Random Fourier Features For Operator-Valued Kernels

Romain Brault ROMAIN.BRAULT@TELECOM-PARISTECH.FR
IBISC, Université d'Évry val d'Essonne
LTCI, CNRS, Télécom ParisTech
46 rue Barrault, Paris, 75684 cedex 13, France

Markus Heinonen MARKUS.O.HEINONEN@AALTO.FI
Department of Information and Computer Science, Aalto University
FI-00076 Aalto, PO Box 15400, Finland

Florence d'Alché-Buc FLORENCE.DALCHE@TELECOM-PARISTECH.FR
LTCI, CNRS, Télécom ParisTech
Université Paris-Saclay
46 rue Barrault, Paris, 75684 cedex 13, France

Editors: Robert J. Durrant and Kee-Eung Kim

Supplementary Materials

A. Uniform error bound for decomposable ORFF

Rahimi and Recht (2007) proved the uniform convergence of Random Fourier Feature (RFF) approximation for a scalar shift invariant kernel.

Theorem 1 (Uniform error bound for RFF, Rahimi and Recht (2007)) *Let $\mathcal{C}$ be a compact of subset of $\mathbb{R}^d$ of diameter $|\mathcal{C}|$. Let k a shift invariant kernel, differentiable with a bounded first derivative and μ its normalized inverse Fourier transform. Let D the dimension of the Fourier feature vectors. Then, for the mapping $\tilde{\phi}$ described in section 3, we have :*

$$\mathbb{P}\left\{\sup_{x,z\in\mathcal{C}}\left\|\tilde{k}(x,z)-k(x,z)\right\|_2\geq\epsilon\right\}\leq 2^8\left(\frac{d\sigma|\mathcal{C}|}{\epsilon}\right)^2\exp\left(-\frac{\epsilon^2 D}{4(d+2)}\right) \tag{S1}$$

From theorem 1, we deduce the following corollary about the uniform convergence of the ORFF approximation of the decomposable kernel. We recall that: for a given pair $(x,z)\in\mathcal{C}^2$, $\tilde{K}(x,z)=\tilde{\Phi}(x)^*\tilde{\Phi}(z)=A\tilde{k}(x,z)$ and $K_0(x-z)=AE_\mu[\tilde{k}(x,z)]$.

Corollary 2 (Uniform error bound for decomposable ORFF) *Let $\mathcal{C}$ be a compact of subset of $\mathbb{R}^d$ of diameter $|\mathcal{C}|$. K_{dec} is a decomposable kernel built from a $p\times p$ positive semi-definite matrix A and k, a shift invariant and differentiable kernel whose first derivative is bounded. Let $\tilde{k}$ the Random Fourier approximation for the scalar-valued kernel k.*

$$\mathbb{P}\left\{\sup_{x,z\in\mathcal{C}}\left\|\tilde{K}(x,z)-K(x,z)\right\|_2\geq\epsilon\right\}\leq 2^8\left(\frac{d\sigma\|A\|_2|\mathcal{C}|}{\epsilon}\right)^2\exp\left(-\frac{\epsilon^2 D}{4\|A\|_2^2(d+2)}\right)$$

Proof. The proof directly extends 1 given by Rahimi and Recht (2007). Since

$$\sup_{x,z\in\mathcal{C}}\left\|\tilde{K}(x,z)-K(x,z)\right\|_2=\sup_{x,z\in\mathcal{C}}\|A\|_2.\left|\tilde{k}(x,z)-k(x,z)\right|$$

and then, taking $\epsilon'=\|A\|_2\epsilon$ gives the following result for all positive ϵ':

$$\mathbb{P}\left\{\sup_{x,z\in\mathcal{C}}\left\|A(\tilde{k}(x,z)-k(x,z))\right\|_2\geq\epsilon'\right\}\leq 2^8\left(\frac{d\sigma\|A\|_2|\mathcal{C}|}{\epsilon'}\right)^2\exp\left(-\frac{\epsilon'^2D}{4\|A\|_2^2(d+2)}\right)$$

■

Note that a similar corollary could have been obtained from the recent result of Sutherland and Schneider (2015) who refined the bound proposed by Rahimi and Recht (2007) by using a Bernstein concentration inequality instead of the Hoeffding inequality.

B. Proof of theorem 10

.

We recall the notations $\delta=x-z$, $\tilde{K}(x,z)=\tilde{\Phi}(x)\tilde{\Phi}(z)$, $\tilde{K}^j(x,z)=\Phi_x(\omega_j)\Phi_z(\omega_j)$ and $K_0(\delta)=K(x,z)$. For the sake of simplicity, we use throughout the proof the quantities:

$$\begin{aligned}F(\delta)&:=\tilde{K}(x,z)-K(x,z)\\F^j(\delta)&:=(\tilde{K}^j(x,z)-K(x,z))/D\end{aligned}$$

Compared to the scalar case, the proof follows the same scheme as the one described in (Rahimi and Recht, 2007; Sutherland and Schneider, 2015) but requires to consider matrix norms and appropriate matrix concentration inequality. The main feature of theorem 10 is that it covers the case of bounded ORFF as well as unbounded ORFF: in the case of bounded ORFF, a Bernstein inequality for matrix concentration such that the one proved in Mackey et al. (2014) (Corollary 5.2) or the formulation of Tropp (2012) recalled in Koltchinskii et al. (2013) is suitable. However some kernels like the curl and the div-free kernels do not have bounded $\left\|F^j\right\|_2$ but exhibit F^j with subexponential tails. Therefore, we use a Bernstein matrix concentration inequality adapted for random matrices with subexponential norms.

B.1. Epsilon-net

Let $\mathcal{D}_\mathcal{C}=\{x-z|x,z\in\mathcal{C}\}$ with diameter at most $2|\mathcal{C}|$ where $|\mathcal{C}|$ is the diameter of $\mathcal{C}$. Since $\mathcal{C}$ is supposed compact, so is $\mathcal{D}_\mathcal{C}$. It is then possible to find an ϵ-net covering $\mathcal{D}_\mathcal{C}$ with at most $T=(4|\mathcal{C}|/r)^d$ balls of radius r.

Let us call $\delta_i, i=1,\ldots,T$ the center of the i-th ball, also called anchor of the ϵ-net. Denote L_F the Lipschitz constant of F. Let $\|.\|_2$ be the ℓ_2 norm on $\mathcal{L}(\mathbb{R}^p)$, that is the spectral norm. We introduce the following lemma:

Lemma 3 *$\forall\delta\in\mathcal{D}_\mathcal{C}$, if (1): $L_F\leq\frac{\epsilon}{2r}$ and (2): $\|F(\delta_i)\|_2\leq\frac{\epsilon}{2}$,for all $0<i<T$, then $\|F(\delta)\|_2\leq\epsilon$.*

Proof. $\|F(\delta)\|_2 = \|F(\delta) - F(\delta_i) + F(\delta_i)\|_2 \leq \|F(\delta) - F(\delta_i)\|_2 + \|F(\delta_i)\|_2$, for all $0 < i < T$. Using the Lipschitz continuity of F we have $\|F(\delta) - F(\delta_i)\|_2 \leq L_F\|\delta - \delta_i\|_2 \leq rL_F$ hence $\|F(\delta)\|_2 \leq rL_F + \|F(\delta_i)\|_2$. ■

To apply the lemma, we must bound the Lipschitz constant of the matrix-valued function F (condition (1)) and $\|F(\delta_i)\|_2$, for all $i = 1, \ldots, T$ as well (condition (2)).

B.2. Regularity condition

We first establish that $\frac{\partial}{\partial\delta}\mathbb{E}\tilde{K}(\delta) = \mathbb{E}\frac{\partial}{\partial\delta}\tilde{K}(\delta)$. Since $\tilde{K}$ is a finite dimensional matrix-valued function, we verify the integrability coefficient-wise, following Sutherland and Schneider (2015)'s demonstration. Namely, without loss of generality we show

$$\left[\frac{\partial}{\partial\delta}\mathbb{E}\tilde{K}(\delta)\right]_{lm} = \mathbb{E}\frac{\partial}{\partial\delta}\left[\tilde{K}(\delta)\right]_{lm}$$

where $[A]_{lm}$ denotes the l-th row and m-th column element of the matrix A.

Proposition 4 (Differentiation under the integral sign) *Let $\mathcal{X}$ be an open subset of $\mathbb{R}^d$ and Ω be a measured space. Suppose that the function $f : \mathcal{X} \times \Omega \to \mathbb{R}$ verifies the following conditions:*

- *$f(x, \omega)$ is a measurable function of ω for each x in $\mathcal{X}$.*
- *For almost all ω in Ω, the derivative $\partial f(x, \omega)/\partial x_i$ exists for all x in $\mathcal{X}$.*
- *There is an integrable function $\Theta : \Omega \to \mathbb{R}$ such that $|\partial f(x, \omega)/\partial x_i| \leq \Theta(\omega)$ for all x in $\mathcal{X}$.*

Then

$$\frac{\partial}{\partial x_i}\int_\Omega f(x, \omega)d\omega = \int_\Omega \frac{\partial}{\partial x_i} f(x, \omega)d\omega.$$

Define the function $\tilde{G}^{i,l,m}_{x,y}(t, \omega) : \mathbb{R} \times \Omega \to \mathbb{R}$ by $\tilde{G}^{i,l,m}_{x,y}(t, \omega) = \left[\tilde{K}(x + te_i - y)\right]_{lm} = \left[\tilde{G}^{i}_{x,y}(t, \omega)\right]_{lm}$, where e_i is the i-th standard basis vector. Then $\tilde{G}^{i,l,m}_{x,y}$ is integrable w.r.t. ω since

$$\int_\Omega \tilde{G}^{i,l,m}_{x,y}(t, \omega)d\omega = \mathbb{E}\left[\tilde{K}(x + te_i - y)\right]_{lm} = [K(x + te_i - y)]_{lm} < \infty.$$

Additionally for any ω in Ω, $\partial/\partial t\tilde{G}^{i,l,m}_{x,y}(t, \omega)$ exists and satisfies

$$\begin{aligned}
\mathbb{E}\left|\frac{\partial}{\partial t}\tilde{G}^{i,l,m}_{x,y}(t, \omega)\right| &= \mathbb{E}\left|\frac{1}{D}\sum_{j=1}^{D} A(\omega)_{lm}\left(\sin\langle y, \omega_j\rangle\frac{\partial}{\partial t}\sin(\langle x, \omega_j\rangle + t\omega_{ij}) + \cos\langle y, \omega_j\rangle\frac{\partial}{\partial t}\cos(\langle x, \omega_j\rangle + t\omega_{ij})\right)\right| \\
&= \mathbb{E}\left|\frac{1}{D}\sum_{j=1}^{D} A(\omega)_{lm}\left(\omega_{ji}\sin\langle y, \omega_j\rangle\sin(\langle x, \omega_j\rangle + t\omega_{ji}) - \omega_{ji}\cos\langle y, \omega_j\rangle\cos(\langle x, \omega_j\rangle + t\omega_{ji})\right)\right| \\
&\leq \mathbb{E}\left[\frac{1}{D}\sum_{j=1}^{D}|A(\omega)_{lm}\omega_{ji}\sin\langle y, \omega_j\rangle\sin(\langle x, \omega_j\rangle + t\omega_{ji})| + |A(\omega)_{lm}\omega_{ji}\cos\langle y, \omega_j\rangle\cos(\langle x, \omega_j\rangle + t\omega_{ji})|\right] \\
&\leq \mathbb{E}\left[\frac{1}{D}\sum_{j=1}^{D} 2|A(\omega)_{lm}\omega_{ji}|\right].
\end{aligned}$$

Hence

$$\mathbb{E}\left|\frac{\partial}{\partial t}\tilde{G}^i_{x,y}(t,\omega)\right| \leq 2\mathbb{E}\left[\|\omega \otimes A(\omega)\|_1\right].$$

which is assumed to exist since in finite dimensions all norms are equivalent and $\mathbb{E}_\mu\left[\|\omega\|_2^2\|A(\omega)\|_2^2\right]$ is assume to exists. Thus applying proposition 4 we have $\left[\frac{\partial}{\partial\delta_i}\mathbb{E}\tilde{K}(\delta)\right]_{lm} = \mathbb{E}\frac{\partial}{\partial\delta_i}\left[\tilde{K}(\delta)\right]_{lm}$ The same holds for y by symmetry. Combining the results for each component x_i and for each element lm, we get that $\frac{\partial}{\partial\delta}\mathbb{E}\tilde{K}(\delta) = \mathbb{E}\frac{\partial}{\partial\delta}\tilde{K}(\delta)$.

B.3. Bounding the Lipschitz constant

Since F is differentiable, $L_F = \left\|\frac{\partial F}{\partial\delta}(\delta^*)\right\|_2$ where $\delta^* = \arg\max_{\delta\in\mathcal{D}_\mathcal{C}}\left\|\frac{\partial F}{\partial\delta}(\delta)\right\|_2$.

$$\begin{aligned}
\mathbb{E}_{\mu,\delta^*}\left[L_F^2\right] &= \mathbb{E}_{\mu,\delta^*}\left\|\frac{\partial\tilde{K}}{\partial\delta}(\delta^*) - \frac{\partial K_0}{\partial\delta}(\delta^*)\right\|_2^2 \\
&\leq \mathbb{E}_{\delta^*}\left[\mathbb{E}_\mu\left\|\frac{\partial\tilde{K}}{\partial\delta}(\delta^*)\right\|_2^2 - 2\left\|\frac{\partial K_0}{\partial\delta}(\delta^*)\right\|_2\mathbb{E}_\mu\left\|\frac{\partial\tilde{K}}{\partial\delta}(\delta^*)\right\|_2 + \left\|\frac{\partial K_0}{\partial\delta}(\delta^*)\right\|_2^2\right]
\end{aligned}$$

Using Jensen's inequality $\left\|\mathbb{E}_\mu\frac{\partial\tilde{K}}{\partial\delta}(\delta^*)\right\|_2 \leq \mathbb{E}_\mu\left\|\frac{\partial\tilde{K}}{\partial\delta}(\delta^*)\right\|_2$ and $\frac{\partial}{\partial\delta}\mathbb{E}\tilde{K}(\delta) = \mathbb{E}\frac{\partial}{\partial\delta}\tilde{K}(\delta)$. (see section B.2), $\mathbb{E}_\mu\frac{\partial\tilde{K}}{\partial\delta}(\delta^*) = \frac{\partial}{\partial\delta}\mathbb{E}_\mu\tilde{K}(\delta^*) = \frac{\partial K_0}{\partial\delta}(\delta^*)$ thus

$$\begin{aligned}
\mathbb{E}_{\mu,\delta^*}\left[L_F^2\right] &\leq \mathbb{E}_{\delta^*}\left[\mathbb{E}_\mu\left\|\frac{\partial\tilde{K}}{\partial\delta}(\delta^*)\right\|_2^2 - 2\left\|\frac{\partial K_0}{\partial\delta}(\delta^*)\right\|_2^2 + \left\|\frac{\partial K_0}{\partial\delta}(\delta^*)\right\|_2^2\right] \\
&= \mathbb{E}_{\mu,\delta^*}\left\|\frac{\partial\tilde{K}}{\partial\delta}(\delta^*)\right\|_2^2 - \mathbb{E}_{\delta^*}\left\|\frac{\partial K_0}{\partial\delta}(\delta^*)\right\|_2^2 \\
&\leq \mathbb{E}_{\mu,\delta^*}\left\|\frac{\partial\tilde{K}}{\partial\delta}(\delta^*)\right\|_2^2 \\
&= \mathbb{E}_{\mu,\delta^*}\left\|\frac{\partial}{\partial\delta^*}\cos\langle\delta^*,\omega\rangle A(\omega)\right\|_2^2 \\
&= \mathbb{E}_{\mu,\delta^*}\|-\omega\sin(\langle\delta^*,\omega\rangle)\otimes A(\omega)\|_2^2 \\
&\leq \mathbb{E}_\mu\left[\|\omega\|_2^2\|A(\omega)\|_2^2\right] := \sigma_p^2
\end{aligned}$$

Eventually applying Markov's inequality yields

$$\mathbb{P}\left\{L_F \geq \frac{\epsilon}{2r}\right\} = \mathbb{P}\left\{L_F^2 \geq \left(\frac{\epsilon}{2r}\right)^2\right\} \leq \sigma_p^2\left(\frac{2r}{\epsilon}\right)^2. \tag{S2}$$

B.4. Bounding F on a given anchor point δ_i

To bound $\|F(\delta_i)\|_2$, Hoeffding inequality devoted to matrix concentration Mackey et al. (2014) can be applied. We prefer here to turn to tighter and refined inequalities such

as Matrix Bernstein inequalities (Sutherland and Schneider (2015) also pointed that for the scalar case). If we had bounded ORFF, we could use the following Bernstein matrix concentration inequality proposed in Tropp (2012); Koltchinskii et al. (2013).

Theorem 5 (Bounded non-commutative Bernstein) *Verbatim from Theorem 3 of Koltchinskii et al. (2013), consider a sequence $(X_j)_{j=1}^D$ of D independent Hermitian $p \times p$ random matrices that satisfy $\mathbb{E}X_j = 0$ and suppose that for some constant $U > 0$, $\|X_j\|_2 \leq U$ for each index j. Denote $B = \left\|\mathbb{E}[X_1^2 + \ldots X_D^2]\right\|_2$. Then, for all $\epsilon \geq 0$,*

$$\mathbb{P}\left\{\left\|\sum_{j=1}^{D} X_j\right\| \geq \epsilon\right\} \leq p \exp\left(-\frac{\epsilon^2}{2B + 2U\epsilon/3}\right)$$

However, to cover the general case including unbounded ORFFs like curl and div-free ORFFs, we choose a version of Bernstein matrix concentration inequality proposed in Koltchinskii et al. (2013) that allow to consider matrices are not uniformly bounded but have subexponential tails. In the following we use the notion of Orlicz norm to bound random variable by their tail behavior rather than their value.

Definition 6 (Orlicz norm) *We follow the definition given by Koltchinskii et al. (2013). Let $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ be a non-decreasing convex function with $\psi(0) = 0$. For a random variable X on a measured space $(\Omega, \mathcal{T}(\Omega), \mu)$, $\|X\|_\psi := \inf\{C > 0 \mid \mathbb{E}[\psi(|X|/C)] \leq 1\}$.*

We now fix $\psi(t) = \psi_1(t) = \exp(t) - 1$. We also introduce two technical lemmas related to Orlicz norm. The first one relates the ψ_1-Orlicz norm to the moment generating function (MGF).

Lemma 7 *Let X be a random variable with a strictly monotonic moment-generating function. We have $\|X\|_{\psi_1}^{-1} = MGF_{|X|}^{-1}(2)$.*

Proof. We have

$$\|X\|_{\psi_1} = \inf\{C > 0 \mid \mathbb{E}[\exp(|X|/C)] \leq 2\} \quad = \frac{1}{\sup\left\{C > 0 \;\middle|\; \mathrm{MGF}_{|X|}(C) \leq 2\right\}}$$

X has strictly monotonic moment-generating thus $C^{-1} = \mathrm{MGF}_{|X|}^{-1}(2)$. Hence $\|X\|_{\psi_1}^{-1} = \mathrm{MGF}_{|X|}^{-1}(2)$. ■

The second lemma gives the Orlicz norm of a positive constant.

Lemma 8 *If $a \in \mathbb{R}_+$ then $\|a\|_{\psi_1} = \frac{a}{\ln(2)} < 2a$.*

Proof. We consider a as a positive constant random variable, whose MGF is $\mathrm{MGF}_a(t) = \exp(at)$. From lemma 7, $\|a\|_{\psi_1} = \frac{1}{\mathrm{MGF}_X^{-1}(2)}$. Then $\mathrm{MGF}_{|a|}^{-1}(2) = \frac{\ln(2)}{|a|}$, $a \neq 0$. If $a = 0$ then $\|a\|_{\psi_1} = 0$ by definition of a norm. Thus $\|a\|_{\psi_1} = \frac{a}{\ln(2)}$. ■

We now turn our attention to Koltchinskii et al.'s theorem to bound F with high probability on the anchors.

Theorem 9 (Unbounded non-commutative Bernstein) *Verbatim from Theorem 4 of Koltchinskii et al. (2013). Consider a sequence $(X_j)_{j=1}^D$ of D independent Hermitian $p \times p$ random matrices, such that $\mathbb{E}X_j = 0$ for all $j = 1, \dots, D$. Define*

$$F := \sum_{j=1}^{D} X_j \quad \text{and} \quad B := \left\| \mathbb{E}\left[\sum_{j=1}^{D} X_j^2 \right] \right\|_2 .$$

Suppose that,

$$M = 2 \max_{1 \le j \le D} \left\| \|X_j\|_2 \right\|_\psi$$

Let $\Delta \in \left]0; \frac{2}{e-1}\right[$ *and*

$$\bar{U} := M \log\left(\frac{2}{\Delta} \frac{M^2}{B^2} + 1 \right)$$

Then, for $\epsilon \bar{U} \le (e-1)(1+\Delta)B$,

$$\mathbb{P}\left\{ \|F\|_2 \ge \epsilon \right\} \le 2p \exp\left(-\frac{\epsilon^2}{2(1+\Delta)B + 2\epsilon\bar{U}/3} \right) \tag{S3}$$

and for $\epsilon \bar{U} > (e-1)(1+\Delta)B$,

$$\mathbb{P}\left\{ \|F\|_2 \ge \epsilon \right\} \le 2p \exp\left(-\frac{\epsilon}{(e-1)\bar{U}} \right). \tag{S4}$$

Let $\psi = \psi_1$. To use this theorem, we set: $X_j = F^j(\delta_i)$. We have indeed: $\mathbb{E}_\mu[F^j(\delta_i)] = 0$ since $\tilde{K}(\delta_i)$ is the Monte-Carlo approximation of $K_0(\delta_i)$ and the matrices $F^j(\delta_i)$ are Hermitian. We assume we can bound all the Orlicz norms of the $F^j(\delta_i) = \frac{1}{D}(\tilde{K}^j(\delta_i) - K_0(\delta_i))$. In the following we use a constant m such that $m_i = DM$. Using lemma 8 and the sub-additivity of the $\|.\|_2$ and $\|.\|_{\psi_1}$ norm,

$$\begin{aligned}
m_i &= 2D \max_{1 \le j \le D} \left\| \left\| F^j(\delta_i) \right\|_2 \right\|_{\psi_1} \\
&\le 2 \max_{1 \le j \le D} \left\| \left\| \tilde{K}^j(\delta_i) \right\|_2 \right\|_{\psi_1} + 2 \| \|K_0(\delta_i)\|_2 \|_{\psi_1} \\
&< 4 \max_{1 \le j \le D} \left\| \|A(\omega_j)\|_2 \right\|_{\psi_1} + 4 \|K_0(\delta_i)\|_2 \\
&= 4(\| \|A(\omega)\|_2 \|_{\psi_1} + \|K_0(\delta_i)\|_2)
\end{aligned}$$

We define $\bar{u}_i = D\bar{U}$ and $b_i = DB$. Then $\bar{u}_i$ can be re-written using m_i and D:

$$\bar{u}_i = m_i \log\left(\frac{2}{\Delta} \frac{m_i^2}{b_i^2} + 1 \right) \text{ and } b_i = D \left\| \mathbb{E}_\mu \left[\sum_{j=1}^{D} F^j(\delta_i)^2 \right] \right\|_2 = D \mathbb{V}_\mu \left[\tilde{K}(\delta_i)^2 \right].$$

Then, we get for all $i = 1, \dots, T$:

$$\mathbb{P}\left\{ \|F(\delta_i)\|_2 \ge \epsilon \right\} \le 2p \begin{cases} \exp\left(-\frac{D\epsilon^2}{2(1+\Delta)b_i + 2\epsilon\bar{u}_i/3} \right) & \text{if } \epsilon\bar{u}_i \le (e-1)(1+\Delta)b_i, \\ \exp\left(-\frac{D\epsilon}{(e-1)\bar{u}_i} \right) & \text{otherwise.} \end{cases} \tag{S5}$$

To simplify the equation we take $\Delta = 1$, thus

$$\mathbb{P}\left\{\|F(\delta_i)\|_2 \geq \epsilon\right\} \leq 2p \begin{cases} \exp\left(-\frac{D\epsilon^2}{4b_i + 2\epsilon\bar{u}_i/3}\right) & \text{if } \epsilon\bar{u}_i \leq (e-1)2b_i, \\ \exp\left(-\frac{D\epsilon}{(e-1)\bar{u}_i}\right) & \text{otherwise.} \end{cases} \tag{S6}$$

To unify the bound on each anchor we define two constant:

$$m = 4(\|\|A(\omega)\|_2\|_{\psi_1} + \sup_{\delta\in\mathcal{D}_\mathcal{C}} \|K_0(\delta)\|_2) \geq \max_{i=1,\dots T} m_i$$
$$b = \sup_{\delta\in\mathcal{D}_\mathcal{C}} D\mathbb{V}_\mu\left[\tilde{K}(\delta)\right] \geq \max_{i=1,\dots T} b_i.$$

B.5. Union bound

Then take the union bound over the centers of the ϵ-net:

$$\mathbb{P}\left\{\bigcup_{i=1}^{D}\|F(\delta_i)\|_2 \geq \frac{\epsilon}{2}\right\} \leq 2Tp \begin{cases} \exp\left(-\frac{\epsilon^2 D}{8\left(2b + \frac{2\epsilon}{6}\bar{u}\right)}\right) & \text{if } \epsilon\bar{u} \leq (e-1)2b \\ \exp\left(-\frac{\epsilon D}{2(e-1)\bar{u}}\right) & \text{otherwise.} \end{cases} \tag{S7}$$

B.5.1. Optimizing over r

Combining eq. (S7) and eq. (S2) yields

$$\mathbb{P}\left\{\sup_{\delta\in\mathcal{D}_\mathcal{C}}\|F(\delta)\|_2 \leq \epsilon\right\} = \mathbb{P}\left\{\|F\|_\infty \leq \epsilon\right\} \geq 1 - \kappa_1 r^{-d} - \kappa_2 r^2,$$

with

$$\kappa_2 = 4\sigma_p^2\epsilon^{-2} \quad \text{and} \quad \kappa_1 = 2p(4|\mathcal{C}|)^d \begin{cases} \exp\left(-\frac{\epsilon^2 D}{16\left(b + \frac{\epsilon}{6}\bar{u}\right)}\right) & \text{if } \epsilon\bar{u} \leq 2(e-1)b \\ \exp\left(-\frac{\epsilon D}{2(e-1)\bar{u}}\right) & \text{otherwise.} \end{cases}$$

we choose r such that $d\kappa_1 r^{-d-1} - 2\kappa_2 r = 0$, i.e. $r = \left(\frac{d\kappa_1}{2\kappa_2}\right)^{\frac{1}{d+2}}$. Eventually let $C'_d = \left(\left(\frac{d}{2}\right)^{\frac{-d}{d+2}} + \left(\frac{d}{2}\right)^{\frac{2}{d+2}}\right)$. The bound becomes

$$\begin{aligned}
\mathbb{P}\left\{\|F\|_\infty \geq \epsilon\right\} &\leq C'_d \kappa_1^{\frac{2}{d+2}} \kappa_2^{\frac{d}{d+2}} \\
&= C'_d \left(4\sigma_p^2\epsilon^{-2}\right)^{\frac{d}{d+2}} \left(2p(4|\mathcal{C}|)^d \begin{cases} \exp\left(-\frac{\epsilon^2 D}{16\left(b + \frac{\epsilon}{6}\bar{u}\right)}\right) & \text{if } \epsilon\bar{u} \leq 2(e-1)b \\ \exp\left(-\frac{\epsilon D}{2(e-1)\bar{u}}\right) & \text{otherwise} \end{cases}\right)^{\frac{2}{d+2}} \\
&= pC'_d 2^{\frac{2+4d+2d}{d+2}} \left(\frac{\sigma_p|\mathcal{C}|}{\epsilon}\right)^{\frac{2d}{d+2}} \begin{cases} \exp\left(-\frac{\epsilon^2}{8(d+2)\left(b + \frac{\epsilon}{6}\bar{u}\right)}\right) & \text{if } \epsilon\bar{u} \leq 2(e-1)b \\ \exp\left(-\frac{\epsilon}{(d+2)(e-1)\bar{u}}\right) & \text{otherwise} \end{cases} \\
&= pC'_d 2^{\frac{6d+2}{d+2}} \left(\frac{\sigma_p|\mathcal{C}|}{\epsilon}\right)^{\frac{2}{1+2/d}} \begin{cases} \exp\left(-\frac{\epsilon^2}{8(d+2)\left(b + \frac{\epsilon}{6}\bar{u}\right)}\right) & \text{if } \epsilon\bar{u} \leq 2(e-1)b \\ \exp\left(-\frac{\epsilon}{(d+2)(e-1)\bar{u}}\right) & \text{otherwise.} \end{cases}
\end{aligned}$$

Conclude the proof by taking $C_{d,p} = pC'_d 2^{\frac{6d+2}{d+2}}$.

C. Proof of proposition 12

We recall the notations $\delta = x - z$, $\tilde{K}(x,z) = \tilde{\Phi}(x)\tilde{\Phi}(z)$, $\tilde{K}^j(x,z) = \Phi_x(\omega_j)\Phi_z(\omega_j)$ and $K_0(\delta) = K(x,z)$.

Proof. We fix $\delta \in \mathcal{D}_\mathcal{C}$. From the definition of the variance, using the fact that the vectors ω_j are i.i.d. random variables,

$$\begin{aligned}
\mathbb{V}_\mu\left[\tilde{K}_0(\delta)\right] &= \mathbb{E}_\mu\left[\frac{1}{D}\sum_{j=1}^{D}\tilde{K}^j(\delta) - K_0(\delta)\right]^2 \\
&= \frac{1}{D^2}\sum_{j=1}^{D}\mathbb{E}_\mu\left[\left(\tilde{K}^j(\delta) - K_0(\delta)\right)^2\right] \\
&= \frac{1}{D}\mathbb{E}_\mu\left[\left(\tilde{K}^j(\delta)^2 - \tilde{K}^j(\delta)K_0(\delta) - K_0(\delta)\tilde{K}^j(\delta) + K_0(\delta)^2\right)\right].
\end{aligned}$$

From the definition of $\tilde{K}^j$, $\mathbb{E}_\mu[\tilde{K}^j(\delta)] = K_0(\delta)$, which leads to

$$\left\|\mathbb{V}_\mu\left[\tilde{K}_0(\delta)\right]\right\|_2 = \frac{1}{D}\left\|\mathbb{E}_\mu\left(\tilde{K}^j(\delta)^2 - K_0(\delta)^2\right)\right\|_2.$$

A trigonometric identity gives us $(\cos\langle\omega,\delta\rangle)^2 = \frac{1}{2}(\cos\langle\omega,2\delta\rangle + \cos\langle\omega,0\rangle)$

$$\begin{aligned}
\left\|\mathbb{V}_\mu\left[\tilde{K}_0(\delta)\right]\right\|_2 &= \frac{1}{D}\left\|\frac{1}{2}\mathbb{E}_\mu\left[(\cos\langle\omega,2\delta\rangle + \cos\langle\omega,0\rangle)A(\omega)^2\right] - K_0(\delta)^2\right\|_2 \\
&= \frac{1}{2D}\left\|\mathbb{E}_\mu\left[(\cos\langle\omega,2\delta\rangle + \cos\langle\omega,0\rangle)A(\omega)^2\right] - \frac{2}{D}K_0(\delta)^2\right\|_2.
\end{aligned}$$

Moreover, we write the expectation of a matrix product coefficient-wise: $\forall \ell, m \in \{1,\dots,p\}$,

$$\begin{aligned}
&\mathbb{E}_\mu\left[\left(\cos\langle\omega,2\delta\rangle A(\omega)^2\right)\right]_{\ell m} = \\
&\qquad\sum_{r=1}^{p}\mathbb{E}_\mu\left[\cos\langle\omega,2\delta\rangle A(\omega)\right]_{\ell r}\mathbb{E}_\mu\left[A(\omega)\right]_{rm} + \sum_{r=1}^{p}\mathrm{Cov}_\mu[\cos\langle\omega,2\delta\rangle A(\omega)_{\ell r}, A(\omega)_{rm}].
\end{aligned}$$

Thus,

$$\mathbb{E}_\mu\left[\left(\cos\langle\omega,2\delta\rangle A(\omega)^2\right)\right] = \mathbb{E}_\mu[\cos\langle\omega,2\delta\rangle A(\omega)]\mathbb{E}_\mu[A(\omega)] + \Sigma^{\cos} = K_0(2\delta)\mathbb{E}_\mu[A(\omega)] + \Sigma^{\cos}$$

where the random matrix $\Sigma^{\cos}$ is defined by: $\Sigma^{\cos}_{\ell m} = \sum_{r=1}^{p}\mathrm{Cov}_\mu[\cos\langle\omega,2\delta\rangle A(\omega)_{\ell r}, A(\omega)_{rm}]$. Similarly, we get $\mathbb{E}_\mu\left[\cos\langle\omega,0\rangle A(\omega)^2\right] = K_0(0)\mathbb{E}_\mu\left[A(\omega)\right] + \Sigma^{\cos}$. Therefore,

$$\begin{aligned}
\left\|\mathbb{V}_\mu\left[\tilde{K}_0(\delta)\right]\right\|_2 &= \frac{1}{2D}\left\|(K_0(2\delta) + K_0(0))\mathbb{E}_\mu[A(\omega)] - 2K_0(\delta)^2 + 2\Sigma^{\cos}\right\|_2 \\
&\leq \frac{1}{2D}\left[\left\|\left(K_0(2\delta) + K_0(0)\right)\mathbb{E}_\mu[A(\omega)] - 2K_0(\delta)^2\right\|_2 + 2\|\mathbb{V}_\mu[A(\omega)]\|_2\right],
\end{aligned}$$

using $\|\Sigma^{\cos}\|_2 \leq \|\mathbb{V}_\mu[A(\omega)]\|_2$. ∎

D. Additional information and results

D.1. Implementation detail

For each $\omega_j \sim \mu$, let $B(\omega_j)$ be a p by p' matrix such that $B(\omega_j)B(\omega_j)^* = A(\omega_j)$. In practice, making a prediction $y = h(x)$ using directly the formula $h(x) = \tilde{\Phi}(x)^*\theta$ is prohibitive. Indeed, if $\Phi(x) = \bigoplus_{j=1}^{D} \exp(-i\langle x, \omega_j\rangle)B(\omega_j)^*B(\omega_j)$, it would cost $O(Dp'p)$ operation to make a prediction, since $\tilde{\Phi(x)}$ is a Dp' by p matrix.

D.1.1. Minimizing Eq. 11 in the main paper

Recall we want to minimize

$$\theta_* = \underset{\theta\in\mathbb{R}^{p'D}}{\arg\min} \left\| \tilde{\phi}(X)^*\theta - Y \right\|^2 + \lambda\|\theta\|^2. \tag{S8}$$

The idea is to replace the expensive computation of the matrix-vector product by $\tilde{\Phi}(X)^*\theta$ by a cheaper linear operator P_x such that $\tilde{\Phi}(X)^*\theta = P_x\theta$. In other word, we minimize:

$$\theta_* = \underset{\theta\in\mathbb{R}^{p'D}}{\arg\min} \|P_X\theta - Y\|^2 + \lambda\|\theta\|^2. \tag{S9}$$

Among many possibilities of solving eq. (S9), we focused on two types of methods:

i) Gradient based methods: to solve eq. (S9) one can iterate $\theta_{t+1} = \theta_t - \eta_t(P_X^*(P_X\theta_t - y) + \lambda\theta_t)$. replace P_X by P_{x_t}, where x_t is a random sample of X at iteration T to perform a stochastic gradient descent.

ii) Linear methods: since the optimization problem defined in eq. (S9) is convex, one can find a solution to the first order condition, namely θ_* such that $(P_X^*P_X)\theta_* = P_X^*y$. Many Iterative solvers able to solve such linear system are available, such as Sonneveld and van Gijzen (2008) or Fong and Saunders (2011).

D.1.2. Defining efficient linear operators

Decomposable kernel Recall that for the decomposable kernel $K_0(\delta) = k_0(\delta)A$ where k_0 is a scalar shift-invariant kernel, $A(\omega_j) = A = BB^*$ and

$$\tilde{\Phi}(x) = \bigoplus_{j=1}^{D} \exp(-i\langle x, \omega_j\rangle)B^* = \tilde{\phi}(x) \otimes B^*$$

where $\tilde{\phi}(x) = \bigoplus_{j=1}^{D} \exp(-i\langle x, \omega_j\rangle)$ is the RFF corresponding to the scalar kernel k_0. Hence $h(x)$ can be rewritten $h(x) = (\tilde{\phi}(x) \otimes B^*)^*\Theta = \mathrm{vec}(\tilde{\phi}(x)^*\Theta B^*)$, where Θ is a D by p' matrix such that $\mathrm{vec}(\Theta) = \Theta$. Eventually we define the following linear (in θ) operator: $P_x^{\mathrm{dec}} : \theta \mapsto \mathrm{vec}(\tilde{\phi}(x)^*\Theta B^*)$. Then $h(x) = P_x^{\mathrm{dec}}\theta = \tilde{\Phi}(x)^*\theta$. Using this formulation, it only costs $O(Dp'+p'p)$ operations to make a prediction. If $B = I_d$ it reduces to $O(Dp)$. Moreover this formulation cuts down memory consumption from $O(Dp'p)$ to $O(D + p'p)$.

Curl-free kernel For the Gaussian curl-free kernel we have, $K_0(\delta) = -\nabla\nabla^T k_0(\delta)$ and the associated feature map is $\Phi(x) = \bigoplus_{j=1}^{D} \exp(-i\langle x, \omega_j\rangle)\omega_j^*$. In the same spirit we can define a linear operator

$$P_x^{\text{curl}} : \theta \mapsto \text{vec}\left(\sum_{j=1}^{D} \tilde{\phi}(x)_j^* \Theta_j \omega_j\right),$$

such that $h(x) = P_x^{\text{curl}}\theta = \tilde{\Phi}(x)^*\theta$. Here the computation time for a prediction is $O(Dp)$ and uses $O(D)$ memory.

Div-free kernel For the Gaussian divergence-free kernel, $K_0(\delta) = (\nabla\nabla^T - I\Delta)k_0(\delta)$ and $\Phi(x) = \bigoplus_{j=1}^{D} \exp(-i\langle x, \omega_j\rangle)(I - \omega_j^*\omega_j)^{1/2}$. Hence, we can define a linear operator

$$P_x^{\text{div}} : \theta \mapsto \text{vec}\left(\sum_{j=1}^{D} \tilde{\phi}(x)_j^* \Theta_j (I_d - \omega_j\omega_j^*)^{1/2}\right),$$

such that $h(x) = P_x^{\text{div}}\theta = \tilde{\Phi}(x)^*\theta$. Here the computation time for a prediction is $O(Dp^2)$ and uses $O(Dp^2)$ memory.

Feature map	$P_x : \theta \mapsto$	$P_x^* : y \mapsto$	$P_x^* P_x : \theta \mapsto$
$\Phi^{dec}(x)$	$\text{vec}(\tilde{\phi}(x)^*\Theta B^*)$	$\text{vec}(\tilde{\phi}(x) y^* B)$	$\text{vec}(\tilde{\phi}(x)\tilde{\phi}(x)^*\Theta B^* B)$
$\Phi^{curl}(x)$	$\text{vec}\left(\sum_{j=1}^{D} \tilde{\phi}(x)_j^*\Theta_j\omega_j\right)$	$\bigoplus_{j=1}^{D} \tilde{\phi}(x)_j y^*\omega_j$	$\text{vec}\left(\tilde{\phi}(x)\tilde{\phi}(x)^*\left(\bigoplus_{j=1}^{D}\Theta_j\|\omega_j\|^2\right)\right)$
$\Phi^{div}(x)$	$\text{vec}\left(\sum_{j=1}^{D} \tilde{\phi}(x)_j^*\Theta_j(I_d - \omega_j\omega_j^*)^{1/2}\right)$	$\bigoplus_{j=1}^{D} \tilde{\phi}(x)_j y^*(I_d - \omega_j\omega_j^*)^{1/2}$	$\text{vec}\left(\tilde{\phi}(x)\tilde{\phi}(x)^*\left(\bigoplus_{j=1}^{D}\Theta_j(I_d - \omega_j\omega_j^*)\right)\right)$

Table 1: fast-operator for different Feature maps.

Feature map	P_x	P_x^*	$P_x^* P_x$
$\Phi^{dec}(x)$	$O(Dp + pp')$	$O(Dp' + pp')$	$O(D^2 + Dp'^2)$
$\Phi^{curl}(x)$	$O(Dp)$	$O(Dp)$	$O(D^2p + Dp)$
$\Phi^{div}(x)$	$O(Dp^2)$	$O(Dp^2)$	$O(D^2p + Dp^2)$

Table 2: Time complexity to compute different Feature maps with fast-operators (for one point x).

D.2. Simulated dataset

Approximation, synthetic data: We trained both an ORFF and an OVK model on synthetic data from $\mathbb{R}^{20} \to \mathbb{R}^4$ as described in Audiffren and Kadri (2013). In this dataset, inputs $(x_1, \ldots, x_{20})$ are generated independently and uniformly over $[0, 1]$ and the different output are computed as follows. Let $\phi(x) = (x_1^2, x_4^2, x_1x_2, x_3x_5, x_2, x_4, 1)$ and (w_i) denotes the iid copies of a 7 dimensional Gaussian distribution with zero mean and covariance equal to Diag(0.5;0.25;0.1;0.05;0.15;0.1;0.15). Then, the outputs of the different tasks are generated as $y_i = w_i\phi(x)$. We use this dataset with $p = 4$, 10^5 instances and for the train set

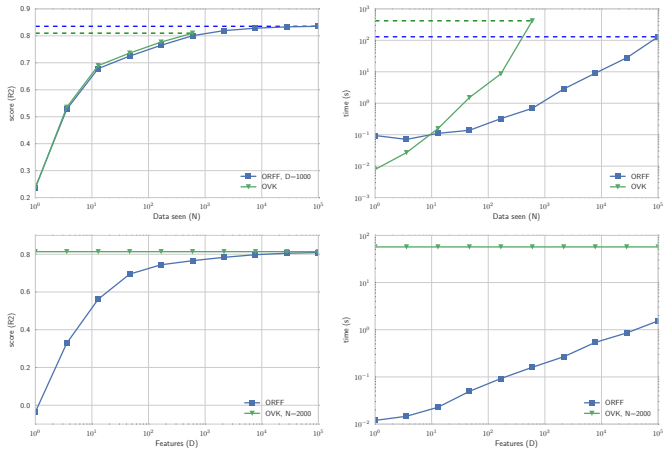

Figure S1: Decomposable kernel on synthetic data: R2 score vs number of data in the train set (N)

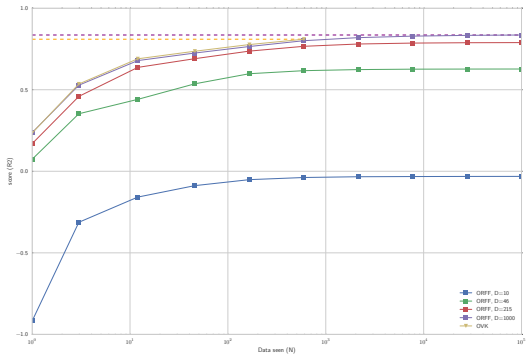

Figure S2: Decomposable kernel on synthetic data: R2 score vs number of data in the train set(N) for different number for different number of random samples (D).

and also 10^5 instances for the test set. In this setting we solved the optimisation problem for both ORFF and OVK using a L-BFGS-B. Figure S1 top row shows that for a fixed number of instance in the train set, OVK performs better than ORFF in terms of accuracy (R2). However ORFF scales better than OVK w.r.t. the number of data. ORFF is able to process more data than OVK in the same time and thus reach a better accuracy for a given amount of time. Bottom row shows that ORFF tends to reach OVK's accuracy for a fixed number of data when the number of features increase.

Acknowledgments

R. Brault was funded by University of Évry (PhD grant numbered 76391). The authors are grateful to Maxime Sangnier for his relevant comments.

References

J. Audiffren and H. Kadri. Online learning with multiple operator-valued kernels. *arXiv preprint arXiv:1311.0222*, 2013.

D. Chin-Lung Fong and M. Saunders. Lsmr: An iterative algorithm for sparse least-squares problems. *SIAM Journal on Scientific Computing*, 33(5):2950–2971, 2011.

V. Koltchinskii et al. A remark on low rank matrix recovery and noncommutative bernstein type inequalities. In *From Probability to Statistics and Back: High-Dimensional Models and Processes*, pages 213–226. Institute of Mathematical Statistics, 2013.

L. Mackey, M. Jordan, R.I. Chen, B. Farrel, and J. Tropp. Matrix concentration inequalities via the method of exchangeable pairs. *Annals of Probability*, 42:3:906–945, 2014.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *NIPS 2007*, pages 1177–1184, 2007.

P. Sonneveld and M. B. van Gijzen. Idr (s): A family of simple and fast algorithms for solving large nonsymmetric systems of linear equations. *SIAM Journal on Scientific Computing*, 31(2):1035–1062, 2008.

D. J. Sutherland and J. G. Schneider. On the error of random fourier features. In *Proc. of UAI 2015, July 12-16, 2015, Amsterdam, The Netherlands*, pages 862–871, 2015.

J. A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.