APPENDIX: The Geometry of Random Features

We present here proofs of all the theoretical results presented in the main body of the paper. Following the proofs, in Section 12 we provide additional experimental results.

8 Proof of result from Section 2

8.1 Proof of Lemma 2.1

Proof. Recall that characteristic functions uniquely characterise measures. Let $\mathbf{M} \in O_n$ be arbitrary, and consider the push-forward measure $\mathbf{M}_{\#}\mu$, defined by $\mathbf{M}_{\#}\mu(A) = \mu(\mathbf{M}^{-1}A)$ for all $A \in \mathcal{B}(\mathbb{R}^n)$. Our goal is to show that $\mathbf{M}_{\#}\mu = \mu$. Let $\mathbf{s} \in \mathbb{R}^n$, and observe that

$$\begin{aligned}\int_{\mathbb{R}^n} \exp(i\langle \mathbf{s}, \mathbf{w}\rangle)\mathbf{M}_{\#}\mu(d\mathbf{w}) &= \int_{\mathbb{R}^n} \exp(i\langle \mathbf{s}, \mathbf{M}^{-1}\mathbf{w}\rangle)\mu(d\mathbf{w}) \\ &= \int_{\mathbb{R}^n} \exp(i\langle \mathbf{M}\mathbf{s}, \mathbf{w}\rangle)\mu(d\mathbf{w}) \\ &= \phi(\|\mathbf{M}\mathbf{s}\|) \\ &= \phi(\|\mathbf{s}\|) \\ &= \int_{\mathbb{R}^n} \exp(i\langle \mathbf{s}, \mathbf{w}\rangle)\mu(d\mathbf{w})\,,\end{aligned}$$

so the characteristic functions of μ and $\mu_{\#}\mathbf{M}$ agree everywhere on $\mathbb{R}^n$, and the conclusion follows. □

9 Proof of results from Section 3

We begin by establishing the following result, which will be useful in the proofs of Theorems 3.1 and 9.4.

Proposition 9.1. *The difference in mean squared error* MSE *between the estimator* $\widehat{K}^{\mathrm{iid}}_{m,n}(\mathbf{x},\mathbf{y})$ *and* $\widehat{K}^{\mathrm{ort}}_{m,n}(\mathbf{x},\mathbf{y})$ *for* $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$ *is given by the formula below.*

$$\begin{aligned}&\mathrm{MSE}(\widehat{K}^{\mathrm{iid}}_{m,n}(\mathbf{x},\mathbf{y})) - \mathrm{MSE}(\widehat{K}^{\mathrm{ort}}_{m,n}(\mathbf{x},\mathbf{y})) = \\ &\frac{m-1}{m}\left(\mathbb{E}\left[\cos(R_1\|\mathbf{z}\|\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle)\right]^2 - \mathbb{E}\left[\cos(\sqrt{R_1^2+R_2^2}\|\mathbf{z}\|\,\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle)\right]\right)\end{aligned}$$

where $\mathbf{w}_1, \mathbf{w}_2$ *are independent samples from the corresponding probabilistic measure* μ_K *and* m *is the number of random features used.*

Proof. From now on we will often drop index n while referring to the probability measure μ_n if from the context it will be always clear what the dimesionality under consideration is. By independence of the $(\mathbf{w}_p)_{p=1}^m$ in the case of the estimator $\widehat{K}^{\mathrm{iid}}_{m,n}(\mathbf{z})$, we have

$$\mathrm{MSE}(\widehat{K}^{\mathrm{iid}}_{m,n}(\mathbf{z})) = \frac{1}{m}\mathrm{Var}(\cos(\langle \mathbf{w}_1, \mathbf{z}\rangle)).$$

Considering the analogous quantity for $\widehat{K}^{\mathrm{ort}}_{m,n}(\mathbf{z})$, we note that it differs from the expression above by the sum of $m(m-1)$ equal covariance terms. The covariance term is of the form

$$\frac{1}{m^2}\mathbb{E}\left[\cos(\langle \mathbf{w}^{\mathrm{ort}}_1, \mathbf{z}\rangle)\cos(\langle \mathbf{w}^{\mathrm{ort}}_2, \mathbf{z}\rangle)\right] - \mathbb{E}\left[\cos(\langle \mathbf{w}^{\mathrm{ort}}_1, \mathbf{z}\rangle)\right]^2\,, \tag{16}$$

where $\mathbf{w}^{\mathrm{ort}}_1, \mathbf{w}^{\mathrm{ort}}_2$ are both marginally distributed according to μ, and are conditioned to be almost-surely orthogonal. Consider the first term of Equation (16). We use the product-to-sum trigonometric identity

$$\cos(a)\cos(b) = \frac{1}{2}(\cos(a+b) + \cos(a-b)) \qquad \forall a, b \in \mathbb{R}\,,$$

to obtain that

$$\begin{aligned}&\mathbb{E}\left[\cos(\langle \mathbf{w}_1^{\mathrm{ort}}, \mathbf{z}\rangle)\cos(\langle \mathbf{w}_2^{\mathrm{ort}}, \mathbf{z}\rangle)\right]\\ =&\frac{1}{2}\left(\mathbb{E}\left[\cos(\langle \mathbf{w}_1^{\mathrm{ort}} + \mathbf{w}_2^{\mathrm{ort}}, \mathbf{z}\rangle)\right] + \mathbb{E}\left[\cos(\langle \mathbf{w}_1^{\mathrm{ort}} - \mathbf{w}_2^{\mathrm{ort}}, \mathbf{z}\rangle)\right]\right) \, .\end{aligned}$$

Note that $\mathbf{w}_1^{\mathrm{ort}} + \mathbf{w}_2^{\mathrm{ort}} \overset{d}{=} \mathbf{w}_1^{\mathrm{ort}} - \mathbf{w}_2^{\mathrm{ort}}$, so it is sufficient to deal with the first term in the final expression above. Since μ is rotationally-invariant, each $\mathbf{w}$ drawn from μ can be decomposed as

$$\mathbf{w} = R\mathbf{v}, \tag{17}$$

where $\mathbf{v} \sim \mathrm{Unif}(S^{n-1})$, and independently, R is a scalar random variable drawn from the distribution of the norm induced by μ. The key observation now is that we have a similar decomposition to Equation (17) for $\mathbf{w}_1^{\mathrm{ort}} + \mathbf{w}_2^{\mathrm{ort}}$; indeed, we have

$$\mathbf{w}_1^{\mathrm{ort}} + \mathbf{w}_2^{\mathrm{ort}} = \sqrt{R_1^2 + R_2^2}\mathbf{v} \, ,$$

with $\mathbf{v} \sim \mathrm{Unif}(S^{n-1})$, and independently, R_1 and R_2 are the norms of $\mathbf{w}_1^{\mathrm{ort}}$ and $\mathbf{w}_2^{\mathrm{ort}}$ respectively (the norm of the sum is given by this form due to almost-sure orthogonality and Pythagoras' theorem). The covariance term can therefore be written

$$\frac{1}{m^2}\left(\mathbb{E}\left[\cos\left(\sqrt{R_1^2 + R_2^2}\langle \mathbf{v}, \mathbf{z}\rangle\right)\right] - \mathbb{E}\left[\cos\left(R_1\langle \mathbf{v}, \mathbf{z}\rangle\right)\right]^2\right) \, , \tag{18}$$

which completes the proof. □

9.1 Proof of Theorem 3.1

By Proposition 9.1, the statement of the theorem is equivalent to showing that the following term is negative:

$$\mathbb{E}\left[\cos\left(\sqrt{R_1^2 + R_2^2}\|\mathbf{z}\|\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle\right)\right] - \mathbb{E}\left[\cos(R_1\|\mathbf{z}\|\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle)\right]^2 \, . \tag{19}$$

We may regard this is a function $f : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}$ of $\|\mathbf{z}\|$, noting that the value of the expectations does not depend on $\widehat{\mathbf{z}}$, as it appears only in the inner product with the random unit vector $\mathbf{v}$, which has an isotropic distribution. We will write $z = \|\mathbf{z}\|$ for the argument of f in what follows for convenience:

$$f(z) = \mathbb{E}\left[\cos\left(\sqrt{R_1^2 + R_2^2}z\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle\right)\right] - \mathbb{E}\left[\cos(R_1 z\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle)\right]^2 \, , \qquad z \in \mathbb{R}_{\geq 0}$$

Observe trivially that $f(0) = 0$. We will show that f is decreasing in a neighbourhood around 0, from which the statement of the theorem immediately follows.

First observe that f is well-defined on all of $\mathbb{R}_{\geq 0}$, since the expectations are of bounded, measureable functions of random variables. A priori, it is not clear that f is differentiable, but we will see that by the dominated convergence theorem, if the random variable R has a finite k^{th} moment, for some $k \in \mathbb{N}$, then f is k times differentiable everywhere, and moreover, the k^{th} derivative is continuous. Specifically, recall the following corollary of the dominated convergence theorem:

Proposition 9.2. *Let μ be a probability measure, and let $g : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ be such that*

- *$x \mapsto g(t, x)$ is in $L^1(\mu)$ for all t*
- *$t \mapsto g(t, x)$ is differentiable for all x*
- *For some function $h \in L^1(\mu)$, we have*

$$\left|\frac{\partial g}{\partial t}(t, x)\right| \leq h(x) \qquad \forall t, x \in \mathbb{R}$$

Then

$$\frac{d}{dt}\mathbb{E}_{X\sim\mu}\left[g(t, X)\right] = \mathbb{E}\left[\frac{\partial g}{\partial t}(t, X)\right] \, .$$

By the assumption that R has a finite 4th moment, we have $R, R^2, R^3, R^4 \in L^1(\mu)$, and we may therefore use these as the dominating functions in Proposition 9.2 to establish fourth-order differentiability of the expectation $\mathbb{E}\left[\cos(R_1 z\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle)\right]$. We note further that since $\sqrt{R_1^2 + R_2^2} \leq R_1 + R_2$ almost-surely, we may use $(R_1 + R_2)^k$ for $k = 1, \ldots, 4$ as dominating functions in Proposition 9.2 to establish the fourth-order differentiability of the expectation $\mathbb{E}\left[\cos\left(\sqrt{R_1^2 + R_2^2} z\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle\right)\right]$. We therefore derive that f is 4 times differentiable, and we obtain the following (by Taylor's theorem with Lagrange's remainder):

$$f(h) = f(0) + hf'(0) + \frac{h^2}{2!} f^{(2)}(0) + \frac{h^3}{3!} f^{(3)}(0) + \frac{h^4}{4!} f^{(4)}(s)\,.$$

Direct computation leveraging Proposition 9.2 yields

$$\begin{aligned}
f^{(1)}(z) =& - \mathbb{E}\left[\sqrt{R_1^2 + R_2^2}\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle \sin\left(z\sqrt{R_1^2 + R_2^2}\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle\right)\right] + 2\mathbb{E}\left[\cos(zR_1\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle)\right]\mathbb{E}\left[R_1\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle \sin(zR_1\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle)\right] \\
f^{(2)}(z) =& - \mathbb{E}\left[(R_1^2 + R_2^2)\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle^2 \cos\left(z\sqrt{R_1^2 + R_2^2}\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle\right)\right] - 2\mathbb{E}\left[R_1\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle \sin(zR_1\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle)\right]^2 \\
&+ 2\mathbb{E}\left[\cos(zR_1\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle)\right]\mathbb{E}\left[R_1^2\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle^2 \cos(zR_1\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle)\right] \\
f^{(3)}(z) =& \mathbb{E}\left[(R_1^2 + R_2^2)^{3/2}\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle^3 \sin\left(z\sqrt{R_1^2 + R_2^2}\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle\right)\right] \\
&- 4\mathbb{E}\left[R_1\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle \sin(zR_1\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle)\right]\mathbb{E}\left[R_1^2\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle^2 \cos(zR_1\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle)\right] \\
&- 2\mathbb{E}\left[R_1\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle \sin(zR_1\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle)\right]\mathbb{E}\left[R_1^2\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle^2 \cos(zR_1\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle)\right] \\
&- 2\mathbb{E}\left[\cos(zR_1\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle)\right]\mathbb{E}\left[R_1^3\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle^3 \sin(zR_1\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle)\right] \\
f^{(4)}(z) =& \mathbb{E}\left[(R_1^2 + R_2^2)^2\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle^4 \cos\left(z\sqrt{R_1^2 + R_2^2}\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle\right)\right] \\
&- 6\mathbb{E}\left[R_1^2\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle^2 \cos(zR_1\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle)\right]^2 \\
&+ 6\mathbb{E}\left[R_1\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle \sin(zR_1\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle)\right]\mathbb{E}\left[R_1^3\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle^3 \sin(zR_1\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle)\right] \\
&+ 2\mathbb{E}\left[R_1\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle \sin(zR_1\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle)\right]\mathbb{E}\left[R_1^3\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle^3 \sin(zR_1\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle)\right] \\
&- 2\mathbb{E}\left[\cos(zR_1\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle)\right]\mathbb{E}\left[R_1^4\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle^4 \cos(zR_1\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle)\right]\ .
\end{aligned}$$

Directly substituing $z = 0$ into these expressions, we obtain

$$f(0) = f'(0) = f^{(2)}(0) = f^{(3)}(0) = 0\,, \qquad f^{(4)}(0) = \mathbb{E}\left[R_1^2\right]^2 \left(2\mathbb{E}\left[\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle^4\right] - 6\mathbb{E}\left[\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle^2\right]^2\right).$$

To establish the sign of $f^{(4)}(0)$, we compute the expectations $\mathbb{E}\left[\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle^4\right]$, $\mathbb{E}\left[\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle^2\right]$ directly. Firstly, note that $\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle$ can be written $\cos(\theta)$, where θ is the angle a uniformly random direction makes with a fixed direction in $\mathbb{R}^n$. by considering hyperspherical coordinates, the density of the angle on the interval $[0, \pi]$ is deduced to be

$$\frac{\sin^{n-2}(\theta)}{\int_0^\pi \sin^{n-2}(\theta')d\theta'}\,.$$

Therefore, we have

$$\begin{aligned}
\mathbb{E}\left[\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle^2\right] &= \frac{\int_0^\pi \cos^2(\theta)\sin^{n-2}(\theta)d\theta}{\int_0^\pi \sin^{n-2}(\theta)d\theta} = \frac{\sqrt{\pi}\frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n}{2})} - \sqrt{\pi}\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2}+1)}}{\sqrt{\pi}\frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n}{2})}} = \frac{1}{n} \\
\mathbb{E}\left[\langle \mathbf{v}, \widehat{\mathbf{z}}\rangle^4\right] = \frac{\int_0^\pi \cos^4(\theta)\sin^{n-2}(\theta)d\theta}{\int_0^\pi \sin^{n-2}(\theta)d\theta} &= \frac{\sqrt{\pi}\frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n}{2})} - 2\sqrt{\pi}\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2}+1)} + \sqrt{\pi}\frac{\Gamma(\frac{n+3}{2})}{\Gamma(\frac{n}{2}+2)}}{\sqrt{\pi}\frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n}{2})}} = \frac{3}{n(n+2)}\,,
\end{aligned}$$

which yields $f^{(4)}(0) = 6\mathbb{E}\left[R_1^2\right]^2 \left(\frac{1}{n(n+2)} - \frac{1}{n^2}\right) < 0$.

Finally, again by applying the dominated convergence theorem to each expectation in the expression above for $f^{(4)}(z)$, we obtain that this function is continuous. Hence, we have:

$$f(h) = \frac{h^4}{4!} f^{(4)}(s)\,,$$

for some $s \in (0, h)$, and by continuity of $f^{(4)}$, for sufficiently small h, the right-hand side above is negative, completing the proof.

9.2 Proof of Theorem 3.3

We prove the following theorem from which Theorem 3.3 follows. This result provides us also with the explicit gap between the MSEs from Theorem 3.1 if the tails of the corresponding distributions are not too heavy (for instance distributions for Gaussian or Poisson-Bessel kernels).

Definition 9.3. *Let $\eta_\mu(k)$ be defined as follows:*

$$\eta_\mu(k) = \max_{l_1+...+l_s=k} \prod_{j=1,...,s} \mathbb{E}[w_i^{2l_j}],$$

where $(w_1, ..., w_n)^\top \sim \mu$ and the maximum goes over all positive integer-sets $\{l_1, ..., l_s\}$ such that $l_1 + ... + l_s = k$.

Theorem 9.4. *Consider a family of radial basis function kernels K on $\mathbb{R}^n \times \mathbb{R}^n$ defined as $K(\mathbf{x}, \mathbf{y}) = \phi(\|\mathbf{x} - \mathbf{y}\|)$ for some fixed positive definite radial basis function ϕ that is not parametrized by data dimensionality n. Denote the associated probabilistic measures as $\{\mu_n\}$. Denote by $\widehat{K}^{\mathrm{iid}}_{m,n}$ its random feature map based estimator applying state-of-the-art independent sampling from μ with m samples (random features) and by $\widehat{K}^{\mathrm{ort}}_{m,n}$ its estimator based on the random orthogonal feature map mechanism. Assume that there exists some $c > 0$ such that $\mathrm{M}_{\mu_n}(2k, 2n) \leq \frac{2(n-1)(n+1)...(n+2k-3)k!}{c^k}$ or $\eta(k) \leq \frac{k!}{(2c)^k}$ for $k = 1, 2, ...$ Then the following holds for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ satisfying $\|\mathbf{z}\| = \|\mathbf{x} - \mathbf{y}\| < \sqrt{2c}$:*

$$\mathrm{MSE}(\widehat{K}^{\mathrm{iid}}_{m,n}(\mathbf{x}, \mathbf{y})) - \mathrm{MSE}(\widehat{K}^{\mathrm{ort}}_{m,n}(\mathbf{x}, \mathbf{y})) \geq \frac{m-1}{m} \tau_\mu(\|\mathbf{x} - \mathbf{y}\|, n, c), \tag{20}$$

where $\tau_\mu(\|\mathbf{x} - \mathbf{y}\|, n, c) = \|\mathbf{x} - \mathbf{y}\|^4 \frac{\mathrm{M}^2_{\mu_n}(2,n)}{2n^2(n+2)} - \frac{\|\mathbf{x}-\mathbf{y}\|^6}{8}(\frac{\mathrm{M}_{\mu_n}(2,n)\mathrm{M}_\mu(4,n)}{n^2(n+2)} + \frac{6}{c^3(1-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2c})})$. Now assume that there exist some $c, K(c) > 0$, $\xi : \mathbb{N} \rightarrow \mathbb{R}$ such that $\mathrm{M}_{\mu_n}(2k, 2n) \leq (n-1)(n+1) \cdot ... \cdot (n+2k-3)\xi(k)$ or $\eta(k) \leq \xi(k)$ and $|\xi(k)| \leq \frac{k!}{(2c)^k}$ for $k > K(c)$. Assume furthermore that the corresponding sequence of Fourier measures is concentrated. Then the following holds for $\|\mathbf{z}\| = \|\mathbf{x} - \mathbf{y}\| < \sqrt{\frac{c}{8}}$:

$$\mathrm{MSE}(\widehat{K}^{\mathrm{iid}}_{m,n}(\mathbf{x}, \mathbf{y})) - \mathrm{MSE}(\widehat{K}^{\mathrm{ort}}_{m,n}(\mathbf{x}, \mathbf{y})) = \frac{m-1}{m}(\frac{1}{8n}\Psi_K(\mathbf{z}) + \kappa(\|\mathbf{z}\|, n, c)), \tag{21}$$

where Ψ_K is defined as in Equation 4 and $|\kappa(\|\mathbf{z}\|, n, c)| \leq (\frac{1}{2n^2} + \frac{3}{n^3} + \frac{5g^2(n)}{2n} + \frac{1}{nh'(n)})\frac{2}{1-\frac{8}{c}\|\mathbf{z}\|^2} - \frac{6}{n^2}(\sum_{k=0}^{K(c)} \frac{\|\mathbf{z}\|^{2k}\xi(k)}{k!} + \frac{\|\mathbf{z}\|^2}{c-\|\mathbf{z}\|^2})$ for some functions $g(n), h'(n)$ such that: $g(n) = o_n(1)$ and $h'(n) = \omega_n(1)$.

Proof. As before, we will use Proposition 9.1. We start by proving the first part of the statement.

Part I: Proof of Inequality 20 We use the expression for the difference in MSEs derived in Proposition 9.1. Denote $\widehat{\mathbf{z}} = \frac{\mathbf{z}}{\|\mathbf{z}\|_2}$. Using the series expansion of $\cos(x)$, we get:

$$\mathbb{E}\left[\cos(\sqrt{R_1^2 + R_2^2}\mathbf{v}^\top \mathbf{z})\right] = \mathbb{E}\left[\sum_{k=0}^{\infty} \frac{\|\mathbf{z}\|^{2k}(-1)^k(R_1^2 + R_2^2)^k \langle \mathbf{v}, \widehat{\mathbf{z}} \rangle^{2k}}{(2k)!}\right]. \tag{22}$$

Similarly,

$$\mathbb{E}\left[\cos(R_i \mathbf{v}_i^\top \mathbf{z})\right] = \mathbb{E}\left[\sum_{k=0}^{\infty} \frac{\|\mathbf{z}\|^{2k}(-1)^k(R_i)^{2k} \langle \mathbf{v}, \widehat{\mathbf{z}} \rangle^{2k}}{(2k)!}\right] \tag{23}$$

for $i = 1, 2$. Denote: $A(k, n) = \mathbb{E}[\langle \mathbf{v}, \widehat{\mathbf{z}} \rangle^k]$, where $v \sim \mathrm{Unif}(S^{n-1})$ and $S(k, m) = \mathrm{M}_\mu(2k, m) = \mathbb{E}[(\sum_{i=1}^m w_i^2)^k]$, where $(w_1, ..., w_m)^\top \sim \mu_m$.

Let us start by calculating $A(2k, n)$.

Lemma 9.5. *The following is true:*

$$A(2k,n) = \frac{(n-2)(n-4)\cdot ... \cdot \delta(n=2)}{(n-3)(n-5)\cdot ... \cdot \gamma(n=2)} \cdot \frac{(2k-1)!!}{(n-1)(n+1)...(n+2k-3)} \cdot \frac{(n+2k-3)(n+2k-5)... \cdot \gamma(n=2)}{(n+2k-2)(n+2k-4)... \cdot \delta(n=2)}, \tag{24}$$

where: $\delta(n=2)=2$ *if* $n=2$ *and* $\delta(n=2)=1$ *otherwise and:* $\gamma(n=2)=1$ *if* $n=2$ *and* $\gamma(n=2)=2$ *otherwise. In particular, the following is true:*

$$|A(2k,n)| \leq \frac{(2k-1)!!}{(n-1)(n+1)\cdot ... \cdot (n+2k-3)}.$$

Proof. Note that:

$$A(k,n) = \frac{1}{\int_0^\pi \sin^{n-2}(\theta)d\theta}\int_0^\pi \cos^k(\theta)\sin^{n-2}(\theta)d\theta, \tag{25}$$

where the formula comes from the well known fact that the density function $p_n(\theta)$ of the angle θ between a vector $\mathbf{r} \in \mathbb{R}^n$ chosen uniformly at random from the unit sphere and some fixed vector $\mathbf{q} \in \mathbb{R}^n$ is of the form:

$$p_n(\theta) = \frac{1}{\int_0^\pi \sin^{n-2}(\theta)d\theta}\sin^{n-2}(\theta). \tag{26}$$

Denote $F(k,n) = \int_0^\pi \cos^k(\theta)\sin^n(\theta)d\theta$. We have:

$$\int_0^\pi \cos^k(\theta)\sin^n(\theta)d\theta = \int_0^\pi \cos^{k-1}(\theta)\sin^n(\theta)(\sin(\theta))'d\theta = \\ \cos^{k-1}(\theta)\sin^{n+1}(\theta)|_0^\pi - \int_0^\pi \sin(\theta)((k-1)\cos^{k-2}(\theta)(-\sin(\theta))\sin^n(\theta) + n\cos^k(\theta)\sin^{n-1}(\theta))d\theta \tag{27}$$

That leads us to the recursive formula on $F(k,n)$ which is:

$$F(k,n) = \frac{k-1}{n+1}F(k-2,n+2). \tag{28}$$

Thus we get:

$$F(n,2k) = \frac{(2k-1)!!}{(n+1)(n+3)\cdot ... \cdot (n+2k-1)}\int_0^\pi \sin^{n+2k}(\theta)d\theta. \tag{29}$$

Again, by partial differentiation formula we get:

$$\int_0^\pi \sin^n(x)dx = -\frac{1}{n}\sin^{n-1}(x)\cos(x)|_0^\pi + \frac{n-1}{n}\int_0^\pi \sin^{n-2}(x)dx = \frac{n-1}{n}\int_0^\pi \sin^{n-2}(x)dx. \tag{30}$$

Thus we get:

$$A(2k,n) = \frac{1}{\frac{n-3}{n-2}\cdot\frac{n-5}{n-4}\cdot ...} \cdot \frac{(2k-1)!!}{(n-1)(n+1)\cdot ... \cdot (n+2k-3)} \cdot \frac{n+2k-3}{n+2k-2} \cdot \frac{n+2k-5}{n+2k-4} \cdot ... \tag{31}$$

That, after simplification, proves the first part of the statement. The second part is implied by the above formula on $A(2k,n)$. □

The following is true:

$$\mathbb{E}[\cos(\sqrt{R_1^2+R_2^2}\mathbf{v}^\top\mathbf{z})] = 1 - \frac{\|\mathbf{z}\|^2}{2!}S(1,2n)A(2,n) + \frac{\|\mathbf{z}\|^4}{4!}S(2,2n)A(4,n) + \\ \mathbb{E}[\sum_{k=3}^{\infty}\frac{\|\mathbf{z}\|^{2k}(-1)^k(R_1^2+R_2^2)^k\langle\mathbf{v},\widehat{\mathbf{z}}\rangle^{2k}}{(2k)!}]. \tag{32}$$

Similarly, we have

$$\begin{aligned}\mathbb{E}[\cos(R_1\mathbf{v}_1^\top\mathbf{z})]\mathbb{E}[\cos(R_2\mathbf{v}_2^\top\mathbf{z})] = \\ (1-\frac{\|\mathbf{z}\|^2}{2!}S(1,n)A(2,n)+\frac{\|\mathbf{z}\|^4}{4!}S(2,n)A(4,n)+\mathbb{E}[\sum_{k=3}^{\infty}\frac{\|\mathbf{z}\|^{2k}(-1)^k(R_1)^{2k}\langle\mathbf{v},\widehat{\mathbf{z}}\rangle^{2k}}{(2k)!}])\times \\ (1-\frac{\|\mathbf{z}\|^2}{2!}S(1,n)A(2,n)+\frac{\|\mathbf{z}\|^4}{4!}S(2,n)A(4,n)+\mathbb{E}[\sum_{k=3}^{\infty}\frac{\|\mathbf{z}\|^{2k}(-1)^k(R_2)^{2k}\langle\mathbf{v},\widehat{\mathbf{z}}\rangle^{2k}}{(2k)!}]).\end{aligned} \tag{33}$$

Therefore we have:

$$\begin{aligned}\mathbb{E}[\cos(R_1\mathbf{v}_1^\top\mathbf{z})]\mathbb{E}[\cos(R_2\mathbf{v}_2^\top\mathbf{z})] = \\ 1-\frac{\|\mathbf{z}\|^2}{2!}2S(1,n)A(2,n)+\frac{\|\mathbf{z}\|^4}{4!}(2S(2,n)A(4,n)+6S^2(1,n)A^2(2,n))+ \\ 2\mathbb{E}[\sum_{k=3}^{\infty}\frac{\|\mathbf{z}\|^{2k}(-1)^k(R_1)^{2k}\langle\mathbf{v},\widehat{\mathbf{z}}\rangle^{2k}}{(2k)!}]\mathbb{E}[\cos(R_1\mathbf{v}_1^\top\mathbf{z})]-\frac{\|\mathbf{z}\|^6}{24}S(1,n)S(2,n)A(2,n)A(4,n)+ \\ \frac{\|\mathbf{z}\|^8}{24^2}S^2(2,n)A^2(4,n).\end{aligned} \tag{34}$$

Let us focus now on the infinite sums that appear in the expressions above. For any given $N \geq 3$, we have (by the dominated convergence theorem):

$$\begin{aligned}|\mathbb{E}[\sum_{k=3}^{N}\frac{\|\mathbf{z}\|^{2k}(-1)^k(R_i)^{2k}\langle\mathbf{v}_i,\widehat{\mathbf{z}}\rangle^{2k}}{(2k)!}]| \leq \sum_{k=3}^{N}\frac{\|\mathbf{z}\|^{2k}\mathbb{E}[(R_i)^{2k}]\mathbb{E}[\langle\mathbf{v}_i,\widehat{\mathbf{z}}\rangle^{2k}]}{(2k)!} \\ \leq \sum_{k=3}^{N}\frac{\|\mathbf{z}\|^{2k}\mathbb{E}[(R_1^2+R_2^2)^k]\mathbb{E}[\langle\mathbf{v},\widehat{\mathbf{z}}\rangle^{2k}]}{(2k)!} \leq \sum_{k=3}^{N}\|\mathbf{z}\|^{2k}\frac{S(k,2n)A(2k,n)}{(2k)!} \\ = \sum_{k=3}^{N}\|\mathbf{z}\|^{2k}\frac{\mathrm{M}_\mu(2k,2n)A(2k,n)}{(2k)!}\end{aligned} \tag{35}$$

for $i = 1, 2$, where we use the fact that random variables R_i and $\langle\mathbf{v}_i,\widehat{\mathbf{z}}\rangle^{2k}$ are independent.

Now note that: $(2k-1)!! \cdot 2 \cdot 4.. \cdot 2k = (2k)!$. Thus we have: $(2k-1)!! = \frac{(2k)!}{2^k k!}$. Therefore, from the previously derived bound on $|A(2k,n)|$, we get:

$$\frac{\mathrm{M}_\mu(2k,2n)A(2k,n)}{(2k)!} \leq \frac{\mathrm{M}_\mu(2k,2n)}{2^k k!(n-1)(n+1)...(n+2k-3)}. \tag{36}$$

If the first condition on $\mathrm{M}_\mu(2k,2n)$ from the statement of the theorem is satisfied, then we get:

$$\frac{\mathrm{M}_\mu(2k,2n)A(2k,n)}{(2k)!} \leq \frac{2}{(2c)^k}. \tag{37}$$

Thus in this scenario we obtain

$$\begin{aligned}|\mathbb{E}[\sum_{k=3}^{N}\frac{\|\mathbf{z}\|^{2k}(-1)^k(R_i)^{2k}\langle\mathbf{v}_i,\widehat{\mathbf{z}}\rangle^{2k}}{(2k)!}]| \leq \sum_{k=3}^{N}\frac{\|\mathbf{z}\|^{2k}\mathbb{E}[(R_i)^{2k}\langle\mathbf{v}_i,\widehat{\mathbf{z}}\rangle^{2k}]}{(2k)!} \\ \leq \sum_{k=3}^{N}\frac{\|\mathbf{z}\|^{2k}\mathbb{E}[(R_1^2+R_2^2)^k]\mathbb{E}[\langle\mathbf{v},\widehat{\mathbf{z}}\rangle^{2k}]}{(2k)!} \leq 2\sum_{k=3}^{N}\|\mathbf{z}\|^{2k}(\frac{1}{2c})^k = \frac{2\|\mathbf{z}\|^6}{(2c)^3}\sum_{k=0}^{N-3}(\frac{\|\mathbf{z}\|^2}{2c})^k\end{aligned} \tag{38}$$

for $i = 1, 2$. In particular, we conclude that for $\|\mathbf{z}\| < \sqrt{2c}$ all the infinite sums introduced above all well-defined and we have:

$$|\mathbb{E}[\sum_{k=3}^{N}\frac{\|\mathbf{z}\|^{2k}(-1)^k(R_i)^{2k}\langle\mathbf{v}_i,\widehat{\mathbf{z}}\rangle^{2k}}{(2k)!}]|, |\mathbb{E}[\sum_{k=3}^{N}\frac{\|\mathbf{z}\|^{2k}(-1)^k(R_1^2+R_2^2)^k\langle\mathbf{v},\widehat{\mathbf{z}}\rangle^{2k}}{(2k)!}]| \leq \frac{2\|\mathbf{z}\|^6}{(2c)^3}\frac{1}{1-\frac{\|\mathbf{z}\|^2}{2c}} \tag{39}$$

for $i = 1, 2$.

Notice that we also have the following trivial upper bound on $S(k, 2n)$:

$$S(k, 2n) \leq (2n)^k \eta(k). \tag{40}$$

Thus we get:

$$\frac{\mathrm{M}_{\mu}(2k, 2n)A(2k, n)}{(2k)!} \leq \frac{(2n)^k \eta(k)}{2^k k!(n-1)(n+1)...(n+2k-3)} \leq \frac{2\eta(k)}{k!}. \tag{41}$$

Thus if the condition on $\eta(k)$ is satisfied, we get:

$$\frac{\mathrm{M}_{\mu}(2k, 2n)A(2k, n)}{(2k)!} \leq \frac{2}{(2c)^k}. \tag{42}$$

Thus, as in the previous case, we get inequalities (39) for $\|\mathbf{z}\| \leq \sqrt{2c}$.

Thus we conclude that for $\|\mathbf{z}\| \leq \sqrt{2c}$ we have:

$$\begin{aligned} \rho \geq 1 - \frac{\|\mathbf{z}\|^2}{2!} 2S(1,n)A(2,n) + \frac{\|\mathbf{z}\|^4}{4!}(2S(2,n)A(4,n) + 6S^2(1,n)A^2(2,n)) + \\ 2\mathbb{E}[\sum_{k=3}^{\infty} \frac{\|\mathbf{z}\|^{2k}(-1)^k (R_1)^{2k} \langle \mathbf{v}, \widehat{\mathbf{z}} \rangle^{2k}}{(2k)!}]\mathbb{E}[\cos(R_1 \mathbf{v}_1^{\top} \mathbf{z})] - \frac{\|\mathbf{z}\|^6}{24} S(1,n)S(2,n)A(2,n)A(4,n) + \\ \frac{\|\mathbf{z}\|^8}{24^2} S^2(2,n)A^2(4,n) - (1 - \frac{\|\mathbf{z}\|^2}{2!} S(1,2n)A(2,n) + \frac{\|\mathbf{z}\|^4}{4!} S(2,2n)A(4,n) + \\ \mathbb{E}[\sum_{k=3}^{\infty} \frac{\|\mathbf{z}\|^{2k}(-1)^k (R_1^2 + R_2^2)^k \langle \mathbf{v}, \widehat{\mathbf{z}} \rangle^{2k}}{(2k)!}]) \end{aligned} \tag{43}$$

Notice that we have: $S(1, 2n) = 2S(1, n)$. Thus we get:

$$\begin{aligned} \rho \geq \frac{\|\mathbf{z}\|^4}{4!}(2S(2,n)A(4,n) + 6S^2(1,n)A^2(2,n) - S(2,2n)A(4,n)) - \\ \frac{\|\mathbf{z}\|^6}{24} S(1,n)S(2,n)A(2,n)A(4,n) + \frac{\|\mathbf{z}\|^8}{24^2} S^2(2,n)A^2(4,n) + \\ 2\mathbb{E}[\sum_{k=3}^{\infty} \frac{\|\mathbf{z}\|^{2k}(-1)^k (R_1)^{2k} \langle \mathbf{v}, \widehat{\mathbf{z}} \rangle^{2k}}{(2k)!}]\mathbb{E}[\cos(R_1 \mathbf{v}_1^{\top} \mathbf{z})] - \mathbb{E}[\sum_{k=3}^{\infty} \frac{\|\mathbf{z}\|^{2k}(-1)^k (R_1^2 + R_2^2)^k \langle \mathbf{v}, \widehat{\mathbf{z}} \rangle^{2k}}{(2k)!}]. \end{aligned} \tag{44}$$

Since $S(2, 2n) = 2S(2, n) + 2S^2(1, n)$, we obtain:

$$\begin{aligned} \rho \geq \frac{\|\mathbf{z}\|^4}{12} S^2(1,n)(3A^2(2,n) - A(4,n)) - \frac{\|\mathbf{z}\|^6}{24} S(1,n)S(2,n)A(2,n)A(4,n) \\ + \frac{\|\mathbf{z}\|^8}{24^2} S^2(2,n)A^2(4,n) + 2\mathbb{E}[\sum_{k=3}^{\infty} \frac{\|\mathbf{z}\|^{2k}(-1)^k (R_1)^{2k} \langle \mathbf{v}, \widehat{\mathbf{z}} \rangle^{2k}}{(2k)!}]\mathbb{E}[\cos(R_1 \mathbf{v}_1^{\top} \mathbf{z})] \\ - \mathbb{E}[\sum_{k=3}^{\infty} \frac{\|\mathbf{z}\|^{2k}(-1)^k (R_1^2 + R_2^2)^k \langle \mathbf{v}, \widehat{\mathbf{z}} \rangle^{2k}}{(2k)!}]. \end{aligned} \tag{45}$$

Now we can use derived earlier upper bounds on the absolute values of the infinite sums under considerations and we obtain:

$$\begin{aligned} \rho \geq \frac{\|\mathbf{z}\|^4}{12} S^2(1,n)(3A^2(2,n) - A(4,n)) - \frac{\|\mathbf{z}\|^6}{24} S(1,n)S(2,n)A(2,n)A(4,n) \\ + \frac{\|\mathbf{z}\|^8}{24^2} S^2(2,n)A^2(4,n) - \frac{6\|\mathbf{z}\|^6}{(2c)^3} \frac{1}{1 - \frac{\|\mathbf{z}\|^2}{2c}} \geq \\ \frac{\|\mathbf{z}\|^4}{12} S^2(1,n)(\frac{3}{n^2} - \frac{3}{n(n+2)}) - \frac{\|\mathbf{z}\|^6}{24} \frac{3\mathrm{M}_{\mu}(2,n)\mathrm{M}_{\mu}(4,n)}{n^2(n+2)} - \frac{6\|\mathbf{z}\|^6}{(2c)^3} \frac{1}{1 - \frac{\|\mathbf{z}\|^2}{2c}} \geq \\ \|\mathbf{z}\|^4 \frac{\mathrm{M}_{\mu}^2(2,n)}{2n^2(n+2)} - \frac{\|\mathbf{z}\|^6}{8}(\frac{\mathrm{M}_{\mu}(2,n)\mathrm{M}_{\mu}(4,n)}{n^2(n+2)} + \frac{6}{c^3(1 - \frac{\|\mathbf{z}\|^2}{2c})}). \end{aligned} \tag{46}$$

Applying Proposition 9.1, we complete the proof of the first part of the theorem.

Part II: Proof of Equality 21

We will borrow notation and several observations from the first part of the proof. Note first that

$$\begin{aligned}\mathbb{E}[\cos(R_1\mathbf{v}^\top\mathbf{z})]\mathbb{E}[\cos(R_2\mathbf{v}^\top\mathbf{z})] = \mathbb{E}[\cos(R_1\mathbf{v}_1^\top\mathbf{z})]\mathbb{E}[\cos(R_2\mathbf{v}_2^\top\mathbf{z})] = \\ \mathbb{E}[\cos(R_1\mathbf{v}_1^\top\mathbf{z})\cos(R_2\mathbf{v}_2^\top\mathbf{z})] = \mathbb{E}[\frac{1}{2}\cos((R_1\mathbf{v}_1+R_2\mathbf{v}_2)^\top z)+\frac{1}{2}\cos((R_1\mathbf{v}_1-R_2\mathbf{v}_2)^\top\mathbf{z})] = \\ \mathbb{E}[\cos((R_1\mathbf{v}_1+R_2\mathbf{v}_2)^\top\mathbf{z})],\end{aligned} \tag{47}$$

where $\mathbf{v}_1,\mathbf{v}_2 \sim \mathbf{v}$, $\mathbf{v}_1,\mathbf{v}_2$ are independent and the last equality follows from the fact that $R_1\mathbf{v}_1+R_2\mathbf{v}_2 \sim R_1\mathbf{v}_1-R_2\mathbf{v}_2$. Note that $R_1\mathbf{v}_1+R_2\mathbf{v}_2 \sim \widehat{R}\mathbf{v}$ for some $\widehat{R}$.

Thus the negative covariance term ρ is of the form

$$\rho = \sum_{k=0}^{\infty}\frac{\|\mathbf{z}\|^{2k}(-1)^k\mathbb{E}[(\mathbf{v}^\top\widehat{\mathbf{z}})^{2k}]}{(2k)!}\mathbb{E}[\widehat{R}^{2k}] - \sum_{k=0}^{\infty}\frac{\|\mathbf{z}\|^{2k}(-1)^k\mathbb{E}[(\mathbf{v}^\top\widehat{\mathbf{z}})^{2k}]}{(2k)!}\mathbb{E}[R^{2k}], \tag{48}$$

where $R^2 = R_1^2 + R_2^2$. We can replace the expectations of the sum with the sum of expectations in the expressions above by the assumptions of the theorem, the convergence analysis given by us in the previous part and a simple observation that:

$$\widehat{R}^2 = R_1^2 + R_2^2 + 2R_1R_2\cos(\mathbf{v}_1,\mathbf{v}_2) \leq R_1^2 + R_2^2 + 2R_1R_2 \leq 2(R_1^2+R_2^2) = 2R^2. \tag{49}$$

Thus the upper bounds on the absolute values on the infinite sums considered now contain one additional multiplicative factor of $2^{2k} = 4^k$ in comparison to these that were considered in the first part of the proof. That does not affect the convergence if $\|\mathbf{z}\| < \sqrt{\frac{2c}{4}} = \sqrt{\frac{c}{2}}$.

Thus the negative covariance term is of the form

$$\rho = \sum_{k=0}^{\infty}\frac{\|\mathbf{z}\|^{2k}(-1)^kA(2k,n)}{(2k)!}(\mathbb{E}[\widehat{R}^{2k}] - \mathbb{E}[R^{2k}]). \tag{50}$$

Note that $\widehat{R}^2 = R^2 + 2R_1R_2\langle\mathbf{v}_1,\mathbf{v}_2\rangle$ and that $\langle\mathbf{v}_1,\mathbf{v}_2\rangle \overset{d}{=} \langle\mathbf{v},\widehat{\mathbf{z}}\rangle$. Thus we get:

$$\rho = \sum_{k=0}^{\infty}\frac{\|\mathbf{z}\|^{2k}(-1)^kA(2k,n)}{(2k)!}\alpha_k, \tag{51}$$

where

$$\alpha_k = \sum_{i=1}^{k}\binom{k}{i}\mathbb{E}[(R^2)^{k-i}(2R_1R_2)^i]A(i,n). \tag{52}$$

Note that since $A(2,n) = \frac{1}{n}$ and furthermore $A(i,n) = 0$ for odd i, we get:

$$\alpha_k = \frac{k(k-1)}{2n}\mathbb{E}[(R^2)^{k-2}(2R_1R_2)^2] + \beta_k, \tag{53}$$

where:

$$\beta_k = \sum_{j=2}^{\lfloor\frac{k}{2}\rfloor}\binom{k}{2j}\mathbb{E}[(R^2)^{k-2j}(2R_1R_2)^{2j}]A(2j,n). \tag{54}$$

Now note that $A(2j,n)$ is a non-increasing function of j (from the definition of $A(k,n)$) and furthermore $\mathbb{E}[(R^2)^{k-2j}(2R_1R_2)^{2j}] \leq \mathbb{E}[(R^2)^{k-2j}(R^2)^j] = \mathbb{E}[(R^2)^k]$.

Thus we get:

$$0 \leq \beta_k \leq 2^kA(4,n)\mathbb{E}[(R^2)^k]. \tag{55}$$

Since $A(4,n)=\frac{3}{n(n+2)}$, we get:

$$0 \leq \beta_k \leq \frac{3 \cdot 2^k}{n^2}\mathbb{E}[R^{2k}]. \tag{56}$$

Thus we get:

$$\mathbb{E}[\widehat{R}^{2k}] - \mathbb{E}[R^{2k}] = \frac{k(k-1)}{2n}\mathbb{E}[(R^2)^{k-2}(2R_1R_2)^2] + \beta_k. \tag{57}$$

Therefore we obtain:

$$\begin{gathered}\rho = \sum_{k=0}^{\infty} \frac{\|\mathbf{z}\|^{2k}(-1)^k A(n,2k)}{(2k)!}\frac{k(k-1)}{2n}\mathbb{E}[(R^2)^{k-2}(2R_1R_2)^2] + \frac{3}{n^2}\Lambda(\|\mathbf{z}\|) = \\ \frac{1}{2n}\sum_{k=0}^{\infty} \frac{\|\mathbf{z}\|^{2k}(-1)^k A(n,2k)}{(2k)!}k(k-1)\mathbb{E}[(R^2)^{k-2}(2R_1R_2)^2] + \frac{3}{n^2}\Lambda(\|\mathbf{z}\|).\end{gathered} \tag{58}$$

where $\Lambda(\|\mathbf{z}\|)$ satisfies

$$|\Lambda(\|\mathbf{z}\|)| \leq \sum_{k=0}^{\infty} \frac{\|\mathbf{z}\|^{2k} A(n,2k)}{(2k)!}2^k\mathbb{E}[R^{2k}] \tag{59}$$

Therefore by the similar analysis as before, and from the definition of $K(c)$, we have:

$$|\Lambda(\|\mathbf{z}\|)| \leq 2\sum_{k=0}^{K(c)} \frac{\|\mathbf{z}\|^{2k}\xi(k)}{k!} + 2\sum_{k=K(c)+1}^{\infty} \|\mathbf{z}\|^{2k}(\frac{1}{2c})^k 2^k. \tag{60}$$

As before, we obtain this inequality if the condition regarding $\mathrm{M}_{\mu}(2k,2n)$ or the inequality regarding $\eta(k)$ is satisfied. We conclude that the following is true:

$$|\Lambda(\|\mathbf{z}\|)| \leq 2\sum_{k=0}^{K(c)} \frac{\|\mathbf{z}\|^{2k}\xi(k)}{k!} + \frac{2\|\mathbf{z}\|^2}{c-\|\mathbf{z}\|^2}. \tag{61}$$

Denote

$$\lambda_k = \frac{\mathbb{E}[R^{2k-4}(2R_1R_2)^2]}{\mathbb{E}[\widehat{R}^{2k}]} - 1. \tag{62}$$

Note that

$$\rho = \frac{1}{2n}\sum_{k=0}^{\infty} \frac{\|\mathbf{z}\|^{2k}(-1)^k A(2k,n)}{(2k)!}k(k-1)\mathbb{E}[\widehat{R}^{2k}] + A + \frac{3}{n^2}\Lambda(\|\mathbf{z}\|), \tag{63}$$

where

$$A = \frac{1}{2n}\sum_{k=0}^{\infty} \frac{\|\mathbf{z}\|^{2k}(-1)^k A(2k,n)}{(2k)!}k(k-1)\mathbb{E}[\widehat{R}^{2k}]\lambda_k. \tag{64}$$

We have:

$$\lambda_k = \frac{\mathbb{E}[R^{2k-4}(2R_1R_2)^2]}{\mathbb{E}[R^{2k}]}\frac{\mathbb{E}[R^{2k}]}{\mathbb{E}[\widehat{R}^{2k}]} - 1. \tag{65}$$

Note first that

$$|\frac{\mathbb{E}[\widehat{R}^{2k}]}{\mathbb{E}[R^{2k}]} - 1| = |\frac{\mathbb{E}[\widehat{R}^{2k}] - \mathbb{E}[R^{2k}]}{\mathbb{E}[R^{2k}]}| \leq \frac{k(k-1)}{2n} + \frac{3 \cdot 2^k}{n^2} \tag{66}$$

where the last inequality comes from Equation 57 and the inequality $R^2 = R_1^2 + R_2^2 \geq 2R_1R_2$. Now let us consider expression $\frac{\mathbb{E}[R^{2k-4}(2R_1R_2)^2]}{\mathbb{E}[R^{2k}]}$. Notice first that:

$$\frac{\mathbb{E}[R^{2k-4}(2R_1R_2)^2]}{\mathbb{E}[R^{2k}]} \leq 1, \tag{67}$$

where the inequality comes from the fact that $R_1^2 + R_2^2 \geq 2R_1R_2$.

We know that there exist functions $g(n) = o_n(1)$ and $h(n) = \omega_n(1)$ such that

$$\mathbb{P}[|\|\mathbf{w}\|_2^2 - \mathrm{M}_\mu(2,n)| \geq \mathrm{M}_\mu(2,n)g(n)] \leq \frac{1}{h(n)}, \quad \mathbf{w} \sim \mu \in \mathcal{M}(\mathbb{R}^n)$$

With this concentration bound in hand, we now denote by Ω_n the event that: $|\|\mathbf{w}_i^{\mathrm{ort}}\|_2^2 - \mathrm{M}_\mu(2,n)| \leq \mathrm{M}_\mu(2,n)g(n)$ for $i=1$ and $i=2$. Note that $\mathbb{P}[\Omega_n] \geq 1 - \frac{2}{h(n)}$. Denote $\Omega = \mathbb{R}^n \times \mathbb{R}^n$. We have:

$$\begin{gathered}\mathbb{E}[\frac{R^{2k-4}(2R_1R_2)^2}{\mathbb{E}[R^{2k}]}] = \int_\Omega \frac{R^{2k-4}(2R_1R_2)^2}{\mathbb{E}[R^{2k}]}\mu(d(R_1,R_2)) = \\ \int_{\Omega_n} \frac{R^{2k-4}(2R_1R_2)^2}{\mathbb{E}[R^{2k}]}\mu(d(R_1,R_2)) + \int_{\Omega \backslash \Omega_n} \frac{R^{2k-4}(2R_1R_2)^2}{\mathbb{E}[R^{2k}]}\mu(d(R_1,R_2))\end{gathered} \tag{68}$$

Our first observation is that

$$\int_{\Omega \backslash \Omega_n} \frac{R^{2k-4}(2R_1R_2)^2}{\mathbb{E}[R^{2k}]}\mu(d(R_1,R_2)) \leq \frac{2}{h'(n)}. \tag{69}$$

for some function $h'(n) = \omega_n(1)$. This comes from the form of the function we are integrating (note that its expected value is bounded by 1) and from the fact that a measure of the set over which we are integrating is at most $\frac{2}{h(n)}$ for $h(n) = \omega_n(1)$. Without loss of generality we can assume that $h(n) \geq 4$. Now let us focus on

$$\int_{\Omega_n} \frac{R^{2k-4}(2R_1R_2)^2}{\mathbb{E}[R^{2k}]}\mu(d(R_1,R_2)). \tag{70}$$

Now note than on Ω_n we have: $R_i^2 \in [\mathrm{M}_\mu(2,n)(1-g(n)), \mathrm{M}_\mu(2,n)(1+g(n))]$ for $i = 1,2$. Thus we obtain

$$\frac{|R^4 - (2R_1R_2)^2|}{R_1^4} = \frac{|(R_1^2+R_2^2)^2 - 4R_1^2R_2^2|}{R_1^4} = |(1 - \frac{R_2^2}{R_1^2})^2| \leq (1 - \frac{1+g(n)}{1-g(n)})^2 = \frac{4g^2(n)}{(1-g(n))^2} \tag{71}$$

Therefore we get:

$$|R^4 - (2R_1R_2)^2| \leq R_1^4 \cdot \frac{4g^2(n)}{(1-g(n))^2}. \tag{72}$$

Therefore we get:

$$\begin{aligned}\int_{\Omega_n} \frac{R^{2k-4}(2R_1R_2)^2}{\mathbb{E}[R^{2k}]}\mu(d(R_1,R_2)) \geq \int_{\Omega_n} \frac{R^{2k}}{\mathbb{E}[R^{2k}]}\mu(d(R_1,R_2)) - T(n)\int_{\Omega_n} \frac{R^{2k-4}(R_1)^4}{\mathbb{E}[R^{2k}]}\mu(d(R_1,R_2)) \\ \geq \int_{\Omega_n} \frac{R^{2k}}{\mathbb{E}[R^{2k}]}\mu(d(R_1,R_2)) - T(n),\end{aligned} \tag{73}$$

where $T(n) = \frac{4g^2(n)}{(1-g(n))^2}$ and the last inequality comes from the fact that $R_1^2 \leq R^2$. Therefore we obtain:

$$\int_{\Omega_n} \frac{R^{2k-4}(2R_1R_2)^2}{\mathbb{E}[R^{2k}]}\mu(d(R_1,R_2)) \geq 1 - \int_{\Omega \backslash \Omega_n} \frac{R^{2k}}{\mathbb{E}[R^{2k}]}\mu(d(R_1,R_2)) - T(n). \tag{74}$$

Now note that

$$\int_{\Omega \backslash \Omega_n} \frac{R^{2k}}{\mathbb{E}[R^{2k}]}\mu(d(R_1,R_2)) \leq \frac{2}{h'(n)}. \tag{75}$$

for some $h'(n) = \omega_n(1)$. The analysis is the same as before. As before, without loss of generality we can assume that $h(n) \geq 4$. Therefore, by combining Equality 68, Inequality 74 and Inequality 75, we get:

$$1 - \frac{4g^2(n)}{(1-g^2(n))} - \frac{2}{h(n)} \leq \mathbb{E}[\frac{R^{2k-4}(2R_1R_2)^2}{\mathbb{E}[R^{2k}]}] \leq 1 \tag{76}$$

Since $g(n) = o_n(1)$, without loss of generality we can assume that $g(n) \leq \frac{1}{3}$. Therefore we obtain the following bound on $|\lambda_k|$:

$$\begin{aligned}|\lambda_k| \leq (\frac{k(k-1)}{2n} + \frac{3 \cdot 2^k}{n^2}) + (\frac{4g^2(n)}{1-g^2(n)} + \frac{2}{h'(n)}) + \\ (\frac{k(k-1)}{2n} + \frac{3 \cdot 2^k}{n^2}) \cdot (\frac{4g^2(n)}{1-g^2(n)} + \frac{2}{h'(n)})\end{aligned} \tag{77}$$

From the assumptions on $g(n)$ and $h'(n)$ we get: $0 \leq \frac{4g^2(n)}{1-g^2(n)} + \frac{2}{h'(n)} \leq 1$. We obtain:

$$|\lambda_k| \leq \frac{k(k-1)}{n} + \frac{6 \cdot 2^k}{n^2} + 5g^2(n) + \frac{2}{h'(n)}. \tag{78}$$

Now, we are ready to find an upper bound on the term $|A|$ from Equation 63. We have:

$$\begin{aligned}|A| \leq \frac{1}{2n^2}\sum_{k=0}^{\infty}\frac{\|\mathbf{z}\|^{2k}A(2k,n)}{(2k)!}k^2(k-1)^2\mathbb{E}[\widehat{R}^{2k}] + \frac{6}{2n^3}\sum_{k=0}^{\infty}\frac{\|\mathbf{z}\|^{2k}A(2k,n)}{(2k)!}k(k-1)2^k\mathbb{E}[\widehat{R}^{2k}] \\ +(5g^2(n) + \frac{2}{h'(n)})\frac{1}{2n}\sum_{k=0}^{\infty}\frac{\|\mathbf{z}\|^{2k}A(2k,n)}{(2k)!}k(k-1)\mathbb{E}[\widehat{R}^{2k}]\end{aligned} \tag{79}$$

Now, by the same analysis as before, from the assumptions of the theorem and from the observation that $2^k \geq k(k-1)$, we obtain for $\|\mathbf{z}\| < \sqrt{\frac{c}{8}}$:

$$\begin{aligned}\sum_{k=0}^{\infty}\frac{\|\mathbf{z}\|^{2k}A(2k,n)}{(2k)!}k^2(k-1)^2\mathbb{E}[\widehat{R}^{2k}], \sum_{k=0}^{\infty}\frac{\|\mathbf{z}\|^{2k}A(2k,n)}{(2k)!}k(k-1)2^k\mathbb{E}[\widehat{R}^{2k}], \\ \sum_{k=0}^{\infty}\frac{\|\mathbf{z}\|^{2k}A(2k,n)}{(2k)!}k(k-1)\mathbb{E}[\widehat{R}^{2k}] \leq \\ 2\sum_{k=0}^{\infty}\|\mathbf{z}\|^{2k}(\frac{1}{2c})^k \cdot 4^k \cdot (2^k)^2 \leq \frac{2}{1-\frac{8}{c}\|\mathbf{z}\|^2}.\end{aligned} \tag{80}$$

Therefore we obtain:

$$|A| \leq (\frac{1}{2n^2} + \frac{3}{n^3} + \frac{5g^2(n)}{2n} + \frac{1}{nh'(n)}) \cdot \frac{2}{1-\frac{8}{c}\|\mathbf{z}\|^2}. \tag{81}$$

Therefore, using Equation 63 and the derived earlier bound on $|\Lambda(\|\mathbf{z}\|)|$, we conclude that

$$\begin{aligned}\rho \geq \frac{1}{2n}\sum_{k=0}^{\infty}\frac{\|\mathbf{z}\|^{2k}(-1)^kA(2k,n)}{(2k)!}k(k-1)\mathbb{E}[\widehat{R}^{2k}] - (\frac{1}{2n^2} + \frac{3}{n^3} + \frac{5g^2(n)}{2n} + \frac{1}{nh'(n)}) \cdot \frac{2}{1-\frac{8}{c}\|\mathbf{z}\|^2} \\ -\frac{3}{n^2}(2\sum_{k=0}^{K(c)}\frac{\|\mathbf{z}\|^{2k}\xi(k)}{k!} + \frac{2\|\mathbf{z}\|^2}{c-\|\mathbf{z}\|^2}).\end{aligned} \tag{82}$$

Note that

$$\sum_{k=0}^{\infty}\frac{\|\mathbf{z}\|^{2k}(-1)^kA(2k,n)}{(2k)!}\mathbb{E}[\widehat{R}^{2k}] = \phi^2(\|\mathbf{z}\|). \tag{83}$$

Note that:

$$\frac{d(\phi^2(x))}{dx}|_{x=\|\mathbf{z}\|} = \sum_{k=0}^{\infty}\frac{2k\|\mathbf{z}\|^{2k-1}(-1)^kA(2k,n)}{(2k)!}\mathbb{E}[\widehat{R}^{2k}] \tag{84}$$

and similarly

$$\frac{d^2(\phi^2(x))}{dx^2}|_{x=\|\mathbf{z}\|} = \sum_{k=0}^{\infty} \frac{2k(2k-1)\|\mathbf{z}\|^{2k-2}(-1)^k A(2k,n)}{(2k)!}\mathbb{E}[\widehat{R}^{2k}] \tag{85}$$

Since $k(k-1) = \frac{2k(2k-1)-2k}{4}$, we get:

$$\begin{gathered}\rho \geq \frac{1}{8n}((\|\mathbf{z}\|^2 \frac{d^2(\phi^2(x))}{dx^2})_{|x=\|\mathbf{z}\|} - (\|\mathbf{z}\|\frac{d(\phi^2(x))}{dx})_{|x=\|\mathbf{z}\|}) \\ -(\frac{1}{2n^2}+\frac{3}{n^3}+\frac{5g^2(n)}{2n}+\frac{1}{nh'(n)})\cdot\frac{2}{1-\frac{8}{c}\|\mathbf{z}\|^2} - \frac{3}{n^2}(2\sum_{k=0}^{K(c)}\frac{\|\mathbf{z}\|^{2k}\xi(k)}{k!}+\frac{2\|\mathbf{z}\|^2}{c-\|\mathbf{z}\|^2}).\end{gathered} \tag{86}$$

As before, by applying Proposition 9.1, we complete the proof of the second part of the theorem.

□

9.3 Proof of Theorem 3.5

Proof. Note that from Theorem 3.7, we knot that $\phi(z) = \psi(z^2)$ for some completely monotone function ψ. Thus we obtain:

$$\frac{d\phi^2(x)}{dx} = 4x\psi(x^2)\frac{d\psi(y)}{dy}\bigg|_{y=x^2}, \tag{87}$$

and

$$\frac{d^2\phi^2(x)}{dx^2} = 4\left[\psi(x^2)\frac{d\psi(y)}{dy}\bigg|_{y=x^2} + 2x^2(\frac{d\psi(y)}{dy}\bigg|_{y=x^2})^2 + 2x^2\psi(x^2)\frac{d^2\psi(y)}{dy^2}\bigg|_{y=x^2}\right]. \tag{88}$$

Therefore we obtain:

$$\Psi_K(z) = 8z^4\left[(\frac{d\psi(y)}{dy}\bigg|_{y=z^2})^2 + \psi(z^2)\frac{d^2\psi(y)}{dy^2}\bigg|_{y=z^2}\right]. \tag{89}$$

That completes the proof since every completely monotone function is nonnegative and convex. □

9.4 Proof of Proposition 3.9

This result follows first by recalling the result of Proposition 9.1, namely that:

$$\begin{gathered}\mathrm{MSE}(\widehat{K}^{\mathrm{iid}}_{m,n}(\mathbf{x},\mathbf{y})) - \mathrm{MSE}(\widehat{K}^{\mathrm{ort}}_{m,n}(\mathbf{x},\mathbf{y})) = \\ \frac{m-1}{m}\left(\mathbb{E}\left[\cos(R_1\|\mathbf{z}\|\langle\mathbf{v},\widehat{\mathbf{z}}\rangle)\right]^2 - \mathbb{E}\left[\cos(\sqrt{R_1^2+R_2^2}\|\mathbf{z}\|\,\langle\mathbf{v},\widehat{\mathbf{z}}\rangle)\right]\right)\end{gathered} \tag{90}$$

Note that both expectations appearing in the expression above have the form

$$\mathbb{E}\left[\cos(A\langle\mathbf{v},\widehat{\mathbf{z}}\rangle)\right] \ ,$$

for some non-negative scalar random variable A. We rewrite this as a nested conditional expectation over the uniform direction:

$$\mathbb{E}\left[\mathbb{E}\left[\cos(A\langle\mathbf{v},\widehat{\mathbf{z}}\rangle)|A\right]\right] \ .$$

Next, we recall from the proof of Proposition 9.1 that the random variable $\langle\mathbf{v},\widehat{\mathbf{z}}\rangle$ may be written $\cos(\theta)$, for a random angle θ distributed on $[0,\pi]$ with density

$$\frac{\sin^{n-2}(\theta)}{\int_0^\pi \sin^{n-2})\theta' d\theta'} \ .$$

Therefore, we can write:

$$\mathbb{E}\left[\mathbb{E}\left[\cos(A\langle\mathbf{v},\widehat{\mathbf{z}}\rangle)|A\right]\right] = \mathbb{E}\left[\frac{\int_0^\pi \cos(A\cos(\theta))\sin^{n-2}(\theta)d\theta}{\int_0^\pi \sin^{n-2}(\theta')d\theta'}\right]$$

For the integral in the denominator of the fraction, we recall that

$$\int_0^{\pi} \sin^{n-2}(\theta)d\theta = \frac{\pi\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n}{2})} \,.$$

For the integral in the numerator, we use the Poisson-Bessel identity:

$$J_\nu(w) = \left(\frac{w}{2}\right)^{\nu} \frac{2}{\sqrt{\pi}\Gamma(\nu - \frac{1}{2})} \int_0^{\pi/2} \cos(w\cos(t))\sin^{2\nu}(t)dt \,.$$

Substituting these expressions into Equation (90) yields the statement of the proposition.

10 Proof of result in Section 4

10.1 Proof of Theorem 4.2

Proof. We will heavily rely on the observations made in the proof of Theorem 9.4. In particular, we use notation from the proof of Theorem 9.4. Fix m and n. Let us take some random feature based smooth estimator $\widehat{K}_{\mathrm{smooth},n}$ and denote by μ_{smooth} the corresponding probabilistic measure used to create the full sample for estimation. Take two vectors $\mathbf{w}_i = R_1\mathbf{v}_1$ and $\mathbf{w}_j = R_2\mathbf{v}_2$, where $\|\mathbf{v}_1\| = \|\mathbf{v}_2\| = 1$, from the full set of vectors $\mathbf{w}$ sampled from μ_{smooth}. Consider the negative covariance term ρ_{ort} for the orthogonal estimator, as in the proof of Theorem 9.4 and the negative covariance term ρ_{smooth} for $\widehat{K}_{\mathrm{smooth},n}$. Note that it suffices to show that:

$$\rho_{\mathrm{diff}} = \rho_{\mathrm{ort}} - \rho_{\mathrm{smooth}} \geq 0 \tag{91}$$

for n large enough. We have:

$$\rho_{\mathrm{diff}} = \sum_{k=0}^{\infty} \frac{\|\mathbf{z}\|^{2k}(-1)^k A(2k,n)}{(2k)!}(\frac{1}{2}(\mathbb{E}[V_1^{2k}] + \mathbb{E}[V_2^{2k}]) - \mathbb{E}[R^{2k}]), \tag{92}$$

where

$$V_1^2 = R_1^2 - 2R_1R_2\mathbf{v}_1^\top\mathbf{v}_2 + R_2^2 = R^2 - 2R_1R_2\mathbf{v}_1^\top\mathbf{v}_2, \tag{93}$$

and

$$V_1^2 = R_1^2 + 2R_1R_2\mathbf{v}_1^\top\mathbf{v}_2 + R_2^2 = R^2 + 2R_1R_2\mathbf{v}_1^\top\mathbf{v}_2, \tag{94}$$

where $R^2 = R_1^2 + R_2^2$.

Denote $\alpha_k = \frac{1}{2}(\mathbb{E}[V_1^{2k}] + \mathbb{E}[V_2^{2k}]) - \mathbb{E}[R^{2k}]$. Note that we have:

$$\alpha_k = \frac{1}{2}\sum_{i=0}^{k} \mathbb{E}[(R^2)^{k-i}(-2R_1R_2\mathbf{v}_1^\top\mathbf{v}_2)^i] + \frac{1}{2}\sum_{i=0}^{k} \mathbb{E}[(R^2)^{k-i}(2R_1R_2\mathbf{v}_1^\top\mathbf{v}_2)^i] - \mathbb{E}[R^{2k}]. \tag{95}$$

From the assumption that samples' lengths are chosen independently from their directions, we get:

$$\alpha_k = \frac{1}{2}\sum_{i=0}^{k} \mathbb{E}[(R^2)^{k-i}(-2R_1R_2)^i]\mathbb{E}[(\mathbf{v}_1^\top\mathbf{v}_2)^i] + \frac{1}{2}\sum_{i=0}^{k} \mathbb{E}[(R^2)^{k-i}(2R_1R_2)^i]\mathbb{E}[(\mathbf{v}_1^\top\mathbf{v}_2)^i] - \mathbb{E}[R^{2k}]. \tag{96}$$

Now by the same analysis as in the proof of Theorem 9.4, and from the smooth property we get:

$$\alpha_k = \frac{k(k-1)}{2}\mathbb{E}[R^{2k-4}(2R_1R_2)^2]\mathbb{E}[(\mathbf{v}_1^\top\mathbf{v}_2))^2] + \cdot 2^k \cdot \mathbb{E}[R^{2k}]q(n)\mathbb{E}[(\mathbf{v}_1^\top\mathbf{v}_2))^2]. \tag{97}$$

Note that for a fixed k, from the property of function q we see that as n goes to infinity, the dominating term in the expression above is the first one. Now we can repeat the analysis from the second part of the proof of Theorem 9.4 with factor $\frac{1}{n}$ in the first expression from the formula on α_k replaced by $\mathbb{E}[(\mathbf{v}_1^\top\mathbf{v}_2))^2]$ and factor $\frac{3}{n^2}$ in the second expression from the same formula replaced by $q(n)\mathbb{E}[(\mathbf{v}_1^\top\mathbf{v}_2))^2]$. By then we see that asymptotically, as n is large enough, the sign of ρ_{diff} is is the same as the sign of $\lambda(\|\mathbf{z}\|)$. But the latter one has to be positive, since otherwise, the orthogonal estimator would not be superior over the default one based on independent sampling. Therefore we have: $\rho_{\mathrm{diff}} \geq 0$ and that, according to our previous observations, completes the proof. □

11 Proof of result from Section 5

11.1 Proof of Theorem 5.4

Proof. Denote $\mathbf{K} + \lambda N \mathbf{I}_N = \mathbf{V}^\top \mathbf{\Sigma}^2 \mathbf{V}$, where an orthonormal matrix $\mathbf{V} \in \mathbf{R}^{N \times N}$ and a diagonal matrix $\mathbf{\Sigma} \in \mathbf{R}^{N \times N}$ define the eigendecomposition of $\mathbf{K} + \lambda N \mathbf{I}_N$. Following Avron et al. (2017), we notice that in order to prove the desired spectral bound, it suffices to show that:

$$\|\mathbf{\Sigma}^{-1}\mathbf{V}\widehat{\mathbf{K}}\mathbf{V}^\top\mathbf{\Sigma}^{-1} - \mathbf{\Sigma}^{-1}\mathbf{V}\mathbf{K}\mathbf{V}^\top\mathbf{\Sigma}^{-1}\|_2 \leq a. \tag{98}$$

From basic properties of the spectral norm $\|\|_2$ and the Frobenius norm $\|\|_F$ we have:

$$\mathbb{P}[\|\mathbf{\Sigma}^{-1}\mathbf{V}\widehat{\mathbf{K}}\mathbf{V}^\top\mathbf{\Sigma}^{-1} - \mathbf{\Sigma}^{-1}\mathbf{V}\mathbf{K}\mathbf{V}^\top\mathbf{\Sigma}^{-1}\|_2 > a] \leq \mathbb{P}[\|\mathbf{\Sigma}^{-1}\mathbf{V}\|_2\|\widehat{\mathbf{K}} - \mathbf{K}\|_F\|\mathbf{V}^\top\mathbf{\Sigma}^{-1}\|_2 > a] \tag{99}$$

The latter probability is equal to $p = \mathbb{P}[\|\widehat{\mathbf{K}} - \mathbf{K}\|_F^2 > \frac{a^2}{\|\mathbf{\Sigma}^{-1}\mathbf{V}\|_2^2 \cdot \|\mathbf{V}^\top\mathbf{\Sigma}^{-1}\|_2^2}]$.

Now note that $\mathbb{E}[\|\widehat{\mathbf{K}} - \mathbf{K}\|_F^2]$ is equal to $\sum_{i,j \in \{1,\ldots,N\}} \mathrm{MSE}(\widehat{\mathrm{K}}(\mathbf{x}_i, \mathbf{x}_j))$ (from the definition of the mean squared error). Furthermore, since $\mathbf{V}$ is an isometry matrix, we have: $\|\mathbf{\Sigma}^{-1}\mathbf{V}\|_2^2 \leq \frac{1}{\sigma_{\min}}$ and $\|\mathbf{V}^\top\mathbf{\Sigma}^{-1}\|_2^2 \leq \frac{1}{\sigma_{\min}}$. Now, we use Markov's inequality to get:

$$\mathbb{P}[\|\widehat{\mathbf{K}} - \mathbf{K}\|_F^2 > \frac{a^2}{\|\mathbf{\Sigma}^{-1}\mathbf{V}\|_2^2 \cdot \|\mathbf{V}^\top\mathbf{\Sigma}^{-1}\|_2^2}] < \frac{\mathbb{E}[\|\widehat{\mathbf{K}} - \mathbf{K}\|_F^2]\|\mathbf{\Sigma}^{-1}\mathbf{V}\|_2^2\|\mathbf{V}^\top\mathbf{\Sigma}^{-1}\|_2^2}{a^2}, \tag{100}$$

substitute the above formula for $\mathbb{E}[\|\widehat{\mathbf{K}} - \mathbf{K}\|_F^2]$ as well as the upper bounds on $\|\mathbf{\Sigma}^{-1}\mathbf{V}\|_2^2$ and $\|\mathbf{V}^\top\mathbf{\Sigma}^{-1}\|_2^2$ and the result follows. □

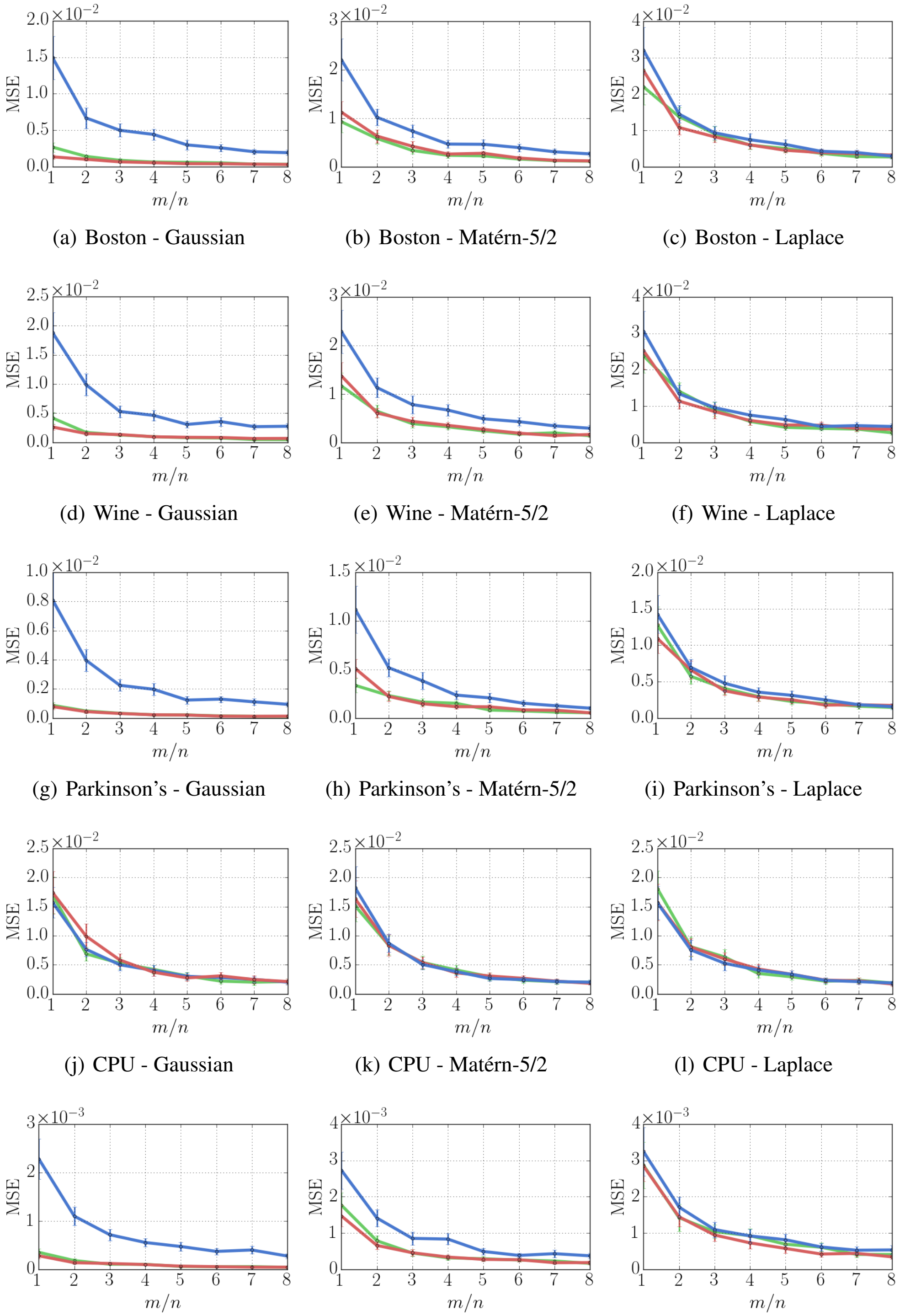
11.2 Proof of Theorem 5.5

Proof. The theorem follows straightforwardly from Theorem 5.4, Lemma 5.3 and the observation that for $0 \leq \Delta < 1$ we have: $\frac{\Delta}{1+\Delta} \leq \frac{1}{2}$ and furthermore, $\mathrm{rank}(\widehat{\mathbf{K}}) \leq m$. The latter one is true, since if a kernel is approximated by m-dimensional feature maps, then the corresponding kernel matrix can be written as a product of two matrices of sizes $N \times m$ and $m \times N$ respectively. □

12 Additional experimental results

12.1 Pointwise kernel and Gram matrix estimation

Here, we provide results for the pointwise kernel and Gram matrix estimation experiments described in Section 6.1 for a larger range of UCI regression datasets. The results for pointwise estimation are given in Figure 9, and the results for Gram matrix estimation are displayed in Figure 10; see the caption for full details.



(a) Boston - Gaussian

(b) Boston - Matérn-5/2

(c) Boston - Laplace

(d) Wine - Gaussian

(e) Wine - Matérn-5/2

(f) Wine - Laplace

(g) Parkinson's - Gaussian

(h) Parkinson's - Matérn-5/2

(i) Parkinson's - Laplace

(j) CPU - Gaussian

(k) CPU - Matérn-5/2

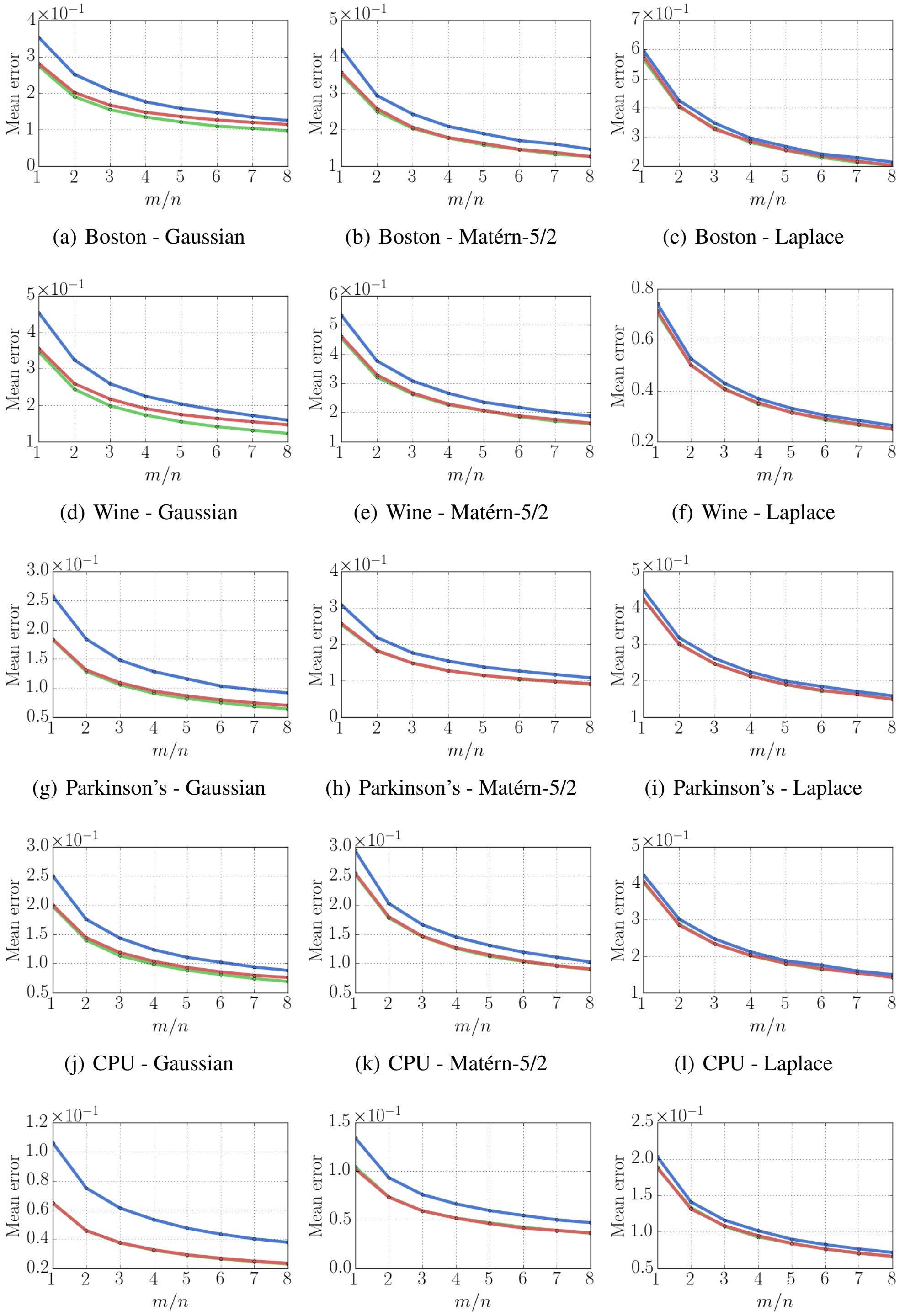
(l) CPU - Laplace

(m) Insurance Company - Gaussian

(n) Insurance Company - Matérn-5/2

(o) Insurance Company - Laplace

Figure 9: MSE for pointwise kernel estimation for a variety of UCI datasets and kernels. Two randomly selected datapoints from each dataset are chosen, and the kernel evaluated at these points is estimated. Estimators are iid random features (blue), orthogonal random features (green) and approximate Hadamard-Rademacher random features (red). In several plots, the red and green curves lie on top of one another.



(a) Boston - Gaussian

(b) Boston - Matérn-5/2

(c) Boston - Laplace

(d) Wine - Gaussian

(e) Wine - Matérn-5/2

(f) Wine - Laplace

(g) Parkinson's - Gaussian

(h) Parkinson's - Matérn-5/2

(i) Parkinson's - Laplace

(j) CPU - Gaussian

(k) CPU - Matérn-5/2

(l) CPU - Laplace

(m) Insurance Company - Gaussian

(n) Insurance Company - Matérn-5/2

(o) Insurance Company - Laplace

Figure 10: Normalized Frobenius norm error for Gram matrix estimation for a variety of UCI datasets and kernels. Estimators are iid random features (blue), orthogonal random features (green) and approximate Hadamard-Rademacher random features (red). In several plots, the red and green curves lie on top of one another.

12.2 Gaussian process regression experiments

In this section we give full results for the Gaussian process regression experiments described in Section 6.2 for a larger range of UCI regression datasets. We report KL divergence against predictions obtained from exact inference (i.e. GP regression without random feature approximation), RMSE prediction error, and wall-clock runtimes; we report the mean and (a bootstrapped estimate of the standard error of this estimate in parentheses) of each of these quantities across 10 runs of the experiment. Experiments were run on a cluster without full control of other processes running on the cluster; timing results should therefore be interpreted cautiously. We emphasise also that a fully-optimised fast Hadamard transform was not used in these experiments, and that the runtime of SORF methods may therefore be an underestimate of the achievable runtimes for these methods.

We observe that the structured methods, ORF and SORF, typically outperform on KL measures. On RMSE, there is little consistent advantage. On timing, we expect that the fast Hadamard transform for rapidly computing matrix multiplications will enable SORF to perform best when dimensionality is high. We do not observe that here, which we believe is due to the use of highly optimized code for (regular) dense matrix multiplication. Runtime performance does start to improve as the dimensionality increases, see results for `insurancecompany`.

Dataset: `boston`. Kernel: Gaussian. Metric: KL divergence against exact GP predictions.

m/n	1	2	3	4	5	6	7	8
IID	104.2 (10.0)	34.21 (1.5)	15.6 (0.87)	11.05 (0.73)	7.946 (0.66)	6.936 (0.41)	5.403 (0.37)	4.349 (0.15)
ORF	**100.4 (5.6)**	**26.62 (1.5)**	**15.1 (1.1)**	**8.707 (0.42)**	7.648 (0.57)	**4.994 (0.21)**	**4.493 (0.3)**	3.832 (0.24)
SORF	108.9 (12.0)	32.29 (2.9)	16.25 (1.3)	10.15 (0.73)	**7.037 (0.31)**	5.783 (0.37)	5.091 (0.37)	**3.501 (0.26)**

Dataset: `boston`. Kernel: Gaussian. Metric: RMSE on test set.

m/n	1	2	3	4	5	6	7	8
IID	**0.54 (0.02)**	0.48 (0.01)	**0.43 (0.008)**	0.4 (0.01)	0.4 (0.008)	0.39 (0.009)	**0.37 (0.005)**	0.38 (0.006)
ORF	0.59 (0.01)	**0.44 (0.008)**	0.43 (0.009)	**0.39 (0.006)**	0.4 (0.01)	**0.38 (0.009)**	0.38 (0.004)	0.38 (0.006)
SORF	0.6 (0.02)	0.5 (0.02)	0.44 (0.009)	0.41 (0.008)	**0.39 (0.008)**	0.4 (0.005)	0.39 (0.004)	**0.36 (0.005)**

Dataset: `boston`. Kernel: Gaussian. Metric: Run time (in seconds).

m/n	1	2	3	4	5	6	7	8
IID	0.00317 (0.0017)	0.00274 (5.5e-05)	0.00409 (2.8e-05)	0.00556 (2.6e-05)	0.00732 (5e-05)	0.00948 (0.00018)	0.0139 (0.0005)	0.0192 (0.00022)
ORF	0.00293 (3e-05)	0.00532 (1.8e-05)	0.00784 (1.9e-05)	0.0137 (0.00064)	0.0196 (8.7e-05)	0.0218 (0.00031)	0.0234 (0.00036)	0.0287 (0.0013)
SORF	0.00217 (2.7e-05)	0.00373 (1.1e-05)	0.00545 (6.7e-05)	0.00858 (0.0008)	0.00949 (0.00016)	0.0154 (0.00059)	0.0196 (0.00019)	0.0229 (0.00018)

Dataset: `boston`. Kernel: Laplace. Metric: KL divergence against exact GP predictions.

m/n	1	2	3	4	5	6	7	8
IID	337.2 (19.0)	126.4 (4.1)	69.66 (3.6)	50.99 (1.7)	35.94 (1.3)	29.81 (1.1)	23.67 (1.1)	18.16 (0.95)
ORF	299.5 (17.0)	**117.7 (3.1)**	**68.4 (2.6)**	**44.25 (1.7)**	39.21 (2.2)	**27.12 (0.99)**	23.79 (1.7)	**16.78 (0.78)**
SORF	**298.3 (7.6)**	121.1 (2.5)	70.56 (1.9)	47.88 (1.5)	**33.45 (1.1)**	27.88 (1.8)	**20.32 (1.1)**	19.87 (0.75)

Dataset: `boston`. Kernel: Laplace. Metric: RMSE on test set.

m/n	1	2	3	4	5	6	7	8
IID	0.69 (0.04)	0.56 (0.02)	0.51 (0.01)	0.48 (0.01)	0.48 (0.009)	0.46 (0.006)	**0.46 (0.01)**	0.45 (0.009)
ORF	0.65 (0.04)	0.54 (0.02)	0.51 (0.01)	0.48 (0.01)	**0.45 (0.01)**	0.45 (0.01)	0.46 (0.007)	0.44 (0.01)
SORF	**0.62 (0.02)**	**0.53 (0.01)**	**0.49 (0.02)**	**0.47 (0.01)**	0.45 (0.01)	**0.45 (0.01)**	0.48 (0.01)	**0.43 (0.01)**

Dataset: `boston`. Kernel: Laplace. Metric: Run time (in seconds).

m/n	1	2	3	4	5	6	7	8
IID	0.0017 (1.5e-05)	0.00301 (0.00011)	0.00491 (0.0004)	0.00603 (6.2e-05)	0.00785 (6.4e-05)	0.00967 (2.5e-05)	0.0163 (0.00051)	0.02 (0.00022)
ORF	0.00363 (0.00038)	0.00571 (0.00011)	0.00822 (4.3e-05)	0.0158 (0.00038)	0.0193 (9.4e-05)	0.0214 (9.3e-05)	0.023 (0.00025)	0.0295 (0.00084)
SORF	0.00224 (9.8e-06)	0.00392 (9.3e-05)	0.00567 (9.1e-05)	0.00821 (0.0006)	0.0102 (0.00043)	0.0166 (0.00034)	0.0195 (0.00014)	0.0224 (9.4e-05)

Dataset: `boston`. Kernel: Matérn-5/2. Metric: KL divergence against exact GP predictions.

m/n	1	2	3	4	5	6	7	8
IID	160.3 (19.0)	47.88 (2.6)	25.87 (1.3)	18.61 (1.2)	12.71 (0.88)	9.837 (0.56)	8.072 (0.45)	7.082 (0.42)
ORF	**123.2 (6.3)**	**41.66 (1.3)**	**21.78 (0.89)**	**16.66 (0.85)**	12.24 (0.5)	**9.43 (0.49)**	**7.726 (0.34)**	6.594 (0.18)
SORF	166.4 (21.0)	44.74 (3.1)	25.14 (0.91)	16.89 (1.1)	**11.44 (0.38)**	10.25 (0.59)	7.73 (0.38)	**6.493 (0.32)**

Dataset: `boston`. Kernel: Matérn-5/2. Metric: RMSE on test set.

m/n	1	2	3	4	5	6	7	8
IID	0.63 (0.02)	0.49 (0.008)	0.45 (0.01)	0.43 (0.006)	**0.4 (0.007)**	**0.39 (0.006)**	**0.39 (0.009)**	0.39 (0.006)
ORF	**0.57 (0.02)**	0.47 (0.02)	**0.42 (0.006)**	**0.42 (0.008)**	0.41 (0.008)	0.41 (0.007)	0.39 (0.009)	**0.38 (0.005)**
SORF	0.61 (0.04)	**0.47 (0.02)**	0.44 (0.01)	0.43 (0.01)	0.42 (0.008)	0.42 (0.008)	0.4 (0.009)	0.39 (0.01)

Dataset: `boston`. Kernel: Matérn-5/2. Metric: Run time (in seconds).

m/n	1	2	3	4	5	6	7	8
IID	0.00274 (0.0011)	0.00271 (5.9e-05)	0.00407 (1.7e-05)	0.00565 (5.8e-05)	0.00734 (2.9e-05)	0.00952 (0.00025)	0.0144 (0.00081)	0.0191 (0.00015)
ORF	0.00319 (0.00024)	0.00544 (2.4e-05)	0.00803 (8.4e-05)	0.0128 (0.00035)	0.0193 (8.3e-05)	0.0218 (0.00052)	0.0236 (0.0004)	0.0288 (0.001)
SORF	0.00277 (0.00059)	0.00366 (1.5e-05)	0.0054 (7e-05)	0.00798 (0.0006)	0.00976 (0.00044)	0.014 (0.00066)	0.0193 (0.00014)	0.0222 (0.00013)

Dataset: `cpu`. Kernel: Gaussian. Metric: KL divergence against exact GP predictions.

m/n	1	2	3	4	5	6	7	8
IID	10640.0 (360.0)	3541.0 (130.0)	**1723.0 (44.0)**	1111.0 (21.0)	770.6 (22.0)	575.0 (14.0)	456.5 (12.0)	367.6 (7.1)
ORF	**10620.0 (310.0)**	**3252.0 (90.0)**	1778.0 (40.0)	**1078.0 (23.0)**	**737.1 (18.0)**	**551.9 (8.5)**	**421.6 (5.1)**	**338.9 (8.0)**
SORF	11200.0 (470.0)	3386.0 (120.0)	1801.0 (58.0)	1153.0 (43.0)	805.3 (13.0)	598.7 (14.0)	479.8 (12.0)	363.5 (7.8)

Dataset: `cpu`. Kernel: Gaussian. Metric: RMSE on test set.

m/n	1	2	3	4	5	6	7	8
IID	**0.69 (0.02)**	0.57 (0.01)	0.51 (0.01)	0.48 (0.007)	0.44 (0.007)	**0.43 (0.007)**	**0.41 (0.008)**	0.41 (0.007)
ORF	0.7 (0.01)	**0.56 (0.02)**	**0.5 (0.01)**	**0.47 (0.007)**	**0.44 (0.009)**	0.43 (0.006)	0.42 (0.006)	**0.39 (0.006)**
SORF	0.72 (0.01)	0.59 (0.01)	0.52 (0.01)	0.5 (0.01)	0.47 (0.007)	0.45 (0.007)	0.43 (0.005)	0.41 (0.005)

Dataset: `cpu`. Kernel: Gaussian. Metric: Run time (in seconds).

m/n	1	2	3	4	5	6	7	8
IID	0.0408 (0.0008)	0.0753 (0.0009)	0.112 (0.00037)	0.156 (0.00058)	0.2 (0.00088)	0.256 (0.0013)	0.305 (0.00074)	0.364 (0.00069)
ORF	0.0434 (0.00026)	0.0805 (0.00088)	0.125 (0.0012)	0.171 (0.00066)	0.218 (0.00071)	0.282 (0.00052)	0.334 (0.001)	0.393 (0.0012)
SORF	0.0413 (0.00081)	0.0748 (0.00082)	0.115 (0.00039)	0.158 (0.00083)	0.204 (0.0011)	0.263 (0.00098)	0.312 (0.0012)	0.37 (0.00069)

Dataset: `cpu`. Kernel: Laplace. Metric: KL divergence against exact GP predictions.

m/n	1	2	3	4	5	6	7	8
IID	32220.0 (820.0)	12710.0 (150.0)	8092.0 (110.0)	**5571.0 (64.0)**	4354.0 (55.0)	3593.0 (53.0)	3069.0 (52.0)	2600.0 (36.0)
ORF	31460.0 (710.0)	**12530.0 (140.0)**	**7861.0 (63.0)**	5720.0 (58.0)	4400.0 (54.0)	**3581.0 (53.0)**	2991.0 (36.0)	2614.0 (21.0)
SORF	**31170.0 (700.0)**	12550.0 (100.0)	8001.0 (160.0)	5735.0 (40.0)	**4317.0 (28.0)**	3593.0 (41.0)	**2983.0 (53.0)**	**2527.0 (46.0)**

Dataset: `cpu`. Kernel: Laplace. Metric: RMSE on test set.

m/n	1	2	3	4	5	6	7	8
IID	0.76 (0.03)	0.56 (0.01)	**0.48 (0.01)**	0.44 (0.008)	0.42 (0.01)	**0.36 (0.007)**	0.36 (0.008)	0.36 (0.004)
ORF	0.74 (0.01)	0.53 (0.02)	0.48 (0.01)	**0.41 (0.01)**	**0.4 (0.01)**	0.37 (0.007)	**0.35 (0.008)**	**0.35 (0.009)**
SORF	**0.7 (0.02)**	**0.52 (0.01)**	0.48 (0.01)	0.44 (0.01)	0.42 (0.01)	0.38 (0.01)	0.36 (0.008)	0.35 (0.008)

Dataset: `cpu`. Kernel: Laplace. Metric: Run time (in seconds).

m/n	1	2	3	4	5	6	7	8
IID	0.0427 (0.00057)	0.0801 (0.0006)	0.121 (0.00086)	0.169 (0.00085)	0.214 (0.00069)	0.272 (0.0016)	0.325 (0.00078)	0.385 (0.001)
ORF	0.0458 (0.0003)	0.0847 (0.00054)	0.133 (0.0011)	0.183 (0.0007)	0.23 (0.00072)	0.297 (0.001)	0.349 (0.00074)	0.413 (0.0019)
SORF	0.0431 (0.00021)	0.0801 (0.00055)	0.122 (0.0013)	0.169 (0.0011)	0.217 (0.00086)	0.29 (0.012)	0.326 (0.0016)	0.39 (0.0014)

Dataset: `cpu`. Kernel: Matérn-5/2. Metric: KL divergence against exact GP predictions.

m/n	1	2	3	4	5	6	7	8
IID	14100.0 (500.0)	4806.0 (130.0)	2574.0 (38.0)	1784.0 (41.0)	1276.0 (15.0)	979.7 (19.0)	791.1 (18.0)	**646.0 (9.4)**
ORF	**13750.0 (400.0)**	4653.0 (120.0)	2578.0 (48.0)	**1684.0 (20.0)**	1227.0 (21.0)	**946.1 (13.0)**	**773.0 (13.0)**	666.2 (12.0)
SORF	14460.0 (580.0)	**4649.0 (82.0)**	**2482.0 (62.0)**	1693.0 (21.0)	**1202.0 (18.0)**	954.4 (14.0)	773.4 (7.8)	675.5 (6.9)

Dataset: `cpu`. Kernel: Matérn-5/2. Metric: RMSE on test set.

m/n	1	2	3	4	5	6	7	8
IID	0.7 (0.01)	0.57 (0.02)	0.49 (0.008)	0.45 (0.008)	0.44 (0.005)	0.4 (0.008)	0.4 (0.006)	0.38 (0.006)
ORF	**0.7 (0.01)**	**0.54 (0.01)**	0.47 (0.009)	**0.43 (0.009)**	0.43 (0.01)	0.4 (0.008)	0.38 (0.01)	**0.37 (0.007)**
SORF	0.71 (0.03)	0.54 (0.01)	**0.47 (0.01)**	0.44 (0.006)	**0.39 (0.009)**	**0.4 (0.007)**	**0.38 (0.01)**	0.39 (0.01)

Dataset: `cpu`. Kernel: Matérn-5/2. Metric: Run time (in seconds).

m/n	1	2	3	4	5	6	7	8
IID	0.041 (0.00084)	0.0737 (0.00072)	0.115 (0.0006)	0.158 (0.0008)	0.201 (0.00084)	0.259 (0.0015)	0.312 (0.0015)	0.369 (0.00062)
ORF	0.0449 (0.00031)	0.0805 (0.00089)	0.127 (0.0008)	0.173 (0.00064)	0.218 (0.0011)	0.286 (0.00081)	0.337 (0.00062)	0.397 (0.0013)
SORF	0.0415 (0.0016)	0.0748 (0.001)	0.117 (0.00074)	0.162 (0.0011)	0.205 (0.00072)	0.265 (0.0012)	0.316 (0.0013)	0.376 (0.0017)

Dataset: `insurancecompany`. Kernel: Gaussian. Metric: KL divergence against exact GP predictions.

m/n	1	2	3	4	5	6	7	8
IID	329.1 (2.6)	144.9 (2.3)	90.02 (1.8)	68.0 (1.3)	51.9 (0.88)	42.84 (0.7)	36.57 (0.39)	31.15 (0.62)
ORF	**292.3 (5.7)**	132.3 (1.2)	**83.06 (1.3)**	**60.39 (1.2)**	49.0 (0.85)	**39.28 (0.87)**	**33.74 (0.5)**	**28.24 (0.36)**
SORF	295.4 (5.2)	**131.2 (2.5)**	83.47 (0.9)	60.84 (1.0)	**48.12 (1.0)**	39.79 (0.9)	33.82 (0.54)	29.86 (0.49)

Dataset: `insurancecompany`. Kernel: Gaussian. Metric: RMSE on test set.

m/n	1	2	3	4	5	6	7	8
IID	0.99 (0.001)	0.98 (0.0009)	0.98 (0.0005)	0.98 (0.0007)	0.98 (0.0008)	**0.98 (0.0005)**	0.98 (0.0006)	0.98 (0.0006)
ORF	**0.99 (0.0007)**	0.98 (0.001)	0.98 (0.0008)	**0.98 (0.0005)**	0.98 (0.0007)	0.98 (0.0008)	**0.98 (0.0006)**	0.98 (0.0006)
SORF	0.99 (0.001)	**0.98 (0.001)**	**0.98 (0.001)**	0.98 (0.0006)	**0.98 (0.0005)**	0.98 (0.0006)	0.98 (0.0006)	**0.98 (0.0007)**

Dataset: `insurancecompany`. Kernel: Gaussian. Metric: Run time (in seconds).

m/n	1	2	3	4	5	6	7	8
IID	0.141 (0.00066)	0.316 (0.00087)	0.524 (0.0012)	0.809 (0.0012)	1.13 (0.0022)	1.57 (0.0025)	2.09 (0.0018)	2.84 (0.0037)
ORF	0.183 (0.0019)	0.393 (0.0012)	0.634 (0.00079)	0.953 (0.0018)	1.32 (0.0022)	1.79 (0.0028)	2.36 (0.0085)	3.15 (0.0038)
SORF	0.143 (0.001)	0.315 (0.00065)	0.526 (0.0014)	0.813 (0.0051)	1.13 (0.0023)	1.58 (0.003)	2.1 (0.002)	2.84 (0.0032)

Dataset: `insurancecompany`. Kernel: Laplace. Metric: KL divergence against exact GP predictions.

m/n	1	2	3	4	5	6	7	8
IID	1297.0 (28.0)	618.1 (10.0)	**378.9 (4.9)**	292.8 (3.2)	233.6 (3.9)	189.9 (6.0)	161.9 (2.0)	139.2 (2.2)
ORF	1310.0 (23.0)	**592.0 (12.0)**	396.3 (5.9)	285.8 (5.1)	235.0 (3.2)	189.0 (3.7)	**156.1 (2.9)**	**137.3 (2.2)**
SORF	**1294.0 (27.0)**	600.9 (17.0)	385.2 (5.8)	**278.9 (3.2)**	**227.0 (2.9)**	**188.3 (3.4)**	165.2 (2.7)	137.9 (2.8)

Dataset: `insurancecompany`. Kernel: Laplace. Metric: RMSE on test set.

m/n	1	2	3	4	5	6	7	8
IID	0.99 (0.002)	**0.99 (0.001)**	1.0 (0.002)	0.99 (0.001)	**0.99 (0.001)**	0.99 (0.001)	0.99 (0.002)	**0.99 (0.001)**
ORF	**0.99 (0.002)**	0.99 (0.002)	**0.99 (0.002)**	1.0 (0.001)	0.99 (0.001)	**0.99 (0.002)**	**0.99 (0.001)**	0.99 (0.001)
SORF	0.99 (0.001)	1.0 (0.002)	1.0 (0.001)	**0.99 (0.002)**	0.99 (0.001)	0.99 (0.002)	0.99 (0.001)	0.99 (0.002)

Dataset: `insurancecompany`. Kernel: Laplace. Metric: Run time (in seconds).

m/n	1	2	3	4	5	6	7	8
IID	0.154 (0.0011)	0.33 (0.00087)	0.554 (0.001)	0.852 (0.0019)	1.19 (0.0016)	1.64 (0.0047)	2.16 (0.0012)	2.9 (0.0015)
ORF	0.196 (0.00076)	0.412 (0.0011)	0.672 (0.0062)	1.0 (0.0017)	1.38 (0.0027)	1.86 (0.0029)	2.43 (0.0027)	3.2 (0.0022)
SORF	0.153 (0.0011)	0.333 (0.00087)	0.556 (0.0013)	0.854 (0.0013)	1.19 (0.0043)	1.64 (0.0022)	2.16 (0.0025)	2.92 (0.0043)

Dataset: `insurancecompany`. Kernel: Matérn-5/2. Metric: KL divergence against exact GP predictions.

m/n	1	2	3	4	5	6	7	8
IID	589.9 (6.4)	277.1 (6.3)	185.7 (4.1)	138.6 (2.3)	109.8 (2.6)	88.9 (2.1)	79.41 (1.8)	69.9 (1.1)
ORF	561.7 (4.8)	**257.9 (6.2)**	**174.6 (2.8)**	129.5 (2.2)	**101.3 (2.1)**	87.41 (2.1)	**74.41 (1.5)**	65.53 (1.1)
SORF	**552.3 (7.4)**	264.8 (4.8)	183.0 (3.4)	**129.1 (2.1)**	104.5 (1.3)	**83.74 (1.7)**	75.23 (1.6)	**64.51 (1.2)**

Dataset: `insurancecompany`. Kernel: Matérn-5/2. Metric: RMSE on test set.

m/n	1	2	3	4	5	6	7	8
IID	0.99 (0.002)	0.99 (0.0008)	**0.98 (0.0008)**	**0.98 (0.001)**	**0.98 (0.0009)**	0.98 (0.001)	0.98 (0.001)	**0.98 (0.0008)**
ORF	0.99 (0.001)	**0.98 (0.0009)**	0.99 (0.001)	0.99 (0.001)	0.98 (0.0007)	**0.98 (0.0009)**	**0.98 (0.001)**	0.98 (0.0008)
SORF	**0.99 (0.001)**	0.99 (0.002)	0.99 (0.0008)	0.98 (0.0008)	0.98 (0.0007)	0.98 (0.001)	0.98 (0.0006)	0.98 (0.001)

Dataset: `insurancecompany`. Kernel: Matérn-5/2. Metric: Run time (in seconds).

m/n	1	2	3	4	5	6	7	8
IID	0.143 (0.00072)	0.315 (0.00044)	0.528 (0.0015)	0.818 (0.0013)	1.14 (0.0026)	1.59 (0.0013)	2.11 (0.0056)	2.82 (0.004)
ORF	0.184 (0.00071)	0.392 (0.00058)	0.642 (0.0011)	0.969 (0.0015)	1.33 (0.0029)	1.81 (0.0017)	2.36 (0.0033)	3.12 (0.0031)
SORF	0.145 (0.0007)	0.316 (0.0013)	0.533 (0.003)	0.818 (0.00078)	1.14 (0.0021)	1.59 (0.0032)	2.1 (0.0032)	2.83 (0.0059)

Dataset: `parkinson`. Kernel: Gaussian. Metric: KL divergence against exact GP predictions.

m/n	1	2	3	4	5	6	7	8
IID	7421.0 (360.0)	3148.0 (150.0)	1632.0 (70.0)	1061.0 (52.0)	848.7 (38.0)	670.7 (20.0)	541.6 (16.0)	521.0 (21.0)
ORF	**6854.0 (300.0)**	**2692.0 (96.0)**	1550.0 (51.0)	**964.2 (38.0)**	798.7 (28.0)	629.7 (21.0)	529.1 (19.0)	429.0 (16.0)
SORF	7537.0 (310.0)	2776.0 (95.0)	**1548.0 (60.0)**	1041.0 (55.0)	**767.5 (24.0)**	**623.0 (22.0)**	**514.5 (24.0)**	**427.8 (15.0)**

Dataset: `parkinson`. Kernel: Gaussian. Metric: RMSE on test set.

m/n	1	2	3	4	5	6	7	8
IID	0.87 (0.004)	0.84 (0.004)	0.81 (0.004)	0.79 (0.004)	0.79 (0.003)	0.78 (0.003)	0.77 (0.002)	0.78 (0.002)
ORF	**0.87 (0.005)**	**0.83 (0.004)**	**0.81 (0.003)**	**0.79 (0.002)**	**0.78 (0.003)**	**0.77 (0.004)**	0.77 (0.002)	**0.77 (0.003)**
SORF	0.88 (0.002)	0.84 (0.003)	0.81 (0.003)	0.79 (0.002)	0.79 (0.002)	0.78 (0.002)	**0.77 (0.002)**	0.77 (0.002)

Dataset: `parkinson`. Kernel: Gaussian. Metric: Run time (in seconds).

m/n	1	2	3	4	5	6	7	8
IID	0.0332 (0.00088)	0.0586 (0.00049)	0.091 (0.00042)	0.127 (0.0008)	0.164 (0.00066)	0.203 (0.0007)	0.246 (0.0013)	0.302 (0.0015)
ORF	0.0378 (0.00048)	0.0674 (0.00078)	0.103 (0.00067)	0.143 (0.00093)	0.18 (0.00077)	0.227 (0.00083)	0.273 (0.0011)	0.334 (0.0016)
SORF	0.0356 (0.00044)	0.0612 (0.00073)	0.0944 (0.001)	0.13 (0.00087)	0.168 (0.00076)	0.207 (0.00074)	0.249 (0.0012)	0.311 (0.00096)

Dataset: `parkinson`. Kernel: Laplace. Metric: KL divergence against exact GP predictions.

m/n	1	2	3	4	5	6	7	8
IID	18120.0 (420.0)	8041.0 (170.0)	**4966.0 (140.0)**	3668.0 (100.0)	**2677.0 (56.0)**	2224.0 (44.0)	1855.0 (33.0)	1554.0 (31.0)
ORF	**17860.0 (350.0)**	8145.0 (220.0)	5159.0 (170.0)	3478.0 (68.0)	2713.0 (77.0)	2231.0 (58.0)	**1727.0 (27.0)**	**1501.0 (33.0)**
SORF	17970.0 (350.0)	**7760.0 (170.0)**	5162.0 (100.0)	**3448.0 (97.0)**	2765.0 (83.0)	**2217.0 (49.0)**	1941.0 (54.0)	1541.0 (31.0)

Dataset: `parkinson`. Kernel: Laplace. Metric: RMSE on test set.

m/n	1	2	3	4	5	6	7	8
IID	**0.9 (0.005)**	**0.85 (0.006)**	**0.83 (0.007)**	0.82 (0.005)	0.81 (0.006)	0.8 (0.005)	0.79 (0.004)	**0.77 (0.004)**
ORF	0.91 (0.006)	0.86 (0.005)	0.83 (0.005)	0.82 (0.002)	**0.8 (0.002)**	**0.79 (0.003)**	**0.78 (0.005)**	0.78 (0.003)
SORF	0.91 (0.007)	0.86 (0.004)	0.85 (0.002)	**0.81 (0.004)**	0.81 (0.004)	0.79 (0.005)	0.79 (0.003)	0.78 (0.002)

Dataset: `parkinson`. Kernel: Laplace. Metric: Run time (in seconds).

m/n	1	2	3	4	5	6	7	8
IID	0.0375 (0.00081)	0.0661 (0.00077)	0.0997 (0.00094)	0.14 (0.00096)	0.18 (0.00066)	0.22 (0.0012)	0.262 (0.0013)	0.317 (0.0013)
ORF	0.0402 (0.00083)	0.0708 (0.001)	0.109 (0.00069)	0.152 (0.001)	0.197 (0.0012)	0.252 (0.012)	0.292 (0.0013)	0.352 (0.00071)
SORF	0.0388 (0.00056)	0.0685 (0.00085)	0.103 (0.0015)	0.141 (0.0016)	0.182 (0.00057)	0.224 (0.0011)	0.265 (0.00091)	0.328 (0.00089)

Dataset: `parkinson`. Kernel: Matérn-5/2. Metric: KL divergence against exact GP predictions.

m/n	1	2	3	4	5	6	7	8
IID	12070.0 (410.0)	**4711.0 (110.0)**	2882.0 (77.0)	2053.0 (81.0)	1531.0 (58.0)	1364.0 (47.0)	1047.0 (33.0)	**911.2 (24.0)**
ORF	**10540.0 (320.0)**	4787.0 (160.0)	2853.0 (100.0)	2070.0 (36.0)	**1497.0 (60.0)**	1235.0 (29.0)	**1042.0 (44.0)**	950.1 (33.0)
SORF	11520.0 (280.0)	5035.0 (130.0)	**2742.0 (85.0)**	**1999.0 (71.0)**	1604.0 (58.0)	**1214.0 (28.0)**	1090.0 (36.0)	915.2 (36.0)

Dataset: `parkinson`. Kernel: Matérn-5/2. Metric: RMSE on test set.

m/n	1	2	3	4	5	6	7	8
IID	0.88 (0.006)	**0.83 (0.003)**	**0.8 (0.003)**	0.79 (0.004)	**0.77 (0.004)**	0.77 (0.003)	0.76 (0.003)	**0.75 (0.003)**
ORF	**0.87 (0.003)**	0.83 (0.002)	0.81 (0.004)	**0.79 (0.003)**	0.78 (0.004)	**0.77 (0.003)**	**0.76 (0.004)**	0.76 (0.003)
SORF	0.88 (0.004)	0.84 (0.003)	0.8 (0.004)	0.79 (0.003)	0.78 (0.003)	0.77 (0.002)	0.77 (0.002)	0.76 (0.003)

Dataset: `parkinson`. Kernel: Matérn-5/2. Metric: Run time (in seconds).

m/n	1	2	3	4	5	6	7	8
IID	0.0331 (0.00078)	0.0594 (0.00028)	0.0925 (0.00076)	0.128 (0.00043)	0.168 (0.0005)	0.207 (0.0013)	0.248 (0.00094)	0.304 (0.0013)
ORF	0.0379 (0.00047)	0.0656 (0.00048)	0.103 (0.00048)	0.144 (0.00038)	0.185 (0.00083)	0.228 (0.0011)	0.274 (0.0011)	0.336 (0.0012)
SORF	0.0337 (0.00087)	0.0602 (0.00035)	0.0934 (0.00068)	0.13 (0.0005)	0.17 (0.0011)	0.209 (0.00078)	0.252 (0.00098)	0.313 (0.00064)

Dataset: wine. Kernel: Gaussian. Metric: KL divergence against exact GP predictions.

m/n	1	2	3	4	5	6	7	8
IID	30940.0 (820.0)	11470.0 (180.0)	6677.0 (100.0)	4362.0 (59.0)	3083.0 (25.0)	2415.0 (34.0)	1906.0 (28.0)	**1529.0 (12.0)**
ORF	**28010.0 (550.0)**	**11020.0 (120.0)**	**6313.0 (110.0)**	**4192.0 (47.0)**	**3023.0 (57.0)**	**2337.0 (22.0)**	**1847.0 (31.0)**	1546.0 (26.0)
SORF	31700.0 (680.0)	12520.0 (300.0)	7011.0 (210.0)	4635.0 (72.0)	3330.0 (57.0)	2514.0 (39.0)	1979.0 (33.0)	1665.0 (27.0)

Dataset: wine. Kernel: Gaussian. Metric: RMSE on test set.

m/n	1	2	3	4	5	6	7	8
IID	0.86 (0.005)	0.82 (0.002)	0.82 (0.002)	0.81 (0.002)	0.8 (0.002)	0.8 (0.002)	0.8 (0.002)	**0.79 (0.002)**
ORF	**0.85 (0.003)**	**0.82 (0.002)**	**0.81 (0.003)**	**0.8 (0.003)**	**0.8 (0.002)**	**0.8 (0.002)**	**0.79 (0.002)**	0.79 (0.002)
SORF	0.86 (0.004)	0.83 (0.002)	0.82 (0.002)	0.81 (0.002)	0.8 (0.002)	0.8 (0.002)	0.8 (0.001)	0.8 (0.002)

Dataset: wine. Kernel: Gaussian. Metric: Run time (in seconds).

m/n	1	2	3	4	5	6	7	8
IID	0.0183 (0.00064)	0.0399 (0.00069)	0.0529 (0.00049)	0.0693 (0.0006)	0.0897 (0.001)	0.11 (0.00085)	0.126 (0.00099)	0.148 (0.00066)
ORF	0.0198 (0.00045)	0.0419 (0.00058)	0.0567 (0.00077)	0.0747 (0.00068)	0.097 (0.0007)	0.118 (0.00094)	0.137 (0.00091)	0.162 (0.00066)
SORF	0.0194 (8.6e-05)	0.0411 (0.00026)	0.055 (0.00095)	0.0715 (0.00076)	0.0911 (0.00063)	0.112 (0.00059)	0.132 (0.0008)	0.152 (0.0011)

Dataset: wine. Kernel: Laplace. Metric: KL divergence against exact GP predictions.

m/n	1	2	3	4	5	6	7	8
IID	198300.0 (1500.0)	96220.0 (810.0)	62890.0 (850.0)	47220.0 (550.0)	38770.0 (460.0)	31470.0 (450.0)	27440.0 (400.0)	23730.0 (270.0)
ORF	**197400.0 (1700.0)**	**95470.0 (570.0)**	**62140.0 (440.0)**	**46260.0 (550.0)**	**37030.0 (360.0)**	**31290.0 (230.0)**	**27180.0 (290.0)**	**23130.0 (260.0)**
SORF	198600.0 (2400.0)	96500.0 (960.0)	64060.0 (740.0)	46880.0 (530.0)	38490.0 (400.0)	32190.0 (390.0)	27260.0 (330.0)	23470.0 (280.0)

Dataset: wine. Kernel: Laplace. Metric: RMSE on test set.

m/n	1	2	3	4	5	6	7	8
IID	**0.88 (0.01)**	0.84 (0.006)	**0.82 (0.002)**	0.82 (0.004)	0.81 (0.002)	0.8 (0.003)	0.8 (0.002)	0.8 (0.003)
ORF	0.89 (0.01)	**0.83 (0.004)**	0.82 (0.003)	**0.81 (0.002)**	0.8 (0.003)	**0.8 (0.001)**	0.8 (0.003)	**0.8 (0.002)**
SORF	0.88 (0.007)	0.84 (0.004)	0.82 (0.004)	0.81 (0.003)	**0.8 (0.002)**	0.8 (0.003)	**0.8 (0.002)**	0.8 (0.003)

Dataset: wine. Kernel: Laplace. Metric: Run time (in seconds).

m/n	1	2	3	4	5	6	7	8
IID	0.0198 (0.00018)	0.0394 (0.00064)	0.0586 (0.00047)	0.0773 (0.00041)	0.0959 (0.00072)	0.117 (0.00094)	0.14 (0.00066)	0.162 (0.0011)
ORF	0.0205 (3.5e-05)	0.0415 (0.00051)	0.0625 (0.00061)	0.0833 (0.0008)	0.104 (0.00057)	0.129 (0.00073)	0.152 (0.0012)	0.174 (0.00087)
SORF	0.0202 (0.00014)	0.0411 (0.00038)	0.0592 (0.00067)	0.0785 (0.00067)	0.101 (0.00093)	0.12 (0.00086)	0.143 (0.0011)	0.165 (0.0011)

Dataset: wine. Kernel: Matérn-5/2. Metric: KL divergence against exact GP predictions.

m/n	1	2	3	4	5	6	7	8
IID	62860.0 (1200.0)	26610.0 (240.0)	16740.0 (160.0)	12030.0 (130.0)	9344.0 (50.0)	**7519.0 (110.0)**	**6424.0 (51.0)**	**5483.0 (67.0)**
ORF	**57660.0 (250.0)**	**26110.0 (190.0)**	**16390.0 (110.0)**	**12020.0 (100.0)**	**9344.0 (100.0)**	7584.0 (120.0)	6474.0 (39.0)	5561.0 (66.0)
SORF	60670.0 (760.0)	26500.0 (260.0)	16810.0 (140.0)	12300.0 (160.0)	9557.0 (71.0)	7708.0 (42.0)	6495.0 (97.0)	5494.0 (43.0)

Dataset: wine. Kernel: Matérn-5/2. Metric: RMSE on test set.

m/n	1	2	3	4	5	6	7	8
IID	0.87 (0.005)	0.83 (0.003)	0.81 (0.001)	0.81 (0.002)	0.8 (0.002)	0.8 (0.002)	0.8 (0.002)	0.79 (0.002)
ORF	**0.84 (0.003)**	0.82 (0.003)	**0.81 (0.003)**	0.81 (0.002)	**0.8 (0.002)**	0.8 (0.002)	0.79 (0.002)	0.79 (0.002)
SORF	0.86 (0.005)	**0.82 (0.004)**	0.81 (0.002)	**0.8 (0.002)**	0.8 (0.002)	**0.8 (0.001)**	**0.79 (0.002)**	**0.79 (0.002)**

Dataset: wine. Kernel: Matérn-5/2. Metric: Run time (in seconds).

m/n	1	2	3	4	5	6	7	8
IID	0.0188 (0.00024)	0.038 (0.00072)	0.0553 (0.00045)	0.0732 (0.00066)	0.0933 (0.00091)	0.113 (0.00073)	0.131 (0.00076)	0.153 (0.0011)
ORF	0.02 (0.00018)	0.0407 (0.00064)	0.0588 (0.00087)	0.0781 (0.00081)	0.0974 (0.00074)	0.122 (0.00092)	0.143 (0.00067)	0.165 (0.0011)
SORF	0.0192 (0.00041)	0.0402 (0.00048)	0.0574 (0.0008)	0.0746 (0.00057)	0.0944 (0.00064)	0.113 (0.0012)	0.135 (0.00071)	0.156 (0.00088)