KAMA-NNs: low-dimensional rotation based neural networks: Supplementary Material

A Proof of Theorem 1 and Theorem 2

Consider the Kernel estimators $\mathrm{URN}_f(\mathbf{x},\mathbf{y})$ and $\mathrm{ORN}_f(\mathbf{x},\mathbf{y})$, corresponding to the unstructured-Gaussian and orthogonal-Gaussian pointwise non-linear kernels, constructed from m random features. The two estimators are unbiased, i.e. $\mathbb{E}[\mathrm{URN}_f(\mathbf{x},\mathbf{y})] = \mathbb{E}[\mathrm{ORN}_f(\mathbf{x},\mathbf{y})] = K_f(\mathbf{x},\mathbf{y})$, while the mean-squared errors differ and equal,

$$\begin{aligned}
\mathrm{MSE}[\mathrm{URN}_f(\mathbf{x},\mathbf{y})] &= \mathbb{E}\left[\mathrm{URN}_f(\mathbf{x},\mathbf{y})\right]^2 - K_f(\mathbf{x},\mathbf{y})^2 \\
&= \mathbb{E}_{\mathbf{w}\sim\mathbf{G}}\left[\frac{1}{m}\sum_{i=1}^{m} f(\mathbf{w}_i^\top\mathbf{x})f(\mathbf{w}_i^\top\mathbf{y})\right]^2 - K_f(\mathbf{x},\mathbf{y})^2 \\
&= \mathbb{E}_{\mathbf{w}\sim\mathbf{G}}\left[\frac{1}{m^2}\sum_{i,j=1}^{m} f(\mathbf{w}_i^\top\mathbf{x})f(\mathbf{w}_i^\top\mathbf{y})f(\mathbf{w}_j^\top\mathbf{x})f(\mathbf{w}_j^\top\mathbf{y})\right] - K_f(\mathbf{x},\mathbf{y})^2 \\
&= \mathbb{E}_{\mathbf{w}\sim\mathbf{G}}\left[\frac{1}{m} f(\mathbf{w}_1^\top\mathbf{x})f(\mathbf{w}_1^\top\mathbf{y})f(\mathbf{w}_1^\top\mathbf{x})f(\mathbf{w}_1^\top\mathbf{y})\right] \\
&\quad + \mathbb{E}_{\mathbf{w}\sim\mathbf{G}}\left[\frac{m-1}{m} f(\mathbf{w}_1^\top\mathbf{x})f(\mathbf{w}_1^\top\mathbf{y})f(\mathbf{w}_2^\top\mathbf{x})f(\mathbf{w}_2^\top\mathbf{y})\right] - K_f(\mathbf{x},\mathbf{y})^2\,,
\end{aligned} \tag{S1}$$

and,

$$\begin{aligned}
\mathrm{MSE}[\mathrm{ORN}_f(\mathbf{x},\mathbf{y})] &= \mathbb{E}_{\mathbf{w}\sim\mathbf{G}_{\mathrm{ort}}}\left[\frac{1}{m} f(\mathbf{w}_1^\top\mathbf{x})f(\mathbf{w}_1^\top\mathbf{y})f(\mathbf{w}_1^\top\mathbf{x})f(\mathbf{w}_1^\top\mathbf{y})\right] \\
&\quad + \mathbb{E}_{\mathbf{w}\sim\mathbf{G}_{\mathrm{ort}}}\left[\frac{m-1}{m} f(\mathbf{w}_1^\top\mathbf{x})f(\mathbf{w}_1^\top\mathbf{y})f(\mathbf{w}_2^\top\mathbf{x})f(\mathbf{w}_2^\top\mathbf{y})\right] - K_f(\mathbf{x},\mathbf{y})^2\,.
\end{aligned} \tag{S2}$$

Here we have used the permutation symmetry of the matrix distributions to express the MSE in terms of a term depending on only a single row ($\mathbf{w}_1$) and a term depending on two rows ($\mathbf{w}_1$ and $\mathbf{w}_2$). Because the rows of $\mathbf{G}_{\mathrm{ort}}$ are marginally Gaussian, the terms that depend only on $\mathbf{w}_1$ are identical for both $\mathbf{G}$ and $\mathbf{G}_{\mathrm{ort}}$. Therefore we have that the difference in MSE, $\Delta\mathrm{MSE} \equiv \mathrm{MSE}[\mathrm{URN}_f(\mathbf{x},\mathbf{y})] - \mathrm{MSE}[\mathrm{ORN}_f(\mathbf{x},\mathbf{y})]$, is given by,

$$\Delta\mathrm{MSE} = \frac{m-1}{m}\left(\mathbb{E}_{\mathbf{w}\sim\mathbf{G}}\left[f(\mathbf{w}_1^\top\mathbf{x})f(\mathbf{w}_1^\top\mathbf{y})f(\mathbf{w}_2^\top\mathbf{x})f(\mathbf{w}_2^\top\mathbf{y})\right] - \mathbb{E}_{\mathbf{w}\sim\mathbf{G}_{\mathrm{ort}}}\left[f(\mathbf{w}_1^\top\mathbf{x})f(\mathbf{w}_1^\top\mathbf{y})f(\mathbf{w}_2^\top\mathbf{x})f(\mathbf{w}_2^\top\mathbf{y})\right]\right). \tag{S3}$$

The matrix distributions of $\mathbf{G}$ and $\mathbf{G}_{\mathrm{ort}}$ are rotationally-invariant. Therefore we can choose a coordinate system in which $\mathbf{x}$ and $\mathbf{y}$ lie in the first two directions. Specifically, let

$$(\mathbf{e}_1,\mathbf{e}_2) = (\mathbf{x},\mathbf{y})\begin{pmatrix}\mathbf{x}^\top\mathbf{x} & \mathbf{x}^\top\mathbf{y}\\ \mathbf{y}^\top\mathbf{x} & \mathbf{y}^\top\mathbf{y}\end{pmatrix}^{-\frac{1}{2}}, \tag{S4}$$

be the first two coordinate directions. Here we assume that $\mathbf{x}$ is not collinear with $\mathbf{y}$ in order that the inverse matrix square root exists. The collinear case follows from a similar argument or by taking the collinear limit of the result we now derive. In this new coordinate system, the functional dependence of $\Delta\mathrm{MSE}$ reduces to four variables, we which capture in a new function, g, to ease the notation,

$$\begin{aligned}
g(w_{11},w_{12},w_{21},w_{22}) &\equiv f(\mathbf{w}_1^\top\mathbf{x})f(\mathbf{w}_1^\top\mathbf{y})f(\mathbf{w}_2^\top\mathbf{x})f(\mathbf{w}_2^\top\mathbf{y}) \\
&= f(w_{11}\mathbf{e}_1^\top\mathbf{x} + w_{12}\mathbf{e}_2^\top\mathbf{x})f(w_{11}\mathbf{e}_1^\top\mathbf{y} + w_{12}\mathbf{e}_2^\top\mathbf{y})f(w_{21}\mathbf{e}_1^\top\mathbf{x} + w_{22}\mathbf{e}_2^\top\mathbf{x})f(w_{21}\mathbf{e}_1^\top\mathbf{y} + w_{22}\mathbf{e}_2^\top\mathbf{y})\,.
\end{aligned} \tag{S5}$$

Finally we rewrite $\Delta\mathrm{MSE}$ in terms of integrals,

$$\Delta\mathrm{MSE} = \frac{m-1}{m}\left(I_{\mathbf{G}}(g) - I_{\mathbf{G}_{\mathrm{ort}}}(g)\right), \tag{S6}$$

where the integrals we seek to evaluate are,

$$\begin{aligned} I_{\mathbf{G}}(g) &= \int ds_1 ds_2 dt_1 dt_2 \frac{e^{-\frac{1}{2}(s_1^2+s_2^2+t_1^2+t_2^2)}}{4\pi^2}\, g(s_1,s_2,t_1,t_2) \\ I_{\mathbf{G}_{\text{ort}}}(g) &= \int d\chi_1 d\chi_2\, \frac{e^{-\frac{1}{2}(\chi_1^2+\chi_2^2)}}{2^{d-2}\Gamma(\frac{d}{2})^2}\chi_1^{d-1}\chi_2^{d-1}\int_{\mathbb{O}_d} dU\; g(\chi_1 u_{11}, \chi_1 u_{12}, \chi_2 u_{21}, \chi_2 u_{22})\,. \end{aligned} \tag{S7}$$

In writing the expression for $I_{\mathbf{G}_{\text{ort}}}(g)$, we have used the fact that $\mathbf{G}_{\text{ort}}$ may be regarded as an orthogonal matrix whose rows have been scaled by χ-distributed random variables with d degrees of freedom. Next, we observe that the integrand depends on only two rows of U. Integrating over the remaining $d-2$ rows induces a marginal distribution over the Stiefel manifold $V_2(\mathbb{R}^d) = \{V \in \mathbb{R}^{2,d} \,|\, VV^\top = I_2\}$,

$$\begin{aligned} I_{\mathbf{G}_{\text{ort}}}(g) &= \int d\chi_1 d\chi_2\, \frac{e^{-\frac{1}{2}(\chi_1^2+\chi_2^2)}}{2^{d-2}\Gamma(\frac{d}{2})^2}\chi_1^{d-1}\chi_2^{d-1}\int_{V_2(\mathbb{R}^d)} dV\; g(\chi_1 v_{11}, \chi_1 v_{12}, \chi_2 v_{21}, \chi_2 v_{22}) \\ &= \frac{(d-2)(d-3)}{4\pi^2}\int d\chi_1 d\chi_2\, \frac{e^{-\frac{1}{2}(\chi_1^2+\chi_2^2)}}{2^{d-2}\Gamma(\frac{d}{2})^2}\chi_1^{d-1}\chi_2^{d-1}\int_{TT^\top<I_2} dT\; \det(I_2 - TT^\top)^{\frac{d-5}{2}} g(\chi_1 t_{11}, \chi_1 t_{12}, \chi_2 t_{21}, \chi_2 t_{22})\,, \end{aligned} \tag{S8}$$

where in the second line we have marginalized out $d-2$ rows of V, and $TT^\top < I_2$ denotes those matrices $T \in \mathbb{R}^{2,2}$ for which $I_2 - TT^\top$ is positive definite. Next we apply the following change of variables,

$$\begin{pmatrix} \chi_1 \\ \chi_2 \\ t_{11} \\ t_{12} \\ t_{21} \\ t_{22} \end{pmatrix} = \begin{pmatrix} \sqrt{ds^2 + s_1^2 + s_2^2} \\ \sqrt{dt^2 + t_1^2 + t_2^2} \\ \frac{s_1}{\sqrt{ds^2+s_1^2+s_2^2}} \\ \frac{s_2}{\sqrt{ds^2+s_1^2+s_2^2}} \\ \frac{t_1}{\sqrt{dt^2+t_1^2+t_2^2}} \\ \frac{t_2}{\sqrt{dt^2+t_1^2+t_2^2}} \end{pmatrix}, \tag{S9}$$

which gives,

$$I_{\mathbf{G}_{\text{ort}}}(g) = \int ds_1 ds_2 dt_1 dt_2 \frac{e^{-\frac{1}{2}(s_1^2+s_2^2+t_1^2+t_2^2)}}{4\pi^2}\, \rho(d,s_1,s_2,t_1,t_2) g(s_1,s_2,t_1,t_2)\,, \tag{S10}$$

where the function ρ encodes the effect of orthogonality and equals,

$$\begin{aligned} \rho(d,s_1,s_2,t_1,t_2) = \frac{(d-2)(d-3)d^{d-2}}{2^{d-2}\Gamma(\frac{d}{2})^2}\int ds dt\; e^{-\frac{d}{2}(s^2+t^2)} s^{d-3} t^{d-3} \\ \times \sqrt{1+\frac{s_1^2+s_2^2}{ds^2}}\sqrt{1+\frac{t_1^2+t_2^2}{dt^2}}\left[1 - \frac{s_1 t_1 + s_2 t_2}{d^2 s^2 t^2}\right]_+^{\frac{1}{2}(d-5)}\,. \end{aligned} \tag{S11}$$

The large-d asymptotics of ρ can be obtained using the saddle point method, also known as Laplace's method. As $d \to \infty$, the integrand decays exponentially in d, and is maximized when $s = t = 1$, which lies in the interior of the integration region as $d \to \infty$. We can therefore expand the integrand around this point and obtain an asymptotic expansion in d by evaluating Gaussian integrals. The result is,

$$\begin{aligned} \rho(d,s_1,s_2,t_1,t_2) &= 1 - \frac{2 - s_1^2 - s_2^2 - t_1^2 - t_2^2 + s_1^2 t_1^2 + s_2^2 t_2^2 + 2 s_1 s_2 t_1 t_2}{2d} + \mathcal{O}\left(\frac{1}{d^2}\right) \\ &\equiv 1 - \frac{\rho_1(s_1,s_2,t_1,t_2)}{d} + \mathcal{O}\left(\frac{1}{d^2}\right). \end{aligned} \tag{S12}$$

Altogether, we have,

$$\Delta\text{MSE} = \frac{m-1}{m}\int ds_1 ds_2 dt_1 dt_2 \frac{e^{-\frac{1}{2}(s_1^2+s_2^2+t_1^2+t_2^2)}}{4\pi^2}\, \frac{\rho_1(s_1,s_2,t_1,t_2)}{d} g(s_1,s_2,t_1,t_2) + \mathcal{O}\left(\frac{1}{d^2}\right). \tag{S13}$$

Next we prove the non-negativity of ΔMSE by rewriting it in terms of differential operators acting on $K_f(\mathbf{x},\mathbf{y})$. To this end, consider the following change of variables,

$$\begin{aligned}\mathbf{z}_1 \equiv \begin{pmatrix} z_{11} \\ z_{12} \end{pmatrix} &= \begin{pmatrix} s_1 \mathbf{e}_1^\top \mathbf{x} + s_2 \mathbf{e}_2^\top \mathbf{x} \\ s_1 \mathbf{e}_1^\top \mathbf{y} + s_2 \mathbf{e}_2^\top \mathbf{y} \end{pmatrix} = (\mathbf{X}^\top \mathbf{X})^{-\frac{1}{2}} \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} \\ \mathbf{z}_2 \equiv \begin{pmatrix} z_{21} \\ z_{22} \end{pmatrix} &= \begin{pmatrix} t_1 \mathbf{e}_1^\top \mathbf{x} + t_2 \mathbf{e}_2^\top \mathbf{x} \\ t_1 \mathbf{e}_1^\top \mathbf{y} + t_2 \mathbf{e}_2^\top \mathbf{y} \end{pmatrix} = (\mathbf{X}^\top \mathbf{X})^{-\frac{1}{2}} \begin{pmatrix} t_1 \\ t_2 \end{pmatrix},\end{aligned} \tag{S14}$$

where $\mathbf{X} = (\mathbf{x}, \mathbf{y})$. In these variables, ΔMSE can be written as,

$$\begin{aligned}\Delta\mathrm{MSE} = \frac{1}{d}\frac{m-1}{m} \int dz_{11} dz_{12} \frac{e^{-\frac{1}{2}\mathbf{z}_1^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{z}_1}}{\sqrt{\det\left(2\pi \mathbf{X}^\top \mathbf{X}\right)}} f(z_{11}) f(z_{12}) \int dz_{21} dz_{22} \frac{e^{-\frac{1}{2}\mathbf{z}_2^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{z}_2}}{\sqrt{\det\left(2\pi \mathbf{X}^\top \mathbf{X}\right)}} f(z_{21}) f(z_{22}) \\ \times \left[1 - \frac{1}{2}\mathbf{z}_1^\top \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{z}_1 - \frac{1}{2}\mathbf{z}_2^\top \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{z}_2 + \frac{1}{2}\left(\mathbf{z}_1^\top \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{z}_2\right)^2\right] + \mathcal{O}\left(\frac{1}{d^2}\right).\end{aligned} \tag{S15}$$

Observe that the each additive term of the integrand can be factorized into a product of integrals depending only on $\mathbf{z}_1$ or $\mathbf{z}_2$. Each such term can be expressed in terms of derivatives of $K_f(\mathbf{x},\mathbf{y})$ by noting that,

$$\begin{aligned}K_f(\mathbf{x},\mathbf{y}) &= \int dz_{11} dz_{12} \frac{e^{-\frac{1}{2}\mathbf{z}_1^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{z}_1}}{\sqrt{\det\left(2\pi \mathbf{X}^\top \mathbf{X}\right)}} f(z_{11}) f(z_{12}) \\ &= \int dz_{21} dz_{22} \frac{e^{-\frac{1}{2}\mathbf{z}_2^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{z}_2}}{\sqrt{\det\left(2\pi \mathbf{X}^\top \mathbf{X}\right)}} f(z_{21}) f(z_{22}),\end{aligned} \tag{S16}$$

and exchanging the order of integration and derivatives with respect to $\theta \equiv (\theta_1, \theta_2, \theta_3)^\top = (\mathbf{x}^\top \mathbf{x}, \mathbf{x}^\top \mathbf{y}, \mathbf{y}^\top \mathbf{y})^\top$. In particular, some straightforward algebra gives,

$$\Delta\mathrm{MSE} = \frac{1}{d}\frac{m-1}{m}\left[(\theta_1 \partial_1 K_f + \theta_2 \partial_2 K_f + \theta_3 \partial_3 K_f)^2 + \frac{\theta_1\theta_3 - \theta_2^2}{2}\left((\partial_2 K_f)^2 - 4\partial_1 K_f \partial_3 K_f\right)\right] + \mathcal{O}\left(\frac{1}{d^2}\right), \tag{S17}$$

where $\partial_i \equiv \partial/\partial\theta_i$. This representation is possible because the rotational symmetry of $\mathbf{G}$ and $\mathbf{G}_{\mathrm{ort}}$ requires that $K_f(\mathbf{x},\mathbf{y})$ be a function of the three rotationally-invariant quantities θ_1, θ_2, and θ_3. Moreover, also by using this symmetry, it is easy to check that,

$$\Delta\mathrm{MSE} = \frac{1}{d}\frac{m-1}{m}\,\mathrm{tr}\,\mathbf{R}^2 + \mathcal{O}\left(\frac{1}{d^2}\right), \tag{S18}$$

where the entries of $\mathbf{R}$ are given by,

$$R_{ij} = R_{ji} = \frac{1}{2}\left(x_i \frac{\partial K_f(\mathbf{x},\mathbf{y})}{\partial x_j} + y_i \frac{\partial K_f(\mathbf{x},\mathbf{y})}{\partial y_j}\right). \tag{S19}$$

Because $\mathbf{R}$ is a symmetric matrix, its eigenvalues are real, and therefore $\mathrm{tr}\,\mathbf{R}^2 \geq 0$, which proves that asymptotically,

$$\Delta\mathrm{MSE} \geq 0. \tag{S20}$$

The inequality is strict for most inputs $\mathbf{x}$ and $\mathbf{y}$. In order to reveal the conditions under which equality can hold, we observe that because $\mathbf{R}$ is a real symmetric matrix, $\mathrm{tr}\,\mathbf{R}^2 = 0$ implies that $\mathbf{R} = 0$. Therefore the following additional equations must be satisfied in order that $\Delta\mathrm{MSE} = 0$ asymptotically,

$$\begin{aligned}0 = \mathrm{tr}\,\mathbf{R} &= \theta_1 \partial_1 K_f + 2\theta_2 \partial_2 K_f + \theta_3 \partial_3 K_f \\ 0 = \mathbf{x}^\top \mathbf{R} \mathbf{x} &= \theta_1^2 \partial_1 K_f + 2\theta_1\theta_2 \partial_2 K_f + \theta_2^2 \partial_3 K_f \\ 0 = \mathbf{x}^\top \mathbf{R} \mathbf{y} &= \theta_1\theta_2 \partial_1 K_f + (\theta_1\theta_3 + \theta_2^2)\partial_2 K_f + \theta_2\theta_3 \partial_3 K_f \\ 0 = \mathbf{y}^\top \mathbf{R} \mathbf{y} &= \theta_2^2 \partial_1 K_f + 2\theta_2\theta_3 \partial_2 K_f + \theta_3^2 \partial_3 K_f.\end{aligned} \tag{S21}$$

A solution to these equations requires one of the following conditions on the inputs (in terms of θ) and the kernel (in terms of its derivatives $\partial_i K_f$):

1. $\theta_1 = \theta_2 = \theta_3 = 0$
2. $\theta_1 = \theta_2 = \partial_3 K_f = 0$
3. $\theta_2 = \theta_3 = \partial_1 K_f = 0$
4. $\partial_1 K_f = \partial_2 K_f = \partial_3 K_f = 0$
5. $\theta_1\theta_3 = \theta_2^2$, $\theta_1\partial_1 + \theta_3\partial_3 = 0$, $\partial_2\theta_2 = 0$

If $\mathbf{x} \neq 0$, $\mathbf{y} \neq 0$, $\nabla_{\mathbf{x}} K_f \neq 0$, $\nabla_{\mathbf{y}} K_f \neq 0$, and $\mathbf{x}$ and $\mathbf{y}$ are not collinear, then none of these equations can be satisfied and therefore we have, asymptotically,

$$\Delta\text{MSE} \geq O(\frac{1}{d})\,. \tag{S22}$$

That proves Theorem 1. To prove Theorem 2, we use the proof of Theorem 1 in [Pillai and Smith, 2015] showing that for every $a, b, \epsilon > 0$ there exists a constant $C(b) > 0$ such that if $k > \max(C(b)d\log(d), (5a+6+\frac{1}{2}+2\epsilon)d\log(d))$ then:

$$\|\mu_{\text{HAAR}} - \mu_{\text{KAC}}\|_{\text{TV}} \leq d^{2a+2}(1 - \frac{1}{2d})^{(5a+5)d\log(d)} + \frac{1}{d^{4(a+1)}} + \frac{2}{d^{\epsilon}} + 6000d^{2-\frac{2(a-1)}{5}} + d^{6-\frac{b}{3}}, \tag{S23}$$

where $\mu_{\mathcal{D}}$ stands for the probabilistic measure related to the probabilistic distribution $\mathcal{D}$, KAC is a distribution of a vector $\mathbf{K}\mathbf{e}_1$ for Kac's random walk matrix $\mathbf{K}$ using k Givens random rotations and HAAR stands for the Haar distribution on the sphere. Notice that from that theorem we get:

$$|\text{MSE}(\text{KRN}_f^k(\mathbf{x}, \mathbf{y})) - \text{MSE}(\text{ORN}_f(\mathbf{x}, \mathbf{y}))| = o(\frac{1}{d}) \tag{S24}$$

for $k = C \cdot d\log(d)$ and constant $C > 0$ large enough. Combining this with Theorem 1, we complete the proof of Theorem 2.

B Proof of Theorem 3

Theorem 4.

$$F(\mathbf{M}_0) - F(\mathbf{M}_t) \geq \sum_{u=0}^{t-1} \frac{\|\nabla F(\mathbf{M}_u)\|^2}{2d(d-1)B}$$

The proof of this theorem is an adapted version of the coordinate descent procedure proposed by [Patrascu and Necoara, 2015]. We start by enunciating the following lemma:

Lemma 1 (Lemma 1 from [Shalit and Chechik, 2014])**.** *If $f : \mathbb{R} \to \mathbb{R}$ is periodic and differentiable, having period 2π, and B Lipschitz derivative f'. It follows that for all $\Theta \in [-\pi, \pi]$: $f(\Theta) \leq f(0) + \Theta f'(0) + \frac{B}{2}\Theta^2$*

As a consequence, for all I, J the function $F(\mathbf{M}_{t-1}\mathbf{G}_{I,J}^{\Theta}) \leq F(\mathbf{M}_{t-1}\mathbf{G}_{I,J}^{0}) - \frac{F'(\mathbf{M}_{t-1}\mathbf{G}_{I,J}^{0})^2}{2B}$, the minimizer of the expression above. By definition of the algorithm, the indices of the givens rotation chosen are the ones minimizing, among all the

Recall the Frobenius norm of the riemanninan gradient of F satisfies:

$$\begin{aligned}\|\nabla F(\mathbf{M}_{t-1})\|_F^2 &= \sum_{1\leq I<J\leq d} 2F'(\mathbf{M}_{t-1}\mathbf{G}_{I,J}^{0})^2 \\ &:= \sum_{1\leq I<J\leq d} 2F'_{I,J}(0)^2\end{aligned}$$

The second equality follows because for all I, J pairs, $\mathbf{G}_{I,J}^{0} = \mathbb{I}$ and we define $F_{I,J}(\Theta) = F(\mathbf{M}_{t-1}\mathbf{G}_{I,J}^{\Theta})$.

Since Algorithm 1 picks the optimal I, J pair, the descent gain from $t-1$ to t is at least the average of the gradient's $\binom{d}{2}$ directions:

$$F(\mathbf{M}_{t-1}) - F(\mathbf{M}_t) \geq \max_{I,J} \frac{F'_{I,J}(0)^2}{2B} \geq \frac{\|\nabla F(\mathbf{M}_{t-1})\|_F^2}{2Bd(d-1)}$$

The result follows.

C Experiments

C.1 Implementation Details

All algorithms are implemented in TensorFlow [Abadi et al., 2016] based on OpenAI baselines [Dhariwal et al., 2017]. We use the default parameter settings for PPO and TRPO algorithms. For PPO, the clipping rate is $\epsilon = 0.2$, learning rate $\alpha \in \{3 \cdot 10^{-4}, 3 \cdot 10^{-5}\}$ and. For TRPO, the trust region size is $\epsilon = 0.01$. The batchsize for update is $n = 2048$ for all algorithms.

For PPO, the policy is parameterized as a neural network with 2 hidden layers each with 64 hidden units and tanh non-linear activation in between layers. The last layer does not have non-linear activation. The value function is parameterized as a neural network with similar architecture.

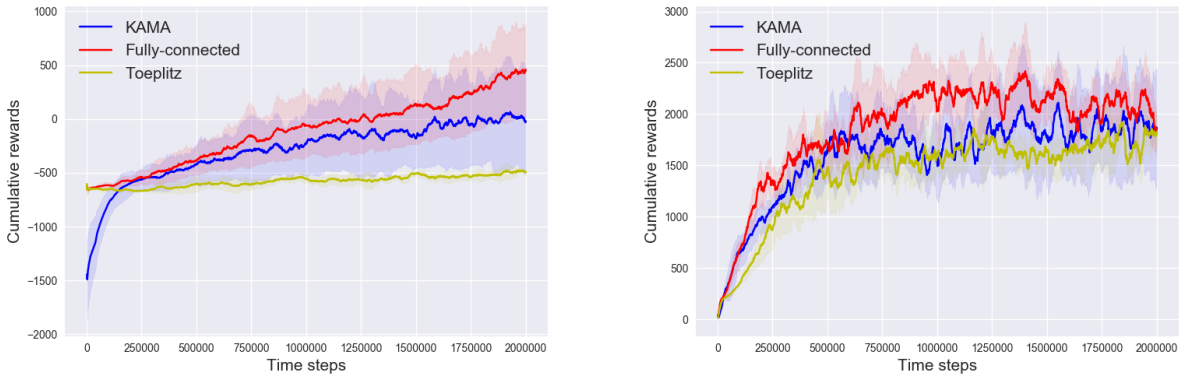
For TRPO, the policy is parameterized as a neural network with 2 hidden layers each with 32 hidden units and tanh non-linear activation in between layers. The last layer does not have non-linear activation. The value function is parameterized as a neural network with similar architecture.

All environments are from OpenAI gym [Brockman et al., 2016, Todorov et al., 2012] and Roboschool [Schulman et al., 2017].

C.2 Additional Experiments

We have shown that KAMA-NN policies achieve substantial compression of parameters while maintaining good performance. By varying the number of Givens rotations K in the structured matrix, we can avoid two undesirable extremes: fully-connected policies (good performance, but many parameters) and Toeplitz policies (significant compression, but bad performance since the model is rigid). Below we show that we can achieve further compression for KAMA-NNs for our RL tasks by reducing the number of rotations in the first and third structured matrices. We let $K_{1,3}$ be the number of rotations in the first and third structured matrix and K_2 be the number of rotations in the second structured matrix.

While in previous experiments we uses $K = K_{1,3} = K_2 = 200$ for PPO and $K = K_{1,3} = K_2 = 100$ for TRPO, now we use $K_{1,3} = 100, K_2 = 200$ for PPO and $K_{1,3} = 50, K_2 = 100$ for TRPO.



(a) **TRPO-HalfCheetah** (b) **TRPO-Hopper**

Figure S1: Illustration of KAMA-NNs policies on MuJoCo benchmark tasks with PPO/TRPO. KAMA-NNs policies with varying number of Givens rotations in the first/third matrix and the second matrix is compared against an unstructured policy and Toeplitz policy. For each task we train the policy with PPO/TRPO for a fixed number of steps and we show the mean $\pm$ std performance. Vertical axis is the cumulative reward and horizontal axis stands for the # of time steps.