Orthogonal Random Features: Explicit Forms and Sharp Inequalities

Anonymous authors
Paper under double-blind review

Abstract

Random features have been introduced to scale up kernel methods via randomization techniques. In particular, random Fourier features and orthogonal random features were used to approximate the popular Gaussian kernel. The former is performed by a random Gaussian matrix and leads exactly to the Gaussian kernel after averaging. In this work, we analyze the bias and the variance of the kernel approximation based on orthogonal random features which makes use of Haar orthogonal matrices. We provide explicit expressions for these quantities using normalized Bessel functions, showing that—contrary to what is commonly thought—orthogonal random features does not approximate the Gaussian kernel but a Bessel kernel. We also derive sharp exponential bounds supporting the view that orthogonal random features are less dispersed than random Fourier features.

1 Introduction

Since their introduction over fifteen years ago in the seminal paper by Rahimi & Recht (2007), random features have become an important subject of research in the field of machine learning (see the review article by Liu et al., 2021). The primary motivation behind introducing them is to reduce the computation and storage requirements of kernel methods—one of the most popular machine learning approaches (Yang et al., 2012; Le et al., 2013; Pennington et al., 2015; Chamakh et al., 2020; Han et al., 2022; Likhosherstov et al., 2022). They have also been used in over-parameterized settings and as a tool for generating and testing hypotheses on the generalization of deep learning (Jacot et al., 2018; Belkin et al., 2019; Yehudai & Shamir, 2019; Jacot et al., 2020; Liu et al., 2022; Mei & Montanari, 2022).

Random Fourier features (RFF) are undoubtedly the most common and widely used random feature method for kernel approximation (Rahimi & Recht, 2007). This approach applies to radial basis function kernels—a large class of kernel functions. It is based on Bochner's theorem (Bochner, 1932; Rudin, 1962), which establishes a one-to-one correspondence between continuous positive-definite functions and the Fourier transform of probability measures. In particular, if φ is a real positive definite function on $\mathbb{R}$, i.e. the inequality $\sum_{i,j=1}^{n} \alpha_i \alpha_j \varphi(x_i - x_j) \geq 0$ holds for any $n \geq 1$ and $x_1, \ldots, x_n, \alpha_1, \ldots, \alpha_n \in \mathbb{R}$, and if $k(x, y) := \varphi(\|x - y\|)$ is a translation-invariant and radial kernel on $\mathbb{R}^d \times \mathbb{R}^d$, then there exists a non-negative measure μ such that

$$k(x, y) := \varphi(\|x - y\|) = \int_{\mathbb{R}^d} \exp(iw^\top (x - y)) d\mu(w), \tag{1}$$

where $\top$ stands for the transpose operation, and μ is invariant under orthogonal transformations. RFF consist in approximating the kernel k by the following one defined by:

$$\tilde{k}(x, y) := \tilde{\phi}(x)^\top \tilde{\phi}(y), \quad \forall x, y, \in \mathbb{R}^d, \tag{2}$$

where

$$\tilde{\phi}(x) := \frac{1}{\sqrt{p}} (\sin(w_1^\top x), \ldots, \sin(w_p^\top x), \cos(w_1^\top x), \ldots, \cos(w_p^\top x))^\top, \tag{3}$$

and $w_1, \dots, w_p \in \mathbb{R}^d$ are sampled from the distribution μ. Indeed, one has

$$\tilde{\phi}(x)^\top \tilde{\phi}(y) = \frac{1}{p} \sum_{j=1}^{p} \cos(w_j^\top (x - y)),$$

so that the expected value of $\tilde{k}(x, y)$ fits exactly the integral displayed in the RHS of equation 2. The imaginary part of the integral in equation 1 vanishes because the kernel function is radial. In particular, if $w_1, \dots, w_p$ are centered Gaussian vectors $\mathcal{N}(0, \mathrm{Id})$, then RFF approximate the well-known Gaussian kernel:

$$\mathbb{E}_{w \sim \mathcal{N}(0,\mathrm{Id})}[\tilde{k}(x, y)] = e^{-\|x-y\|^2/2}. \tag{4}$$

Orthogonal random features (ORF) is a variant of RFF that uses a random orthogonal matrix O instead of the Gaussian matrix W (Yu et al., 2016). It has shown empirically and theoretically that ORF estimators are more accurate than standard mechanisms based on i.i.d sampling (Yu et al., 2016; Choromanski et al., 2017; 2018; 2019). In this paper, we build on this line of research and provide an analytic characterization of the bias and of the variance of ORF using normalized Bessel functions of the first kind. Specifically, we make the following contributions:

- We give explicit forms of the bias and of the variance of ORF in the case where the random orthogonal matrix O is drawn from the Haar measure.
- We derive sharp exponential bounds for these two quantities that are much tighter than the already known ones.
- We prove that the variance of ORF is less than the one of RFF in an interval whose length grows linearly with the data dimension.
- We corroborate our theoretical findings with numerical experiments, supporting previous works showing the beneficial effect of orthogonality on random features.

The rest of the paper is organized as follows. Section 2 lays out notations needed for the statement of our main results, the latter being subsequently presented in Section 3. Numerical validation of them is provided in Section 4, while Section 5 concludes the paper. Proofs of all results are deferred to appendices.

2 Notation and preliminaries

Let p, d, be positive integers such that $2 \le p \le d$ and take a Haar $d \times d$ orthogonal matrix O. For the reader's convenience, recall that the Haar measure on the orthogonal group is the unique left and right invariant measure and that a Haar orthogonal matrix may be obtained using the Gram-Schmidt procedure applied to a Gaussian matrix G. In particular, the columns (and rows) of O are uniformly distributed on the sphere (for further details see Meckes 2019, Chapter 1).

We denote by $\tilde{\phi}_{ORF}$ the random features of ORF computed using equation 3 with $w_1, \dots, w_p$ being columns of O. We also use the notation $\tilde{\phi}_{RFF}$ for the random features of RFF when $w_1, \dots, w_p$ are columns of a Gaussian matrix G (i.e., a random matrix whose entries are independent and centered Gaussian random variables). The approximate kernels obtained using ORF and RFF will be denoted by $\tilde{k}_{ORF}(x, y)$ and $\tilde{k}_{RFF}(x, y)$, respectively (i.e., $\tilde{k}_{ORF}(x, y) := \tilde{\phi}_{ORF}(x)^\top \tilde{\phi}_{ORF}(y)$ and $\tilde{k}_{RFF}(x, y) := \tilde{\phi}_{RFF}(x)^\top \tilde{\phi}_{RFF}(y)$). In order to simplify the exposition and without loss of generality, we will assume throughout this paper that the bandwidth of the Gaussian kernel σ is equal to 1. In this respect and for sake of completeness, let us recall the expressions of the bias and the variance of RFF.

Theorem 1 (Bias and variance of $\tilde{\mathbf{k}}_{\mathbf{RFF}}(\mathbf{x}, \mathbf{y})$)**.** *Let $\tilde{k}_{RFF}(x, y)$ be the RFF-based approximate kernel computed with p random vectors in $\mathbb{R}^d$. Then its expectation and its variance are given by*

$$\mathbb{E}[\tilde{k}_{RFF}(x, y)] = \exp\left(-\frac{\|x - y\|^2}{2}\right), \tag{5}$$

and

$$V[\tilde{k}_{RFF}(x,y)] = \frac{1}{2p}\left(1 - \exp{(-\frac{\|x-y\|^2}{2})}\right)^2, \tag{6}$$

respectively.

Proof. See Yu et al. (2016, Lemma 1). □

It is worth noting that the equality equation 5 remains valid if one replaces the Gaussian matrix G by the product SO, where O is a Haar orthogonal matrix and S is a diagonal matrix whose entries are independent and ξ-distributed random variables with d degrees of freedom (Yu et al., 2016, Theorem 1). However, it fails when $w_1, \ldots, w_p$ are columns of O. We will see in the next section how the bias and the variance change and behave in this case.

3 Main results

In this section, we state the main results of this paper and comment on our findings. We start with an explicit expression of the bias of ORF-based kernel approximation.

Theorem 2 (Bias of $\tilde{\mathbf{k}}_{\mathbf{ORF}}(\mathbf{x},\mathbf{y})$)**.** *Let $\tilde{k}_{ORF}(x,y)$ be the ORF-based approximate kernel computed with p random vectors in $\mathbb{R}^d$. Then its expectation reads:*

$$\mathbb{E}[\tilde{k}_{ORF}(x,y)] = j_{d/2-1}(z), \quad z := \|x-y\|, \tag{7}$$

where $j_{d/2-1}(\cdot)$ is the normalized Bessel function of the first kind defined by:

$$j_{d/2-1}(z) = \sum_{n\geq 0} \frac{(-1)^n \Gamma(d/2)}{n!\Gamma(n+(d/2))}\left(\frac{z}{2}\right)^{2n}, \tag{8}$$

with $\Gamma(\cdot)$ being the Gamma function.

Proof. See Appendix A. □

Theorem 2 shows that ORF approximate the kernel defined by the normalized Bessel function of the first kind (Watson, 1995; Shishkina & Sitnik, 2020, Chapter 1). This function is oscillating in contrast to the so-called Matern kernel given by the modified Bessel function of the second kind (see Equation 12 in Genton 2001). Moreover, the absolute value of the former admits a polynomial decay to zero as its argument becomes large while the latter decays exponentially.

To address the question of how the bias of ORF behaves compared to that of RFF, we use Theorem 2 to prove the following result.

Proposition 3. *For all $x, y \in \mathbb{R}^d$, let $z := \|x-y\|$ and define*

$$\begin{aligned} b_d &:= 2^{1/4}d^{3/4}\sqrt{1 - \frac{4}{2\sqrt{2}d^{3/2} - d}}, \quad d \geq 2, \\ c_d &:= \left(\frac{d^2}{4} - 1\right)^{1/2}\sqrt{1 - \frac{8}{d^2 - 2d - 4}}, \quad d \geq 5. \end{aligned}$$

Then for all $z \in [0, \max(b_d, c_d)]$, we have

$$e^{-z^2/2} \leq \mathbb{E}[\tilde{k}_{ORF}(x,y)] \leq e^{-z^2/(2d)}, \tag{9}$$

where we convent that $\max(b_d, c_d) = b_d$ *when $2 \leq d \leq 4$. The upper bound is valid up to the second zero of $j_{d/2-1}$.*

Proof. See Appendix B. □

For fixed $d \geq 5$, the constants b_d and c_d are the values of the increasing function

$$f_d : u \mapsto u\sqrt{1 - \frac{4}{2u^2 - d}}, \quad u \geq \frac{d+4}{2},$$

at two lower bounds of the first zero of $j_{(d/2)-1}$; see eqs. equation 18 and equation 19 below. Moreover, $c_d > b_d$ for sufficiently large d ($d \gtrsim 35$) and offers therefore a linear growth of the interval $[0, c_d]$ compared to $d^{3/4}$ for $[0, b_d]$. As a matter of fact, the inequalities displayed in equation 9 hold true for a large range of z provided that d is large as well. As we shall see later from the proof of Proposition 3, the lower bound even holds true in a larger interval which almost reaches the first zero of $j_{(d/2)-1}$ as d becomes large. Note in passing that $j_{(d/2)-1}$ becomes negative once vanishing at its first zero so that our lower bound is quite sharp.

As to the upper bound, it clearly remains valid up to the second zero of $j_{(d/2)-1}$ since the latter is non positive between its first and its second zeroes. Note also that the lower and the upper bounds correspond to Gaussian kernels with standard deviations equal to one and to $\sqrt{d}$ respectively.

On the other hand, we may equivalently write

$$0 \leq \mathbb{E}[\tilde{k}_{ORF}(x,y)] - \mathbb{E}[\tilde{k}_{RFF}(x,y)] \leq e^{-z^2/(2d)} - e^{-z^2/2}, \quad \forall z \in [0, \max(b_d, c_d)], \tag{10}$$

which considerably improves Theorem 2 in Yu et al. (2016). Indeed, our upper bound decays exponentially fast while the one given in Yu et al. (2016) admits an exponential growth. This growth is due to the fact that the triangular inequality used in the proof there kills the oscillations of the normalized Bessel function $j_{d/2-1}$ and leads to the normalized modified Bessel function of the first kind which is known to grow exponentially (Watson, 1995).

It is also worth mentioning that for large enough d, the differences between $\mathbb{E}[\tilde{k}_{ORF}(x,y)]$ and the exponential bounds displayed in equation 9 become very small for large z ($z \gg d$) since $|j_{(d/2)-1}|$ has a polynomial decay to zero.

We now state our second main result providing an explicit closed expression of the variance of ORF by means of normalized Bessel functions of the first kind.

Theorem 4 (Variance of $\tilde{\mathbf{k}}_{\mathbf{ORF}}(\mathbf{x},\mathbf{y})$). *Let $\tilde{k}_{ORF}(x,y)$ be the ORF-based approximate kernel built out of p random vectors in $\mathbb{R}^d$. Then its variance is given by:*

$$V[\tilde{k}_{ORF}(x,y)] = \frac{1}{p}\left\{\frac{1 + j_{(d/2)-1}(2z)}{2} + (p-1)j_{(d/2)-1}(\sqrt{2}z) - p\left(j_{(d/2)-1}(z)\right)^2\right\}, \tag{11}$$

where we recall the notation $z = \|x - y\|$.

Proof. See Appendix C. □

Remark 5. *When $p = 1$, the variance reduces to*

$$V\left(K(x,y)\right) = \frac{1 + j_{(d-2)/2}(2z)}{2} - \left(j_{(d-2)/2}(z)\right)^2.$$

The non negativity of the RHS follows also from the inequality:

$$[j_\nu(x) + j_\nu(y)]^2 \leq [1 + j_\nu(x+y)][1 + j_\nu(x-y)]$$

valid for any $\nu > -1/2$ and any $x, y \in \mathbb{R}$ (Neuman, 2004).

Using equation 9, this variance can be bounded as follows for all $z \in [0, \max(b_d, c_d)]$:

$$\frac{1 + e^{-2z^2}}{2p} + \frac{p-1}{p}e^{-z^2} - e^{-z^2/d} \leq V[\tilde{k}_{ORF}(x,y)] \leq \frac{1 + e^{-2z^2/d}}{2p} + \frac{p-1}{p}e^{-z^2/d} - e^{-z^2}. \tag{12}$$

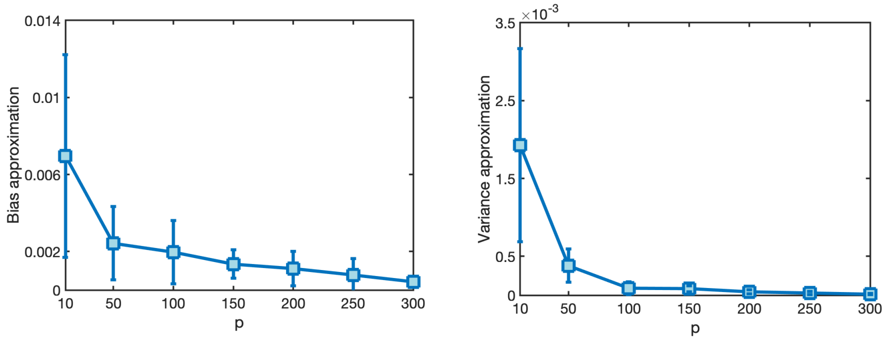

Figure 1: The absolute difference between theoretical and empirical values of the bias and the variance of ORF for different values of the number of random features p. **Left:** $\left|M_{emp} - \mathbb{E}[\tilde{k}_{ORF}(x,y)]\right|$. **Right:** $\left|V_{emp} - V[\tilde{k}_{ORF}(x,y)]\right|$. The bias and variance of $\tilde{k}_{ORF}(x,y)$, $\mathbb{E}[\tilde{k}_{ORF}(x,y)]$ and $V[\tilde{k}_{ORF}(x,y)]$, are computed using the explicit closed expressions provided in Theorems 2 and 4. M_{emp} and V_{emp} are the empirical bias and variance, respectively.

Similarly to equation 10, the upper bound in equation 12 is much sharper than the one in Yu et al. (2016, Theorem 2). However, it does not give information about how the variance of ORF compares to that of RFF. The following proposition provides an answer to this question.

Proposition 6. *For $d \geq 2$, denote*

$$\begin{aligned} \alpha_d &:= \left(\frac{d}{2}\right)^{3/4}, \\ \beta_d &:= \frac{1}{2}\sqrt{\frac{d^2}{4} - 1}. \end{aligned}$$

Then for any $x, y \in \mathbb{R}^d$ and any $z = ||x - y|| \in [0, \max(\alpha_d, \beta_d)]$, we have

$$V[\tilde{k}_{ORF}(x,y)] \leq V[\tilde{k}_{RFF}(x,y)]. \tag{13}$$

Proof. See Appendix D. □

Proposition 6 shows that $\tilde{k}_{ORF}$ is less dispersed than $\tilde{k}_{RFF}$ when the norm difference between data points z lies within an interval whose length is linear in the data dimension d when the latter is sufficiently large. This is in agreement with previous results (Choromanski et al., 2017; 2018; 2019) though holding in a small z-neighborhood of zero.

Another striking feature of this proposition is its independence of the number of random features p. Actually, its proof (see Appendix D below) shows that the covariance of $(\cos(w_i^T(x-y)), \cos(w_j^T(x-y)), i \neq j$, is negative in $[0, \sqrt{2}\max(\alpha_d, \beta_d)]$. As a matter of fact, one is left with bounding the variance of a single mode $\cos(w_1^T(x-y))$. In the same vein, the proof of this proposition shows that the inequality equation 13 remains valid only in a slightly larger interval where $j_{(d/2)-1}(\sqrt{2}z)$ is negative.

4 Numerical illustrations

In this section we provide experimental results on synthetic and real data that corroborate our theoretical findings.

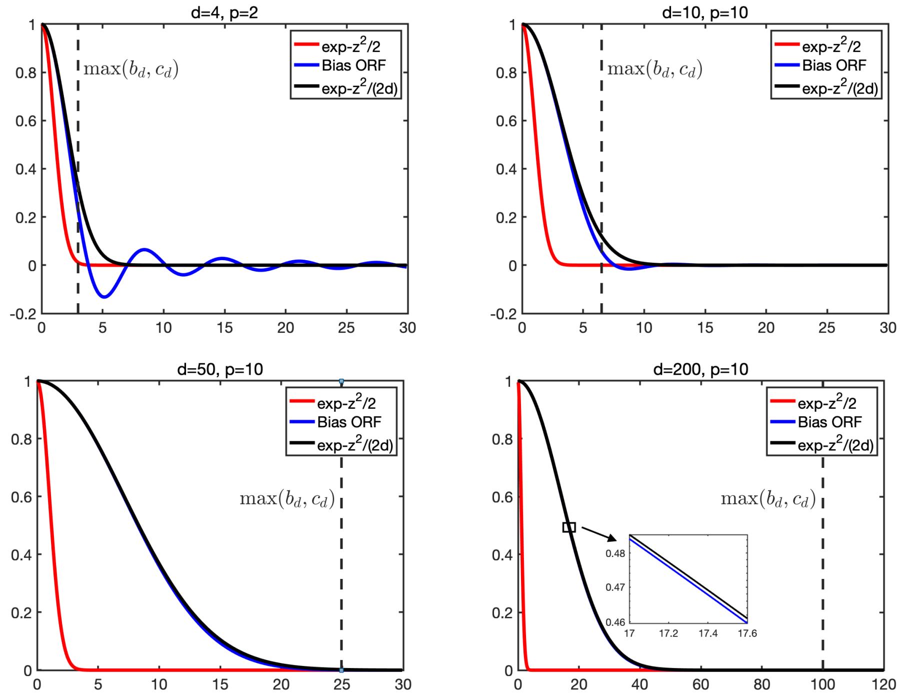

Figure 2: The bias of $\tilde{k}_{ORF}(x, y)$ and bounds of Proposition 3 as a function of $z := \|x - y\|$.

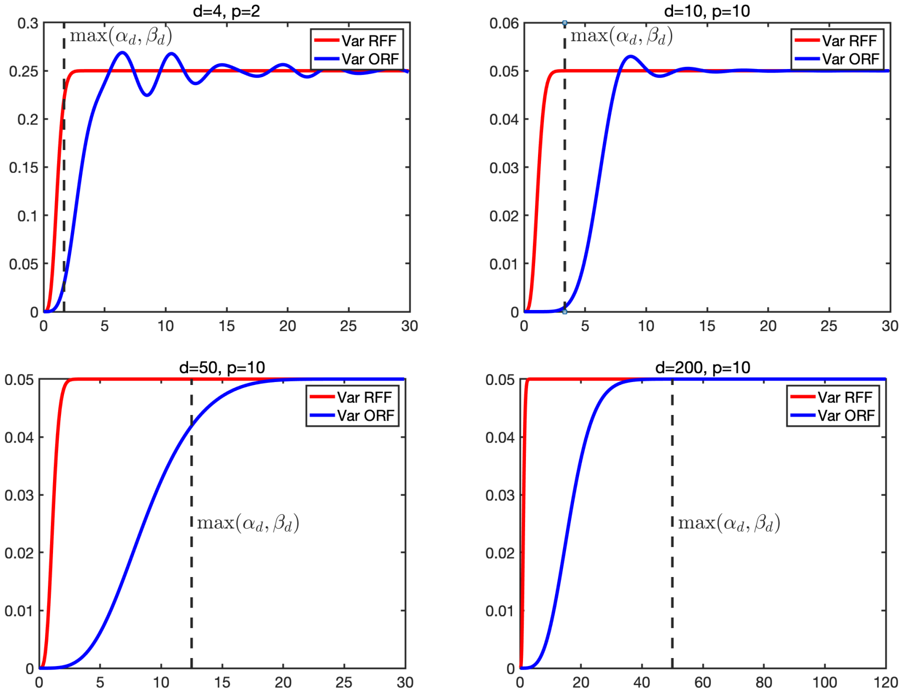

Figure 3: The variance of $\tilde{k}_{ORF}(x, y)$ and $\tilde{k}_{RFF}(x, y)$ as a function of $z := \|x - y\|$.

Table 1: Dataset statistics. d is the dimension of the features.

Datasets	d	# data points
Ionosphere[1]	34	351
Ovariancancer[2]	100	216
Campaign[3]	62	41,188
Backdoor[4]	196	95,329

4.1 Synthetic data results

We generate synthetic data with dimension $d = 300$ and varying values of the random features $p = \{10, 50, 100, 150, 200, 250, 300\}$. The data are randomly generated from a normal distribution with zero and unit variance. We compute $M_{emp} := \frac{1}{s}\sum_{l=1}^{s} \tilde{k}_l(x,y)$ and $V_{emp} := \frac{1}{s}\sum_{l=1}^{s}(\tilde{k}_l(x,y) - M_{emp})^2$, the empirical bias and variance of $\tilde{k}_{ORF}$ respectively. Each kernel $\tilde{k}_l$ is computed using a random Haar orthogonal matrix O^l, i.e., $\tilde{k}_l(x,y) = \langle \tilde{\phi}_l(x), \tilde{\phi}_l(y) \rangle$ where $\tilde{\phi}_l(x) = \frac{1}{\sqrt{p}}(\sin(\langle w_1^l, x\rangle), \ldots, \sin(\langle w_p^l, x\rangle), \cos(\langle w_1^l, x\rangle), \ldots, \cos(\langle w_p^l, x\rangle)^\top$ and $w_1^l, \ldots, w_p^l$ are the columns of O^l. The experiment is repeated 10 times with different random seeds. Figure 1 shows the approximation errors $\left|M_{emp} - \mathbb{E}[\tilde{k}_{ORF}(x,y)]\right|$ and $\left|V_{emp} - V[\tilde{k}_{ORF}(x,y)]\right|$ for $s = 50$ and for different values of p. The mean and variance of $\tilde{k}_{ORF}(x,y)$, $\mathbb{E}[\tilde{k}_{ORF}(x,y)]$ and $V[\tilde{k}_{ORF}(x,y)]$, are computed using the explicit closed expressions provided in Theorems 2 and 4. As can be seen, the mean and variance approximation errors are very small, which are in agreement with our results.

Figure 2 shows the bias of $\tilde{k}_{ORF}(x,y)$ and the bounds of Proposition 3 as a function of $z = \|x - y\|$. It illustrates that inequalities in equation 9 hold for any $z \in [0, \max(b_d, c_d)]$. Figure 3 depicts the variance of $\tilde{k}_{ORF}$ and $\tilde{k}_{RFF}$. It confirms that ORF has smaller variance compared to the standard RFF, as claimed in Proposition 6.

4.2 Real data results

We also conduct experiments on real-world datasets to confirm our theoretical findings. The number of feature dimension and data samples for each dataset are provided in Table 1. Figure 4 compares the mean squared error (MSE) of ORF and RFF, i.e., $\|K - \tilde{K}\|_F^2/n^2$ where $K := [k(x_i, x_j)]_{i,j=1}^n$ is the Bessel or Gaussian kernel matrix and $\tilde{K}$ is its approximation via ORF or RFF, respectively. Note that MSE corresponds to the empirical variance. The Gaussian kernel bandwidth σ is set as the average distance between all pairs of data points, i.e., $\sigma = \sqrt{1/n^2 \sum_{i,j=1}^n \|x_i - x_j\|^2}$, and the experiment is repeated five times with different random seeds. ORF often achieves lower MSE than RFF.

5 Conclusion

In this paper, we provided explicit closed expressions of the bias and of the variance of ORF by means of normalized Bessel functions of the first kind. We also derived exponential bounds that improve previously known ones. In particular, we proved that the variance of ORF is less than the one of RFF when the norm difference between data points lies in an interval of length $O(d)$, d being the data dimension.

[3] Ionosphere data from the UCI machine learning repository: `https://archive.ics.uci.edu/dataset/52/ionosphere`.

[4] Ovarian cancer data (Conrads et al., 2004): `https://fr.mathworks.com/help/stats/sample-data-sets.html`.

[5] Campaign data is a data set of direct bank marketing campaigns via phone calls (Pang et al., 2019): `https://github.com/GuansongPang/ADRepository-Anomaly-detection-datasets#numerical-datasets`.

[6] Backdoor attack detection data extracted from the UNSW-NB 15 dataset (Moustafa & Slay, 2015): `https://github.com/GuansongPang/ADRepository-Anomaly-detection-datasets#numerical-datasets`.

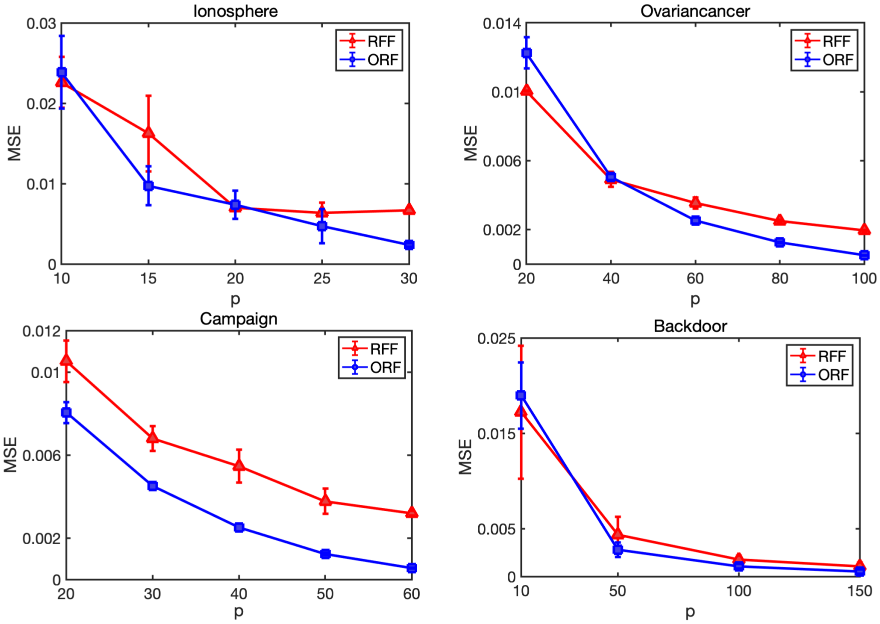

Figure 4: Mean squared error (MSE) between the kernel matrix approximated by ORF or RFF and the full kernel matrix computed by the Bessel or the Gaussian kernel, for different values of the number of random features p.

References

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Salomon Bochner. *Vorlesungen über Fouriersche Integrale*. Akademische Verlagsgesellschaft, 1932.

Linda Chamakh, Emmanuel Gobet, and Zoltán Szabó. Orlicz random Fourier features. *The Journal of Machine Learning Research*, 21(1):5739–5775, 2020.

Krzysztof Choromanski, Mark Rowland, and Adrian Weller. The unreasonable effectiveness of structured random orthogonal embeddings. *Advances in Neural Information Processing Systems*, 30, 2017.

Krzysztof Choromanski, Mark Rowland, Tamás Sarlós, Vikas Sindhwani, Richard Turner, and Adrian Weller. The geometry of random features. In *International Conference on Artificial Intelligence and Statistics*, pp. 1–9, 2018.

Krzysztof Choromanski, Mark Rowland, Wenyu Chen, and Adrian Weller. Unifying orthogonal monte carlo methods. In *International Conference on Machine Learning*, pp. 1203–1212, 2019.

Thomas P. Conrads, Vincent A. Fusaro, Sally Ross, Don Johann, Vinodh Rajapakse, Ben A. Hitt, Seth M. Steinberg, Elise C. Kohn, David A. Fishman, Gordon Whitely, et al. High-resolution serum proteomic features for ovarian cancer detection. *Endocrine-related cancer*, 11(2):163–178, 2004.

Pedro Freitas. Bessel quotients and robin eigenvalues. *Pacific Journal of Mathematics*, 315(1):75–87, 2021.

Marc G. Genton. Classes of kernels for machine learning: a statistics perspective. *Journal of machine learning research*, 2:299–312, 2001.

Izrail Solomonovich Gradshteyn and Iosif Moiseevich Ryzhik. *Table of integrals, series, and products*. Academic press, 2014.

Insu Han, Amir Zandieh, and Haim Avron. Random Gegenbauer features for scalable kernel methods. In *International Conference on Machine Learning*, pp. 8330–8358, 2022.

EK Ifantis and PD Siafarikas. Inequalities involving bessel and modified bessel functions. *Journal of mathematical analysis and applications*, 147(1):214–227, 1990.

Mourad EH Ismail and Martin E Muldoon. On the variation with respect to a parameter of zeros of bessel and q-bessel functions. *Journal of mathematical analysis and applications*, 135(1):187–207, 1988.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, 31, 2018.

Arthur Jacot, Berfin Simsek, Francesco Spadaro, Clément Hongler, and Franck Gabriel. Implicit regularization of random feature models. In *International Conference on Machine Learning*, pp. 4631–4640, 2020.

Chandra Mohan Joshi and S. K. Bissu. Inequalities for some special functions. *Journal of computational and applied mathematics*, 69(2):251–259, 1996.

Quoc Le, Tamás Sarlós, Alex Smola, et al. Fastfood-approximating kernel expansions in loglinear time. In *International Conference on Machine Learning*, pp. 244–252, 2013.

Valerii Likhosherstov, Krzysztof Choromanski, Kumar Avinava Dubey, Frederick Liu, Tamas Sarlos, and Adrian Weller. Chefs' random tables: Non-trigonometric random features. *Advances in Neural Information Processing Systems*, 35:34559–34573, 2022.

Fanghui Liu, Xiaolin Huang, Yudong Chen, and Johan Suykens. Random features for kernel approximation: A survey on algorithms, theory, and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7128–7148, 2021.

Fanghui Liu, Johan Suykens, and Volkan Cevher. On the double descent of random features models trained with SGD. *Advances in Neural Information Processing Systems*, 35, 2022.

Elizabeth S. Meckes. *The random matrix theory of the classical compact groups*, volume 218. Cambridge University Press, 2019.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

Nour Moustafa and Jill Slay. UNSW-NB15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set). In *2015 military communications and information systems conference (MilCIS)*, pp. 1–6, 2015.

Edward Neuman. Inequalities involving bessel functions of the first kind. *J. Ineq. Pure and Appl. Math*, 5 (4), 2004.

Guansong Pang, Chunhua Shen, and Anton van den Hengel. Deep anomaly detection with deviation networks. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 353–362, 2019.

Jeffrey Pennington, Felix X. Yu, and Sanjiv Kumar. Spherical random features for polynomial kernels. *Advances in Neural Information Processing Systems*, 28, 2015.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

Walter Rudin. *Fourier Analysis on Groups.* Wiley, 1962.

Elina Shishkina and Sergei Sitnik. *Transmutations, singular and fractional differential equations with applications to mathematical physics.* Academic Press, 2020.

George N. Watson. *A Treatise on the Theory of Bessel Functions.* Cambridge University Press, 1995.

Tianbao Yang, Yu-Feng Li, Mehrdad Mahdavi, Rong Jin, and Zhi-Hua Zhou. Nyström method vs random Fourier features: A theoretical and empirical comparison. *Advances in Neural Information Processing Systems*, 25, 2012.

Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Felix X. Yu, Ananda Theertha Suresh, Krzysztof Choromanski, Daniel Holtmann-Rice, and Sanjiv Kumar. Orthogonal random features. *Advances in Neural Information Processing Systems*, 29, 2016.

A Proof of Theorem 2

Proof. By linearity of the expectation and since $w_j, 1 \leq j \leq p$, are uniformly distributed on the sphere, we obviously have:

$$\mathbb{E}[\tilde{k}_{ORF}(x,y)] = \mathbb{E}[\cos(w_1^T(x-y)].$$

Moreover, we can find an orthogonal matrix $O_{x,y}$ such that

$$O_{x,y}(x-y) = ||x-y||e_1,$$

where e_1 is the first vector of the canonical basis of $\mathbb{R}^d$. Since the uniform measure on the sphere is invariant by orthogonal transformation, we further get:

$$\mathbb{E}[\tilde{k}_{ORF}(x,y)] = \mathbb{E}[\cos(||x-y||w_{11})],$$

where w_{11} is the first coordinate of w_1. This real random variable follows the beta distribution whose density is given by:

$$\frac{\Gamma(d/2)}{\sqrt{\pi}\Gamma((d-1)/2)}(1-u^2)^{(d-3)/2}, \quad -1 < u < 1.$$

Theorem 2 follows from the integral representation of the normalized Bessel function of the first kind:

$$j_\nu(z) = \frac{\Gamma(\nu+1)}{\sqrt{\pi}\Gamma(\nu+(1/2))}\int_{-1}^{1}\cos(zu)(1-u^2)^{\nu-1/2}, \quad \nu > -1/2.$$

B Proof of Proposition 3

Proof The normalized Bessel function $j_{d/2-1}$ admits an infinite number of positive simple zeros increasing to infinity:

$$0 < a_{d,1} < a_{d,2} < \ldots$$

As a matter of fact, $j_{d/2-1}$ has the following infinite product representation (Watson (1995), p.498):

$$j_{(d/2)-1}(z) = \prod_{j=1}^{\infty}\left(1 - \frac{z^2}{(a_{d,j})^2}\right). \tag{14}$$

Moreover, the first Rayleigh sum is given by Watson (1995, p. 502):

$$\sum_{j\geq 1}\frac{1}{(a_{d,j})^2} = \frac{1}{2d}. \tag{15}$$

Consequently, the inequality $1-u \leq e^{-u}, u \in [0,1]$ shows that if $z \leq a_{d,1}$ then

$$j_{(d/2)-1}(z) \leq e^{-\sum_{j\geq 1} z^2/(a_{d,j})^2} = e^{-z^2/(2d)} \tag{16}$$

yielding the upper bound. The latter remains true in the interval $[a_{d,1}, a_{d,2}]$ since the Bessel function is non positive there.

As to the lower bound, we first differentiate the function

$$h_d : z \mapsto e^{z^2/2} j_{(d/2)-1}(z), \quad 0 \leq z \leq a_{d,1},$$

and note that

$$(j_{(d/2)-1})'(z) = -\frac{z}{d} j_{(d/2)}(z).$$

It follows that

$$h_d'(z) = \frac{e^{z^2/2} j_{(d/2)-1}(z)}{d} \left(d - \frac{j_{(d/2)}(z)}{j_{(d/2)-1}(z)} \right).$$

From the Mittag-Leffler expansion (Ifantis & Siafarikas, 1990, eq. 2.9):

$$\frac{j_{(d/2)}(z)}{j_{(d/2)-1}(z)} = 2d \sum_{m=1}^{\infty} \frac{1}{a_{d,m}^2 - z^2}$$

the equation $h_d'(z) = 0, 0 < z < a_{d,1}$, is equivalent to

$$\sum_{m=1}^{\infty} \frac{1}{a_{d,m}^2 - z^2} = 1/2.$$

Since the LHS of the last equality is increasing in the z-variable and tends to $+\infty$ as $z \to a_{d,1}$, then the first Rayleigh sum equation 15 implies the existence of one and only one solution $z_0(d)$ to the equation $h_d'(z) = 0$ in $(0, a_{d,1})$. Consequently, $h_d(z) \geq 1$ for any $z \in [0, z_0(d)]$.

Now, in order to get a more precise information on $z_0(d)$, we appeal to the following inequality which readily follows from Theorem 2.1 in Freitas (2021):

$$\sum_{m=1}^{\infty} \frac{1}{a_{d,m}^2 - z^2} \leq \frac{1}{a_{d,1}^2 - z^2} + \frac{d}{4a_{d,1}^2},$$

to see that

$$\frac{1}{2} \leq \frac{1}{a_{d,1}^2 - [z_0(d)]^2} + \frac{d}{4a_{d,1}^2},$$

or equivalently,

$$z_0(d) \geq a_{d,1} \sqrt{1 - \frac{4}{2a_{d,1}^2 - d}} = f_d(a_{d,1}). \tag{17}$$

Finally, we recall the following lower bounds (Ismail & Muldoon, 1988, eq. 5.4):

$$a_{d,1} > \sqrt{2d} \left(\frac{d}{2} \right)^{1/4} = 2^{1/4} d^{3/4}, \quad d \geq 2, \tag{18}$$

and (Watson, 1995, eq. 5, p. 486)

$$a_{d,1} > \sqrt{\frac{d^2}{4} - 1}, \quad d \geq 2. \tag{19}$$

Since f_d is increasing for fixed d, then

$$z_0(d) > \max \left(f(2^{1/4} d^{3/4}) = b_d, f\left(\frac{d^2}{4} - 1 \right) = c_d \right)$$

which completes the proof.

Remark 7. *If we use the following inequality (Joshi & Bissu, 1996, eq. 2.6):*

$$1 - \frac{z^2}{2d} \leq j_{(d/2)-1}(z),$$

we can prove that the lower bound holds in the interval $[0, \sqrt{d}]$.

C Proof of Theorem 4

Proof Let us consider the variance of $K(x, y)$:

$$\begin{aligned} V\left(\tilde{k}_{ORF}(x,y)\right) &= \frac{1}{p^2} V\left[\sum_{j=1}^{p} \cos(w_j^T(x-y))\right] \\ &= \frac{1}{p} V\left[\cos(w_1^T(x-y))\right] + \frac{p-1}{p} \operatorname{cov}\left[\cos(w_1^T(x-y)), \cos(w_2^T(x-y))\right]. \end{aligned}$$

Now,

$$\begin{aligned} V\left[\cos(w_1^T(x-y))\right] &= \mathbb{E}[\cos^2(w_1^T(x-y))] - \left[\mathbb{E}[\cos(w_1^T(x-y))]\right]^2 \\ &= \frac{1}{2} + \frac{j_{(d/2)-1}(2z)}{2} - \left(j_{(d/2)-1}(z)\right)^2. \end{aligned}$$

As to the covariance term, the invariance of the Haar distribution entails:

$$\begin{aligned} \operatorname{cov}\left[\cos(w_1^T(x-y)), \cos(w_2^T(x-y))\right] &= \mathbb{E}[\cos(w_1^T(x-y))\cos(w_2^T(x-y))] - \left(j_{(d-2)/2}(z)\right)^2 \\ &= \mathbb{E}[\cos(w_{11}z)\cos(w_{12}z)] - \left(j_{(d-2)/2}(z)\right)^2 \\ &= \frac{1}{2}\left\{\mathbb{E}[\cos((w_{11}+w_{12})z)] + \mathbb{E}[\cos((w_{11}-w_{12})z)]\right\} \\ &- \left(j_{(d/2)-1}(z)\right)^2 \end{aligned}$$

where w_{11} and w_{12} are the first coordinates of the column vectors w_1 and w_2. But (w_{11}, w_{12}) is the first row of the Haar orthogonal matrix O which is uniformly distributed on S^{d-1}. Consequently, the joint distribution of (w_{11}, w_{12}) is given by the following probability density:

$$\frac{d-2}{2\pi}(1-u^2-v^2)^{(d/2)-2}\mathbf{1}_{\{u^2+v^2<1\}}$$

with respect to Lebesgue measure $du\,dv$, whence

$$\operatorname{cov}\left[\cos(w_1^T(x-y)), \cos(w_2^T(x-y))\right] = \mathbb{E}[\cos((w_{11}+w_{12})z)] - \left(j_{(d/2)-1}(z)\right)^2.$$

Moving to polar coordinates:

$$w_{11} = r\cos(\theta), \quad w_{12} = r\sin(\theta),$$

it follows that

$$\mathbb{E}[\cos((w_{11}+w_{12})z)] = \frac{d-2}{2\pi}\int_0^1\int_0^{2\pi}[\cos(\cos\theta+\sin\theta)rz]r(1-r^2)^{(d/2)-2}drd\theta. \tag{20}$$

Expanding further the cosine into power series, we are left with the following two integrals:

$$\int_0^1 r^{2j+1}(1-r^2)^{(d/2)-2}dr = \frac{\Gamma(j+1)\Gamma((d/2)-1)}{2\Gamma(j+(d/2))}, \quad j \geq 0, \tag{21}$$

and (equation 3.66.1.2., p. 405 in Gradshteyn & Ryzhik, 2014):

$$\int_0^{2\pi}(\cos\theta+\sin\theta)^{2j}d\theta = 2\pi\frac{2^j(2j-1)!!}{(2j)!!}$$

where $(2j)!! = (2j)(2j-2)\cdots(2)$ is the double factorial and likewise $(2j-1)!! = (2j-1)(2j-3)\cdots(3)(1)$. Writing

$$(2j)!! = 2^j j!, \quad (2j-1)!! = \frac{(2j)!}{2^j j!},$$

we equivalently get:

$$\int_0^{2\pi} (\cos\theta + \sin\theta)^{2j} d\theta = 2\pi \frac{(2j)!}{2^j (j!)^2}. \tag{22}$$

Gathering equation 20, equation 21 and equation 22, we end up with the expression:

$$\begin{aligned}\mathbb{E}[\cos((w_{11} + w_{12})z)] &= (d-2)\Gamma((d-2)/2) \sum_{j\geq 0} \frac{(-1)^j z^{2j}}{2^j j! \Gamma(j + (d/2))} \\ &= j_{(d/2)-1}(\sqrt{2}z),\end{aligned}$$

where we used the formula $(d-2)\Gamma((d-2)/2) = 2\Gamma(d/2)$. Finally

$$\begin{aligned}V\left(\tilde{k}_{ORF}(x,y)\right) &= \frac{1}{p}\left\{\frac{1 + j_{(d/2)-1}(2z)}{2} - \left(j_{(d/2)-1}(z)\right)^2\right\} + \frac{p-1}{p}\left\{j_{(d/2)-1}(\sqrt{2}z) - \left(j_{(d/2)-1}(z)\right)^2\right\} \\ &= \frac{1}{p}\left\{\frac{1 + j_{(d/2)-1}(2z)}{2} + (p-1) j_{(d/2)-1}(\sqrt{2}z) - p\left(j_{(d/2)-1}(z)\right)^2\right\},\end{aligned}$$

as desired.

D Proof of Proposition 6

Proof The infinite product equation 14 and the inequality

$$1 - 2u \leq (1-u)^2, \quad u \geq 0,$$

implies that the covariance term computed above is non positive:

$$(p-1)\left[j_{(d/2)-1}(\sqrt{2}z) - \left(j_{(d/2)-1}(z)\right)^2\right] \leq 0$$

on the interval $[0, a_{d,1}/\sqrt{2}]$. Consequently,

$$V\left(K(x,y)\right) \leq \frac{1}{p}\left[\frac{1 + j_{(d/2)-1}(2z)}{2} - \left(j_{(d/2)-1}(z)\right)^2\right].$$

Similarly,

$$(1 - 4u) \leq (1-u)^4, \quad u \geq 0,$$

holds since the discriminant of

$$u^2 - 4u + 6, \quad u \geq 0,$$

is negative. Appealing again to equation 14 entails:

$$j_{(d/2)-1}(2z) \leq \left[j_{(d/2)-1}(z)\right]^4$$

on the interval $[0, a_{d,1}/2]$. Keeping in mind the lower bound derived in Proposition 3 which remains valid in $[0, z_0(d)]$ (see the proof), we get on the interval $[0, \inf(z_0(d), a_{d,1}/2)]$

$$\begin{aligned}\frac{1 + j_{(d/2)-1}(2z)}{2} - \left(j_{(d/2)-1}(z)\right)^2 &\leq \frac{1 + \left[j_{(d/2)-1}(z)\right]^4 - 2\left(j_{(d/2)-1}(z)\right)^2}{2} \\ &= \frac{\left[1 - \left(j_{(d/2)-1}(z)\right)^2)\right]^2}{2} \\ &\leq \frac{[1 - e^{-z^2}]^2}{2}.\end{aligned}$$

Furthermore, by the virtue of equation 17, we are led to compare

$$\sqrt{1 - \frac{4}{2a_{d,1}^2 - d}}$$

and the value 1/2. In this respect, the lower bound given in Ismail & Muldoon (1988), eq. (5.4), shows that

$$\sqrt{1 - \frac{4}{2a_{d,1}^2 - d}} > 1/2,$$

whence we infer that $z_0(d) > a_{d,1}/2$. Consequently, $V[\tilde{k}_{ORF}(x,y)] \leq V[\tilde{k}_{RFF}(x,y)]$ holds true for any $z = \|x - y\| \in [0, a_{d,1}/2]$. Recalling equation 18 and equation 19, we are done.