
SchoenbAt: Rethinking Attention with Polynomial basis

Yuhan Guo^{1,2}, Lizhong Ding^{1,3}, Yuwan Yang¹, Xuewei Guo¹

¹Beijing Institute of Technology

²3220231188@bit.edu.cn, ³lizhong.ding@outlook.com

Abstract

Kernelized attention extends the attention mechanism by modeling sequence correlations through kernel functions, making significant progresses in optimizing attention. Under the guarantee of harmonic analysis theory, kernel functions can be expanded with basis functions, inspiring random feature-based approaches to enhance the efficiency of kernelized attention while maintaining predictive performance. However, current random feature-based works are limited to the Fourier basis expansions under Bochner’s theorem. We propose Schoenberg’s theorem-based attention (SchoenbAt), which approximates dot-product kernelized attention with the polynomial basis under Schoenberg’s theorem via random Maclaurin features and applies a two-stage regularization to constrain the input space and restore the output scale, acting as a drop-in replacement of dot-product kernelized attention. Our theoretical proof of the unbiasedness and concentration error bound of SchoenbAt supports its efficiency and accuracy as a kernelized attention approximation, which is also empirically validated under various random feature dimensions. Evaluations on real-world datasets demonstrate that SchoenbAt significantly enhances computational speed while preserving competitive performance in terms of precision, outperforming several efficient attention methods.

1 Introduction

In recent years, attention mechanisms [1] have made significant advancements in several areas of machine learning including natural language processing [2, 3, 4], computer vision [5, 6, 7, 8] and bioinformatics [9, 10, 11], particularly with the success of the Transformer architecture [12] in Large-Language Models [13, 14, 15, 16]. Kernelized attention [17, 18] is a widely studied variant of attention [19, 20, 21], which generalizes the function capturing sequence dependencies to a kernel function, thereby offering new perspectives for broadening the horizons of research [22, 23, 24].

The expansion of the kernel function is one of the key techniques for efficient kernelized attention. As a classical theoretical result in harmonic analysis, Bochner’s theorem [25] guarantees that a *shift-invariant kernel* is an expansion with Fourier bases, motivating random Fourier feature methods [26, 27, 28, 29, 30, 31] which approximate the kernel function through nonlinear random mappings, significantly improving computational efficiency while maintaining accuracy [32, 33, 34, 35, 36, 37]. However, the attention mechanism inherently models sequence dependencies through a *dot-product kernel*, and works based on Fourier expansions require additional computation to transform the dot-product kernel into a shift-invariant kernel, limiting their practical utility.

We observe that the *polynomial basis expansion of dot-product kernelized attention* can be theoretically guaranteed by Schoenberg’s theorem [38], which motivates our proposed Schoenberg’s theorem-based Attention mechanism (SchoenbAt). Starting from Mercer’s theorem [39], we reinterpret kernelized attention through the lens of basis function expansions, establishing a unified theoretical framework. Harmonic analysis provides that kernel functions can be expressed as ex-

pansions over specific basis functions, with Bochner’s theorem and Schoenberg’s theorem yielding Fourier and polynomial basis representations, respectively. In this work, we introduce a novel polynomial basis expansion for kernelized attention. Building on the unbiased approximation of dot-product kernels using Random Maclaurin Features (**RMF**) [40], we design Random Maclaurin Feature Attention (**RMFA**) as the core component of SchoenbAt, which reduces the computational complexity of kernelized attention by restructuring its computational graph and fully leveraging computational parallelism. Since Schoenberg’s theorem requires the input space of RMFA to be bounded, SchoenbAt incorporates a Pre-Post Scaling Batch Normalization (**ppSBN**) mechanism, inspired by Batch Normalization [41]. This mechanism scales the input with trainable parameters to satisfy the theoretical constraints of Schoenberg’s theorem, while also restoring the original output magnitude, ensuring that SchoenbAt remains an unbiased estimator of the kernelized attention when the input space is unconstrained. We provide a rigorous proof of the unbiasedness and establish an approximation error bound as a function of the random feature dimension, laying a solid theoretical foundation for SchoenbAt. Our numerical experiments report the approximation errors of SchoenbAt with five different kernel functions across varying random feature dimensions, as well as the speedup ratios on simulated data of different lengths. To evaluate the performance of SchoenbAt on real-world datasets, we applied SchoenbAt to a basic Transformer [21] and compared it with various effective attention methods on the Long Range Arena (LRA) benchmark [42], which includes five tasks covering both natural language processing (Text [43], ListOps [44], Retrieval [45]) and computer vision (Pathfinder [46], Image [47]). Our results show that SchoenbAt delivers substantial acceleration and maintains competitive accuracy across diverse domains.

2 Background

Notation: In this paper, we denote matrices in uppercase bold letters like $\mathbf{X}$ and vectors in lowercase letters like x . Unless specifically mentioned, our vectors $x \in \mathbb{R}^d$, and the input to the attention mechanism, namely the query, key, and value matrices, is denoted as $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{n \times d}$. We discuss a kernel function $\mathcal{K} : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ in its corresponding reproducing kernel Hilbert space $\mathcal{H}$. A function defined on a vector space $\mathbb{R}^d$ can also accept tensor inputs of arbitrary dimensions, acting on the last dimension. Let $\ell_\rho(0, \xi)$ denotes the set $\{\|\mathbf{X}\|_\rho \leq \xi\}$.

2.1 Kernelized Attention

Derived from the widely used and fundamental softmax attention [1], which models the correlation between $\mathbf{Q}$ and $\mathbf{K}$ with $\exp(\cdot)$, constructing a convex combination of $\mathbf{V}$ as the output:

$$\text{attn}_{\text{Softmax}}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \sum_{i=1}^n \frac{\exp(\mathbf{Q}\mathbf{K}_i^\top / \sqrt{d})\mathbf{V}_i}{\sum_{j=1}^n \exp(\mathbf{Q}\mathbf{K}_j^\top / \sqrt{d})\mathbf{1}_d^\top},$$

where $\mathbf{1}_d = [1, 1, \dots, 1]^\top \in \mathbb{R}^d$ and $\frac{\mathbf{X}}{\mathbf{Y}}$ means element-wise division, kernelized attention [17, 18] proposes to generalize the function modeling sequence similarity into a kernel function, resulting in a linear combination of $\mathbf{V}$ weighted by the normalized $\mathcal{K}(\mathbf{Q}, \mathbf{K})$:

$$\text{attn}_{\mathcal{K}}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \sum_{i=1}^n \frac{\mathcal{K}(\mathbf{Q}\mathbf{K}_i^\top / \sqrt{d})\mathbf{V}_i}{\sum_{j=1}^n \mathcal{K}(\mathbf{Q}\mathbf{K}_j^\top / \sqrt{d})\mathbf{1}_d^\top}.$$

When $\mathcal{K}(\cdot) = \exp(\cdot)$, kernelized attention degenerates into softmax attention. Kernelized attention transforms the computation of attention into kernel functions, offering another perspective for studying attention. For example, efficient attention based on Nyström method views the attention matrices as kernel matrices and approximates them by low-rank approximations [48, 49, 20, 21].

2.2 Kernel Expansions in Fourier Basis

We study kernelized attention from the perspective of basis functions [50, 51, 52, 53]. As a classical result in harmonic analysis, Bochner’s theorem [25] ensures that shift-invariant kernels can be expanded as a Fourier transform of a non-negative measure $m : \mathbb{R}^d \rightarrow \mathbb{R}_{\geq 0}$:

$$\mathcal{K}(x - y) = \int_{\mathbb{R}^d} m(\omega) e^{i\omega^\top (x-y)} d\omega, m(\cdot) \geq 0. \quad (1)$$

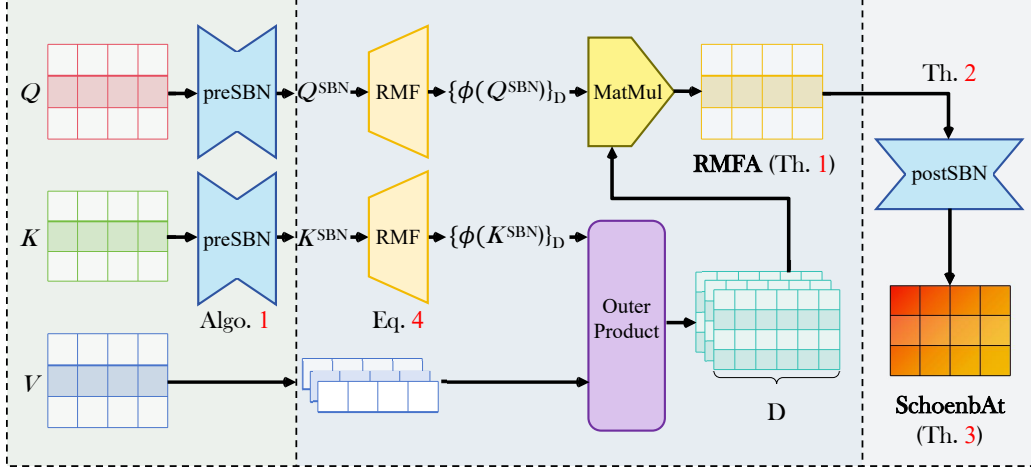


Figure 1: The module design structure of SchoenbAt.

Equation (1) allows one to represent a shift-invariant kernel as an expansion with Fourier bases, where the measure $m(\cdot)$ can be seen as the weight for each Fourier basis. Rahimi (2007) proposed a random feature sampling method [26, Algorithm 1] that unbiasedly estimates the Fourier transform when the weights are probability measures, thus providing an approximation of a shift-invariant kernel.

By properly scaling the input and representing the exponential function by a Gaussian kernel as

$$\begin{aligned} \exp(QK^\top/\sqrt{d}) &= \exp\left(\frac{Q^\top Q + K^\top K}{2\sqrt{d}} + C\right) \exp\left(-\frac{\|Q - K\|^2}{2\sqrt{d}}\right) \\ &= \exp\left[(Q^\top Q + K^\top K)/2\sqrt{d} + C\right] \mathcal{K}(Q - K), \end{aligned}$$

where $C \in \{X | X \in \mathbb{R}^{d \times d} \wedge \text{tr}(X) = 0\}$, a series of previous works [32, 33, 37] have optimized kernelized attention based on Bochner’s theorem at the cost of introducing additional computations.

3 SchoenbAt

In this section, we first attempt to understand the expansion of kernel functions through a general expression that applies to both continuous and discrete domains, where Bochner’s theorem and Schoenberg’s theorem serve as specific instances. Based on the polynomial basis expansion of dot-product kernelized attention, we design SchoenbAt which employs the RMFA and ppSBN mechanisms to improve computational efficiency while maintaining performance. We derive theoretical results analyzing the unbiasedness and approximation error of SchoenbAt to guarantee its validity.

3.1 Harmonic Expansions of SchoenbAt

The kernel function of kernelized attention admits a general expansion form. Let λ be an eigenvalue of $\mathcal{K}(\cdot)$ and $\psi(\cdot)$ be the corresponding eigenfunction, according to Mercer’s theorem [39], $\{\sqrt{\lambda_i}\psi_i\}_{i=1}^\infty$ form an orthogonal function basis for the Hilbert space $\mathcal{H}$.

In most cases, it is difficult to describe $\psi(\cdot)$ explicitly [53]. A harmonic analysis theorem ensures an expansion of a kernel function as a weighted combination of basis functions, thereby driving the research and development of kernelized attention under its theoretical foundation.

We propose that, for a given domain Ω , a kernel function can be expanded as a unified weighted combination of basis functions, which is represented as

$$\mathcal{K}(\delta) = \int_{\Omega} H(u)G(\delta, u)du, \quad (2)$$

where $H(\cdot)$ represents the non-negative weights and $G(\cdot, \cdot)$ represents the basis functions.

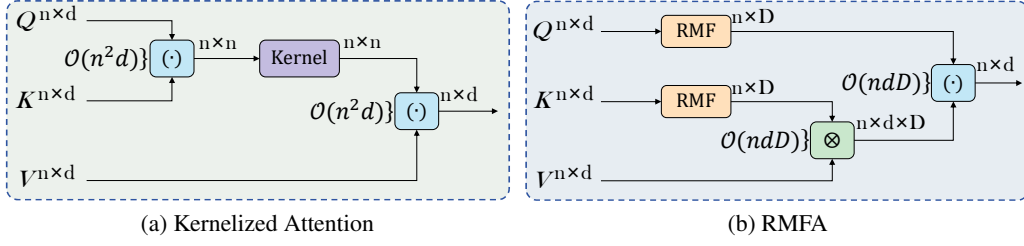


Figure 2: Computation graphs for kernelized attention and RMFA. In each figure, the data on the left represents the input to the attention layer. Here, operators $(\cdot)$ and $\otimes$ respectively denote matrix multiplication and outer product. Each step’s main computation cost is marked on the left side of the operators, and the dimensions of data and intermediate results are indicated with superscripts.

When $\delta = x - y$, the integration domain Ω is $\mathbb{R}^d$, the weights $H(\cdot)$ are taken as a non-negative measure $m(\cdot)$, and the basis functions $G(\delta, u) = \exp(iu^T \delta)$, Equation (2) degenerates into Equation (1), which is the expansion form of shift-invariant kernels with Fourier bases.

Schoenberg’s theorem [38] characterizes positive definite functions on the unit sphere in a Hilbert space, stating that a dot-product kernel $\mathcal{K}(\langle x, y \rangle) = \mathcal{K}(x, y) : \ell_2(0, 1) \times \ell_2(0, 1) \rightarrow \mathbb{R}$ is positive definite if and only if $\mathcal{K}(\cdot)$ is an analytic function admitting a Maclaurin expansion with only non-negative coefficients, written as

$$\mathcal{K}(\langle x, y \rangle) = \sum_{i=0}^{\infty} a_i \langle x, y \rangle^i, a_i \geq 0.$$

We note that if a discrete function $f(z)$ is defined on $\mathbb{Z}$, such as the n -th term of the Maclaurin expansion of a dot-product kernel, its summation can be expressed as $\int_{\mathbb{R}} f(\lfloor u \rfloor) du = \sum_{z=0}^{\infty} f(z)$. Thus, under the guarantee of Schoenberg’s theorem, let the integration domain $\Omega = \mathbb{Z}$, the weights $H(\cdot)$ be the non-negative Maclaurin coefficients $H(u) = \frac{1}{u!} \mathcal{K}^{(u)}(0)$, and the basis function $G(\delta, u) = \delta^u$ be polynomial basis, we obtain another instance of Equation (2) as

$$\mathcal{K}(\langle x, y \rangle) = \int_{\mathbb{Z}} \frac{1}{u!} \mathcal{K}^{(u)}(0) \langle x, y \rangle^u du, \quad (3)$$

where $\mathcal{K}^{(u)}(x)$ is the u -th derivative of $\mathcal{K}(\cdot)$ at point x . Schoenberg’s theorem guarantees the equivalence between analytic functions with non-negative Maclaurin coefficients and dot-product kernels, where the coefficients represent the weights of the polynomial bases.

3.2 Module Design of SchoenbAt

We observe that kernel function $\mathcal{K}(\mathbf{Q}\mathbf{K}^\top / \sqrt{d})$ in kernelized attention is inherently a dot-product kernel. Based on Equation (3), we expand dot-product kernelized attention and express it in the following form:

$$\text{attn}_{\mathcal{K}}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \sum_{i=1}^n \frac{\int_{\mathbb{Z}} \frac{1}{u!} \mathcal{K}^{(u)}(0) (\mathbf{Q}\mathbf{K}^\top / \sqrt{d})^u du \mathbf{V}_i}{\sum_{j=1}^n \int_{\mathbb{Z}} \frac{1}{u!} \mathcal{K}^{(u)}(0) (\mathbf{Q}\mathbf{K}^\top / \sqrt{d})^u du \mathbf{1}_d^\top}.$$

The polynomial expansion of $\mathcal{K}(\langle x, y \rangle)$ can be approximated by the inner product of RMF [40] vectors $\Phi_{\mathcal{K}}(x)$ and $\Phi_{\mathcal{K}}(y)$ in arbitrary D s:

$$\mathbb{E} [\Phi_{\mathcal{K}}(x) \Phi_{\mathcal{K}}^\top(y)] = \int_{\mathbb{Z}} \frac{1}{u!} \mathcal{K}^{(u)}(0) \langle x, y \rangle^u du, \quad (4)$$

where $\Phi_{\mathcal{K}}(x) = \sqrt{1/D} [\phi_1(x), \dots, \phi_D(x)]$, each $\phi_i(x) = \sqrt{a_N p^{N+1}} \prod_{j=1}^N \langle \omega_j, x \rangle$ is independent of i , $p > 1$ is a predetermined constant, N is sampled from $\mathbb{N}$ with $\mathbb{P}[N = \eta] = \frac{1}{p^{\eta+1}}$, a_N is the N th coefficient of the Maclaurin expansion of the kernel function $\mathcal{K}(\cdot)$, and $\omega_j \in \mathbb{R}^d$ is a Rademacher vector independent of j .

$\mathcal{K}$	$f(\langle x, y \rangle)$	a_N
exp	$\exp(\langle x, y \rangle)$	$1/N!$
inv	$1/(1 - \langle x, y \rangle)$	1
logi	$1 - \log(1 - \langle x, y \rangle)$	$\frac{1}{\min(1, N)}$
trigh	$\sinh(\langle x, y \rangle) + \cosh(\langle x, y \rangle)$	$1/N!$
sqrt	$2 - \sqrt{1 - \langle x, y \rangle}$	$\frac{\max(1, 2N-3)}{2^N N!}$

Table 1: The dot-product kernels we study in this work along with their non-negative Maclaurin coefficients, where $\mathcal{K}(\cdot) = f(\langle x, y \rangle)$.

Algorithm 1 ppSBN in SchoenbAt

Input: attention input $\mathbf{Q}, \mathbf{K}, \mathbf{V}$

Parameter: trainable parameter β, γ ; hyperparameter ε

Output: attention output att

- 1: $\mathbf{Q}', \mathbf{K}' \leftarrow \frac{\mathbf{Q} - \mu_{\mathbf{Q}}}{\sqrt{\sigma_{\mathbf{Q}} + \varepsilon}}, \frac{\mathbf{K} - \mu_{\mathbf{K}}}{\sqrt{\sigma_{\mathbf{K}} + \varepsilon}}$
 - 2: $\mathbf{Q}^{\text{SBN}}, \mathbf{K}^{\text{SBN}} \leftarrow \frac{\mathbf{Q}'}{\|\mathbf{Q}'\|_2}, \frac{\mathbf{K}'}{\|\mathbf{K}'\|_2}$
 - 3: $att \leftarrow \text{RMFA}_{\mathcal{K}}(\mathbf{Q}^{\text{SBN}}, \mathbf{K}^{\text{SBN}}, \mathbf{V})$
 - 4: $att \leftarrow \gamma \cdot att^{\beta}$
-

We establish the relationship between dot-product kernelized attention and the inner product of RMF through the polynomial basis expansion guaranteed by Schoenberg’s theorem, thereby designing RMFA, a core component of SchoenbAt.

Theorem 1. Suppose $\mathcal{K}(\mathbf{Q}\mathbf{K}^{\top}/\sqrt{d})$ is a dot-product kernel $\mathcal{K}(\mathbf{Q}/d^{\frac{1}{4}}, \mathbf{K}/d^{\frac{1}{4}})$ with only non-negative Maclaurin coefficients, $\Phi_{\mathcal{K}}(\cdot)$ defines a random Maclaurin feature map of $\mathcal{K}$, the output of the attention can be estimated by RMFA.

$$\text{attn}_{\mathcal{K}}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) \approx \frac{\Phi_{\mathcal{K}}(\mathbf{Q}/d^{\frac{1}{4}}) \sum_{i=1}^n \left[\Phi_{\mathcal{K}}^{\top}(\mathbf{K}_i/d^{\frac{1}{4}}) \right] \otimes \mathbf{V}_i}{\Phi_{\mathcal{K}}(\mathbf{Q}/d^{\frac{1}{4}}) \sum_{j=1}^n \left[\Phi_{\mathcal{K}}^{\top}(\mathbf{K}_j/d^{\frac{1}{4}}) \right] \mathbf{1}_d^{\top}} = \text{RMFA}_{\mathcal{K}}(\mathbf{Q}, \mathbf{K}, \mathbf{V}),$$

where $x \otimes y \in \mathbb{R}^{d \times d}$ represents the outer product of vectors.

The proof is given in Appendix A.1. Figure 2 illustrates the attention computation paths for the original kernelized attention and RFMA which factors out $\Phi_{\mathcal{K}}(\mathbf{Q}/d^{\frac{1}{4}})$. In Figure 2a, the direct computation of the attention matrix results in a time complexity of $O(n^2d)$. In contrast, the computation graph shown in Figure 2b changes the time complexity to $O(ndD)$, which can be further decreased by parallelizing the outer product. In most cases, where $n \gg D$, RMFA in SchoenbAt can significantly reduce the computational overhead of attention. In this paper, we study five dot product kernels with only non-negative Maclaurin coefficients as examples of SchoenbAt, which are shown in Table 1, where kernel $\exp(\cdot)$ corresponds to Softmax attention. It is worth noting that the domains of dot-product kernels $\text{inv}(\cdot)$, $\text{logi}(\cdot)$, and $\text{sqrt}(\cdot)$ are restricted to inputs less than 1.

According to Schoenberg’s theorem, when the input space of kernelized attention is constrained to $\ell_2(0, 1)$, RMFA serves as an unbiased estimate of the dot-product kernelized attention. However, in practical applications, the input space of the attention module is uncontrolled, and the output will be strongly affected when the inputs are scaled. Inspired by the idea of reshaping data expectations and variances through Batch Normalization [41], SchoenbAt addresses this issue through a two-stage regularization mechanism called pre-post Scaling Batch Normalization.

As shown in Algorithm 1, SchoenbAt first normalizes and scales $\mathbf{Q}, \mathbf{K}$ to ensure that $\mathbf{Q}^{\text{SBN}}, \mathbf{K}^{\text{SBN}} \in \ell_2(0, 1)$, then feeds them into RMFA for computation. The mean value matrices $\mu_{\mathbf{Q}}, \mu_{\mathbf{K}}$ and variance matrices $\sigma_{\mathbf{Q}}, \sigma_{\mathbf{K}}$ are unsqueezed from vectors, and ε is the all- ε s matrix. The output of RMFA is rescaled by trainable parameters β and γ , serving as the final output of SchoenbAt. We can mathematically demonstrate that when approximating attn_{exp} , the ppSBN mechanism can restore the original output while ensuring that the input space remains constrained.

Theorem 2. Suppose $\mathbf{Q}^{\text{SBN}}, \mathbf{K}^{\text{SBN}}$ denote $\mathbf{Q}, \mathbf{K}$ after scaling and batch normalization, we have

$$\text{RMFA}_{\text{exp}}(\mathbf{Q}^{\text{SBN}}, \mathbf{K}^{\text{SBN}}, \mathbf{V}) \approx \frac{1}{t} \left[\frac{1}{s} \odot \text{RMFA}_{\text{exp}}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) \right]^{\frac{1}{r}},$$

where $\odot$ denotes element-wise product, $r = \|\mathbf{Q}'\|_2 \|\mathbf{K}'\|_2 \sqrt{(\sigma_{\mathbf{Q}} + \varepsilon)(\sigma_{\mathbf{K}} + \varepsilon)}$ depends on the data, $t = \sum_{i=1}^n \exp \left[\left(\mathbf{Q}\mathbf{K}_i^{\top} - \mu_{\mathbf{Q}}\mathbf{K}_i^{\top} \right) / r\sqrt{d} \right] \mathbf{1}_n^{\top} / \sum_{i=1}^n \exp \left[\left(\mathbf{Q}\mathbf{K}_i^{\top} \mathbf{1}_n^{\top} - \mu_{\mathbf{Q}}\mathbf{K}^{\top} \right) / r\sqrt{d} \right]$ and $s = \frac{\|\exp(\mathbf{Q}\mathbf{K}^{\top}/\sqrt{d})\|_1^{\frac{1}{r}}}{\|\exp(\mathbf{Q}\mathbf{K}^{\top}/\sqrt{d})\|_1} V^{r-1}$.

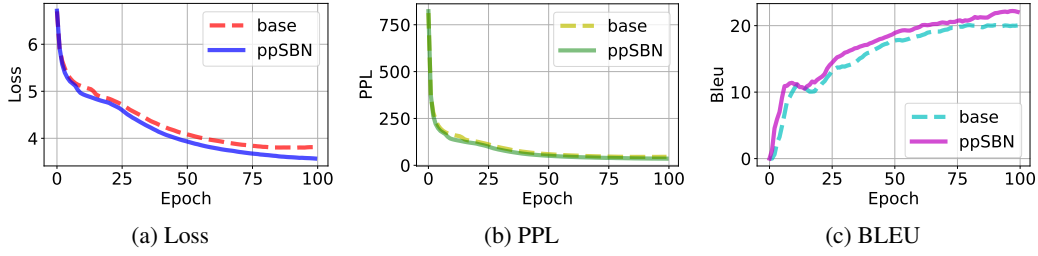


Figure 3: The loss, perplexity (lower is better), and Bleu scores (higher is better) of the traditional Transformer with and without ppSBN across training epochs on machine translation task of Multi30k dataset. In each plot, solid lines represent the Transformer with ppSBN, while dashed lines represent the Transformer without ppSBN.

We give the proof of Theorem 2 in Appendix A.2. The trainable parameters γ and β in Algorithm 1 fit the $s \odot t^r$ and r in Theorem 2, thus preserving the scale of data. In addition, combining Theorem 1 and 2, we can obtain that the approximate equivalence is in the sense of expectation, i.e.

$$\mathbb{E} [\text{RMFA}_{\text{exp}}(\mathbf{Q}^{\text{SBN}}, \mathbf{K}^{\text{SBN}}, \mathbf{V})] = \mathbb{E} \left[\frac{1}{t} \left[\frac{1}{s} \odot \text{RMFA}_{\text{exp}}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) \right]^{\frac{1}{r}} \right]. \quad (5)$$

We conducted a toy experiment to verify whether the parameters γ and β in ppSBN can be appropriately trained. Figure 3 shows the loss, perplexity, and Bleu scores of a machine translation experiment on the Multi30K dataset [54], where we used the traditional Transformer [12] as the base model to avoid the influence of randomness induced by RMFA and incorporated the ppSBN mechanism before and after the softmax attention layer for comparison. The model with the ppSBN mechanism slightly outperforms the base model in all three metrics, demonstrating that γ and β can be trained end-to-end with other parameters in a few epochs without degrading model performance.

Let $\overrightarrow{\text{SBN}}(\mathbf{X}) : \mathbf{X} \rightarrow \mathbf{X}^{\text{SBN}}$ represents the pre-SBN in algorithm 1, and $\overleftarrow{\text{SBN}}(\mathbf{X}; \gamma, \beta) : \mathbf{X} \rightarrow (\gamma \mathbf{X})^\beta$ denotes the post-SBN, we provide SchoenbAt as

$$\text{SchoenbAt}_{\mathcal{K}}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \overleftarrow{\text{SBN}}(\text{RMFA}_{\mathcal{K}}(\overrightarrow{\text{SBN}}(\mathbf{Q}), \overrightarrow{\text{SBN}}(\mathbf{K}), \mathbf{V}); \gamma, \beta).$$

As shown in Figure 1, the input and output of SchoenbAt are the same as those of kernelized attention, making it a drop-in replacement for the attention mechanism in machine learning architectures.

3.3 Theoretical Analyses of SchoenbAt

Since the use of RMF introduces stochasticity into the approximation process, rigorous theoretical analysis is required to ensure the reliability and effectiveness of the proposed SchoenbAt mechanism. As an initial guarantee, we establish the unbiasedness formally proven in Appendix A.3.

Theorem 3. *Assuming that the parameters γ and β are properly trained to restore the attention output, $\Phi_{\mathcal{K}}(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^D$ defines a random Maclaurin feature map for a dot-product kernel $\mathcal{K}(\cdot)$, then for any attention inputs $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{n \times d}$, we have $\mathbb{E}[\text{SchoenbAt}_{\mathcal{K}}(\mathbf{Q}, \mathbf{K}, \mathbf{V})] = \text{attn}_{\mathcal{K}}(\mathbf{Q}, \mathbf{K}, \mathbf{V})$.*

Theorem 3 provides an important theoretical result, indicating that as the number of computations approaches infinity, the expectation of SchoenbAt converges to dot-product kernelized attention. This result establishes that, in the asymptotic case, SchoenbAt is an unbiased estimator of kernelized attention, offering a strong foundation for its approximation effectiveness. However, in practical learning tasks, SchoenbAt is computed only once during each forward pass, meaning that the approximation is subject to finite sampling and computational constraints, which impose strict requirements on estimation stability.

To address this issue, we derive an upper bound on the approximation error of SchoenbAt under the assumption that the norm of $\mathbf{V}$ is bounded. This bound quantitatively characterizes the discrepancy between the output of SchoenbAt and that of the exact kernelized attention, thereby providing a theoretical measure of the approximation quality.

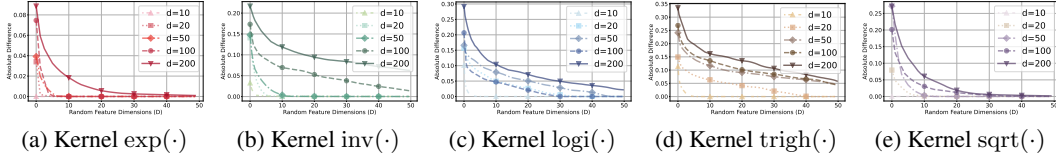


Figure 4: The average absolute differences between SchoenbAt and kernelized attention on five dot-product kernels. Plots of different colors represent different dimensions of attention input.

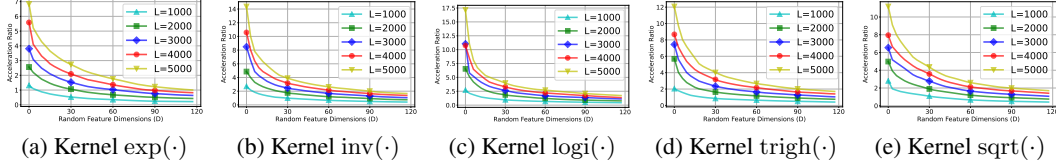


Figure 5: The speedup ratio of SchoenbAt relative to kernelized attention on five dot-product kernels. Plots of different colors represent different sequence lengths of attention input.

Theorem 4. Suppose attention inputs $|\mathbf{V}_{ij}| \leq S$, assuming that the parameters γ and β are properly trained to restore the attention output, $\Phi_{\mathcal{K}}(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^D$ defines a random Maclaurin feature map for a dot-product kernel $\mathcal{K}(\cdot)$. Let data $\mathcal{D} = (\mathbf{Q}, \mathbf{K}, \mathbf{V})$, then for any $\epsilon > 0$, we have

$$\mathbb{P}(|\text{SchoenbAt}_{\mathcal{K}}(\mathcal{D}) - \text{attn}_{\mathcal{K}}(\mathcal{D})| > \epsilon) \leq 2D \exp\left(-\frac{D\epsilon^2}{2S^2d^2}\right).$$

Detailed proof of Theorem 4 is given in Appendix A.4. We have demonstrated that the approximation performance of SchoenbAt theoretically improves as D increases and S or d decreases. Specifically, increasing D leads to a more accurate approximation of kernelized attention. On the other hand, reducing S or d helps to minimize the scale of data which determines the cost of approximation.

In practical applications, the dimensionality of random features D can be adjusted flexibly to balance the approximation accuracy and computational speed. For example, a higher value of D may be suitable for scenarios where precision is critical, while a smaller D may be employed in real-time applications where speed is more important than perfect accuracy. This trade-off allows SchoenbAt to be efficiently deployed across a wide range of machine learning tasks with varying performance and resource requirements.

4 Experiments

In this section, we report a numerical experiment testing the speedup and error between SchoenbAt and kernelized attention, validations on the LRA benchmark, and an ablation study. We implement RMF based on Wacker [28] in each experiment and set the hyperparameter $p = 2$. Our code is available at anonymous.4open.science/r/SchoenbAt-D6CE.

4.1 Approximation Error and Speedup

In our first numerical experiment, we test the approximation error of SchoenbAt on five dot-product kernels in Table 1. We sample $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ from a $100 \times d$ -dimensional standard Gaussian distribution, where d varies from 10 to 200, and compute SchoenbAt with D varying from 10 to 50. The parameters γ and β are set to the ideally trained values. The experiment under each setting is repeated 100 times, and the error incurred by randomness is calculated as the average absolute difference between kernelized attention and SchoenbAt.

From Figure 4, one can easily observe that as D increases, the approximation error of SchoenbAt decreases rapidly, with the approximation error for SchoenbAt_{exp} being the smallest and decreasing the fastest, while the errors for SchoenbAt_{logi} and SchoenbAt_{trigh} are larger, with SchoenbAt_{trigh} exhibiting the slowest rate of decrease. In each case, the increase in the original data dimension d

Model	RF	Time						Accuracy (%)					
		T	L	R	P	I	Avg.	T	L	R	P	I	Avg.
Softmax	✗	1.000	1.000	1.000	1.000	1.000	1.000	63.31	37.30	75.04	68.64	38.69	56.59
Reformer	✗	0.181	0.229	0.330	0.851	0.484	0.415	64.76	23.99	73.36	50.29	37.23	49.92
Nystromformer	✗	0.158	0.188	0.280	0.484	0.482	0.318	65.66	37.75	69.81	50.33	37.21	52.15
Skyformer	✗	0.398	0.544	0.405	1.003	0.853	0.641	65.25	20.82	70.46	51.92	36.65	49.02
Informer	✗	0.161	0.227	0.422	1.021	0.807	0.528	62.63	22.58	69.77	50.39	35.72	48.21
Bigbird	✗	0.311	0.632	0.502	1.152	0.421	0.603	64.34	37.65	75.05	50.09	37.47	52.92
Cosformer	✗	0.111	0.195	0.198	0.265	0.271	0.208	63.84	23.49	75.32	50.27	33.72	49.32
RFA	✓	0.783	0.836	0.900	0.727	0.866	0.822	65.22	37.15	77.84	59.91	36.09	55.24
Performer	✓	0.349	0.412	0.294	0.623	0.455	0.426	63.79	18.80	75.69	50.37	31.98	48.12
SchoenbAt _{exp}	✓	0.076	0.236	0.105	0.322	0.214	0.191	64.12	38.21	70.75	72.35	36.23	56.33
SchoenbAt _{inv}	✓	0.209	0.193	0.194	0.412	0.268	0.255	63.54	38.56	79.37	68.32	38.33	57.62
SchoenbAt _{logi}	✓	0.167	0.192	0.106	0.437	0.289	0.238	64.62	39.77	70.35	71.42	38.35	56.90
SchoenbAt _{trigh}	✓	0.219	0.232	0.533	0.434	0.254	0.334	64.23	39.92	70.71	71.39	38.35	56.92
SchoenbAt _{sqrt}	✓	0.189	0.211	0.214	0.321	0.229	0.233	64.25	38.66	79.36	72.32	38.49	58.61

Table 2: Experimental results on three LRA benchmark tasks. We compared softmax attention, efficient attention improvements, and our method: SchoenbAt, where **RF** indicates whether it is a random feature method. We report the training time (normalized to Softmax) and accuracy on the test set, where **T, L, R, P, I** respectively represent Text, Listops, Retrieval, Pathfinder, Image tasks, and **Avg.** is the average across the three tasks. The data in **red** and **blue** respectively represent the best performance among all methods and random feature methods.

leads to a higher approximation error for SchoenbAt, which still achieves a good approximation with a random feature dimension D that is lower than the data dimension d . This experimental result confirms our Theorem 4.

Our second numerical experiment evaluates the speedup of SchoenbAt with five dot-product kernels relative to kernelized attention. The Q , K , and V were drawn from a standard Gaussian distribution with $d = 50$, and the number of attention heads was set to 8. The sequence length L varied between 1000 and 5000, while the random feature dimension D ranged from 2 to 120. Each experiment under the same configuration was repeated 100 times, and the speedup was calculated as the ratio of the time consumption of kernelized attention to that of SchoenbAt.

Figure 5 illustrates a similar trend across all experiments: the speedup of SchoenbAt becomes more pronounced as L increases but diminishes as D grows. The speedup for SchoenbAt_{logi} is notably higher than the others, while SchoenbAt_{exp} shows the least speedup. When $L \gg D$, SchoenbAt achieves remarkable acceleration, aligning with our complexity analysis.

4.2 Validations on LRA Benchmark

In our experiments on real-world datasets, we replaced the attention mechanism in Transformer with SchoenbAt, Nyström-based attention [55, 20, 21], linear attention [49, 56, 57], and random feature-based attention [32, 33], comparing their performance on the LRA benchmark [42]. The LRA is a standard test used to assess and compare different Transformer models’ ability to handle long sequence data. We organized our experiments around five tasks from this benchmark, consisting of natural language processing tasks: byte-level **Text** classification, long **Listops**, byte-level document **Retrieval**, and computer vision tasks: long-range **Pathfinder**, pixel-level **Image** classification.

We conducted all experiments on one NVIDIA RTX A6000 48G, working on a baseline model implemented by Chen [21] with PyTorch. Our model has an embedding dimension of 64, a hidden dimension of 128, 2 layers, and uses 2 attention heads. The batch size is selected depending on the task: 16 for Text and Retrieval, 32 for Listops, 128 for Pathfinder and 256 for Image. The default random projection dimension of each random feature-based method is set to 128, and the ϵ for ppSBN is set to 10^{-13} . Each training session involves 1000 steps of initialization and 10000 steps of optimization. All methods in our experiments share the same hyperparameters of the Transformer, and we report the average time consumption and accuracy over 50 repetitions of training.

Table 2 presents the experimental results on the LRA benchmark, and memory cost is reported in Appendix B Table 4. In all five LRA tasks, SchoenbAt with five different dot-product kernels significantly improved computational efficiency while maintaining competitive accuracy. Notably,

Ablations	base	base+RMFA					base+ppSBN	base+RMFA+ppSBN				
		exp($\cdot$)	inv($\cdot$)	logi($\cdot$)	trigh($\cdot$)	sqrt($\cdot$)		exp($\cdot$)	inv($\cdot$)	logi($\cdot$)	trigh($\cdot$)	sqrt($\cdot$)
Time	1.000	0.264	0.267	0.264	0.264	0.265	0.785	0.230	0.232	0.230	0.242	0.233
Accuracy (%)	65.58	63.04	62.47	63.11	63.42	63.47	65.1	71.54	70.53	71.25	71.6	71.65

Table 3: Ablation study on LRA-Text, reporting normalized training time and prediction accuracy under different settings, illustrating the individual and combined effects of RMFA and ppSBN.

it outperformed other methods by a large margin in both the ListOps and Pathfinder tasks in terms of accuracy. Although SchoenbAt achieved less acceleration than Cosformer on the Pathfinder task, it yielded a substantial accuracy improvement: from 50.27% to 71.16% on average across the five kernels. Although SchoenbAt does not achieve the global optimum, it performs comparably to the best method in terms of runtime on the ListOps task and accuracy on the Text task, while significantly outperforming it on the other metric. Overall, SchoenbAt_{exp} achieved the highest average speedup, while SchoenbAt_{sqrt} attained the best average accuracy.

The evaluation on the LRA benchmark demonstrates SchoenbAt’s capability in modeling both natural language dependencies and pixel-wise relationships. In addition, the above experimental results reveal a phenomenon: for each task, SchoenbAt with different kernel functions exhibits significantly varying performance. We believe this is because SchoenbAt approximates kernelized attention modeling sequence correlations with different kernel functions. We recommend using SchoenbAt_{exp} in most cases, but other dot-product kernels can be substituted depending on the application scenario.

4.3 Ablation Study

To evaluate the individual contributions of RMFA and ppSBN in a controlled setting, we chose the Text task from LRA for ablation as it is the most representative and general-purpose sequence modeling benchmark across the dataset. Each experimental result is obtained as the average from leave-one-out cross-validation. We use standard softmax attention as the base model, and consider the following four configurations:

- Base: The original Transformer with softmax attention, serving as a baseline.
- Base+RMFA: Replacing softmax attention with RMFA with dot-product kernels in Table 1, measuring the effect of kernel approximation alone.
- Base+ppSBN: Inserts ppSBN around softmax attention to examine its standalone impact.
- Base+RMFA+ppSBN (SchoenbAt): Our full method combining both RMFA with dot-product kernels in Table 1 and ppSBN.

This setup allows us to assess how RMFA contributes to efficiency and how ppSBN enhances the accuracy and stability of the approximation.

We report computational speedup (relative to the Base model) and prediction accuracy in Table 3 to comprehensively evaluate each configuration. Using RMFA alone accelerates training but leads to a drop in accuracy, validating the approximation property claimed in Theorem 1. In contrast, applying ppSBN alone yields a marginal improvement in efficiency while largely preserving or even improving accuracy, consistent with Theorem 2. SchoenbAt combines the strengths of both RMFA and ppSBN, achieving substantial speedup along with superior accuracy.

5 Conclusion

Under the guidance of Schoenberg’s theorem, we expand dot-product kernelized attention and propose SchoenbAt as an efficient approximation. We rigorously establish its unbiasedness and derive approximation guarantees, both of which are validated through numerical simulations and empirical results on the LRA benchmark.

This work introduces a harmonic analysis-based theoretical framework for understanding and constructing kernel expansions in attention mechanisms. Future research may build upon this foundation by exploring alternative basis expansions or developing generalized random feature sampling techniques to approximate the unified form of kernelized attention more effectively.

References

- [1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.
- [2] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. In *NeurIPS* 32, 2019.
- [3] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.
- [4] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *NeurIPS* 35, 2022.
- [5] Yizeng Han, Dongchen Han, Zeyu Liu, Yulin Wang, Xuran Pan, Yifan Pu, Chao Deng, Junlan Feng, Shiji Song, and Gao Huang. Dynamic perceiver for efficient visual recognition. In *ICCV*, 2023.
- [6] Yizeng Han, Yifan Pu, Zihang Lai, Chaofei Wang, Shiji Song, Junfeng Cao, Wenhui Huang, Chao Deng, and Gao Huang. Learning to weight samples for dynamic early-exiting networks. In *ECCV*, 2022.
- [7] Xuran Pan, Tianzhu Ye, Dongchen Han, Shiji Song, and Gao Huang. Contrastive language-image pre-training with knowledge graphs. *NeurIPS* 35, 2022.
- [8] Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. In *CVPR*, 2024.
- [9] Wei Wang, Shitong Wan, Hu Jin, Dong Liu, Hongjun Zhang, Yun Zhou, and Xianfang Wang. ResaPred: A deep residual network with self-attention to predict protein flexibility. In *TCBB*, 2025.
- [10] Guanghui Li, Ziyang Cao, Cheng Liang, Qiu Xiao, and Jiawei Luo. MCHAN: Prediction of human microbe-drug associations based on multiview contrastive hypergraph attention network. *Current Bioinformatics*, 20(1):70–86, 2025.
- [11] Mingguang Shi and Xuefeng Li. Addressing scalability and managing sparsity and dropout events in single-cell representation identification with ZIGACL. *Briefings in Bioinformatics*, 26(1):bbae703, 2025.
- [12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS* 30, 2017.
- [13] Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5):1122–1136, 2023.
- [14] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv:2412.15115*, 2024.
- [15] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv:2412.19437*, 2024.
- [16] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv:2412.15115*, 2024.
- [17] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are Rnns: fast autoregressive transformers with linear attention. In *ICML*, 2020.
- [18] Yao-Hung Hubert Tsai, Shaojie Bai, Makoto Yamada, Louis-Philippe Morency, and Ruslan Salakhutdinov. Transformer dissection: An unified understanding for transformer’s attention via the lens of kernel. In *EMNLP-IJCNLP*, 2019.
- [19] Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv:2006.04768*, 2020.
- [20] Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. Nyströmformer: A nyström-based algorithm for approximating self-attention. In *AAAI*, 2021.
- [21] Yifan Chen, Qi Zeng, Heng Ji, and Yun Yang. Skyformer: Remodel self-attention with gaussian kernel and Nyström method. In *NeurIPS* 34, 2021.
- [22] Yingyi Chen, Qinghua Tao, Francesco Tonin, and Johan Suykens. Primal-attention: Self-attention through asymmetric kernel svd in primal representation. In *NeurIPS* 36, 2024.

- [23] Hao Li, Yang Nan, Javier Del Ser, and Guang Yang. Large-kernel attention for 3D medical image segmentation. *Cognitive Computation*, 16(4):2063–2077, 2024.
- [24] Genyun Sun, Zhaojie Pan, Aizhu Zhang, Xiuping Jia, Jinchang Ren, Hang Fu, and Kai Yan. Large kernel spectral and spatial attention networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023.
- [25] Walter Rudin. *Fourier Analysis on Groups*. Interscience Publishers, 1962.
- [26] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *NeurIPS 20*, 2007.
- [27] Luigi Carratino, Alessandro Rudi, and Lorenzo Rosasco. Learning with SGD and random features. In *NeurIPS 31*, 2018.
- [28] Jonas Wacker and Maurizio Filippone. Local random feature approximations of the Gaussian kernel. *Procedia Computer Science*, 207:987–996, 2022.
- [29] Nicolas Langrené, Xavier Warin, and Pierre Gruet. New random projections for isotropic kernels using stable spectral distributions. *arXiv:2411.02770*, 2024.
- [30] Kyohei Atarashi, Subhransu Maji, and Satoshi Oyama. Random feature maps for the itemset kernel. In *AAAI*, 2019.
- [31] Andrew Cotter, Joseph Keshet, and Nathan Srebro. Explicit approximations of the Gaussian kernel. *arXiv:1109.4603*, 2011.
- [32] Hao Peng, Nikolaos Pappas, Dani Yogatama, Roy Schwartz, Noah Smith, and Lingpeng Kong. Random feature attention. In *ICLR*, 2021.
- [33] Krzysztof Marcin Choromanski, Valerii Likhoshesterov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Benjamin Belanger, Lucy J Colwell, and Adrian Weller. Rethinking attention with performers. In *ICLR*, 2021.
- [34] Zailin Ma, Jiansheng Yang, and Yaodong Yang. Random feature models with learnable activation functions. *arXiv:2411.19468*, 2024.
- [35] Yingxu Wang, Long Chen, Tianjun Li, and Yifei Tian. Random maclaurin feature-based fuzzy clustering. In *ICSS*, 2020.
- [36] Ivano Lauriola, Michele Donini, and Fabio Aiolli. Learning dot-product polynomials for multiclass problems. In *ESANN*, 2017.
- [37] Duke Nguyen, Aditya Joshi, and Flora D. Salim. Spectraformer: A unified random feature framework for transformer. *arXiv:2405.15310*, 2024.
- [38] Isaac Jacob Schoenberg. Positive definite functions on spheres. *Duke Mathematical Journal*, 9(1):96–108, 1942.
- [39] James Mercer. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society A*, 209:415–446, 1909.
- [40] Purushottam Kar and Harish Karnick. Random feature maps for dot product kernels. In *AISTATS*, 2012.
- [41] Sergey Ioffe and Christian Szegedy. Batch normalization: accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.
- [42] Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Don Metzler. Long range arena : A benchmark for efficient transformers. In *ICLR*, 2021.
- [43] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *ACL*, 2011.
- [44] Nikita Nangia and Samuel Bowman. ListOps: A diagnostic dataset for latent tree learning. In *NAACL*, 2018.
- [45] Dragomir R. Radev, Pradeep Muthukrishnan, and Vahed Qazvinian. The ACL Anthology network corpus. In *NLP4DL*, 2009.

- [46] R. Houtkamp and P. R. Roelfsema. Parallel and serial grouping of image elements in visual perception. *Journal of Experimental Psychology: Human Perception and Performance*, 36(6):1443–1459, 2010.
- [47] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.
- [48] Jiachen Lu, Junge Zhang, Xiatian Zhu, Jianfeng Feng, Tao Xiang, and Li Zhang. Softmax-free linear transformers. *International Journal of Computer Vision*, 132(8):3355–3374, 2024.
- [49] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wan Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *AAAI*, 2020.
- [50] Gabor Horvath. CMAC neural network as an SVM with B-spline kernel functions. In *ITC*, 2003.
- [51] Youngmin Cho and Lawrence Saul. Kernel methods for deep learning. In *NeurIPS 22*, 2009.
- [52] Li Zhang, Weida Zhou, and Licheng Jiao. Wavelet support vector machine. *IEEE Transactions on Systems, Man, and Cybernetics*, 34(1):34–39, 2004.
- [53] Ingo Steinwart, Don Hush, and Clint Scovel. An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels. *IEEE Transactions on Information Theory*, 52(10):4635–4643, 2006.
- [54] Desmond Elliott, Stella Frank, Khalil Sima’an, and Lucia Specia. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, 2016.
- [55] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *ICLR*, 2020.
- [56] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. In *NeurIPS 33*, 2020.
- [57] Zhen Qin, Weixuan Sun, Hui Deng, Dongxu Li, Yunshen Wei, Baohong Lv, Junjie Yan, Lingpeng Kong, and Yiran Zhong. cosformer: Rethinking softmax in attention. *arXiv preprint arXiv:2202.08791*, 2022.

Appendix

A Proofs

A.1 Proof of Theorem 1

Proof. We begin with

$$\text{attn}_{\mathcal{K}}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \sum_{i=1}^n \frac{\int_{\mathbb{Z}} \frac{1}{u!} \mathcal{K}^{(u)}(0) (\mathbf{Q} \mathbf{K}^{\top} / \sqrt{d})^u du \mathbf{V}_i}{\sum_{j=1}^n \int_{\mathbb{Z}} \frac{1}{u!} \mathcal{K}^{(u)}(0) (\mathbf{Q} \mathbf{K}^{\top} / \sqrt{d})^u du \mathbf{1}_d^{\top}},$$

by Equation (4) we have $\int_{\mathbb{Z}} \frac{1}{u!} \mathcal{K}^{(u)}(0) (\mathbf{Q} \mathbf{K}^{\top} / \sqrt{d})^u du \approx \Phi_{\mathcal{K}}(\mathbf{Q} / d^{\frac{1}{4}}) \Phi_{\mathcal{K}}^{\top}(\mathbf{K}_i / d^{\frac{1}{4}})$, thus

$$\begin{aligned} \text{attn}_{\mathcal{K}}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) &\approx \sum_{i=1}^n \frac{[\Phi_{\mathcal{K}}(\mathbf{Q} / d^{\frac{1}{4}}) \Phi_{\mathcal{K}}^{\top}(\mathbf{K}_i / d^{\frac{1}{4}})] \mathbf{V}_i}{\sum_{j=1}^n \Phi_{\mathcal{K}}(\mathbf{Q} / d^{\frac{1}{4}}) \Phi_{\mathcal{K}}^{\top}(\mathbf{K}_j / d^{\frac{1}{4}}) \mathbf{1}_d^{\top}} \\ &= \frac{\Phi_{\mathcal{K}}(\mathbf{Q} / d^{\frac{1}{4}}) \sum_{i=1}^n [\Phi_{\mathcal{K}}^{\top}(\mathbf{K}_i / d^{\frac{1}{4}})] \otimes \mathbf{V}_i}{\Phi_{\mathcal{K}}(\mathbf{Q} / d^{\frac{1}{4}}) \sum_{j=1}^n [\Phi_{\mathcal{K}}^{\top}(\mathbf{K}_j / d^{\frac{1}{4}})] \mathbf{1}_d^{\top}} = \text{RMFA}_{\mathcal{K}}(\mathbf{Q}, \mathbf{K}, \mathbf{V}). \end{aligned}$$

□

A.2 Proof of Theorem 2

Proof. By Theorem 1 we have $\text{RMFA}_{\text{exp}}(\mathbf{Q}^{\text{SBN}}, \mathbf{K}^{\text{SBN}}, \mathbf{V}) \approx \text{attn}_{\text{exp}}(\mathbf{Q}^{\text{SBN}}, \mathbf{K}^{\text{SBN}}, \mathbf{V})$, let $r = \|\mathbf{Q}'\|_2 \|\mathbf{K}'\|_2 \sqrt{(\sigma_{\mathbf{Q}} + \varepsilon)(\sigma_{\mathbf{K}} + \varepsilon)}$ depends on the data, $t = \frac{\sum_{i=1}^n \exp[(\mathbf{Q} \mathbf{K}_i^{\top} - \mu_{\mathbf{Q}} \mathbf{K}_i^{\top}) / r \sqrt{d}] \mathbf{1}_n^{\top}}{\sum_{i=1}^n \exp[(\mathbf{Q} \mathbf{K}_i^{\top} \mathbf{1}_n^{\top} - \mu_{\mathbf{Q}} \mathbf{K}^{\top}) / r \sqrt{d}]}$ and $s = \frac{\|\exp(\mathbf{Q} \mathbf{K}^{\top} / \sqrt{d})\|_{\frac{1}{r}}}{\|\exp(\mathbf{Q} \mathbf{K}^{\top} / \sqrt{d})\|_1} V^{r-1}$, we have

$$\begin{aligned} \text{RMFA}_{\text{exp}}(\mathbf{Q}^{\text{SBN}}, \mathbf{K}^{\text{SBN}}, \mathbf{V}) &\approx \text{Softmax} \left[\frac{(\mathbf{Q} - \mu_{\mathbf{Q}})(\mathbf{K} - \mu_{\mathbf{K}})^{\top}}{\|\mathbf{Q}'\|_2 \|\mathbf{K}'\|_2 \sqrt{(\sigma_{\mathbf{Q}} + \varepsilon)(\sigma_{\mathbf{K}} + \varepsilon)}} \Big/ \sqrt{d} \right] \mathbf{V} \\ &= \frac{\sum_{i=1}^n \exp[(\mathbf{Q} \mathbf{K}_i^{\top} \mathbf{1}_n^{\top} - \mu_{\mathbf{Q}} \mathbf{K}^{\top}) / r \sqrt{d}]}{\sum_{i=1}^n \exp[(\mathbf{Q} \mathbf{K}_i^{\top} - \mu_{\mathbf{Q}} \mathbf{K}_i^{\top}) / r \sqrt{d}] \mathbf{1}_n^{\top}} \frac{\exp(\mathbf{Q} \mathbf{K}^{\top} / r \sqrt{d}) \mathbf{V}}{\sum_{i=1}^n \exp(\mathbf{Q} \mathbf{K}_i^{\top} / r \sqrt{d}) \mathbf{1}_d^{\top}} \\ &= \frac{1}{t} \sqrt[r]{\frac{\|\exp(\mathbf{Q} \mathbf{K}^{\top} / \sqrt{d})\|_1}{\|\exp(\mathbf{Q} \mathbf{K}^{\top} / \sqrt{d})\|_{\frac{1}{r}}}} \mathbf{V}^{r-1} \odot \frac{\exp(\mathbf{Q} \mathbf{K}^{\top} / \sqrt{d}) \mathbf{V}}{\sum_{i=1}^n \exp(\mathbf{Q} \mathbf{K}_i^{\top} \mathbf{1}_d^{\top} / \sqrt{d})} \\ &= \frac{1}{t} \sqrt[r]{\frac{1}{s}} \odot \text{attn}_{\text{exp}}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) \approx \frac{1}{t} \left[\frac{1}{s} \odot \text{RMFA}_{\text{exp}}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) \right]^{\frac{1}{r}}, \end{aligned}$$

Combinign Theorem 1 and 2, we can get

$$\begin{aligned} \mathbb{E} [\text{RMFA}_{\text{exp}}(\mathbf{Q}^{\text{SBN}}, \mathbf{K}^{\text{SBN}}, \mathbf{V})] &= \mathbb{E} [\text{attn}_{\text{exp}}(\mathbf{Q}^{\text{SBN}}, \mathbf{K}^{\text{SBN}}, \mathbf{V})] \\ &= \mathbb{E} \left[\frac{1}{t} \sqrt[r]{\frac{1}{s}} \odot \text{attn}_{\text{exp}}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) \right] = \mathbb{E} \left[\frac{1}{t} \left[\frac{1}{s} \odot \text{RMFA}_{\text{exp}}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) \right]^{\frac{1}{r}} \right]. \end{aligned}$$

□

A.3 Proof of Theorem 3

Proof. We begin with the expectation-wise equality of ppSBN, as given in Equation (5),

$$\mathbb{E}[\text{SchoenbAt}_{\mathcal{K}}(\mathbf{Q}, \mathbf{K}, \mathbf{V})] = \mathbb{E} \left[\gamma \text{RMFA}_{\mathcal{K}}(\overrightarrow{\text{SBN}}(\mathbf{Q}), \overrightarrow{\text{SBN}}(\mathbf{K}), \mathbf{V})^\beta \right] = \mathbb{E}[\text{RMFA}_{\mathcal{K}}(\mathbf{Q}, \mathbf{K}, \mathbf{V})].$$

With the fact that $d \geq 1$, we can ensure that $\mathbf{Q}/d^{\frac{1}{4}}, \mathbf{K}/d^{\frac{1}{4}} \in \ell_2(0, 1)$, then we have

$$\begin{aligned} \mathbb{E}[\text{RMFA}_{\mathcal{K}}(\mathbf{Q}, \mathbf{K}, \mathbf{V})] &= \sum_{i=1}^n \frac{\mathbb{E} \left[\frac{1}{D} \sum \phi(\mathbf{Q}/d^{\frac{1}{4}}) \phi^\top(\mathbf{K}_i/d^{\frac{1}{4}}) \right] \mathbf{V}_i}{\sum_{j=1}^n \mathbb{E} \left[\frac{1}{D} \sum \phi(\mathbf{Q}/d^{\frac{1}{4}}) \phi^\top(\mathbf{K}_j/d^{\frac{1}{4}}) \right] \mathbf{1}_d^\top} \\ &= \sum_{i=1}^n \frac{\mathbb{E} \left[\frac{1}{D} \sum \phi(\mathbf{Q}/d^{\frac{1}{4}}) \phi^\top(\mathbf{K}_i/d^{\frac{1}{4}}) \right] \mathbf{V}_i}{\sum_{j=1}^n \mathbb{E} \left[\frac{1}{D} \sum \phi(\mathbf{Q}/d^{\frac{1}{4}}) \phi^\top(\mathbf{K}_j/d^{\frac{1}{4}}) \right] \mathbf{1}_d^\top} = \sum_{i=1}^n \frac{\mathbb{E}_N \left[a_N p^{N+1} \mathbb{E}_\omega \left(\prod_{l=1}^N \langle \omega_l, \mathbf{Q} \rangle \langle \omega_l, \mathbf{K}_i \rangle / \sqrt{d} \right) \right] \mathbf{V}_i}{\sum_{j=1}^n \mathbb{E}_N \left[a_N p^{N+1} \mathbb{E}_\omega \left(\prod_{l=1}^N \langle \omega_l, \mathbf{Q} \rangle \langle \omega_l, \mathbf{K}_j \rangle / \sqrt{d} \right) \right] \mathbf{1}_d^\top} \\ &= \sum_{i=1}^n \frac{\sum_{\eta=0}^\infty a_\eta \frac{1}{p^{\eta+1}} p^{\eta+1} (\mathbf{Q} \mathbf{K}_i^\top / \sqrt{d})^\eta \mathbf{V}_i}{\sum_{j=1}^n \left[\sum_{\eta=0}^\infty a_\eta \frac{1}{p^{\eta+1}} p^{\eta+1} (\mathbf{Q} \mathbf{K}_j^\top / \sqrt{d})^\eta \right] \mathbf{1}_d^\top} = \sum_{i=1}^n \frac{\mathcal{K}(\mathbf{Q} \mathbf{K}_i^\top / \sqrt{d}) \mathbf{V}_i}{\sum_{j=1}^n \mathcal{K}(\mathbf{Q} \mathbf{K}_j^\top / \sqrt{d}) \mathbf{1}_d^\top} = \text{attn}_{\mathcal{K}}(\mathbf{Q}, \mathbf{K}, \mathbf{V}). \end{aligned}$$

□

A.4 Proof of Theorem 4

Proof. Since the randomness in SchoenbAt only affects the attention matrix, and the attention mechanism computes a linear combination of $\mathbf{V}$, we have

$$\forall |\mathbf{V}_{ij}| \leq S, \text{SchoenbAt}_{\mathcal{K}}(\mathcal{D})_{ij} \in [-Sd, Sd],$$

then, by Hoeffding's inequality,

$$\begin{aligned} \mathbb{P}(|\text{SchoenbAt}_{\mathcal{K}}(\mathcal{D}) - \text{attn}_{\mathcal{K}}(\mathcal{D})| > \epsilon) &= \mathbb{P}(|\text{RMFA}_{\mathcal{K}}(\mathcal{D}) - \text{attn}_{\mathcal{K}}(\mathcal{D})| > \epsilon) \\ &= \mathbb{P} \left(\exists \phi_t : \left| \sum_{i=1}^n \frac{\left[\phi_t(\mathbf{Q}/d^{\frac{1}{4}}) \phi_t(\mathbf{K}_i/d^{\frac{1}{4}}) \right] \mathbf{V}_i}{\sum_{j=1}^n \phi_t(\mathbf{Q}/d^{\frac{1}{4}}) \phi_t(\mathbf{K}_j/d^{\frac{1}{4}}) \mathbf{1}_n^\top} - \text{attn}_{\mathcal{K}}(\mathcal{D}) \right| > \epsilon \right) \\ &\leq \sum_{t=1}^D \mathbb{P} \left(\left| \sum_{i=1}^n \frac{\left[\phi_t(\mathbf{Q}/d^{\frac{1}{4}}) \phi_t(\mathbf{K}_i/d^{\frac{1}{4}}) \right] \mathbf{V}_i}{\sum_{j=1}^n \phi_t(\mathbf{Q}/d^{\frac{1}{4}}) \phi_t(\mathbf{K}_j/d^{\frac{1}{4}}) \mathbf{1}_n^\top} - \text{attn}_{\mathcal{K}}(\mathcal{D}) \right| > \epsilon \right) \\ &\leq D \cdot 2 \exp \left[-\frac{2D^2 \epsilon^2}{D(2Sd)^2} \right] = 2D \exp \left(-\frac{D \epsilon^2}{2S^2 d^2} \right). \end{aligned}$$

□

B Experiment Compute Resources

Model	T	L	R	P	I	Avg.
Softmax	18092	4878	11284	6058	11790	10420
SchoenbAt	4452	1696	3758	2062	3636	3121
Ratio (%)	0.246	0.348	0.333	0.340	0.308	0.315

Table 4: Memory consumption of SchoenbAt compared to Softmax attention in Experiment 4.2. Since SchoenbAt with the five dot-product kernels consumes the same amount of memory, we represent them using a single row in the table.