Simplex Random Features

Isaac Reid [1] **Krzysztof Choromanski** [*2 3] **Valerii Likhosherstov** [1] **Adrian Weller** [*1 4]

Abstract

We present *Simplex Random Features* (SimRFs), a new random feature (RF) mechanism for unbiased approximation of the softmax and Gaussian kernels by geometrical correlation of random projection vectors. We prove that SimRFs provide the smallest possible mean square error (MSE) on unbiased estimates of these kernels among the class of weight-independent geometrically-coupled positive random feature (PRF) mechanisms, substantially outperforming the previously most accurate *Orthogonal Random Features* (ORFs, Yu et al., 2016) at no observable extra cost. We present a more computationally expensive *SimRFs+* variant, which we prove is asymptotically optimal in the broader family of weight-dependent geometrical coupling schemes (which permit correlations between random vector directions and norms). In extensive empirical studies, we show consistent gains provided by SimRFs in settings including pointwise kernel estimation, nonparametric classification and scalable Transformers (Choromanski et al., 2020).[1]

1. Introduction

Embedding methods, which project feature vectors into a new space, are ubiquitous in machine learning. The canonical example is the Johnson-Lindenstrauss Transform (JLT) (Johnson, 1984; Dasgupta et al., 2010; Kane & Nelson, 2014; Kar & Karnick, 2012), where a collection of high-dimensional points is embedded in a much lower dimensional space whilst (approximately) preserving their metric relationships, e.g. distances and dot-products. Another application is found in kernel approximation (Liu et al., 2022; Yang et al., 2014; Pennington et al., 2015; Li et al., 2010), where the nonlinear similarity measure (kernel) in the original space is translated to a linear kernel in the latent space. For example, a kernel $K(\cdot,\cdot):\mathbb{R}^d\times\mathbb{R}^d\rightarrow\mathbb{R}$ can be approximated using so-called *random features* (RFs): randomised nonlinear transformations $\phi(\cdot):\mathbb{R}^d\rightarrow\mathbb{R}^{d'}$ constructed such that

$$K(\boldsymbol{x},\boldsymbol{y})=\mathbb{E}[\widehat{K}(\boldsymbol{x},\boldsymbol{y})],\ \text{where}\ \widehat{K}(\boldsymbol{x},\boldsymbol{y})\stackrel{\text{def}}{=}\phi(\boldsymbol{x})^\top\phi(\boldsymbol{y}). \tag{1}$$

Provided K is stationary, meaning $K(\boldsymbol{x},\boldsymbol{y})=K(\boldsymbol{x}-\boldsymbol{y})$, we can use Bochner's theorem to write

$$K(\boldsymbol{x}-\boldsymbol{y})=\int_{\mathbb{R}^d}p(\boldsymbol{w})e^{i\boldsymbol{w}^\top(\boldsymbol{x}-\boldsymbol{y})}\mathrm{d}^d\boldsymbol{w}, \tag{2}$$

where $p(\mathbf{w})$ is the Fourier transform of K. If K is positive semidefinite, $p(\mathbf{w})$ is non-negative so we can treat it as a probability density. This invites Monte Carlo (MC) sampling, yielding *Random Fourier Features* (RFFs) of the following form, where vectors $\boldsymbol{w}_i$ are sampled from $p(\boldsymbol{w})$, m is their number and $\odot$ denotes concatenation (Rahimi & Recht, 2007; 2008):

$$\phi_{\mathrm{RFF}}(\boldsymbol{z})\stackrel{\text{def}}{=}\sqrt{\frac{1}{m}}(\odot_{i=1}^m[\sin(\boldsymbol{w}_i^\top\boldsymbol{z}),\cos(\boldsymbol{w}_i^\top\boldsymbol{z})])^\top. \tag{3}$$

Furthermore, if K is a Gaussian kernel, defined by

$$K_{\mathrm{gauss}}(\boldsymbol{x},\boldsymbol{y})\stackrel{\text{def}}{=}\exp(-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|_2^2}{2}), \tag{4}$$

random vectors $\boldsymbol{w}_i$ are sampled from the multivariate Gaussian distribution $\mathcal{N}(0,\mathbf{I}_d)$. Another kernel, of key interest in Transformer architectures (Vaswani et al., 2017; Choromanski et al., 2020), is the so-called *softmax kernel*:

$$K_{\mathrm{smax}}(\boldsymbol{x},\boldsymbol{y})\stackrel{\text{def}}{=}\exp(\boldsymbol{x}^\top\boldsymbol{y}). \tag{5}$$

Since $K_{\mathrm{gauss}}(\boldsymbol{x},\boldsymbol{y})=K_{\mathrm{smax}}(\boldsymbol{x},\boldsymbol{y})\exp(-\frac{x^2}{2}-\frac{y^2}{2})$, RF mechanisms for the Gaussian kernel can be readily converted into the corresponding mechanism for softmax and vice versa (Likhosherstov et al., 2022). Our results will hence apply to both settings. For brevity, we will mostly refer to K_{gauss}.

However, as noted in (Choromanski et al., 2020), RFFs lead to unstable training of implicit linear-attention Transformers.

*Equal senior co-leads [1]University of Cambridge [2]Google [3]Columbia University [4]Alan Turing Institute. Correspondence to: Isaac Reid <ir337@cam.ac.uk>, Krzysztof Choromanski <kchoro@google.com>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

[1]Code is available at https://github.com/isaac-reid/simplex_random_features.

The authors address this by proposing *Positive Random Features* (PRFs), defined by

$$K_{\text{gauss}}(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{E}[\phi_{\text{PRF}}(\boldsymbol{x})^\top \phi_{\text{PRF}}(\boldsymbol{y})], \tag{6}$$

where for $\boldsymbol{w}_1, ..., \boldsymbol{w}_m \sim \mathcal{N}(0, \mathbf{I}_d)$,

$$\phi_{\text{PRF}}(\boldsymbol{z}) \stackrel{\text{def}}{=} \sqrt{\frac{1}{m}} \exp(-\|\boldsymbol{z}\|_2^2)(\odot_{i=1}^m [\exp(\boldsymbol{w}_i^\top \boldsymbol{z})])^\top. \tag{7}$$

The straightforward implementation of PRFs (and RFFs) draws $\boldsymbol{w}_i$ independently – a strategy we refer to as *IIDRFs*. However, the isotropy of the Gaussian distribution permits us to entangle different $\boldsymbol{w}_i$ to be exactly orthogonal[2] whilst preserving the Gaussian marginal distributions $\boldsymbol{w}_i \sim \mathcal{N}(0, \mathbf{I}_d)$ (Yu et al., 2016). This mechanism is referred to as *Orthogonal Random Features* (ORFs), and is an example of a *weight-independent geometrically-coupled RF mechanism*.

Definition 1.1. Consider the random vectors $\{\boldsymbol{w}_i | i \leq m\} \subset \mathbb{R}^d$, which can be described by norms $w_i = \|\boldsymbol{w}_i\|_2$ and directions $\widehat{\boldsymbol{w}}_i = \frac{\boldsymbol{w}_i}{\|\boldsymbol{w}_i\|_2}$. An RF mechanism is described as *geometrically-coupled* if the norms of random vectors $\{w_i\}$ are independent, but the directions $\{\widehat{\boldsymbol{w}}_i\}$ are permitted to be correlated with one another and with the norms $\{w_i\}$. Such a coupling is *weight-independent* under the further restriction that directions $\{\widehat{\boldsymbol{w}}_i\}$ are independent of the norms $\{w_i\}$.

Unless otherwise stated, all coupling mechanisms considered in this work will be geometrical. ORFs provide a lower mean squared error (MSE) on Gaussian kernel approximation than IIDRFs (Yu et al., 2016; Choromanski et al., 2020), though for RFFs only at asymptotically large d. ORFs are used in a broad range of applications including kernel ridge regression and Transformers. In the latter case, they offer linear (cf. quadratic) space- and time-complexity of the attention module, enabling efficient long-range attention modelling as part of the so-called Performer architecture (Choromanski et al., 2020). Sec. 2 details further applications beyond Gaussian and softmax kernel estimation. Recently Likhosherstov et al. (2022) showed that further MSE reduction (for fixed m and preserving unbiasedness) can be achieved by collecting light data statistics. RFs can also be applied with more computationally expensive preprocessing to improve accuracy in downstream tasks (Trokicic & Todorovic, 2019), but they no longer approximate the Gaussian kernel.

However, the following question remains open: *do ORFs provide the lowest possible MSE on unbiased estimates of the Gaussian kernel among the class of weight-independent geometrically-coupled PRF mechanisms?*

[2]All $\boldsymbol{w}_i$ can be orthogonal if $m \leq d$. If $m > d$ we construct ensembles of independent orthogonal blocks.

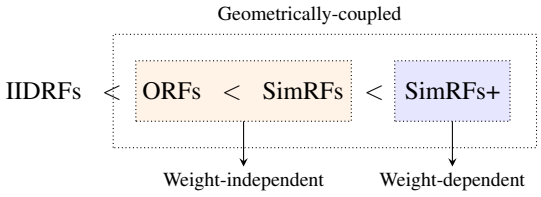

Figure 1. Schematic of performance of RF mechanisms described in this manuscript. SimRFs and SimRFs+ are novel.

Here, we comprehensively answer this question, finding that ORFs are *not* optimal. We derive the optimal mechanism, coined *Simplex Random Features* (SimRFs), and show that it substantially outperforms ORFs at close to no extra computational cost. We also consider the broader family of weight-*dependent* geometrically-coupled PRFs, where random vector directions $\{\widehat{\boldsymbol{w}}_i\}$ can be correlated with norms $\{w_i\}$, and present a *SimRFs+* variant which we prove is asymptotically optimal in this more general class. Our empirical studies demonstrate the consistent gains provided by SimRFs in diverse settings, including pointwise kernel estimation, nonparametric classification and scalable Transformers (Choromanski et al., 2020).

In more detail, our principal contributions are as follows:

1. In Sec. 3, we introduce SimRFs and prove that they provide the lowest kernel estimator MSE of any weight-independent geometrically-coupled PRF mechanism, outperforming the previously most accurate ORFs. We demonstrate that a fast, simple scheme applying minor alterations to SimRFs yields SimRFs+: a marginally better weight-dependent mechanism. See Fig. 1.

2. In Sec. 4, we provide novel theoretical results to add insight to the discussion in Sec. 3. They may be of independent interest. We derive the first non-asymptotic **closed-form** formulae for the MSE for PRFs in the IIDRF, ORF and SimRF settings, and show how it is straightforward to generalise some of these forms to RFFs. This allows us to precisely quantify how much the kernel estimator MSE can be suppressed by geometrical coupling. We also compare the time- and space-complexities of the different PRF mechanisms and describe a faster, approximate implementation.

3. In Sec. 5, we support our theoretical results with comprehensive experiments, demonstrating the superiority of SimRFs over ORFs and IIDRFs. We empirically confirm that they offer lower kernel estimator MSE, and find that this translates to better downstream performance in nonparametric classification tasks (Sec. 5.3) and scalable Transformers (Sec. 5.4).

Proofs not provided in the main body are in Appendix A.

2. Related Work

The literature on *structured* RFs, where random vectors are conditionally dependent, is extensive (Ailon & Chazelle, 2009; Liberty et al., 2011; Ailon & Liberty, 2013; Le et al., 2013; Yu et al., 2017). ORFs were first proposed for nonlinear kernel estimation in (Yu et al., 2016), where the authors derived strict asymptotic gains from ORFs compared to IIDRFs when using RFFs for Gaussian kernel approximation. We refer to this phenomenon – the supression of kernel estimator MSE when random features are conditioned to be orthogonal – as *the orthogonality gap*.

Further progress towards an understanding of the orthogonality gap was provided in (Choromanski et al., 2018), where the authors introduced and studied the so-called *charm property* of stationary kernels. However, a rigorous mathematical analysis in the non-asymptotic setting remained out of reach. In (Choromanski et al., 2017), the authors showed the superiority of ORFs over IIDRFs for angular kernel estimation in any d (not just asymptotic) and conducted an extensive analysis of the linear (dot-product) kernel, but they did not address stationary kernels. The authors of (Lin et al., 2020) used the lens of *determinantal point processes* and the negative dependence property (Kulesza & Taskar, 2012) to explore the efficacy of ORFs.

ORFs are used with PRFs in Performers (Choromanski et al., 2020; Schlag et al., 2021; Luo et al., 2021; Likhosherstov et al., 2021; Chowdhury et al., 2021; Xiao et al., 2022): a recently-proposed class of efficient Transformer (Kitaev et al., 2020; Roy et al., 2021) that can be applied to ultra-long sequences or to expedite inference on regular-size sequences.

3. Simplex Random Features (SimRFs)

In this section, we describe our core contributions.

We begin by presenting Simplex Random Features (SimRFs). In analogy to the square orthogonal block, we define the so-called *simplex block*, consisting of d d-dimensional random vectors $\{\boldsymbol{w}_i | i \leq d\}$. In practical applications where $m > d$ random features are needed, multiple simplex blocks are constructed independently.

Instead of being orthogonal, the rows of the simplex block point towards the vertices of a $d-1$-dimensional simplex embedded in d-dimensional space, subtending angles $\theta = \arccos(-\frac{1}{d-1})$. The entire simplex (or, equivalently, the vector it operates on) is randomly rotated to preserve isotropy, and the rows are independently renormalised by weights $w_i \sim \chi_d$ such that they are marginally Gaussian. Explicitly, we define the simplex block $\mathbf{W}_{\text{simp}} \in \mathbb{R}^{d \times d}$ by

$$\mathbf{W}_{\text{simp}} = \mathbf{DSR} \tag{8}$$

where $\mathbf{D} \in \mathbb{R}^{d \times d} = \text{diag}(w_i)$ with w_i sampled from a χ_d-distribution. $\mathbf{R} \in \mathbb{R}^{d \times d}$ is a random orthogonal matrix drawn from Haar measure on $\text{O}(d)$, the group of orthogonal matrices in $\mathbb{R}^{d \times d}$, constructed e.g by Gram-Schmidt orthogonalisation of an unstructured Gaussian matrix (Yu et al., 2016). The rows $\boldsymbol{s}_i$ of the simplex projection matrix $\mathbf{S} \in \mathbb{R}^{d \times d}$ are given by the unit vectors

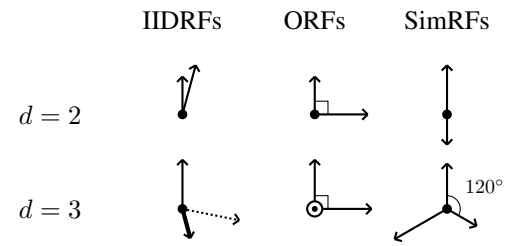

Figure 2. Schematic of different geometrical couplings for small d. Dotted lines have a component into the plane of the paper, thick lines have a component out, and $\odot$ is purely out (i.e. perpendicular to the paper's plane). With IIDRFs, the respective orientations of vectors are chosen independently. With ORFs, we condition the vectors to be perpendicular. With SimRFs, they subtend angles $\theta = \arccos(-\frac{1}{d-1})$. Intuitively, conditioning the vectors to subtend fixed, obtuse angles means they 'explore' $\mathbb{R}^d$ better, suppressing the kernel estimator MSE. All norms are drawn independently from a χ_d-distribution.

$$\boldsymbol{s}_i = \begin{cases} \sqrt{\frac{d}{d-1}}\mathbf{e}_i - \frac{\sqrt{d}+1}{(d-1)^{3/2}}(1, ..., 1, 0)^\top & \text{for } 1 \leq i < d \\ \frac{1}{\sqrt{d-1}}(1, 1, ..., 1, 0)^\top & \text{for } i = d \end{cases} \tag{9}$$

which are manifestly normalised and subtend obtuse angles. Fig. 2 visualises the different geometrical couplings of IIDRFs, ORFs and SimRFs in low data dimensionality d.

3.1. RF-Conformity and SimRFs vs ORFs

Recalling again that the Gaussian and softmax kernels are readily interchanged, we focus on K_{gauss} without loss of generality. We begin by defining the *RF-conformity*.

Definition 3.1. The RF-conformity, $\rho(\boldsymbol{x}, \boldsymbol{y})$, is given by

$$\rho(\boldsymbol{x}, \boldsymbol{y}) \stackrel{\text{def}}{=} \frac{\Gamma(\frac{d}{2})}{m(m-1)} \sum_{i,j \neq i} \mathbb{E}_{w_{ij}} \left(\sum_{k=0}^{\infty} \frac{v^{2k} w_{ij}^{2k}}{2^{2k} k! \Gamma(k + \frac{d}{2})} \right), \tag{10}$$

with $w_{ij} = \|\boldsymbol{w}_i + \boldsymbol{w}_j\|_2$, $v = \|\boldsymbol{x} + \boldsymbol{y}\|_2$ for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, Γ the Gamma-function and m the no. random vectors $\boldsymbol{w}_i$.

$\rho(\boldsymbol{x}, \boldsymbol{y})$ depends on correlations induced between random vector directions. It is bigger when random vectors point in similar directions, 'exploring' $\mathbb{R}^d$ less effectively. In Appendix A.1, we prove the following important result.

Theorem 3.2 (MSE depends on RF-conformity)**.** *For PRFs, the* MSE *of the unbiased estimator* $\widehat{K}(\boldsymbol{x}, \boldsymbol{y})$ *is given by*

$$\text{MSE}(\widehat{K}) = \frac{e^{-2x^2-2y^2}}{m} \Big((e^{2v^2} - e^{v^2}) + (m-1)(\rho(\boldsymbol{x}, \boldsymbol{y}) - e^{v^2}) \Big). \tag{11}$$

That is, the MSE is an increasing function of the RF-conformity.

For any w_i, w_j, SimRFs give strictly smaller values of w_{ij} than ORFs because the random vectors subtend a bigger angle. Explicitly, $w_{ij} = (w_i^2 + w_j^2 + 2w_i w_j \cos\theta)^{1/2}$ is smaller when $\cos\theta = -\frac{1}{d-1}$ (SimRFs) compared to when $\cos\theta = 0$ (ORFs). This leads to smaller values of $\rho(\boldsymbol{x}, \boldsymbol{y})$, which immediately implies the following important result.

Corollary 3.3 (SimRFs outperform ORFs)**.** *For PRFs, the kernel estimator MSE obtained with SimRFs is strictly lower than with ORFs for arbitrary data dimensionality d.*

In fact, we are able to make the following substantially stronger statement, proved in Appendix A.2.

Theorem 3.4 (SimRFs optimal for weight-independent geometrical coupling)**.** *Supposing that d random vector norms $\{w_i | i \leq d\}$ are i.i.d., SimRFs constitute the* ***best possible weight-independent geometrical coupling mechanism****, giving the lowest possible PRF kernel estimator MSE.*

3.2. SimRFs+

Now we consider the broader family of weight-*dependent* geometrical coupling mechanisms, where random vector directions $\{\hat{\boldsymbol{w}}_i\}$ are permitted to be correlated with norms $\{w_i\}$. In particular, given d vectors $\{\boldsymbol{w}_i\}$ of known norms (from d draws of χ_d), we would like to arrange them in d-dimensional space in order to minimise the sum[3]

$$\rho(\boldsymbol{x}, \boldsymbol{y}) = \frac{\Gamma(\frac{d}{2})}{m(m-1)} \sum_{i,j\neq i} \left(\sum_{k=0}^{\infty} \frac{v^{2k} w_{ij}^{2k}}{2^{2k} k! \Gamma(k + \frac{d}{2})} \right). \quad (12)$$

One brute-force approach is to parameterise each of the d random vector directions in hyperspherical coordinates and use an off-the-shelf numerical optimiser (e.g. scipy.optimize). This is prohibitively slow, and moreover the solution has data-dependence via $v = \|\boldsymbol{x} + \boldsymbol{y}\|_2$ which frustrates the method's scalability: the optimisation needs to be carried out pairwise for every $(\boldsymbol{x}, \boldsymbol{y})$, which undermines our ability to quickly evaluate $\widehat{K}(\boldsymbol{x}, \boldsymbol{y}) = \phi(\boldsymbol{x})^\top \phi(\boldsymbol{y})$ for any given pair of input vectors. However, the numerical approach does benchmark the lowest possible RF-conformity that can be achieved with weight-dependent geometrical coupling.

The generic *analytic* minimisation of Eq. 12 is challenging, and solutions will suffer the same v-dependence described above, so we instead consider a tractable approximation. Dropping constant prefactors for clarity, the first few terms from Eq. 10 are given by:

$$\sum_{i,j\neq i} \mathbb{E}_{w_{ij}} \left(\frac{1}{\Gamma(\frac{d}{2})} + \frac{v^2 w_{ij}^2}{4\Gamma(\frac{d}{2}+1)} + \frac{v^4 w_{ij}^4}{32\Gamma(\frac{d}{2}+2)} + ... \right)$$
$$= \frac{1}{\Gamma(\frac{d}{2})} \sum_{i,j\neq i} 1 + \tau \left(1 + \frac{v^2}{8} \frac{\Gamma(\frac{d}{2}+1)}{\Gamma(\frac{d}{2}+2)} \frac{\mathbb{E}(w_{ij}^4)}{\mathbb{E}(w_{ij}^2)} + ... \right) \quad (13)$$

with $\tau = \frac{\Gamma(\frac{d}{2}) v^2 \mathbb{E}(w_{ij}^2)}{4\Gamma(\frac{d}{2}+1)}$. The precise value of $\frac{\mathbb{E}(w_{ij}^4)}{\mathbb{E}(w_{ij}^2)}$ will depend on the geometrical coupling scheme employed, but for the types we have considered we generally expect it to scale as $\sim d$, with some constant prefactor[4]. Therefore the sum in Eq. 10 can be approximated by:

$$\frac{1}{\Gamma(\frac{d}{2})} \sum_{i,j\neq i} 1 + \frac{\Gamma(\frac{d}{2}) v^2 \mathbb{E}(w_{ij}^2)}{4\Gamma(\frac{d}{2}+1)} \left(1 + \mathcal{O}(v^2) + ...\right). \quad (14)$$

In the limit of small v, this invites us to truncate the sum at $k = 1$, dropping the $\mathcal{O}(v^2)$ terms. Omitting additive constants, we are left with the approximate objective

$$\tilde{\rho}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\Gamma(d/2) v^2}{4m(m-1)\Gamma(1+d/2)} \sum_{i,j\neq i} w_{ij}^2, \quad (15)$$

the physical analogue of which is the Heisenberg Hamiltonian with different coupling constants between different spin pairs. This is exactly minimised by

$$\boldsymbol{w}_i = -\frac{\sum_{j\neq i} \boldsymbol{w}_j}{\|\sum_{j\neq i} \boldsymbol{w}_j\|_2} w_i \qquad i = 1, ..., d \quad (16)$$

where each random vector points away from the resultant of all the others (see Appendix A.3 for details). Fig. 3 captures this essential difference between SimRFs and SimRFs+: in the latter case, vectors with larger norms subtend bigger angles. Empirically, we find that the iterative update scheme

$$\boldsymbol{w}_i \leftarrow -\frac{\sum_{j\neq i} \boldsymbol{w}_j}{\|\sum_{j\neq i} \boldsymbol{w}_j\|_2} w_i \quad (17)$$

converges to Eq. 16 quickly (after a small number of passes through the set of d vectors), especially if we initialise in the near-optimal simplex geometry. Conveniently, the solution has no v-dependence and is therefore scalable: the optimisation needs to be carried out for every draw of weights $\{w_i\}$ but *not* every pair of data points $(\boldsymbol{x}, \boldsymbol{y})$. We refer to this mechanism of weight-dependent geometrical coupling as *SimRFs+*, and emphasise that it is asymptotically optimal (in the sense of minimising $\rho(\boldsymbol{x}, \boldsymbol{y})$) in the $v \ll 1$ limit.

[3] We remove the expectation value because, given a fixed set of norms, assigning any probability mass to suboptimal configurations will increase the RF-conformity in expectation – that is, the best geometrical coupling between vectors of known magnitudes $\{w_i\}$ is deterministic.

[4] For example, with orthogonal coupling $\frac{\mathbb{E}(w_{ij}^4)}{\mathbb{E}(w_{ij}^2)} = \frac{\mathbb{E}(w_i^4 + w_j^4 + 2w_i^2 w_j^2)}{\mathbb{E}(w_i^2 + w_j^2)} = \frac{\mathbb{E}(w_i^4)}{\mathbb{E}(w_i^2)} + \mathbb{E}(w_i^2) = 2\frac{\Gamma(\frac{d}{2}+2)}{\Gamma(\frac{d}{2}+1)} + 2\frac{\Gamma(\frac{d}{2}+1)}{\Gamma(\frac{d}{2})} \sim d$, where we took moments of the χ_d distribution. We can perform similar analyses in the i.i.d. and simplex cases.

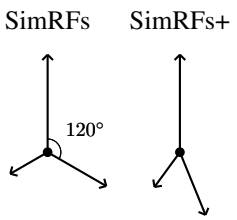

Figure 3. With SimRFs, random vectors are geometrically correlated such that all pairs subtend an equal angle $\theta = \arccos(-\frac{1}{d-1})$. With SimRFs+, random vectors with bigger norms subtend bigger angles, guaranteeing smaller kernel estimator MSE when v is sufficiently small.

Fig. 4 compares the RF-conformity of the mechanisms we have considered, as well as the outcome of the inefficient numerical optimisation. The additional benefits of weight-dependent coupling are marginal: SimRFs+ access only slightly lower conformity than SimRFs at the expense of an extra optimisation step of time-complexity $\mathcal{O}(d^3)$. This gives context to the excellent performance of SimRFs; they can compete with members of a much broader class at a fraction of the computational cost. We also note that the minimisation of the truncated objective (SimRFs+) is a good approximation to the minimisation of the true objective ('numerically optimised'), accessing comparably small values of ρ. Informally, SimRFs+ are close to optimal among the class of weight-dependent geometrically-coupled PRF mechanisms.

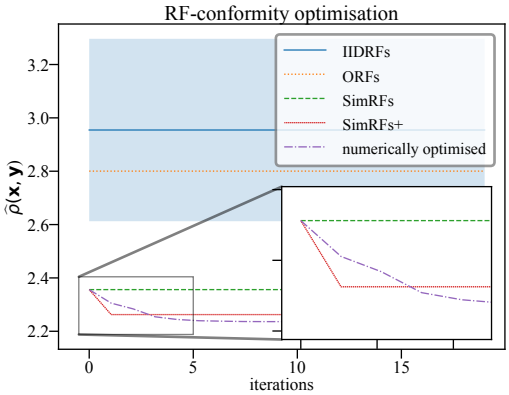

Figure 4. Comparison of the RF-conformity defined in Eq. 10 (lower is better) for a *single random draw* of norms $\{w_i\}$, $v = \|\boldsymbol{x}+\boldsymbol{y}\|_2 = 1$ and $d = 6$. IIDRFs, ORFs, SimRFs and SimRFs+ are implemented as described in the main text. 'Numerically optimised' uses an off-the-shelf numerical optimiser to arrange vectors to minimise the RF-conformity: a scheme which is too computationally inefficient to be practical but benchmarks the lowest possible value. Any improvements above SimRFs using weight-dependent geometrical coupling are marginal. The IIDRF value is averaged over 100 random couplings of fixed weights, and the shaded region gives 1 standard deviation.

4. From ORFs to SimRFs: the Theory

This section provides more detailed theoretical analysis to add insight to the results of Sec. 3. It can safely be omitted on a quick reading. We derive analytic expressions for the RF-conformity $\rho(\boldsymbol{x}, \boldsymbol{y})$, and therefore the kernel estimator MSE, for IIDRFs, ORFs and SimRFs. This allows us to quantitatively compare the performance of different coupling mechanisms. As before, we specialise to K_{gauss}. Detailed proofs are provided in Appendix A.

We have seen that RF-conformity depends on an expectation value over $w_{ij} = \|\boldsymbol{w}_i + \boldsymbol{w}_j\|_2$. This motivates us to begin with the following auxiliary lemma.

Lemma 4.1 (IIDRF conformity)**.** *When random vectors $\boldsymbol{w}_i, \boldsymbol{w}_j \in \mathbb{R}^d$ are i.i.d. (IIDRFs), the probability distribution $p(w_{ij})$ with $w_{ij} = \|\boldsymbol{w}_i + \boldsymbol{w}_j\|_2$ is given by*

$$p_{i.i.d.}(w_{ij}) = \frac{w_{ij}^{d-1} e^{-w_{ij}^2/4}}{2^{d-1}\Gamma(\frac{d}{2})} \tag{18}$$

which induces an RF-conformity

$$\rho_{\text{IIDRF}}(\boldsymbol{x}, \boldsymbol{y}) = e^{v^2} \tag{19}$$

where $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ and $v = \|\boldsymbol{x} + \boldsymbol{y}\|_2$.

Now we make the following important observation.

Lemma 4.2 (PDF for vectors subtending θ)**.** *Supposing random vectors $\boldsymbol{w}_i, \boldsymbol{w}_j$ are marginally Gaussian but are conditioned to subtend a fixed angle θ, the probability distribution $p_\theta(w_{ij})$, is given by*

$$\frac{w^{2d-1}}{2^{d-2}\Gamma(\frac{d}{2})^2} \int_{\phi=0}^{\pi/2} d\phi (\sin\phi\cos\phi)^{d-1} \frac{e^{-\frac{w^2}{2(1+\sin 2\phi\cos\theta)}}}{(1+\sin 2\phi\cos\theta)^d}. \tag{20}$$

ORFs and SimRFs correspond to special instances of this with $\cos\theta = 0$ and $\cos\theta = -\frac{1}{d-1}$, respectively. It is instructive to observe that, in the orthogonal case, the distribution reduces to the χ_{2d}-distribution. The probability distribution $p_\theta(w_{ij})$ induces an RF-conformity

$$\rho_\theta(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2^{d-1}\Gamma(\frac{d}{2})} \int_0^\pi \mathrm{d}\phi (\sin\phi)^{d-1} \cdot \sum_{k=0}^{\infty} \frac{v^{2k}(1+\sin\phi\cos\theta)^k}{2^k k! \Gamma(k+\frac{d}{2})} \Gamma(k+d). \tag{21}$$

Inspecting the form closely, we see that every term in the sum over k is proportional to the integral

$$\int_0^\pi \mathrm{d}\phi (\sin\phi)^{d-1}(1+\sin\phi\cos\theta)^k \tag{22}$$

which is strictly smaller for $\cos\theta < 0$ compared to $\cos\theta = 0$ (since $\sin\phi$ is nonnegative everywhere in the domain).

Since every term in the sum is positive, we immediately conclude that for PRFs the conformity of SimRFs is strictly smaller than ORFs, and hence the MSE is smaller. We already derived this in Sec. 3, but are now also able to provide the following closed forms.

Theorem 4.3 (ORF and SimRF conformity closed forms)**.** *For PRFs with $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, the RF-conformity of ORFs is*

$$\rho_{\mathrm{ORF}}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\Gamma(\frac{d}{2})}{\Gamma(d)} \sum_{k=0}^{\infty} \frac{v^{2k}}{2^k k!} \frac{\Gamma(k+d)}{\Gamma(k+\frac{d}{2})} \tag{23}$$

whereas the RF-conformity of SimRFs is

$$\rho_{\mathrm{SimRF}}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sqrt{\pi}}{\Gamma(\frac{d}{2}) 2^{d-1}} \sum_{k=0}^{\infty} \frac{\Gamma(k+d)}{\Gamma(k+\frac{d}{2})} \frac{v^{2k}}{2^k} \cdot \sum_{p=0}^{k} \left(-\frac{1}{d-1}\right)^p \frac{\Gamma(\frac{d+p}{2})}{\Gamma(\frac{d+p+1}{2})} \frac{1}{(k-p)!p!}. \tag{24}$$

These results are novel. They permit the **first analytic characterisation of the difference in kernel estimator MSE between IIDRFs, ORFs and SimRFs**. We make one further observation.

Corollary 4.4 (ORFs always outperform IIDRFs)**.** *In the PRF setting, the orthogonality gap (difference in kernel estimator MSE between IIDRFs and ORFs) is given by*

$$\Delta\mathrm{MSE}(\widehat{K}(\mathbf{x}, \mathbf{y})) = e^{-2x^2-2y^2} \frac{m-1}{m} \cdot \left(e^{v^2} - \frac{\Gamma(d/2)}{\Gamma(d)} \sum_{k=0}^{\infty} \frac{v^{2k}}{k!} \frac{\Gamma(k+d)}{\Gamma(k+d/2)} \right) \tag{25}$$

where $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, $v = \|\boldsymbol{x}+\boldsymbol{y}\|_2$ and $m \leq d$ is the number of random vectors. This is positive everywhere.

The sign of this orthogonality gap was first reported in (Choromanski et al., 2020) but without an accompanying closed form.

Plotting each of derived probability distributions $p(w_{ij})$ (Eq. 18 and Eq. 20, taking $\cos\theta = 0$ and $\cos\theta = -\frac{1}{d-1}$) and noting from Eq. 10 that the RF-conformity depends on the expectation value of the monotonically increasing function $f(w_{ij}, v) = \Gamma(\frac{d}{2}) \sum_{k=0}^{\infty} \frac{v^{2k} w_{ij}^{2k}}{2^{2k} k! \Gamma(k+\frac{d}{2})}$, the intuitive reason for the relative efficacy of SimRFs, ORFs and IIDRFs becomes clear: conformity is penalised by tails at large w_{ij}, which we suppress with geometrical coupling (Fig. 5).

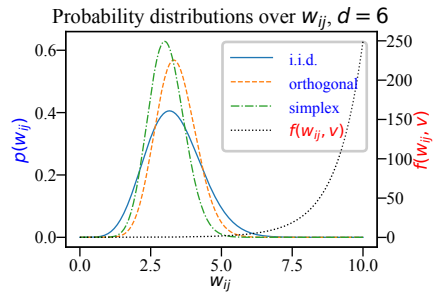

Figure 5. Probability distributions over the random variable $w_{ij} = \|\boldsymbol{w}_i + \boldsymbol{w}_j\|_2$ for IIDRFs, ORFs and SimRFs. The RF-conformity depends on the expectation of a monotonically increasing function $f(w_{ij})$. With PRFs, geometrical coupling decreases this by reducing the probability mass at large w_{ij}.

4.1. Extension to RFFs

We briefly note that, with minimal work, the preceding results for PRFs can be modified to consider RFFs. For example, the following is true.

Theorem 4.5 (RFF orthogonality gap)**.** *In the RFF setting, the orthogonality gap (difference in kernel estimator MSE between IIDRFs and ORFs) is given by*

$$\Delta\mathrm{MSE}(\widehat{K}(\boldsymbol{x}, \boldsymbol{y})) = \frac{m-1}{m} \left(e^{-z^2} - \frac{\Gamma(d/2)}{\Gamma(d)} \sum_{k=0}^{\infty} \frac{(-z^2)^k}{2^k k!} \frac{\Gamma(k+d)}{\Gamma(k+d/2)} \right) \tag{26}$$

where $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, $z = \|\boldsymbol{x}-\boldsymbol{y}\|_2$ and $m \leq d$ is the number of random vectors.

To the best of our knowledge, this result is also novel. The expression does not admit the same simple analysis as the PRF form (25) because successive terms in the sum oscillate in sign, but a cursory numerical analysis reveals that the MSE of ORFs is smaller than IIDRFs up to some threshold $z_{\mathrm{crit}}(d)$, the value of which diverges as $d \to \infty$. Taylor expanding our exact result in $\frac{1}{d}$ reproduces the following.

Corollary 4.6 (RFF asymptotic MSE ratio, Yu et al. (2016))**.** *The ratio of ORF to IIDRF kernel estimator MSE is given by*

$$\frac{MSE(\widehat{K}_{ORF})}{MSE(\widehat{K}_{IIDRF})} = 1 - (m-1) \left(\frac{e^{-z^2} z^4}{d(1-e^{-z^2})^2} + \mathcal{O}\left(\frac{1}{d^2}\right) \right), \tag{27}$$

where $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, $z = \|\boldsymbol{x}-\boldsymbol{y}\|_2$ and $m \leq d$ is the number of random features.

The negative subleading term shows that the RFF orthogonality gap is positive everywhere when $d \to \infty$.

4.2. Implementation, Complexity and Fast SimRFs

The replacement of ORFs with SimRFs is straightforward: instead of calculating random projections $\mathbf{W}\boldsymbol{x}$ using the

orthogonal block $\mathbf{W}_{\text{ort}} = \mathbf{DR}$, we use the *simplex block* $\mathbf{W}_{\text{simp}} = \mathbf{DSR}$, with the matrices $\mathbf{D}, \mathbf{S}, \mathbf{R} \in \mathbb{R}^{d \times d}$ and the object $\boldsymbol{x} \in \mathbb{R}^d$ defined at the beginning of Sec. 3. By choosing the order of computation $\mathbf{D}(\mathbf{S}(\mathbf{R}\boldsymbol{x}))$, we can avoid the $\mathcal{O}(d^3)$ time complexity of computing matrix-matrix products. Both $\mathbf{D}$ and $\mathbf{S}$ support matrix-vector multiplication of time complexity $\mathcal{O}(d)$ (see Appendix B.2.1). Generically, the time complexity to sample the random orthogonal matrix $\mathbf{R}$ is $\mathcal{O}(d^3)$ and the matrix-vector multiplication $\mathbf{R}\boldsymbol{x}$ is $\mathcal{O}(d^2)$. However, following exactly the same tricks as with ORFs, it is possible to replace $\mathbf{R}$ with a proxy $\widetilde{\mathbf{R}}$ which is *approximately* sampled from the orthogonal group according to Haar measure and which supports fast matrix-vector multiplication: for example, $\mathbf{HD}$-product matrices (Choromanski et al., 2017) or products of Givens random rotations (Dao et al., 2019). Then the time-complexity will be limited by the computation $\widetilde{\mathbf{R}}\boldsymbol{x}$ which is subquadratic by construction (e.g. $\mathcal{O}(d \log d)$ for the examples above). We refer to this mechanism as *fast SimRFs*, and show its excellent experimental performance in Appendix B.2.2.

SimRFs+ are implemented by $\mathbf{W}_{\text{simp+}} = \mathbf{DS}'\mathbf{R}$, where $\mathbf{S}'$ is obtained from $\mathbf{S}$ according to the $\mathcal{O}(d^3)$ iterative optimisation scheme defined in Eq. 17. This will dominate the scaling of time-complexity if we apply fast SimRFs+.

Table 1. Time complexities of RF-mechanisms and their fast variants.

	Time-complexity		
	ORFs	SimRFs	SimRFs+
Regular	$\mathcal{O}(d^3)$	$\mathcal{O}(d^3)$	$\mathcal{O}(d^3)$
Fast	$\mathcal{O}(d \log d)$	$\mathcal{O}(d \log d)$	$\mathcal{O}(d^3)$

For all regular schemes, the space complexity to store $\mathbf{R}$ is $\mathcal{O}(d^2)$. For fast ORFs and fast SimRFs, the space complexity becomes $\mathcal{O}(d)$ because we no longer need to explicitly store $\widetilde{\mathbf{R}}$, just the d weights $\{w_i\}$ from χ_d. But the space complexity of fast SimRFs+ is still $\mathcal{O}(d^2)$ since all vectors must be stored during the optimisation step.

It is clear that **SimRFs are essentially equal in computational cost to ORFs**, and in Sec. 5 we will see that they often perform substantially better in downstream tasks. Meanwhile, SimRFs+ are mostly of academic interest.

5. Experiments

Here we report the outcomes of an extensive empirical evaluation of SimRFs for PRFs, demonstrating their superiority over IIDRFs and ORFs in a variety of settings. Technical details are reported in Appendix B. The section is organised as follows: (a) in Sec. 5.1 we plot the derived MSE expressions for IIDRFs, ORFs and SimRFs; (b) in Sec. 5.2 we verify that SimRFs permit higher-quality kernel matrix approximation by considering the Frobenius norm of the difference between the true and approximated Gram matrices; (c) in Sec. 5.3 we compare the performance of the different RF mechanisms on nonparametric classification tasks using kernel regression; (d) in Sec. 5.4 we compare the RF mechanisms for approximation of the attention module in vision Performer-Transformers.

5.1. Comparison of MSE Between RF Mechanisms

We begin by plotting the MSE of the PRF estimator $\widehat{K}$ with IIDRFs, ORFs and SimRFs, given by Eq. 10 with the RF-confirmities 19, 23 and 24. We note that the *ratio* of the MSE of any pair of RF mechanisms only depends in the data $\boldsymbol{x}, \boldsymbol{y}$ via $v = \|\boldsymbol{x} + \boldsymbol{y}\|_2$, so it is natural to plot $\text{MSE}_{\text{ORF}}/\text{MSE}_{\text{IIDRF}}$ and $\text{MSE}_{\text{SimRF}}/\text{MSE}_{\text{IIDRF}}$ as a function of v – see Fig. 6. We take $d = 64$ which is standard in Transformer applications.

SimRFs always outperform ORFs and IIDRFs, but the size of the improvement depends sensitively on the data. SimRFs are particularly effective compared to ORFs and IIDRFs when estimating kernel evaluations at small v. This can be understood from their respective Taylor expansions. For both IIDRFs and ORFs, the MSE goes as $\text{MSE}_{\text{IIDRF,ORF}} = v^2 + \mathcal{O}(v^4)$. Meanwhile, for SimRFs, $\text{MSE}_{\text{SimRF}} = v^2 \left(1 - \frac{\sqrt{\pi}\Gamma(d+1)\Gamma(\frac{d}{2}+\frac{1}{2})}{\Gamma(\frac{d}{2})\Gamma(\frac{d}{2}+1)^2 2^d}\right) + \mathcal{O}(v^4)$. For $d = 64$, the SimRF v^2 prefactor evaluates to 0.0078 which is manifestly substantially smaller than 1.

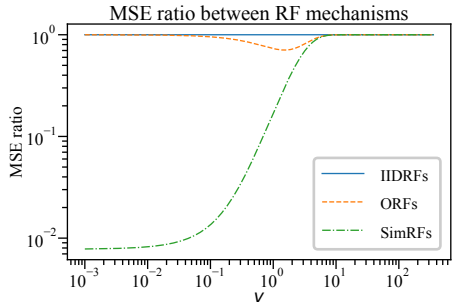

Figure 6. Analytic form the the MSE ratio of the PRF kernel estimator $\widehat{K}$ for different couplings, plotted as a function of $v = \|\boldsymbol{x} + \boldsymbol{y}\|_2$. Smaller values indicate lower MSE and are hence better. SimRFs always perform the best, followed by ORFs then IIDRFs. The size of the improvement depends on the data; it is bigger at smaller v.

5.2. Quality of Gram Matrix Approximation

Another straightforward task is to directly compare the quality approximation of the Gram matrix $\widehat{\mathbf{K}}$ with the different RF mechanisms. We can quantify this using the Frobenius norm between the exact and approximated matrices $\sum_{i=1}^{N}\sum_{j=1}^{N}(\mathbf{K}_{ij} - \widehat{\mathbf{K}}_{ij})^2$, where $\mathbf{K}_{ij} \stackrel{\text{def}}{=} K_{\text{gauss}}(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and $\widehat{\mathbf{K}}_{ij}$ is the corresponding low-rank decomposition. For demonstration purposes, we randomly generate $N = 64$

data points of dimensionality $d = 64$ according to the distribution $\boldsymbol{x}_i \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$. We take $\sigma = 0.1$. Fig. 7 shows the results; the quality of Gram matrix approximation improves with the number of features, and is better with SimRFs than ORFs and IIDRFs.

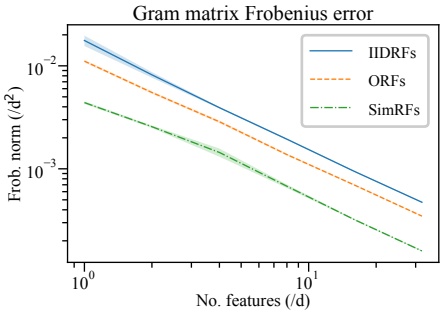

Figure 7. Frobenius norm between the true and approximated Gram matrices (lower is better) using different RF mechanisms and a different number of random features. More features give a better approximation, and SimRFs consistently outperform ORFs and IIDRFs. The data is of dimensionality $d = 64$ and we take $N = 64$ points, generated normally with $\sigma = 0.1$. The shading gives one standard deviation on estimates of the mean.

5.3. Nonparametric Classification Using Kernel Regression

Here we demonstrate how reduced kernel estimator MSE translates to better performance in downstream classification tasks. We use 8 different datasets retrieved from the UCI Machine Learning Repository (Dua & Graff, 2017a), each consisting of L training data $\{(\boldsymbol{x}, \boldsymbol{y})\}$ and test data $\{(\boldsymbol{x}', \boldsymbol{y}')\}$. The objects are d-dimensional vectors $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$ and their labels are one-hot encoded $\boldsymbol{y}, \boldsymbol{y}' \in \mathbb{R}^n$. We predict the label distribution of a test object using kernel regression with the Gaussian kernel, $\boldsymbol{y}'_{\text{pred}} = \sum_{i=1}^{L} K(\sigma\boldsymbol{x}', \sigma\boldsymbol{x}^{(i)})\boldsymbol{y}^{(i)} / \sum_{i=1}^{L} K(\sigma\boldsymbol{x}', \sigma\boldsymbol{x}^{(i)})$. We then predict a class by taking the greatest argument of $\boldsymbol{y}'_{\text{pred}}$. We measure accuracy by the proportion of correct label predictions across the test-set. The $\sigma > 0$ hyperparameter is tuned for good PRF performance on a validation dataset; see Appendix B.1 for detailed discussion. Fig. 8 presents the results, plotting classification accuracy against the number of random features used. The size of the benefit accrued from using SimRFs depends on the data (as we noted in Sec. 5.1) and in the limit of large m performance tends towards the exact kernel result. SimRFs consistently perform best.

5.3.1. SimRFs+ for Nonparametric Classification

Table 2 compares the classification accuracies achieved with SimRFs and SimRFs+ on the task detailed above, using $m = d$ random features. As suggested in Sec. 3 (see in particular Fig. 4), SimRFs are already close to optimal and

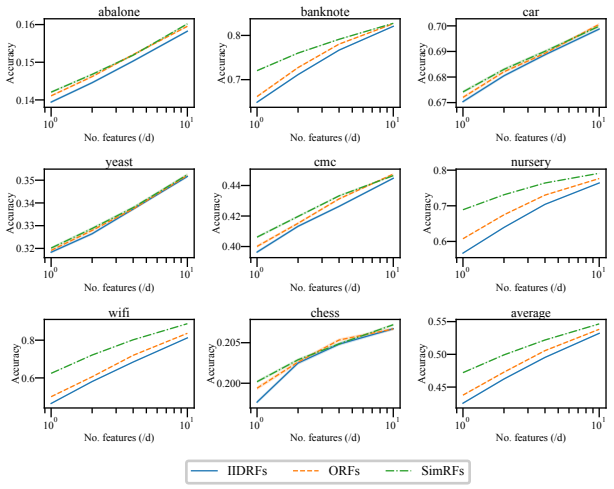

Figure 8. Nonparametric classification using kernel regression for a variety of datasets (Dua & Graff, 2017a; Nash et al., 1994; Dua & Graff, 2017b; Bohanec & Rajkovic, 1988; Horton & Nakai, 1996; Lim et al., 2000; Olave et al., 1989; Dua & Graff, 2017c), where the Gaussian kernel is approximated with different RFs. Plots show mean classification accuracy vs the number of random features used to approximate the kernel (/d, the dimensionality of the objects $\boldsymbol{x}$). Shading gives the standard deviation on the estimates of the mean. SimRFs consistently perform best.

any gain provided by using SimRFs+ is marginal. Moreover, improvements tend to occur where v is small so truncating the objective series expansion at $k = 1$ is reasonable.

Table 2. Classification accuracies from kernel regression with SimRFs and SimRFs+, using random features of length $m = d$. $\bar{v}$ records the mean (σ-scaled) value of v in each dataset. Note that both variants substantially outperform ORFs on every dataset.

Data set	$\bar{v}$	Classification accuracy SimRFs	SimRFs+
abalone	1.7	**0.1421±0.0002**	**0.1419±0.0002**
banknote	2.6	**0.7229±0.0012**	0.7132±0.0012
car	5.0	**0.6754±0.0004**	**0.6751±0.0004**
yeast	3.1	**0.3202±0.0004**	**0.3208±0.0004**
cmc	2.0	0.4047±0.0005	**0.4065±0.0005**
nursery	1.4	0.6874±0.0005	**0.6917±0.0004**
wifi	0.8	0.6314±0.0018	**0.6473±0.0018**
chess	2.3	**0.2000±0.0001**	**0.2000±0.0001**

5.4. SimRFs-Performers: Scalable Attention for Transformers

PRFs were first introduced in (Choromanski et al., 2020) in order to accurately approximate the softmax attention module of Transformers – an architecture coined the *Performer*. This technique for kernelising the attention mechanism, which identifies complex dependencies between the elements of an input sequence, permits linear (c.f. quadratic) space- and time-complexity without assuming restrictive priors such as sparsity and low-rankness. Performers offer

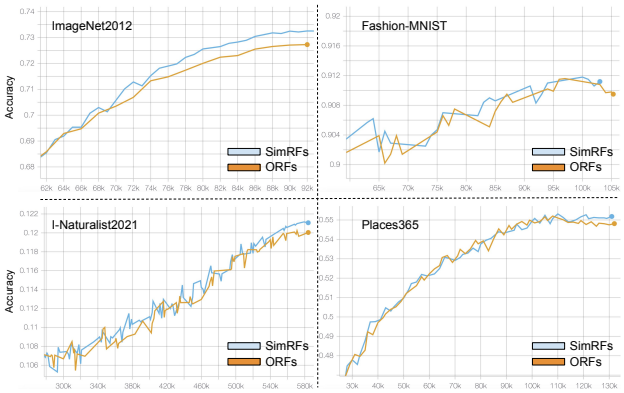

Figure 9. Accuracy comparison (higher is better) of the SimRFs-Performer and the regular ORFs-Performer. Tests are on four image classification tasks: (a) ImageNet2012, (b) Fashion-MNIST, (c) I-Naturalist2021, (d) Places365. x-axis is training epochs.

competitive results across a range of tasks (Tay et al., 2021), including vision modeling (Yuan et al., 2021; Horn et al., 2021) and speech (Liutkus et al., 2021).

Since Performers apply the ORF variant of PRFs, it is natural to expect that the SimRFs mechanism, which gives provably lower kernel estimator MSE, will be more effective. We refer to this architecture as the *SimRFs-Performer*, and show that it outperforms the regular ORFs-Performer.

We focus on the 'performised' versions of Vision Transformers (ViTs) (Dosovitskiy et al., 2021) and consider four datasets: (a) ImageNet2012 (Deng et al., 2009) (1K classes, 1.2M training images, 100K test set); (b) Fashion-MNIST (Xiao et al., 2017) (10 classes, 60K training images, 10K test set); (c) I_naturalist2021 (Horn et al., 2018) (10K classes, 2.7M training images, 500K test set) and (d) Places365 (Zhou et al., 2018) (365 classes, 1.8M training images, 328K test set). These are often used to benchmark ViTs.

In all four experiments, we use a ViT with 12 layers, 12 heads, mlp_dim equal to 3072, a dropout rate of 0.1 and no attention dropout. We use the adam optimiser with weight decay equal to 0.1 and batch size $\mathrm{bs} = 4096$, trained for 300 epochs on the TPU architecture. We apply 130 random vectors to approximate the softmax attention kernel with PRFs, testing both the ORF and SimRF coupling mechanisms.

The results, comparing ORFs and SimRFs for approximating attention, are presented in Fig. 9. The SimRFs-Performer often achieves gains over the regular ORFs-Performer – and is certainly never worse – for no observable extra cost. The exact difference depends on the data distribution (see. Sec. 5.1) and the importance of MSE reduction for that particular task; if some other factor is bottlenecking Performer accuracy, then improving the approximation of the attention matrix cannot provide gains. Nonetheless, for some of the tested datasets the difference is substantial: for instance, on $\mathrm{ImageNet2012}$, which is frequently used to benchmark new Transformer variants, the SimRFs-Performer saturates at an accuracy which is greater than the regular ORFs-Performer by **0.5%**. It is remarkable that such a large gain can be accrued with a single drop-in matrix multiplication at no observable computational cost, without any architectural or ViT-specific changes.

6. Conclusion

We have introduced *Simplex Random Features* (SimRFs), a new mechanism for unbiased approximation of the Gaussian and softmax kernels. By correlating the directions of random vectors in the ensemble, we access lower kernel estimator MSE than the previously predominant *Orthogonal Random Features* (ORFs): a fact we have verified both theoretically and empirically via extensive experiments. We have shown that the suppressed MSE of SimRFs compared to ORFs often permits better performance in downstream applications, including in nonparametric classification and scalable Transformer training. However, the size of the gain depends on the data distribution and whether the quality of kernel approximation is currently bottlenecking model performance. We have proved that SimRFs constitute the best weight-independent geometrically-coupled PRF mechanism, with further marginal improvements available in some regimes from a weight-dependent SimRFs+ variant. Finally, through our detailed quantitative analysis of the different RF mechanisms, we have derived novel closed-form results for ORFs, precisely formalising qualitative and asymptotic findings previously reported in the literature.

7. Relative Contributions and Acknowledgements

IR developed the SimRF and SimRF+ mechanisms, proved all theoretical results, and ran the pointwise kernel evaluation, Frobenius norm and nonparametric classification experiments. KC designed and ran the Performer experiments, and was crucially involved in all aspects of the work throughout. AW and VL provided helpful discussion and feedback on drafts.

IR acknowledges support from a Trinity College External Studentship. VL acknowledges support from the Cambridge Trust and DeepMind. AW acknowledges support from a Turing AI Fellowship under grant EP/V025279/1 and the Leverhulme Trust via CFI.

References

Ailon, N. and Chazelle, B. The fast johnson–lindenstrauss transform and approximate nearest neighbors. *SIAM J. Comput.*, 39(1):302–322, 2009. doi: 10.1137/060673096. URL https://doi.org/10.1137/060673096.

Ailon, N. and Liberty, E. An almost optimal unrestricted fast johnson-lindenstrauss transform. *ACM Trans. Algorithms*, 9(3):21:1–21:12, 2013. doi: 10.1145/2483699.2483701. URL https://doi.org/10.1145/2483699.2483701.

Bohanec, M. and Rajkovic, V. Knowledge acquisition and explanation for multi-attribute decision making. In *8th intl workshop on expert systems and their applications*, pp. 59–78. Avignon France, 1988. URL https://kt.ijs.si/MarkoBohanec/pub/Avignon88.pdf.

Bojarski, M., Choromanska, A., Choromanski, K., Fagan, F., Gouy-Pailler, C., Morvan, A., Sakr, N., Sarlos, T., and Atif, J. Structured adaptive and random spinners for fast machine learning computations. In *Artificial intelligence and statistics*, pp. 1020–1029. PMLR, 2017. URL http://proceedings.mlr.press/v54/bojarski17a/bojarski17a.pdf.

Choromanski, K., Rowland, M., Sarlós, T., Sindhwani, V., Turner, R. E., and Weller, A. The geometry of random features. In Storkey, A. J. and Pérez-Cruz, F. (eds.), *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, volume 84 of *Proceedings of Machine Learning Research*, pp. 1–9. PMLR, 2018. URL http://proceedings.mlr.press/v84/choromanski18a.html.

Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, L., et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020. URL https://openreview.net/pdf?id=Ua6zuk0WRH.

Choromanski, K. M., Rowland, M., and Weller, A. The unreasonable effectiveness of structured random orthogonal embeddings. *Advances in neural information processing systems*, 30, 2017. URL https://arxiv.org/abs/1703.00864.

Chowdhury, S. P., Solomou, A., Dubey, A., and Sachan, M. On learning the transformer kernel. *CoRR*, abs/2110.08323, 2021. URL https://arxiv.org/abs/2110.08323.

Dao, T., Gu, A., Eichhorn, M., Rudra, A., and Ré, C. Learning fast algorithms for linear transforms using butterfly factorizations. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1517–1527. PMLR, 2019. URL http://proceedings.mlr.press/v97/dao19a.html.

Dasgupta, A., Kumar, R., and Sarlós, T. A sparse johnson: Lindenstrauss transform. In Schulman, L. J. (ed.), *Proceedings of the 42nd ACM Symposium on Theory of Computing, STOC 2010, Cambridge, Massachusetts, USA, 5-8 June 2010*, pp. 341–350. ACM, 2010. doi: 10.1145/1806689.1806737. URL https://doi.org/10.1145/1806689.1806737.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pp. 248–255. IEEE Computer Society, 2009. doi: 10.1109/CVPR.2009.5206848. URL https://doi.org/10.1109/CVPR.2009.5206848.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Dua, D. and Graff, C. UCI machine learning repository, 2017a. URL http://archive.ics.uci.edu/ml.

Dua, D. and Graff, C. Banknote authentication dataset, UCI machine learning repository, 2017b. URL http://archive.ics.uci.edu/ml.

Dua, D. and Graff, C. Chess (king-rook vs. king) dataset, UCI machine learning repository, 2017c. URL http://archive.ics.uci.edu/ml.

Faris, W. Radial functions and the fourier transform, 2008. URL http://www.math.arizona.edu/~faris/methodsweb/hankel.pdf. Accessed: 20-10-2022.

Horn, G. V., Aodha, O. M., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., and Belongie, S. J. The inaturalist species classification and detection dataset. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 8769–8778.

Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00914. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Van_Horn_The_INaturalist_Species_CVPR_2018_paper.html.

Horn, M., Shridhar, K., Groenewald, E., and Baumann, P. F. M. Translational equivariance in kernelizable attention. *CoRR*, abs/2102.07680, 2021. URL https://arxiv.org/abs/2102.07680.

Horton, P. and Nakai, K. A probabilistic classification system for predicting the cellular localization sites of proteins. In *Ismb*, volume 4, pp. 109–115, 1996. URL https://pubmed.ncbi.nlm.nih.gov/8877510/.

Johnson, W. B. Extensions of lipschitz mappings into a hilbert space. *Contemp. Math.*, 26:189–206, 1984. URL https://doi.org/10.1090/conm/026/737400.

Kane, D. M. and Nelson, J. Sparser johnson-lindenstrauss transforms. *J. ACM*, 61(1):4:1–4:23, 2014. doi: 10.1145/2559902. URL https://doi.org/10.1145/2559902.

Kar, P. and Karnick, H. Random feature maps for dot product kernels. In Lawrence, N. D. and Girolami, M. A. (eds.), *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2012, La Palma, Canary Islands, Spain, April 21-23, 2012*, volume 22 of *JMLR Proceedings*, pp. 583–591. JMLR.org, 2012. URL http://proceedings.mlr.press/v22/kar12.html.

Kitaev, N., Kaiser, L., and Levskaya, A. Reformer: The efficient transformer. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=rkgNKkHtvB.

Kulesza, A. and Taskar, B. Determinantal point processes for machine learning. *Found. Trends Mach. Learn.*, 5 (2-3):123–286, 2012. doi: 10.1561/2200000044. URL https://doi.org/10.1561/2200000044.

Le, Q. V., Sarlós, T., and Smola, A. J. Fastfood - computing hilbert space expansions in loglinear time. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pp. 244–252. JMLR.org, 2013. URL http://proceedings.mlr.press/v28/le13.html.

Li, F., Ionescu, C., and Sminchisescu, C. Random fourier approximations for skewed multiplicative histogram kernels. In Goesele, M., Roth, S., Kuijper, A., Schiele, B., and Schindler, K. (eds.), *Pattern Recognition - 32nd DAGM Symposium, Darmstadt, Germany, September 22-24, 2010. Proceedings*, volume 6376 of *Lecture Notes in Computer Science*, pp. 262–271. Springer, 2010. doi: 10.1007/978-3-642-15986-2_27. URL https://doi.org/10.1007/978-3-642-15986-2_27.

Liberty, E., Ailon, N., and Singer, A. Dense fast random projections and lean walsh transforms. *Discret. Comput. Geom.*, 45(1):34–44, 2011. doi: 10.1007/s00454-010-9309-5. URL https://doi.org/10.1007/s00454-010-9309-5.

Likhosherstov, V., Choromanski, K. M., Davis, J. Q., Song, X., and Weller, A. Sub-linear memory: How to make performers slim. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 6707–6719, 2021. URL https://doi.org/10.48550/arXiv.2012.11346.

Likhosherstov, V., Choromanski, K., Dubey, A., Liu, F., Sarlos, T., and Weller, A. Chefs' random tables: Non-trigonometric random features. In *NeurIPS*, 2022. URL https://doi.org/10.48550/arXiv.2205.15317.

Lim, T.-S., Loh, W.-Y., and Shih, Y.-S. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine learning*, 40(3):203–228, 2000. URL https://doi.org/10.1023/A:1007608224229.

Lin, H., Chen, H., Choromanski, K. M., Zhang, T., and Laroche, C. Demystifying orthogonal monte carlo and beyond. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://doi.org/10.48550/arXiv.2005.13590.

Liu, F., Huang, X., Chen, Y., and Suykens, J. A. K. Random features for kernel approximation: A survey on algorithms, theory, and beyond. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10):7128–7148, 2022. doi: 10.1109/TPAMI.2021.3097011. URL https://doi.org/10.1109/TPAMI.2021.3097011.

Liutkus, A., Cífka, O., Wu, S., Simsekli, U., Yang, Y., and Richard, G. Relative positional encoding for transformers with linear complexity. In Meila, M. and Zhang,

T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7067–7079. PMLR, 2021. URL http://proceedings.mlr.press/v139/liutkus21a.html.

Luo, S., Li, S., Cai, T., He, D., Peng, D., Zheng, S., Ke, G., Wang, L., and Liu, T. Stable, fast and accurate: Kernelized attention with relative positional encoding. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 22795–22807, 2021. URL https://doi.org/10.48550/arXiv.2106.12566.

Nash, W. J., Sellers, T. L., Talbot, S. R., Cawthorn, A. J., and Ford, W. B. The population biology of abalone (haliotis species) in tasmania. i. blacklip abalone (h. rubra) from the north coast and islands of bass strait. *Sea Fisheries Division, Technical Report*, 48:p411, 1994. URL https://www.researchgate.net/publication/287546509_7he_Population_Biology_of_Abalone_Haliotis_species_in_Tasmania_I_Blacklip_Abalone_H_rubra_from_the_North_Coast_and_Islands_of_Bass_Strait.

Olave, M., Rajkovic, V., and Bohanec, M. An application for admission in public school systems. *Expert Systems in Public Administration*, 1:145–160, 1989. URL https://www.academia.edu/16670755/An_application_for_admission_in_public_school_systems.

Pennington, J., Yu, F. X., and Kumar, S. Spherical random features for polynomial kernels. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 1846–1854, 2015. URL https://proceedings.neurips.cc/paper/2015/file/f7f580e11d00a75814d2ded41fe8e8fe-Paper.pdf.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007. URL https://people.eecs.berkeley.edu/~brecht/papers/07.rah.rec.nips.pdf.

Rahimi, A. and Recht, B. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L. (eds.), *Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, 2008*, pp. 1313–1320. Curran Associates, Inc., 2008. URL https://people.eecs.berkeley.edu/~brecht/papers/08.rah.rec.nips.pdf.

Roy, A., Saffar, M., Vaswani, A., and Grangier, D. Efficient content-based sparse attention with routing transformers. *Trans. Assoc. Comput. Linguistics*, 9:53–68, 2021. doi: 10.1162/tacl_a_00353. URL https://doi.org/10.1162/tacl_a_00353.

Schlag, I., Irie, K., and Schmidhuber, J. Linear transformers are secretly fast weight programmers. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9355–9366. PMLR, 2021. URL http://proceedings.mlr.press/v139/schlag21a.html.

Tay, Y., Dehghani, M., Abnar, S., Shen, Y., Bahri, D., Pham, P., Rao, J., Yang, L., Ruder, S., and Metzler, D. Long range arena : A benchmark for efficient transformers. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=qVyeW-grC2k.

Trokicic, A. and Todorovic, B. Randomized nyström features for fast regression: An error analysis. In Ciric, M., Droste, M., and Pin, J. (eds.), *Algebraic Informatics - 8th International Conference, CAI 2019, Niš, Serbia, June 30 - July 4, 2019, Proceedings*, volume 11545 of *Lecture Notes in Computer Science*, pp. 249–257. Springer, 2019. doi: 10.1007/978-3-030-21363-3_21. URL https://doi.org/10.1007/978-3-030-21363-3_21.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017. URL http://arxiv.org/abs/1708.07747.

Xiao, X., Zhang, T., Choromanski, K., Lee, T. E., Francis, A. G., Varley, J., Tu, S., Singh, S., Xu, P., Xia, F., Persson, S. M., Kalashnikov, D., Takayama, L., Frostig, R., Tan, J., Parada, C., and Sindhwani, V. Learning model predictive controllers with real-time attention for real-world navigation. *CoRL 2022*, abs/2209.10780, 2022. doi: 10.48550/arXiv.2209.10780. URL https://doi.org/10.48550/arXiv.2209.10780.

Yang, J., Sindhwani, V., Avron, H., and Mahoney, M. W. Quasi-monte carlo feature maps for shift-invariant kernels. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, volume 32 of *JMLR Workshop and Conference Proceedings*, pp. 485–493. JMLR.org, 2014. URL http://proceedings.mlr.press/v32/yangb14.html.

Yu, F. X., Bhaskara, A., Kumar, S., Gong, Y., and Chang, S. On binary embedding using circulant matrices. *J. Mach. Learn. Res.*, 18:150:1–150:30, 2017. URL http://jmlr.org/papers/v18/15-619.html.

Yu, F. X. X., Suresh, A. T., Choromanski, K. M., Holtmann-Rice, D. N., and Kumar, S. Orthogonal random features. *Advances in neural information processing systems*, 29, 2016. URL https://doi.org/10.48550/arXiv.1610.09072.

Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z., Tay, F. E. H., Feng, J., and Yan, S. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 538–547. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00060. URL https://doi.org/10.1109/ICCV48922.2021.00060.

Zhou, B., Lapedriza, À., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6):1452–1464, 2018. doi: 10.1109/TPAMI.2017.2723009. URL https://doi.org/10.1109/TPAMI.2017.2723009.

A. Supplementary Proofs and Discussion

In this appendix we provide further discussion and proofs of results stated in the main text.

A.1. Proof of Theorem 3.2 (MSE Depends on RF-conformity)

We begin by deriving a form for the kernel estimator MSE in the PRF setting, showing how it depends upon the so-called *RF-conformity* defined in Eq. 10.

From the definitions in Eq. 6 and Eq. 7, it follows that

$$\widehat{K} = \phi(\boldsymbol{x})^\top \phi(\boldsymbol{y}) = \frac{e^{-x^2-y^2}}{m} \sum_{i=1}^{m} e^{\boldsymbol{w}_i^\top(\boldsymbol{x}+\boldsymbol{y})} = \frac{e^{-x^2-y^2}}{m} \sum_{i=1}^{m} b_i \tag{28}$$

where m is the number of random features and $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$. We introduced b_i, where

$$b_i \stackrel{\text{def}}{=} e^{\boldsymbol{w}_i^\top \boldsymbol{v}}, \tag{29}$$

with $\boldsymbol{v} = \boldsymbol{x} + \boldsymbol{y} \in \mathbb{R}^d$. Here, $i = 1, ..., m$ enumerates the random features. It is straightforward to show that this is an unbiased estimator of the Gaussian kernel $K(\boldsymbol{x}, \boldsymbol{y}) = \exp(-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|_2^2}{2})$ when $\boldsymbol{w}_i$ are sampled from $\mathcal{N}(0, \mathbf{I}_d)$; in particular, we find that $\mathbb{E}(b_i) = e^{\frac{v^2}{2}}$. After some algebra, we can also show that

$$\text{MSE}(\widehat{K}) = \frac{e^{-2x^2-2y^2}}{m} \left((e^{2v^2} - e^{v^2}) + (m-1)(\frac{1}{m(m-1)} \sum_i \sum_{i \neq j} \mathbb{E}[b_i b_j] - e^{v^2}) \right). \tag{30}$$

Now consider the correlation term $\frac{1}{m(m-1)} \sum_i \sum_{i \neq j} \mathbb{E}[b_i b_j] = \frac{1}{m(m-1)} \sum_i \sum_{i \neq j} \mathbb{E}[e^{(\boldsymbol{w}_i + \boldsymbol{w}_j)^\top \boldsymbol{v}}]$ more carefully. Evidently, we care about the probability distribution over the random variable $\boldsymbol{w}_i + \boldsymbol{w}_j$, denoted compactly by $\boldsymbol{w}_{ij}$. For all couplings we consider, the random vectors $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$ are marginally isotropic (a necessary condition to be marginally Gaussian) and their resultant $\boldsymbol{w}_{ij}$ will also be marginally isotropic. This permits us to rewrite the expectation value using the *Hankel transform* (Faris, 2008):

$$\begin{aligned} \mathbb{E}[e^{\boldsymbol{w}_{ij}^\top \boldsymbol{v}}] &= \int_{\mathbb{R}_d} \mathrm{d}^d \boldsymbol{w}_{ij} p(\boldsymbol{w}_{ij}) e^{\boldsymbol{w}_{ij}^\top \boldsymbol{v}} \\ &= \Gamma(d/2) 2^{\frac{d}{2}-1} \int_0^\infty \mathrm{d}w_{ij} p(w_{ij}) (i w_{ij} v)^{1-\frac{d}{2}} J_{\frac{d}{2}-1}(i w_{ij} v) \end{aligned} \tag{31}$$

where $J_{\frac{d}{2}-1}$ is a Bessel function of the first kind[5]. Importantly, we are integrating over a single variable: the norm of the resultant random vector, $w_{ij} = \|\boldsymbol{w}_i + \boldsymbol{w}_j\|_2$. The probability distribution $p(w_{ij})$ will depend on whether the random vectors are i.i.d. or exhibit geometrical coupling, even though the marginal distributions are identical (Gaussian) in every case.

Recalling the Taylor expansion

$$J_\alpha(z) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k! \Gamma(k+\alpha+1)} \left(\frac{z}{2}\right)^{2k+\alpha}, \tag{32}$$

we can rewrite the correlation term as

$$\mathbb{E}[b_i b_j] = \Gamma\left(\frac{d}{2}\right) \mathbb{E}_{w_{ij}} \left(\sum_{k=0}^{\infty} \frac{v^{2k} w_{ij}^{2k}}{2^{2k} k! \Gamma(k + \frac{d}{2})} \right). \tag{33}$$

Inserting this into Eq. 74, this immediately yields the important result:

$$\text{MSE}(\widehat{K}) = \frac{e^{-2x^2-2y^2}}{m} \left((e^{2v^2} - e^{v^2}) + (m-1)(\rho(\boldsymbol{x}, \boldsymbol{y}) - e^{v^2}) \right), \tag{34}$$

[5] In fact, given the purely imaginary argument, the function $I_\alpha(x) = i^{-\alpha} J_\alpha(ix)$ is referred to as the *modified* Bessel function.

where we defined the *RF-conformity*

$$\rho(\boldsymbol{x},\boldsymbol{y}) \stackrel{\text{def}}{=} \frac{\Gamma(\frac{d}{2})}{m(m-1)} \sum_i \sum_{j \neq i} \mathbb{E}_{w_{ij}} \left(\sum_{k=0}^{\infty} \frac{v^{2k} w_{ij}^{2k}}{2^{2k} k! \Gamma(k+\frac{d}{2})} \right), \tag{35}$$

as in Eq. 10 of the main text. Summations run from $i = 1$ to m, the number of random features. The MSE is manifestly an increasing function of $\rho(\boldsymbol{x},\boldsymbol{y})$, which itself depends sensitively on the any correlations induced between random vectors via $p(w_{ij})$. It is clear that any coupling mechanisms that reduce values of $w_{ij} = \|\boldsymbol{w}_i + \boldsymbol{w}_j\|_2$, e.g. by conditioning that random vectors point away from one another, will suppress $\rho(\boldsymbol{x},\boldsymbol{y})$. The RF-conformity will form a core consideration in the discussion that follows.

A.2. Proof of Theorem 3.4 (SimRFs Optimal for Weight-Independent Geometrical Coupling)

Here, we prove the central result that, supposing the weights $w_i = \|\boldsymbol{w_i}\|_2, i = 1, ..., d$ are i.i.d. (in our case from χ_d), SimRFs constitute the best possible weight-independent geometrical coupling scheme. Recall again that by 'weight-independent' we mean that vector directions $\{\hat{\boldsymbol{w}}_i\}$ are independent of norms $\{w_i\}$, though directions can still be correlated among themselves. Our choice of geometrical coupling will not depend on each particular draw of norms; we just use the fact that all w_i are identically distributed.

We begin by proving the following simpler auxiliary lemma.

Lemma A.1 (SimRFs optimal for equal norms)**.** *Suppose that, instead of being sampled from a χ_d distribution, we condition that $\boldsymbol{w}_i \in \mathbb{R}^d$ for $i = 1, ..., d$ all have equal lengths w. Then $\rho(\boldsymbol{x},\boldsymbol{y})$ is minimised when the ensemble exhibits simplex geometrical coupling.*

Proof: given the set of vector norms $w_i = w$ with $i = 1, ..., d$, we would like to know how to choose the angles θ_{ij} subtended between each pair $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$ to minimise the RF-conformity $\rho(\boldsymbol{x},\boldsymbol{y})$. It is immediately obvious that we should choose θ_{ij} deterministically rather than probabilistically, because assigning probability mass to suboptimal configurations will always increase the expectation value (that is, $p(w_{ij}|w_i, w_j = w) = \delta(w_{ij} - \sqrt{2w^2(1+\cos\theta_{ij})})$, with δ the delta function). So the task is to choose $\{\theta_{ij}\}$ to minimise

$$\rho(\boldsymbol{x},\boldsymbol{y}) = \frac{\Gamma(\frac{d}{2})}{m(m-1)} \sum_i \sum_{j \neq i} \sum_{k=0}^{\infty} \frac{v^{2k} w^{2k}}{2^{2k} k! \Gamma(k+\frac{d}{2})} \|\widehat{\boldsymbol{w}}_i + \widehat{\boldsymbol{w}}_j\|_2^{2k} = \sum_i \sum_{j \neq i} f(\|\widehat{\boldsymbol{w}}_i + \widehat{\boldsymbol{w}}_j\|_2^2) \tag{36}$$

where we defined the increasing convex function

$$f(\|\widehat{\boldsymbol{w}}_i + \widehat{\boldsymbol{w}}_j\|_2^2) \stackrel{\text{def}}{=} \frac{\Gamma(\frac{d}{2})}{m(m-1)} \sum_{k=0}^{\infty} \frac{v^{2k} w^{2k}}{2^{2k} k! \Gamma(k+\frac{d}{2})} \|\widehat{\boldsymbol{w}}_i + \widehat{\boldsymbol{w}}_j\|_2^{2k}. \tag{37}$$

It follows from Jensen's inequality that

$$\sum_i \sum_{j \neq i} f(\|\widehat{\boldsymbol{w}}_i + \widehat{\boldsymbol{w}}_j\|_2^2) \geq m(m-1) f\left(\frac{\sum_i \sum_{j \neq i} \|\widehat{\boldsymbol{w}}_i + \widehat{\boldsymbol{w}}_j\|_2^2}{m(m-1)} \right) \tag{38}$$

with equality when $\|\widehat{\boldsymbol{w}}_i + \widehat{\boldsymbol{w}}_j\|_2$ is identical for every i, j, i.e. all random vectors subtend equal angles. Since f is increasing,

$$\begin{aligned} \sum_i \sum_{j \neq i} \|\widehat{\boldsymbol{w}}_i + \widehat{\boldsymbol{w}}_j\|_2^2 &= \sum_i \sum_{j \neq i} 2 + 2\widehat{\boldsymbol{w}}_i^\top \widehat{\boldsymbol{w}}_j \\ &= 2m(m-1) + 2\sum_i \widehat{\boldsymbol{w}}_i^\top \sum_{j \neq i} \widehat{\boldsymbol{w}}_j \\ &= 2m(m-2) + 2(\sum_i \widehat{\boldsymbol{w}}_i)^\top (\sum_j \widehat{\boldsymbol{w}}_j) \\ &= 2m(m-2) + 2\|\sum_i \widehat{\boldsymbol{w}}_i\|_2^2 \geq 2m(m-2) \end{aligned} \tag{39}$$

with equality achieved when $\sum_i \widehat{\boldsymbol{w}}_i = 0$. Therefore,

$$\rho(\boldsymbol{x}, \boldsymbol{y}) = \sum_i \sum_{j \neq i} f(\|\widehat{\boldsymbol{w}}_i + \widehat{\boldsymbol{w}}_j\|_2^2) \geq m(m-1) f\left(\frac{2(m-2)}{m-1}\right). \tag{40}$$

This shows that the conformity is minimised when we have that i) all vectors $\widehat{\boldsymbol{w}}_i$ subtend equal angles, and ii) $\sum_i \widehat{\boldsymbol{w}}_i = 0$. This is nothing other than the geometry of a $d-1$-dimensional simplex embedded in d-dimensional space, as described by the basis vectors defined in Eq. 9. □

Armed with the result of Lemma A.1, we now consider the more general setting where $\{w_i\}$ are i.i.d. random variables but draws are not generically identical.

We begin with the observation that, if random variables $w_1, ..., w_d$ are i.i.d., the joint distribution $p(w_1, w_2, ..., w_d) = p(w_1) \cdot p(w_2) \cdot ... \cdot p(w_d)$ is invariant under permutation of w_i. This is because the joint distribution factorises into d identical functions, though more general joint distributions with this property exist. Intuitively, for every given draw of weights $\{w_1, ..., w_d\}$, there are $d! - 1$ other draws of equal probability given by the permutations $\{w_{P_1}, ..., w_{P_d}\}$ where $P \in S_d$, the symmetric group on d letters. Therefore, the RF conformity can be expressed as

$$\begin{aligned}
\rho(\boldsymbol{x}, \boldsymbol{y}) &= \frac{\Gamma(\frac{d}{2})}{m(m-1)} \int \mathrm{d}w_1 \mathrm{d}w_2 ... \mathrm{d}w_d p(w_1, ..., w_d) \sum_i \sum_{j \neq i} \sum_{k=0}^{\infty} \frac{v^{2k} w_{ij}^{2k}}{2^{2k} k! \Gamma(k + \frac{d}{2})} \\
&= \frac{\Gamma(\frac{d}{2})}{m(m-1)} \int \mathrm{d}w_1 \mathrm{d}w_2 ... \mathrm{d}w_d \frac{p(w_1, ..., w_d)}{d!} \\
&\cdot \sum_{P \in S_d} \sum_i \sum_{j \neq i} \sum_{k=0}^{\infty} \frac{v^{2k}}{2^{2k} k! \Gamma(k + \frac{d}{2})} \left(w_{P_i}^2 \hat{\boldsymbol{w}}_i^\top \hat{\boldsymbol{w}}_i + w_{P_j}^2 \hat{\boldsymbol{w}}_j^\top \hat{\boldsymbol{w}}_j + 2 w_{P_i} w_{P_j} \hat{\boldsymbol{w}}_i^\top \hat{\boldsymbol{w}}_j \right)^k
\end{aligned} \tag{41}$$

where we wrote $p(w_1, ..., w_d) = \frac{1}{d!} \sum_{P \in S_n} p(w_{P_1}, w_{P_2}, ..., w_{P_d})$ then relabelled the integration variables. Here, we have permuted the random vector norms w_i but *not* the directions $\hat{\boldsymbol{w}}_i$. We would like to obtain the geometry $\{\hat{\boldsymbol{w}}_i\}$ that minimises $\rho(\boldsymbol{x}, \boldsymbol{y})$, subject to the condition that the normalisations of the unit vectors $\hat{\boldsymbol{w}}_i^\top \hat{\boldsymbol{w}}_i = 1$ are fixed. Since the integrand is nonnegative everywhere, we minimise the sum in the final line of Eq. 41, namely

$$\begin{aligned}
&\sum_{P \in S_d} \sum_i \sum_{j \neq i} f\left(w_{P_i}^2 + w_{P_j}^2 + 2 w_{P_i} w_{P_j} \hat{\boldsymbol{w}}_i^\top \hat{\boldsymbol{w}}_j \right) \\
&= \sum_{P \in S_d} \sum_i \sum_{j \neq i} \sum_{k=0}^{\infty} \frac{v^{2k}}{2^{2k} k! \Gamma(k + \frac{d}{2})} \left(w_{P_i}^2 + w_{P_j}^2 + 2 w_{P_i} w_{P_j} \hat{\boldsymbol{w}}_i^\top \hat{\boldsymbol{w}}_j \right)^k
\end{aligned} \tag{42}$$

where f is once again convex and positive definite. Relabelling summation variables then using Jensen's inequality, we can write this as

$$\sum_{P \in S_d} \sum_i \sum_{j \neq i} f\left(w_i^2 + w_j^2 + 2 w_i w_j \hat{\boldsymbol{w}}_{P_i}^\top \hat{\boldsymbol{w}}_{P_j} \right) \geq d! \sum_i \sum_{j \neq i} f\left(\frac{\sum_{P \in S_d} w_i^2 + w_j^2 + 2 w_i w_j \hat{\boldsymbol{w}}_{P_i}^\top \hat{\boldsymbol{w}}_{P_j}}{d!} \right) \tag{43}$$

with equality when $\hat{\boldsymbol{w}}_{P_i}^\top \hat{\boldsymbol{w}}_{P_j}$ is identical for every permutation – that is, when all the random vectors subtend identical angles. With this in mind, we write the Lagrangian as

$$\mathcal{L} = \sum_{P \in S_d} \sum_i \sum_{j \neq i} \sum_{k=0}^{\infty} \frac{v^{2k}}{2^{2k} k! \Gamma(k + \frac{d}{2})} \left(w_{P_i}^2 \hat{\boldsymbol{w}}_i^\top \hat{\boldsymbol{w}}_i + w_{P_j}^2 \hat{\boldsymbol{w}}_j^\top \hat{\boldsymbol{w}}_j + 2 w_{P_i} w_{P_j} \hat{\boldsymbol{w}}_i^\top \hat{\boldsymbol{w}}_j \right)^k - \sum_i \lambda_i (\hat{\boldsymbol{w}}_i^\top \hat{\boldsymbol{w}}_i - 1). \tag{44}$$

Differentiating wrt $\hat{\boldsymbol{w}}_i$,

$$\sum_{P \in S_d} \sum_{j \neq i} \sum_{k=0}^{\infty} \frac{v^{2k} k w_{P_i P_j}^{2k-2}}{2^{2k} k! \Gamma(k + \frac{d}{2})} \left(w_{P_i}^2 \hat{\boldsymbol{w}}_i + w_{P_i} w_{P_j} \hat{\boldsymbol{w}}_j \right) - \lambda_i \boldsymbol{w}_i = 0 \qquad i = 1, ..., d. \tag{45}$$

where we used that $\hat{\boldsymbol{w}}_{P_i}^\top \hat{\boldsymbol{w}}_{P_j} = \hat{\boldsymbol{w}}_i^\top \hat{\boldsymbol{w}}_j$ to take

$$w_{P_i}^2 \hat{\boldsymbol{w}}_i^\top \hat{\boldsymbol{w}}_i + w_{P_j}^2 \hat{\boldsymbol{w}}_j^\top \hat{\boldsymbol{w}}_j + 2 w_{P_i} w_{P_j} \hat{\boldsymbol{w}}_i^\top \hat{\boldsymbol{w}}_j = w_{P_i P_j} = w_{P_i}^2 \hat{\boldsymbol{w}}_{P_i}^\top \hat{\boldsymbol{w}}_{P_i} + w_{P_j}^2 \hat{\boldsymbol{w}}_{P_j}^\top \hat{\boldsymbol{w}}_{P_j} + 2 w_{P_i} w_{P_j} \hat{\boldsymbol{w}}_{P_i}^\top \hat{\boldsymbol{w}}_{P_j}. \tag{46}$$

Eq. 45 implies that

$$\hat{\boldsymbol{w}}_i \propto -\sum_{j\neq i}\sum_{k=0}^{\infty}\frac{v^{2k}k}{2^{2k}k!\Gamma(k+\frac{d}{2})}\left(\sum_{P\in S_d} w_{P_iP_j}^{2k-2}w_{P_i}w_{P_j}\right)\hat{\boldsymbol{w}}_j \qquad i=1,...,d \tag{47}$$

with the proportionality constant fixed by the normalisation of $\hat{\boldsymbol{w}}_i$. Crucially, since we are summing over all permutations S_d of the d labels, the term in parentheses $\left(\sum_{P\in S_d} w_{P_iP_j}^{2k-2}w_{P_i}w_{P_j}\right)$ is identical for every i, j. This immediately implies that

$$\hat{\boldsymbol{w}}_i \propto -\sum_{j\neq i}\hat{\boldsymbol{w}}_j \qquad i=1,...,d. \tag{48}$$

Subject to the further constraint that all $\hat{\boldsymbol{w}}_i$ subtend equal angles, this is uniquely given by the simplex geometry described by the basis vectors in Eq. 9. That is, supposing the vector norms w_i are i.i.d. and that the geometrical coupling is weight-independent, SimRFs give the lowest possible MSE in the PRF setting. □

An intuitive explanation of this result is as follows. In Lemma A.1, we observed that SimRFs are optimal if all vector norms w_i are equal. Supposing norms are not equal but are identically distributed, any geometrical coupling scheme that is better for some particular draw of norms $\{w_i\}$ will be worse for some of the (equally probable) label permutations $\{w_{P_i}\}, P \in S_d$. The effect of summing over all the permutations is the same as collapsing all the distributions over w_i to a single, identical value.

A.3. Derivation of Eq. 16 (SimRFs+ Geometry Minimises the Truncated RF-Conformity Objective)

Here, we show that the SimRFs+ geometrical coupling mechanism (Eq. 16) minimises the truncated approximation to the RF-conformity $\tilde{\rho}(\boldsymbol{x}, \boldsymbol{y})$ (Eq. 15). Writing a Lagrangian using the truncated sum and differentiating, it is straightforward to find that

$$\sum_{j\neq i}\left(\frac{v^2}{4\Gamma(1+\frac{d}{2})}(\boldsymbol{w}_i+\boldsymbol{w}_j)\right)-\lambda_i\boldsymbol{w}_i=0 \qquad i=1,...,d \tag{49}$$

with the Lagrange multipliers λ_i fixed by the (known) normalisations of $\boldsymbol{w}_i$. Should such a geometry exist, this will be solved by

$$\boldsymbol{w}_i \propto -\sum_{j\neq i}\boldsymbol{w}_j \qquad i=1,...,d. \tag{50}$$

Note that, on account of the truncation of the objective, we do *not* need to make any assumptions about the vector norms or angles subtended being equal to reach this conclusion. It is straightforward to convince oneself that such a geometry always exists for any set of norms: if one norm w_i exceeds the sum of all the others, Eq. 50 is trivially satisfied by arranging the vector of maximum norm to be antialigned with all the rest; if this is not the case, it is always possible to arrange the vectors such that they sum to 0, i.e. form a closed loop. Then $\boldsymbol{w}_i = -\sum_{j\neq i}\boldsymbol{w}_j$, which satisfies Eq. 50. We conclude that the SimRFs+ geometry

$$\boldsymbol{w}_i = -\frac{\sum_{j\neq i}\boldsymbol{w}_j}{\|\sum_{j\neq i}\boldsymbol{w}_j\|_2}w_i \qquad i=1,...,d \tag{51}$$

minimises $\tilde{\rho}(\boldsymbol{x}, \boldsymbol{y})$.

We briefly note that Eq. 51 does not actually define one unique geometrical coupling, but empirically the iterative update scheme in Eq. 17 always finds a good solution when initialised in the simplex geometry.

A.4. Proof of Lemma 4.1 (IIDRF Conformity)

In this appendix, we derive the probability distribution $p(w_{ij})$ over $w_{ij} = \|\boldsymbol{w}_i + \boldsymbol{w}_j\|_2$ in the case that all $\boldsymbol{w}_i$ follow independent Gaussian distributions $\mathcal{N}(0, \mathbf{I}_d)$, and use it to evaluate the corresponding IIDRF conformity $\rho(\boldsymbol{x}, \boldsymbol{y})$.

In the i.i.d. case, each component of the vector $\boldsymbol{w}_i + \boldsymbol{w}_j$ is the sum of two standard normal distributions, $\mathcal{N}(0, 1)$. This gives another normal distribution with twice the variance, $\mathcal{N}(0, 2)$, which leads simply to the generalised χ_d distribution

$$p(w_{ij}) = \frac{w_{ij}^{d-1}e^{-w_{ij}^2/4}}{2^{d-1}\Gamma(\frac{d}{2})}. \tag{52}$$

Considering the definition of $\rho(\boldsymbol{x}, \boldsymbol{y})$ in Eq. 10, it is straightforward to calculate

$$\rho\left(\boldsymbol{x}, \boldsymbol{y}\right) = \Gamma(\frac{d}{2}) \int_0^\infty \mathrm{d}w \frac{w^{d-1} e^{-\frac{w^2}{4}}}{2^{d-1}\Gamma(\frac{d}{2})} \sum_{k=0}^{\infty} \frac{v^{2k} w^{2k}}{2^{2k} k! \Gamma(k+\frac{d}{2})} = \sum_{k=0}^{\infty} \frac{v^{2k}}{k!} = e^{v^2}, \tag{53}$$

as reported in the main text. We used the fact that all w_{ij} follow the same distribution and suppressed the ij subscripts for notational clarity. To perform the integral over w, we used the identity $\int_{w=0}^{\infty} \mathrm{d}w w^{2z-1} e^{-\frac{w^2}{2}} = 2^{z-1}\Gamma(z)$. □

This result is obtained more quickly by noting that, following the notation in Sec. A.1, $\rho(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{m(m-1)} \sum_i \sum_{i \neq j} \mathbb{E}[e^{(\boldsymbol{w}_i + \boldsymbol{w}_j)^\top \boldsymbol{v}}] = \mathbb{E}[e^{\boldsymbol{w}_1^\top \boldsymbol{v}}]\mathbb{E}[e^{\boldsymbol{w}_2^\top \boldsymbol{v}}]$. We used the fact that $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$ are independent and that all $\boldsymbol{w}_{ij}$ follow the same distribution (then choosing $i = 1$ and $j = 2$ wlg). We have already seen that $\mathbb{E}[e^{\boldsymbol{w}_1^\top \boldsymbol{v}}] = e^{\frac{v^2}{2}}$ (in fact the condition for unbiased estimation of $\widehat{K}$), which immediately yields $\rho(\boldsymbol{x}, \boldsymbol{y}) = e^{v^2}$. But the approach using $p(w_{ij})$ is a good warmup for the theorems that follow and will permit a more unified account.

A.5. Proof of Lemma 4.2 (PDF for Vectors Subtending θ)

In this appendix, we derive the form of Eq. 20, the probability distribution of $w_{ij} = \|\boldsymbol{w}_i + \boldsymbol{w}_j\|_2$ if $\boldsymbol{w}_i, \boldsymbol{w}_j \in \mathbb{R}^d$ are marginally Gaussian vectors conditioned to subtend a fixed angle θ. Later, the special cases of $\theta = \frac{\pi}{2}$ (orthogonal) and $\theta = \arccos(-\frac{1}{d-1})$ (simplex) will be of particular interest.

Clearly $w^2 = w_i^2 + w_j^2 + 2w_i w_j \cos\theta$, with weight magnitudes $w_{i,j} \sim \chi_d$ (we have suppressed the ij subscript, replacing w_{ij} by w, to minimise notational clutter). Diagonalising the quadratic form, we see that a constant w surface will trace out an ellipse in (w_i, w_j) space with semi-major (-minor) axis lengths $\frac{w}{\sqrt{1 \pm \cos(\theta)}}$. Now

$$p(w < w') = \int_{\mathcal{A}} p_\chi(w_i) p_\chi(w_j) \mathrm{d}w_i \mathrm{d}w_j \tag{54}$$

where p_χ denotes the χ_d distribution obeyed by $w_{i,j}$ and $\mathcal{A}$ denotes the area in the positive quadrant bounded by an ellipse of constant $w = w'$ (recall that $w_{i,j} \geq 0$ since these are vector magnitudes). Expressing this in polar coordinates,

$$p(w < w') = \int_{\phi=0}^{\pi/2} \mathrm{d}\phi \int_{r=0}^{\frac{w'}{\sqrt{1+\sin(2\phi)\cos(\theta)}}} \mathrm{d}r r p_\chi(r\cos\phi) p_\chi(r\sin\phi). \tag{55}$$

Differentiating wrt w' to get the pdf,

$$\begin{aligned} p(w) &= \int_{\phi=0}^{\pi/2} \mathrm{d}\phi \frac{w}{\sqrt{1+\sin(2\phi)\cos(\theta)}} p_\chi(\frac{w\cos\phi}{\sqrt{1+\sin(2\phi)\cos(\theta)}}) p_\chi(\frac{w\sin\phi}{1+\sin(2\phi)\cos(\theta)}) \\ &= \frac{w^{2d-1}}{2^{d-2}\Gamma(\frac{d}{2})^2} \int_{\phi=0}^{\pi/2} \mathrm{d}\phi (\sin\phi\cos\phi)^{d-1} \frac{e^{-\frac{w^2}{2(1+\sin 2\phi\cos\theta)}}}{(1+\sin 2\phi\cos\theta)^d}, \end{aligned} \tag{56}$$

as reported in Eq. 20 of the main text. □

As an aside, it is instructive to set $\theta = \pi/2$ and inspect the form of $p(w)$. Doing so, we arrive at the integral

$$\begin{aligned} p(w) = \frac{w^{2d-1}}{2^{d-2}\Gamma(\frac{d}{2})^2} \int_{\phi=0}^{\pi/2} \mathrm{d}\phi (\sin\phi\cos\phi)^{d-1} e^{-\frac{w^2}{2}} &= \frac{w^{2d-1}}{2^{2d-2}\Gamma(\frac{d}{2})^2} \int_{\phi=0}^{\pi} \mathrm{d}\phi (\sin\phi)^{d-1} e^{-\frac{w^2}{2}} \\ &= \sqrt{\pi} \frac{w^{2d-1}}{2^{2d-2}\Gamma(\frac{d}{2})\Gamma(\frac{d}{2}+\frac{1}{2})} e^{-\frac{w^2}{2}}. \end{aligned} \tag{57}$$

Recalling the Legendre duplication formula, $\sqrt{\pi}\Gamma(2z) = 2^{2z-1}\Gamma(z)\Gamma(z+\frac{1}{2})$, this reduces to

$$p(w) = \frac{w^{2d-1} e^{-w^2/2}}{2^{d-1}\Gamma(d)}. \tag{58}$$

This is nothing other than the χ-distribution with $2d$ degrees of freedom. This makes intuitive sense because, since $\boldsymbol{w}_i$ and $\boldsymbol{w_j}$ are orthogonal, it follows that $w^2 = w_i^2 + w_j^2$. Now $w_{i,j}$ follow χ_d distributions (square root of sum of squares of d standard normal variates), so w must be a square root of sum of squares of $2d$ standard normal variates – that is, a χ_{2d} distribution.

A.6. Proof of Theorem 4.3 (ORF and SimRF Conformity Closed Forms)

Here, we derive the RF-conformities $\rho(\boldsymbol{x}, \boldsymbol{y})$ of the ORF and SimRF variants.

Recall the form of $\rho(\boldsymbol{x}, \boldsymbol{y})$, defined in Eq. 10 and reproduced here for convenience:

$$\rho(\boldsymbol{x}, \boldsymbol{y}) = \frac{\Gamma(\frac{d}{2})}{m(m-1)} \sum_i \sum_{j \neq i} \mathbb{E}_{w_{ij}} \left(\sum_{k=0}^{\infty} \frac{v^{2k} w_{ij}^{2k}}{2^{2k} k! \Gamma(k + \frac{d}{2})} \right). \tag{59}$$

Use the probability distribution for two marginally Gaussian weights conditioned to subtend an angle θ,

$$p(w_{ij}) = \frac{w_{ij}^{2d-1}}{2^{d-2}\Gamma(\frac{d}{2})^2} \int_{\phi=0}^{\pi/2} \mathrm{d}\phi (\sin\phi\cos\phi)^{d-1} \frac{e^{-\frac{w_{ij}^2}{2(1+\sin 2\phi\cos\theta)}}}{(1+\sin 2\phi\cos\theta)^d}, \tag{60}$$

where $w_{ij} = \|\boldsymbol{w}_i^2 + \boldsymbol{w}_j^2\|_2$ with $i \neq j$ (see Lemma 4.2 and the accompanying proof in Sec. A.5). Since all w_{ij} follow the same distribution, the sums give a multiplicative factor of $m(m-1)$ that cancels with the denominator. Now we have

$$\rho_\theta(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k=0}^{\infty} \frac{v^{2k}}{2^{2k} k! 2^{d-2} \Gamma(\frac{d}{2}) \Gamma(k + \frac{d}{2})} \int_{w=0}^{\infty} \mathrm{d}w \int_{\phi=0}^{\frac{\pi}{2}} \mathrm{d}\phi w^{2k+2d-1} (\sin\phi\cos\phi)^{d-1} \frac{e^{-\frac{w^2}{2(1+\sin 2\phi\cos\theta)}}}{(1+\sin 2\phi\cos\theta)^d}. \tag{61}$$

Changing variables $w \to w\sqrt{1+\sin 2\phi\cos\theta}$ and doing the integral over w,

$$\rho_\theta(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k=0}^{\infty} \frac{v^{2k}\Gamma(k+d)}{2^k k! 2^{d-2} \Gamma(\frac{d}{2}) \Gamma(k + \frac{d}{2})} \int_{\phi=0}^{\frac{\pi}{2}} \mathrm{d}\phi (\sin 2\phi)^{d-1} (1+\sin 2\phi\cos\theta)^k. \tag{62}$$

Finally, changing variables $\phi \to \frac{\phi}{2}$ and rearranging, we arrive at

$$\rho_\theta(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2^{d-1}\Gamma(\frac{d}{2})} \int_0^{\pi} \mathrm{d}\phi (\sin\phi)^{d-1} \cdot \sum_{k=0}^{\infty} \frac{v^{2k}(1+\sin\phi\cos\theta)^k}{2^k k! \Gamma(k + \frac{d}{2})} \Gamma(k+d) \tag{63}$$

as reported in Eq. 21 of the main text.

Now we substitute in the values of θ corresponding to the particular cases of ORFs and SimRFs.

1) ORFs: $\cos\theta = 0$

Note that

$$\frac{1}{\Gamma(\frac{d}{2})} \int_{\phi=0}^{\pi} \mathrm{d}\phi (\sin\phi)^{d-1} = \sqrt{\pi} \frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d}{2} + \frac{1}{2}) \Gamma(\frac{d}{2})} = 2^{d-1} \frac{\Gamma(\frac{d}{2})}{\Gamma(d)} \tag{64}$$

where we used the identity $\int_0^{\pi} \mathrm{d}x \sin^d x = \frac{\sqrt{\pi}\Gamma(\frac{d}{2}+\frac{1}{2})}{\Gamma(\frac{d}{2}+1)}$ and the Legendre duplication formula. It follows immediately that

$$\rho_{\text{ORF}}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\Gamma(\frac{d}{2})}{\Gamma(d)} \sum_{k=0}^{\infty} \frac{v^{2k}}{2^k k!} \frac{\Gamma(k+d)}{\Gamma(k + \frac{d}{2})}. \tag{65}$$

We could have obtained this more directly using the χ_{2d} distribution (see discussion at the end of Sec. A.5), but leaving θ unspecified for as long as possible permits a more direct comparison with SimRFs.

2) SimRFs: $\cos\theta = -\frac{1}{d-1}$

Carrying out the binomial expansion,

$$\begin{aligned} \int_0^{\pi} \mathrm{d}\phi (\sin\phi)^{d-1} \left(1 - \frac{\sin\phi}{d-1}\right)^k &= \sum_{p=0}^{k} \frac{k!}{(k-p)!p!} \left(-\frac{1}{d-1}\right)^p \int_0^{\pi} \mathrm{d}\phi (\sin\phi)^{d+p-1} \\ &= \sum_{p=0}^{k} \frac{k!}{(k-p)!p!} \left(-\frac{1}{d-1}\right)^p \sqrt{\pi} \frac{\Gamma(\frac{d+p}{2})}{\Gamma(\frac{d+p+1}{2})}. \end{aligned} \tag{66}$$

Substituting this in, we immediately arrive at

$$\rho_{\mathrm{SimRF}}(\boldsymbol{x},\boldsymbol{y}) = \frac{\sqrt{\pi}}{\Gamma(\frac{d}{2})2^{d-1}} \sum_{k=0}^{\infty} \frac{\Gamma(k+d)}{\Gamma(k+\frac{d}{2})} \frac{v^{2k}}{2^k} \sum_{p=0}^{k} \left(-\frac{1}{d-1}\right)^p \frac{\Gamma(\frac{d+p}{2})}{\Gamma(\frac{d+p+1}{2})} \frac{1}{(k-p)!p!} \tag{67}$$

which we have seen is smaller than $\rho_{\mathrm{ORF}}(\boldsymbol{x},\boldsymbol{y})$. □

A.7. Proof of Corollary 4.4 (ORFs Always Outperform IIDRFs)

Here we derive an analytic expression for the orthogonality gap (difference in kernel estimator MSE between the IIDRF and ORF mechanisms) and show that it is positive everywhere.

From Eq. 11, we immediately have that

$$\Delta\mathrm{MSE}(\widehat{K}(\mathbf{x},\mathbf{y})) = e^{-2(x^2+y^2)} \frac{m-1}{m} \left(\rho_{\mathrm{IIDRF}}(\boldsymbol{x},\boldsymbol{y}) - \rho_{\mathrm{ORF}}(\boldsymbol{x},\boldsymbol{y})\right). \tag{68}$$

Inserting the respective RF-conformities from Eqs. 19 and 23,

$$\begin{aligned}\Delta\mathrm{MSE}(\widehat{K}(\mathbf{x},\mathbf{y})) &= e^{-2(x^2+y^2)} \frac{m-1}{m} \left(e^{v^2} - \frac{\Gamma(\frac{d}{2})}{\Gamma(d)} \sum_{k=0}^{\infty} \frac{v^{2k}}{2^k k!} \frac{\Gamma(k+d)}{\Gamma(k+\frac{d}{2})}\right) \\ &= e^{-2x^2-2y^2} \frac{m-1}{m} \sum_{k=0}^{\infty} \frac{v^{2k}}{k!} \left(1 - \frac{(k+d-1)!}{(d-1)!} \frac{(d-2)!!}{(2k+d-2)!!}\right),\end{aligned} \tag{69}$$

where !! denotes the double factorial. We can write the term in parentheses as

$$1 - \frac{d}{d} \cdot \frac{d+1}{d+2} \cdot \ldots \cdot \frac{d+k-2}{d+2(k-2)} \cdot \frac{d+k-1}{d+2(k-1)} > 0 \tag{70}$$

so the series expansion is positive. It follows that the kernel estimator MSE with ORFs is upper bounded but that of IIDRFs. □

A.8. Proof of Theorem 4.5 (RFF Orthogonality Gap)

Here, we demonstrate how, with minor modifications, many of the stated results for PRFs can be translated to RFFs.

Recall the RFF definition, stated in Eq. 3 of the main text and reproduced here for convenience.

$$\phi_{\mathrm{RFF}}(\boldsymbol{z}) \stackrel{\mathrm{def}}{=} \sqrt{\frac{1}{m}} (\odot_{i=1}^{m} [\sin(\boldsymbol{w}_i^\top \boldsymbol{z}), \cos(\boldsymbol{w}_i^\top \boldsymbol{z})])^\top. \tag{71}$$

Now we have that

$$\widehat{K} = \phi(\boldsymbol{x})^\top \phi(\boldsymbol{y}) = \frac{1}{m} \sum_{i=1}^{m} \cos \boldsymbol{w}_i^\top (\boldsymbol{x}-\boldsymbol{y}) = \frac{1}{m} \sum_{i=1}^{m} a_i \tag{72}$$

where we defined

$$a_i \stackrel{\mathrm{def}}{=} \cos(\boldsymbol{w}_i^\top \boldsymbol{z}) \tag{73}$$

and let $\boldsymbol{z} = \boldsymbol{x} - \boldsymbol{y}$. It is straightforward to show that $\widehat{K}$ is an unbiased estimator of the Gaussian kernel $e^{-\frac{\|\boldsymbol{x}-\boldsymbol{y}]|_2^2}{2}}$ when we sample $\boldsymbol{w}_i \sim \mathcal{N}(0, \mathbf{I}_d)$; that is, $\mathbb{E}[a_i] = e^{-\frac{z^2}{2}}$. After some work, we also have that

$$\mathrm{MSE}(\widehat{K}) = \frac{1}{m} \left(\frac{(1-e^{-z^2})^2}{2} + (m-1)(\frac{1}{m(m-1)} \sum_i \sum_{i \neq j} \mathbb{E}[a_i a_j] - e^{-z^2}) \right). \tag{74}$$

The object of interest (which is precisely the analogue of $\rho(\boldsymbol{x},\boldsymbol{y})$ but for RFFs) is $\sum_i \sum_{i\neq j} \mathbb{E}[a_i a_j] = \sum_i \sum_{i \neq j} \mathbb{E}[\cos \boldsymbol{w}_i^\top \boldsymbol{z} \cos \boldsymbol{w}_j^\top \boldsymbol{z}]$. It will vary depending on any geometrical coupling scheme employed.

From elementary trigonometry, $\cos \boldsymbol{w}_i^\top \boldsymbol{z} \cos \boldsymbol{w}_j^\top \boldsymbol{z} = \frac{1}{2}(\cos\big((\boldsymbol{w}_i + \boldsymbol{w}_j)^\top \boldsymbol{z}\big) + \cos\big((\boldsymbol{w}_i - \boldsymbol{w}_j)^\top \boldsymbol{z}\big)$. It is also simple to convince oneself that, when the random vectors $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$ are (a) i.i.d. or (b) conditioned to be orthogonal, the distributions of the two random variables $\boldsymbol{w}_i + \boldsymbol{w}_j$ and $\boldsymbol{w}_i - \boldsymbol{w}_j$ are identical. As such, we can just consider the single random variable $\boldsymbol{w}_{ij} = \boldsymbol{w}_i + \boldsymbol{w}_j$ wlg. Then we have that

$$\begin{aligned}\mathbb{E}[\cos \boldsymbol{w}_{ij}^\top \boldsymbol{z}] &= \int_{\mathbb{R}_d} p(\boldsymbol{w}_{ij}) e^{-i\boldsymbol{w}_{ij}^\top \boldsymbol{z}} \mathrm{d}^d \boldsymbol{w}_{ij} \\ &= \Gamma(d/2) 2^{\frac{d}{2}-1} \int_0^\infty \mathrm{d}w_{ij} p(w_{ij}) (w_{ij} z)^{1-\frac{d}{2}} J_{\frac{d}{2}-1}(w_{ij} z)\end{aligned} \tag{75}$$

where we have used that the probability distribution $p(\boldsymbol{w}_{ij})$ is real regardless of whether the random vectors are i.i.d. or orthogonal, and written the expression as a Hankel transform (Faris, 2008). Note that we do *not* consider the simplex coupling case, where the random variables $\boldsymbol{w}_i + \boldsymbol{w}_j$ and $\boldsymbol{w}_i - \boldsymbol{w}_j$ will follow different distributions.

Carefully comparing with Eq. 31 in Sec. A.1, we observe that **the expression is identical to the PRF case, but instead taking** $v \to -iz$. This means that we can obtain all the previously stated IIDRF and ORF results in the RFF setting with minimal extra work. For instance, inspecting Eq. 25, we can immediately state the RFF orthogonality gap (difference in kernel estimator MSE between IIDRFs and ORFs) reported in Theorem 4.5:

$$\Delta\mathrm{MSE}(\widehat{K}(\mathbf{x}, \mathbf{y})) = \frac{m-1}{m}\left(e^{-z^2} - \frac{\Gamma(d/2)}{\Gamma(d)} \sum_{k=0}^{\infty} \frac{(-z^2)^k}{2^k k!} \frac{\Gamma(k+d)}{\Gamma(k+d/2)}\right). \tag{76}$$

(Note that we also dropped the exponential prefactor $e^{-2x^2-2y^2}$, originating from the definition of PRFs (7) where it is needed to keep kernel estimation unbiased). □

A.9. Proof of Corollary 4.6 (RFF Asymptotic MSE Ratio)

Here we derive Eq. 27, the ratio of ORF to IIDRF kernel estimator MSEs in the $d \to \infty$ limit. This was first reported in (Yu et al., 2016), and is included here to show consistency with our more general (finite d) closed forms.

Considering the discussion in Sec. A.8, it is straightforward to reason that the ratio of MSEs is given by

$$\frac{\mathrm{MSE}_{\mathrm{ORF}}}{\mathrm{MSE}_{\mathrm{IIDRF}}} = 1 + \frac{2(m-1)}{(1+e^{-z^2})^2}(\mathbb{E}_{\mathrm{ort}}(a_1 a_2) - e^{-z^2}) \tag{77}$$

where $a_i = e^{-i\boldsymbol{w}_i^\top \boldsymbol{z}}$ and the expectation is being taken over the random variable $w_{12} = \|\boldsymbol{w}_1 + \boldsymbol{w}_2\|_2$, with $\boldsymbol{w}_{1,2}$ conditioned to be orthogonal (see e.g. Eq. 58 for the appropriate probability distribution). From the discussion in Sec. A.8, the term in parentheses on the right can be written as the series expansion

$$\sum_{k=0}^{\infty} \frac{(-z^2)^k}{k!}\left(\frac{1}{2^k}\frac{\Gamma(k+d)\Gamma(\frac{d}{2})}{\Gamma(k+\frac{d}{2})\Gamma(d)} - 1\right). \tag{78}$$

Recalling Stirling's formula for the asymptotic form of the Gamma function,

$$\lim_{x\to\infty} \Gamma(x+1) = \sqrt{2\pi x} e^{-x} x^x, \tag{79}$$

we can rewrite term

$$\begin{aligned}\lim_{d\to\infty} \frac{1}{2^k}\frac{\Gamma(k+d)\Gamma(\frac{d}{2})}{\Gamma(k+\frac{d}{2})\Gamma(d)} &= \frac{1}{2^k}\sqrt{\frac{(k+d-1)(\frac{d}{2}-1)}{(k+\frac{d}{2}-1)(d-1)}}\frac{(k+d-1)^{m+d-1}(\frac{d}{2}-1)^{\frac{d}{2}-1}}{(k+\frac{d}{2}-1)^{k+\frac{d}{2}-1}(d-1)^{d-1}} \\ &= \sqrt{\frac{(k+d-1)(\frac{d}{2}-1)}{(k+\frac{d}{2}-1)(d-1)}} \cdot \frac{1}{2^k}\left(\frac{k+d-1}{k+\frac{d}{2}-1}\right)^k \cdot \frac{\left(1+\frac{k}{d-1}\right)^{d-1}}{\left(1+\frac{k}{\frac{d}{2}-1}\right)^{\frac{d}{2}-1}}.\end{aligned} \tag{80}$$

We can Taylor expand each of the constituent components,

$$\sqrt{\frac{(k+d-1)(\frac{d}{2}-1)}{(k+\frac{d}{2}-1)(d-1)}} \simeq 1 - \frac{k}{2d} \tag{81}$$

$$\frac{1}{2^k}\left(\frac{k+d-1}{k+\frac{d}{2}-1}\right)^k \simeq 1-\frac{k(k-1)}{d} \tag{82}$$

$$\frac{\left(1+\frac{k}{d-1}\right)^{d-1}}{\left(1+\frac{k}{\frac{d}{2}-1}\right)^{\frac{d}{2}-1}} \simeq 1+\frac{k^2}{2d} \tag{83}$$

which combine to yield

$$\frac{1}{2^k}\frac{\Gamma(k+d)\Gamma(\frac{d}{2})}{\Gamma(k+\frac{d}{2})\Gamma(d)} \simeq 1-\frac{k(k-1)}{2d}. \tag{84}$$

It follows that

$$\sum_{k=0}^{\infty}\frac{(-z^2)^k}{k!}\left(\frac{1}{2^k}\frac{\Gamma(k+d)\Gamma(\frac{d}{2})}{\Gamma(k+\frac{d}{2})\Gamma(d)}-1\right) = -\sum_{k=0}^{\infty}\frac{(-z^2)^k}{k!}\frac{k(k-1)}{2d} = -\frac{z^4}{2d}e^{-z^2}. \tag{85}$$

Putting this into Eq. 77,

$$\frac{\text{MSE}(\widehat{K}_{\text{ORF}})}{\text{MSE}(\widehat{K}_{\text{IIDRF}})} = 1-(m-1)\left(\frac{e^{-z^2}z^4}{d(1-e^{-z^2})^2}+\mathcal{O}\left(\frac{1}{d^2}\right)\right), \tag{86}$$

as reported in Eq. 27 and (Yu et al., 2016). Importantly, the negative sign of the subleading term means that, in the RFF setting, ORFs will always outperform IIDRFs when $d \to \infty$. □

B. Experimental Details

In this appendix, we provide further experimental details to supplement the discussion in Sec. 5.

B.1. Choosing σ

Here we elaborate on Sec 5.3, where we report tuning the hyperparameter σ with a validation dataset. In particular, given some fixed dataset $(\boldsymbol{x}, \boldsymbol{y})$ and a Gaussian kernel $K(\boldsymbol{x}, \boldsymbol{x}') = e^{-\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|_2^2}{2}}$, we apply the scalar transformation $(\boldsymbol{x}, \boldsymbol{y}) \to (\sigma\boldsymbol{x}, \boldsymbol{y})$ with $\sigma \in \mathbb{R}^+$ to optimise the IIDRF performance. There are two (potentially competing) factors to consider:

1. σ implicitly controls the smoothness of the the kernel K that we are approximating. Multiplying the data by σ is equivalent to rescaling the kernel characteristic lengthscale by $\frac{1}{\sigma}$, i.e. taking $K(\boldsymbol{x}, \boldsymbol{x}') = e^{-\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|_2^2}{2}} \to e^{-\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|_2^2}{2/\sigma^2}}$. This will change the classifier accuracy even when using the exact kernel.

2. The kernel estimator variance has some dependence on the data $(\boldsymbol{x}, \boldsymbol{x}')$ – consider e.g. any of the results in Sec. 4, or Fig. 6. Roughly speaking, PRFs tend to perform worse at large σ (equivalent to a sharply varying kernel).

In order to navigate a possible tradeoff between these factors and pick a suitable value for σ, we tune by running a coarse search optimising classifier accuracy on a validation set. This is sensible because we are specifically interested in comparing the performance of SimRFs, ORFs and IIDRFs in settings where kernel approximation with random features is already effective. Fig. 10 shows the results; we choose the value of σ at which the i.i.d. PRF classification accuracy (orange solid line) peaks. As we have suggested, this does *not* generically coincide with where the exact kernel performance (blue dotted line) peaks.

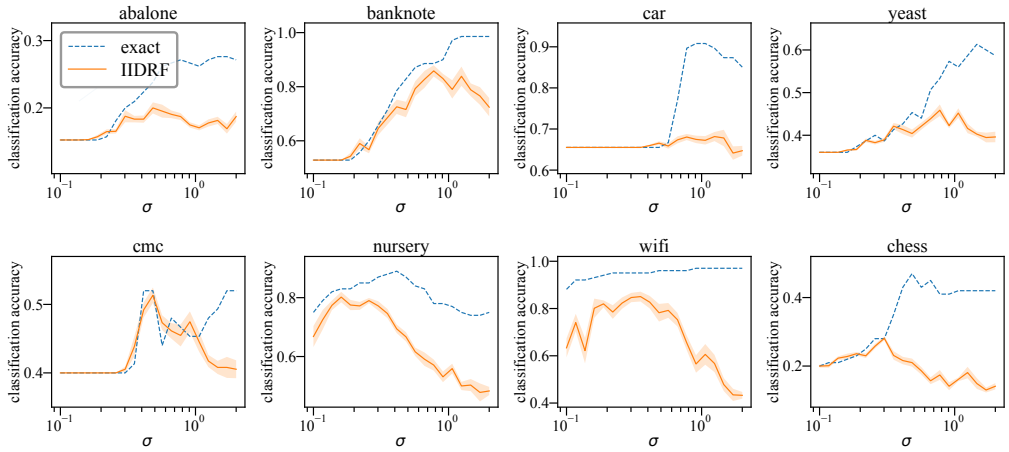

Figure 10. Plots showing classification accuracy vs the data rescaling factor σ for each of the nonparametric classification validation datasets, including both the exact (dotted blue) and IIDRF (solid orange) kernels. They are used for tuning the σ hyperparameter. The RFs are of dimensionality $m = 10d$, with d the data dimensionality, and the shaded region gives one standard deviation on the estimate of the mean classification accuracy over $N = 10$ samples. During this coarse σ search phase, the large nursery and chess datasets are restricted to 1000 training examples and 100 test examples for speed.

There is also a broader question about the relationship between the kernel estimator MSE and the performance in downstream tasks. The fact that SimRFs consistently outperform ORFs and IIDRFs in a variety of nonparametric classification tasks and when used to approximate the attention module in Performers confirms that lower variance on kernel estimates often helps performance in applications. But making rigorous mathematical statements about the effect of estimator variance – especially, why its importance differs between tasks – is complicated and is left as an open question.

B.2. Fast SimRFs: Further Discussion and Experimental Results

In this appendix, we provide more detailed discussion of fast SimRFs (Sec. 4.2) and demonstrate a simple implementation on the nonparametric classification task. Experiments will be closely related to those described in Sec. 5.3 so readers are advised to review this section first.

Recall the definition of the simplex block,

$$\mathbf{W}_{\text{simp}} = \mathbf{DSR}, \tag{87}$$

where $\mathbf{D} \in \mathbb{R}^{d \times d} = \text{diag}(w_i)$ with w_i sampled from a χ_d-distribution. $\mathbf{R} \in \mathbb{R}^{d \times d}$ is a random orthogonal matrix drawn from Haar measure on $\text{O}(d)$, the group of orthogonal matrices in $\mathbb{R}^{d \times d}$. The rows $\boldsymbol{s}_i$ of the simplex projection matrix $\mathbf{S} \in \mathbb{R}^{d \times d}$ are given by the simplex unit vectors, defined in Eq. 9 and reproduced below for convenience.

$$\boldsymbol{s}_i = \begin{cases} \sqrt{\frac{d}{d-1}}\mathbf{e}_i - \frac{\sqrt{d}+1}{(d-1)^{3/2}}(1, ..., 1, 0)^\top & \text{for } 1 \leq i < d \\ \frac{1}{\sqrt{d-1}}(1, 1, ..., 1, 0)^\top & \text{for } i = d. \end{cases} \tag{88}$$

Recall further that, with fast SimRFs, we replace the matrix $\mathbf{R}$ by an orthogonal proxy $\widetilde{\mathbf{R}}$ that is only approximately sampled from Haar measure, but which supports fast matrix-vector multiplication.

B.2.1. S Supports Fast Matrix-Vector Multiplication

We begin by showing that the matrix $\mathbf{S}$ supports fast matrix-vector multiplication, as stated in the main body. The following simple algorithm of time-complexity $\mathcal{O}(d)$ calculates $\mathbf{S}\boldsymbol{x}$, with $\boldsymbol{x} \in \mathbb{R}^d$.

Algorithm 1 Fast matrix-vector multiplication with $\mathbf{S}$

Input: object vector $\boldsymbol{x} \in \mathbb{R}^d$ with components $x_i, i = 1, ..., d$
Output: 'simplex projection' vectors $\boldsymbol{y} = \mathbf{S}\boldsymbol{x} \in \mathbb{R}^d$
Main:
$y_d = \frac{1}{\sqrt{d-1}} \sum_{i=1}^{d-1} x_i$
for $i = 1$ **to** $d-1$ **do**
  $y_i = \sqrt{\frac{d}{d-1}} x_i - \frac{\sqrt{d}+1}{d-1} y_d$
end for

We note that, in downstream applications such as the SimRFs-Performer, the time taken by this extra simplex projection (whether using the fast implementation or not) is typically dwarfed by other computational requirements. It is rarely observable. That said, the extra time cost compared to ORFs is technically nonzero and constitutes the only real weakness of SimRFs.

B.2.2. Implementation using $\mathbf{HD}$-Product Matrices

To demonstrate one possible implementation, we use the so-called $\mathbf{HD}$-product matrices, formed by multiplication of $k \in \mathbb{N}$ $\mathbf{HD}$ blocks,

$$\widetilde{\mathbf{R}} = \prod_{i=1}^{k} \mathbf{HD}_i^{(\mathcal{R})}. \tag{89}$$

Here, $\mathbf{H}$ is the normalised Hadamard matrix, defined by the recursive relation

$$\mathbf{H}_1 = (1), \qquad \mathbf{H}_i = \frac{1}{\sqrt{2}} \begin{pmatrix} \mathbf{H}_{i-1} & \mathbf{H}_{i-1} \\ \mathbf{H}_{i-1} & -\mathbf{H}_{i-1} \end{pmatrix} \quad \text{for } i > 1, \tag{90}$$

and $\mathbf{D}_i^{(\mathcal{R})} = \text{diag}(d_i)$ with $d_i \sim \text{Unif}(\{\pm 1\})$, i.i.d. Rademacher random variables. $\mathbf{HD}$-blocks have previously received attention for dimensionality reduction (Ailon & Chazelle, 2009), locally-sensitive hashing methods (Bojarski et al., 2017) and kernel approximation (Choromanski et al., 2017), where they exhibit good computational and statistical properties. Importantly, given some vector $\boldsymbol{x} \in \mathbb{R}^d$, the matrix-vector product $\mathbf{H}\boldsymbol{x}$ can be computed in time $\mathcal{O}(d \log d)$ via the fast Walsh-Hadamard transform.

We report the results of the nonparametric classification tasks described in Sec. 5.3, now inserting $\mathbf{HD}$-product matrices with $k = 3$ in place of $\mathbf{R}$. With $m = d$ random features, we observe that using fast SimRFs and fast ORFs does not substantially change the accuracy of nonparametric classification. Results are frequently identical to the regular case.

Table 3. Classification accuracies from kernel regression with IIDRFs, ORFs and SimRFs, where we include both regular and fast implementations in the latter two cases. Replacing the random orthogonal matrix $\mathbf{R}$ (sampled from Haar measure) with a structured $\mathbf{HD}$-product does not change the accuracy of nonparametric classification.

Dataset	Classification accuracy				
	IIDRFs	ORFs		SimRFs	
		Regular	Fast	Regular	Fast
abalone	0.1432 ± 0.0003	0.1445 ± 0.0003	0.1447 ± 0.0003	0.1455 ± 0.0003	0.1462 ± 0.0003
banknote	0.6441 ± 0.0024	0.6612 ± 0.0025	0.6596 ± 0.0024	0.7196 ± 0.0019	0.7296 ± 0.0017
car	0.6768 ± 0.0006	0.6788 ± 0.0006	0.6784 ± 0.0006	0.6797 ± 0.0006	0.6800 ± 0.0006
yeast	0.3187 ± 0.0006	0.3193 ± 0.0006	0.3171 ± 0.0006	0.3187 ± 0.0006	0.3195 ± 0.0006
cmc	0.4088 ± 0.0009	0.4149 ± 0.0009	0.4159 ± 0.0009	0.4206 ± 0.0008	0.4222 ± 0.0008
nursery	0.5870 ± 0.0013	0.6213 ± 0.0019	0.6193 ± 0.0019	0.7030 ± 0.0008	0.7037 ± 0.0008
wifi	0.4914 ± 0.0026	0.5224 ± 0.0025	0.5310 ± 0.0024	0.6509 ± 0.0027	0.6533 ± 0.0027
chess	0.2011 ± 0.0002	0.2017 ± 0.0002	0.2016 ± 0.0002	0.2021 ± 0.0002	0.2021 ± 0.0002

Note that Hadamard matrices are defined such that $\mathbf{H} \in \mathbb{R}^{d \times d}$ with $d = 2^i, i \in \mathbb{N}$ a non-negative integer. Given data of some arbitrary dimensionality, we have to 'pad' each object vector $\boldsymbol{x}$ with 0s such that its length is a power of 2. This

accounts for the small discrepancies between the results reported in Table 3 and accuracies in Fig. 8, where vectors are not padded so $m = d$ is smaller. Results in each column of the table are for the same effective d so the regular and fast mechanisms can be safely compared.