ADAPTING LINEAR ATTENTION TO QUERY KEY DISTRIBUTIONS

John Blackwelder
Department of Computer Science
Columbia University
New York, NY 10027, USA
{jwb2168}@columbia.edu

ABSTRACT

I adapt the *complex exponential* estimator (CEXP), first introduced in Choromanski et al. (2022), to the setting where query and key vectors are assumed to be sampled from two different multivariate normal distributions. I find that many of the strong theoretical results derived in Choromanski et al. (2022) and Choromanski et al. (2021) apply to CEXP due to the similarities between the CEXP and PRF softmax kernel estimators, and I use these to obtain concentration results that take the query and key distributions into account. I also derive a simple heuristic for choosing the parameters of the CEXP estimator and empirically demonstrate that this new method leads to improved mean squared error and max relative error results.

1 INTRODUCTION & RELATED WORK

Consider the *softmax kernel* function SM : $\mathbb{R}^{d \times d} \rightarrow \mathbb{R}$ defined as follows:

$$\mathrm{SM}(\mathbf{x}, \mathbf{y}) = \exp(\mathbf{x}^T \mathbf{y}).$$

As described in Choromanski et al. (2021), being able to efficiently estimate the softmax kernel using feature functions $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^m$ such that $\mathrm{SM}(\mathbf{x}, \mathbf{y}) \approx \phi(\mathbf{x})^T \phi(\mathbf{y})$ allows for transformer model variants that don't need to explicitly generate the attention matrix, a quadratic time operation that causes bottlenecks as the number of tokens increases. The *positive random features* (PRF) estimator introduced in this paper uses the following random feature function to estimate the softmax kernel:

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

where $\omega_i \sim N(0, \mathbf{I}_d)$. First introduced in Choromanski et al. (2022), the *complex exponential* estimator is a simple generalization of PRF defined as follows:

$$\widehat{\mathrm{SM}}^{cexp}_{\mathbf{A},m}(\mathbf{x}, \mathbf{y}) = \Psi^m_{\mathbf{A}}(\mathbf{x})^T \Psi^m_{\mathbf{A}^{-T}}(\mathbf{y}),$$

where $\Psi^m_{\mathbf{M}}(\mathbf{u}) = \frac{1}{\sqrt{m}} \exp(-\frac{\|\mathbf{Mu}\|^2}{2})(\exp(\omega_1^T \mathbf{Mu}), \ldots, \exp(\omega_m^T \mathbf{Mu}))$, $\mathbf{A}$ is a real invertible matrix, and ω_i is defined as above (note that CEXP as originally defined allows for complex valued $\mathbf{A}$, but I will only consider the case where $\mathbf{A} \in \mathbb{R}^{d \times d}$ here). Both of these estimators were proven to be unbiased when they were introduced.

1.1 THE HYPERBOLIC COMPLEX EXPONENTIAL ESTIMATOR

Notice that as a direct consequence of the proof of lemma 1 from Choromanski et al. (2021), the following alteration to CEXP also results in an unbiased softmax kernel estimator:

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

I will refer to the variant of CEXP using $\Psi^{m+}_{\mathbf{M}}$ defined above as the hyperbolic complex exponential estimator (HCEXP).

2 THE ALGORITHM

Suppose we wish to create the softmax attention matrix over the query values $\mathbf{x}_1, \ldots, \mathbf{x}_s \in \mathbb{R}^d$ and key values $\mathbf{y}_1, \ldots, \mathbf{y}_s \in \mathbb{R}^d$, which we assume are sampled from two potentially different multivariate normal distributions. Begin by obtaining component-wise estimates of the mean and variance parameters, $\hat{\mu_{\mathbf{x}}}_i, \hat{\mu_{\mathbf{y}}}_i, \hat{\sigma}^2_{\mathbf{x}_i}, \hat{\sigma}^2_{\mathbf{y}_i}$. This can be done using the sample means and the unbiased variance estimator. Using these parameter estimates, obtain the real diagonal matrix $\mathbf{A}$ either by using the method outlined in the appendix of Hybrid Random Features or the heuristic derived in the next section. Finally, after choosing a value for m that is as large as time and space constraints allow for, the ijth entry of the attention matrix $\mathbf{T}$ can be estimated as follows:

$$\mathbf{T}_{ij} \approx \widehat{\mathrm{SM}}^{hcexp}_{\mathbf{A},m}(\mathbf{x}_i, \mathbf{y}_j).$$

3 THEORETICAL RESULTS

Due to the similarities between Positive Random Features and Complex Exponential estimators, when $\mathbf{A}$ is real and invertible, the mean squared errors of the Complex Exponential estimators can be found using derivations mirroring those in the Performers paper. This results in the following lemma:

Lemma 3.1. *When $\mathbf{A}$ is a real invertible matrix, the mean squared errors of the estimators are as follows:*

$$MSE(\widehat{SM}^{cexp}_{\mathbf{A},m}(\mathbf{x},\mathbf{y})) = \frac{1}{m}\exp(\|\mathbf{A}\mathbf{x} + \mathbf{A}^{-T}\mathbf{y}\|^2)SM^2(\mathbf{x},\mathbf{y})(1 - \exp(-\|\mathbf{A}\mathbf{x} + \mathbf{A}^{-T}\mathbf{y}\|^2)),$$

$$MSE(\widehat{SM}^{hcexp}_{\mathbf{A},m}(\mathbf{x},\mathbf{y})) = \frac{1}{2m}\exp(\|\mathbf{A}\mathbf{x} + \mathbf{A}^{-T}\mathbf{y}\|^2)SM^2(\mathbf{x},\mathbf{y})(1 - \exp(-\|\mathbf{A}\mathbf{x} + \mathbf{A}^{-T}\mathbf{y}\|^2))^2.$$

The key fact that allows the above statements to be true is that $(\mathbf{A}\mathbf{x})^T(\mathbf{A}^{-\mathbf{T}}\mathbf{y}) = \mathbf{x}^{\mathbf{T}}(\mathbf{A}^{\mathbf{T}}\mathbf{A}^{-\mathbf{T}})\mathbf{y} = \mathbf{x}^T\mathbf{y}$, which in turn implies that $\mathrm{SM}(\mathbf{x},\mathbf{y}) = \mathrm{SM}(\mathbf{A}\mathbf{x}, \mathbf{A}^{-\mathbf{T}}\mathbf{y})$. As a result, the Complex Exponential Estimators can be thought of as their corresponding Positive Random Features estimators applied to the transformed input $\mathbf{x}' = \mathbf{A}\mathbf{x}, \mathbf{y}' = \mathbf{A}^{-\mathbf{T}}\mathbf{y}$, which explains the similar looking mean squared error results. The hyperbolic variant of the Complex Exponential estimator is clearly superior to the original, since as before with Positive Random Features, the squared rightmost term results in strictly smaller mean squared error values, even when making m smaller to account for the added functions contained within the random feature map. The ensuing analysis will therefore focus on this variant.

Lemma 3.2. *Let $C \subseteq \mathbb{R}^d$ be some region and $\mathbf{A} \in \mathbb{R}^{d\times d}$ an invertible matrix defined such that $\mathbf{x}, \mathbf{y} \in C$ implies both $\|\mathbf{A}\mathbf{x}\|$ and $\|\mathbf{A}^{-T}\mathbf{y}\| \leq r$. With max relative error defined as in Hybrid Random Features and $W(r) = \exp(2r^2)(1 - \exp(4r^2))$, it follows that*

$$\epsilon_C(\widehat{SM}^{hcexp}_{\mathbf{A},m}(\mathbf{x},\mathbf{y})) = \frac{1}{\sqrt{2m}}W(r).$$

The proof mirrors that of Lemma 3.4 from Hybrid Random Features. Consider now the case where $\mathbf{X} \sim N(\mu_{\mathbf{x}}, \mathbf{\Sigma}_{\mathbf{x}})$ and $\mathbf{Y} \sim N(\mu_{\mathbf{y}}, \mathbf{\Sigma}_{\mathbf{y}})$ are multivariate normal random vectors. It follows that $\mathbf{A}\mathbf{X} \sim N(\mathbf{A}\mu_{\mathbf{x}}, \mathbf{A}\mathbf{\Sigma}_{\mathbf{x}}\mathbf{A}^T)$ and $\mathbf{A}^{-T}\mathbf{Y} \sim N(\mathbf{A}^{-\mathbf{T}}\mu_{\mathbf{y}}, \mathbf{A}^{-T}\mathbf{\Sigma}_{\mathbf{y}}\mathbf{A}^{-1})$. From now on these parameters will be referred to as $\mu_{\mathbf{A}\mathbf{x}}$, $\mathbf{\Sigma}_{\mathbf{A}\mathbf{x}}$, $\mu_{\mathbf{A}^{-\mathbf{T}}\mathbf{y}}$, and $\mathbf{\Sigma}_{\mathbf{A}^{-\mathbf{T}}\mathbf{y}}$.

Theorem 3.3. *Suppose $\mathbf{A} \in \mathbb{R}^{d\times d}$ is an invertible matrix, $\mathbf{x}_1, \ldots, \mathbf{x}_s$ and $\mathbf{y}_1, \ldots, \mathbf{y}_s$ are sampled i.i.d from $N(\mu_{\mathbf{x}}, \mathbf{\Sigma}_{\mathbf{x}})$ and $N(\mu_{\mathbf{y}}, \mathbf{\Sigma}_{\mathbf{y}})$ respectively, and $\mu_{\mathbf{A}\mathbf{x}}$, $\mathbf{\Sigma}_{\mathbf{A}\mathbf{x}}, \mu_{\mathbf{A}^{-\mathbf{T}}\mathbf{y}}, \mathbf{\Sigma}_{\mathbf{A}^{-\mathbf{T}}\mathbf{y}}$ are defined as above. Define $\lambda_{Ax}, \lambda_{A^{-T}y}$ as the largest eigenvalues of $\mathbf{\Sigma}_{Ax}, \mathbf{\Sigma}_{\mathbf{A}^{-\mathbf{T}}\mathbf{y}}$ respectively, let $\chi^2_d(\mathrm{p})$ be the quantile function for probability p of the chi-squared distribution with d degrees of freedom, and let*

$$r = \max\Big(\sqrt{\lambda_{Ax} * \chi^2_d(\sqrt[2s]{p})} + \|\mu_{\mathbf{A}\mathbf{x}}\|, \sqrt{\lambda_{A^{-T}y} * \chi^2_d(\sqrt[2s]{p})} + \|\mu_{\mathbf{A}^{-\mathbf{T}}\mathbf{y}}\|\Big).$$

Then with probability at least p, the largest observed relative error for $\widehat{SM}^{hcexp}_{\mathbf{A},m}(\mathbf{x}_i, \mathbf{y}_j)$ across all $i, j \in \{1, \ldots, s\}$ is less than or equal to $\frac{1}{\sqrt{2m}}W(r)$.

Proof. With probability q, the following inequality holds for $\mathbf{x} \in \mathbb{R}^d$ randomly sampled from $N(\mu_{\mathbf{x}}, \mathbf{\Sigma_x})$:

$$(\mathbf{Ax} - \mu_{\mathbf{Ax}})^T \mathbf{\Sigma}_{Ax}^{-1}(\mathbf{Ax} - \mu_{\mathbf{Ax}}) \leq \chi_d^2(q),$$

where $\chi_d^2(q)$ is the quantile function for probability q of the chi-squared distribution with d degrees of freedom Siotani (1964). Applying Rayleigh quotient bounds results in

$$\frac{1}{\lambda_{Ax}} \|\mathbf{Ax} - \mu_{\mathbf{Ax}}\|^2 \leq (\mathbf{Ax} - \mu_{\mathbf{Ax}})^T \mathbf{\Sigma}_{Ax}^{-1}(\mathbf{Ax} - \mu_{\mathbf{Ax}}) \leq \chi_d^2(q),$$

where λ_{Ax} is the largest eigenvalue in $\mathbf{\Sigma}_{Ax}$. Thus, with probability greater than or equal to q,

$$\|\mathbf{Ax} - \mu_{\mathbf{Ax}}\|^2 \leq \lambda_{Ax} * \chi_d^2(q).$$

The same argument shows that, with $\lambda_{A^{-T}y}$ being the largest eigenvalue in $\mathbf{\Sigma}_{\mathbf{A^{-T}y}}$, the following holds with probability at least q:

$$\|\mathbf{A^{-T}y} - \mu_{\mathbf{A^{-T}y}}\|^2 \leq \lambda_{A^{-T}y} * \chi_d^2(q).$$

Thus, with probability at least q^2, both of the above inequalities hold. For samples of size s, $2s$ inequalities of this form must hold, which occurs with probability at least q^{2s}. Applying the triangle inequality to these results in

$$\|\mathbf{Ax}\| \leq \|\mathbf{Ax} - \mu_{\mathbf{Ax}}\| + \|\mu_{\mathbf{Ax}}\| \leq \sqrt{\lambda_{Ax} * \chi_d^2(q)} + \|\mu_{\mathbf{Ax}}\|,$$

and

$$\|\mathbf{A^{-T}y}\| \leq \sqrt{\lambda_{A^{-T}y} * \chi_d^2(q)} + \|\mu_{\mathbf{A^{-T}y}}\|.$$

Applying the previous lemma to these bounds concludes the proof.

□

Note that in the case where the true mean and covariance parameters are unknown, a similar result can be obtained using Hotelling's T-squared distribution, which approaches χ_d^2 as the sample size increases. This result indicates that in order to choose A effectively, one should consider both the means and covariances of the query and key distributions. Consider the case where $\mu_{\mathbf{x}_i}$ is very close to 0 while $|\mu_{\mathbf{y}_i}|$ is considerably larger. Setting $\mathbf{A}_{ii} = \frac{\sqrt{|\mu_{\mathbf{y}_i}|}}{\sqrt{|\mu_{\mathbf{x}_i}|}}$ as described in the appendix of Hybrid Random Features is optimal when $\mathbf{A}$ is restricted to a real diagonal matrix and $\mathbf{x} = \mu_{\mathbf{x}}, \mathbf{y} = \mu_{\mathbf{y}}$. However, this means that $\mathbf{A}_{ii}$ gets arbitrarily large as $\mu_{\mathbf{x}_i}$ approaches 0, as does the ith entry along the diagonal of $\mathbf{\Sigma_{Ax}}$. Thus, any variance from the mean in the ith component of $\mathbf{x}$ will be magnified and potentially result in inaccurate softmax estimates. This motivates a heuristic for choosing $\mathbf{A}$ that takes the variance of the components of $\mathbf{x}$ and $\mathbf{y}$ into account.

Lemma 3.4. *When* $\mathbf{A}$ *is restricted to being a real diagonal matrix, the matrix* $\mathbf{A}$ *which minimizes* $\mathbb{E}_{\mathbf{x},\mathbf{y}}[\|\mathbf{Ax}\|^2 + \|\mathbf{A}^{-T}\mathbf{y}\|^2]$ *is defined by setting*

$$\mathbf{A_{ii}} = \sqrt[4]{\frac{(\sigma_{yi}^2 + \mu_{yi}^2)}{(\sigma_{xi}^2 + \mu_{xi}^2)}}.$$

Proof.

$$\mathbb{E}[\|Ax\|^2] = Tr(\mathbf{A\Sigma_x A^T}) + (\mathbf{A}\mu_{\mathbf{x}})^T(\mathbf{A}\mu_{\mathbf{x}}).$$

$$Tr(\mathbf{\Sigma_x A}^2) = \sigma_{x1}^2 a_{11}^2 + \cdots + \sigma_{xd}^2 a_{dd}^2,$$

and

$$(\mathbf{A}\mu_{\mathbf{x}})^T(\mathbf{A}\mu_{\mathbf{x}}) = \mu_{x1}^2 a_{11}^2 + \cdots + \mu_{xd}^2 a_{dd}^2.$$

Similar results are obtained for $\mathbb{E}[\|A^{-T}y\|^2]$ albeit with a_{ii}^{-2} replacing a_{ii}^2. Summing these together, computing the gradient and setting it to 0, and then solving results in the above.

□

These two ways of defining $\mathbf{A}$ result in the same entries in the case where variance equals 0 (i.e., $\mathbf{x_i}$ is a constant), highlighting the similarities between the methods. Minimizing the expectation in the lemma does not necessarily minimize the expected mean squared error or expected relative error of the estimator over the $\mathbf{x}$ and $\mathbf{y}$ distributions. To do so, one would need to minimize a much more complicated function with respect to $\mathbf{A}$, and an exact solution may not have a closed form expression. In the following section, we validate experimentally that this new method of choosing $\mathbf{A}$ seems to result in better estimators.

4 Experimental Results

Positive Random Features, Complex Exponential Features without variance adjustment, and Complex Exponential Features with variance adjustment are tested on a randomly created dataset containing 1000 (x, y) pairs (Only $\widehat{\mathrm{SM}}(\mathbf{x}_i, \mathbf{y}_i)$ is calculated for each i) with $m = 1024$ and $d = 50$. Some outliers don't appear. $\mathbf{x}_1, \ldots, \mathbf{x}_{1000}$ are sampled i.i.d from $N(\mu_{\mathbf{x}}, \mathbf{\Sigma_x})$, where $\mu_{\mathbf{x}j} \sim_{i.i.d} Laplace(0, 50)$ and $\mathbf{\Sigma_x}$ is a diagonal matrix with i.i.d entries sampled from $Gamma(.02, 1)$. The sample $\mathbf{y}_1, \ldots, \mathbf{y}_{1000}$ is created similarly, albeit with $Gamma(.01, 1)$ used for the entries of the diagonal covariance matrix. Finally, the samples are rescaled in order to ensure their norms fall into a reasonable range. Inspired by figure 9 in the appendix of Hybrid Random Features, the $\mathbf{x}$ are encouraged to have norms around 5 and the $\mathbf{y}$ are encouraged to have norms around 0.5. The sample means and covariance matrices are used to calculate $\mathbf{A}$ in order to reflect how the algorithm would perform when the true parameters are unknown.

Figure 1 contains scatterplots which demonstrate how the estimators perform relative to one another. Due to the extremely challenging dataset, Positive Random Features performs poorly. Complex Exponential Features without variance adjustment performs much better, as there is some semblance of a linear relationship between the true softmax values and predicted softmax values. Finally, Complex Exponential Features with the variance adjustment described in the previous section clearly performs the best on this particular dataset, as a strong linear relationship holds for softmax values less than 2.

I then run these estimators on 20 different randomly generated datasets, this time setting $m = 1024$ and making each dataset contain 250 $(\mathbf{x}, \mathbf{y})$ pairs, and I calculate the mean squared error and max relative error for each dataset and estimator combination.

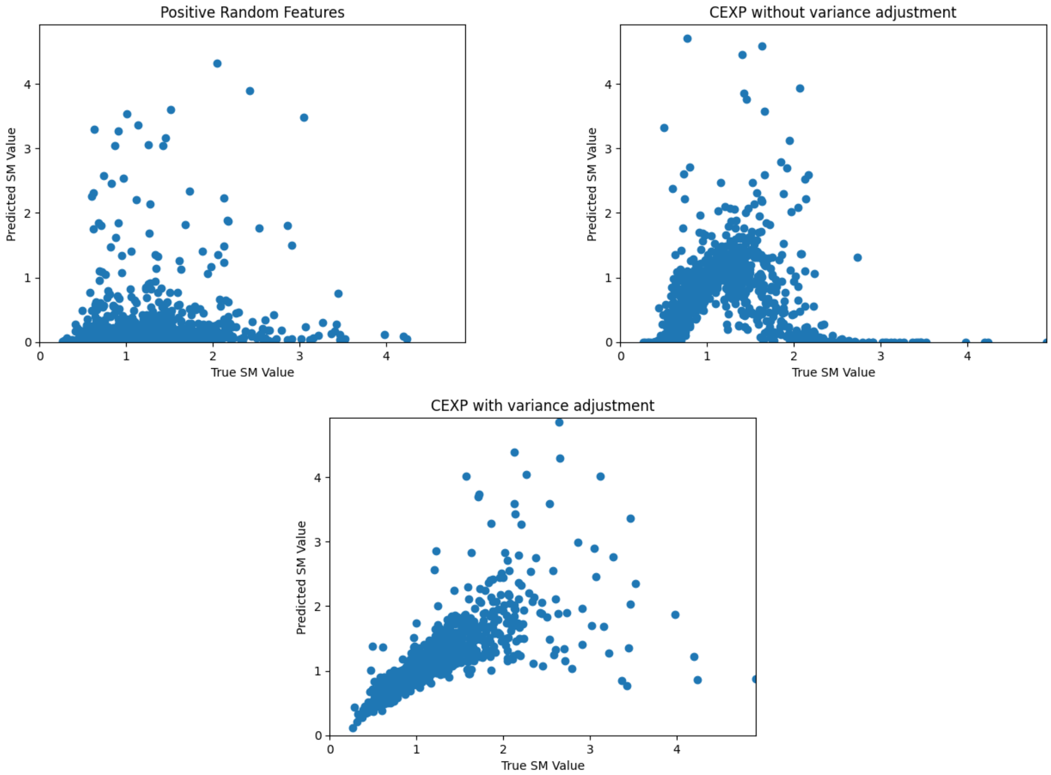

Figure 1: Positive Random Features, Complex Exponential Features without variance adjustment, and Complex Exponential Features with variance adjustment tested on a randomly created dataset containing 1000 (x, y) pairs.

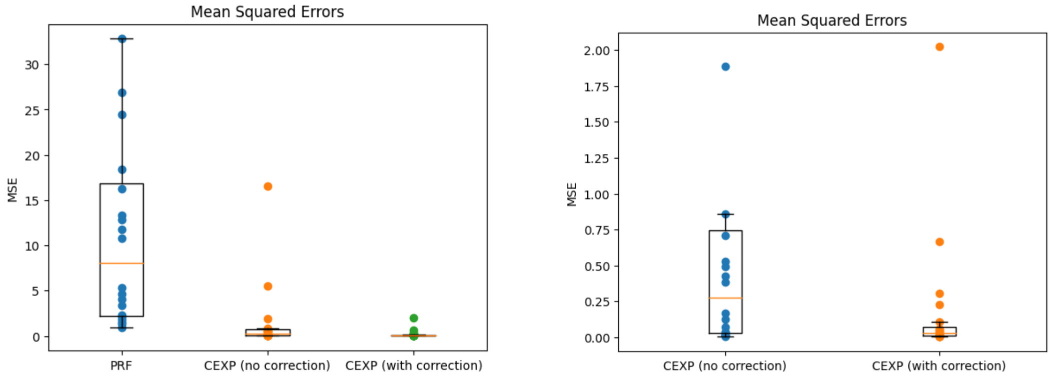

Figure 2: Boxplots showing the mean squared errors of the estimators across 20 randomly generated datasets.

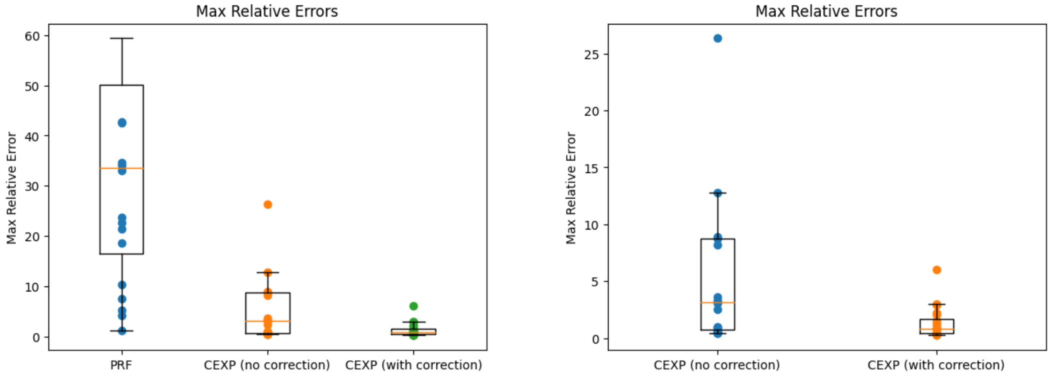

Figure 3: Boxplots showing the max relative errors of the estimators across 20 randomly generated datasets.

5 Conclusion

The complex exponential estimator, and by extension hyperbolic complex exponential estimator, is a promising method for estimating the softmax kernel and could potentially be used in efficient transformer models based on linear attention. I obtained concentration results for HCEXP when queries and keys are assumed to be normally distributed, and I introduced a new heuristic for choosing the entries of $\mathbf{A}$. I showed empirically that using sample variance statistics in addition to component-wise sample means when computing $\mathbf{A}$ leads to better approximations. Other methods for defining $\mathbf{A}$, like allowing it to contain complex values or using numerical optimization, were not discussed here but could potentially be fruitful areas of research.

6 Acknowledgments

Thank you to Professor Choromanski for a great semester of Machine Learning and High Dimensional Data Mining and for his guidance on this project!

References

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser David Belanger, Lucy Colwell, and Adrian Weller. Rethinking attention with performers. *ICLR*, 2021.

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, and Adrian Weller. Hybrid random features. *ICLR*, 2022.

Minoru Siotani. Tolerance regions for a multivariate normal population. *Annals of the Institute of Statistical Mathematics*, 1964.