Google AI

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

UNIVERSITY OF CAMBRIDGE

Hybrid Random Features

Krzysztof Choromanski*, Haoxian Chen*, Han Lin*, Yuanzhe Ma*, Arijit Sehanobish*,
Deepali Jain, Michael S Ryoo, Jake Varley, Andy Zeng,
Valerii Likhosherstov, Dmitry Kalashnikov, Vikas Sindhwani, Adrian Weller

** Equal Contribution*
Correspondence to kchoro@google.com

ICLR

Random Features for Softmax Kernel Approximation

- Softmax Kernel

$$\mathrm{SM}(\mathbf{x}, \mathbf{y}) \stackrel{\mathrm{def}}{=} \exp(\mathbf{x}^\top \mathbf{y}), \qquad \text{where } \mathbf{x} \in \mathbb{R}^d, \mathbf{y} \in \mathbb{R}^d$$

- Random Features Approximation for Softmax Kernel

$$\widehat{\mathrm{SM}}_m(\mathbf{x}, \mathbf{y}) = \phi_m(\mathbf{x})^\top \phi_m(\mathbf{y}), \qquad \text{where } \phi : \mathbb{R}^d \to \mathbb{R}^{2m},\ m : \text{number of random features}$$

Random Features for Softmax Kernel Approximation

- Softmax Kernel

$$\mathrm{SM}(\mathbf{x}, \mathbf{y}) \stackrel{\mathrm{def}}{=} \exp(\mathbf{x}^\top \mathbf{y}), \qquad \text{where } \mathbf{x} \in \mathbb{R}^d, \mathbf{y} \in \mathbb{R}^d$$

- Random Features Approximation for Softmax Kernel

$$\widehat{\mathrm{SM}}_m(\mathbf{x}, \mathbf{y}) = \phi_m(\mathbf{x})^\top \phi_m(\mathbf{y}), \qquad \text{where } \phi : \mathbb{R}^d \to \mathbb{R}^{2m},\ m : \text{number of random features}$$

- Trigonometric Random Features (Rahimi & Recht, 2007)

$$\phi_m^{\mathrm{trig}}(\mathbf{u}) = \frac{1}{\sqrt{m}} \exp\left(\frac{\|\mathbf{u}\|^2}{2}\right) \left(\sin(\omega_1^\top \mathbf{u}), ..., \sin(\omega_m^\top \mathbf{u}), \cos(\omega_1^\top \mathbf{u}), ..., \cos(\omega_m^\top \mathbf{u})\right)^\top$$

- Positive Random Features (FAVOR+) (Choromanski et al., 2021)

$$\phi_m^{++}(\mathbf{u}) = \frac{1}{\sqrt{2m}} \exp\left(-\frac{\|\mathbf{u}\|^2}{2}\right) \left(\exp(\omega_1^\top \mathbf{u}), ..., \exp(\omega_m^\top \mathbf{u}), \exp(-\omega_1^\top \mathbf{u}), ..., \exp(-\omega_m^\top \mathbf{u})\right)^\top$$

where $\omega_1, ..., \omega_m \overset{\mathrm{iid}}{\sim} \mathcal{N}(0, \mathbf{I}_d)$

MSEs for Trigonometric and Positive RFs

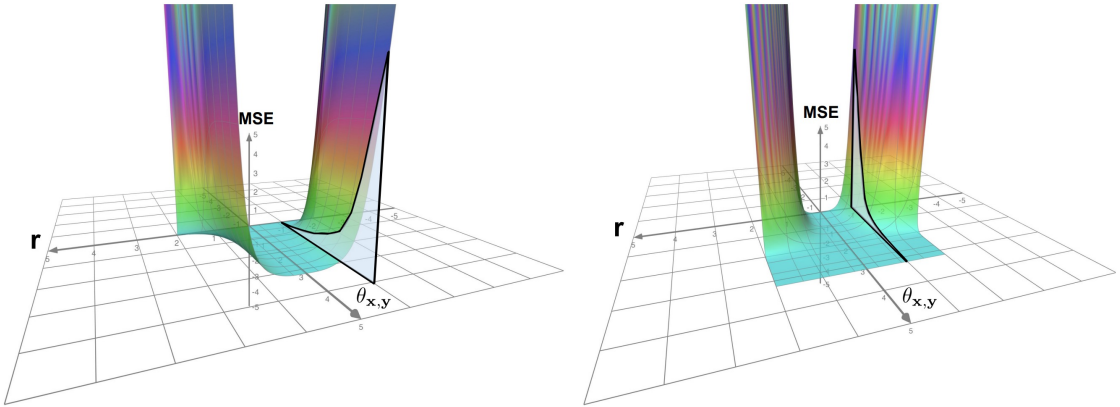
Lemma 2.2 (positive versus trigonometric RFs)**.** *Take* $\Delta = \mathbf{x} - \mathbf{y}$, $\mathbf{z} = \mathbf{x} + \mathbf{y}$, $f_1(u) = (2m)^{-1}\exp(u^2)\mathrm{SM}^{-2}(\mathbf{x},\mathbf{y})$, $f_2(u) = (2m)^{-1}\exp(u^2)\mathrm{SM}^{2}(\mathbf{x},\mathbf{y})$, $f_3(u) = (1-\exp(-u^2))^2$. *The MSEs of these estimators are:*

$$\mathrm{MSE}(\widehat{\mathrm{SM}}_m^{\mathrm{trig}}(\mathbf{x},\mathbf{y})) = f_1(\|\mathbf{z}\|_2)f_3(\|\Delta\|_2),\ \mathrm{MSE}(\widehat{\mathrm{SM}}_m^{++}(\mathbf{x},\mathbf{y})) = f_2(\|\mathbf{z}\|_2)f_3(\|\mathbf{z}\|_2). \quad (7)$$

MSEs for Trigonometric and Positive RFs

Lemma 2.2 (positive versus trigonometric RFs). *Take* $\Delta = \mathbf{x} - \mathbf{y}$, $\mathbf{z} = \mathbf{x} + \mathbf{y}$, $f_1(u) = (2m)^{-1}\exp(u^2)\mathrm{SM}^{-2}(\mathbf{x},\mathbf{y})$, $f_2(u) = (2m)^{-1}\exp(u^2)\mathrm{SM}^{2}(\mathbf{x},\mathbf{y})$, $f_3(u) = (1-\exp(-u^2))^2$. *The MSEs of these estimators are:*

$$\mathrm{MSE}(\widehat{\mathrm{SM}}_m^{\mathrm{trig}}(\mathbf{x},\mathbf{y})) = f_1(\|\mathbf{z}\|_2)f_3(\|\Delta\|_2),\ \mathrm{MSE}(\widehat{\mathrm{SM}}_m^{++}(\mathbf{x},\mathbf{y})) = f_2(\|\mathbf{z}\|_2)f_3(\|\mathbf{z}\|_2). \quad (7)$$



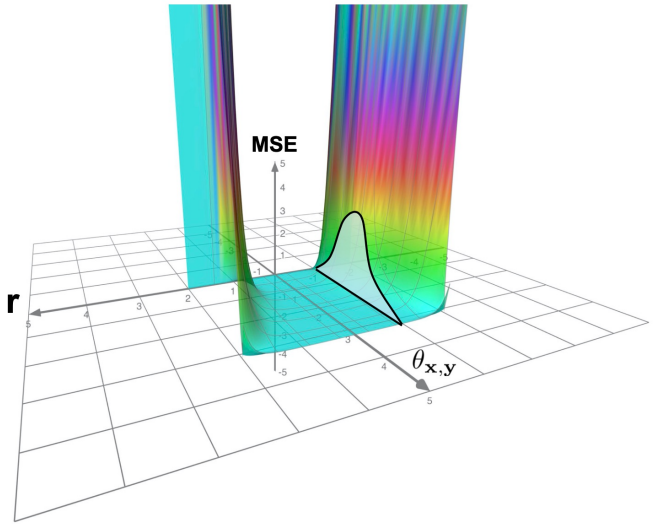
MSE for $\widehat{\mathrm{SM}}^{\mathrm{trig}}$ MSE for $\widehat{\mathrm{SM}}^{++}$

MSEs are given as functions of: an angle $\theta_{\mathbf{x},\mathbf{y}} \in [0,\pi]$ between $\mathbf{x}$ and $\mathbf{y}$ and r (symmetrized along 0 for length axis) . For each plot, we marked in grey its slice for a fixed r.

Our Motivation for Hybrid Random Features (HRFs)

- Question:

  Can we create a hybrid estimator which combines the advantage of $\widehat{\mathrm{SM}}^{\mathrm{trig}}$ and $\widehat{\mathrm{SM}}^{++}$ to make MSE goes to zero for both $\theta_{\mathbf{x},\mathbf{y}} \to 0$ and $\theta_{\mathbf{x},\mathbf{y}} \to \pi$?



Hybrid Random Feature Estimator of SM(x,y)

Denote by $\mathcal{E} = (\widehat{\mathrm{SM}}^{k}(\mathbf{x}, \mathbf{y}))_{k=1}^{p+1}$ a list of estimators of $\mathrm{SM}(\mathbf{x}, \mathbf{y})$ (the so-called *base estimators*) and by $\Lambda = (\widehat{\lambda}^{k}(\mathbf{x}, \mathbf{y}))_{k=1}^{p}$ a list of estimators of $\{\lambda^{k}(\mathbf{x}, \mathbf{y})\}_{k=1}^{p}$ for some functions $\lambda^{k} : \mathbb{R}^{d} \times \mathbb{R}^{d} \to [0, 1]$, constructed independently from $\mathcal{E}$. Take the following estimator of $\mathrm{SM}(\mathbf{x}, \mathbf{y})$:

$$\widehat{\mathrm{SM}}^{\mathcal{E},\Lambda}(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{p} \widehat{\lambda}^{k}(\mathbf{x}, \mathbf{y}) \widehat{\mathrm{SM}}^{k}(\mathbf{x}, \mathbf{y}) + \left(1 - \sum_{k=1}^{p} \widehat{\lambda}^{k}(\mathbf{x}, \mathbf{y})\right) \widehat{\mathrm{SM}}^{p+1}(\mathbf{x}, \mathbf{y})$$

Hybrid Random Feature Estimator of SM(x,y)

Denote by $\mathcal{E} = (\widehat{\mathrm{SM}}^{k}(\mathbf{x},\mathbf{y}))_{k=1}^{p+1}$ a list of estimators of $\mathrm{SM}(\mathbf{x},\mathbf{y})$ (the so-called *base estimators*) and by $\Lambda = (\widehat{\lambda}^{k}(\mathbf{x},\mathbf{y}))_{k=1}^{p}$ a list of estimators of $\{\lambda^{k}(\mathbf{x},\mathbf{y})\}_{k=1}^{p}$ for some functions $\lambda^{k} : \mathbb{R}^{d} \times \mathbb{R}^{d} \to [0,1]$, constructed independently from $\mathcal{E}$. Take the following estimator of $\mathrm{SM}(\mathbf{x},\mathbf{y})$:

$$\widehat{\mathrm{SM}}^{\mathcal{E},\Lambda}(\mathbf{x},\mathbf{y}) = \sum_{k=1}^{p} \widehat{\lambda}^{k}(\mathbf{x},\mathbf{y})\widehat{\mathrm{SM}}^{k}(\mathbf{x},\mathbf{y}) + \left(1 - \sum_{k=1}^{p} \widehat{\lambda}^{k}(\mathbf{x},\mathbf{y})\right)\widehat{\mathrm{SM}}^{p+1}(\mathbf{x},\mathbf{y})$$

- Base Estimators: $\widehat{\mathrm{SM}}^{k}(\mathbf{x},\mathbf{y}) = (\phi_{1,m}^{k}(\mathbf{x}))^{\top}\phi_{2,m}^{k}(\mathbf{y})$

where $\phi_{j,m}^{k}(\mathbf{u}) = \frac{1}{\sqrt{m}}\phi_{j,m}^{1,k}(\mathbf{u}) \star \ldots \star \phi_{j,m}^{t_k,k}(\mathbf{u})$, $t_k > 0$ and $\phi_{j,m}^{1,k}, \ldots, \phi_{j,m}^{t_k,k} : \mathbb{R}^{d} \to \mathbb{R}^{m}$

Hybrid Random Feature Estimator of SM(x,y)

Denote by $\mathcal{E} = (\widehat{\mathrm{SM}}^k(\mathbf{x},\mathbf{y}))_{k=1}^{p+1}$ a list of estimators of $\mathrm{SM}(\mathbf{x},\mathbf{y})$ (the so-called *base estimators*) and by $\Lambda = (\widehat{\lambda}^k(\mathbf{x},\mathbf{y}))_{k=1}^{p}$ a list of estimators of $\{\lambda^k(\mathbf{x},\mathbf{y})\}_{k=1}^{p}$ for some functions $\lambda^k : \mathbb{R}^d \times \mathbb{R}^d \to [0,1]$, constructed independently from $\mathcal{E}$. Take the following estimator of $\mathrm{SM}(\mathbf{x},\mathbf{y})$:

$$\widehat{\mathrm{SM}}^{\mathcal{E},\Lambda}(\mathbf{x},\mathbf{y}) = \sum_{k=1}^{p} \widehat{\lambda}^k(\mathbf{x},\mathbf{y})\widehat{\mathrm{SM}}^k(\mathbf{x},\mathbf{y}) + \left(1 - \sum_{k=1}^{p} \widehat{\lambda}^k(\mathbf{x},\mathbf{y})\right) \widehat{\mathrm{SM}}^{p+1}(\mathbf{x},\mathbf{y})$$

- Base Estimators: $\widehat{\mathrm{SM}}^k(\mathbf{x},\mathbf{y}) = (\phi_{1,m}^k(\mathbf{x}))^\top \phi_{2,m}^k(\mathbf{y})$

  where $\phi_{j,m}^k(\mathbf{u}) = \frac{1}{\sqrt{m}}\phi_{j,m}^{1,k}(\mathbf{u}) \star \ldots \star \phi_{j,m}^{t_k,k}(\mathbf{u})$, $t_k > 0$ and $\phi_{j,m}^{1,k}, \ldots, \phi_{j,m}^{t_k,k} : \mathbb{R}^d \to \mathbb{R}^m$

- Lambda Coefficients: $\lambda^k(\mathbf{x},\mathbf{y}) = a_k + \mathbb{E}_{\tau\sim\Omega}[\sum_{i=1}^{l_k} f_{1,k}^i(\mathbf{x},\tau) f_{2,k}^i(\mathbf{y},\tau)]$

  where scalar $a_k \in \mathbb{R}$, distribution $\Omega \in \mathcal{P}(\mathbb{R}^d)$, mappings $\xi_k^i, \eta_k^i : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$

- Lambda Coefficients Estimators: $\widehat{\lambda}_n^k(\mathbf{x},\mathbf{y}) = a_k + (\rho_{1,n}^k(\mathbf{x}))^\top \rho_{2,n}^k(\mathbf{y})$

  where $\rho_{j,n}^k(\mathbf{u}) = \frac{1}{\sqrt{n}}\rho_{j,n}^{1,k}(\mathbf{u}) \star \ldots \star \rho_{j,n}^{l_k,k}(\mathbf{u})$, $\rho_{j,n}^{i,k}(\mathbf{u}) = (f_{j,k}^i(\mathbf{u},\tau_1), \ldots, f_{j,k}^i(\mathbf{u},\tau_n))^\top$, $\tau_1, \ldots, \tau_n \sim \Omega$

HRF Estimator and Its Time Complexity

Lemma 2.3. *The HRF estimator* $\widehat{\mathrm{SM}}^{\mathrm{hyb}}_{m,n}(\mathbf{x},\mathbf{y})$ *satisfies* $\widehat{\mathrm{SM}}^{\mathrm{hyb}}_{m,n}(\mathbf{x},\mathbf{y}) = \Psi_1(\mathbf{x})^\top \Psi_2(\mathbf{y})$, *where* Ψ_j *for* $j \in \{1,2\}$ *is given as* $\Psi_j(\mathbf{z}) = \Psi_j^1(\mathbf{z}) \star \Psi_j^2(\mathbf{z}) \star \Psi_j^3(\mathbf{z}) \star \Psi_j^4(\mathbf{z})$ *and:*

$$\Psi_j^1(\mathbf{z}) = \prod_{k=1,\ldots,p}^{\star} \sqrt{\frac{a_k}{m}} \phi_{j,m}^{1,k}(\mathbf{z}) \star \ldots \star \phi_{j,m}^{t_k,k}(\mathbf{z})$$

$$\Psi_j^2(\mathbf{z}) = \frac{1}{\sqrt{mn}} \prod_{k=1,\ldots,p}^{\star} \prod_{i,j \in \{1,\ldots,l_k\}\times\{1,\ldots,t_k\}}^{\star} \rho_{j,n}^{i,k}(\mathbf{z}) \otimes \phi_{j,m}^{j,k}(\mathbf{z})$$

$$\Psi_j^3(\mathbf{z}) = \sqrt{\frac{1-\sum_{k=1}^{p} a_k}{m}} \phi_{j,m}^{1,p+1}(\mathbf{z}) \star \ldots \star \phi_{j,m}^{t_{p+1},p+1}(\mathbf{z})$$

$$\Psi_j^4(\mathbf{z}) = \frac{\mathbf{i}}{\sqrt{mn}} \prod_{k=1,\ldots,p}^{\star} \prod_{i,j \in \{1,\ldots,l_k\}\times\{1,\ldots,t_{p+1}\}}^{\star} \rho_{j,n}^{i,k}(\mathbf{z}) \otimes \phi_{j,m}^{j,p+1}(\mathbf{z})$$

- Computational Gains:

  The resulting $\Theta(mn)$-dimensional random feature map $\Psi(\mathbf{z})$ can be constructed in time $O(nd + md + mn)$ as opposed to $O(mnd)$ as it is the case for the regular estimator.

Bipolar HRF Estimators

- Take $\mathcal{E} = (\widehat{\mathrm{SM}}^{++}(\mathbf{x},\mathbf{y}), \widehat{\mathrm{SM}}^{\mathrm{trig}}(\mathbf{x},\mathbf{y}))$:

$$\widehat{\mathrm{SM}}^{\mathrm{hyb}}_{m,n}(\mathbf{x},\mathbf{y}) = \widehat{\lambda}_n(\mathbf{x},\mathbf{y})\widehat{\mathrm{SM}}^{++}_m(\mathbf{x},\mathbf{y}) + (1-\widehat{\lambda}_n(\mathbf{x},\mathbf{y}))\widehat{\mathrm{SM}}^{\mathrm{trig}}_m(\mathbf{x},\mathbf{y})$$

Bipolar HRF Estimators

- Take $\mathcal{E} = (\widehat{\mathrm{SM}}^{++}(\mathbf{x},\mathbf{y}), \widehat{\mathrm{SM}}^{\mathrm{trig}}(\mathbf{x},\mathbf{y}))$:

$$\widehat{\mathrm{SM}}^{\mathrm{hyb}}_{m,n}(\mathbf{x},\mathbf{y}) = \widehat{\lambda}_n(\mathbf{x},\mathbf{y})\widehat{\mathrm{SM}}^{++}_m(\mathbf{x},\mathbf{y}) + (1-\widehat{\lambda}_n(\mathbf{x},\mathbf{y}))\widehat{\mathrm{SM}}^{\mathrm{trig}}_m(\mathbf{x},\mathbf{y})$$

- Angular Lambda Coefficients:

$$\lambda(\mathbf{x},\mathbf{y}) = \tfrac{\theta_{\mathbf{x},\mathbf{y}}}{\pi} \qquad\qquad \widehat{\lambda}(\mathbf{x},\mathbf{y}) = \frac{1}{2} + \rho_n(\mathbf{x})^\top \rho_n(\mathbf{y})$$

$$\text{where } \rho_{1,n}(\mathbf{z}) = \rho_{2,n}(\mathbf{z}) = \rho_n(\mathbf{z}) \overset{\text{def}}{=} \tfrac{\mathbf{i}}{\sqrt{2n}}(\mathrm{sgn}(\tau_1^\top \mathbf{z}), ..., \mathrm{sgn}(\tau_n^\top \mathbf{z}))^\top \text{ for } \tau_1, ..., \tau_n \sim \mathcal{N}(0, \mathbf{I}_d)$$

- Gaussian Lambda Coefficients

MSE of Bipolar HRF Estimators

Theorem 3.1 (MSE of the bipolar hybrid estimator)**.** *Take the bipolar hybrid estimator* $\widehat{\mathrm{SM}}_{m,n}^{\mathrm{hyb}}(\mathbf{x},\mathbf{y})$*, where* $\widehat{\mathrm{SM}}_{m}^{\mathrm{trig}}(\mathbf{x},\mathbf{y})$ *and* $\widehat{\mathrm{SM}}_{m}^{++}(\mathbf{x},\mathbf{y})$ *are chosen independently i.e. their random projections are chosen independently (note that we always assume that* $\widehat{\lambda}_n(\mathbf{x},\mathbf{y})$ *is constructed independently from* $\widehat{\mathrm{SM}}_{m}^{\mathrm{trig}}(\mathbf{x},\mathbf{y})$ *and* $\widehat{\mathrm{SM}}_{m}^{++}(\mathbf{x},\mathbf{y})$*). Then the following holds:*

$$\mathrm{MSE}(\widehat{\mathrm{SM}}_{m,n}^{\mathrm{hyb}}(\mathbf{x},\mathbf{y})) = \mathbb{E}[\widehat{\lambda}_n^2(\mathbf{x},\mathbf{y})]\mathrm{MSE}(\widehat{\mathrm{SM}}_{m}^{++}(\mathbf{x},\mathbf{y})) + \mathbb{E}[(1-\widehat{\lambda}_n(\mathbf{x},\mathbf{y}))^2]\mathrm{MSE}(\widehat{\mathrm{SM}}_{m}^{\mathrm{trig}}(\mathbf{x},\mathbf{y}))$$

MSE of Bipolar HRF Estimators

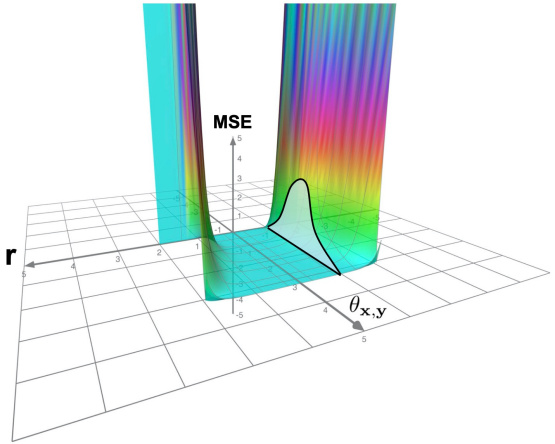
Theorem 3.1 (MSE of the bipolar hybrid estimator)**.** *Take the bipolar hybrid estimator* $\widehat{\mathrm{SM}}_{m,n}^{\mathrm{hyb}}(\mathbf{x},\mathbf{y})$*, where* $\widehat{\mathrm{SM}}_{m}^{\mathrm{trig}}(\mathbf{x},\mathbf{y})$ *and* $\widehat{\mathrm{SM}}_{m}^{++}(\mathbf{x},\mathbf{y})$ *are chosen independently i.e. their random projections are chosen independently (note that we always assume that* $\widehat{\lambda}_n(\mathbf{x},\mathbf{y})$ *is constructed independently from* $\widehat{\mathrm{SM}}_{m}^{\mathrm{trig}}(\mathbf{x},\mathbf{y})$ *and* $\widehat{\mathrm{SM}}_{m}^{++}(\mathbf{x},\mathbf{y})$*). Then the following holds:*

$$\mathrm{MSE}(\widehat{\mathrm{SM}}_{m,n}^{\mathrm{hyb}}(\mathbf{x},\mathbf{y})) = \mathbb{E}[\widehat{\lambda}_n^2(\mathbf{x},\mathbf{y})]\mathrm{MSE}(\widehat{\mathrm{SM}}_{m}^{++}(\mathbf{x},\mathbf{y})) + \mathbb{E}[(1-\widehat{\lambda}_n(\mathbf{x},\mathbf{y}))^2]\mathrm{MSE}(\widehat{\mathrm{SM}}_{m}^{\mathrm{trig}}(\mathbf{x},\mathbf{y}))$$

Lemma 3.2. *For the angular hybrid estimator the following holds:*

$$\mathbb{E}[\widehat{\lambda}_n^2(\mathbf{x},\mathbf{y})] = \frac{\theta_{\mathbf{x},\mathbf{y}}}{\pi}\left(\frac{\theta_{\mathbf{x},\mathbf{y}}}{\pi} - \frac{\theta_{\mathbf{x},\mathbf{y}}}{n\pi} + \frac{1}{n}\right), \mathbb{E}[\widehat{\lambda}_n(\mathbf{x},\mathbf{y})] = \frac{\theta_{\mathbf{x},\mathbf{y}}}{\pi}.$$

The variance of the angular hybrid estimator is zero for both $\theta_{\mathbf{x},\mathbf{y}} = 0$ and $\theta_{\mathbf{x},\mathbf{y}} = \pi$ if inputs $\mathbf{x}, \mathbf{y}$ have the same length.



Max Relative Error for Angular Hybrid Estimator

Definition 3.3 (Relative Error). *Denote by* $S(r)$ *a sphere centered at* 0 *and of radius* r. *For* $\mathbf{x}, \mathbf{y} \in S(r)$ *and such that* $\theta = \theta_{\mathbf{x},\mathbf{y}}$, *define* $\epsilon_{\theta,r}(\widehat{\mathrm{SM}}) \stackrel{\mathrm{def}}{=} \epsilon_{\mathbf{x},\mathbf{y}}(\widehat{\mathrm{SM}}) = \frac{\sqrt{\mathrm{MSE}(\widehat{\mathrm{SM}}(\mathbf{x},\mathbf{y}))}}{\mathrm{SM}(\mathbf{x},\mathbf{y})}$.

Max Relative Error for Angular Hybrid Estimator

Definition 3.3 (Relative Error). *Denote by $S(r)$ a sphere centered at 0 and of radius r. For $\mathbf{x},\mathbf{y} \in S(r)$ and such that $\theta = \theta_{\mathbf{x},\mathbf{y}}$, define* $\epsilon_{\theta,r}(\widehat{\mathrm{SM}}) \stackrel{\mathrm{def}}{=} \epsilon_{\mathbf{x},\mathbf{y}}(\widehat{\mathrm{SM}}) = \frac{\sqrt{\mathrm{MSE}(\widehat{\mathrm{SM}}(\mathbf{x},\mathbf{y}))}}{\mathrm{SM}(\mathbf{x},\mathbf{y})}$.

Lemma 3.4. *The following holds:*

$$\epsilon_{\theta,r}(\widehat{\mathrm{SM}}_m^{\mathrm{trig}}) = \epsilon_{\pi-\theta,r}(\widehat{\mathrm{SM}}_m^{++}) = \frac{1}{\sqrt{2m}} \exp(2r^2 \sin^2(\frac{\theta}{2})) \left(1 - \exp(-4r^2 \sin^2(\frac{\theta}{2}))\right),$$

and consequently for $W(r) = \exp(2r^2)\left(1 - \exp(-4r^2)\right)$:

$$\epsilon_{S(r)}(\widehat{\mathrm{SM}}_m^{\mathrm{trig}}) = \epsilon_{S(r)}(\widehat{\mathrm{SM}}_m^{++}) = \lim_{\theta\to\pi} \epsilon_{\theta,r}(\widehat{\mathrm{SM}}_m^{\mathrm{trig}}) = \lim_{\theta\to 0} \epsilon_{\theta,r}(\widehat{\mathrm{SM}}_m^{++}) = \sqrt{\frac{1}{2m} W(r)}$$

Max Relative Error for Angular Hybrid Estimator

Definition 3.3 (Relative Error). *Denote by $S(r)$ a sphere centered at 0 and of radius r. For $\mathbf{x},\mathbf{y} \in S(r)$ and such that $\theta = \theta_{\mathbf{x},\mathbf{y}}$, define* $\epsilon_{\theta,r}(\widehat{\mathrm{SM}}) \stackrel{\mathrm{def}}{=} \epsilon_{\mathbf{x},\mathbf{y}}(\widehat{\mathrm{SM}}) = \frac{\sqrt{\mathrm{MSE}(\widehat{\mathrm{SM}}(\mathbf{x},\mathbf{y}))}}{\mathrm{SM}(\mathbf{x},\mathbf{y})}$.

Lemma 3.4. *The following holds:*

$$\epsilon_{\theta,r}(\widehat{\mathrm{SM}}_m^{\mathrm{trig}}) = \epsilon_{\pi-\theta,r}(\widehat{\mathrm{SM}}_m^{++}) = \frac{1}{\sqrt{2m}} \exp(2r^2 \sin^2(\frac{\theta}{2})) \left(1 - \exp(-4r^2 \sin^2(\frac{\theta}{2}))\right),$$

and consequently for $W(r) = \exp(2r^2)\left(1 - \exp(-4r^2)\right)$:

$$\epsilon_{S(r)}(\widehat{\mathrm{SM}}_m^{\mathrm{trig}}) = \epsilon_{S(r)}(\widehat{\mathrm{SM}}_m^{++}) = \lim_{\theta\to\pi} \epsilon_{\theta,r}(\widehat{\mathrm{SM}}_m^{\mathrm{trig}}) = \lim_{\theta\to 0} \epsilon_{\theta,r}(\widehat{\mathrm{SM}}_m^{++}) = \sqrt{\frac{1}{2m}} W(r)$$

Theorem 3.5. *The max-relative-error of the angular hybrid estimator for the inputs $\mathbf{x},\mathbf{y}$ on the sphere $S(r)$ of radius $r \geq 1$ satisfies for $W(r) = \exp(2r^2)\left(1 - \exp(-4r^2)\right)$:*

$$\epsilon_{S(r)}(\widehat{\mathrm{SM}}_{m,n}^{\mathrm{anghyb}}) \leq \frac{1}{r}\sqrt{\frac{1}{2m}} W(r) \sqrt{\frac{1}{\pi} - \frac{1}{n\pi} + \frac{1}{n\sqrt{\pi}}}$$

Furthermore, $\lim_{\theta\to 0} \frac{\epsilon_{\theta,r}(\widehat{\mathrm{SM}}_{m,n}^{\mathrm{anghyb}})}{\sqrt{\theta}} = \lim_{\theta\to\pi} \frac{\epsilon_{\theta,r}(\widehat{\mathrm{SM}}_{m,n}^{\mathrm{anghyb}})}{\sqrt{\theta-\pi}} = \sqrt{\frac{1}{2\pi mn}} W(r).$

Pointwise Softmax Kernel Estimation

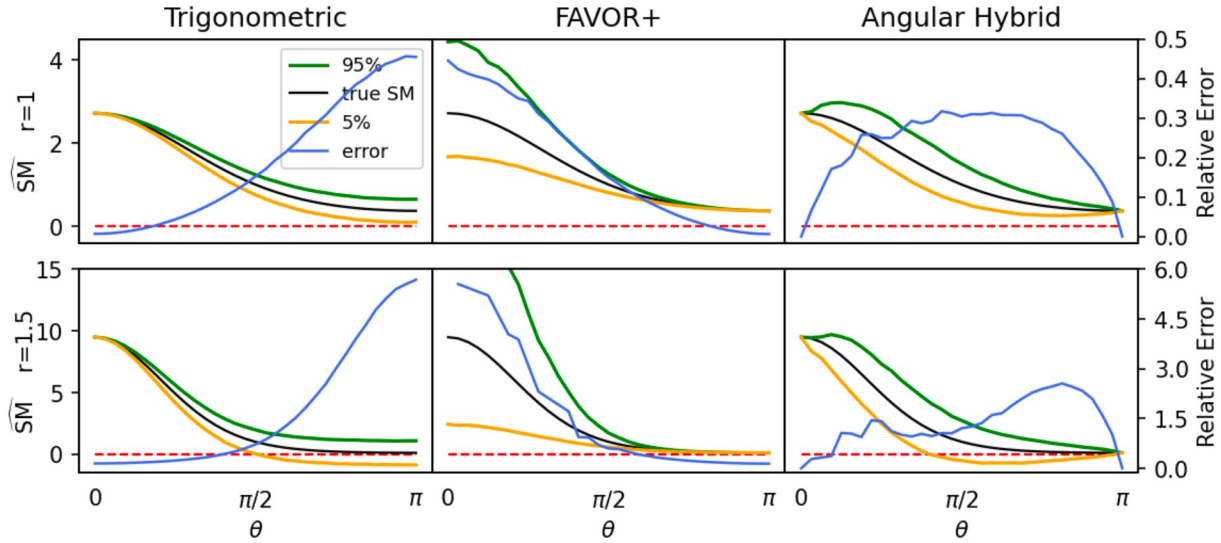


Figure 2: Pointwise estimation of $\mathrm{SM}(\mathbf{x}, \mathbf{y})$ for the same-length 64-dim inputs ($r = 1.0$ and $r = 1.5$) and various angles $\theta_{\mathbf{x},\mathbf{y}}$. Red-dotted lines are for marking zero-level. We used $s = 10000$ estimated softmax values in each subplot. The true value and the 5^{th} and 95^{th} quantile estimated values are shown by the left y-axis, and the empirical relative errors are shown by the right y-axis. Trigonometric estimator and FAVOR+ applied 128 random features. To make fair comparison, for the hybrid variant the configuration leading to the similar number of FLOPS operations per random feature map creation was applied. Similar gains as for the angular are obtained by the Gaussian hybrid variant.

Experiment 1: Language Modeling

- Language modeling task (Rawat et al., 2019)

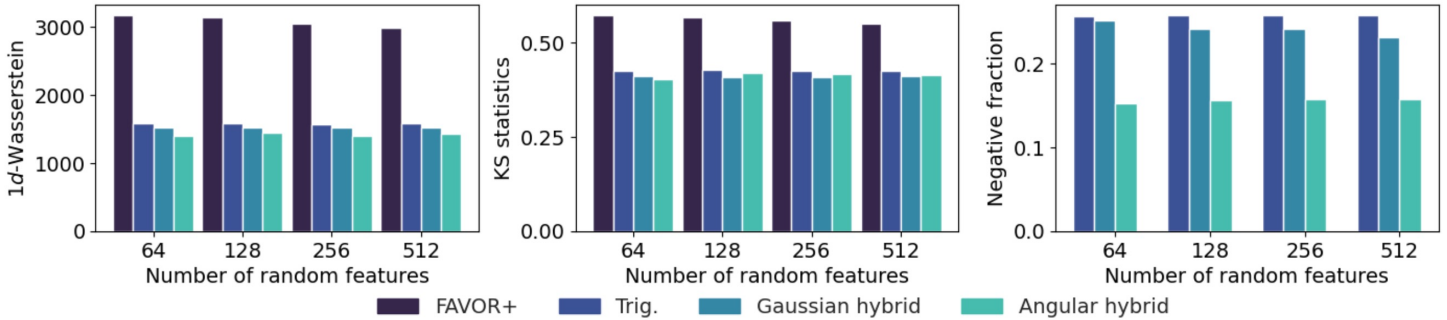


Figure 3: Statistical metrics measuring softmax matrix approximation quality on PennTree Bank. For standard estimators, the number of random features are $64, 128, 256, 512$. To make fair comparison, for the hybrid variants, the configurations leading to the similar number of FLOPS operations per random feature map creation were applied. Negative fractions were not reported for FAVOR+ since by definition they are equal to zero.

Experiment 2: Training Speech Models with HRF-Conformers-Performers

- Speech models with LibriSpeech ASR corpus (Panayotov et al., 2015)

Table 2: Comparison of WERs of Conformer-Transducer applying different RF-mechanisms for the implicit attention. For methods other than clustering-based HRFs (HRF-C), numbers next to method names define the values of m or (m, n). Method HRF-A stands for the angular hybrid variant. Numbers next to HRF-C correspond to the number of clusters constructed in the query and key space respectively. HRF-C uses 64 random features. We also report standard deviations averaged over 10 different training runs.

	HRF-C(3,3)	HRF-C(2, 3)	HRF-C(3,2)	HRF-C(2,2)	HRF-A(16,8)
WER	$\mathbf{1.72 \pm 0.02}\%$	$1.75 \pm 0.03\%$	$1.83 \pm 0.03\%$	$1.85 \pm 0.04\%$	$2.03 \pm 0.08\%$
	HRF-A(8, 8)	FAVOR+ 432	FAVOR+ 256	Trig 432	Trig 256
WER	$2.05 \pm 0.05\%$	$2.65 \pm 0.06\%$	$2.77 \pm 0.04\%$	$3.12 \pm 0.05\%$	$3.3 \pm 0.06\%$

Experiment 3: Downstream Robotics Experiments

- Step-stone locomotion and robotic-arm manipulation tasks (Choromanski et al., 2021a)

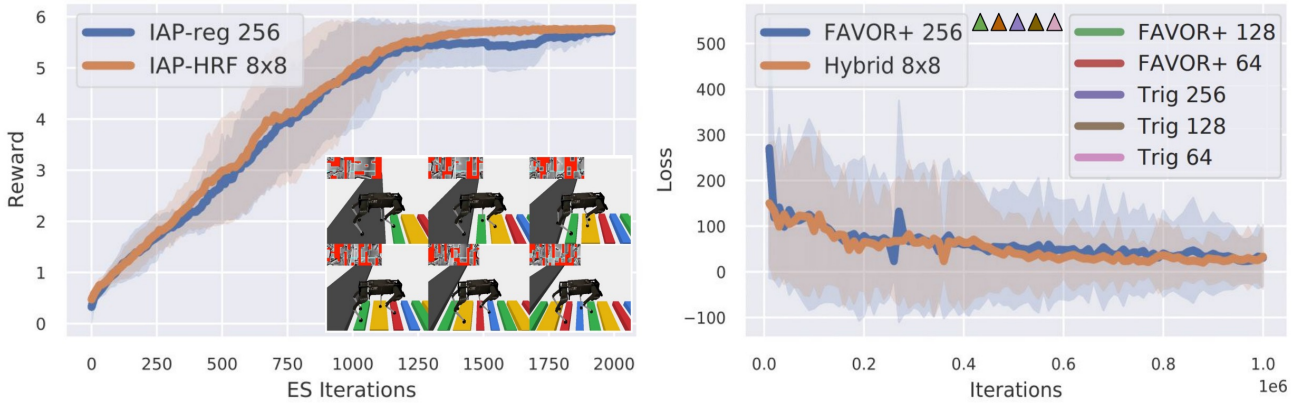


Figure 4: **Left**: Step-stone locomotion task. Comparison of the training curves for: the IAP using angular hybrid estimator of $m = n = 8$ and IAP applying regular FAVOR+ mechanism from (Choromanski et al., 2021b) with $m = 256$. Both final policies are of similar quality, yet HRF-method requires **3x+** fewer FLOPS to run its trained policy. The visualization of the HRF policy in action and its attention (with filtered out pixels marked in red) is in the bottom right corner. **Right:** Similar setting (and conclusions) but for the robotic-arm manipulation task. The additional five regular RF-configurations did not train by producing Nan loss due to large variance of the underlying softmax kernel estimators.