The Combined Softmax Estimator

Neil Kachappilly, Makoto Powers

December 2024

Abstract

Building on previous work in hybrid random features, we extend the work on angular hybrid models (AHMs), specifically designed for linearizing angular kernels in restricted settings where all input vectors share the same magnitude. AHMs leverage hybrid random feature constructions to adaptively optimize the quality of kernel approximations, providing accurate estimates within well-defined regions of interest. By specializing to angular kernels, AHMs extend Bochner's Theorem for restricted-magnitude inputs and refine compositional random feature mechanisms to achieve superior performance. These extensions offer strong theoretical guarantees, including unbiased kernel approximations and provably smaller worst-case relative errors compared to existing methods. Specifically, we look to utilize the abilities of AHMs to combine existing softmax kernel estimators to form a hybrid estimator that represents the best of its parts. Through extensive experiments, we validate the effectiveness of AHMs in this task, showcasing their capabilities in pointwise kernel estimation. The results highlight AHMs' ability to deliver high-quality approximations and robustness in constrained angular kernel environments.

1 Introduction & Related Work

Consider the attention kernel $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ and two of its estimators $\widehat{K}_1(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x})\phi(\mathbf{y})^T$ and $\widehat{K}_2(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x})\psi(\mathbf{y})^T$ for specific randomized maps such that the former estimator is very accurate for estimating small kernel values, corresponding to large angles between the input vectors, and the latter for estimating large kernel values, corresponding to small angles between input vectors. The classic attention kernel is the softmax kernel and the corresponding classic random feature map mechanism is the trigonometric feature map, obtained from Bochner's Theorem applied to the Gaussian kernel (Rahimi & Recht, 2007):

$$\phi_m^{trig}(\mathbf{u}) = \frac{1}{\sqrt{m}} \exp(\frac{||\mathbf{u}||}{2})(\sin(\omega_1^T \mathbf{u}), \ldots, \sin(\omega_1^T \mathbf{u}), \cos(\omega_1^T)\mathbf{u}), \ldots, \cos(\omega_m^T \mathbf{u}))^T \tag{1}$$

where m stands for the number of random features and $\omega_1, \ldots, \omega_m \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{I}_d)$.
While a commonly used method for linearizing the softmax kernel, it has been shown to fail in some important applications of scalable kernel methods, such as implicit-attention Transformer architectures called Performers (Choromanski et al., 2021b). Notably, the given mechanism results in high variance as kernel values get close to zero, a key issue considering that many attention matrices' entries in Transformers' models are very small and any kernel estimator needs to be accurate here in order to be practically implemented. Thus, this estimator falls under the category of $\widehat{K}_2$ as described above. The solution, fitting our $\widehat{K}_1$, was presented via an alternate random feature map mechanism, named FAVOR+ (Choromanski et al., 2021b). Here, a new positive random feature map for unbiased softmax kernel estimation was defined as:

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

While this solves the issue of estimating small softmax kernel, it also gives rise to a new issue: high variance for larger kernel values. Thus, while crucial to improving the performance of random features in Transformers

training, the new positive random features mechanism fails in the crucial cases of large entries in the attention matrices of Transformers. Such a scenario calls for the development of a new estimator, one that can combine the best from both worlds, as being particularly accurate for both small and large kernel values regime.
Through this paper, we hope to demonstrate that this is indeed possible as we present the *combined softmax estimator* or CSE. Our approach entails designing a third estimator defined as a linear combination of both the previously defined estimators, producing high accuracy and low variance in both extreme regions of interest. We provide detailed theoretical analysis of our estimator, explaining the key logic behind the coefficient of the linear combination and verifying the reduced variance with formulas for the mean squared error (MSE). Note that the MSE can be interpreted as equivalent to variance for the purposes of this paper since we are stricly limiting our scope to unbiased estimators. Further, we will make the assumption that our kernel acts on vectors taken from the sphere of given radius R.
Related Work: While there is significant literature on the different random feature map mechanisms for softmax kernel estimation, our work is primarily related to existing work on trigonometric random features (Rahimi & Recht, 2007), positive random features (Choromanski et al., 2021b), and hybrid random features (Choromanski et al., 2022). We aim to build on the work on trigonometric and positive random features by best utilizing both, constructing a hybrid akin to the special case presented in the hybrid random features paper. We narrow our scope to just this case and provide theoretical and empirical results supporting its advantage over using the estimators individually, with theoretical guarantees for vectors drawn from the sphere with radius r.

2 Combined Softmax Estimator

2.1 Key Definitions

We start by introducing some basic definitions and results inherited from previous work, crucial to understanding the presentation of our work to follow.
Definition 2.1 (Kernel with a Random Feature Map Representation). *A kernel function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ admits a random feature (RF) map representation if it can be written as*

$$K(x, y) = \mathbb{E}_{\omega \sim \Omega} \sum_{i=1}^{l} [\xi_i(\mathbf{x}, \omega)\xi_i(\mathbf{y}, \omega] \tag{3}$$

for some $\xi_i : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, and where ω is sampled from some probabilistic distribution $\Omega \in \mathbb{P}(\mathbb{R}^d)$. The corresponding random feature map, for a given $m \in \mathbb{N}$, is defined as

$$\phi_m(\mathbf{u}) = \frac{1}{\sqrt{m}} \phi_m^1(\mathbf{u}) \star \cdots \star \phi_m^l(\mathbf{u}) \in \mathbb{R}^{ml} \tag{4}$$

where $\phi_m^i = (\xi_i(\mathbf{u}, \omega_1), \ldots, \xi_i(\mathbf{u}, \omega_m))^T$, $\star$ stands for vertical concatenation, and $\omega_1, \ldots, \omega_m \overset{iid}{\sim} \Omega$.
Random feature maps can then be used as an unbiased estimator for the corresponding kernel using the below equation:

$$\widehat{K}(\mathbf{x}, \mathbf{y}) = \phi_m(\mathbf{x})^T \phi_m(\mathbf{y}) \tag{5}$$

We then define the trigonometric random feature ϕ_m^{trig} and the positive random features ϕ_m^+ by plugging into equation 4. ϕ_m^{trig} can be obtained by applying $l = 2, \xi_1(\mathbf{u}, \omega) = \sin(\omega^T \mathbf{u}), \xi_2(\mathbf{u}, \omega) = \cos(\omega^T \mathbf{u})$. As for ϕ_m^+, we can take $l = 2, \xi_1(\mathbf{u}, \omega) = \frac{1}{\sqrt{2}} \exp(\omega^T \mathbf{u}), \xi_2(\mathbf{u}, \omega) = \frac{1}{\sqrt{2}} \exp(-\omega^T \mathbf{u})$.
From (Choromanski et al., 2021b), we also have the MSE for both of these estimators, confirming our earlier claims. These results show that the trigonometric estimator has small MSE for large kernel values and large MSE for small kernel values while the opposite is true for the positive random features estimator. Let $\widehat{\mathrm{SM}}_m^{trig}(\mathbf{x}, \mathbf{y})$ be the trigonometric kernel estimator and $\widehat{\mathrm{SM}}_m^{++}(\mathbf{x}, \mathbf{y})$ be the estimator applying positive

random features. Denote: $\mathbf{z} = \mathbf{x} + \mathbf{y}$ and $\Delta = \mathbf{x} - \mathbf{y}$.
Lemma 2.2 For $\omega \sim \mathcal{N}(0, \mathbf{I}_d)$, *the following is true*:

$$\mathrm{MSE}(\widehat{\mathrm{SM}}_m^{trig}(\mathbf{x}, \mathbf{y})) = \frac{1}{2m} \exp(||\mathbf{z}||^2)\mathrm{SM}^{-2}(\mathbf{x}, \mathbf{y})(1 - (\exp(-||\Delta||^2))^2 \tag{6}$$

$$\mathrm{MSE}(\widehat{\mathrm{SM}}_m^{++}(x, y)) = \frac{1}{m} \exp(\|\mathbf{z}\|^2)\mathrm{SM}^2(\mathbf{x}, \mathbf{y})(1 - \exp(-\|\mathbf{z}\|^2)) \tag{7}$$

Thus, for $\mathrm{SM}(\mathbf{x}, \mathbf{y}) \to 0$ we have: $\mathrm{MSE}(\widehat{\mathrm{SM}}_m^{\mathrm{trig}}(\mathbf{x}, \mathbf{y})) \to \infty$ and $\mathrm{MSE}(\widehat{\mathrm{SM}}_m^{+}(\mathbf{x}, \mathbf{y})) \to 0$, consistent with our expectations.

2.2 The Combined Softmax Estimator

We are now ready to present our method of combining the two estimators discussed thus far into a third estimator that outperforms both $\widehat{\mathrm{SM}}^{+}$ and $\widehat{\mathrm{SM}}^{\mathrm{trig}}$ in our restricted setting with input vectors drawn from the sphere of radius R. From our previous work, we know that $\widehat{\mathrm{SM}}_m^{++}$ becomes perfect for $\theta = \pi$ and $\widehat{\mathrm{SM}}_m^{\mathrm{trig}}$ becomes perfect for $\theta = 0$. Thus, we choose a simple linear dependence of λ on θ to guarantee vanishing variance for both the critical values: $\theta = 0$ and $\theta = \pi$. Let $\alpha_{\mathbf{x},\mathbf{y}}$ be the angle between input vectors $\mathbf{x}, \mathbf{y}$. From the Goemans-Williamson algorithm (Goemans & Williamson, 2004)

$$1 - \frac{2\alpha_{\mathbf{x},\mathbf{y}}}{\pi} = \mathbb{E}[\tau(\mathbf{x})\tau(\mathbf{y}^T)] \tag{8}$$

with the definition

$$\tau(\mathbf{x}) = \frac{1}{\sqrt{m}}(\mathrm{sgn}(\omega_1 \mathbf{z}^T), \ldots, \mathrm{sgn}(\omega_m \mathbf{z}^T)) \tag{9}$$

It follows that

$$\frac{\widehat{\alpha_{\mathbf{x},\mathbf{y}}}}{\pi} = \frac{1}{2} - \frac{1}{2}\tau(\mathbf{x})\tau(\mathbf{y}^T) \tag{10}$$

is an unbiased estimator of $\alpha_{\mathbf{x},\mathbf{y}}$ (details in Appendix). We define the combined Softmax estimator $\widehat{\mathrm{SM}}^{\mathrm{hyb}}$ as a linear combination of the $\widehat{\mathrm{SM}}^{+}$ and $\widehat{\mathrm{SM}}^{\mathrm{trig}}$, with the coefficient of the linear combination being an affine function of the angle between input vectors $\mathbf{x}, \mathbf{y}$ and the magnitude of the input vectors $||\mathbf{x}|| = ||\mathbf{y}|| = r$, so with the notation $\lambda(\mathbf{x}, \mathbf{y}) = \frac{\alpha_{\mathbf{x},\mathbf{y}}}{\pi}$ and $\hat{\lambda}(\mathbf{x}, \mathbf{y}) = \frac{\widehat{\alpha_{\mathbf{x},\mathbf{y}}}}{\pi}$ we have $\widehat{\mathrm{SM}}_{m,n}^{\mathrm{hyb}}$:

$$\widehat{\mathrm{SM}}_{m,n}^{\mathrm{hyb}}(\mathbf{x}, \mathbf{y}) = \hat{\lambda}_n(\mathbf{x}, \mathbf{y})\widehat{\mathrm{SM}}_m^{++}(\mathbf{x}, \mathbf{y}) + \big(1 - \hat{\lambda}_n(\mathbf{x}, \mathbf{y})\big)\widehat{\mathrm{SM}}_m^{\mathrm{trig}}(\mathbf{x}, \mathbf{y}) \tag{11}$$

As we show in section 3 and in our experiments, this estimator is accurate in approximating both small and large softmax kernel values. For our restricted case when input vectors are of fixed length, its variance is zero for both the smallest ($\theta_{\mathbf{x},\mathbf{y}} = \pi$) and largest ($\theta_{\mathbf{x},\mathbf{y}} = 0$) softmax kernel values (see: Fig. 1).

3 Theoretical Guarantees

Theorem 3.1 (MSE of the bipolar hybrid estimator). Consider 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 $\hat{\lambda}_n(\mathbf{x}, \mathbf{y})$ is constructed independently from $\widehat{\mathrm{SM}}_m^{\mathrm{trig}}(\mathbf{x}, \mathbf{y})$ and $\widehat{\mathrm{SM}}_m^{++}(\mathbf{x}, \mathbf{y})$). Then we have:

$$\mathrm{MSE}(\widehat{\mathrm{SM}}_{m,n}^{\mathrm{hyb}}(\mathbf{x}, \mathbf{y})) = \mathbb{E}[\hat{\lambda}_n^2(\mathbf{x}, \mathbf{y})]\mathrm{MSE}(\widehat{\mathrm{SM}}_m^{++}(\mathbf{x}, \mathbf{y})) + \mathbb{E}[(1 - \hat{\lambda}_n(\mathbf{x}, \mathbf{y}))^2]\mathrm{MSE}(\widehat{\mathrm{SM}}_m^{\mathrm{trig}}(\mathbf{x}, \mathbf{y})) \tag{12}$$

Furthermore, if $\widehat{\mathrm{SM}}_m^{\mathrm{trig}}(\mathbf{x},\mathbf{y})$ and $\widehat{\mathrm{SM}}_m^{++}(\mathbf{x},\mathbf{y})$ apply the exact same sets of random projections, the mean squared error of the hybrid estimator is further reduced. In particular:

$$\mathrm{MSE}(\widehat{\mathrm{SM}}_{m,n}^{\mathrm{hyb}}(\mathbf{x},\mathbf{y})) = \mathbb{E}[\hat{\lambda}_n^2(\mathbf{x},\mathbf{y})]\mathrm{MSE}(\widehat{\mathrm{SM}}_m^{++}(\mathbf{x},\mathbf{y})) \tag{13}$$

$$+ \mathbb{E}[(1-\hat{\lambda}_n(\mathbf{x},\mathbf{y}))^2]\mathrm{MSE}(\widehat{\mathrm{SM}}_m^{\mathrm{trig}}(\mathbf{x},\mathbf{y})) \tag{14}$$

$$- \frac{2}{m}\mathrm{SM}^2(\mathbf{x},\mathbf{y})(1-\cos(\|\mathbf{x}\|_2^2 - \|\mathbf{y}\|_2^2))\mathbb{E}[\hat{\lambda}_n(\mathbf{x},\mathbf{y})(1-\hat{\lambda}_n(\mathbf{x},\mathbf{y}))] \tag{15}$$

The exact formula on $\mathrm{MSE}(\widehat{\mathrm{SM}}_m^{++}(\mathbf{x},\mathbf{y}))$ and $\mathrm{MSE}(\widehat{\mathrm{SM}}_m^{\mathrm{trig}}(\mathbf{x},\mathbf{y}))$ is given in Lemma 2.2.

Lemma 3.2. For our combined softmax hybrid estimator the following holds:

$$\mathbb{E}[\widehat{\lambda}_n^2(\mathbf{x},\mathbf{y})] = \frac{\theta_{\mathbf{x},\mathbf{y}}}{\pi}\left(\frac{\theta_{\mathbf{x},\mathbf{y}}}{\pi} - \frac{\theta_{\mathbf{x},\mathbf{y}}}{n\pi} + \frac{1}{n}\right) \quad \mathbb{E}[\widehat{\lambda}_n(\mathbf{x},\mathbf{y})] = \frac{\theta_{\mathbf{x},\mathbf{y}}}{\pi} \tag{16}$$

We observe again that 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. We introduce one more definition.

Definition 3.3. Assume that the inputs to the estimators are taken from some given bounded set $C \subseteq \mathbb{R}^d$. For a given estimator $\widehat{\mathrm{SM}}$ on feature vectors $\mathbf{x},\mathbf{y} \in C$, we define its max-relative-error with respect to C as

$$\epsilon_C(\widehat{\mathrm{SM}}) = \max_{\mathbf{x},\mathbf{y}\in C} \epsilon_{\mathbf{x},\mathbf{y}}(\widehat{\mathrm{SM}}) \tag{17}$$

where

$$\epsilon_{\mathbf{x},\mathbf{y}}(\widehat{\mathrm{SM}}) = \frac{\sqrt{\mathrm{MSE}(\widehat{\mathrm{SM}}(\mathbf{x},\mathbf{y}))}}{\widehat{\mathrm{SM}}(\mathbf{x},\mathbf{y})} \tag{18}$$

Let by $S(r)$ a sphere centered at 0 of radius r. Define

$$\epsilon_{\theta,r}(\widehat{\mathrm{SM}}) \stackrel{\mathrm{def}}{=} \epsilon_{\mathbf{x},\mathbf{y}}(\widehat{\mathrm{SM}}) \tag{19}$$

for $\mathbf{x},\mathbf{y} \in S(r)$ and such that $\theta = \theta_{\mathbf{x},\mathbf{y}}$ (note that the mean squared errors of the considered estimators depend only on the angle $\theta_{\mathbf{x},\mathbf{y}}$ for $\mathbf{x},\mathbf{y}$ chosen from a fixed sphere).

Lemma 3.4. The following holds:

$$\epsilon_{\theta,r}\left(\widehat{\mathrm{SM}}_m^{\mathrm{trig}}\right) = \epsilon_{\pi-\theta,r}\left(\widehat{\mathrm{SM}}_m^{++}\right) = \frac{1}{\sqrt{2m}}\exp\left(2r^2\sin^2\left(\frac{\theta}{2}\right)\right)\left(1-\exp\left(-4r^2\sin^2\left(\frac{\theta}{2}\right)\right)\right) \tag{20}$$

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

$$\epsilon_{S(r)}\left(\widehat{\mathrm{SM}}_m^{\mathrm{trig}}\right) = \epsilon_{S(r)}\left(\widehat{\mathrm{SM}}_m^{++}\right) = \lim_{\theta\to\pi}\epsilon_{\theta,r}\left(\widehat{\mathrm{SM}}_m^{\mathrm{trig}}\right) = \lim_{\theta\to 0}\epsilon_{\theta,r}\left(\widehat{\mathrm{SM}}_m^{++}\right) = \sqrt{\frac{1}{2m}W(r)} \tag{21}$$

In the next theorem, we show that the max-relative-error of the angular hybrid estimator scales as $\frac{1}{r}\exp(2r^2)$ in the length r of its inputs as opposed to $\exp(2r^2)$ as it is the case for $\widehat{\mathrm{SM}}_m^{\mathrm{trig}}$ and $\widehat{\mathrm{SM}}_m^{++}$ (see: Lemma 3.4). Furthermore, the max-relative-error scales as $\sqrt{\theta}$ and $\sqrt{\pi-\theta}$ as $\theta \to 0$ and $\theta \to \pi$ respectively, in particular goes to 0 in both critical cases. This is not true for $\widehat{\mathrm{SM}}_m^{\mathrm{trig}}$ nor for $\widehat{\mathrm{SM}}_m^{++}$.

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

$$\epsilon_{S(r)}\left(\widehat{\mathrm{SM}}_{m,n}^{\mathrm{anghyb}}\right) \leq \frac{1}{r}\sqrt{\frac{1}{2m}W(r)}\sqrt{\frac{1}{\pi} - \frac{1}{n\pi} + \frac{1}{n\sqrt{\pi}}} \tag{22}$$

Furthermore,

$$\lim_{\theta\to 0}\frac{\epsilon_{\theta,r}\left(\widehat{\mathrm{SM}}_{m,n}^{\mathrm{anghyb}}\right)}{\sqrt{\theta}} = \lim_{\theta\to\pi}\frac{\epsilon_{\theta,r}\left(\widehat{\mathrm{SM}}_{m,n}^{\mathrm{anghyb}}\right)}{\sqrt{\pi-\theta}} = \sqrt{\frac{1}{2\pi mn}W(r)} \tag{23}$$

4 Experiments

In this section we conduct experiments to demonstrate the performance of the angular hybrid estimator against trigonometric random feature and FAVOR+ estimators. We measure the empirical relative errors of the different estimators for different angles $\theta_{\mathbf{x},\mathbf{y}}$. Note that unless mentioned otherwise, these experiments are conducted with 128 random features of dimensionality 64 each and the inputs being drawn from a sphere of size $r = 1$.

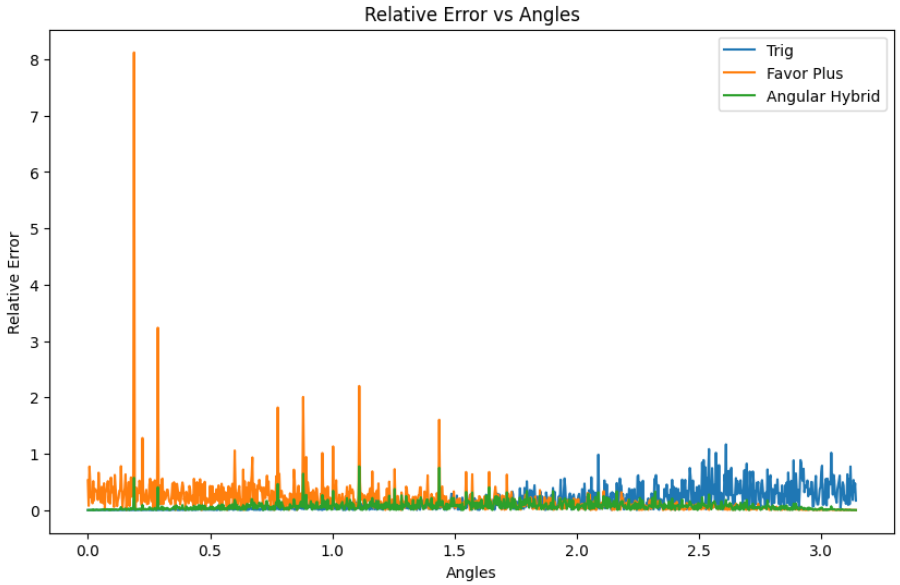

Figure 1: Relative Error of Estimation vs Angle for Trig, FAVOR+, and Angular Hybrid Estimators

We see that the combined Softmax estimator has lower relative error than $\widehat{\mathrm{SM}}^{\mathrm{trig}}$ and $\widehat{\mathrm{SM}}^{++}$ for all values of θ.

When varying the normed distance $||\mathbf{x}|| = ||\mathbf{y}|| = r$, we see that the decay in performance is consistent with our theoretical results. Notably, the relative error increases alongside the increasing radius r. This motivates further investigation into a coefficient for the linear combination that depends on both θ and r.

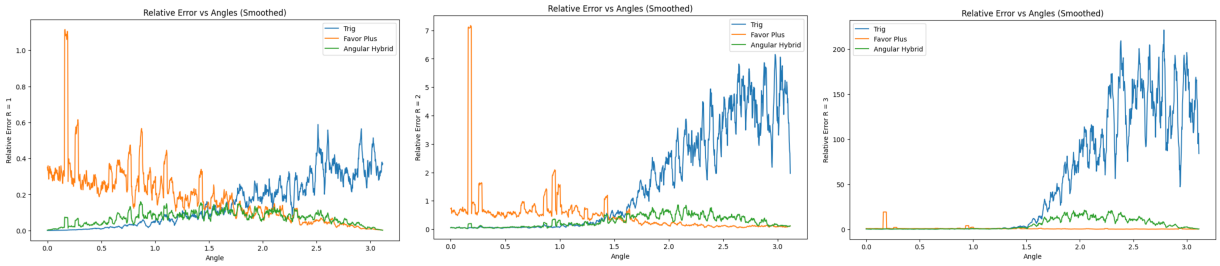

Figure 2: Relative Error of Estimation vs Angle for Trig, FAVOR+, and Angular Hybrid Estimators for $r = 1, 2, 3$.

We also note that the trig estimator exhibits significantly worsening performance at large angles than FAVOR+ does at small angles as r increases. It seems that the trig error scales up much larger as compared to the FAVOR+ and Hybrid. We also note that the FAVOR+ estimator appears to be most prone to outliers with especially egrergious outliers visible at r = 1 and r = 2. The hybrid is relatively immune to both of

these shortcomings of the other estimators, allowing for a significantly more robust mechanism.

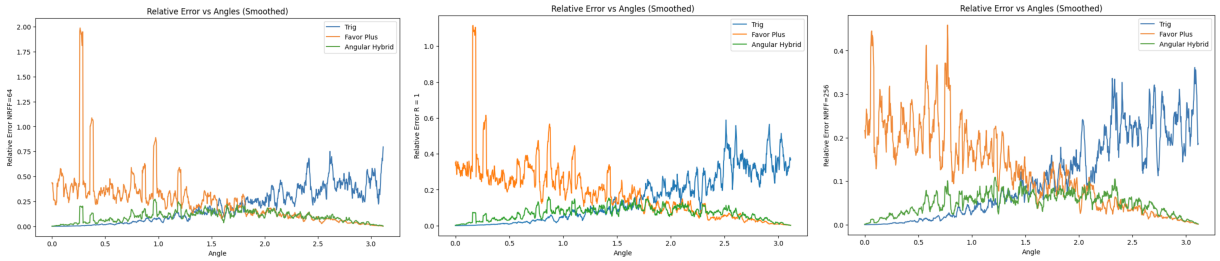

Figure 3: Relative Error of Estimation vs Angle for Trig, FAVOR+, and Angular Hybrid Estimators for $NRFF = 64, 128, 256$.

Our experiments also show that the number of random features does not seem to overly affect the distribution of the errors barring reducing the drastic outlier kernel score produced by the FAVOR+ algorithm for small angles, as visible in the graphs. With 256 features, the outlier is essentially eliminated.

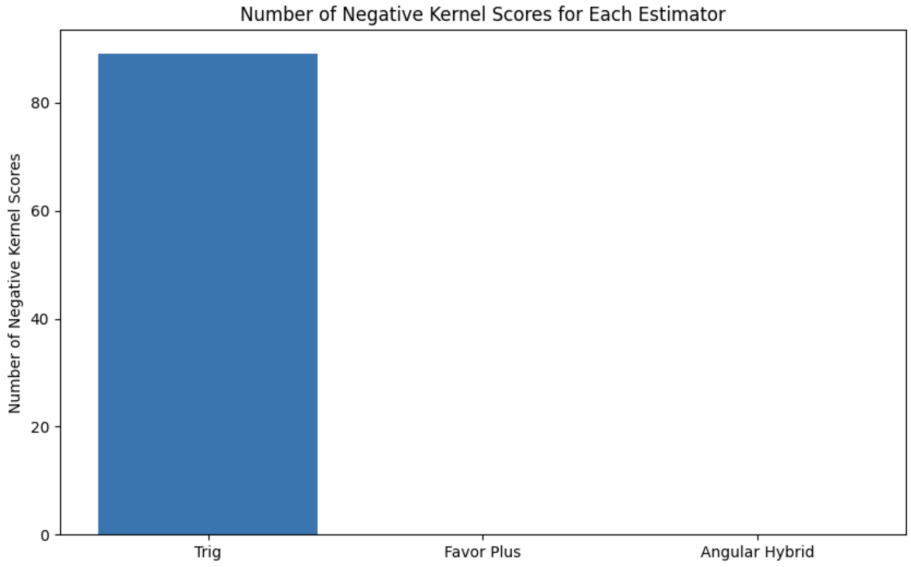

Figure 4: Number of Negative Kernel Scores for Trig, FAVOR+, and Angular Hybrid Estimators (10,000 estimations).

Another interesting to note during the course of the experimentation is the empirical results regarding the concern for the presence of negative kernel scores. Negative kernel scores are a major issue with the trigonometric estimator as applying feature maps with negative dimension values leads to unstable behaviors, especially when it comes to estimating kernel scores that are close to zero (large angles). Since the hybrid incorporates the trig estimator, it naturally follows that the same concern is in play but practical results show that even with 89 negative scores from the trig estimator through 10,000 iterations, the angular hybrid yields 0 negative scores, an indicator that this theoretical concern may be less relevant than expected in actual practice.

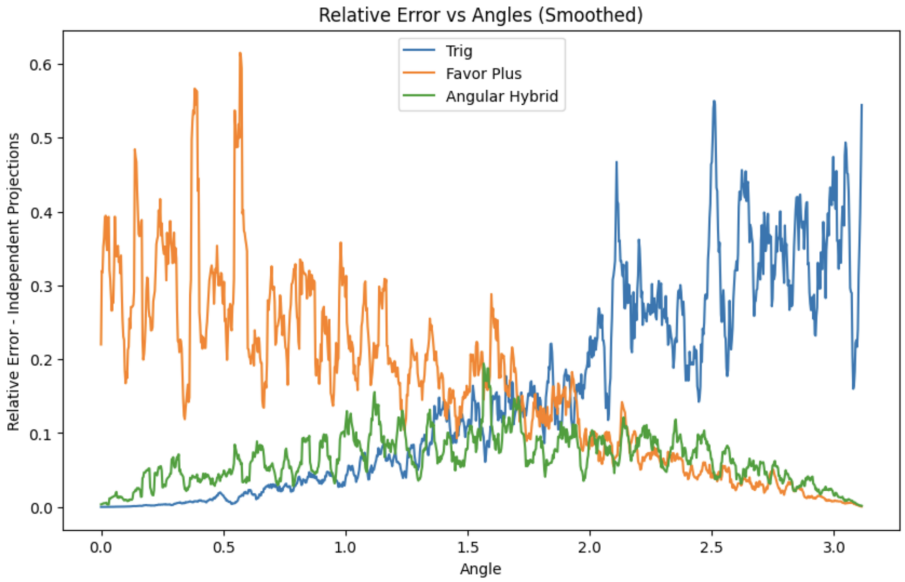

Figure 5: The Impact of Independent Random Projections for each Estimator

Finally, the previous experiments were all conducted using the same set of random projections for the trigonometric estimator, the FAVOR+ estimator, and in the angular kernel. To observe the effects of using independent sets of random projections instead, we assigned each estimator its own set, leading to minimal change in results. As shown above, the only notable change in the distribution of scores for each estimator is the removal of the outliers in the FAVOR+ kernel score distribution, similar to the distribution of the 256 random features case. This could warrant a deeper dive into the theoretical implications of using independent random projections vs. the same ones.

5 Conclusion

In this work, we extended the theory and application of angular hybrid models (AHMs) to address the challenges of linearizing angular kernels in scenarios where input vectors share the same magnitude. By leveraging hybrid random feature constructions and refining compositional random feature mechanisms, AHMs demonstrated the ability to adaptively optimize kernel approximations, offering precise and robust estimates within restricted regions of interest.

Through these advancements, we established strong theoretical guarantees, including unbiased kernel approximations and significantly reduced worst-case relative errors compared to existing methods. By extending Bochner's Theorem for restricted-magnitude inputs, AHMs further reinforced the theoretical foundation for angular kernel approximations, pushing the boundaries of kernel estimation techniques.

Our experimental results validated the efficacy of AHMs across a range of pointwise kernel estimation tasks, highlighting their capacity to deliver high-quality approximations and maintain robustness in constrained angular kernel environments. These findings demonstrate the potential of AHMs as a practical and efficient tool for tackling problems in machine learning, signal processing, and other domains reliant on angular similarity measures.

Future work can build on this foundation by exploring the applicability of AHMs in broader settings, such as anisotropic angular kernels or datasets with varying magnitudes, and extending their utility to large-scale applications.

References

[1] Y. Cho and L. K. Saul, "Analysis and extension of arc-cosine kernels for large margin classification," *Technical Report CS2012-0972*, Department of Computer Science and Engineering, University of California, San Diego, 2012.

[2] K. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlos, P. Hawkins, J. Davis, A. Mohiuddin, L. Kaiser, D. Belanger, L. Colwell, and A. Weller, "Rethinking Attention with Performers," 2022.

[3] K. M. Choromanski, H. Lin, H. Chen, A. Sehanobish, Y. Ma, D. Jain, J. Varley, A. Zeng, M. S. Ryoo, V. Likhosherstov, D. Kalashnikov, V. Sindhwani, and A. Weller, "Hybrid Random Features," in *International Conference on Learning Representations*, 2022.

[4] M. X. Goemans and D. P. Williamson, "Approximation algorithms for MAX-3-CUT and other problems via complex semidefinite programming," *J. Comput. Syst. Sci.*, vol. 68, no. 2, pp. 442–470, 2004.

[5] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3–6, 2007*, J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, Eds., pp. 1177–1184, Curran Associates, Inc., 2007.

[6] G. Blanc and S. Rendle, "Adaptive sampled softmax with kernel based sampling," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10–15, 2018*, J. G. Dy and A. Krause, Eds., vol. 80 of *Proceedings of Machine Learning Research*, pp. 589–598, PMLR, 2018.

6 Appendix

Proof of Theorem 3.1

The following holds:

$$\begin{aligned}\mathrm{Var}(\widehat{\mathrm{SM}}^{\mathrm{hyb}}_{m,n}(\mathbf{x},\mathbf{y})) &= \mathrm{Var}(\hat{\lambda}_n(\theta)\widehat{\mathrm{SM}}^{++}_m(\mathbf{x},\mathbf{y})) + \mathrm{Var}((1-\hat{\lambda}_n(\theta))\widehat{\mathrm{SM}}^{\mathrm{trig}}_m(\mathbf{x},\mathbf{y})) \\ &\quad + 2\mathrm{Cov}(\hat{\lambda}_n(\theta)\widehat{\mathrm{SM}}^{++}_m(\mathbf{x},\mathbf{y}),(1-\hat{\lambda}_n(\theta))\widehat{\mathrm{SM}}^{\mathrm{trig}}_m(\mathbf{x},\mathbf{y})) \end{aligned} \tag{24}$$

Focusing on the covariance term, and for the sake of simplicity let $\hat{\lambda} = \hat{\lambda}_n$. Assume first that $\widehat{\mathrm{SM}}^{\mathrm{trig}}_m(\mathbf{x},\mathbf{y})$ and $\widehat{\mathrm{SM}}^{++}_m(\mathbf{x},\mathbf{y})$ are chosen independently. Then we have:

$$\begin{aligned}\mathrm{Cov}(\hat{\lambda}(\theta)\widehat{\mathrm{SM}}^{++}_m(\mathbf{x},\mathbf{y}),(1-\hat{\lambda}(\theta))\widehat{\mathrm{SM}}^{\mathrm{trig}}_m(\mathbf{x},\mathbf{y})) &= \mathbb{E}[\hat{\lambda}(\theta)(1-\hat{\lambda}(\theta))\widehat{\mathrm{SM}}^{++}_m(\mathbf{x},\mathbf{y})\widehat{\mathrm{SM}}^{\mathrm{trig}}_m(\mathbf{x},\mathbf{y})] \\ &\quad - \mathbb{E}[\hat{\lambda}(\theta)\widehat{\mathrm{SM}}^{++}_m(\mathbf{x},\mathbf{y})]\mathbb{E}[(1-\hat{\lambda}(\theta))\widehat{\mathrm{SM}}^{\mathrm{trig}}_m(\mathbf{x},\mathbf{y})] \\ &= (\mathbb{E}[\hat{\lambda}(\theta)(1-\hat{\lambda}(\theta))] - \mathbb{E}[\hat{\lambda}(\theta)]\mathbb{E}[(1-\hat{\lambda}(\theta))])\mathbb{E}[\widehat{\mathrm{SM}}^{++}_m(\mathbf{x},\mathbf{y})]\mathbb{E}[\widehat{\mathrm{SM}}^{\mathrm{trig}}_m(\mathbf{x},\mathbf{y})] \\ &= -(\widehat{\mathrm{SM}}(\mathbf{x},\mathbf{y}))^2\mathrm{Var}(\hat{\lambda}(\theta)) \end{aligned} \tag{25}$$

Now assume that $\widehat{\mathrm{SM}}^{\mathrm{trig}}_m(\mathbf{x},\mathbf{y})$ and $\widehat{\mathrm{SM}}^{++}_m(\mathbf{x},\mathbf{y})$ use the exact same random projections. Using a similar analysis, we get:

$$\begin{aligned}\mathrm{Cov}(\hat{\lambda}(\theta)\widehat{\mathrm{SM}}^{++}_m(\mathbf{x},\mathbf{y}),(1-\hat{\lambda}(\theta))\widehat{\mathrm{SM}}^{\mathrm{trig}}_m(\mathbf{x},\mathbf{y})) &= \mathbb{E}[\widehat{\mathrm{SM}}^{++}_m(\mathbf{x},\mathbf{y})\widehat{\mathrm{SM}}^{\mathrm{trig}}_m(\mathbf{x},\mathbf{y})]\mathbb{E}[\hat{\lambda}(\theta)(1-\hat{\lambda}(\theta))] \\ &\quad - (\widehat{\mathrm{SM}}(\mathbf{x},\mathbf{y}))^2\mathbb{E}[\hat{\lambda}(\theta)]\mathbb{E}[(1-\hat{\lambda}(\theta))] \end{aligned} \tag{26}$$

It is no longer true, however, that

$$\mathbb{E}[\widehat{\mathrm{SM}}^{++}_m(\mathbf{x},\mathbf{y})\widehat{\mathrm{SM}}^{\mathrm{trig}}_m(\mathbf{x},\mathbf{y})] = \mathbb{E}[\widehat{\mathrm{SM}}^{++}_m(\mathbf{x},\mathbf{y})]\mathbb{E}[\widehat{\mathrm{SM}}^{\mathrm{trig}}_m(\mathbf{x},\mathbf{y})] = (\widehat{\mathrm{SM}}(\mathbf{x},\mathbf{y}))^2 \tag{27}$$

since $\widehat{\mathrm{SM}}^{++}_m(\mathbf{x},\mathbf{y})$ and $\widehat{\mathrm{SM}}^{\mathrm{trig}}_m(\mathbf{x},\mathbf{y})$ are no longer independent. Thus, in order to compute

$$\mathbb{E}[\widehat{\mathrm{SM}}^{++}_m(\mathbf{x},\mathbf{y})\widehat{\mathrm{SM}}^{\mathrm{trig}}_m(\mathbf{x},\mathbf{y})]\mathbb{E}[\hat{\lambda}(\theta)(1-\hat{\lambda}(\theta))] \tag{28}$$

we let $\omega_1,\ldots,\omega_m \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,I_d)$ the random projections sampled to construct both $\widehat{\mathrm{SM}}^{++}_m(\mathbf{x},\mathbf{y})$ and $\widehat{\mathrm{SM}}^{\mathrm{trig}}_m(\mathbf{x},\mathbf{y})$. Denote:

$$Y_i = \cosh(\omega_i^\top(\mathbf{x}+\mathbf{y})) \quad \text{and} \quad \widehat{\mathrm{SM}}^{++}_m(\mathbf{x},\mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}\|^2+\|\mathbf{y}\|^2}{2}\right)\frac{Y_1+\cdots+Y_m}{m} \tag{29}$$

Letting $Z_i = \cosh(\omega_i^\top(\mathbf{x}-\mathbf{y}))$, we have $\widehat{\mathrm{SM}}^{\mathrm{trig}}_m(\mathbf{x},\mathbf{y})$:

$$\widehat{\mathrm{SM}}^{\mathrm{trig}}_m(\mathbf{x},\mathbf{y}) = \exp\left(\frac{\|\mathbf{x}\|^2+\|\mathbf{y}\|^2}{2}\right)\frac{Z_1+\cdots+Z_m}{m} \tag{30}$$

then we can then rewrite $\mathbb{E}[\widehat{\mathrm{SM}}^{++}_m(\mathbf{x},\mathbf{y})\widehat{\mathrm{SM}}^{\mathrm{trig}}_m(\mathbf{x},\mathbf{y})]$ as:

$$\begin{aligned}\mathbb{E}[\widehat{\mathrm{SM}}^{++}_m(\mathbf{x},\mathbf{y})\widehat{\mathrm{SM}}^{\mathrm{trig}}_m(\mathbf{x},\mathbf{y})] &= \frac{1}{m^2}\left[\sum_{i\neq j}\mathbb{E}[Y_iZ_j] + \sum_{i=1}^m \mathbb{E}[Y_iZ_i]\right] \\ &= \frac{1}{m^2}\left[\binom{m}{2}(\widehat{\mathrm{SM}}(\mathbf{x},\mathbf{y}))^2 + m\mathbb{E}[\cosh(\omega^\top(\mathbf{x}+\mathbf{y}))\cosh(\omega^\top(\mathbf{x}-\mathbf{y}))]\right] \end{aligned} \tag{31}$$

where $\omega \sim \mathcal{N}(0, I_d)$ where equality follows from the unbiasedness of $\widehat{\mathrm{SM}}_m^{\mathrm{trig}}(\mathbf{x}, \mathbf{y})$ and $\widehat{\mathrm{SM}}_m^{++}(\mathbf{x}, \mathbf{y})$ and the fact that different ω_i are chosen independently. We still need to compute the following term:

$$\rho = \mathbb{E}[\cosh(\omega^\top(\mathbf{x}+\mathbf{y}))\cos(\omega^\top(\mathbf{x}-\mathbf{y}))]. \tag{32}$$

Note first that:

$$\rho = \mathbb{E}[\exp(\omega^\top(\mathbf{x}+\mathbf{y}))\cos(\omega^\top(\mathbf{x}-\mathbf{y}))] \tag{33}$$

since $-\omega \sim \mathcal{N}(0, I_d)$ and cos is an even function. Denote $z = \mathbf{x}+\mathbf{y}+i(\mathbf{x}-\mathbf{y})$. We have:

$$\begin{aligned}
\mathbb{E}[\exp(\omega^\top(\mathbf{x}+\mathbf{y}))\cos(\omega^\top(\mathbf{x}-\mathbf{y}))] &= \mathrm{Re}\left[\mathbb{E}[\exp(\omega^\top z)]\right] \\
&= \mathrm{Re}\left[\prod_{i=1}^{d}\exp\left(\frac{z_i^2}{2}\right)\right] \\
&= \mathrm{Re}\left[\exp\left(\sum_{j=1}^{d}\frac{(x_j+y_j)^2+2i(x_j^2-y_j^2)-(x_j-y_j)^2}{2}\right)\right] \\
&= (\widehat{\mathrm{SM}}(\mathbf{x}, \mathbf{y}))^2\cos(\|\mathbf{x}\|^2 - \|\mathbf{y}\|^2)
\end{aligned} \tag{34}$$

Thus, we conclude that:

$$\mathbb{E}[\widehat{\mathrm{SM}}_m^{++}(\mathbf{x}, \mathbf{y})\widehat{\mathrm{SM}}_m^{\mathrm{trig}}(\mathbf{x}, \mathbf{y})] = \left(1 - \frac{1}{m}\right)(\widehat{\mathrm{SM}}(\mathbf{x}, \mathbf{y}))^2 + \frac{1}{m}(\widehat{\mathrm{SM}}(\mathbf{x}, \mathbf{y}))^2\cos(\|\mathbf{x}\|^2 - \|\mathbf{y}\|^2) \tag{35}$$

Therefore, we get the formulae for the covariance term in both settings: when random projections of $\widehat{\mathrm{SM}}_m^{\mathrm{trig}}(\mathbf{x}, \mathbf{y})$ and $\widehat{\mathrm{SM}}_m^{++}(\mathbf{x}, \mathbf{y})$ are shared (variant II) and when they are not (variant I). The following is true for $Z = \frac{1}{m}(1 - \cos(\|\mathbf{x}\|^2 - \|\mathbf{y}\|^2))\mathbb{E}[\hat{\lambda}(\theta)(1 - \hat{\lambda}(\theta))]$:

$$\mathrm{Cov}(\hat{\lambda}(\theta)\widehat{\mathrm{SM}}_m^{++}(\mathbf{x}, \mathbf{y}), (1 - \hat{\lambda}(\theta))\widehat{\mathrm{SM}}_m^{\mathrm{trig}}(\mathbf{x}, \mathbf{y})) = \begin{cases} -(\widehat{\mathrm{SM}}(\mathbf{x}, \mathbf{y}))^2\mathrm{Var}(\hat{\lambda}(\theta)) & \text{for variant I} \\ -(\widehat{\mathrm{SM}}(\mathbf{x}, \mathbf{y}))^2(\mathrm{Var}(\hat{\lambda}(\theta)) + Z) & \text{for variant II} \end{cases} \tag{36}$$

We also have the following:

$$\begin{aligned}
\mathrm{Var}(\hat{\lambda}(\theta)\widehat{\mathrm{SM}}_m^{++}(\mathbf{x}, \mathbf{y})) &= \mathbb{E}[\hat{\lambda}(\theta)^2]\mathbb{E}[(\widehat{\mathrm{SM}}_m^{++}(\mathbf{x}, \mathbf{y}))^2] - (\mathbb{E}[\hat{\lambda}(\theta)\widehat{\mathrm{SM}}_m^{++}(\mathbf{x}, \mathbf{y})])^2 \\
&= \mathbb{E}[\hat{\lambda}(\theta)^2](\mathrm{MSE}(\widehat{\mathrm{SM}}_m^{++}(\mathbf{x}, \mathbf{y})) + (\widehat{\mathrm{SM}}(\mathbf{x}, \mathbf{y}))^2) - (\mathbb{E}[\hat{\lambda}(\theta)]^2(\widehat{\mathrm{SM}}(\mathbf{x}, \mathbf{y}))^2) \\
&= (\widehat{\mathrm{SM}}(\mathbf{x}, \mathbf{y}))^2\mathrm{Var}(\hat{\lambda}(\theta)) + \mathbb{E}[\hat{\lambda}(\theta)^2]\mathrm{MSE}(\widehat{\mathrm{SM}}_m^{++}(\mathbf{x}, \mathbf{y}))
\end{aligned} \tag{37}$$

Similarly:

$$\mathrm{Var}((1 - \hat{\lambda}(\theta))\widehat{\mathrm{SM}}_m^{\mathrm{trig}}(\mathbf{x}, \mathbf{y})) = (\widehat{\mathrm{SM}}(\mathbf{x}, \mathbf{y}))^2\mathrm{Var}(\hat{\lambda}(\theta)) + \mathbb{E}[(1 - \hat{\lambda}(\theta))^2]\mathrm{MSE}(\widehat{\mathrm{SM}}_m^{\mathrm{trig}}(\mathbf{x}, \mathbf{y})) \tag{38}$$

By putting the derived formulae for the above variance terms as well as the covariance terms back in Equation 24, we complete the proof of the theorem (note that the mean squared error of the hybrid estimator is its variance since it is unbiased).

Proof of Lemma 3.2

Proof. The formula for the expectation is implied by the formula (1) from Sec. 2.1 of Cho & Saul (2012) for the zeroth-order arc-cosine kernel $k_0(\mathbf{x}, \mathbf{y})$:

$$k_0(\mathbf{x}, \mathbf{y}) = 1 - \frac{\theta_{\mathbf{x},\mathbf{y}}}{\pi}. \tag{39}$$

It also follows from the Goemans-Williamson algorithm (Goemans & Williamson, 2004). We will now derive the formula for $c = \mathbb{E}[\hat{\lambda}^2(\mathbf{x}, \mathbf{y})]$. Since λ depends only on the angle $\theta = \theta_{\mathbf{x},\mathbf{y}}$, we will refer to it as $\lambda(\theta)$ and to its estimator as $\hat{\lambda}(\theta)$. Denote:

$$\phi_n^{\text{ang}}(z) = \frac{1}{\sqrt{n}}\big(\text{sgn}(\tau_1^\top z), \ldots, \text{sgn}(\tau_n^\top z)\big)^\top \tag{40}$$

where $\tau_1, \ldots, \tau_n$ are independent random vectors. Denote:

$$X_i = \phi_n^{\text{ang}}(\mathbf{x})[i]\phi_n^{\text{ang}}(\mathbf{y})[i]. \tag{41}$$

We have:

$$\hat{\lambda}(\theta) = \frac{1}{2}\left(1 - \frac{1}{n}\sum_{i=1}^{n} X_i\right). \tag{42}$$

Note first that by the construction of $\hat{\lambda}(\theta)$, we have:

$$\mathbb{E}[\hat{\lambda}(\theta)] = \frac{\theta}{\pi} \quad \text{and thus:} \quad \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = 1 - \frac{2\theta}{\pi} \tag{43}$$

Therefore, we conclude that:

$$\begin{aligned}
c &= \frac{1}{4}\mathbb{E}\left[1 - 2\sum_{i=1}^{n} X_i + \left(\sum_{i=1}^{n} X_i\right)^2\right] \\
&= \frac{1}{4}\left(1 - 2\left(1 - \frac{2\theta}{\pi}\right) + \sum_{i=1}^{n}\mathbb{E}[X_i^2] + \sum_{i\neq j}\mathbb{E}[X_i]\mathbb{E}[X_j]\right) \\
&= \frac{1}{4}\left(1 - 2\left(1 - \frac{2\theta}{\pi}\right) + n\cdot\frac{1}{n^2} + n(n-1)\cdot\frac{1}{n^2}\left(1 - \frac{2\theta}{\pi}\right)^2\right) \\
&= \frac{1}{4}\left(\frac{4\theta}{\pi} + \left(1 - \frac{1}{n}\right)\left(1 - \frac{2\theta}{\pi}\right)^2\right) \\
&= \frac{\theta}{\pi}\left(\frac{\theta}{\pi} - \frac{\theta}{n\pi} + \frac{1}{n}\right)
\end{aligned}$$

Explicit Formula for the MSE of the Angular Hybrid Estimator

Consider the angular hybrid estimator $\widehat{\text{SM}}_{m,n}^{\text{anghyb}}(\mathbf{x}, \mathbf{y})$, where $\widehat{\text{SM}}_m^{\text{trig}}(\mathbf{x}, \mathbf{y})$ and $\widehat{\text{SM}}_m^{++}(\mathbf{x}, \mathbf{y})$ are chosen independently, i.e., their random projections are chosen independently (note that we always assume that $\hat{\lambda}(\theta)$ is chosen independently from $\widehat{\text{SM}}_m^{\text{trig}}(\mathbf{x}, \mathbf{y})$ and $\widehat{\text{SM}}_m^{++}(\mathbf{x}, \mathbf{y})$). Then the following holds:

$$\begin{aligned}
\text{MSE}(\widehat{\text{SM}}_{m,n}^{\text{anghyb}}(\mathbf{x}, \mathbf{y})) = \frac{\theta}{\pi}\left(\frac{\theta}{\pi} - \frac{\theta}{n\pi} + \frac{1}{n}\right)\frac{1}{2m}\exp(\|z\|^2)\widehat{\text{SM}}^2(\mathbf{x}, \mathbf{y})(1 - \exp(-\|z\|^2))^2 \\
+ \left(1 - \frac{\theta}{\pi}\right)\left(1 - \frac{\theta}{\pi} + \frac{\theta}{n\pi}\right)\frac{1}{2m}\exp(\|z\|^2)\widehat{\text{SM}}^{-2}(\mathbf{x}, \mathbf{y})(1 - \exp(-\|\Delta\|^2))^2
\end{aligned} \tag{44}$$

Here, $\Delta = \mathbf{x} - \mathbf{y}$ and $z = \mathbf{x} + \mathbf{y}$. If $\widehat{\mathrm{SM}}_m^{\mathrm{trig}}(\mathbf{x}, \mathbf{y})$ and $\widehat{\mathrm{SM}}_m^{++}(\mathbf{x}, \mathbf{y})$ apply the exact same sets of random projections, the mean squared error of the hybrid estimator is further reduced, in particular:

$$\begin{aligned}\mathrm{MSE}(\widehat{\mathrm{SM}}_{m,n}^{\mathrm{anghyb}}(\mathbf{x}, \mathbf{y})) = \frac{\theta}{\pi}\left(\frac{\theta}{\pi} - \frac{\theta}{n\pi} + \frac{1}{n}\right)\frac{1}{2m}\exp(\|z\|^2)\widehat{\mathrm{SM}}^2(\mathbf{x}, \mathbf{y})(1 - \exp(-\|z\|^2))^2 \\ + \left(1 - \frac{\theta}{\pi}\right)\left(1 - \frac{\theta}{\pi} + \frac{\theta}{n\pi}\right)\frac{1}{2m}\exp(\|z\|^2)\widehat{\mathrm{SM}}^{-2}(\mathbf{x}, \mathbf{y})(1 - \exp(-\|\Delta\|^2))^2 \\ - \frac{2}{m}\widehat{\mathrm{SM}}^2(\mathbf{x}, \mathbf{y})(1 - \cos(\|\mathbf{x}\|^2 - \|\mathbf{y}\|^2))\frac{\theta}{\pi}\left(1 - \frac{1}{n} - \frac{\theta}{\pi} + \frac{\theta}{n\pi}\right) \quad (45)\end{aligned}$$

Thus, if $\|\mathbf{x}\|_2 = \|\mathbf{y}\|_2 = r$, then independent of whether the same sets of random projections are applied or not, we get:

$$\begin{aligned}\mathrm{MSE}(\widehat{\mathrm{SM}}_{m,n}^{\mathrm{anghyb}}(\mathbf{x}, \mathbf{y})) = \frac{\theta}{\pi}\left(\frac{\theta}{\pi} - \frac{\theta}{n\pi} + \frac{1}{n}\right)\frac{1}{2m}\exp(8r^2\cos^2(\frac{\theta}{2}) - 2r^2) \\ \cdot \frac{\theta}{\pi}\left(1 - \frac{1}{n} - \frac{\theta}{\pi} + \frac{\theta}{n\pi}\right)\frac{1}{2m}\exp(2r^2)(1 - \exp(-4r^2\sin^2(\frac{\theta}{2}))^2) \quad (46)\end{aligned}$$

Proof of Theorem 3.5

Proof. Note first that from the derived above formula of the MSE of the combined softmax estimator and the definition of the max-relative-error, we obtain:

$$\epsilon_{S(r)}\left(\widehat{\mathrm{SM}}_{m,n}^{\mathrm{anghyb}}\right) = \frac{\exp(r^2)}{\sqrt{2m}}\sqrt{\max_{\theta\in[0,\pi]} h_r(\theta)} \quad (47)$$

where:

$$h_r(\theta) = a_r(\theta) + a_r(\pi - \theta) \quad (48)$$

and $a_r(\theta)$ is defined as:

$$a_r(\theta) = \frac{\theta}{\pi}\left(\frac{\theta}{\pi} - \frac{\theta}{n\pi} + \frac{1}{n}\right)\exp(2r^2\cos(\theta))\left(1 - \exp(-4r^2\cos^2\left(\frac{\theta}{2}\right))\right)^2 \quad (49)$$

Therefore, we have:

$$\epsilon_{S(r)}\left(\widehat{\mathrm{SM}}_{m,n}^{\mathrm{anghyb}}\right) \leq \frac{\exp(r^2)}{\sqrt{m}}\sqrt{\max_{\theta\in[0,\pi]} a_r(\theta)} \quad (50)$$

Notice that:

$$a_r(\theta) \leq b_r(\theta)\left(1 - \exp(-4r^2)\right)^2 \quad (51)$$

Where:

$$b_r(\theta) = b_r^1(\theta) + b_r^2(\theta) \quad (52)$$

$$b_r^1(\theta) = \left(1 - \frac{1}{n}\right)\frac{\theta^2}{\pi^2}\exp(2r^2\cos(\theta)) \quad (53)$$

$$b_r^2(\theta) = \frac{1}{n}\frac{\theta}{\pi}\exp(2r^2\cos(\theta)) \quad (54)$$

Therefore:

$$\max_{\theta\in[0,\pi]} b_r(\theta) \leq \max_{\theta\in[0,\pi]} b_r^1(\theta) + \max_{\theta\in[0,\pi]} b_r^2(\theta) \quad (55)$$

Denote: $b^1 = \max_{\theta\in[0,\pi]} b_r^1(\theta)$ and $b^2 = \max_{\theta\in[0,\pi]} b_r^2(\theta)$. Note that:

$$\frac{db_r^1(\theta)}{d\theta} = \exp(2r^2\cos(\theta))\left(1-\frac{1}{n}\right)\frac{2\theta}{\pi^2}(1-r^2\theta\sin(\theta)) \tag{56}$$

$$\frac{db_r^2(\theta)}{d\theta} = \exp(2r^2\cos(\theta))\frac{1}{n\pi}(1-2r^2\theta\sin(\theta)) \tag{57}$$

Thus, from the properties of the function $\theta \to \theta\sin(\theta)$ and the fact that $r \geq 1$, we conclude that both derivatives are first non-negative, then non-positive, and then non-negative, and that the unique local maximum on the interval $[0,\pi]$ is achieved for $\theta \leq \frac{\pi}{2}$. Note also that $b_r^1(\theta), b_r^1(\pi) \geq 0$ and $b_r^2(0) = b_r^2(0) = 0, b_r^2(\pi) = (1-\frac{1}{n}\exp(-2r^2), b_r^2(\pi) = \frac{1}{n}\exp(-2r^2)$. We conclude that the global maximum for $b_r(\theta)$ on the interval $[0,\pi]$ for $i = 1,2$ is achieved either in its unique local maximum on that interval or for $\theta = \pi$. Let us consider first $b_r^1(\theta)$. In its local maximum on $[0,\pi]$, we have:

$$\theta^*\sin(\theta^*) = \frac{1}{r^2} \tag{58}$$

Since $\theta \leq \sin(\theta)\cdot\frac{\pi}{2}$ on $[0,\frac{\pi}{2}]$, we get:

$$(\theta^*)^2 \leq \frac{\pi}{2r^2} \tag{59}$$

$$\theta^* \leq \sqrt{\frac{\pi}{2r^2}} \tag{60}$$

Therefore:

$$b_r^1(\theta^*) \leq \left(1-\frac{1}{n}\right)\frac{1}{2\pi r^2}\exp(2r^2) \geq b_r^1(\pi) \tag{61}$$

We thus conclude that:

$$\max_{\theta\in[0,\pi]} b_r^1(\theta) \leq \left(1-\frac{1}{n}\right)\frac{1}{2\pi r^2}\exp(2r^2) \tag{62}$$

By the completely analogous analysis applied to b_r^2, we obtain:

$$\max_{\theta\in[0,\pi]} b_r^2(\theta) \leq \frac{1}{2n\sqrt{\pi}r^2}\exp(2r^2) \tag{63}$$

Now, using Equation 48, Equation 49, and Equation 56, we obtain:

$$\epsilon_{S(r)}\left(\widehat{\mathrm{SM}}_{m,n}^{\mathrm{anghyb}}\right) \leq \frac{\exp(2r^2)}{\sqrt{2m}r}\left(1-\exp(-4r^2)\right)\left(\frac{1}{\pi}-\frac{1}{n\pi}+\frac{1}{n\sqrt{\pi}}\right) \tag{64}$$

and that completes the first part of the proof (proof of Inequality 27). The equations on the limits are directly implied by the fact that:

$$\epsilon_{S(r)}\left(\widehat{\mathrm{SM}}_{m,n}^{\mathrm{anghyb}}\right) = \frac{\exp(2r^2)}{\sqrt{2m}}\sqrt{h_r(\theta)} \tag{65}$$