mathbf{y}]J_m + I_m\right)\right]^2 = \mathrm{tr}\left[\log\left(\frac{1}{\gamma m}J_mK[\mathbf{y}]J_m + I_m\right)\right]^2. \tag{18}$$

Combining the expressions in Eqs. (14), (15), (16), (17), (18), we obtain

$$\lim_{D\to\infty} ||\log(C_{\hat{\Phi}_D(\mathbf{x})} + \gamma I_D) - \log(C_{\hat{\Phi}_D(\mathbf{y})} + \gamma I_D)||^2_F = ||\log(C_{\Phi(\mathbf{x})} + \gamma I_{\mathcal{H}}) - \log(C_{\Phi(\mathbf{y})} + \gamma I_{\mathcal{H}})||^2_{\mathrm{eHS}}.$$

This completes the proof of the Theorem. □

2. Further information on the Hilbert-Schmidt distance between covariance operators

For completeness, in this section we provide the mathematical expression for the Hilbert-Schmidt distance between two RKHS covariance operators. This was used for carrying out the corresponding experiments on the Fish dataset in the main paper. Let K be a positive definite kernel on an arbitrary non-empty set $\mathcal{X}$ and $\mathcal{H}_K$ be its corresponding RKHS. Let $C_{\Phi(\mathbf{x})}$ and $C_{\Phi(\mathbf{y})}$ be the covariance operators corresponding to two $n \times m$ data matrices $\mathbf{x}$ and $\mathbf{y}$, respectively, sampled from $\mathcal{X}$. Following [2], let $K[\mathbf{x}]$, $K[\mathbf{y}]$, and $K[\mathbf{x}, \mathbf{y}]$ denote the $m \times m$ Gram matrices defined by

$$(K[\mathbf{x}])_{ij} = K(x_i, x_j), \quad (K[\mathbf{y}])_{ij} = K(y_i, y_j), \quad (K[\mathbf{x}, \mathbf{y}])_{ij} = K(x_i, y_j), \quad 1 \leq i, j \leq m.$$

Then the Gram matrices and the covariance operators are related by

$$\Phi(\mathbf{x})^T\Phi(\mathbf{x}) = K[\mathbf{x}], \quad \Phi(\mathbf{y})^T\Phi(\mathbf{y}) = K[\mathbf{y}], \quad \Phi(\mathbf{x})^T\Phi(\mathbf{y}) = K[\mathbf{x}, \mathbf{y}].$$

Here $\Phi(\mathbf{x})^T$ denotes the transpose of $\Phi(\mathbf{x})$ in the case $\dim(\mathcal{H}_K) < \infty$ and the adjoint operator of $\Phi(\mathbf{x})$ in the case $\dim(\mathcal{H}_K) = \infty$.

Lemma 7. *The Hilbert-Schmidt distances between two RKHS covariance operators $C_{\Phi(\mathbf{x})}$ and $C_{\Phi(\mathbf{y})}$ is given by*

$$||C_{\Phi(\mathbf{x})} - C_{\Phi(\mathbf{y})}||^2_{\mathrm{HS}} = \frac{1}{m^2}\langle J_mK[\mathbf{x}], K[\mathbf{x}]J_m\rangle_F - \frac{2}{m^2}\langle J_mK[\mathbf{x},\mathbf{y}], K[\mathbf{x},\mathbf{y}]J_m\rangle_F + \frac{1}{m^2}\langle J_mK[\mathbf{y}], K[\mathbf{y}]J_m\rangle_F. \tag{19}$$

Proof of Lemma 7. By definition of the Hilbert-Schmidt norm and property of the trace operation, we have

$$||C_{\Phi(\mathbf{x})} - C_{\Phi(\mathbf{y})}||^2_{\mathrm{HS}} = \left\|\frac{1}{m}\Phi(\mathbf{x})J_m\Phi(\mathbf{x})^T - \frac{1}{m}\Phi(\mathbf{y})J_m\Phi(\mathbf{y})^T\right\|^2_{\mathrm{HS}}$$
$$= \frac{1}{m^2}||\Phi(\mathbf{x})J_m\Phi(\mathbf{x})^T||^2_{\mathrm{HS}} - \frac{2}{m^2}\langle\Phi(\mathbf{x})J_m\Phi(\mathbf{x})^T, \Phi(\mathbf{y})J_m\Phi(\mathbf{y})^T\rangle_{\mathrm{HS}} + \frac{1}{m^2}||\Phi(\mathbf{y})J_m\Phi(\mathbf{y})^T||^2_{\mathrm{HS}}$$
$$= \frac{1}{m^2}\mathrm{tr}[\Phi(\mathbf{x})J_m\Phi(\mathbf{x})^T\Phi(\mathbf{x})J_m\Phi(\mathbf{x})^T] - \frac{2}{m^2}\mathrm{tr}[\Phi(\mathbf{x})J_m\Phi(\mathbf{x})^T\Phi(\mathbf{y})J_m\Phi(\mathbf{y})^T] + \frac{1}{m^2}\mathrm{tr}[\Phi(\mathbf{y})J_m\Phi(\mathbf{y})^T\Phi(\mathbf{y})J_m\Phi(\mathbf{y})^T]$$
$$= \frac{1}{m^2}\mathrm{tr}[(K[\mathbf{x}]J_m)^2 - 2K[\mathbf{y},\mathbf{x}]J_mK[\mathbf{x},\mathbf{y}]J_m + (K[\mathbf{y}]J_m)^2]$$
$$= \frac{1}{m^2}[\langle J_mK[\mathbf{x}], K[\mathbf{x}]J_m\rangle_F - 2\langle J_mK[\mathbf{x},\mathbf{y}], K[\mathbf{x},\mathbf{y}]J_m\rangle_F + \langle J_mK[\mathbf{y}], K[\mathbf{y}]J_m\rangle_F].$$

This completes the proof of the lemma. □

3. Further information on the Quasi-random Fourier features

Consider again the expression of the kernel $K(x, y) = k(x - y)$ by Bochner's theorem

$$
\begin{aligned}
k(x-y) &= \int_{\mathbb{R}^n} e^{-i\langle \omega, x-y\rangle} d\rho(\omega) \\
&= \int_{\mathbb{R}^n} \rho(\omega)\phi_\omega(x)\overline{\phi_\omega(y)}d\omega, \text{ where } \phi_\omega(x) = e^{-i\langle\omega, x\rangle}.
\end{aligned} \tag{20}
$$

The Random Fourier feature maps arise from the Monte-Carlo approximation of the integral in Eq. (20), using a *random* set of points ω_j's sampled according to the distribution ρ. In this section, we describe in more detail the Quasi-random Fourier features approach proposed recently by [4]. This approach is based on the methodology of Quasi-Monte Carlo integration [1], in which the ω_j's are *deterministic* points arising from a *low-discrepancy* sequence in $[0, 1]^n$ (see below for more details).

Assume that the distribution ρ in Eq. (20) has the product form $\rho(\omega) = \prod_{j=1}^n \rho_j(\omega_j)$. Assume that each component cumulative distribution function $\psi_j(x_j) = \int_{-\infty}^{x_j} \rho_j(z_j)dz_j$ is strictly increasing, so that the inverse functions $\psi_j^{-1} : [0, 1] \to \mathbb{R}$ are all well-defined. Let $\psi : \mathbb{R}^n \to [0, 1]^n$ be defined by $\psi(x) = \psi(x_1, \ldots, x_n) = (\psi_1(x_1), \ldots, \psi_n(x_n))$. Then its inverse function $\psi^{-1} : [0, 1]^n \to \mathbb{R}^n$ is well-defined and is given component-wise by $\psi^{-1}(z) = \psi^{-1}(z_1, \ldots, z_n) = (\psi_1^{-1}(z_1), \ldots, \psi_n^{-1}(z_n))$.

With the change of variable $\omega = \psi^{-1}(t)$, the integral in Eq. (20) becomes

$$
\int_{\mathbb{R}^n} e^{-i\langle\omega, x-y\rangle}\rho(\omega)d\omega = \int_{[0,1]^n} e^{-i\langle\psi^{-1}(t), x-y\rangle}dt. \tag{21}
$$

Instead of approximating the left hand side of Eq. (21) using a random set of points $\{\omega_j\}_{j=1}^D$ in $\mathbb{R}^n$ sampled according to ρ, in the Quasi-Monte Carlo approach, one approximates the right hand side using a deterministic, low-discrepancy sequence of points $\{t_j\}_{j=1}^D$ in $[0, 1]^n$. This sequence gives rise to a deterministic sequence

$$
\omega_j = \psi^{-1}(t_j), \quad 1 \le j \le D, \tag{22}
$$

from which we construct the Fourier feature map as described by Eqs. (23), (24), and (25),

$$
\cos(W^T x) = (\cos(\langle\omega_1, x\rangle), \ldots, \cos(\langle\omega_D, x\rangle))^T \in \mathbb{R}^D, \tag{23}
$$

$$
\sin(W^T x) = (\sin(\langle\omega_1, x\rangle), \ldots, \sin(\langle\omega_D, x\rangle))^T \in \mathbb{R}^D. \tag{24}
$$

$$
\hat{\Phi}_D(x) = \frac{1}{\sqrt{D}}(\cos(W^T x); \sin(W^T x)) \in \mathbb{R}^{2D}, \tag{25}
$$

just as in the case of random Fourier features. In our experiments, $\{t_j\}_{j=1}^D$ is a Halton sequence, whose implementation is readily available in MATLAB[1].

3.1. Low-discrepancy sequences

In this section, we briefly review the concept of *low-discrepancy sequences* in Quasi-Monte Carlo methods. For a comprehensive treatment, we refer to [3]. Let $n \in \mathbb{N}$ be fixed. Let $I^n = [0, 1)^n$ and denote its closure by $\overline{I}^n = [0, 1]^n$. For an integrable function f in $\overline{I}^n$, we consider the approximation

$$
\int_{\overline{I}^n} f(u)du \approx \frac{1}{N}\sum_{j=1}^N f(x_j) \tag{26}
$$

using a deterministic set of points $P = (x_1, \ldots, x_N)$, which are part of an infinite sequence $(x_j)_{j\in\mathbb{N}}$ in $\overline{I}^n$, such that the integration error satisfies

$$
\lim_{N\to\infty} \left| \frac{1}{N}\sum_{j=1}^N f(x_j) - \int_{\overline{I}^n} f(u)du \right| = 0. \tag{27}
$$

[1] http://www.mathworks.com/help/stats/quasi-random-numbers.html

This convergence can be measured via the concept of *discrepancy* as follows. Let N be fixed. For an arbitrary set $B \subset \overline{I}^n$, define the counting function

$$A(B; P) = \sum_{j=1}^{N} \chi_B(x_j), \tag{28}$$

where χ_B denotes the characteristic function for B. Thus $A(B; P)$ denotes the number of points in P that lie in the set B.

Let $\mathcal{B}$ be a non-empty family of Lebesgue-measurable subsets of $\overline{I}^n$. The discrepancy of the set P with respect to $\mathcal{B}$ is then defined by

$$D_N(\mathcal{B}; P) = \sup_{B \in \mathcal{B}} \left| \frac{A(B; P)}{N} - \mathrm{vol}(B) \right|, \tag{29}$$

with $\mathrm{vol}(B)$ denoting the volume of B with respect to the Lebesgue measure.

The *star discrepancy* $D_N^*(P)$ is defined by

$$D_N^*(P) = D_N(\mathcal{J}^*; P), \tag{30}$$

where $\mathcal{J}^*$ denotes the family of all subintervals of I^n of the form $\prod_{j=1}^{n}[0, x_j)$. The star discrepancy and the integration error are related via the Koksma- Hlawka inequality, as follows. Define

$$V(f) = \sum_{k=1}^{n} \sum_{1 \leq i_1 \leq \cdots \leq i_k \leq n} \int_0^1 \cdots \int_0^1 \left| \frac{\partial^k f}{\partial u_{i_1} \ldots \partial u_{i_k}} \right| du_{i_1} \ldots du_{i_k}, \tag{31}$$

which is called the *variation of f on $\overline{I}^n$ in the sense of Hardy-Krause*.

Theorem 3 (**Koksma-Hlawka inequality**). *If f has bounded variation $V(f)$ on $\overline{I}^n$ in the sense of Hardy-Krause, then for any set $(x_1, \ldots, x_N)$ in I^n,*

$$\left| \frac{1}{N} \sum_{j=1}^{N} f(x_j) - \int_{\overline{I}^n} f(u)du \right| \leq V(f) D_N^*(x_1, \ldots, x_N). \tag{32}$$

By Theorem 3, to achieve a small integration error, we need a sequence $(x_j)_{j \in \mathbb{N}}$ with *low discrepancy* $D_N^*(x_1, \ldots, x_N) \to 0$ as $N \to \infty$. Some examples of low-discrepancy sequences are Halton and Sobol' sequences (we refer to [3, 1] for the detailed constructions of these and other sequences). The Halton sequence in particular satisfies $D_N^*(x_1, \ldots, x_N) = C(n)\frac{(\log N)^n}{N}$ for $N \geq 2$.

3.2. The Gaussian case

In this section, we give the explicit expression for the functions ψ and ψ^{-1}, as defined above, in the case of the Gaussian kernel. It suffices for us to consider the one-dimensional setting here, since the multivariate case is defined componentwise using the one-dimensional case. For $K(x, y) = e^{-\frac{(x-y)^2}{\sigma^2}}$, we have $\rho(z) = \frac{\sigma}{2\sqrt{\pi}} e^{-\frac{\sigma^2 z^2}{4}}$. Recall the Gaussian error function erf defined by $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-z^2} dz$ and the complementary Gaussian error function erfc defined by $\mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-z^2} dz = 1 - \mathrm{erf}(x)$. By definition, the cumulative distribution function ψ for ρ is given by

$$\psi(x) = \int_{-\infty}^{x} \rho(z)dz = 1 - \int_x^\infty \rho(z)dz = 1 - \frac{\sigma}{2\sqrt{\pi}} \int_x^\infty e^{-\frac{\sigma^2 z^2}{4}} dz = 1 - \frac{1}{\sqrt{\pi}} \int_{\frac{x\sigma}{2}}^\infty e^{-u^2} du = 1 - \frac{1}{2}\mathrm{erfc}\left(\frac{x\sigma}{2}\right).$$

It follows that the inverse function ψ^{-1} is given by

$$x = \psi^{-1}(t) = \frac{2}{\sigma} \mathrm{erfc}^{-1}(2 - 2t) = \frac{2}{\sigma} \mathrm{erf}^{-1}(2t - 1). \tag{33}$$

References

[1] J. Dick, F. Kuo, and I. Sloan. High-dimensional integration: the quasi-Monte Carlo way. *Acta Numerica*, 22:133–288, 2013. 6, 7

[2] H. Minh, M. San Biagio, and V. Murino. Log-Hilbert-Schmidt metric between positive definite operators on Hilbert spaces. In *NIPS*, 2014. 3, 5

[3] H. Niederreiter. *Random number generation and quasi-Monte Carlo methods*. SIAM, 1992. 6, 7

[4] J. Yang, V. Sindhwani, H. Avron, and M. Mahoney. Quasi-Monte Carlo feature maps for shift-invariant kernels. In *ICML*, pages 485–493, 2014. 6