Chefs' Random Tables: Non-Trigonometric Random Features – Appendix

9 Appendix

We also attach the code for producing: Table 1, Fig. 1 and Fig. 2.

9.1 Orthogonal random projections

The orthogonal random projections mechanism ([15]) is the Monte Carlo method, where samples $\omega_1, ..., \omega_M$, marginally distributed as $\mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$ (thus maintaining unbiasedness of the overall mechanism), are conditioned to form an orthogonal ensemble when $M \leq d$, otherwise samples are partitioned into $d \times d$ independent orthogonal blocks. Orthogonal random projections can be easily constructed form th iid projections via the Gram-Schmidt orthogonalization algorithm (see: [16]).

9.2 Proof of Theorem 3.1

Proof. We rewrite (1) for $f_{\mathrm{GE}}^{(\cdot)}$ and deduce that

$$\mathbb{E}\left(f_{\mathrm{GE}}^{(1)}(\omega, \mathbf{x}) f_{\mathrm{GE}}^{(2)}(\omega, \mathbf{y})\right) = (2\pi)^{-d/2} D^2 \int_{\mathbb{R}^d} \exp(-\|\omega\|^2/2 + 2A\|\omega\|^2 + B\omega^\top(\mathbf{x} + s\mathbf{y}) + C(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2)) d\omega \tag{16}$$

where we rewrite expectation as an integral and use definitions of $f_{\mathrm{GE}}^{(1)}(\omega, \mathbf{x})$, $f_{\mathrm{GE}}^{(2)}(\omega, \mathbf{y})$ and $p_{\mathrm{GE}}(\omega)$. Next, we move out constant terms from the integral and put ω into an elementwise square of difference:

$$\begin{aligned}\int_{\mathbb{R}^d} &\exp(-\|\omega\|^2/2 + 2A\|\omega\|^2 + B\omega^\top(\mathbf{x} + s\mathbf{y}) + C(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2)) d\omega \\ &= \exp\left(\frac{B^2}{2(1-4A)}\|\mathbf{x} + s\mathbf{y}\|^2 + C(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2)\right) \\ &\times \int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}(1-4A)\left[\omega - \frac{B}{1-4A}(\mathbf{x} + s\mathbf{y})\right]^2\right) d\omega\end{aligned} \tag{17}$$

where by $[\cdot]^2$ we denote an elementwise square of the input vector. Next, we use an identity:

$$\int_{\mathbb{R}^d} \exp\left(-\frac{\alpha}{2}\left[\omega - \beta\right]^2\right) d\omega = (2\pi)^{d/2}\left(\sqrt{\alpha}\right)^{-d}. \tag{18}$$

where $\alpha \in \mathbb{C}$, $\mathrm{Re}\,(\alpha) > 0$ ($\alpha = 1 - 4A$ in (9.2)) and $\beta \in \mathbb{C}^d$ ($\beta = (B/(1-4A))(\mathbf{x} + s\mathbf{y})$ in (9.2)). When both α and β are real, (18) is correct since it is integral of the scaled multivariate Gaussian density. Since both the left and the right hand side in (18) are analytic functions of α and β when $\mathrm{Re}\,(\alpha) > 0$, by the identity theorem [56] we conclude that (18) holds when α and β are complex and $\mathrm{Re}\,(\alpha) > 0$. Applying (18) to (17), we deduce that

$$\int_{\mathbb{R}^d} \exp\left(-\frac{1}{2}(1-4A)\left[\omega - \frac{B}{1-4A}(\mathbf{x} + s\mathbf{y})\right]^2\right) d\omega = (2\pi)^{d/2}\left(\sqrt{1-4A}\right)^{-d}. \tag{19}$$

Combining (16, 17, 19) together, we conclude that

$$\mathbb{E}\left(f_{\mathrm{GE}}^{(1)}(\omega, \mathbf{x}) f_{\mathrm{GE}}^{(2)}(\omega, \mathbf{y})\right) = D^2\left(\sqrt{1-4A}\right)^{-d} \exp\left(\frac{B^2}{2(1-4A)}\|\mathbf{x} + s\mathbf{y}\|^2 + C(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2)\right).$$

The right hand side of (9.2) is $K(\mathbf{x}, \mathbf{y})$ if the following conditions are satisfied in addition to $\mathrm{Re}\,(1-4A) > 0$:

$$D^2 = (\sqrt{1-4A})^d, \quad sB^2 = (1-4A), \quad \frac{B^2}{2(1-4A)} + C = -\frac{1}{2}. \tag{20}$$

(20) is satisfied when (5) takes place. The final thing to mention is that $\mathrm{Re}\,(\cdot)$ is a linear operation, and therefore, if (20) is satisfied,

$$\mathbb{E}\mathrm{Re}\left(f_{\mathrm{GE}}^{(1)}(\omega,\mathbf{x})f_{\mathrm{GE}}^{(2)}(\omega,\mathbf{y})\right)=\mathrm{Re}\left(\mathbb{E}\left(f_{\mathrm{GE}}^{(1)}(\omega,\mathbf{x})f_{\mathrm{GE}}^{(2)}(\omega,\mathbf{y})\right)\right)=\mathrm{Re}\,(K(\mathbf{x},\mathbf{y}))=K(\mathbf{x},\mathbf{y}).$$

□

It's possible to use other complex roots in (5) rather than just principal roots. However, in the proof of Theorem 3.2, we will only use (20) and, therefore, the variance is the same when other complex roots are used. We opt for principal roots for simplicity.

9.3 Proof of Theorem 3.2

Proof. We first use $\mathrm{Var}Z=\mathbb{E}Z^2-(\mathbb{E}Z)^2$ which holds for any random variable Z, e.g. $\mathrm{Re}\left(f_{\mathrm{GE}}^{(1)}(\omega,\mathbf{x})f_{\mathrm{GE}}^{(2)}(\omega,\mathbf{y})\right)$:

$$\mathrm{Var}_{p_{\mathrm{GE}}}\mathrm{Re}\left(f_{\mathrm{GE}}^{(1)}(\omega,\mathbf{x})f_{\mathrm{GE}}^{(2)}(\omega,\mathbf{y})\right)=\mathbb{E}\,\mathrm{Re}\left(f_{\mathrm{GE}}^{(1)}(\omega,\mathbf{x})f_{\mathrm{GE}}^{(2)}(\omega,\mathbf{y})\right)^2-\left(\mathbb{E}\,\mathrm{Re}\left(f_{\mathrm{GE}}^{(1)}(\omega,\mathbf{x})f_{\mathrm{GE}}^{(2)}(\omega,\mathbf{y})\right)\right)^2 \tag{21}$$

The second term transforms into $\left(\mathbb{E}\,\mathrm{Re}\left(f_{\mathrm{GE}}^{(1)}(\omega,\mathbf{x})f_{\mathrm{GE}}^{(2)}(\omega,\mathbf{y})\right)\right)^2=K(\mathbf{x},\mathbf{y})^2$. As for the first term, we use $\mathrm{Re}\,(z)=\frac{1}{2}(z+\overline{z})$ and get:

$$\mathbb{E}\,\mathrm{Re}\left(f_{\mathrm{GE}}^{(1)}(\omega,\mathbf{x})f_{\mathrm{GE}}^{(2)}(\omega,\mathbf{y})\right)^2=\frac{1}{4}\mathbb{E}\left(f_{\mathrm{GE}}^{(1)}(\omega,\mathbf{x})f_{\mathrm{GE}}^{(2)}(\omega,\mathbf{y})+\overline{f_{\mathrm{GE}}^{(1)}(\omega,\mathbf{x})f_{\mathrm{GE}}^{(2)}(\omega,\mathbf{y})}\right)^2 \tag{22}$$

We unfold the square of the sum:

$$\mathbb{E}\left(f_{\mathrm{GE}}^{(1)}(\omega,\mathbf{x})f_{\mathrm{GE}}^{(2)}(\omega,\mathbf{y})+\overline{f_{\mathrm{GE}}^{(1)}(\omega,\mathbf{x})f_{\mathrm{GE}}^{(2)}(\omega,\mathbf{y})}\right)^2=\mathbb{E}\Bigg(f_{\mathrm{GE}}^{(1)}(\omega,\mathbf{x})^2f_{\mathrm{GE}}^{(2)}(\omega,\mathbf{y})^2+2|f_{\mathrm{GE}}^{(1)}(\omega,\mathbf{x})|^2|f_{\mathrm{GE}}^{(2)}(\omega,\mathbf{y})|^2+\overline{f_{\mathrm{GE}}^{(1)}(\omega,\mathbf{x})^2f_{\mathrm{GE}}^{(2)}(\omega,\mathbf{y})^2}\Bigg) \tag{23}$$

Further, we observe that, again:

$$f_{\mathrm{GE}}^{(1)}(\omega,\mathbf{x})^2f_{\mathrm{GE}}^{(2)}(\omega,\mathbf{y})^2+\overline{f_{\mathrm{GE}}^{(1)}(\omega,\mathbf{x})^2f_{\mathrm{GE}}^{(2)}(\omega,\mathbf{y})^2}=2\mathrm{Re}\left(f_{\mathrm{GE}}^{(1)}(\omega,\mathbf{x})^2f_{\mathrm{GE}}^{(2)}(\omega,\mathbf{y})^2\right). \tag{24}$$

We use that in (23) and also put expectation inside the sum and $\mathrm{Re}\,(\cdot)$ due to linearity:

$$\begin{aligned}&\mathbb{E}\left(2\mathrm{Re}\left(f_{\mathrm{GE}}^{(1)}(\omega,\mathbf{x})^2f_{\mathrm{GE}}^{(2)}(\omega,\mathbf{y})^2\right)+2|f_{\mathrm{GE}}^{(1)}(\omega,\mathbf{x})|^2|f_{\mathrm{GE}}^{(2)}(\omega,\mathbf{y})|^2\right)\\&=2\mathrm{Re}\left(\mathbb{E}\left(f_{\mathrm{GE}}^{(1)}(\omega,\mathbf{x})^2f_{\mathrm{GE}}^{(2)}(\omega,\mathbf{y})^2\right)\right)+2\mathbb{E}\left(|f_{\mathrm{GE}}^{(1)}(\omega,\mathbf{x})|^2|f_{\mathrm{GE}}^{(2)}(\omega,\mathbf{y})|^2\right)\end{aligned} \tag{25}$$

Denote $f_{\mathrm{GE}}^{(1)}$ and $f_{\mathrm{GE}}^{(2)}$ with parameters A,B,C,D as $f_{A,B,C,D}^{(1)},f_{A,B,C,D}^{(2)}$. Then according to (4),

$$f_{\mathrm{GE}}^{(1)}(\omega,\mathbf{x})^2=f_{2A,2B,2C,D^2}^{(1)}(\omega,\mathbf{x}),\quad f_{\mathrm{GE}}^{(2)}(\omega,\mathbf{x})^2=f_{2A,2B,2C,D^2}^{(2)}(\omega,\mathbf{y}), \tag{26}$$

$$|f_{\mathrm{GE}}^{(1)}(\omega,\mathbf{x})|^2=f_{2\mathrm{Re}(A),2\mathrm{Re}(B),2\mathrm{Re}(C),|D|^2}^{(1)}(\omega,\mathbf{x}), \tag{27}$$

$$|f_{\mathrm{GE}}^{(2)}(\omega,\mathbf{y})|^2=f_{2\mathrm{Re}(A),2\mathrm{Re}(B),2\mathrm{Re}(C),|D|^2}^{(2)}(\omega,\mathbf{y}). \tag{28}$$

By substituting $A,B,C,D\to 2A,2B,2C,D^2$ into (9.2) (it's possible since $\mathrm{Re}\,(1-4(2A))>0$, we compute the first expectation in (25) as:

$$\begin{aligned}\mathbb{E}\left(f_{2A,2B,2C,D^2}^{(1)}(\omega,\mathbf{x})f_{2A,2B,2C,D^2}^{(2)}(\omega,\mathbf{y})\right)&=D^4\left(\sqrt{1-8A}\right)^{-d}\\&\times\exp\left(\frac{4B^2}{2(1-8A)}\|\mathbf{x}+s\mathbf{y}\|^2+2C(\|\mathbf{x}\|^2+\|\mathbf{y}\|^2)\right).\end{aligned}$$

Next, we express B, C, D through A, s using (5):

$$\mathbb{E}\left(f^{(1)}_{2A,2B,2C,D^2}(\omega,\mathbf{x})f^{(2)}_{2A,2B,2C,D^2}(\omega,\mathbf{y})\right) = \left(\sqrt{\frac{(1-4A)^2}{1-8A}}\right)^d$$

$$\times \exp\left(\frac{4s(1-4A)}{2(1-8A)}\|\mathbf{x}+s\mathbf{y}\|^2 - (s+1)(\|\mathbf{x}\|^2+\|\mathbf{y}\|^2)\right)$$

$$= \alpha_1 \exp\left(\alpha_2\|\mathbf{x}+s\mathbf{y}\|^2 - (s+1)\left(\|\mathbf{x}\|^2+\|\mathbf{y}\|^2\right)\right), \tag{29}$$

By substituting $A, B, C, D \rightarrow 2\mathrm{Re}\,(A)\,, 2\mathrm{Re}\,(B)\,, 2\mathrm{Re}\,(C)\,, |D|^2$ into (9.2) (it's possible since $\mathrm{Re}\,(1-8A) > 0$ and, hence, $\mathrm{Re}\,(1-8\mathrm{Re}\,(A)) > 0$), we can compute the second expectation in (25):

$$\mathbb{E}\left(f^{(1)}_{2\mathrm{Re}(A),2\mathrm{Re}(B),2\mathrm{Re}(C),|D|^2}(\omega,\mathbf{x})f^{(2)}_{2\mathrm{Re}(A),2\mathrm{Re}(B),2\mathrm{Re}(C),|D|^2}(\omega,\mathbf{y})\right) = |D|^4\left(\sqrt{1-8\mathrm{Re}\,(A)}\right)^{-d}$$

$$\cdot \exp\left(\frac{(B+\overline{B})^2}{2(1-8\mathrm{Re}\,(A))}\|\mathbf{x}+s\mathbf{y}\|^2 + 2\mathrm{Re}\,(C)\,(\|\mathbf{x}\|^2+\|\mathbf{y}\|^2)\right).$$

Next, we observe that $|D|^4 = D^2\overline{D^2}$, $(B+\overline{B})^2 = B^2+\overline{B^2}+2|B^2|$ and use (5) to express B, C, D through A and C:

$$\mathbb{E}\left(f^{(1)}_{2\mathrm{Re}(A),2\mathrm{Re}(B),2\mathrm{Re}(C),|D|^2}(\omega,\mathbf{x})f^{(2)}_{2\mathrm{Re}(A),2\mathrm{Re}(B),2\mathrm{Re}(C),|D|^2}(\omega,\mathbf{y})\right)$$

$$= \left(\frac{(1-4A)(\overline{1-4A})}{1-8\mathrm{Re}\,(A)}\right)^{d/2}$$

$$\times \exp\left(\frac{s(2-8\mathrm{Re}\,(A))+2|1-4A|}{2(1-8\mathrm{Re}\,(A))}\|\mathbf{x}+s\mathbf{y}\|^2 - (s+1)(\|\mathbf{x}\|^2+\|\mathbf{y}\|^2)\right)$$

$$= \alpha_3 \exp\left(\alpha_4\|\mathbf{x}+s\mathbf{y}\|^2 - (s+1)\left(\|\mathbf{x}\|^2+\|\mathbf{y}\|^2\right)\right) \tag{30}$$

where in the last transition we also take into account that $(1-4A)(\overline{1-4A}) = 1-8\mathrm{Re}\,(A)+16|A|^2$. (21, 22, 23, 24, 25, 26-28, 29, 30) taken together result in (6). ☐

9.4 Proof of Theorem 3.3

Proof. When A is real and $s = +1$, variance (6) has a form:

$$\mathrm{Var}_{p_{\mathrm{GE}}}\left(f^{(1)}_{\mathrm{GE}}(\omega,\mathbf{x})f^{(2)}_{\mathrm{GE}}(\omega,\mathbf{y})\right) = \left(\frac{1-4A}{\sqrt{1-8A}}\right)^d$$

$$\cdot \exp\left(\frac{2(1-4A)}{1-8A}\|\mathbf{x}+\mathbf{y}\|^2 - 2(\|\mathbf{x}\|^2+\|\mathbf{y}\|^2)\right) - K(\mathbf{x},\mathbf{y})^2$$

$$= 2^{-d}\left(\frac{\rho+1}{\sqrt{\rho}}\right)^d \exp\left((1+\rho)\,\|\mathbf{x}+\mathbf{y}\|^2 - 2(\|\mathbf{x}\|^2+\|\mathbf{y}\|^2)\right) - K(\mathbf{x},\mathbf{y})^2$$

where we change the variable $\rho = \frac{1}{1-8A} \in (0, +\infty)$. We see that the minimum of variance with respect to $\rho \in (0, +\infty)$ coincides with the minimum of the logarithm of the first term:

$$g(\rho) = -d\log 2 + d\log(\rho+1) - \frac{d}{2}\log\rho + (1+\rho)\|\mathbf{x}+\mathbf{y}\|^2 - 2(\|\mathbf{x}\|^2+\|\mathbf{y}\|^2).$$

All stationary points ρ^* can be found by setting its derivative to zero:

$$g'(\rho^*) = \frac{d}{\rho^*+1} - \frac{d}{2\rho^*} + \|\mathbf{x}+\mathbf{y}\|^2 = 0.$$

Multiply by $2\rho^*(\rho^*+1) > 0$ and obtain an equivalent quadratic equation:

$$d(\rho^*-1) + 2\rho^*(\rho^*+1)\|\mathbf{x}+\mathbf{y}\|^2 = 0;$$

$$2\|\mathbf{x}+\mathbf{y}\|^2(\rho^*)^2 + (2\|\mathbf{x}+\mathbf{y}\|^2+d)\rho^* - d = 0;$$

$$\rho^*_{1,2} = \frac{1}{4\|\mathbf{x}+\mathbf{y}\|^2}\left(\pm\sqrt{(2\|\mathbf{x}+\mathbf{y}\|^2+d)^2+8d\|\mathbf{x}+\mathbf{y}\|^2} - 2\|\mathbf{x}+\mathbf{y}\|^2 - d\right). \tag{31}$$

The root ρ_2^* of the quadratic equation with "$-$" sign in place of "$\pm$" (31) is a negative number. Since $\|\mathbf{x}+\mathbf{y}\|^2 > 0$, we conclude that the only stationary point is the positive root $\rho^* = \rho_1^* > 0$ with "+" sign in place of "$\pm$".

$g'(\rho)$ is a continuous function with $g'(\rho) \to -\infty$ as $\rho \to +0$ and $g'(\rho) \to \|\mathbf{x}+\mathbf{y}\|^2 > 0$ as $\rho \to +\infty$. There is only one ρ^* such that $g'(\rho^*) = 0$, and therefore for all $\rho < \rho^*$, $g'(\rho) < 0$ and for all $\rho > \rho^*$, $g'(\rho) < 0$. Hence, ρ^* is a global minimum of $g(\rho)$.

Since $g'(1) = \|\mathbf{x}+\mathbf{y}\|^2 > 0$, we also point out that $\rho^* < 1$. □

9.5 Proof of Theorem 3.4

Proof. First, we use $\mathrm{Var}Z = \mathbb{E}Z^2 - (\mathbb{E}Z)^2$ which holds for any random variable Z, e.g. $\mathrm{Re}\left(f_{\text{pois}}(\omega,\mathbf{x})f_{\text{pois}}(\omega,\mathbf{y})\right)$:

$$\mathrm{Var}_{p_{\text{pois}}}\left(f_{\text{pois}}(\omega,\mathbf{x})f_{\text{pois}}(\omega,\mathbf{y})\right) = \mathbb{E}\left(f_{\text{pois}}(\omega,\mathbf{x})^2 f_{\text{pois}}(\omega,\mathbf{y})^2\right) - \left(\mathbb{E}\left(f_{\text{pois}}(\omega,\mathbf{x})f_{\text{pois}}(\omega,\mathbf{y})\right)\right)^2. \tag{32}$$

We know that $\left(\mathbb{E}\left(f_{\text{pois}}(\omega,\mathbf{x})f_{\text{pois}}(\omega,\mathbf{y})\right)\right)^2 = K(\mathbf{x},\mathbf{y})^2$. Since $\omega_1,\dots,\omega_d$ are independent, $f_{\text{pois}}(\omega,\mathbf{x})^2 f_{\text{pois}}(\omega,\mathbf{y})^2$ can be decomposed into a product of d independent random variables:

$$f_{\text{pois}}(\omega,\mathbf{x})^2 f_{\text{pois}}(\omega,\mathbf{y})^2 = \exp(2\lambda d - \|\mathbf{x}\|^2 - \|\mathbf{y}\|^2)\prod_{l=1}^{d}(\mathbf{x}_l\mathbf{y}_l)^{2\omega_l}\lambda^{-2\omega_l}.$$

Its expectation is therefore a product of d independent expectations:

$$\mathbb{E}\left(f_{\text{pois}}(\omega,\mathbf{x})^2 f_{\text{pois}}(\omega,\mathbf{y})^2\right) = \exp(2\lambda d - \|\mathbf{x}\|^2 - \|\mathbf{y}\|^2)\prod_{l=1}^{d}\mathbb{E}(\mathbf{x}_l\mathbf{y}_l)^{2\omega_l}\lambda^{-2\omega_l}.$$

We compute each expectation in the product. First, we rewrite it as a sum:

$$\mathbb{E}(\mathbf{x}_l\mathbf{y}_l)^{2\omega_l}\lambda^{-2\omega_l} = \sum_{k=0}^{\infty}p_k(\mathbf{x}_l\mathbf{y}_l)^{2k}\lambda^{-2k} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}(\mathbf{x}_i\mathbf{y}_i)^{2k}\lambda^{-2k} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\mathbf{x}_l^2\mathbf{y}_l^2\lambda^{-1})^k}{k!}.$$

The last sum is a Taylor expansion of $\exp\left(\frac{\mathbf{x}_l^2\mathbf{y}_l^2}{\lambda}\right)$. So we have:

$$\mathbb{E}\left(f_{\text{pois}}(\omega,\mathbf{x})^2 f_{\text{pois}}(\omega,\mathbf{y})^2\right) = \exp\left(\lambda d + \lambda^{-1}\sum_{l=1}^{d}\mathbf{x}_l^2\mathbf{y}_l^2 - \|\mathbf{x}\|^2 - \|\mathbf{y}\|^2\right).$$

Taking it together with (32) results in (11). □

9.6 Proof of Theorem 3.5

Proof. The proof is similar to Theorem 3.4. First, we use $\mathrm{Var}Z = \mathbb{E}Z^2 - (\mathbb{E}Z)^2$ which holds for any random variable Z, e.g. $\mathrm{Re}\left(f_{\text{geom}}(\omega,\mathbf{x})f_{\text{geom}}(\omega,\mathbf{y})\right)$:

$$\mathrm{Var}_{p_{\text{geom}}}\left(f_{\text{geom}}(\omega,\mathbf{x})f_{\text{pois}}(\omega,\mathbf{y})\right) = \mathbb{E}\left(f_{\text{geom}}(\omega,\mathbf{x})^2 f_{\text{geom}}(\omega,\mathbf{y})^2\right) - \left(\mathbb{E}\left(f_{\text{geom}}(\omega,\mathbf{x})f_{\text{geom}}(\omega,\mathbf{y})\right)\right)^2. \tag{33}$$

We know that $\left(\mathbb{E}\left(f_{\text{pois}}(\omega,\mathbf{x})f_{\text{pois}}(\omega,\mathbf{y})\right)\right)^2 = K(\mathbf{x},\mathbf{y})^2$. Since $\omega_1,\dots,\omega_d$ are independent, $f_{\text{pois}}(\omega,\mathbf{x})^2 f_{\text{pois}}(\omega,\mathbf{y})^2$ can be decomposed into a product of d independent random variables:

$$f_{\text{geom}}(\omega,\mathbf{x})^2 f_{\text{geom}}(\omega,\mathbf{y})^2 = p^{-2d}\exp(-\|\mathbf{x}\|^2 - \|\mathbf{y}\|^2)\prod_{l=1}^{d}(\omega_l!)^{-2}((1-p)^{-1}\mathbf{x}_l\mathbf{y}_l)^{2\omega_l}.$$

Its expectation is therefore a product of d independent expectations:

$$\mathbb{E}\left(f_{\text{geom}}(\omega,\mathbf{x})^2 f_{\text{geom}}(\omega,\mathbf{y})^2\right) = p^{-2d}\exp(-\|\mathbf{x}\|^2 - \|\mathbf{y}\|^2)\prod_{l=1}^{d}\mathbb{E}(\omega_l!)^{-2}((1-p)^{-1}\mathbf{x}_l\mathbf{y}_l)^{2\omega_l}.$$

We compute each expectation in the product. First, we rewrite it as a sum:

$$\begin{aligned}\mathbb{E}(\omega_l!)^{-2}((1-p)^{-1}\mathbf{x}_l\mathbf{y}_l)^{2\omega_l} &= \sum_{k=0}^{\infty} p_k (k!)^{-2}((1-p)^{-1}\mathbf{x}_l\mathbf{y}_l)^{2k} \\ &= p\sum_{k=0}^{\infty}(k!)^{-2}((1-p)^{-1/2}\mathbf{x}_l\mathbf{y}_l)^{2k}.\end{aligned}$$

The last sum is a Taylor expansion of $I_0(2(1-p)^{-1/2}\mathbf{x}_l\mathbf{y}_l) = I_0(2(1-p)^{-1/2}|\mathbf{x}_l\mathbf{y}_l|)$ (I_0 is an even function). So we have:

$$\mathbb{E}\left(f_{\text{geom}}(\omega,\mathbf{x})^2 f_{\text{geom}}(\omega,\mathbf{y})^2\right) = p^{-d}\exp(-\|\mathbf{x}\|^2 - \|\mathbf{y}\|^2)\prod_{l=1}^{d} I_0(2(1-p)^{-\frac{1}{2}}|\mathbf{x}_l\mathbf{y}_l|).$$

Taking it together with (33) results in (13). □

We take absolute values $|\mathbf{x}_l\mathbf{y}_l|$ instead of just $\mathbf{x}_l\mathbf{y}_l$ because the average of $\mathbf{x}_l^{(i)}$ and $\mathbf{y}_l^{(j)}$ would converge to zero due to different signs and wouldn't produce any meaningful statistic.

9.7 Proof of Theorem 4.1

We prove here a much more general result from which Theorem 4.1 follows.

Theorem 9.1. *Consider a random variable X of the form: $X = g(\omega^\top \mathbf{z}, \|\omega\|)$ for some fixed $\mathbf{z} \in \mathbb{R}^d$ and: $g : \mathbb{R} \times \mathbb{R}_{\geq 0} \to \mathbb{R}$, where ω is sampled from the isotropic distribution $\Omega(d)$ with the corresponding distribution of $\|\omega\|$ denoted as $\tilde{\Omega}(d)$. Assume furthermore that for every $y \in \mathbb{R}_{\geq 0}$, function $g_y : \mathbb{R} \to \mathbb{R}$, defined as $g_y(x) = g(x, y)$, satisfies: $g_y(x) = \sum_{k=0}^{\infty} a_k(y)x^k$ for some $a_0(y), a_1(y), ... \geq 0$. Take two unbiased estimators of $K = \mathbb{E}[X]$, defined for $M \leq d$ as:*

$$\widehat{K}_M^{\text{iid}} = \frac{1}{M}\sum_{m=1}^{M} g((\omega_m^{\text{iid}})^\top \mathbf{z}, \|\omega_m\|),\ \widehat{K}_M^{\text{ort}} = \frac{1}{M}\sum_{m=1}^{M} g((\omega_m^{\text{ort}})^\top \mathbf{z}, \|\omega_m\|) \tag{34}$$

for $\omega_1^{\text{iid}}, ..., \omega_M^{\text{iid}} \overset{\text{iid}}{\sim} \Omega(d)$ and the orthogonal ensemble $\omega_1^{\text{ort}}, ..., \omega_M^{\text{ort}} \sim \Omega(d)$ (the orthogonal ensemble can be constructed since $\Omega(d)$ is isotropic). Then:

$$\text{Var}(\widehat{K}_M^{\text{ort}}) \leq \text{Var}(\widehat{K}_M^{\text{iid}}) - (1 - \frac{1}{M})\frac{2}{d+2}F^2(\mathbf{z}), \tag{35}$$

where $F(\mathbf{z}) \overset{\text{def}}{=} \mathbb{E}_{\mathbf{u}\sim\text{Unif}(0,\mathcal{S}^{d-1})}\mathbb{E}_{x\sim\tilde{\Omega}(d)}\left[\tilde{g}(x\mathbf{u}^\top\mathbf{z}, x)\right]$, $\text{Unif}(0, \mathcal{S}^{d-1})$ is the uniform probabilistic distribution on the $(d-1)$-dimensional unit sphere in $\mathbb{R}^d$ and $\tilde{g}(a,b) \overset{\text{def}}{=} \frac{g(a,b)+g(-a,b)}{2} - g(0,b)$.

If we define g as: $g(a,b) = D^2\exp(2Ab^2 + Ba + 2C(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2))$ for $A, B, C \in \mathbb{R}$ (see: Sec. 3.1), take $\Omega = \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$ and $\mathbf{z} = \mathbf{x} + \mathbf{y}$ then $\widehat{K}_M^{\text{iid}}$ and $\widehat{K}_M^{\text{ort}}$ from Theorem 9.1 become the estimators of the Gaussian kernel applying M generalized exponential random features that are either i.i.d or constructed from the orthogonal ensembles. Further, since $\Omega(d)$ is isotropic, $F^2(\mathbf{z})$ only depends on the norm of $\|\mathbf{z}\|$ and can be denoted $F^2(\mathbf{z}) = \mathcal{C}(\|\mathbf{z}\|)$. Therefore, as a corollary we obtain Theorem 4.1.

Proof. We start by factorizing the variance of K_M^{iid} and K_M^{ort} by conditioning on the lengths of the used random samples. We have:

$$\text{Var}(K_M^{\text{iid}}) = \int_{\mathbb{R}\times\ldots\times\mathbb{R}} \text{Var}\left(K^{\text{iid}_m} \mid \{\|\omega_1^{\text{iid}}\| = x_1, ..., \|\omega_m^{\text{iid}}\| = x_M\}\right)\prod_{m=1}^{M}\mathcal{P}(x_m)\cdot dx_1\cdot\ldots\cdot dx_M, \tag{36}$$

and similarly:

$$\text{Var}(K_M^{\text{ort}}) = \int_{\mathbb{R}\times\ldots\times\mathbb{R}} \text{Var}\left(K_M^{\text{ort}} \mid \{\|\omega_1^{\text{ort}}\| = x_1, ..., \|\omega_M^{\text{ort}}\| = x_M\}\right)\prod_{m=1}^{M}\mathcal{P}(x_m)\cdot dx_1\cdot\ldots\cdot dx_M, \tag{37}$$

where $\mathcal{P}$ is the pdf function for the distribution $\tilde{\Omega}(d)$ of the lengths of samples taken from $\Omega(d)$. We use the fact that in both scenarios: iid samples and an orthogonal ensemble, the lengths of vectors ω_i are sampled from the same distribution $\tilde{\Omega}$, independently from their directions and from each other. Therefore we have:

$$\mathrm{Var}(K_M^{\mathrm{iid}}) - \mathrm{Var}(K_M^{\mathrm{ort}}) = \int_{\mathbb{R}\times\ldots\times\mathbb{R}} T(x_1,...,x_M) \prod_{m=1}^{M} \mathcal{P}(x_m)\cdot dx_1 \cdot \ldots \cdot dx_M, \tag{38}$$

where

$$\begin{aligned} T(x_1,...,x_M) = \mathrm{Var}\left(K_M^{\mathrm{iid}} \mid \{\|\omega_1^{\mathrm{iid}}\| = x_1,...,\|\omega_M^{\mathrm{iid}}\| = x_M\}\right) - \\ \mathrm{Var}\left(K_M^{\mathrm{ort}} \mid \{\|\omega_1^{\mathrm{ort}}\| = x_1,...,\|\omega_M^{\mathrm{ort}}\| = x_M\}\right) \end{aligned} \tag{39}$$

Since the lengths of the samples are chosen independently from their directions, we conclude that:

$$\mathrm{Var}\left(K_m^{\mathrm{iid}} \mid \{\|\omega_1^{\mathrm{iid}}\| = x_1,...,\|\omega_m^{\mathrm{iid}}\| = x_M\}\right) = \mathrm{Var}\left(\frac{1}{M}\sum_{m=1}^{M} X_m^{\mathrm{iid}}\right) \tag{40}$$

and

$$\mathrm{Var}\left(K_M^{\mathrm{ort}} \mid \{\|\omega_1^{\mathrm{ort}}\| = x_1,...,\|\omega_M^{\mathrm{ort}}\| = x_m\}\right) = \mathrm{Var}\left(\frac{1}{M}\sum_{m=1}^{M} X_m^{\mathrm{ort}}\right), \tag{41}$$

where $X_m^{\mathrm{iid}} = g_{x_m}((\mathbf{u}_m^{\mathrm{iid}})^\top \mathbf{z})$ and $X_m^{\mathrm{ort}} = g_{x_m}((\mathbf{u}_m^{\mathrm{ort}})^\top \mathbf{z})$, $\{\mathbf{u}_1^{\mathrm{iid}},...,\mathbf{u}_M^{\mathrm{iid}}\}$ are iid samples from the unit-sphere in $\mathbb{R}^d$ and $\{\mathbf{u}_1^{\mathrm{ort}},...,\mathbf{u}_M^{\mathrm{ort}}\}$ is an orthogonal ensemble of samples taken from the unit sphere in $\mathbb{R}^d$.

Thus we have:

$$T(x_1,...,x_M) = \mathrm{Var}(\frac{1}{M}\sum_{m=1}^{M} X_m^{\mathrm{iid}}) - \mathrm{Var}(\frac{1}{M}\sum_{m=1}^{M} X_m^{\mathrm{ort}}) \tag{42}$$

By the similar analysis as in the proof of Theorem 5 in [15], we obtain for $(g_1,...,g_d) \sim \mathcal{N}(0, \mathbf{I}_d)$:

$$\begin{aligned} T(x_1,...,x_M) \geq \frac{2}{(d+2)}\cdot\frac{2}{M^2}\sum_{1\leq i<j\leq M}\sum_{t,u=1}^{\infty} a_{2t}(x_i)a_{2u}(x_j)\|\mathbf{z}\|^{2t+2u}\mathbb{E}[\|\omega\|^{2t}]\mathbb{E}[\|\omega\|^{2u}]\cdot \\ \frac{\mathbb{E}[g_1^{2t}]\mathbb{E}[g_2^{2u}]}{\mathbb{E}[\sqrt{g_1^2+...+g_d^2}^{2t}]\mathbb{E}[\sqrt{g_1^2+...+g_d^2}^{2u}]} = \frac{2}{d+2}\cdot\frac{2}{M^2}\sum_{1\leq i<j\leq M} \\ \left(\sum_{t=1}^{\infty} a_{2t}(x_i)\|\mathbf{z}\|^{2t}\cdot\frac{\mathbb{E}[\|\omega\|^{2t}]\cdot\mathbb{E}[g_1^{2t}]}{\mathbb{E}[\sqrt{g_1^2+...+g_d^2}^{2t}]}\right)\cdot\left(\sum_{t=1}^{\infty} a_{2t}(x_j)\|\mathbf{z}\|^{2t}\cdot\frac{\mathbb{E}[\|\omega\|^{2t}]\cdot\mathbb{E}[g_2^{2t}]}{\mathbb{E}[\sqrt{g_1^2+...+g_d^2}^{2t}]}\right) \\ = \frac{2}{d+2}\cdot\frac{2}{M^2}\sum_{1\leq i<j\leq M} F_{x_i}(\mathbf{z})F_{x_j}(\mathbf{z}), \end{aligned} \tag{43}$$

where $F_x(\mathbf{z}) \overset{\mathrm{def}}{=} \mathbb{E}[\tilde{g}(x\mathbf{u}^\top\mathbf{z}, x)]$, $\tilde{g}(a,b) \overset{\mathrm{def}}{=} \frac{g(a,b)+g(-a,b)}{2} - g(0,b)$ and $\mathbf{u} \sim \mathrm{Unif}(\mathcal{S}^{d-1})$ is taken uniformly at random from the unit $(d-1)$-dimensional sphere in $\mathbb{R}^d$.

We conclude that:

$$\begin{aligned} \mathrm{Var}(K_M^{\mathrm{iid}}) - \mathrm{Var}(K_M^{\mathrm{ord}}) &= \frac{4}{M^2(d+2)}\int_{\mathbb{R}\times\ldots\times\mathbb{R}}\sum_{1\leq i<j\leq M} F_{x_i}(\mathbf{z})F_{x_j}(\mathbf{z})\prod_{i=1}^{M}\mathcal{P}(x_i)\cdot dx_1\cdot\ldots\cdot dx_M \\ &= \frac{4}{M^2(d+2)}\binom{M}{2}\int_{\mathbb{R}\times\mathbb{R}} F_x(\mathbf{z})F_y(\mathbf{z})\mathcal{P}(x)\mathcal{P}(y)dxdy = (1-\frac{1}{M})\frac{2}{d+2}F^2(\mathbf{z}), \end{aligned} \tag{44}$$

where $F(\mathbf{z}) = \mathbb{E}_{\mathbf{u}\sim\mathrm{Unif}(0,\mathcal{S}^{d-1})}\mathbb{E}_{x\sim\tilde{\Omega}(d)}\left[\tilde{g}(\mathbf{u}^\top\mathbf{z}, x)\right]$. That completes the proof. □

9.8 Proof of Theorem 4.2

Proof. To prove the first part of the theorem, we use the following Hoeffding's inequality:

Lemma 9.1 (Hoeffding's Inequality)**.** *Let $X_1, ..., X_M$ be M independent random variables (not necessarily identically distributed) with zero mean. Assume furthermore that: $-a_i \leq X_i \leq b_i$ for $a_i, b_i \geq 0$ for $i = 1, ..., M$. Then the following is true for any $a > 0$:*

$$\mathbb{P}[|\sum_{i=1}^{M} X_i| > a] \leq 2 \cdot \exp\left(-\frac{a^2}{\sum_{i=1}^{N}(a_i + b_i)^2}\right) \tag{45}$$

Note first that we have:

$$\begin{aligned} 0 \leq Z = \exp\left(-\left\|\sqrt{-A}\omega - \frac{B}{2\sqrt{-A}}\mathbf{x}\right\|^2 - \frac{B^2}{4A}\|\mathbf{x}\|^2 + C\|\mathbf{x}\|^2\right) \cdot \\ \exp\left(-\left\|\sqrt{-A}\omega - \frac{B}{2\sqrt{-A}}\mathbf{y}\right\|^2 - \frac{B^2}{4A}\|\mathbf{y}\|_2^2 + C\|\mathbf{y}\|_2^2\right) \leq \exp\left(-\frac{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2}{4A}\right), \end{aligned} \tag{46}$$

where the last inequality follows from taking: $B = \sqrt{1 - 4A}$, $C = -1$.

Denote: $\mathcal{M}(\mathbf{x}, \mathbf{y}) = \exp(-\frac{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2}{4A})(1 - 4A)^{\frac{d}{2}}$. Define: $Y = Z - \mathbb{E}[Z]$. Note that: $\mathbb{E}[Y] = 0$. Furthermore, from Inequality 46, we get: $0 - K(\mathbf{x}, \mathbf{y}) \leq Y \leq \mathcal{M}(\mathbf{x}, \mathbf{y}) - K(\mathbf{x}, \mathbf{y})$. Thus we have: $-a \leq Y \leq b$ for $a = K(\mathbf{x}, \mathbf{y})$, $b = \mathcal{M}(\mathbf{x}, \mathbf{y}) - K(\mathbf{x}, \mathbf{y})$. The following is true:

$$\mathbb{P}[|\widehat{K}^{\mathrm{iid}}_M(\mathbf{x}, \mathbf{y}) - K(\mathbf{x}, \mathbf{y})| \geq \epsilon] = \mathbb{P}\left[\frac{Y_1 + ... + Y_M}{M} \geq \epsilon\right] = \mathbb{P}[|Y_1 + ... + Y_M| \geq M\epsilon], \tag{47}$$

where $Y_1, \ldots, Y_M$ are independent copies of Y. We complete the proof of the first part of the theorem by applying Hoeffding's Inequality for: $X_i = Y_i$, $a_i = a$, $b_i = b$ ($i = 1, ..., M$) and $a = M\epsilon$.

The second part of the theorem follows directly from the exact same method as applied in the proof of Theorem 4.1 (e.g. we condition on the lengths of the sampled vectors ω_i), combined again with the analysis from Theorem 5 in [15], but this time for higher moments. Note that critically, Legendre Transform is well-defined since the corresponding random variables are bounded. The nonnegativity of the Legendre Transform for the inputs from statement of the theorem follows from the standard properties of the Legendre Transform for the inputs $x > \mathbb{E}X$, where X is the corresponding random variable. □

9.9 Proof of Theorem 4.3

Proof. The proof is similar to the proof of Claim 1 from [45]. Note that in the regular attention mechanism, queries and keys are renormalized by the multiplicative factor: $\frac{1}{d^{\frac{1}{4}}}$. Thus denote: $\mathbf{x} = \frac{\mathbf{q}}{d^{\frac{1}{4}}}$ and $\mathbf{y} = \frac{\mathbf{k}}{d^{\frac{1}{4}}}$. Note that: $\|\mathbf{x}\|, \|\mathbf{y}\| \leq \frac{R}{d^{\frac{1}{4}}}$. Consider vector $\mathbf{z} = [\mathbf{x}^\top, \mathbf{y}^\top]^\top \in \mathbb{R}^{2d}$. Note that: $\|\mathbf{z}\|_2 \leq \sqrt{2}\frac{R}{d^{\frac{1}{4}}}$. By the analogous analysis as in Claim 1, we cover the ball $B(0, \sqrt{2}\frac{R}{d^{\frac{1}{4}}}) \subseteq \mathbb{R}^{2d}$ with the ϵ-net of at most $T = (\frac{4\rho}{r})^{2d}$ balls of radius r for $\rho = \sqrt{2}\frac{R}{d^{\frac{1}{4}}}$. If L_f denotes the Lipschitz constant of f, the straightforward calculations lead to:

$$\mathbb{E}[L_f^2] \leq \max_{\mathbf{x},\mathbf{y}} \widehat{\mathcal{M}}^2(\mathbf{x}, \mathbf{y}) \max_{\mathbf{x},\mathbf{y}} \left(2\|\mathbf{x}\|^2 + 2\|\mathbf{y}\|^2 + 4\mathbb{E}[\|\omega\|_2^2]\right), \tag{48}$$

where $\widehat{\mathcal{M}}(\mathbf{x}, \mathbf{y}) = \exp(-\frac{\|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2}{2})\mathcal{M}(\mathbf{x}, \mathbf{y})$, $\mathcal{M}(\mathbf{x}, \mathbf{y})$ is defined as in the proof above and $\omega \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$ (the extra multiplicative term next before $\mathcal{M}(\mathbf{x}, \mathbf{y})$ is needed since now we work with the softmax-kernel which is the rescaled variant of the Gaussian kernel, see: discussion in the paper). Thus we have: $\mathbb{E}[L_f^2] \leq \gamma^2$, where: $\gamma = 2(1 - 4A)^{\frac{d}{2}}\sqrt{\exp(-\frac{3R^2}{A\sqrt{d}})(\frac{R^2}{\sqrt{d}} + d^2)}$. Using Theorem 4.2, we also notice that we can get analogous inequality as Inequality (6) from the proof of Claim 1 in [45], but for: $D = 4M(1 - 4A)^{-\frac{d}{2}} \max_{\mathbf{x},\mathbf{y}} \exp(\frac{3(\|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2)}{2A}) = 4M(1 - 4A)^{-\frac{d}{2}} \exp(\frac{3R^2}{A\sqrt{d}})$. Thus substituting: (a) σ_p with γ, (b) D with $4M\exp(\frac{3R^2}{A\sqrt{d}})$, (c) d with $2d$ and (d) $\mathrm{diam}(\mathcal{M})$ with ρ in the statement of Claim 1, we obtain Theorem 4.3.

□

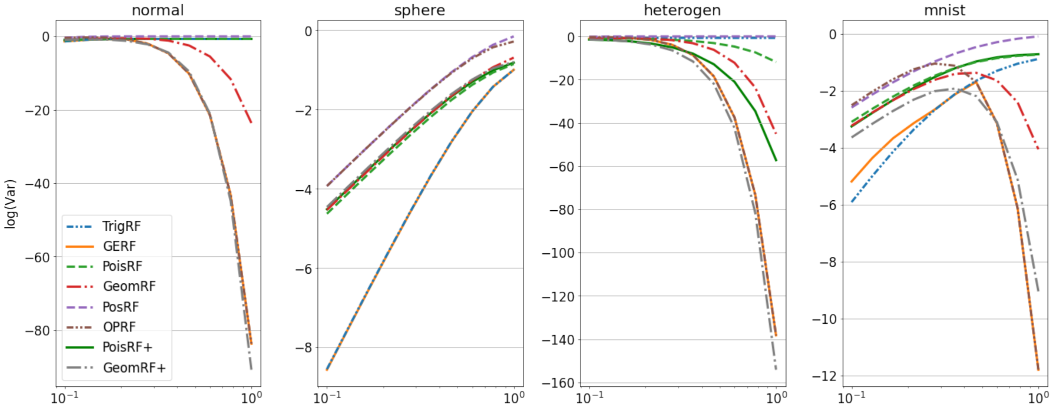

Figure 5: Version of Figure 2 with positive-valued and arbitrary-valued RFs on one plot.

9.10 Additional experimental details

We use NumPy [28] and the free version of Google Colaboratory for running the first two experiments. For the Transformer experiments, we use a TPU cluster and JAX [6] implementation.

9.10.1 Comparing variance of different RFs

We use Brent method [7] with 100 iterations for minimization of p in GeomRF(+) and two L-BFGS-B [65] routines of 50 iterations to minimize A in GERF for $s = -1$ and $+1$ respectively. We reuse these configurations in the non-parametric classification experiment.

We sample pairs of sets $\{\mathbf{x}^{(i)}\}_{1 \leq i \leq L}$, $\{\mathbf{y}^{(j)}\}_{1 \leq j \leq L}$, where $L = 1024$, 5 times. On each pair of sets, we compute the variance of approximating $K(\mathbf{x}^{(i)}, \mathbf{y}^{(j)})$ for all pairs of $\mathbf{x}^{(i)}$ and $\mathbf{y}^{(j)}$. Also, on each pair of sets of $\{\mathbf{x}^{(i)}\}_{1 \leq i \leq L}$, $\{\mathbf{y}^{(j)}\}_{1 \leq j \leq L}$, we compute statistics (8,12,14) and then use them to optimize parameters of the corresponding method. The means and standard deviations are reported for averaging over all pairs of $\mathbf{x}^{(i)}$ and $\mathbf{y}^{(j)}$, over all 5 samples.

For a fair comparison, for real-valued RF mechanisms, we compute the variance assuming that $M = 2$ (the variance is divided by 2), since complex RF mechanisms effectively use real and imaginary part of the number.

Figure 5 is where we put both positive-valued and arbitrary-valued RFs on one plot for comparison. We observe that GERFs and OPRFs have the same log-variance which is explainable since GERFs extend OPRFs.

9.10.2 Non-parametric classification

Experimental details. We randomly split the raw dataset into 90% which is used for training, 5% for tuning σ and 5% for testing. These splits are fixed for all compared methods. σ is tuned on a log-uniform grid of 10 values from 10^{-2} to 10^{2}. For each σ and each method, we average accuracy for 50 seeds used to draw RFs both during validation and testing (for the best σ only). As for the previous experiment, we use M for real-valued RFs and $M/2$ for complex-valued for a fair comparison. We use orthogonal ω's for all GERF-descendant methods.

We use $\epsilon = 10^{-8}$ when making input features positive in PoisRF+ and GeomRF+. $\mathbf{c}$ is inferred from the train set, and we clamp validation/test input features to be at least ϵ to guarantee that they are positive without leaking test data into $\mathbf{c}$.

Numerical optimization of parameters in GERF, GeomRF(+) is performed in the same way as described in Appendix 9.10.1.

Standard deviations. Table 3 reports standard deviations of the test accuracies reported in the main text.

Ablation over M. Table 4 is an ablation over the number of (real-valued) random features M. We see that OPRF consistently outperforms the baselines (TrigRF, PosRF) and also outperforms or is competitive with other methods proposed in the paper. Further, OPRF shows the best performance among positive-valued random features (PosRF, OPRF, PoisRF+, GeomRF+) in all settings. As for the choice of M, we see that performance increases as M grows which is expected. Hence, in practice, a good strategy is to select M as big as the compute budget permits which would alleviate an expensive grid search over M.

Comparison of variances. Table 5 shows average log-variances computed using analytic formulas (8,12,14) for pairs of $\mathbf{x}, \mathbf{y}$ where $\mathbf{x}$ comes from the train set and $\mathbf{y}$ from the test set. We observe that OPRF has the smallest variance with the same value only for GERF (which can be explained since GERF extends OPRF but it's complex-valued, hence we need to use 2 times less features for the comparable amount of computation). The smallest variance for OPRF can be explained by the input data distribution due to which OPRF's variance is smaller in average than other variances.

Effect of orthogonal RFs. Table 6 compares using non-orthogonal and orthogonal variants of TrigRF, PosRF, GERF and OPRF. We observe that the orthogonal variant either doesn't harm, or improves the result in most cases. Further, two positive-valued random features (PosRF and OPRF) benefit from orthogonality when averaged over all benchmarks.

Kernel ridge regression. We run an additional experiment with kernel ridge regression (KRR) [37] instead of kernel regression for predicting logits $\mathbf{r}^*$. In KRR, $\mathbf{r}^*$ is predicted as the result of linear ridge regression on

$$\widehat{\mathbf{x}}^{(i)} = (f^{(1)}(\omega^{(1)}, \mathbf{o}^{(i)}), \dots, f^{(1)}(\omega^{(M)}, \mathbf{o}^{(i)})) \in \mathbb{R}^M$$

if RFs are real-valued and

$$\begin{aligned}\widehat{\mathbf{x}}^{(i)} = (&\mathrm{Re}\left(f^{(1)}(\omega^{(1)}, \mathbf{o}^{(i)})\right), \mathrm{Im}\left(f^{(1)}(\omega^{(1)}, \mathbf{o}^{(i)})\right), \dots, \\ &\mathrm{Re}\left(f^{(1)}(\omega^{(M/2)}, \mathbf{o}^{(i)})\right), \mathrm{Im}\left(f^{(1)}(\omega^{(M/2)}, \mathbf{o}^{(i)})\right)) \in \mathbb{R}^M\end{aligned}$$

if RFs are complex-valued. Similarly, $\widehat{\mathbf{y}}^*$ is defined as

$$\widehat{\mathbf{y}}^* = (f^{(1)}(\omega^{(1)}, \mathbf{o}^*), \dots, f^{(1)}(\omega^{(M)}, \mathbf{o}^*)) \in \mathbb{R}^M$$

if RFs are real-valued and

$$\begin{aligned}\widehat{\mathbf{y}}^* = (&\mathrm{Re}\left(f^{(1)}(\omega^{(1)}, \mathbf{o}^*)\right), -\mathrm{Im}\left(f^{(1)}(\omega^{(1)}, \mathbf{o}^*)\right), \dots, \\ &\mathrm{Re}\left(f^{(1)}(\omega^{(M/2)}, \mathbf{o}^*)\right), -\mathrm{Im}\left(f^{(1)}(\omega^{(M/2)}, \mathbf{o}^*)\right)) \in \mathbb{R}^M\end{aligned}$$

if RFs are complex-valued. Next, $\mathbf{r}^*$ is predicted as

$$(\mathbf{r}^*)^\top = (\widehat{\mathbf{y}}^*)^\top \left(\phi \mathbf{I}_M + \sum_{i=1}^{L} \widehat{\mathbf{x}}^{(i)} (\widehat{\mathbf{x}}^{(i)})^\top\right)^{-1} \sum_{i=1}^{L} \widehat{\mathbf{x}}^{(i)} (\mathbf{r}^{(i)})^\top$$

where $\phi > 0$ is a hyperparameter.

We tune ϕ on a log-uniform grid of 10 values from 10^{-2} to 10^2. For each ϕ and each method, we average accuracy for 50 seeds used to draw RFs both during validation and testing (for the best ϕ only). In the rest, the setup is the same as with kernel classification. Result accuracies and their standard deviations are reported in Tables 7 and 8 respectively. We observe that, again, OPRF shows the best average performance which is also slightly better than for the kernel regression model (58.1 against 57.8, Table 1).

9.10.3 Text

We pretrained on two publicly available datasets (see: Table 10). Following the original BERT training, we mask 15% of tokens in these two datasets, and train to predict the mask. We used the exact same hyperparameter-setup for all the baselines (FAVOR+ [15], ELU [31], ReLU [15]) and FAVOR++. The hyperparameters for pretraining are shown in Table 9. We finetuned on GLUE task, warm-starting with the weights of the pretrained model. The setup is analogous to the one from the original BERT paper.

Table 3: Non-parametric classification, standard deviations.

Dataset	TrigRF	PosRF	GERF	PoisRF	GeomRF	OPRF	PoisRF+	GeomRF+
`abalone`	< 0.05	2.1	1.9	1.8	1.3	1.7	2.9	2.9
`banknote`	< 0.05	3.7	4.3	2.1	3.0	3.4	5.9	7.7
`car`	< 0.05	3.0	2.5	0.0	< 0.05	3.0	< 0.05	1.5
`yeast`	< 0.05	3.2	5.0	6.0	3.4	4.9	< 0.05	2.4
`cmc`	< 0.05	4.0	3.9	4.3	3.4	3.8	5.3	5.2
`nursery`	< 0.05	6.3	3.2	7.2	7.3	6.3	5.6	8.2
`wifi`	< 0.05	6.2	4.1	2.8	2.0	4.1	13.1	9.8
`chess`	< 0.05	1.3	1.2	1.8	2.0	1.2	1.7	1.9

Table 4: Non-parametric classification, ablation over M, average accuracy over all tasks.

M	TrigRF	PosRF	GERF	PoisRF	GeomRF	OPRF	PoisRF+	GeomRF+
16	35.5	46.9	47.4	$\underline{49.0}$	**49.1**	48.5	41.3	43.2
32	35.5	50.5	51.2	$\underline{52.2}$	**52.6**	51.8	44.1	46.5
64	35.5	51.3	54.0	54.2	$\underline{55.2}$	**55.4**	47.0	50.0
128	35.5	54.3	56.1	55.5	$\underline{56.8}$	**57.8**	49.9	52.3
256	35.5	55.6	$\underline{58.1}$	56.5	57.8	**59.7**	51.9	55.0

9.10.4 Speech

Our applied Conformer-Transducer models consisted of $l = 17$ conformer layers. Each attention layer used $h = 4$ heads. The embedding dimensionality was $p = 256$. Dimensions were split equally among heads, leading to $d_{QK} = 64$ dimensions per query/key. Input sequences were of length $L \sim 500$. We applied padding mechanism for all tested variants. The model provides transcribed speech (see also: Table: 11).

9.10.5 Vision

The vision experiments follow Section 4 in the MAE paper, where we use a ViT-Large (Table: 15) and the same setup for training from scratch (Table: 14) and fine-tuning (Table: 13) as for the MAE baseline trained with regular softmax attention (Table: 12). Note that the fine-tuning setup has a shorter schedule which tests the adaptability of low-rank attention variants to the regular softmax attention.

The ablations over sequence lengths are conducted by training from scratch and use ViT-tiny model (Table: 16). Different sequence lengths are derived by adjusting the input size and the patch size which results in different number of patches (Table: 17). Different patch sizes require different sizes of projection layers before converting to tokens with latent representations of the same dimesionality.

9.11 Long Range Arena

In Long range arena, [54] propose a a diverse set of datasets for evaluating the performance of efficient transformer on long sequence tasks. In Table 18 we report the performance of FAVOR++ on three very diverse datasets (ListOps, Text Retrieval, Image Classification). We follow the exact same setup as [54]. We find very similar trend as our previous experiments i.e. FAVOR++ almost always improves the performance of performer showing the significance of designing the kernel for variance reduction.

Table 5: Non-parametric classification, average log-variance over all tasks.

TrigRF	PosRF	GERF	PoisRF	GeomRF	OPRF	PoisRF+	GeomRF+
−0.8	−0.0	**−20.5**	−0.7	−7.0	**−20.5**	36.2	−16.7

Table 6: Non-parametric classification, accuracy with non-orthogonal RFs / orthogonal RFs.

Dataset	TrigRF	PosRF	GERF	OPRF
`abalone`	12.0 / 12.0	15.5 / **16.0**	**17.7** / 17.0	16.7 / **17.1**
`banknote`	66.2 / 66.2	**83.9** / 83.4	93.2 / **92.4**	92.3 / **92.6**
`car`	66.3 / 66.3	68.9 / **69.2**	70.5 / **70.9**	**69.9** / 69.5
`yeast`	29.7 / 29.7	**34.6** / 34.4	42.8 / **42.9**	42.1 / **44.4**
`cmc`	46.6 / 46.6	44.7 / **45.1**	47.4 / **47.8**	**47.3** / 46.3
`nursery`	31.3 / 31.3	73.2 / **77.4**	63.8 / 63.8	75.8 / **78.9**
`wifi`	15.2 / 15.2	84.6 / **88.8**	93.0 / **93.3**	92.1 / **93.3**
`chess`	16.5 / 16.5	19.6 / **20.2**	20.4 / 20.4	**20.4** / 20.2
Average	35.5 / 35.5	53.1 / **54.3**	56.1 / 56.1	57.1 / **57.8**

Table 7: Kernel ridge regression, accuracy.

Dataset	TrigRF	PosRF	GERF	PoisRF	GeomRF	OPRF	PoisRF+	GeomRF+	L
`abalone`	10.1	21.4	21.9	23.2	**25.1**	21.8	16.3	13.6	3758
`banknote`	38.7	99.1	99.8	**100.0**	**100.0**	99.7	90.8	99.2	1233
`car`	36.1	70.7	**70.8**	34.0	39.0	70.5	62.7	67.4	1554
`yeast`	15.9	49.2	51.8	39.4	52.2	**52.6**	5.6	20.8	1334
`cmc`	34.1	46.6	47.3	44.9	**49.4**	47.8	37.2	40.8	1324
`nursery`	27.5	**58.5**	57.4	30.0	31.0	57.7	41.0	46.6	11664
`wifi`	13.7	97.2	**98.1**	92.7	97.1	97.3	36.1	81.6	1799
`chess`	11.0	17.2	16.9	12.5	13.7	**17.4**	12.7	16.3	25249
Average	23.4	57.5	58.0	47.1	50.9	**58.1**	37.8	48.3	N/A

Table 8: Kernel ridge regression, standard deviation.

Dataset	TrigRF	PosRF	GERF	PoisRF	GeomRF	OPRF	PoisRF+	GeomRF+
`abalone`	3.1	1.8	1.8	2.1	1.8	1.6	3.2	3.4
`banknote`	4.9	1.1	0.6	0.2	0.2	0.7	4.0	1.2
`car`	5.3	3.3	4.4	5.8	6.2	4.2	8.1	3.5
`yeast`	3.7	4.0	3.6	6.9	3.9	3.6	10.2	9.6
`cmc`	5.5	3.3	4.2	4.6	4.2	3.6	6.5	4.9
`nursery`	2.0	3.4	3.1	2.7	3.0	2.8	5.4	6.7
`wifi`	5.2	1.9	1.1	2.9	1.4	1.3	4.0	7.6
`chess`	0.8	1.7	1.3	1.6	1.2	1.4	3.8	2.6

Table 9: Hyperparameters for the base models for pre-training for the baselines (FAVOR+ [15], ELU [31] and ReLU [15]) and FAVOR++.

Parameter	Value
# of heads	12
# of hidden layers	12
Hidden layer size	768
# of tokens	512
Batch size	256
M	256
Pretrain Steps	$1M$
Loss	MLM
Activation layer	gelu
Dropout prob	0.1
Attention dropout prob	0.1
Optimizer	Adam
Learning rate	10^{-4}
Compute resources	8 × 8 TPUv3

Table 10: Dataset used for pre training.

Dataset	# tokens	Avg. doc len.
Books [66]	1.0B	37K
Wikipedia	3.1B	592

Table 11: Hyperparameters for trained Speech models.

Parameter	Value
# of heads	4
# of hidden layers	17
Hidden layer size	256
# of tokens	512
Batch size	256
Activation layer	gelu
Dropout prob	0.1
Optimizer	Adam
Learning rate	10^{-4}
Compute resources	8 × 8 TPUv3

Table 12: Hyperparameters for Vision pre-training setting.

Parameter	Value
Batch size	4096
Optimizer	AdamW
Base Learning rate	$1.5e^{-4}$
Weight decay	0.05
Optimizer momentum	β_1, β_2 = 0.9, 0.95
Learning rate schedule	cosine decay
Warm up epochs	40
Augmentation	RandomResizedCrop
Compute resources	8 × 8 TPUv3

Table 13: Hyperparameters for Vision End-to-End fine-tuning setting.

Parameter	Value
Batch size	1024
Optimizer	AdamW
Base Learning rate	$1e^{-3}$
Layer-wise lr decay	0.75
Weight decay	0.05
Optimizer momentum	β_1, β_2 = 0.9, 0.999
Learning rate schedule	cosine decay
Warm up epochs	5
Training epochs	50
Augmentation	RandomAug (9, 0.5)
Label smoothing	0.1
Mixup	0.8
CutMix	1.0
Droppath	0.1
Compute resources	8 × 8 TPUv3

Table 14: Hyperparameters for Vision - training from scratch setting.

Parameter	Value
Batch size	4096
Optimizer	AdamW
Base Learning rate	$1e^{-4}$
Layer-wise lr decay	0.75
Weight decay	0.3
Optimizer momentum	β_1, β_2 = 0.9, 0.999
Learning rate schedule	cosine decay
Warm up epochs	20
Training epochs	200
Augmentation	RandomAug (9, 0.5)
Label smoothing	0.1
Mixup	0.8
CutMix	1.0
Droppath	0.2
Exp moving avg	0.9999
Compute resources	8 × 8 TPUv3

Table 15: Hyperparameters for Vision model - ViT Large.

Parameter	Value
# of heads	16
# of layers	24
Hidden layer size	1024

Table 16: Hyperparameters for Vision model - ViT tiny.

Parameter	Value
# of heads	3
# of layers	12
Hidden layer size	192

Table 17: ViT sequence length (# patches) and image input mapping.

Patches	Image input size
8x8	224
16x16	224
32x32	224
40x40	240
44x44	220

Table 18: Experimental results on the *LRA* benchmark. The best model is in boldface and the second best is underlined. Accuracy scores for all baseline models are from [54] (Table 6 in Appendix E.1). Here, L refers to the sequence length, K refers to the size of a local window and $B \ll L$ is a model specific parameter.

Model	**Complexity**	**ListOps** 2K	**Retrieval** 4K	**Image** 1K
Softmax Transformer [55]	$O(L^2)$	36.38	57.46	42.44
Synthesizer [52]	$O(L^2)$	36.50	54.67	41.61
Sinkhorn [53]	$O((L/B)^2)$	34.20	53.83	41.23
Sparse Transformer [10]	$O(L\sqrt{L})$	35.78	$\underline{59.59}$	**44.24**
Reformer [32]	$O(L \log L)$	36.30	53.40	38.07
Local Attention [43]	$O(LK)$	15.95	53.39	41.46
Longformer [2]	$O(LK)$	36.03	56.89	42.22
Linformer [58]	$O(L)$	35.49	52.27	38.56
BigBird [64]	$O(LK)$	$\underline{37.08}$	59.29	40.83
LinearElu [31]	$O(L)$	17.15	53.09	42.34
Performer [15]	$O(L)$	36.00	53.82	$\underline{42.77}$
FAVOR++	$O(L)$	**42.65**	**60.40**	39.47