Random Feature Maps for Dot Product Kernels
Supplementary Material

Purushottam Kar and Harish Karnick
Indian Institute of Technology Kanpur, INDIA

{purushot,hk}@cse.iitk.ac.in

Abstract

This document contains detailed proofs of theorems stated in the main article entitled *Random Feature Maps for Dot Product Kernels.*

1 Proof of Theorem 1

We first recollect Schoenberg's result in its original form

Theorem 1 (Schoenberg (1942), Theorem 2). *A function $f : [-1, 1] \to \mathbb{R}$ constitutes a positive definite kernel $K : S_\infty \times S_\infty \to \mathbb{R}$, $K : (\mathbf{x}, \mathbf{y}) \mapsto f(\langle \mathbf{x}, \mathbf{y} \rangle)$ iff f is an analytic function admitting a Maclaurin expansion with only non-negative coefficients i.e. $f(x) = \sum\limits_{n=0}^{\infty} a_n x^n, a_n \geq 0, n = 0, 1, 2, \ldots$. Here $S_\infty = \{\mathbf{x} \in \mathcal{H} : \|\mathbf{x}\|_2 = 1\}$ for some Hilbert space $\mathcal{H}$.*

Corollary 2 (Theorem 1 restated). *A function $f : \mathbb{R} \to \mathbb{R}$ constitutes a positive definite kernel $K : \mathcal{B}_2(\mathbf{0}, 1) \times \mathcal{B}_2(\mathbf{0}, 1) \to \mathbb{R}$, $K : (\mathbf{x}, \mathbf{y}) \mapsto f(\langle \mathbf{x}, \mathbf{y} \rangle)$ iff f is an analytic function admitting a Maclaurin expansion with only non-negative coefficients i.e. $f(x) = \sum\limits_{n=0}^{\infty} a_n x^n, a_n \geq 0, n = 0, 1, 2, \ldots$. Here $\mathcal{B}_2(\mathbf{0}, 1) \subset \mathcal{H}$ for some Hilbert space $\mathcal{H}$.*

Proof. To see that the non-negativeness of the coefficients of the Maclaurin expansion is necessary just apply Theorem 1 to points on S_∞. Since $\{\langle \mathbf{x}, \mathbf{y} \rangle : \mathbf{x}, \mathbf{y} \in \mathcal{B}_2(\mathbf{0}, 1)\} = \{\langle \mathbf{x}, \mathbf{y} \rangle : \mathbf{x}, \mathbf{y} \in S_\infty\}$, the result extends to the general case when the points are coming from $\mathcal{B}_2(\mathbf{0}, 1)$. To see that this suffices we make use of some well known facts regarding positive definite kernels (for example refer to Schölkopf and Smola, 2002).

Fact 3. *If $K_n, n \in \mathbb{N}$ are positive definite kernels defined on some common domain then the following statements are true*

1. *$c_m K_m + c_n K_n$ is also a positive definite kernel provided $c_m, c_n \geq 0$.*
2. *$K_m K_n$ is also a positive definite kernel.*
3. *If $\lim\limits_{n \to \infty} K_n = K$ and K is continuous then K is also a positive definite kernel.*

Starting with the fact that the dot product kernel is positive definite on any Hilbert space $\mathcal{H}$, applying Fact 3.1 and Fact 3.2, we get that for every $n \in N$, the kernel $K_n(\mathbf{x}, \mathbf{y}) = \sum\limits_{i=0}^{n} a_i \langle \mathbf{x}, \mathbf{y} \rangle^i$ is positive definite. An application of Fact 3.3 along with the fact that the Maclaurin series converges uniformly within its radius of convergence then proves the result. □

2 Positive definite dot product kernels over finite dimensional spaces

As noted in the main paper, the original result of Schoenberg characterizing functions that yield a positive definite dot product kernel over finite dimensional Euclidean spaces in terms of those admitting positive Gegenbauer expansions is not very useful in practice. This is because of two reasons. Firstly, as we shall show below, functions that have non-negative Gegenbauer expansions include those that yield positive definite kernels only up to a certain dimensionality i.e. these kernels are positive definite up to $\mathbb{R}^{d_0}$ for some fixed d_0 and indefinite on all Euclidean spaces of dimensionality $d > d_0$. Secondly, from an algorithmic perspective, the Gegenbauer expansions do not seem amenable to the type of feature construction methods described in this paper - this is because Gegenbauer polynomials themselves admit negative coefficients.

The result characterizing positive definite functions over Hilbert spaces in terms of positive Maclaurin expansions on the other hand is appealing for the very same reasons - functions satisfying this stronger condition are positive definite over all finite dimensional spaces and the method readily lends itself to feature construction methods.

Lemma 4. *A function $f : \mathbb{R} \to \mathbb{R}$ yields positive definite dot product kernels over all finite dimensional Euclidean spaces iff it yields positive definite dot product kernels over Hilbert spaces.*

Proof. We shall first prove this result for the special case of ℓ_2, the Hilbert space of all square summable sequences. Schoenberg's result (Corollary 2) will then allow us to extend it to all Hilbert spaces. The *if* part follows readily from the observation that ℓ_2 contains all finite dimensional Euclidean spaces as subspaces and the fact that any kernel that is positive definite over a set is positive definite over all its subsets as well.

For the *only if* part consider any set of n points $S = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subset \ell_2$. Clearly there exists an embedding $\Phi : S \to \mathbb{R}^n$ such that for all $i, j \in [n], \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$ (note that the left and the right hand sides are inner products over different spaces). Such an embedding can be constructed, for example, by taking the Cholesky decomposition of the Gram matrix given by the inner product on ℓ_2 (the entries of the Gram matrix are finite by an application of Cauchy-Schwarz inequality).

Consider the matrix $A = [a_{ij}]$ where $a_{ij} = f\left(\langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle\right)$. Since f yields positive definite kernels over all finite dimensional Euclidean spaces, we have $A \succeq 0$. However, by the isometry of the embedding, we have $a_{ij} = f\left(\langle \mathbf{x}_i, \mathbf{x}_j \rangle\right)$. Hence, for any $n < \infty$, for any arbitrary n points, the gram matrix given by $f(\langle \cdot, \cdot \rangle)$ is positive definite (here $\langle \cdot, \cdot \rangle$ is the dot product over ℓ_2). Thus f yields a positive definite kernel over ℓ_2 as well.

To finish off the proof we now use Schoenberg's theorem to extend this result to all Hilbert spaces. If a dot product kernel is positive definite over all finite dimensional spaces then the above argument shows it to be positive definite over ℓ_2. Hence, by Corollary 2, the function f defining this kernel must have a non-negative Maclaurin's expansion. From here on an argument similar to the one used to prove the sufficiency part of Corollary 2 (using Fact 3) can be used to show that this kernel is positive definite over all Hilbert spaces.

On the other hand, if a dot product kernel is positive definite over Hilbert spaces, then we use its positive-definiteness over ℓ_2, along with the argument used in showing the *if* part above, to prove that the kernel is positive definite over all finite dimensional Euclidean spaces. □

An easy application of Corollary 2 then gives us the following result :

Corollary 5. *A function $f : \mathbb{R} \to \mathbb{R}$ yields positive definite kernels over all finite dimensional Euclidean spaces iff it is an analytic function admitting a Maclaurin expansion with only non-negative coefficients.*

However, we note that even functions that have only positive Gegenbauer expansions (and not positive Maclaurin expansions) may admit low dimensional feature maps. This is indicated by the Johnson-Lindenstrauss Lemma (for example see Indyk and Motwani, 1998) that predicts the existence of low-distortion embeddings from arbitrary Hilbert spaces (thus, in particular from the reproducing kernel Hilbert spaces of these kernels) to finite dimensional Euclidean spaces. Interestingly, it is very tempting to view the constructions of Rahimi and Recht (2007) and Vedaldi and Zisserman (2010) (among others) as algorithmic versions of the Johnson-Lindenstrauss Lemma. The challenge in all such cases, however, is to make these constructions explicit, uniform, as well as algorithmically efficient.

3 Proof of Lemma 2

Lemma 6 (Lemma 2 restated)**.** *Let $\boldsymbol{\omega} \in \mathbb{R}^d$ be a vector each of whose coordinates have been chosen pairwise independently using fair coin tosses from the set $\{-1, 1\}$ and consider the feature map $Z : \mathbb{R}^d \to \mathbb{R}$, $Z : \mathbf{x} \mapsto \boldsymbol{\omega}^\top \mathbf{x}$. Then for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, $\underset{\boldsymbol{\omega}}{\mathbb{E}} [Z(\mathbf{x})Z(\mathbf{y})] = \langle \mathbf{x}, \mathbf{y} \rangle$.*

Proof. We have $\underset{\boldsymbol{\omega}}{\mathbb{E}} [Z(\mathbf{x})Z(\mathbf{y})] = \underset{\boldsymbol{\omega}}{\mathbb{E}} \left[\boldsymbol{\omega}^\top \mathbf{x} \cdot \boldsymbol{\omega}^\top \mathbf{y}\right]$

$$\begin{aligned}
&= \underset{\boldsymbol{\omega}}{\mathbb{E}} \left[\left(\sum_{i=1}^{d} \boldsymbol{\omega}_i \mathbf{x}_i \right) \left(\sum_{i=1}^{d} \boldsymbol{\omega}_i \mathbf{y}_i \right) \right] \\
&= \underset{\boldsymbol{\omega}}{\mathbb{E}} \left[\sum_{i=1}^{d} \boldsymbol{\omega}_i^2 \mathbf{x}_i \mathbf{y}_i + \sum_{i \neq j}^{d} \boldsymbol{\omega}_i \boldsymbol{\omega}_j \mathbf{x}_i \mathbf{y}_j \right] \\
&= \sum_{i=1}^{d} \underset{\boldsymbol{\omega}}{\mathbb{E}} \left[\boldsymbol{\omega}_i^2\right] \mathbf{x}_i \mathbf{y}_i + \sum_{i \neq j}^{d} \underset{\boldsymbol{\omega}}{\mathbb{E}} \left[\boldsymbol{\omega}_i\right] \underset{\boldsymbol{\omega}}{\mathbb{E}} \left[\boldsymbol{\omega}_j\right] \mathbf{x}_i \mathbf{y}_j \\
&= \sum_{i=1}^{d} \mathbf{x}_i \mathbf{y}_i + 0 = \langle \mathbf{x}, \mathbf{y} \rangle
\end{aligned}$$

where in the third equality we have used linearity of expectation and the pairwise independence of the different coordinates of $\boldsymbol{\omega}$. The fourth equality is arrived at by using properties of the distribution. Notice that any distribution that is symmetric about zero with unit second moment can be used for sampling the coordinates of $\boldsymbol{\omega}$. This particular choice both simplifies the analysis as well as is easy to implement in practice. □

4 Proof of Lemma 3

Lemma 7 (Lemma 3 restated)**.** *Let $Z : \mathbb{R}^d \to \mathbb{R}$ be the feature map constructed above. Then for all $\mathbf{x}, \mathbf{y} \in \Omega$, we have $\mathbb{E}[Z(\mathbf{x})Z(\mathbf{y})] = K(\mathbf{x}, \mathbf{y})$ where the expectation is over the internal randomness of the feature map.*

Proof. We have $\mathbb{E}[Z(\mathbf{x})Z(\mathbf{y})]$

$$\begin{aligned}
&= \underset{N}{\mathbb{E}} \left[\underset{\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_N}{\mathbb{E}} [Z(\mathbf{x})Z(\mathbf{y})] \middle| N \right] \\
&= \underset{N}{\mathbb{E}} \left[a_N p^{N+1} \underset{\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_N}{\mathbb{E}} \left[\prod_{j=1}^{N} \boldsymbol{\omega}_j^\top \mathbf{x} \prod_{j=1}^{N} \boldsymbol{\omega}_j^\top \mathbf{y} \right] \right] \\
&= \underset{N}{\mathbb{E}} \left[a_N p^{N+1} \left(\underset{\boldsymbol{\omega}}{\mathbb{E}} \left[\boldsymbol{\omega}^\top \mathbf{x} \cdot \boldsymbol{\omega}^\top \mathbf{y}\right] \right)^N \right] \\
&= \underset{N}{\mathbb{E}} \left[a_N p^{N+1} \langle \mathbf{x}, \mathbf{y} \rangle^N \right] \\
&= \sum_{n=0}^{\infty} \frac{1}{p^{n+1}} \cdot a_n p^{n+1} \langle \mathbf{x}, \mathbf{y} \rangle^n \\
&= K(\mathbf{x}, \mathbf{y}).
\end{aligned}$$

where the first step uses the fact that the index N and the vectors $\boldsymbol{\omega}_i$ are chosen independently, the fourth step uses the fact that the vectors $\boldsymbol{\omega}_i$ are chosen independently among themselves and the fifth step uses Lemma 2. □

5 Proof of Lemma 4

Lemma 8 (Lemma 4 restated)**.** *For all $\mathbf{x}, \mathbf{y} \in \Omega$, we have $|Z(\mathbf{x})Z(\mathbf{y})| \leq pf(pR^2)$.*

Proof. Since $Z(\mathbf{x})Z(\mathbf{y}) = a_N p^{N+1} \prod_{j=1}^{N} \boldsymbol{\omega}_j^\top \mathbf{x} \prod_{j=1}^{N} \boldsymbol{\omega}_j^\top \mathbf{y}$, by Hölder's inequality we have, for all j, $\left|\boldsymbol{\omega}_j^\top \mathbf{x}\right| \leq \left\|\boldsymbol{\omega}_j\right\|_\infty \left\|\mathbf{x}\right\|_1 \leq R$ since every coordinate of $\boldsymbol{\omega}_j$ is either 1 or -1 and $\mathbf{x} \in \Omega \subseteq \mathcal{B}_1(\mathbf{0}, R)$. A similar result holds for $\left|\boldsymbol{\omega}_j^\top \mathbf{y}\right|$ as well. Thus we have $|Z(\mathbf{x})Z(\mathbf{y})| \leq a_N p^{N+1} R^{2N} \leq p \cdot \sum_{n=0}^{\infty} a_n p^n R^{2n} = pf(pR^2)$. □

6 Proof of Lemma 5

Lemma 9 (Lemma 5 restated)**.** *If a bivariate function f defined over a domain $\Omega \subseteq \mathbb{R}^d$ is L-Lipschitz in both its arguments then for every $\mathbf{x}, \mathbf{y} \in \Omega$,* $\sup_{\substack{\mathbf{x}' \in \mathcal{B}_2(\mathbf{x}, r) \cap \Omega \\ \mathbf{y}' \in \mathcal{B}_2(\mathbf{y}, r) \cap \Omega}} |f(\mathbf{x}, \mathbf{y}) - f(\mathbf{x}', \mathbf{y}')| \leq 2Lr$.

Proof. We have $|f(\mathbf{x}, \mathbf{y}) - f(\mathbf{x}', \mathbf{y}')| \leq |f(\mathbf{x}, \mathbf{y}) - f(\mathbf{x}, \mathbf{y}')| + |f(\mathbf{x}, \mathbf{y}') - f(\mathbf{x}', \mathbf{y}')| \leq L \cdot \|\mathbf{y} - \mathbf{y}'\| + L \cdot \|\mathbf{x} - \mathbf{x}'\| \leq 2Lr$ where in the second step we have used the fact that $\mathbf{x}, \mathbf{y}' \in \Omega$. □

7 Proof of Lemma 6

Lemma 10 (Lemma 6 restated)**.** *We have*

$$\begin{aligned} \sup_{\mathbf{x},\mathbf{y} \in \Omega} \|\nabla_{\mathbf{x}} K(\mathbf{x}, \mathbf{y})\| &\leq f'(R^2) \\ \sup_{\mathbf{x},\mathbf{y} \in \Omega} \|\nabla_{\mathbf{y}} K(\mathbf{x}, \mathbf{y})\| &\leq f'(R^2) \end{aligned}$$

Proof. We have $\nabla_{\mathbf{x}} K(\mathbf{x}, \mathbf{y}) = \nabla_{\mathbf{x}} \left(\sum_{n=0}^{\infty} a_n \langle \mathbf{x}, \mathbf{y} \rangle^n \right) = \sum_{n=0}^{\infty} a_n \nabla_{\mathbf{x}} \langle \mathbf{x}, \mathbf{y} \rangle^n = \mathbf{y} \sum_{n=0}^{\infty} n a_n \langle \mathbf{x}, \mathbf{y} \rangle^{n-1}$. Thus we have $\|\nabla_{\mathbf{x}} K(\mathbf{x}, \mathbf{y})\| = \left\| \mathbf{y} \sum_{n=0}^{\infty} n a_n \langle \mathbf{x}, \mathbf{y} \rangle^{n-1} \right\| \leq \sum_{n=0}^{\infty} n a_n |\langle \mathbf{x}, \mathbf{y} \rangle|^{n-1} \leq \sum_{n=0}^{\infty} n a_n (R^2)^{n-1} = f'(R^2)$ where in the third step we have used the fact that $\mathbf{x}, \mathbf{y} \in \Omega \subseteq \mathcal{B}_1(\mathbf{0}, R) \subset \mathcal{B}_2(\mathbf{0}, R)$. Similarly we can show $\sup_{\mathbf{x},\mathbf{y} \in \Omega} \|\nabla_{\mathbf{y}} K(\mathbf{x}, \mathbf{y})\| \leq f'(R^2)$. □

8 Proof of Lemma 7

Lemma 11 (Lemma 7 restated)**.** *We have*

$$\begin{aligned} \sup_{\mathbf{x},\mathbf{y} \in \Omega} \|\nabla_{\mathbf{x}} (Z_1(\mathbf{x}) Z_1(\mathbf{y}))\| &\leq p^2 R \sqrt{d} f'(pR^2) \\ \sup_{\mathbf{x},\mathbf{y} \in \Omega} \|\nabla_{\mathbf{y}} (Z_1(\mathbf{x}) Z_1(\mathbf{y}))\| &\leq p^2 R \sqrt{d} f'(pR^2) \end{aligned}$$

Proof. Since $\langle \mathbf{Z}(\mathbf{x}), \mathbf{Z}(\mathbf{y}) \rangle = \frac{1}{D} \sum_{i=1}^{D} Z_i(\mathbf{x}) Z_i(\mathbf{y})$ and $\nabla_{\mathbf{x}} \langle \mathbf{Z}(\mathbf{x}), \mathbf{Z}(\mathbf{y}) \rangle = \frac{1}{D} \sum_{i=1}^{D} \nabla_{\mathbf{x}} (Z_i(\mathbf{x}) Z_i(\mathbf{y}))$ we have $\|\nabla_{\mathbf{x}} \langle \mathbf{Z}(\mathbf{x}), \mathbf{Z}(\mathbf{y}) \rangle\| \leq \frac{1}{D} \sum_{i=1}^{D} \|\nabla_{\mathbf{x}} (Z_i(\mathbf{x}) Z_i(\mathbf{y}))\|$ by triangle inequality. Since all the Z_i feature maps are identical it would be sufficient to bound $\|\nabla_{\mathbf{x}} (Z_1(\mathbf{x}) Z_1(\mathbf{y}))\|$ and by the above calculation, the same bound would hold for $\|\nabla_{\mathbf{x}} \langle \mathbf{Z}(\mathbf{x}), \mathbf{Z}(\mathbf{y}) \rangle\|$ as well. Let $Z_1 : \mathbf{x} \mapsto \sqrt{a_N p^{N+1}} \prod_{j=1}^{N} \boldsymbol{\omega}_j^\top \mathbf{x}$ for some $N \leq k$.

Thus we can bound the quantity $\nabla_{\mathbf{x}} (Z_1(\mathbf{x}) Z_1(\mathbf{y}))$ as $\nabla_{\mathbf{x}} \left(a_N p^{N+1} \prod_{j=1}^{N} \boldsymbol{\omega}_j^\top \mathbf{x} \prod_{j=1}^{N} \boldsymbol{\omega}_j^\top \mathbf{y} \right)$ which simplifies to $\left(a_N p^{N+1} \prod_{j=1}^{N} \boldsymbol{\omega}_j^\top \mathbf{y} \right) \nabla_{\mathbf{x}} \left(\prod_{j=1}^{N} \boldsymbol{\omega}_j^\top \mathbf{x} \right)$ and further to $\left(a_N p^{N+1} \prod_{j=1}^{N} \boldsymbol{\omega}_j^\top \mathbf{y} \right) \sum_{j=1}^{N} \left(\prod_{i \neq j} \boldsymbol{\omega}_i^\top \mathbf{x} \right) \boldsymbol{\omega}_i$.

We note that for any $\boldsymbol{\omega}$ chosen, $\|\boldsymbol{\omega}\| = \sqrt{d}$. Moreover, as we have seen before, for any $\boldsymbol{\omega}$, $\sup_{\mathbf{x}\in\Omega} \left|\boldsymbol{\omega}^\top \mathbf{x}\right| \leq R$ by Hölder's inequality. Thus we can bound $\|\nabla_{\mathbf{x}} \left(Z_1(\mathbf{x})Z_1(\mathbf{y})\right)\|$ as

$$\begin{aligned}
& \left\|\left(a_N p^{N+1} \prod_{j=1}^{N} \boldsymbol{\omega}_j^\top \mathbf{y}\right) \sum_{j=1}^{N} \left(\prod_{i\neq j} \boldsymbol{\omega}_i^\top \mathbf{x}\right) \boldsymbol{\omega}_i\right\| \\
= \; & a_N p^{N+1} \left(\prod_{j=1}^{N} \left|\boldsymbol{\omega}_j^\top \mathbf{y}\right|\right) \left\|\sum_{j=1}^{N} \left(\prod_{i\neq j} \boldsymbol{\omega}_i^\top \mathbf{x}\right) \boldsymbol{\omega}_i\right\| \\
\leq \; & a_N p^{N+1} \left(\prod_{j=1}^{N} \left|\boldsymbol{\omega}_j^\top \mathbf{y}\right|\right) \sum_{j=1}^{N} \left(\prod_{i\neq j} \left|\boldsymbol{\omega}_i^\top \mathbf{x}\right|\right) \|\boldsymbol{\omega}_i\| \\
\leq \; & a_N p^{N+1} R^N \sum_{j=1}^{N} R^{N-1}\sqrt{d} = N a_N p^{N+1} R^{2N-1}\sqrt{d} \\
\leq \; & p^2 R\sqrt{d} \sum_{n=0}^{\infty} n a_n (pR^2)^{n-1} = p^2 R\sqrt{d} f'(pR^2)
\end{aligned}$$

where we have used the triangle inequality in the third step. Similarly we can show $\sup_{\mathbf{x},\mathbf{y}\in\Omega} \|\nabla_{\mathbf{y}} \left(Z_1(\mathbf{x})Z_1(\mathbf{y})\right)\| \leq p^2 R\sqrt{d} f'(pR^2)$. □

9 Alternate Feature Maps with Reduced Randomness Usage

Suppose we have a positive definite dot product kernel K defined on a domain $\Omega \subset \mathcal{B}_1(\mathbf{0}, R)$ in some Euclidean space $\mathbb{R}^d$ by a function $f(x) = \sum_{n=0}^{\infty} a_n x^n$. If we choose $k = k(\epsilon, R)$ such that $\sum_{n=0}^{k} a_n R^{2n} = f(R^2) - \epsilon$ (or select some set $S \subset \mathbb{N} \cup \{0\}$ such that $\sum_{n\in S} a_n R^{2n} = f(R^2) - \epsilon$ and $|S| = k$) and create a new kernel $\tilde{K}(\mathbf{x}, \mathbf{y}) = \sum_{n=0}^{k} a_n \langle \mathbf{x}, \mathbf{y}\rangle^n$, then the residual error $R_k = \sup_{\mathbf{x},\mathbf{y}\in\Omega} \left|\tilde{K}(\mathbf{x}, \mathbf{y}) - K(\mathbf{x}, \mathbf{y})\right| = \sup_{\mathbf{x},\mathbf{y}\in\Omega} \left|\sum_{i=k+1}^{\infty} a_n \langle \mathbf{x}, \mathbf{y}\rangle^n\right| \leq \sum_{i=k+1}^{\infty} a_n R^{2n} \leq \epsilon$ since $\Omega \subset \mathcal{B}_1(\mathbf{0}, R) \subset \mathcal{B}_2(\mathbf{0}, R)$ and $\sum_{n=0}^{\infty} a_n R^{2n} = f(R^2)$. Thus for all $\mathbf{x}, \mathbf{y} \in \Omega$, $K(\mathbf{x}, \mathbf{y}) - \epsilon \leq \tilde{K}(\mathbf{x}, \mathbf{y}) \leq K(\mathbf{x}, \mathbf{y}) + \epsilon$. Since $\tilde{K}$ also satisfies the conditions of Corollary 2, one can now obtain ϵ_1-accurate feature maps for $\tilde{K}$ using the techniques mentioned above and those feature maps would provide an $(\epsilon + \epsilon_1)$-accurate estimate to K.

10 Designing Feature Maps for Compositional Kernels

We are given a compositional kernel K_{co} defined as $K_{\text{co}}(\mathbf{x}, \mathbf{y}) = K_{\text{dp}}(K(\mathbf{x}, \mathbf{y}))$ for some positive definite dot product kernel K_{dp} and an arbitrary positive definite kernel K for which we wish to provide random feature maps. We assume that we have black-box access to a (possibly randomized) feature map selection routine $\mathcal{A}$ which when invoked, returns a feature map $W : \mathbb{R}^d \to \mathbb{R}$ for K. We first formally state the assumptions made about the kernel K and the feature maps returned by $\mathcal{A}$:

1. K is defined over some domain $\Omega \subset \mathbb{R}^d$.
2. K is bounded i.e. we have $\sup_{\mathbf{x},\mathbf{y}\in\Omega} |K(\mathbf{x}, \mathbf{y})| \leq C_K$ for some $C_K \in \mathbb{R}^+$.
3. K is Lipschitz i.e. we have $\sup_{\mathbf{x},\mathbf{y}\in\Omega} \|\nabla_{\mathbf{x}} K(\mathbf{x}, \mathbf{y})\| \leq L_K$ and $\sup_{\mathbf{x},\mathbf{y}\in\Omega} \|\nabla_{\mathbf{y}} K(\mathbf{x}, \mathbf{y})\| \leq L_K$ for some $L_K \in \mathbb{R}^+$.

Algorithm 1 Random Maclaurin Feature Maps for Compositional Kernels

```
Require: A compositional positive definite kernel $K_{\text{co}}(\mathbf{x},\mathbf{y}) = K_{\text{dp}}(K(\mathbf{x},\mathbf{y})) = f(K(\mathbf{x},\mathbf{y}))$.
Ensure: A randomized feature map $\mathbf{Z} : \mathbb{R}^d \to \mathbb{R}^D$ such that $\langle \mathbf{Z}(\mathbf{x}), \mathbf{Z}(\mathbf{y}) \rangle \approx K_{\text{co}}(\mathbf{x},\mathbf{y})$.
  Obtain the Maclaurin expansion of $f(x) = \sum_{n=0}^{\infty} a_n x^n$ by setting $a_n = \frac{f^{(n)}(0)}{n!}$.
  Fix a value $p > 1$.
  for $i = 1$ to $D$ do
    Choose a non negative integer $N \in \mathbb{N} \cup \{0\}$ with $\mathbb{P}[N = n] = \frac{1}{p^{n+1}}$.
    Get $N$ independent instantiations of the feature map for $K$ from $\mathcal{A}$ as $W_1, \ldots, W_N$.
    Let feature map $Z_i : \mathbf{x} \mapsto \sqrt{a_N p^{N+1}} \prod_{j=1}^{N} W_j(\mathbf{x})$.
  end for
  Output $\mathbf{Z} : \mathbf{x} \mapsto \frac{1}{\sqrt{D}} (Z_1(\mathbf{x}), \ldots, Z_D(\mathbf{x}))$.
```

4. W is an unbiased estimator of K i.e. for all $\mathbf{x},\mathbf{y} \in \Omega$, $\mathbb{E}[W(\mathbf{x})W(\mathbf{y})] = K(\mathbf{x},\mathbf{y})$ where the expectation is over the internal randomness of W.

5. W is a bounded feature map i.e. there exists some $C_W \in \mathbb{R}^+$ such that $\sup_{\mathbf{x}\in\Omega} |W(\mathbf{x})| \leq \sqrt{C_W}$.

6. W is Lipschitz on expectation i.e. for some $L_W \in \mathbb{R}^+$, $\sup_{\mathbf{x}\in\Omega} \mathbb{E}[\|\nabla_{\mathbf{x}} W(\mathbf{x})\|] \leq L_W$.

Our feature map construction algorithm is similar to the one used for dot product kernels. We pick a non-negative integer $N \in \mathbb{N} \cup \{0\}$ with $\mathbb{P}[N = n] = \frac{1}{p^{n+1}}$ for some fixed $p > 1$ and output the feature map $Z : \mathbb{R}^d \to \mathbb{R}$, $Z : \mathbf{x} \mapsto \sqrt{a_N p^{N+1}} \prod_{j=1}^{N} W_j(\mathbf{x})$ where $W_1, \ldots, W_N$ are independent instantiations of the feature map W associated with the kernel K. We concatenate D such feature maps to give our final feature map.

It is clear that on expectation, the product of the feature map values is equal to the value of the kernel i.e. $\mathop{\mathbb{E}}_{N,W_1,\ldots,W_N} [\langle \mathbf{Z}(\mathbf{x}), \mathbf{Z}(\mathbf{y}) \rangle] = K_{comp}(\mathbf{x},\mathbf{y})$ where $\mathbf{Z} : \mathbb{R}^d \to \mathbb{R}^D$, $\mathbf{Z} : \mathbf{x} \mapsto \frac{1}{\sqrt{D}} (Z_1(\mathbf{x}), \ldots, Z_D(\mathbf{x}))$. Yet again we expect that the concatenation of D such feature maps for a large enough D would provide us a close approximation to K_{co} with high probability. For this we first prove that our feature map is bounded.

Lemma 12. *For all* $\mathbf{x},\mathbf{y} \in \Omega$, $|Z(\mathbf{x})Z(\mathbf{y})| \leq pf(pC_W)$.

Proof. $Z(\mathbf{x})Z(\mathbf{y}) = a_N p^{N+1} \prod_{j=1}^{N} W_j(\mathbf{x}) \prod_{j=1}^{N} W_j(\mathbf{x})$. Using the bound on the feature maps we get the inequality $|Z(\mathbf{x})Z(\mathbf{y})| \leq a_N p^{N+1} C_W^N \leq pf(pC_W)$ □

Thus we have for any $\mathbf{x},\mathbf{y} \in \Omega$, $\mathbb{P}[|\langle \mathbf{Z}(\mathbf{x}), \mathbf{Z}(\mathbf{y}) \rangle - K_{\text{co}}(\mathbf{x},\mathbf{y})| \leq \epsilon]$ with probability at least $1 - 2\exp\left(-\frac{D\epsilon^2}{8C_1^2}\right)$ where $C_1 = pf(pC_W)$. We now investigate the Lipschitz properties of K_{co} and our feature map.

Lemma 13. *We have*

$$\begin{aligned} \sup_{\mathbf{x},\mathbf{y}\in\Omega} \|\nabla_{\mathbf{x}} K_{co}(\mathbf{x},\mathbf{y})\| &\leq L_K f'(C_K) \\ \sup_{\mathbf{x},\mathbf{y}\in\Omega} \|\nabla_{\mathbf{y}} K_{co}(\mathbf{x},\mathbf{y})\| &\leq L_K f'(C_K) \end{aligned}$$

Proof. $K_{\text{comp}}(\mathbf{x},\mathbf{y}) = \sum_{n=0}^{\infty} a_n K(\mathbf{x},\mathbf{y})^n$. Thus we have by linearity $\nabla_{\mathbf{x}} K_{\text{comp}}(\mathbf{x},\mathbf{y}) = \sum_{n=0}^{\infty} a_n \nabla_{\mathbf{x}} (K(\mathbf{x},\mathbf{y})^n) = \sum_{n=0}^{\infty} n a_n K(\mathbf{x},\mathbf{y})^{n-1} \nabla_{\mathbf{x}} K(\mathbf{x},\mathbf{y})$ i.e $\|\nabla_{\mathbf{x}} K_{\text{comp}}(\mathbf{x},\mathbf{y})\| \leq \|\nabla_{\mathbf{x}} K(\mathbf{x},\mathbf{y})\| \sum_{n=0}^{\infty} n a_n C_K^{n-1} \leq L_K f'(C_K)$. Similarly we have $\sup_{\mathbf{x},\mathbf{y}\in\Omega} \|\nabla_{\mathbf{y}} K_{\text{co}}(\mathbf{x},\mathbf{y})\| \leq L_K f'(C_K)$. □

We next move on to the Lipschitz properties of $\mathbf{Z}$. Since we have only made assumptions on the expected Lipschtiz properties of W, we would only be able to give guarantees on the expected Lipschitz properties of $\mathbf{Z}$. However, as we shall see, these would be sufficient to provide a uniform convergence guarantee over the entire domain Ω. As before, we find that by linearity of expectation, analyzing the expected Lipschitz properties of a single feature map Z are sufficient to guarantee, on expectation, similar properties for $\mathbf{Z}$ as well.

Lemma 14. *We have*

$$\begin{aligned}\sup_{\mathbf{x},\mathbf{y}\in\Omega} \left\|\nabla_{\mathbf{x}}\left(Z(\mathbf{x})Z(\mathbf{y})\right)\right\| &\leq L_W p^2\sqrt{C_W} f'(pC_W)\\ \sup_{\mathbf{x},\mathbf{y}\in\Omega} \left\|\nabla_{\mathbf{y}}\left(Z(\mathbf{x})Z(\mathbf{y})\right)\right\| &\leq L_W p^2\sqrt{C_W} f'(pC_W)\end{aligned}$$

Proof. Since $Z(\mathbf{x})Z(\mathbf{y}) = a_N p^{N+1}\prod_{j=1}^{N} W_j(\mathbf{x})W_j(\mathbf{y})$, by linearity we can write

$$\nabla_{\mathbf{x}} Z(\mathbf{x})Z(\mathbf{y}) = \left(a_N p^{N+1}\prod_{j=1}^{N} W_j(\mathbf{y})\right)\sum_{j=1}^{N}\left(\prod_{i\neq j} W_i(\mathbf{x})\right)\nabla_{\mathbf{x}} W_j(\mathbf{x})$$

Thus we can then write $\|\nabla_{\mathbf{x}} Z(\mathbf{x})Z(\mathbf{y})\|$ as

$$a_N p^{N+1}\left|\prod_{j=1}^{N} W_j(\mathbf{y})\right|\left\|\sum_{j=1}^{N}\left(\prod_{i\neq j} W_i(\mathbf{x})\right)\nabla_{\mathbf{x}} W_j(\mathbf{x})\right\| \leq a_N p^{N+1} C_W^{\frac{N}{2}}\sum_{j=1}^{N} C_W^{\frac{N-1}{2}}\left\|\nabla_{\mathbf{x}} W_j(\mathbf{x})\right\|$$

which gives us, by linearity of expectation and the bound on the expected Lipschitz properties of the individual estimators,

$$\begin{aligned}\mathbb{E}\left[\|\nabla_{\mathbf{x}} Z(\mathbf{x})Z(\mathbf{y})\|\right] &\leq N a_N p^{N+1} C_W^{N-\frac{1}{2}} L_W\\ &= L_W p^2\sqrt{C_W}\cdot N a_N\left(pC_W\right)^{N-1}\\ &\leq L_W p^2\sqrt{C_W} f'(pC_W)\end{aligned}$$

The other part follows similarly. □

Working as before we find that the error function $\mathcal{E}(\mathbf{x},\mathbf{y}) = \langle\mathbf{Z}(\mathbf{x}),\mathbf{Z}(\mathbf{y})\rangle - K_{\text{co}}(\mathbf{x},\mathbf{y})$ is, on expectation, L_1-Lipschitz for $L_1 = L_K f'(C_K) + L_W p^2\sqrt{C_W} f'(pC_W)$. Hence the probability that the error function will not be $\frac{\epsilon}{2r}$-Lipschitz is less than $\frac{2L_1 r}{\epsilon}$ by an application of Markov's inequality. However if this is not the case then constructing an ϵ-net at scale r over the domain Ω and ensuring that the estimator provides an $\epsilon/2$-approximation at centers of these points would ensure an ϵ-accurate estimation to the kernel on the entire domain Ω. Setting up such a net would require at most $\left(\frac{4R}{r}\right)^d$ centers if $\Omega \subseteq \mathcal{B}_1(\mathbf{0},R)$. Adding the failure probabilities of the estimator not being accurate on the ϵ-net centers to the probability of the error function not being Lipschitz gives us the total error probability of our estimator giving an inaccurate estimate over any point in the domain as $2\left(\frac{4R}{r}\right)^d\exp\left(-\frac{D\epsilon^2}{8C_1^2}\right) + \frac{2L_1 r}{\epsilon}$.

Looking at this quantity as of the form $k_1 r^{-d} + k_2 r$ and setting $r = \left(\frac{k_1}{k_2}\right)^{\frac{1}{d+1}}$ gives us the error probability as $2k_1^{\frac{1}{d+1}} k_2^{\frac{d}{d+1}} \leq \left(\frac{32RL_1}{\epsilon}\right)\exp\left(-\frac{D\epsilon^2}{8C_1^2 d}\right)$ if $\epsilon < 8RL_1$ which gives us the following theorem.

Theorem 15. *Let $\Omega \subseteq \mathcal{B}_1(\mathbf{0},R)$ be a compact subset of $\mathbb{R}^d$ and $K_{co}(\mathbf{x},\mathbf{y}) = K_{dp}(K(\mathbf{x},\mathbf{y}))$ be a compositional kernel defined on Ω satisfying the necessary boundedness and Lipschitz conditions. Assuming we have black-box access to a feature map selection algorithm for K also satisfying the necessary boundedness and Lipschitz conditions, for the feature map $\mathbf{Z}$ defined in Algorithm 1, we have $\mathbb{P}\left[\sup_{\mathbf{x},\mathbf{y}\in\Omega}\left|\langle\mathbf{Z}(\mathbf{x}),\mathbf{Z}(\mathbf{y})\rangle - K_{co}(\mathbf{x},\mathbf{y})\right| > \epsilon\right] \leq \left(\frac{32RL_1}{\epsilon}\right)\exp\left(-\frac{D\epsilon^2}{8C_1^2 d}\right)$ where $C_1 = pf(pC_W)$ and $L_1 = L_K f'(C_K) + L_W p^2\sqrt{C_W} f'(pC_W)$ for some small constant $p > 1$. Moreover, with $D = \Omega\left(\frac{dC_1^2}{\epsilon^2}\log\left(\frac{RL_1}{\epsilon\delta}\right)\right)$, one can ensure the same with probability greater than $1-\delta$.*

References

Isaac Jacob Schoenberg. Positive Definite Functions on Spheres. *Duke Mathematical Journal*, 9(1): 96–108, 1942.

Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, 2002.

Piotr Indyk and Rajeev Motwani. Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality. In *Thirtieth Annual ACM Symposium on Theory of Computing*, pages 604–613, 1998.

Ali Rahimi and Benjamin Recht. Random Features for Large-Scale Kernel Machines. In *Neural Information Processing Systems*, 2007.

Andrea Vedaldi and Andrew Zisserman. Efficient Additive Kernels via Explicit Feature Maps. In *23rd IEEE Conference on Computer Vision and Pattern Recognition*, pages 3539–3546, 2010.