The Geometry of Random Features

Krzysztof Choromanski[*1] **Mark Rowland**[*2] **Tamas Sarlos**[1] **Vikas Sindhwani**[1] **Richard E. Turner**[2] **Adrian Weller**[23]
[1]Google Brain, NY [2]University of Cambridge, UK [3]The Alan Turing Institute, UK

Abstract

We present an in-depth examination of the effectiveness of radial basis function kernel (beyond Gaussian) estimators based on orthogonal random feature maps. We show that orthogonal estimators outperform state-of-the-art mechanisms that use iid sampling under weak conditions for tails of the associated Fourier distributions. We prove that for the case of many dimensions, the superiority of the orthogonal transform can be accurately measured by a property we define called the charm of the kernel, and that orthogonal random features provide optimal (in terms of mean squared error) kernel estimators. We provide the first theoretical results which explain why orthogonal random features outperform unstructured on downstream tasks such as kernel ridge regression by showing that orthogonal random features provide kernel algorithms with better spectral properties than the previous state-of-the-art. Our results enable practitioners more generally to estimate the benefits from applying orthogonal transforms.

1 INTRODUCTION

Kernel methods are a central tool in machine learning, with many applications including classification (SVMs, Cortes and Vapnik, 1995), regression (kernel ridge regression), Gaussian processes (Rasmussen and Williams, 2005), principal component analysis, novelty detection, bioinformatics (graph kernels), predictive state representation and reinforcement learning (Ormoneit and Sen, 2002). An important drawback is poor scalability with the size of the dataset. One approach to address this problem is the popular random feature map method (Rahimi and Recht, 2007), where values of kernels are approximated by dot products of the corresponding random feature maps (RFMs), since compact RFMs lead to much more scalable models.

RFMs can be constructed more efficiently by using structured matrices, but typically at the cost of lower accuracy (Ailon and Chazelle, 2006; Hinrichs and Vybíral, 2011; Vybíral, 2011; Zhang and Cheng, 2013; Choromanski and Sindhwani, 2016; Choromanska et al., 2016; Bojarski et al., 2017). Recent results suggest that in certain settings, structured approaches based on orthogonal transforms outperform iid methods in terms of accuracy (Yu et al., 2016; Choromanski et al., 2017). These techniques also often lead to faster routines for the RFM computation if they can be discretized (Choromanski et al., 2017), yielding triple win improvements in accuracy, speed and space complexity.

These triple win methods have been used so far only in very special scenarios such as Gaussian kernel approximation in the regime of high data dimensionality (Yu et al., 2016), dimensionality reduction with modified Johnson-Lindenstrauss transform, angular kernel approximation (Choromanski et al., 2017) and cross-polytope LSH (Andoni et al., 2015). Little is known about their theoretical guarantees. The question of how broadly they may be applied is an important open problem in theory and in practice.

Until recently no theoretical results showing that with number of random features $m \ll N$, where N stands for data size, one can obtain accurate approximation of the exact kernel method for such tasks as kernel ridge regression or SVM were known. Most of the theoretical results (including all mentioned above) considered pointwise kernel approximation – the question whether these results translate (if at all) to quantities such as small empirical risk for kernel ridge regression in the setting $m \ll N$ was open. One of the first results here was proposed by Avron et al. (2017), but this considered unstructured random features. In this paper, we prove that orthogonal random features for RBF kernels provide strictly better bounds. Further, we show that this is a consequence of a more general observation that kernel algorithms based on orthogonal random features are characterized by better spectral properties than the unstructured ones. We achieve this by combining our novel pointwise guarantees with recent work by Avron et al. (2017).

For a practitioner considering an RFM for her particular kernel application, key questions include: How to evaluate the gains provided by the structured approach (including time for orthogonalization if required)? How do gains depend on the region of interest and the choice of the kernel (the high

[*]Authors contributed equally.

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84.

dimensionality setting is typically more important)? Whether pointwise gains coming from the orthogonal random features imply downstream applications gains?

We answer these questions for the prominent class of *radial basis function kernels* (RBFs), presenting the first general approach to the open problem. Our results include the earlier result of Yu et al. (2016) as a special case. An RBF $K : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ is defined by $K(\mathbf{x}, \mathbf{y}) = \phi_K(\|\mathbf{x} - \mathbf{y}\|)$, for some positive-definite (PD) function $\phi_K : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}$. RBFs include the Gaussian and Matérn kernel, and play an important role in machine learning, leading to well established architectures such as RBF networks. There are deep connections between RBFs and function approximation, regularization theory, density estimation and interpolation in the presence of noise.

We highlight the following contributions:

- In §3: Asymptotic results for fixed $\|\mathbf{x} - \mathbf{y}\|$ and large dimensionality n, and also for fixed n and small $\|\mathbf{x}-\mathbf{y}\|$. In the latter case, we show (under certain conditions) superiority of orthogonal random features relative to iid features for a large class of RBFs defined by bounded fourth moments of the corresponding Fourier distributions. In the former case, we express the benefit of orthogonality in terms of the *charm function* of the RBF at a given point $\mathbf{x} - \mathbf{y}$ (see §3 for details). We draw particular attention to Theorems 3.1 and 3.8 as key theoretical results.
- In §4: We show optimality of the random orthogonal feature method for large classes of RBFs under weak conditions regarding the geometry of the applied random feature map mechanism.
- In §5: We provide guarantees that orthogonal random features for RBFs outperform unstructured ones on such downstream tasks as kernel ridge regression.
- In §6: We explore empirically the benefits from orthogonal features for pointwise kernel approximation, Gram matrix approximation and GP regression.

2 RANDOM FOURIER FEATURES

Since an RBF kernel K is shift-invariant, by Bochner's theorem (Rahimi and Recht, 2007) there exists a finite Borel measure $\mu_K \in \mathcal{M}(\mathbb{R}^n)$ (the Fourier measure associated with K) such that

$$K(\mathbf{x}, \mathbf{y}) = \mathrm{Re}\left(\int_{\mathbb{R}^n} \exp(i\langle \mathbf{w}, \mathbf{x} - \mathbf{y}\rangle)\mu_K(\mathrm{d}\mathbf{w})\right). \quad (1)$$

In Figure 1, we recall several commonly-used RBFs and their corresponding Fourier densities, which will be used throughout the remainder of the paper.

For simplicity, we assume $\mu_K(\mathbb{R}^n) = 1$; the extension to general non-negative finite measures is straightforward.

Name	Positive-definite function
Gaussian	$\sigma^2 \exp\left(-\frac{1}{2\lambda^2} z^2\right)$
Matérn	$\sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} z\right)^\nu K_\nu\left(\sqrt{2\nu} z\right)$

Name	Fourier density
Gaussian	$\frac{\sigma^2}{(2\pi\lambda^2)^{n/2}} \exp\left(-\frac{1}{2\lambda^2}\lVert\mathbf{w}\rVert_2^2\right)$
Matérn	$\frac{\Gamma(\nu+n/2)}{\Gamma(\nu)(2\nu\pi)^{n/2}} \left(1 + \frac{1}{2\nu}\lVert\mathbf{w}\rVert^2\ \right)^{-\nu-p/2}$

Figure 1: Common RBF kernels, their corresponding positive definite functions, and their Fourier transforms.

Bochner's theorem leads to the Monte Carlo scheme for approximating values of RBFs and to the random feature map mechanism, where rows of the random matrix are taken independently at random from distribution μ_K.

Using the identity given by Bochner's theorem (Equation 1), a standard Monte Carlo approximation yields the pointwise kernel estimator

$$\widehat{K}^{\mathrm{iid}}_{m,n}(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{m} \frac{\cos(\langle \mathbf{w}_i, \mathbf{x} - \mathbf{y}\rangle)}{m} = \langle \Phi_{m,n}(\mathbf{x}), \Phi_{m,n}(\mathbf{y})\rangle, \quad (2)$$

where $\Phi_{m,n} : \mathbb{R}^n \rightarrow \mathbb{R}^{2m}$ is a random feature embedding:

$$\Phi_{m,n}(\mathbf{x}) = \left(\frac{1}{\sqrt{m}}\cos(\langle \mathbf{w}_i, \mathbf{x}\rangle), \frac{1}{\sqrt{m}}\sin(\langle \mathbf{w}_i, \mathbf{x}\rangle)\right)_{i=1}^{m},$$

for all $\mathbf{x} \in \mathbb{R}^n$, $(\mathbf{w}_i)_{i=1}^m \overset{\mathrm{iid}}{\sim} \mu_K$. Here m stands for the total number of random features used. Thus, a kernel algorithm applying a non-linear kernel K on a dataset $(\mathbf{x}_i)_{i=1}^N$ can be approximated by using the linear (inner product) kernel with the randomly embedded dataset $(\Phi(\mathbf{x}_i))_{i=1}^N$. The special linear structure of the approximation can be exploited to yield fast training algorithms (Joachims, 2006). There has been much recent work in understanding the errors incurred by random feature approximations (Sutherland and Schneider, 2015), and in speeding up the computation of the random embeddings (Le et al., 2013).

2.1 Geometrically Structured Random Fourier Features

We start by identifying some basic properties of the probability measures μ associated with an RBF. The following lemma demonstrates that a random vector $\mathbf{w}$ drawn from the corresponding Fourier measure $\mu \in \mathcal{M}(\mathbb{R}^n)$ may be decomposed as $\mathbf{w} = R\widehat{\mathbf{v}}$, where $\widehat{\mathbf{v}} \sim \mathrm{Unif}(S^{n-1})$, and $R \geq 0$ is the norm of the random vector $\mathbf{w}$.

Lemma 2.1. *If K is an RBF, then its Fourier transform $\mu \in \mathcal{M}(\mathbb{R}^n)$ is isotropic: $\mu(A) = \mu(\mathbf{M}^{-1}A)$ for all Borel sets A, and all $\mathbf{M} \in O_n$, the orthogonal group on $\mathbb{R}^n$.*

With this decomposition of the distribution of the frequency vectors in hand, we can now consider introducing geometric couplings into the joint distribution over $(\mathbf{w}_i)_{i=1}^m = (R_i \widehat{\mathbf{v}}_i)_{i=1}^m$. In particular, we shall consider couplings of the direction vectors $(\widehat{\mathbf{v}}_i)_{i=1}^m$ so that marginally each direction vector $\widehat{\mathbf{v}}_i$ is distributed uniformly over the sphere S^{n-1}, but the direction vectors are no longer independent. There are many ways in which such a coupling can be constructed; for example, direction vectors could be drawn iteratively, with the distribution of a direction vector given by a parametric distribution (such as a von-Mises-Fisher distribution), with parameters depending on previously drawn directions.

One case of particular interest is when direction vectors are conditioned to be orthogonal, which was recently introduced by Yu et al. (2016) in the case of the Gaussian kernel, defined in greater generality below.

Definition 2.2 (Orthogonal Random Features)**.** *Let $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be an RBF kernel, with associated Fourier measure $\mu_K \in \mathcal{M}(\mathbb{R}^n)$. The orthogonal random feature map $\Phi : \mathbb{R}^n \to \mathbb{R}^{2m}$ of dimension $2m = 2kn$ (for some integer $k \in \mathbb{N}$) associated with K is given by*

$$\Phi_{m,n}^{\mathrm{ort}}(\mathbf{x}) = \left(\frac{1}{\sqrt{m}} \cos(\langle \mathbf{w}_i^l, \mathbf{x} \rangle), \frac{1}{\sqrt{m}} \sin(\langle \mathbf{w}_i^l, \mathbf{x} \rangle) \right)_{l=1,i=1}^{l=k,i=n},$$

where the blocks of frequency vectors $(\mathbf{w}_{1:n}^l)_{l=1}^k$ are independent, and for each frequency vector block, the frequency vectors $\mathbf{w}_1^l, \ldots, \mathbf{w}_n^l$ are marginally distributed according to μ_K, and are jointly almost-surely orthogonal. We denote the corresponding kernel estimator as follows:

$$\begin{aligned} \widehat{K}_{m,n}^{\mathrm{ort}}(\mathbf{x}, \mathbf{y}) &= \sum_{l=1}^{k} \sum_{i=1}^{n} \frac{\cos(\langle \mathbf{w}_i^l, \mathbf{x} - \mathbf{y} \rangle)}{m} \qquad (3) \\ &= \langle \Phi_{m,n}^{\mathrm{ort}}(\mathbf{x}), \Phi_{m,n}^{\mathrm{ort}}(\mathbf{y}) \rangle . \end{aligned}$$

Henceforth we take $k = 1$. The analysis for a number of blocks $k > 1$ is completely analogous.

3 ORTHOGONAL RANDOM FEATURES FOR GENERAL RBFS AND THE CHARM FUNCTION

In this section, we establish asymptotically the benefits of the orthogonal random feature map mechanism for a large class of RBFs $K(\mathbf{x}, \mathbf{y})$. Let $\mathbf{z} = \mathbf{x} - \mathbf{y}$. We focus mainly on two regimes: (i) fixed dimensionality n and small $\|\mathbf{z}\|$; and (ii) fixed $\|\mathbf{z}\|$ and large n. We introduce the *charm* function defined in Equation (4), and explain its role in assessing the accuracy of models based on random feature maps for large n. In particular, we show that for classes of RBFs defined by positive definite functions ϕ that are not parametrized by data dimensionality, charm is always nonnegative. This observation leads to the conclusion that when in this setting, for n large enough, the orthogonal estimator outperforms the iid estimator $\widehat{K}_{m,n}^{\mathrm{iid}}$ across the entire domain provided that the tails of the corresponding Fourier distributions are not too heavy. At the outset, we highlight Theorems 3.1 and 3.8 as key theoretical results.

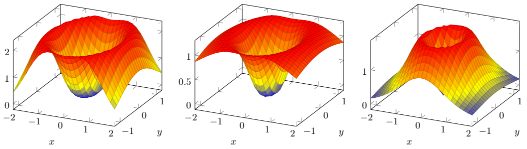


Figure 2: Plots of the charm function Ψ_K for $n = 2$ and different RBFs K. On the left: Gaussian kernel, in the middle: kernel defined by the PD function $\phi(\|\mathbf{z}\|) = \left(1 + \|\mathbf{z}\|^2\right)^{-1/2}$, on the right: kernel defined by the PD function $\phi(\|\mathbf{z}\|) = \left(1 + \|\mathbf{z}\|^2\right)^{-1}$. "Warmer" regions indicate larger gains from applying structured approach. Positive values of Ψ_K imply asymptotic superiority of the structured orthogonal estimator. All charm functions plotted are positive everywhere.

Charm. We shall show that charm plays a crucial role in understanding the behavior of orthogonal transforms for the large dimensionality regime. The charm function Ψ_K of an RBF $K(\mathbf{x}, \mathbf{y}) = \phi_K(\|\mathbf{x} - \mathbf{y}\|)$ is a function $\mathbb{R}^n \to \mathbb{R}$ defined at point $\mathbf{z} = \mathbf{x} - \mathbf{y}$ as follows:

$$\Psi_K(\mathbf{z}) = \|\mathbf{z}\|^2 \left. \frac{d^2 \phi_K^2}{dx^2} \right|_{x=\|\mathbf{z}\|} - \|\mathbf{z}\| \left. \frac{d\phi_K^2}{dx} \right|_{x=\|\mathbf{z}\|} . \qquad (4)$$

We shall see that in the large dimensionality regime, the superiority of orthogonal transforms follows from the positive sign of the charm function across the entire domain. This in turn is a consequence of the intricate connection between classes of positive definite RBFs not parametrized by data dimensionality and *completely monotone* functions. The benefits from using orthogonal transforms in comparison to state-of-the-art can be quantitatively measured by the value of the charm of the kernel at point $\mathbf{z} = \mathbf{x} - \mathbf{y}$ for large data dimensionality. Large charm values (see Figure 2) indicate regions where the mean-squared error defined as: $\mathbf{MSE}(\mathbf{x}, \mathbf{y}) = \mathbb{E}[(\widehat{K}(\mathbf{x}, \mathbf{y}) - K(\mathbf{x}, \mathbf{y}))^2]$, of the orthogonal estimator is significantly smaller than for an iid estimator and thus the geometry of the charm function across the domain gives strong guidance on the accuracy benefits of the structured approach.

3.1 The Landscape for Fixed n and Small $\|\mathbf{x} - \mathbf{y}\|$

Our main result of this section compares the mean squared error (MSE) of the iid random feature estimator based on independent sampling to the MSE of the estimator applying random orthogonal feature maps for small enough $\|\mathbf{x} - \mathbf{y}\|$.

Theorem 3.1. *Let $K : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ be an RBF and let $\mu_K \in \mathcal{M}(\mathbb{R}^n)$ be its associated Fourier measure. Suppose that $\mathbb{E}_{\mu_K}\left[\|\mathbf{w}\|^4\right] < \infty$. Then for sufficiently small $\|\mathbf{x}-\mathbf{y}\|$, we have*

$$\mathrm{MSE}(\widehat{K}^{\mathrm{iid}}_{m,n}(\mathbf{x},\mathbf{y})) > \mathrm{MSE}(\widehat{K}^{\mathrm{ort}}_{m,n}(\mathbf{x},\mathbf{y})).$$

The assumptions of the theorem above are satisfied for many classes of RBFs such as Gaussian, Matérn with smoothness parameter $\nu > 2$, and Poisson-Bessel kernels. In the Appendix we present additional results that give an explicit lower bound on the gap between the MSEs, given additional assumptions on the tail of μ.

3.2 The Landscape for Fixed $\|\mathbf{x}-\mathbf{y}\|$ and Large n

Having established asymptotic results for small $\|\mathbf{x}-\mathbf{y}\|$, we now explore the asymptotic behaviour of orthogonal features for large dimensionality n. We state our main result below, which first requires a preliminary definition.

Definition 3.2. *Let $\mathrm{M}_{\mu_n}(k,n)$ be the k-th moment of the random variable $X = \|\mathbf{w}\|_2$, for $\mathbf{w} \sim \mu_n$, where $\mu_n \in \mathcal{M}(\mathbb{R}^n)$. We say that a sequence of measures $\{\mu_n\}$ is concentrated if $\mathbb{P}[|\|\mathbf{w}\|_2^2 - \mathrm{M}_{\mu_n}(2,n)| \geq \mathrm{M}_{\mu_n}(2,n)g(n)] \leq \frac{1}{h(n)}$ for some $g(n) = o_n(1)$ and $h(n) = \omega_n(1)$.*

Note that the above is a very weak concentration condition regarding second moments, where no exponentially small upper bounds are needed. Now the charm function (4) plays a crucial role. Our key technical result, from which we will deduce several practical corrollaries, is as follows.

Theorem 3.3. *Consider a fixed positive definite radial basis function ϕ, a family of RBF kernels K, where K on $\mathbb{R}^n \times \mathbb{R}^n$ for each $n \in \mathbb{N}$ is defined as $K(\mathbf{x},\mathbf{y}) = \phi(\|\mathbf{x}-\mathbf{y}\|)$ for all $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$, and an associated concentrated sequence of Fourier measures $\{\mu_n\}_{n\in\mathbb{N}}$. Assume also that there exist constant $C > 0$ and $\xi : \mathbb{N} \rightarrow \mathbb{R}$ such that $\mathrm{M}_{\mu_n}(2k,2n) \leq (n-1)(n+1)\cdot \ldots \cdot (n+2k-3)\xi(k)$ and $\frac{|\xi(k)|}{k!} \leq C^k$ for k large enough. Then the following holds for $\|\mathbf{z}\| = \|\mathbf{x}-\mathbf{y}\| < \frac{1}{4\sqrt{C}}$:*

$$\mathrm{MSE}(\widehat{K}^{\mathrm{iid}}_{m,n}(\mathbf{x},\mathbf{y})) - \mathrm{MSE}(\widehat{K}^{\mathrm{ort}}_{m,n}(\mathbf{x},\mathbf{y})) = \frac{m-1}{m}\left(\frac{1}{8n}\Psi_K(\mathbf{z}) + o(n^{-1})\right), \tag{5}$$

where Ψ_K is defined as in Equation (4) and m is the number of random features used. A tight upper bound on the $o(n^{-1})$ term and a strengthened version of the above theorem is given in the Appendix.

Theorem 3.3 leads to many important corollaries, as we show below. In particular, we highlight that the charm function Ψ_K associated with the kernel K is central in determining the relative performance of orthogonal random features and iid features in high dimensions, due to its place in Equation (5). As special cases, Theorem 3.3 implies all earlier theoretical results for orthogonal random features for a Gaussian kernel (Yu et al., 2016).

Corollary 3.4. *If K is a Gaussian kernel then for any fixed $\|\mathbf{z}\| > 0$ and n large enough the orthogonal random feature map outperforms the iid random feature map in MSE. This is implied by the fact that for this kernel, $\mathrm{M}_\mu(2k,2n) = 2^k\frac{(n+k-1)!}{(n-1)!}$ and thus one can take: $\xi(k) = 2^{k+1}$ in the theorem above. Note that from Stirling's formula we get: $k! = k^{k+\frac{1}{2}}e^{-k}(1+o_k(1))$. Thus the assumptions of Theorem 3.3 are satisfied for any fixed $C > 0$. It remains to observe that the charm function is positive for the Gaussian kernel K, since: $\Psi_K(\mathbf{z}) = 4\|\mathbf{z}\|^4 e^{-\|\mathbf{z}\|^2}$ (see Figure 2) and that the sequence of Fourier measures associated with the class of Gaussian kernels is concentrated (standard concentration result, see Chernoff, 2011).*

The fact that charm is nonnegative across the entire domain for the family of Gaussian kernels is not a coincidence. In fact the following is true.

Theorem 3.5 (Positive charm)**.** *Let $\phi : \mathbb{R} \rightarrow \mathbb{R}$ be such that for every $n \in \mathbb{N}$, $K_n : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ defined by $K_n(\mathbf{x},\mathbf{y}) = \phi(\|\mathbf{x}-\mathbf{y}\|)$ is a positive definite kernel. Then for each such K_n, the charm function Ψ_{K_n} is non-negative everywhere.*

The result above (details in the Appendix) uses a subtle connection between positive definite functions ϕ considered above and *completely monotone* functions.

Definition 3.6. *A function $\sigma : [0,+\infty] \rightarrow \mathbb{R}$ which is in $C[0,\infty] \cap C^\infty(0,\infty)$ and which satisfies $(-1)^r\frac{d^r\sigma}{dx^r} \geq 0\ \forall r \in \mathbb{N}_{\geq 0}$, is called completely monotone on $[0,\infty]$.*

The connection is given by the following theorem.

Theorem 3.7 (Schoenberg, 1938)**.** *A function σ is completely monotone on $[0,+\infty]$ iff the function $\phi : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ defined for $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$ as $\phi(\mathbf{x},\mathbf{y}) = \sigma(\|\mathbf{x}-\mathbf{y}\|^2)$ is positive definite for all $n \in \mathbb{N}$.*

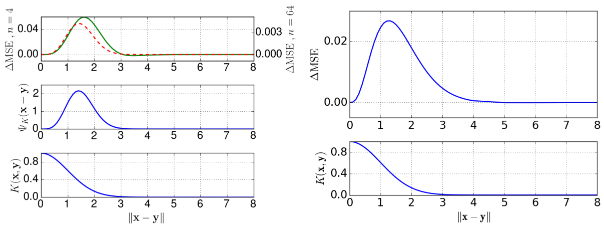
Combining Theorem 3.3 with Theorem 3.5, we obtain the following key result.

Theorem 3.8 (Superiority of the orthogonal transform)**.** *Under the assumptions of Theorem 3.3, for any fixed $z \in \mathbb{R}_{>0}$, for sufficiently large n, for any $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$ such that $\|\mathbf{x}-\mathbf{y}\| = z$,*

$$\mathrm{MSE}(\widehat{K}^{\mathrm{iid}}_{m,n}(\mathbf{x},\mathbf{y})) > \mathrm{MSE}(\widehat{K}^{\mathrm{ort}}_{m,n}(\mathbf{x},\mathbf{y})). \tag{6}$$

3.3 Non-asymptotic Results

Complementing the theoretical asymptotic results presented above, we provide additional analysis of the behavior of



(a) Gaussian kernel (b) Matérn-5/2 kernel

Figure 3: Difference between iid MSE and orthogonal MSE, charm function Ψ_K, and kernel K for Gaussian and Matérn-5/2 kernels for a range of dimensionalities. In the top plot of subplot (a), solid green is $n = 4$, dotted red is $n = 64$.

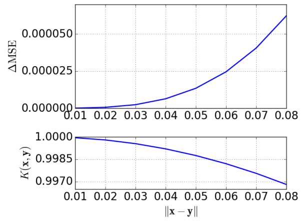


Figure 4: Difference in MSE for orthogonal and iid random features for the Matérn-5/2 kernel over $\mathbb{R}^{64}$, which satisfies the moment condition of Theorem 3.1.

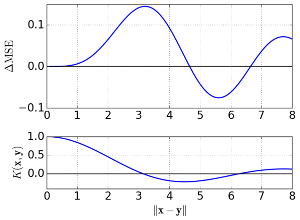


Figure 5: Difference in MSE for orthogonal and iid features for the sinc kernel, which does not correspond to a completely monotone positive definite function.

orthogonal random features in non-asymptotic regimes. The analysis centers on Proposition 3.9, which expresses the difference in MSE between iid and orthogonal random features in terms of univariate integrals, which although generally intractable, can be accurately and efficiently evaluated by deterministic numerical integration.

Proposition 3.9. *For an RBF kernel K on $\mathbb{R}^n$ with Fourier measure μ_K and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, writing $\mathbf{z} = \mathbf{x} - \mathbf{y}$, we have:*

$$\mathrm{MSE}(\widehat{K}^{\mathrm{iid}}_{m,n}(\mathbf{x},\mathbf{y})) - \mathrm{MSE}(\widehat{K}^{\mathrm{ort}}_{m,n}(\mathbf{x},\mathbf{y})) = \frac{m-1}{m}\mathbb{E}_{R_1,R_2}\left[\frac{J_{\frac{n}{2}-1}(\sqrt{R_1^2+R_2^2}\|\mathbf{z}\|)\Gamma(n/2)}{(\sqrt{R_1^2+R_2^2}\|\mathbf{z}\|/2)^{\frac{n}{2}-1}}\right] - \frac{m-1}{m}\mathbb{E}_{R_1}\left[\frac{J_{\frac{n}{2}-1}(R_1\|\mathbf{z}\|)\Gamma(n/2)}{(R_1\|\mathbf{z}\|/2)^{\frac{n}{2}-1}}\right]^2, \tag{7}$$

where R_1, R_2 are independent scalar random variables with the distribution of the norm of a vector drawn from μ_K, and J_α is the Bessel function of the first kind of degree α.

Firstly, In Figure 3, we plot the difference in MSE between iid random features and orthogonal random features for a range of kernels, noting that orthogonal features provide superior MSE across a wide range of values of $\|\mathbf{z}\|$. In the same plots, we show the value of the kernel K and of the charm function Ψ_K, noting that the charm function describes the benefits of orthogonal features accurately, even in the case of low dimensions. In all plots in this section, we write ΔMSE for $\mathrm{MSE}(\widehat{K}^{\mathrm{iid}}_m(\mathbf{x},\mathbf{y})) - \mathrm{MSE}(\widehat{K}^{\mathrm{ort}}_m(\mathbf{x},\mathbf{y}))$, so that $\Delta\mathrm{MSE} > 0$ corresponds to superior performance of orthogonal features over iid features.

Secondly, we illustrate the broad applicability of Theorem 3.1 by plotting the relative performance of orthogonal and iid features for the Matérn-5/2 kernel around the origin, see Figure 4.

Finally, we consider an RBF $K(\mathbf{x},\mathbf{y}) = \phi(\|\mathbf{x}-\mathbf{y}\|)$ which does *not* correspond to a completely monotone function. Let $n = 3$, and consider the Fourier measure μ that puts unit mass uniformly on the sphere $S^2 \subseteq \mathbb{R}^3$. As this is a finite isotropic measure on $\mathbb{R}^3$, there exists a corresponding RBF kernel K, which by performing an inverse Fourier transform can be shown to be

$$K(\mathbf{x},\mathbf{y}) = \sin(\|\mathbf{x}-\mathbf{y}\|)/\|\mathbf{x}-\mathbf{y}\|\,.$$

We term this the sinc kernel. Since the kernel takes on negative values for certain inputs, it does not correspond to a completely monotone function. Given the particular form of the Fourier measure, we may compute the difference in MSEs as given in Proposition 3.9 exactly, which yields

$$\mathrm{MSE}(\widehat{K}^{\mathrm{iid}}_{m,n}(\mathbf{x},\mathbf{y})) - \mathrm{MSE}(\widehat{K}^{\mathrm{ort}}_{m,n}(\mathbf{x},\mathbf{y})) = \frac{2}{3}\left(\frac{\sin(\sqrt{2}\|\mathbf{z}\|)}{\sqrt{2}\|\mathbf{z}\|} - \frac{\sin^2(\|\mathbf{z}\|)}{\|\mathbf{z}\|^2}\right). \tag{8}$$

We plot this function in Figure 5, noting there are large regions where orthogonal features are outperformed by iid features. Thus it may not be possible to relax the requirement in Theorem 3.8 that the pd function ϕ_K corresponds to a completely monotone function, as in Theorem 3.7.

4 OPTIMALITY OF THE RANDOM ORTHOGONAL FEATURE MAP MECHANISM

In this section, we consider unbiased estimators of RBFs introduced in Subsection 2.1. We show that for a significant family of random feature based estimators which we call *smooth*, asymptotically for large n, the orthogonal estimator is optimal in the sense of minimizing mean squared error. We will now identify a particular estimator E with a collection of probabilistic distributions on m-length n-dimensional tuples (each for different dimensionality n and number of random features m), each defining a set of sampled vectors $\mathbf{w}^n_1, ..., \mathbf{w}^n_m$.

Definition 4.1 (smooth estimators)**.** *A random feature based estimator E is* smooth *if for a fixed m, n lengths of directions of sampled vectors are chosen independently and* *furthermore, there exists a function $q : \mathbb{N} \rightarrow \mathbb{R}$ such that $q(x) \xrightarrow{x \rightarrow \infty} 0$ and for sampled vectors $\mathbf{w}_1^n, ..., \mathbf{w}_m^n$ the following is true:*

$$\mathbb{E}[|\cos(\theta_{i,j}^n)|^3] \leq q(n) \cdot \mathbb{E}[|\cos(\theta_{i,j}^n)|^2],$$

where $\theta_{i,j}^n$ is an angle between $\mathbf{w}_i^n$ and $\mathbf{w}_j^n$ and $i \neq j$.

Note that many useful estimators are smooth, including state-of-the-art estimators based on independent sampling, and also structured orthogonal estimators (note that for structured orthogonal estimators we have: $\cos(\theta_{i,j}) = 0$ with probability 1). Further, it is not hard to see that other estimators which can be obtained from general von Mises–Fisher distributions (Navarro et al., 2017) are also smooth. Von Mises-Fisher distributions generalize uniform distributions on the sphere with concentration parameters which are not too large – for example, the first sampled direction might define the mean direction then other directions could be sampled from a von Mises–Fisher distribution with the mean direction determined by the first sample.

We are ready to present our main result of this section, which shows that orthogonal random features are asymptotically optimal for the family of smooth estimators from Definition 4.1.

Theorem 4.2. *Consider a fixed positive definite radial basis function ϕ, a family of RBF kernels K, where K on $\mathbb{R}^n \times \mathbb{R}^n$ for each $n \in \mathbb{N}$ is defined as $K(\mathbf{x}, \mathbf{y}) = \phi(\|\mathbf{x} - \mathbf{y}\|)$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, and an associated concentrated sequence of Fourier measures $\{\mu_n\}_{n \in \mathbb{N}}$. Denote by E_{ort} an orthogonal estimator and by E_{smooth} some smooth estimator. Denote $\mathbf{z} = \mathbf{x} - \mathbf{y}$. Then, under assumptions of Theorem 3.3, for any fixed $\|\mathbf{z}\|$ and n large enough the following is true:*

$$\mathrm{MSE}(\widehat{K}_{m,n}^{\mathrm{ort}}(\mathbf{x}, \mathbf{y})) \leq \mathrm{MSE}(\widehat{K}_{m,n}^{\mathrm{smooth}}(\mathbf{x}, \mathbf{y})), \qquad (9)$$

where $\widehat{K}_{m,n}^{\mathrm{ort}}$ is an instance of E_{ort} for dimensionality n and using m random features (as in Theorem 3.3) and furthermore, $\widehat{K}_{m,n}^{\mathrm{smooth}}$ stands for the analogous instance of E_{smooth}.

5 SUPERIORITY OF ORTHOGONAL RANDOM FEATURES FOR DOWNSTREAM APPLICATIONS

One of the key applications of random feature maps is kernel ridge regression (KRR), where they lead to a scalable version of the algorithm. The KRR algorithm is a subject of intense research since ridge regression is one of the most fundamental machine learning methods that can be kernelized (Avron et al., 2016; Zhang et al., 2015). For this section we will borrow some notation from Avron et al. (2017). In the first subsection we give an overview and in the next one, present our new results.

5.1 Background: Ridge Regression with Approximate Kernel Methods

We must first introduce a few definitions.

Definition 5.1. *We say that a matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$ is a Δ-spectral approximation of another matrix $\mathbf{B} \in \mathbb{R}^{N \times N}$ for $\Delta \in \mathbb{R}_+$ if the following holds:*

$$(1 - \Delta)\mathbf{B} \preceq \mathbf{A} \preceq (1 + \Delta)\mathbf{B}, \qquad (10)$$

where $\mathbf{X} \preceq \mathbf{Y}$ stands for $\mathbf{Y} - \mathbf{X}$ being positive semidefinite.

Definition 5.2. *For a dataset $\mathcal{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ and a given kernel K, we define the kernel matrix $\mathbf{K}$ as*

$$\mathbf{K} = \{K(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j \in \{1,...,N\}}.$$

The random matrix obtained from $\mathbf{K}$ by replacing exact values of the kernel by the approximate values computed with iid features is denoted as $\widehat{\mathbf{K}}^{\mathrm{iid}}$, whereas the matrix where values are replaced by the approximate values computed with orthogonal features is $\widehat{\mathbf{K}}^{\mathrm{ort}}$.

We show that for $N \in \mathbb{N}$, an RBF kernel K (under assumptions of Theorem 3.3), an identity matrix $\mathbf{I}_N \in \mathbb{R}^{N \times N}$ and $\lambda > 0$, matrix $\widehat{\mathbf{K}}^{\mathrm{ort}} + \lambda N \mathbf{I}_N$ provides a strictly tighter spectral approximation of $\mathbf{K} + \lambda N \mathbf{I}_N$ than $\widehat{\mathbf{K}}^{\mathrm{iid}} + \lambda N \mathbf{I}_N$. It was shown by Avron et al. (2017) that the tightness of the spectral approximation of $\mathbf{K} + \lambda N \mathbf{I}_N$ implies accuracy guarantees of random feature based kernel methods on such downstream tasks as kernel ridge regression and kernel k-means clustering; for the reader's convenience we explain this in more detail below on the example of kernel ridge regression. Thus our results on the tightness of spectral approximation of orthogonal versus iid features will immediately imply the superiority of the orthogonal features approach on these downstream tasks. The matrix $\widehat{\mathbf{K}}^{\mathrm{ort}} + \lambda N \mathbf{I}_N$ will be our central object of study in this section.

We consider here the following model of data generation:

$$y_i = f^*(\mathbf{x}_i) + \nu_i\,, \qquad i = 1, \ldots, N\,, \qquad (11)$$

where f^* is the unknown groundtruth function to be learnt, $(y_i)_{i=1}^N$ are values assigned to data points $(\mathbf{x}_i)_{i=1}^N$ and $(\nu_i)_{i=1}^N$ are i.i.d noise terms distributed as mean-zero normal variables with standard deviation σ. The empirical risk of an estimator f of the groundtruth f^* obtained with the use of perturbed groundtruth values y_i is defined as:

$$\mathcal{R}(f) \equiv \mathbb{E}_{\nu_i}\left[\frac{1}{N}\sum_{j=1}^{N}(f(\mathbf{x}_i) - f^*(\mathbf{x}_i))^2\right], \qquad (12)$$

where N is the number of data points. Denote by f_{KRR} the kernel ridge regression estimator based on the groundtruth kernel matrix $\mathbf{K}$ and by $\mathbf{f}^* \in \mathbb{R}^n$ the vector whose j^{th} entry is $f^*(\mathbf{x}_j)$. In (Alaoui and Mahoney, 2015; Bach, 2013) it was proven that: $\mathcal{R}(f_{\mathrm{KRR}}) = \frac{\lambda^2}{N}(\mathbf{f}^*)^\top(\mathbf{K}+\lambda N\mathbf{I}_N)^{-2}\mathbf{f}^* + \frac{\sigma^2}{N}\mathrm{Tr}(\mathbf{K}^2(\mathbf{K}+\lambda N\mathbf{I}_N)^{-2})$, where λ stands for the regularization parameter, and $\mathbf{K}$ and $\mathbf{I}_N$ are as above.

As Avron et al. (2017) notice, the risk bound $\mathcal{R}(f_{\mathrm{KRR}})$ is upper-bounded by $\widehat{\mathcal{R}}_{\mathbf{K}}(\mathbf{f}^*)$, where $\widehat{\mathcal{R}}_{\mathbf{K}}(\mathbf{f}^*)$ is given as: $\widehat{\mathcal{R}}_{\mathbf{K}}(\mathbf{f}^*) \equiv \frac{\lambda(\mathbf{f}^*)^\top}{N}(\mathbf{K}+\lambda N\mathbf{I}_N)^{-1}\mathbf{f}^* + \frac{\sigma^2}{N}s_\lambda(\mathbf{K})$,

for $s_\lambda(K) \equiv \mathrm{Tr}((\mathbf{K}+\lambda N\mathbf{I}_N)^{-1}\mathbf{K})$. The expression $\widehat{\mathcal{R}}_{\mathbf{K}}(\mathbf{f}^*)$ played a crucial role in the analysis of Avron et al. (2017), leading to a compact formula on the upper bound on the risk of the general estimator in terms of the quality of the spectral approximation of $\mathbf{K}+\lambda N\mathbf{I}_N$:

Lemma 5.3. *Consider* KRR *estimator $\widehat{f}$ based on the matrix $\widehat{\mathbf{K}}$ approximating groundtruth kernel matrix $\mathbf{K}$. Assume furthermore that $\widehat{\mathbf{K}}+\lambda N\mathbf{I}_N$ is a Δ-spectral approximation of $\mathbf{K}+\lambda N\mathbf{I}_N$ for some $0 \leq \Delta < 1$ and that $\|\mathbf{K}\|_2 \geq 1$. Then the following is true:*

$$\mathcal{R}(\widehat{f}) \leq \frac{1}{1-\Delta}\widehat{\mathcal{R}}_{\mathbf{K}}(\mathbf{f}^*) + \frac{\Delta}{1+\Delta}\frac{\mathrm{rank}(\widehat{\mathbf{K}})}{N}\sigma^2. \qquad (13)$$

5.2 New Results: Kernal Ridge Regression with Orthogonal Features

We are ready to present our results. For simplicity we will give it for one random block (see: Section 2) however the result can be straightforwardly generalized to any number k of blocks. We show that orthogonal features lead to tighter spectral approximation of $\mathbf{K}+\lambda N\mathbf{I}_N$ for the class of considered RBFs and n large enough. We will borrow notation from the analysis above and Theorem 3.3.

Theorem 5.4. *Subject to the conditions of Theorem 3.3, consider RBFs (in particular Gaussian kernels). Let $\widehat{\Delta}$ denote the smallest positive number such that $\widehat{\mathbf{K}}+\lambda N\mathbf{I}_N$ is a Δ-approximation of $\mathbf{K}+\lambda N\mathbf{I}_N$, where $\widehat{\mathbf{K}}$ is an approximate kernel matrix obtained by using certain random feature map based kernel estimator. Then for any $a > 0$:*

$$\mathbb{P}[\widehat{\Delta} > a] \leq \frac{B}{a^2\sigma^2_{\min}}, \qquad (14)$$

where: $B = \sum_{i,j\in\{1,...,N\}}\mathrm{MSE}(\widehat{K}(\mathbf{x}_i,\mathbf{x}_j))$ and $\sigma_{\min}$ is the smallest singular value of $\mathbf{K}+\lambda N\mathbf{I}_N$. In particular, if if B^{ort} refers to the value of B for the estimator $\widehat{K}^{\mathrm{ort}}$ and B^{iid} to the one for the estimator $\widehat{K}^{\mathrm{iid}}$ then

$$B^{\mathrm{iid}} - B^{\mathrm{ort}} = \frac{m-1}{m}\Big(\frac{1}{8n}\cdot \sum_{i,j\in\{1,...,N\}}\left[\Psi_K(\|\mathbf{x}_i-\mathbf{x}_j\| + o\left(\frac{1}{n}\right)\right]\Big), \qquad (15)$$

where n is the data dimensionality and m is the number of random features used.

Note that for these RBFs, $B^{\mathrm{iid}} > B^{\mathrm{ort}}$ for n large enough, and thus orthogonal random features provide strictly better bound than iid features. To understand better the order of the magnitude of the upper bound on $\mathbb{P}[\widehat{\Delta} > a]$ from Theorem 5.4, it suffices to notice that if a dataset $\mathcal{X} = \{\mathbf{x}_1,...,\mathbf{x}_N\}$ is taken from some bounded region then random feature based estimators under consideration satisfy $\mathrm{MSE}(\widehat{K}(\mathbf{x}_i,\mathbf{x}_j)) = O(\frac{1}{m})$. For a constant $a > 0$ the upper bound is thus of the order $O(\frac{N^2}{m\sigma^2_{\min}})$. For $\lambda N \gg 1$ (which is the case for all practical applications) we have: $\sigma^4_{\min} = \Omega(\lambda^2N^2)$, thus the upper bound on $\mathbb{P}[\widehat{\Delta} > a]$ is of the order of magnitude $O(\frac{1}{m\lambda^2})$. Thus for $\lambda \gg \frac{1}{\sqrt{N}}$ (a reasonable practical choice), it suffices to take $m \ll N$ random features to get an upper bound of order $o(1)$ as $N \to \infty$.

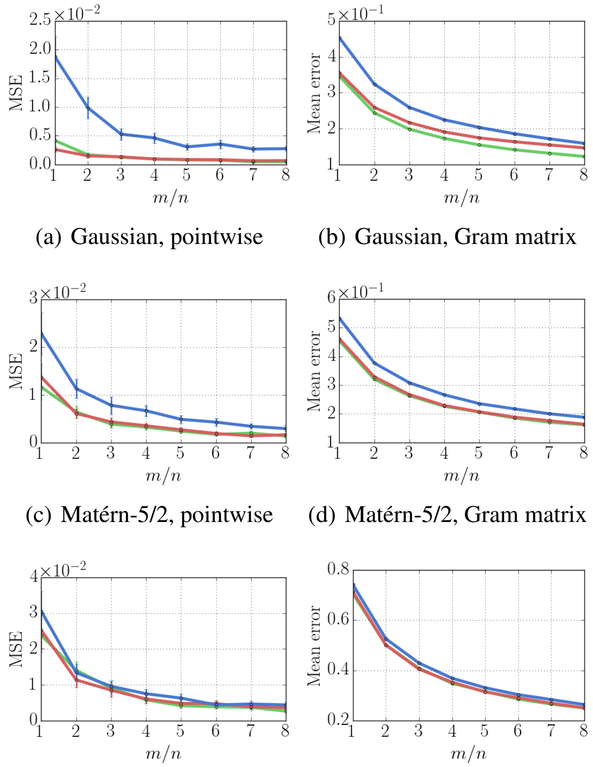
The above result immediately leads to the following regarding risk bounds for kernel ridge regression.

Theorem 5.5. *Under the assumptions of Theorem 5.4, the following holds for the kernel ridge regression risk and any $c > 0$ if m-dimensional random feature maps are used to approximate a kernel: $\mathbb{P}[\mathcal{R}(\widehat{f}) > c] \leq \frac{B}{a_c^2\sigma^2_{\min}}$, where a_c is given as: $a_c = 1 - \frac{\widehat{\mathcal{R}}_{\mathbf{K}}(\mathbf{f}^*)}{c-\frac{m\sigma^2}{2N}}$ and the probability is taken in respect to the random choices of features.*

Note that in the above bound the only term that depends on the choice of the random feature mechanism is B and thus as before, we conclude that orthogonal random features provide strictly stronger guarantees (this time in terms of the empirical risk of the random feature based kernel ridge regression estimator) than iid features. However, as we have noted before, the applications of spectral results given in Theorem 5.4 go beyond kernel ridge regression and can be applied in other kernelized algorithms.

6 EXPERIMENTS

We complement the theoretical results for pointwise kernel approximations in earlier sections with empirical studies of the effectiveness and limits of orthogonal random features in a variety of downstream applications. We also compare against structured orthogonal random features (SORF, first introduced only in the Gaussian case by Yu et al., 2016), where instead of drawing the directions of feature marginally from $\mathrm{Unif}(S^{n-1})$, we use the rows of the random matrix $\mathbf{H}\mathbf{D}_1\mathbf{H}\mathbf{D}_2\mathbf{H}\mathbf{D}_3$. Here, $\mathbf{H}$ is the normalized Hadamard matrix, and $\mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_3$ are iid diagonal matrices with independent $\mathrm{Unif}(\{\pm 1\})$ entries on the diagonals. Such matrices have recently been investigated as approximations to uniform orthogonal matrices, both empirically (Andoni et al., 2015) and analytically (Choromanski et al., 2017). We examine various numbers m of random features, while n is the dimensionality of the data. There is a one-time cost in constructing orthogonal features, which is small in practice.



(a) Gaussian, pointwise

(b) Gaussian, Gram matrix

(c) Matérn-5/2, pointwise

(d) Matérn-5/2, Gram matrix

(e) Laplace, pointwise

(f) Laplace, Gram matrix

Figure 6: Pointwise kernel evaluation MSE (left column) and normalized Frobenius norm error for Gram matrix approximation (right column) for the UCI "wine" dataset for Gaussian (top), Matérn-5/2 (center) and Laplace (bottom) kernels. Estimators are iid random features (blue), orthogonal random features (green) and approximate Hadamard-Rademacher random features (red).

6.1 Pointwise kernel and Gram matrix estimation

In this experiment, we study the estimation, via random feature maps, of kernel Gram matrices. We use MSE as an error measure for pointwise estimation, and normalized Frobenius norm as a measure of error for Gram matrices (so that the error incurred by estimating the Gram matrix $\mathbf{X}$ with the matrix $\widehat{\mathbf{X}}$ is $\|\mathbf{X} - \widehat{\mathbf{X}}\|_{\mathrm{F}}/\|\mathbf{X}\|_{\mathrm{F}}$). Kernel bandwidths are set via the median trick (Yu et al., 2016). We estimate pointwise kernel values and Gram matrices on a variety of full UCI regression datasets; see Figure 6 for examples. We plot the estimated mean Frobenius norm error, and bootstrapped estimates of standard error of the mean error estimates; in Figure 6, these error bars are extremely small. Full results are given in the Appendix, and have similar qualitative behaviour to that shown in Figure 6. Note that the orthogonal and approximate-orthogonal approaches are in general superior to iid random features, and that the improvement in performance is most pronounced for kernels with light-tailed Fourier distributions, as suggested by the theoretical developments in Section 3. Note that the Laplace kernel is a special case of the Matérn kernel in Figure 1.

6.2 Gaussian processes

We consider random feature approximations to Gaussian processes (GPs) for regression, and report (i) KL divergences between approximate predictive distributions versions obtained via random feature approximations against the predictive distribution obtained by an exactly-trained GP, and (ii) predictive RMSE on test sets. Experiments were run on a variety of UCI regression datasets - full experimental details are given in the Appendix. In Figures 7 and 8, results are shown for regression on the Boston housing dataset (Lichman, 2013). We use Gaussian, Matérn-5/2, and Laplace covariance kernels for the GP. Importantly, note that the posterior mean of the Gaussian process exactly corresponds to a kernel ridge regression estimator, so the RMSE results also serve to illustrate the theory in Section 5.

Kernel	Feature map	$m/n = 1$	$m/n = 2$	$m/n = 3$	$m/n = 4$
Gaussian	IID	104.2 (10.0)	34.21 (1.5)	15.6 (0.87)	11.05 (0.73)
	ORF	**100.4 (5.6)**	**26.62 (1.5)**	**15.1 (1.1)**	**8.707 (0.42)**
	SORF	108.9 (12.0)	32.29 (2.9)	16.25 (1.3)	10.15 (0.73)
Matérn-5/2	IID	160.3 (19.0)	47.88 (2.6)	25.87 (1.3)	18.61 (1.2)
	ORF	**123.2 (6.3)**	**41.66 (1.3)**	**21.78 (0.89)**	**16.66 (0.85)**
	SORF	166.4 (21.0)	44.74 (3.1)	25.14 (0.91)	16.89 (1.1)
Laplace	IID	337.2 (19.0)	126.4 (4.1)	69.66 (3.6)	50.99 (1.7)
	ORF	299.5 (17.0)	**117.7 (3.1)**	**68.4 (2.6)**	**44.25 (1.7)**
	SORF	**298.3 (7.6)**	121.1 (2.5)	70.56 (1.9)	47.88 (1.5)

Figure 7: Approximate GP regression results on Boston dataset. Reported numbers are average KL divergence from true posterior, along with bootstrap estimates of standard error (in parentheses).

Kernel	Feature map	$m/n = 1$	$m/n = 2$	$m/n = 3$	$m/n = 4$
Gaussian	IID	**0.54 (0.02)**	0.48 (0.01)	**0.43 (0.008)**	0.4 (0.01)
	ORF	0.59 (0.01)	**0.44 (0.008)**	0.43 (0.009)	**0.39 (0.006)**
	SORF	0.6 (0.02)	0.5 (0.02)	0.44 (0.009)	0.41 (0.008)
Matérn-5/2	IID	0.63 (0.02)	0.49 (0.008)	0.45 (0.01)	0.43 (0.006)
	ORF	**0.57 (0.02)**	0.47 (0.02)	**0.42 (0.006)**	**0.42 (0.008)**
	SORF	0.61 (0.04)	**0.47 (0.02)**	0.44 (0.01)	0.43 (0.01)
Laplace	IID	0.69 (0.04)	0.56 (0.02)	0.51 (0.01)	0.48 (0.01)
	ORF	0.65 (0.04)	0.54 (0.02)	0.51 (0.01)	0.48 (0.01)
	SORF	**0.62 (0.02)**	**0.53 (0.01)**	**0.49 (0.02)**	**0.47 (0.01)**

Figure 8: Approximate GP regression results on Boston dataset. Reported numbers are average test RMSE, along with bootstrap estimates of standard error (in parentheses).

7 CONCLUSION

We have explained the phenomenon of structured random features based on geometric conditions for RBF kernels. We showed the superiority of estimators based on orthogonal random feature maps for a large class of RBFs, substantially extending previously known results. Further, we showed in the high dimensionality regime that superiority comes from the shape of the introduced RBF-related charm function.

Acknowledgements

We thank the anonymous reviewers for helpful comments. MR acknowledges support by the UK Engineering and Physical Sciences Research Council (EPSRC) grant EP/L016516/1 for the University of Cambridge Centre for Doctoral Training, the Cambridge Centre for Analysis. RET is supported by Google as well as EPSRC grants EP/M0269571 and EP/L000776/1. AW acknowledges support from the David MacKay Newton research fellowship at Darwin College, The Alan Turing Institute under EPSRC grant EP/N510129/1 & TU/B/000074, and the Leverhulme Trust via the CFI.

References

N. Ailon and B. Chazelle. Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. In *STOC*, 2006.

A. El Alaoui and M. Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In *NIPS*, 2015.

A. Andoni, P. Indyk, T. Laarhoven, I. P. Razenshteyn, and L. Schmidt. Practical and optimal LSH for angular distance. In *NIPS*, 2015.

H. Avron, K. Clarkson, and D. Woodruff. Faster kernel ridge regression using sketching and preconditioning. *CoRR*, abs/1611.03220, 2016.

H. Avron, M. Kapralov, C. Musco, C. Musco, A. Velingker, and A. Zandieh. Random fourier features for kernel ridge regression: Approximation bounds and statistical guarantees. In *ICML*, 2017.

F. Bach. Sharp analysis of low-rank kernel matrix approximations. In *COLT*, 2013.

M. Bojarski, A. Choromanska, K. Choromanski, F. Fagan, C. Gouy-Pailler, A. Morvan, N. Sakr, T. Sarlos, and J. Atif. Structured adaptive and random spinners for fast machine learning computations. In *AISTATS*, 2017.

Herman Chernoff. Chernoff bound. In *International Encyclopedia of Statistical Science*, pages 242–243, 2011. doi: 10.1007/978-3-642-04898-2_170. URL `https://doi.org/10.1007/978-3-642-04898-2_170`.

A. Choromanska, K. Choromanski, M. Bojarski, T. Jebara, S. Kumar, and Y. LeCun. Binary embeddings with structured hashed projections. In *ICML*, 2016.

K. Choromanski and V. Sindhwani. Recycling randomness with structure for sublinear time kernel expansions. In *ICML*, 2016.

Krzysztof Choromanski, Mark Rowland, and Adrian Weller. The unreasonable effectiveness of structured random orthogonal embeddings. In *NIPS*, 2017.

C. Cortes and V. Vapnik. Support-vector networks. In *Machine Learning*, pages 273–297, 1995.

A. Hinrichs and J. Vybíral. Johnson-Lindenstrauss lemma for circulant matrices. *Random Structures & Algorithms*, 39(3):391–398, 2011.

T. Joachims. Training linear SVMs in linear time. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD, pages 217–226, New York, NY, USA, 2006.

Q. Le, T. Sarlós, and A. Smola. Fastfood - approximating kernel expansions in loglinear time. In *ICML*, 2013.

M. Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

Alexandre K. W. Navarro, Jes Frellsen, and Richard E. Turner. The multivariate generalised von Mises distribution: Inference and applications. In *AAAI*, 2017.

D. Ormoneit and S. Sen. Kernel-based reinforcement learning. *Machine Learning*, 49(2-3):161–178, 2002.

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

C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

I. Schoenberg. Metric Spaces and Completely Monotone Functions. *The Annals of Mathematics*, 39(4):811–841, 1938.

D. Sutherland and J. Schneider. On the error of random Fourier features. In *UAI*, 2015.

J. Vybíral. A variant of the Johnson-Lindenstrauss lemma for circulant matrices. *Journal of Functional Analysis*, 260(4):1096–1105, 2011.

F. Yu, A. Suresh, K. Choromanski, D. Holtmann-Rice, and S. Kumar. Orthogonal random features. In *NIPS*, 2016.

H. Zhang and L. Cheng. New bounds for circulant Johnson-Lindenstrauss embeddings. *CoRR*, abs/1308.6339, 2013.

Y. Zhang, J. Duchi, and M. Wainwright. Divide and conquer kernel ridge regression: a distributed algorithm with minimax optimal rates. *Journal of Machine Learning Research*, 16, 2015.