DATA-AWARE RANDOM FEATURE KERNEL FOR TRANSFORMERS

Amirhossein Farzam, Hossein Mobahi, Nolan Andrew Miller & Luke Sernau
Google DeepMind
a.farzam@duke.edu

ABSTRACT

Transformers excel across domains, yet their quadratic attention complexity poses a barrier to scaling. Random-feature attention, as in *Performers*, can reduce this cost to linear in the sequence length by approximating the softmax kernel with positive random features drawn from an isotropic distribution. In pretrained models, however, queries and keys are typically anisotropic. This induces high Monte Carlo variance in isotropic sampling schemes unless one retrains the model or uses a large feature budget. Importance sampling can address this by adapting the sampling distribution to the input geometry, but complex data-dependent proposal distributions are often intractable. We show that by data aligning the softmax kernel, we obtain an attention mechanism which can both admit a tractable minimal-variance proposal distribution for importance sampling, and exhibit better training stability. Motivated by this finding, we introduce *DARKFormer*, a **D**ata-**A**ware **R**andom-feature **K**ernel trans**former** that features a data-aligned kernel geometry. DARKFormer learns the random-projection covariance, efficiently realizing an importance-sampled positive random-feature estimator for its data-aligned kernel. Empirically, DARKFormer narrows the performance gap with exact softmax attention, particularly in finetuning regimes where pretrained representations are anisotropic. By combining random-feature efficiency with data-aware kernels, DARKFormer advances kernel-based attention in resource-constrained settings.

1 INTRODUCTION

The remarkable success of transformers is largely attributed to self-attention's ability to model long-range dependencies while flexibly adapting to their input. Yet the quadratic complexity of attention mechanisms remains a critical limitation, especially for applications requiring long sequences. Random feature kernel approximations, as in *Performers* (Choromanski et al., 2021), address this by linearly approximating the softmax kernel in a feature space, achieving complexity that is linear in sequence length times the sample size. Under uniform sampling, this is an asymptotically unbiased estimator for the softmax kernel. This uniform sampling strategy, however, leads to high Monte Carlo variance when inputs exhibit anisotropic structure. To achieve low estimation error with anisotropically distributed queries and keys, as is common in real-world data (Godey et al., 2024), the unbiased uniform estimator requires either large feature samples to reduce variance or extensive training to reshape input distributions toward isotropy. Importance sampling offers a principled solution by adapting the sampling distribution to match data geometry, but implementing the optimal scheme depends on input-dependent matrices that are unknown a priori and involves non-trivial computations. Motivated by these challenges, we propose a **D**ata-**A**ware **R**andom-feature **K**ernel for trans**former**s—*DARKFormer*—that learns a data-aligned kernel geometry and is capable of realizing importance sampling automatically. This improves the performance and stability while achieving the promised efficiency of random-feature-kernel-based attention.

Building on the random feature paradigm for efficient attention, DARKFormer replaces the dot product in standard softmax attention with a Mahalanobis inner product, obtaining a data-aware kernel that automatically corrects for anisotropy. The random-feature estimator in DARKFormer learns a kernel geometry matrix Σ (parameterized via $\Sigma = M^\top M$), yielding a linear-time approximation that implicitly implements an importance-sampling scheme and can yield lower-variance estimates. Because the sampling strategy can adapt to the input distribution, queries and keys need not be

isotropically distributed for DARKFormer to yield accurate approximations. It naturally takes more draws where input density is enriched and fewer where it is depleted. We show that this is equivalent to importance sampling in the feature space, allowing DARKFormer to learn a data-aligned sampling geometry without explicitly computing per-sample importance weights. Consequently, DARKFormer enhances approximation under limited feature budgets, improving performance without requiring large samples. Furthermore, sampling random features according to input geometry can improve training stability, as we observe empirically.

DARKFormer is particularly suitable for finetuning regimes, where attention input distributions are dictated by pretrained weights, demanding substantial retraining to produce isotropic query-key distributions. The learned covariance can be interpreted as learning a linear re-embedding. When this covariance matches (or approximates) the inverse input covariance, this re-embedding can whiten the queries and keys. After this whitening, the attention kernel takes the standard softmax form. Our analysis in sections 3 and 4 motivates why aligning the sampling geometry with the representation statistics can reduce Monte Carlo error. We corroborate these findings through experiments on the Gemma model (Gemma Team et al., 2024a;b; 2025). Our empirical observations show that DARKFormer narrows the performance gap with exact attention, enjoys more stable training dynamics, and does not require a large sample, extensive retraining, or thorough hyperparameter tuning to achieve these improvements. Notably, DARKFormer excels especially well in finetuning settings, requiring only a small amount of data to produce robust covariance matrices and adapt to the presence of whitening. Together, these results show that this data-aware framework can achieve the linear complexity of random-feature attention with lower sampling costs via importance sampling, opening new possibilities for resource-efficient transformer architectures in applications where quadratic complexity remains prohibitive, such as long-context modeling and high-resolution vision tasks.

Main Contributions

Importance Sampling via DARKFormer. We introduce *DARKFormer*, which implements data-aligned random feature attention through a learnable covariance matrix, achieving low sample complexity with a tractable proposal distribution.

Variance Optimality and Data-Aligned Sampling. We observe that variance-optimal random-feature estimators require data-aligned sampling.

Efficiently Improves Performance. Experiments on Gemma demonstrate that DARKFormer narrows the performance gap with exact attention, with particular benefits in finetuning scenarios where query-key distributions are anisotropic.

Resource-constrained Finetuning. Our results show that DARKFormer can improve performance without large samples, long training cycles, or extensive hyperparameter tuning. Notably, it does not require training from scratch and is compatible with finetuning from pre-trained weights, making it particularly suitable for resource-constrained environments.

2 Background

Exactly computing the softmax attention entails quadratic cost in sequence length, posing a fundamental challenge for scaling. This has motivated significant research in efficient approximations (Choromanski et al., 2021; Peng et al., 2021; Likhosherstov et al., 2023), with random feature methods offering elegant linear-complexity solutions. We briefly review the attention mechanism and its random feature approximation that forms the foundation for our approach.

Attention Mechanism in Transformers. The fundamental building block of the Transformer architecture is the *attention* mechanism, which captures pairwise interactions between tokens. Consider a sequence of L input tokens, each with a d-dimensional embedding, mapped via learnable projection weights to $Q, K, V \in \mathbb{R}^{L\times d}$, which are referred to as *queries*, *keys*, and *values*, respectively. The classical scaled dot-product attention Vaswani et al. (2017) is then defined by

$$\mathrm{Att}(Q, K, V) = \mathrm{softmax}\Big(\tfrac{1}{\sqrt{d}}\, QK^\top\Big)\, V \;\in\; \mathbb{R}^{L\times d}.$$

The softmax matrix is formed by entry-wise exponentiation of the dot-product kernel, where the (i, j) entry captures similarity between the ith query and the jth key, normalized into a probability distribution over tokens for each query. This yields $L \times L$ attention weights, which, when multiplied by V, produces the attention output. The corresponding time and memory complexities are $\mathcal{O}(L^2\, d)$

and $\mathcal{O}(L^2)$. When L is large, this becomes prohibitive in speed and storage. Various approximations and alternative formulations have been proposed to mitigate this bottleneck.

Random Feature Approximation of Kernels. Rahimi & Recht (2007) introduced random feature expansions as an efficient way to approximate kernel inner products. Utilizing this foundational technique we can obtain efficient approximations of attention by replacing the softmax kernel $\kappa_{\text{SM}}(\mathbf{q}, \mathbf{k}) := \exp(\mathbf{q}^\top \mathbf{k})$ by a random feature expansion [1]. Suppose we wish to approximate a kernel $\kappa(\mathbf{x}, \mathbf{y})$ using an inner product of finite-dimensional feature maps, $\kappa(\mathbf{x}, \mathbf{y}) \approx \phi(\mathbf{x})^\top \phi(\mathbf{y})$. We can construct such $\phi(\cdot)$ by sampling a set of random projections and suitably normalized basis functions. For instance, using trigonometric basis functions, we can construct the Gaussian kernel, $\kappa_{\text{Gauss}}(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{1}{2}\|\mathbf{x} - \mathbf{y}\|^2\right)$, by drawing m samples $\omega_1, \ldots, \omega_m \sim \mathcal{N}(0, I_d)$, defining

$$\phi_\omega^{\text{trig}}(\mathbf{x}) = \frac{h(\mathbf{x})}{\sqrt{m}} \left[\cos\left(\omega_1^\top \mathbf{x}\right),\ \ldots,\ \cos\left(\omega_m^\top \mathbf{x}\right),\ \sin\left(\omega_1^\top \mathbf{x}\right),\ \ldots,\ \sin\left(\omega_m^\top \mathbf{x}\right)\right]^\top,$$

and setting $h(\mathbf{x}) = 1$. Alternatively, since $\kappa_{\text{SM}}(\mathbf{x}, \mathbf{y}) = \exp\left(\frac{1}{2}\|\mathbf{x}\|^2 + \frac{1}{2}\|\mathbf{y}\|^2\right)\kappa_{\text{Gauss}}(\mathbf{x}, \mathbf{y})$, with $h(\mathbf{x}) = \exp\left(\frac{1}{2}\|\mathbf{x}\|^2\right)$ we obtain the softmax kernel. Choromanski et al. (2021) rely on the same principle, but propose an alternate construction, *Positive Random Features* (PRFs), to approximate κ_{SM}, where the random feature map is defined as

$$\phi_\omega^{+}(\mathbf{x}) := \frac{h(\mathbf{x})}{\sqrt{m}} \left[\exp\left(\omega_1^\top \mathbf{x}\right),\ \ldots,\ \exp\left(\omega_m^\top \mathbf{x}\right)\right]^\top, \tag{1}$$

with $h(\mathbf{x}) = \exp\left(-\frac{1}{2}\|\mathbf{x}\|^2\right)$ and again, isotropic iid projection vectors $\omega_1, \ldots, \omega_m \sim \mathcal{N}(0, I_d)$. Importantly, with this construction, Lemma 2.1 holds for PRFs.

Lemma 2.1 (Choromanski et al. (2021))**.** *For* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$\exp\left(\mathbf{x}^\top \mathbf{y}\right) = \mathbb{E}_{\omega \sim \mathcal{N}(0, I_d)}\left[\exp\left(\omega^\top \mathbf{x} - \tfrac{1}{2}\|\mathbf{x}\|^2\right) \exp\left(\omega^\top \mathbf{y} - \tfrac{1}{2}\|\mathbf{y}\|^2\right)\right].$$

Hence, PRFs yield an unbiased approximation for the softmax kernel.

Choromanski et al. (2021) argue that PRFs outperform their trigonometric counterparts, especially where kernel values are small. Building on this, they introduce *Performer*, an efficient transformer architecture that approximates softmax attention using PRF-based random feature maps. Choromanski et al. (2021) enforce orthogonality of $\omega_1, \ldots, \omega_m$ via a Gram-Schmidt orthogonalization, combining PRFs with orthogonal random features to reduce the variance of the empirical estimator. Given a query-key pair $(\mathbf{q}, \mathbf{k})$, the softmax kernel in Performer is then approximated via the empirical estimator of $\mathbb{E}_{\omega \sim \mathcal{N}(0, I_d)}\left[\phi_\omega^{+}(\mathbf{q})^\top \phi_\omega^{+}(\mathbf{k})\right]$.

Efficient Attention via Random Feature Kernels. The random feature map replaces the exponential kernel $\exp(QK^\top)$ with its linear approximation in the feature space $(Q'(K')^\top)$ where the rows of $Q', K' \in \mathbb{R}^{L \times m}$ are the random feature maps of the rows of Q, K. Leveraging this linearity, we can approximate $\text{Att}(Q, K, V)$ without explicitly forming the $L \times L$ matrix, by first computing ${K'}^\top V$ and then left multiplying it by Q'. This ordering results in a time complexity of $\mathcal{O}(Lmd)$, as illustrated in Figure 1. Storing Q' and K' on the other hand, requires $\mathcal{O}(Lm)$ memory, and the largest intermediate matrix multiplication is either $(L \times m)\,(m \times d)$ or $(m \times L)\,(L \times d)$, consuming total memory on the order of $\mathcal{O}\left(\max\{Lm, Ld\}\right)$. Hence, this approximation improves computation time as long as $m < L$ and memory efficiency when $\max\{d, m\} < L$. Performer provides an elegant unbiased approximation of the softmax kernel via PRFs. However, the feature sample sizes required to achieve good performance with these estimators can be so large that it dilutes the computational efficiency. Moreover, these approximations allocate their feature budget isotropically over the input space. For anisotropic

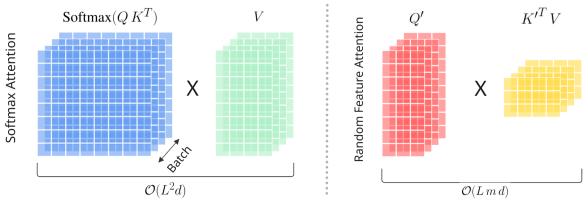

Figure 1: The random feature attention replaces the softmax kernel with a linear approximation in the feature space, reducing the quadratic complexity in sequence length (L) to linear in sequence length times sample size (m).

[1] For mapping this to $\text{Att}(Q, K, V)$, $\mathbf{q}$ and $\mathbf{k}$ in κ_{SM} can be viewed as queries and keys that have absorbed the scaling by $1/\sqrt{d}$. Normalizing this expression then produces softmax attention.

query-key distributions, as observed in practice (Godey et al., 2024), this can waste samples on low-density directions and inflate Monte Carlo variance. These caveats motivate our data-aligned random feature kernel, which improves approximation accuracy at smaller sample sizes by relaxing the isotropy assumption.

3 Variance Reduction via Data-Aware Sampling

An important error measure for random-feature kernel estimators is their Monte Carlo variance, which largely governs the gap with the exact exponential kernel when the number of features m is limited. We recall that importance sampling can achieve the minimum possible variance through a proposal density that depends on the query-key distribution. While importance sampling proposals are known in principle (Sernau et al., 2024), instantiating them for PRF estimators is challenging because the optimal proposal density is difficult to sample from directly. We show that the optimal proposal density is anisotropic whenever the input distribution is anisotropic, motivating anisotropy as an important dimension for improving PRF-based attention. This variance analysis provides the theoretical motivation for DARKFormer, which realizes data-aware sampling through a learned covariance matrix, leading to the improvements we show in Section 4.

Suppose the queries and keys are each iid draws from a distribution $\mathcal{D}$ on $\mathbb{R}^d$ with finite second moments and matching covariance matrices $\Lambda := \mathrm{Cov}(q) = \mathrm{Cov}(k)$. Let $p_I(\omega) = \mathcal{N}(0, I_d)$ be the usual sampling density and let $\psi(\omega)$ be any alternative density with the same support.

Lemma 3.1. *Define the PRF estimator with an importance factor $\frac{p_I(\omega_j)}{\psi(\omega_j)}$ as*

$$\hat{\kappa}_\psi(q,k) := \frac{1}{m}\sum_{j=1}^{m}\Big[\frac{p_I(\omega_j)}{\psi(\omega_j)}\exp\Big(\omega_j^\top q - \tfrac{1}{2}\|q\|^2\Big)\,\exp\Big(\omega_j^\top k - \tfrac{1}{2}\|k\|^2\Big)\Big]. \tag{2}$$

Among all finite-variance ψ, the expected Monte Carlo variance $\mathbb{E}_{q,k\sim\mathcal{D}}\big[\mathrm{Var}_\omega[\hat{\kappa}_\psi(q,k)]\big]$ is minimized by $\psi^*(\omega) \;\propto\; p_I(\omega)\,\sqrt{B_q(\omega)\,B_k(\omega)}$, where $B_x(\omega) := \mathbb{E}_x\left[\exp\left(2\omega^\top x - \|x\|^2\right)\right]$.

Lemma 3.1, proved in Appendix A, shows, for general $\psi(\omega)$, how importance sampling achieves an alignment with the input distribution, via the $B_q(\omega)B_k(\omega)$ term in the optimal $\psi^*(\omega)$, to minimize the Monte Carlo variance. To form a more concrete intuition for the structure of ψ^* and its relationship to the isotropic baseline, consider Gaussian queries and keys, where we can characterize ψ^* in closed form, assuming ψ^* is normalizable, as follows (see Appendix A for the proof).

Theorem 3.2. *Consider $q, k \sim \mathcal{N}(0, \Lambda)$. Then the optimal sampling density ψ^* is a centered Gaussian $\psi^* = \mathcal{N}(0, \Sigma^*)$ with $\Sigma^* = (I_d + 2\Lambda)(I_d - 2\Lambda)^{-1}$ whenever this expression defines a valid covariance. Moreover:*

1. *$\Sigma^* \propto I_d$ if and only if $\Lambda \propto I_d$.*
2. *The expected Monte Carlo variance satisfies*

$$\mathbb{E}_{q,k\sim\mathcal{D}}\big[\mathrm{Var}_\omega[\hat{\kappa}_{\psi^*}(q,k)]\big] \;\le\; \mathbb{E}_{q,k\sim\mathcal{D}}\big[\mathrm{Var}_\omega[\hat{\kappa}_{p_I}(q,k)]\big],$$

with strict inequality for any non-degenerate covariance $\Lambda \neq 0$.

Theorem 3.2 implies that the optimal proposal ψ^* matches the geometry of the inputs. Σ^* inherits the eigenbasis of Λ, and is isotropic if and only if Λ is isotropic. As soon as the queries and keys have non-trivial covariance, the optimal importance-sampled estimator strictly reduces the expected Monte Carlo variance compared to the isotropic baseline. In other words, drawing projection vectors from $\mathcal{N}(0, I_d)$ is never optimal once the input distribution is non-degenerate. However, computing this importance factor depends on the query-key distribution and involves non-trivial matrix operations. This motivates our approach: learning a tractable, data-dependent sampling geometry that can adapt to anisotropic representations. We introduce DARKFormer, a random feature scheme that efficiently realizes an importance-sampled estimator for a data-aligned kernel.

4 Data-Aware Random Feature Kernel

Motivated by the discussion in Section 3, we introduce a **D**ata-**A**ware **R**andom-feature **K**ernel Trans**former** (*DARKFormer*). DARKFormer learns a matrix Σ (parameterized as $\Sigma = M^\top M$) that

defines a data-aligned kernel and serves as the covariance of its random projections, characterizing the sampling geometry of its PRF [2]. This provides a tractable mechanism for adapting random-feature attention to anisotropic inputs, effectively implementing importance sampling, while bypassing explicit computation of per-sample proposal densities. DARKFormer's data-aware approach enhances performance without extensive retraining or large feature samples, and can also improve training stability. These improvements make it particularly amenable to resource-constrained settings, especially for finetuning, where query-key distributions are dictated by pretrained weights.

4.1 Positive Random Features with Covariance Learning

To achieve low estimation error, our proposed PRF leverages an exponential kernel of the form $\exp\big(q^\top \Sigma\, k\big)$, where $\Sigma \succeq 0$ is learned to adapt to the input geometry, as is achieved by a Mahalanobis metric in the kernel's inner product space. When $\Sigma = I_d$, this becomes the standard softmax kernel. For anisotropic or correlated representations, however, the Euclidean dot-product, $q^\top k$ can be miscalibrated because it treats all directions equally and ignores cross-coordinate correlations. This miscalibration can be corrected by measuring similarity in a Mahalanobis geometry induced by Σ. When the inputs have covariance $\Lambda \succ 0$, the whitening can be achieved through $\Sigma = \Lambda^{-1}$.

Mahalanobis Geometry for Attention. Given $\Sigma \succ 0$, let

$$\|x\|_\Sigma^2 \coloneqq x^\top \Sigma x \quad \text{and} \quad \|x-y\|_\Sigma^2 \coloneqq (x-y)^\top \Sigma (x-y),$$

which are the Mahalanobis norm and distance, respectively. Then we can write

$$q^\top \Sigma k = \frac{1}{2}(\|q\|_\Sigma^2 + \|k\|_\Sigma^2 - \|q-k\|_\Sigma^2).$$

As a result, $\exp(q^\top \Sigma k)$ is, up to scaling, a Gaussian kernel in the Mahalanobis distance $\|q-k\|_\Sigma$. Equivalently, writing $\Sigma = M^\top M$ gives $q^\top \Sigma k = (Mq)^\top (Mk)$ and $\|q-k\|_\Sigma = \|Mq - Mk\|$. In other words, this is the standard softmax kernel after the linear re-embedding $x \mapsto Mx$. In particular, if the input covariance is $\Lambda \succ 0$ then choosing $M = \Lambda^{-1/2}$ (so that $\Sigma = \Lambda^{-1}$) whitens the inputs and makes Mq and Mk isotropic. Thus, the Mahalanobis geometry measures distance in units of standard deviation in the principal directions, correcting scale and correlation mismatch under anisotropic representations. We provide a discussion and formal derivations for this in Appendix C.

Random Feature Kernel with a Learned Geometry. Let $\mathbf{x} \in \mathbb{R}^d$ be the kernel input, and suppose $\Sigma \in \mathbb{R}^{d\times d}$ is expressed as $\Sigma = M^\top M$, for a learnable matrix $M \in \mathbb{R}^{r\times d}$. We then write the random feature map as [3]

$$\phi_\Sigma(\mathbf{x}, \tilde{\omega}) \coloneqq \tfrac{h(\tilde{\mathbf{x}})}{\sqrt{m}} \left[\exp\big(\tilde{\omega}_1^\top \mathbf{x}\big),\ \ldots,\ \exp\big(\tilde{\omega}_m^\top \mathbf{x}\big)\right]^\top,$$

where $\tilde{\omega}_1, \ldots, \tilde{\omega}_m \sim \mathcal{N}(0, \Sigma)$, $h(\tilde{\mathbf{x}}) = \exp\big(-\frac{1}{2}\|\tilde{\mathbf{x}}\|^2\big)$, and $\tilde{\mathbf{x}} \coloneqq M\mathbf{x}$. Given a query–key pair $q, k \in \mathbb{R}^d$, the resulting PRF yields an unbiased approximation for $\exp\big(q^\top \Sigma k\big)$ as follows.

$$\exp\big(\mathbf{q}^\top \Sigma \mathbf{k}\big) \;=\; \mathbb{E}_{\tilde{\omega}}\Big[\, \exp\big(\tilde{\omega}^\top \mathbf{q} - \tfrac{1}{2}\mathbf{q}^\top \Sigma \mathbf{q}\big) \exp\big(\tilde{\omega}^\top \mathbf{k} - \tfrac{1}{2}\mathbf{k}^\top \Sigma \mathbf{k}\big)\Big]. \tag{3}$$

The derivation of Equation 3 is provided in Appendix B. In other words, the feature map $\phi_\Sigma(\cdot, \tilde{\omega})$ yields an exponential kernel with a learned geometry characterized by Σ. Learning Σ provides a simple and flexible way to make the attention geometry data-aligned while preserving the positive random-feature structure. In Section 4.2 we connect this covariance learning to importance sampling, which further motivates why it can reduce Monte Carlo error, in light of the discussion in Section 3. This leads to closing the performance gap with exact softmax, and meanwhile, it exhibits improved training stability, as observed in our experiments reported in Section 6.

4.2 Reducing Approximation Error via DARKFormer

We now show that DARKFormer effectively implements an importance sampling strategy. Section 3 shows that a fixed isotropic sampling, treating every query–key direction as equally important, is sub-optimal when the input distribution is anisotropic. Through learning a covariance Σ for

[2] For consistency of notation, we reserve Λ for the covariance of the query–key distribution, as in Section 3.

[3] Here we write the random feature map ϕ_Σ as a function of ω. This is a change of notation compared to Section 2, and only for ease of notation when the map involves Σ.

its data-aware kernel, DARKFormer realizes an importance sampling strategy that avoids explicit computation of per-sample importance weights by coupling the sampler with the learned kernel geometry. This in turn explains why DARKFormer can improve performance without requiring large feature samples or extensive retraining.

Let $p_I(\omega) = \mathcal{N}(0, I_d)$ denote the isotropic sampling distribution of the projection vectors in the PRF (Equation 1) and let us denote its data-aware counterpart by $p_\Sigma(\omega) = \mathcal{N}(0, \Sigma)$. Viewing the change from p_I to p_Σ through the lens of importance sampling reveals that DARKFormer's unweighted estimator is in fact equal in expectation to an isotropic estimator whose samples are re-weighted by an importance factor. This analogy opens the door to known results regarding the optimality of importance sampling estimators, and in particular, makes the results from Section 3 applicable, which point to DARKFormer's capability in implementing a sampling scheme that can reduce estimation error. To see this, let us rewrite the PRF map for $\omega \sim p_\Sigma(\omega)$ as

$$\phi_\Sigma(x, \omega) \;\coloneqq\; \exp\!\big(\omega^\top x - \tfrac{1}{2} x^\top \Sigma\, x\big).$$

Given a query–key pair (q, k) and a feature budget of m (i.e., sample size is m), we define the $w-$weighted empirical mean estimator for the feature kernel estimand in Equation 3 as

$$\hat{\kappa}^{w}_{\Sigma}(q, k) \coloneqq \frac{1}{m} \sum_{j=1}^{m} w(\omega_j)\, \phi_\Sigma(q, \omega_j)\, \phi_\Sigma(k, \omega_j), \tag{4}$$

where the weight can depend on ω_j. Proposition 4.1 shows that sampling ω from $p_\Sigma(\omega)$ in the unweighted estimator is equivalent in expectation to sampling ω from the isotropic distribution $p_I(\omega)$ and applying importance weights (see Appendix B for the proof).

Proposition 4.1. *Let $\Sigma \succ 0$ and $w_\Sigma(\omega) \coloneqq \frac{p_\Sigma(\omega)}{p_I(\omega)}$. For any map $f : \mathbb{R}^d \to \mathbb{R}$ with a finite mean,*

$$\mathbb{E}_{\omega \sim p_\Sigma}\big[f(\omega)\big] \;=\; \mathbb{E}_{\omega \sim p_I}\big[w_\Sigma(\omega)\, f(\omega)\big].$$

Consequently, for every (q, k) and $m \geq 1$, $\mathbb{E}_{\omega_{1:m} \sim p_\Sigma}\left[\hat{\kappa}^{1}_{\Sigma}(q, k)\right] \;=\; \mathbb{E}_{\omega_{1:m} \sim p_I}\left[\hat{\kappa}^{w_\Sigma(\omega)}_{\Sigma}(q, k)\right]$.

This Proposition shows that drawing from p_Σ is equivalent to re-weighted importance sampling from p_I. Hence, by learning the covariance matrix Σ, DARKFormer implicitly implements importance sampling without computing per-sample importance weights. This is especially valuable for finetuning in resource-constrained settings, where query-key distributions are determined by pretrained weights and may require substantial retraining to conform to isotropy assumptions.

5 Related Work

Relevant prior work is cited throughout the manuscript. A comprehensive review is provided in Appendix D, which details the literature on random feature approximations for kernel methods and their adaptation for efficient attention. In Appendix D, we also discuss the positioning of our work, highlighting how it addresses the gap in random feature kernels for efficient attention.

6 Experiments

To investigate DARKFormer's capabilities in empirical settings and validate the improvements we anticipate theoretically, we conduct a series of experiments on training and finetuning a Gemma-based DARKFormer. Replacing the exact softmax in the Gemma model (Gemma Team et al., 2024a) with the PRF-based approximation, we compare a DARKFormer model against its counterpart with isotropic PRF maps, which is a Performer-type model (Choromanski et al., 2021). While DARKFormer achieves this improvement in pretraining as well, since it is primarily intended for resource-scarce settings, our primary focus is on finetuning. Our results demonstrate that the data-aware sampling strategy of DARKFormer translates the theoretical variance reduction benefits into practical performance improvements. In particular, they confirm that: (1) DARKFormer reduces the performance gap with exact attention compared to isotropic PRF; (2) it achieves this improvement without requiring large feature samples or extensive training; and (3) it is particularly advantageous in finetuning scenarios where query-key distributions are dictated by pretrained weights.

Experimental Setup. For these experiments we use a 2-B-parameter Gemma model (Gemma Team et al., 2024a), benchmarked on the C4 dataset (Raffel et al., 2020) for the next-token-prediction

task. While our primary comparison is between a DARKFormer and a Performer-type model against the exact softmax attention, we also include both learned feature kernels (LFK) and simple baselines. In particular, for the LFK, we replace the randomly drawn projections in the PRF (ω in Equation 2) with a trainable vector. The LFK allows the model to freely learn the projections that yield a kernel which minimizes the task loss. Meanwhile the simple baselines, which include a uniform random and a constant attention, benchmark all other models against a transformer-type architecture that learns around the softmax attention, highlighting how much of the change in performance is in fact due to the ability of the feature kernel used to implement an effective attention mechanism. All comparisons are performed under identical training conditions and hyperparameters.

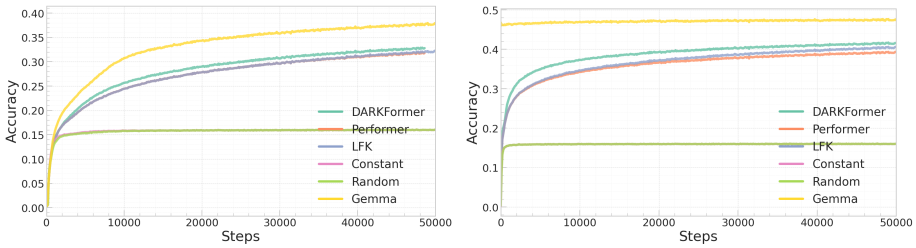

Figure 2: Next token prediction accuracy during pretraining (left) and finetuning (right) of the Gemma-2B model with a DARKFormer (green), a Performer (orange), learned feature kernel (LFK) (blue), a random baseline (yellow), a constant baseline (lime), and an exact softmax attention. The DARKFormer model considerably narrows the gap between the exact softmax and the Performer-type model and also outperforms LFK, especially in finetuning.

Pretraining and Finetuning Performance. Starting our empirical validation with pretraining experiments, we observe that DARKFormer narrows the performance gap with the exact softmax compared with the Performer-type model. This improvement is present throughout the pretraining as shown in Figure 2. Turning to our main focus—finetuning—Figure 2 shows that the DARKFormer model achieves significantly better next-token-prediction accuracy than the Performer model as well as all baselines, especially after the initial phase, but without extensive finetuning. This underscores that our proposed data-aware framework improves random-feature attention in practice, as suggested by our discussion in sections 3 and 4, which highlight the benefits of data-aligned sampling geometries under anisotropic inputs. Comparison with the baselines further indicates that (i) the observed performance is not due to the transformer learning around the PRF kernel; and (ii) the improvement is critically due to the random feature kernel used, as a learned feature kernel needs extensive finetuning to reach a performance on par with DARKFormer.

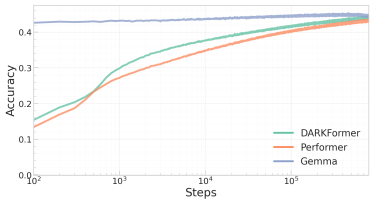

Figure 3: Next token prediction accuracy for finetuning of the Gemma-2B (blue) model with a DARKFormer (green), a Performer (orange), and an exact softmax attention over a long cycle of 650k finetuning steps. Observe that DARKFormer outperforms Performer throughout training despite approximating a novel whitened kernel that is out of distribution for pretrained Gemma. Note that the x-axis is shown on logarithmic scale.

Efficient Finetuning with DARKFormer. Investigating the model's capabilities to adapt to each of the PRF kernels, we repeat our main comparison between the DARKFormer and the Performer-type models as well as Gemma with the exact softmax over extended finetuning. Recall that our proposed data-aligned PRF leads to an improvement in the softmax approximation when the inputs are anisotropic, which is often the case. However, given the capabilities of transformers, they could eventually learn to produce isotropic inputs when the model is trained over a long period with isotropic PRFs. Our observations in Figure 3 confirm this: with sufficiently long finetuning the Performer-based model eventually closes the gap to DARKFormer, but only after many more optimization steps, incurring a significantly higher computational cost. This highlights DARKFormer's advantage in resource-constrained scenarios, reinforcing its benefits for efficient finetuning.

Limited Attention Finetuning. To better disentangle effects due to the model's ability to eventually learn to produce isotropic inputs for the PRF, we investigate partial finetuning, which is also of interest in resource-scarce settings. We freeze all layers of the pretrained transformer except the query, key, and value projections and, for DARKFormer, the trainable PRF covariance. While these projections could in principle adapt toward isotropic q and k, freezing the rest of the network substantially limits their ability to do so. Repeating the finetuning experiment under this setup further reveals DARKFormer's ability to improve performance, without the help of other components of the transformer. In particular, Figure 4 shows that its performance gap with Performer is even more pronounced and does not fade over long partial finetuning, unlike the full finetuning. This provides further evidence on DARKFormer's capability to improve performance with limited resources.

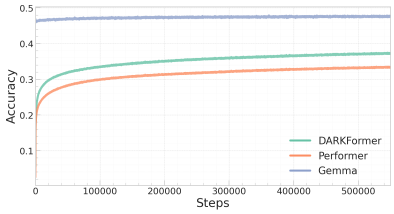

Figure 4: Next token accuracy for finetuning only the q-k-v projection weights and the PRF projection covariance for the case of DARKFormer, in a Gemma-2B model with a DARKFormer (green), a Performer (orange), and an exact softmax (blue) attention over 550k finetuning steps. The DARKFormer-induced improvement is even more pronounced than finetuning the full model, and does not decrease in later finetuning iterations.

Training Stability Across Learning Rates. We also evaluate training stability during finetuning by sweeping the learning rate and examining the resulting loss dynamics. Under identical hyperparameters, Performer frequently exhibits numerical instability, with sharp loss spikes as the learning rate increases. DARKFormer remains stable over the same sweep, with substantially fewer spikes. One driver of instability at large learning rates is noisy stochastic updates that occasionally take loss-increasing steps. Complementarily, the Mahalanobis geometry in DARKFormer implicitly whitens inputs and tempers extreme dot products, keeping the exponential kernel numerically well behaved. Thus, aligning the random-feature sampling geometry with input statistics can reduce Monte Carlo variance and make updates less erratic, matching the reduced spike frequency we observe. These observations suggest that DARKFormer is more robust to learning-rate choice, reducing sensitivity to hyperparameter tuning.

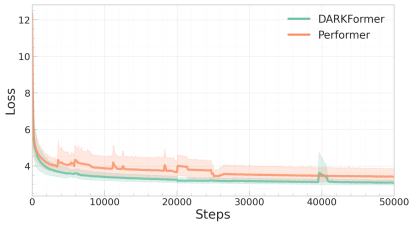

Figure 5: Loss dynamics for finetuning of the Gemma-2B model with DARKFormer (green) and a Performer (orange) with different learning rates. The shaded area marks the variance of loss values across seven different learning rates with the line showing the average. DARKFormer exhibits stable loss minimization in all but the largest learning rate where there is a short instability phase, while Performer shows more frequent instability phases and loss spikes during its finetuning with large learning rates. For visualization purposes the horizontal axis is shown on logarithmic scale.

Takeaways. Collectively, these experiments reinforce DARKFormer's theoretical motivations and practical advantages. DARKFormer narrows the gap to exact attention without large feature samples, adapts rapidly during finetuning, and is effective in resource-constrained settings, supporting the benefits of data-aware PRF kernels. Moreover, the stability results suggest DARKFormer can reduce costly hyperparameter tuning. Overall, these findings position DARKFormer as a practical approach for efficient transformers under constrained resources and anisotropic inputs.

7 Discussion

In this work, we introduced *DARKFormer*, a data-aware random feature kernel for efficient transformer architectures that narrows the performance gap between exact softmax attention and positive random-feature approximations. By learning the covariance of the random projections, DARKFormer implements a data-aligned kernel and a corresponding importance-sampled estimator, reducing approximation error, reducing Monte Carlo variance, and enhancing training stability, with particularly strong benefits when computational resources are scarce. Our empirical results on a Gemma model offer clear evidence for these benefits and highlight DARKFormer's practical ad-

vantages, especially for efficient finetuning. Looking ahead, combining data-aware random feature attention with complementary efficiency techniques and evaluating it across broader domains such as high-resolution inputs, long-context modeling, or on-device training are promising directions. We view DARKFormer as a step toward principled, data-aware attention mechanisms that make efficient transformers more accurate and more practical.

REFERENCES

Ahmed Alaoui and Michael W. Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In Advances in Neural Information Processing Systems, 2015.

Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. Journal of Machine Learning Research, 2017.

Yifan Chen, Qi Zeng, Heng Ji, and Yun Yang. Skyformer: Remodel self-attention with gaussian kernel and nyström method. Advances in Neural Information Processing Systems, 34:2122–2135, 2021.

Krzysztof Choromanski, Mark Rowland, and Adrian Weller. The unreasonable effectiveness of structured random orthogonal embeddings. In Advances in Neural Information Processing Systems, 2017.

Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. In International Conference on Learning Representations, 2021.

Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. arXiv preprint arXiv:2307.08691, 2023.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. Advances in Neural Information Processing Systems, 35:16344–16359, 2022.

Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In Empirical Methods in Natural Language Processing, 2019.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. arXiv preprint arXiv:2403.08295, 2024a.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. arXiv preprint arXiv:2408.00118, 2024b.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. arXiv preprint arXiv:2503.19786, 2025.

Nathan Godey, Éric Clergerie, and Benoît Sagot. Anisotropy is inherent to self-attention in transformers. In European Chapter of the Association for Computational Linguistics, pp. 35–48, 2024.

Praneeth Kacham, Vahab Mirrokni, and Peilin Zhong. Polysketchformer: fast transformers via sketching polynomial kernels. In International Conference on Machine Learning, pp. 22748–22770, 2024.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In International Conference on Machine Learning, 2020.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In International Conference on Learning Representations, 2020.

Quoc V. Le, Tamás Sarlós, and Alex J. Smola. Fastfood: Approximating kernel expansions in loglinear time. In International Conference on Machine Learning, 2013.

Sangho Lee, Hayun Lee, and Dongkun Shin. Proxyformer: Nyström-based linear transformer with trainable proxy tokens. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pp. 13418–13426, 2024.

Dacheng Li, Rulin Shao, Anze Xie, Eric P Xing, Xuezhe Ma, Ion Stoica, Joseph E Gonzalez, and Hao Zhang. DISTFLASHATTN: Distributed memory-efficient attention for long-context llms training. arXiv preprint arXiv:2310.03294, 2023.

Valerii Likhosherstov, Krzysztof Choromanski, Avinava Dubey, Frederick Liu, Tamas Sarlos, and Adrian Weller. Favor#: Sharp attention kernel approximations via new classes of positive random features. arXiv preprint arXiv:2302.00787, 2023.

Avner May, Alireza Bagheri Garakani, Zhiyun Lu, Dong Guo, Kuan Liu, Aurélien Bellet, Linxi Fan, Michael Collins, Daniel Hsu, Brian Kingsbury, et al. Kernel approximation methods for speech recognition. Journal of Machine Learning Research, 20(59):1–36, 2019.

Hao Peng, Nikolaos Pappas, Dani Yogatama, Roy Schwartz, Noah Smith, and Lingpeng Kong. Random feature attention. In International Conference on Learning Representations, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140):1–67, 2020.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. Advances in Neural Information Processing Systems, 20, 2007.

Ali Rahimi and Benjamin Recht. Uniform approximation of functions with random bases. In Allerton Conference on Communication, Control, and Computing, pp. 555–561. IEEE, 2008.

Luke Sernau, Silvano Bonacina, and Rif A Saurous. All random features representations are equivalent. arXiv preprint arXiv:2406.18802, 2024.

Jay Shah, Ganesh Bikshandi, Ying Zhang, Vijay Thakkar, Pradeep Ramani, and Tri Dao. FlashAttention-3: Fast and accurate attention with asynchrony and low-precision. Advances in Neural Information Processing Systems, 37:68658–68685, 2024.

Dougal J. Sutherland and Jeff Schneider. On the error of random fourier features. In Uncertainty in Artificial Intelligence, 2015.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in Neural Information Processing Systems, 30, 2017.

Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. In Advances in Neural Information Processing Systems, 2020.

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 Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pp. 14138–14148, 2021.

Felix Xinnan Yu, Ananda Theertha Suresh, Krzysztof Choromanski, Daniel Holtmann-Rice, and Sanjiv Kumar. Orthogonal random features. In Advances in Neural Information Processing Systems, 2016.

Yu Zheng, Difan Yuan, Sheng Wang, and Lingjie Kong. Efficient attention via control variates. In International Conference on Learning Representations, 2023.

A Optimal Importance Sampling for Random Feature Kernels

In this appendix, we provide the proofs for Lemma 3.1 and Theorem 3.2 from Section 3, which establish the theoretical foundation for variance reduction through importance sampling in random feature kernels. Recall that Lemma 3.1 characterizes the optimal sampling distribution for minimizing the Monte Carlo variance of the positive random feature kernel estimator. Specifically, it shows that the variance is minimized by an importance factor that depends on the input distribution through the term $\sqrt{B_q(\omega)B_k(\omega)}$, as defined in the statement of the theorem. The subsequent Theorem 3.2 applies this result to the Gaussian case, showing that the optimal sampling distribution is isotropic if and only if the input distribution is isotropic. These results motivate learning anisotropic sampling geometries when the input distribution is anisotropic, which informs DARKFormer's use of a learned covariance to obtain a tractable data-aware random-feature scheme.

Proof of Lemma 3.1. Given (q, k), let $Z(q, k, \omega) := \exp\big(\omega^\top q - \frac{1}{2}\|q\|^2\big)\exp\big(\omega^\top k - \frac{1}{2}\|k\|^2\big)$. By Lemma 2.1, $\kappa(q, k) := \mathbb{E}_{p_I}[Z] = \exp(q^\top k)$. Since the summands in Equation 2 are iid, we have

$$\mathrm{Var}_\omega[\hat{\kappa}_\psi(q, k)] = \frac{1}{m}\mathrm{Var}_\psi\Big[\frac{p_I}{\psi}\, Z\Big] \;=\; \frac{1}{m}\Big(\mathbb{E}_\psi[(p_I/\psi)^2 Z^2] \;-\; \kappa(q, k)^2\Big). \tag{5}$$

Taking the expectation over $q, k \sim \mathcal{D}\otimes\mathcal{D}$ and discarding the constant terms, subject to $\int \psi(\omega)\, d\omega = 1$, we can write the variance reduction objective $V(\psi)$ as

$$V(\psi) := \frac{1}{m}\int_{\mathbb{R}^d} \frac{p_I(\omega)^2}{\psi(\omega)}\, B_q(\omega)\, B_k(\omega)\, d\omega. \tag{6}$$

where $B_x(\omega) := \mathbb{E}_x\left[\exp\left(2\omega^\top x - \|x\|^2\right)\right]$. Defining $f(\omega) := p_I(\omega)\sqrt{B_q(\omega)B_k(\omega)} \geq 0$, we can rewrite this as $V(\psi) = \frac{1}{m}\int f(\omega)^2/\psi(\omega)\, d\omega$. For ease of notation, let $g_1(\omega) = f(\omega)/\sqrt{\psi(\omega)}$ and $g_2(\omega) = \sqrt{\psi(\omega)}$, and note that both $g_1(\cdot)$ and $g_2(\cdot)$ are in $L_2(\mathbb{R}^d)$, thus, we can apply Cauchy-Schwarz, which yields

$$\Big(\int f(\omega)\, d\omega\Big)^2 = \Big(\int g_1(\omega)\, g_2(\omega)\, d\omega\Big)^2 \leq \big(\int g_1^2(\omega)\, d\omega\big)\,\big(\int g_2^2(\omega)\, d\omega\big) = \int \tfrac{f^2(\omega)}{\psi(\omega)}\, d\omega, \tag{7}$$

where the last equality is due to $\int \psi(\omega)\, d\omega = 1$. Note that equality holds if and only if g_1 and g_2 are collinear, which requires $\psi \propto f$. Now, by definition $f(\omega)^2 = p_I(\omega)^2 B_q(\omega)B_k(\omega)$, hence $\int f^2/\psi = m\, V(\psi)$. Plugging this in 7 and rearranging gives

$$V(\psi) \;\geq\; \frac{(\int f(\omega)\, d\omega)^2}{m},$$

where $V(\psi)$ attains its minimum when equality holds, i.e., when $\psi \propto f$. Thus, the optimal ψ satisfies $\psi^* \propto p_I(\omega)\sqrt{B_q(\omega)B_k(\omega)} \geq 0$ as required, completing the proof of Lemma 3.1. □

Proof of Theorem 3.2. Let $\mathcal{D} := \mathcal{N}(0, \Lambda)$ and write the eigen-decomposition $\Lambda = U\,\mathrm{diag}(\lambda_1, \ldots, \lambda_d)\, U^\top$ with $\lambda_i \geq 0$ and U orthogonal. Since (q, k) are iid draws from $\mathcal{D}$, the coordinates q'_i, k'_i of $q' := U^\top q$, $k' := U^\top k$ are independent one-dimensional Gaussians $\mathcal{N}(0, \lambda_i)$.

We first analyze $B_q(\omega)$ in this eigenbasis. Writing $\omega' := U^\top\omega$ and using independence across coordinates we obtain

$$\begin{aligned} B_q(\omega) = \mathbb{E}_q\big[\exp\big(2\omega^\top q - \|q\|^2\big)\big] &= \mathbb{E}_{q'}\Big[\exp\big(2{\omega'}^\top q' - \|q'\|^2\big)\Big] \\ &= \prod_{i=1}^d \mathbb{E}_{q'_i}\Big[\exp\big(2\omega'_i q'_i - (q'_i)^2\big)\Big] \\ &= \prod_{i=1}^d c_i\, \exp\big(\beta_i(\omega'_i)^2\big), \end{aligned}$$

where $c_i > 0$ does not depend on ω'_i and $\beta_i := \frac{2\lambda_i}{2\lambda_i+1} > 0$. Hence

$$B_q(\omega) \;\propto\; \exp\Big(\sum_{i=1}^d \beta_i(\omega'_i)^2\Big),$$

or, equivalently, $B_q(\omega) \propto \exp\Big(\frac{1}{2}\omega^\top S\omega\Big)$, where $S := U\ \mathrm{diag}(2\beta_1, \ldots, 2\beta_d)\, U^\top$. The same expression holds for $B_k(\omega)$ since q and k share the same distribution.

By Lemma 3.1, the optimal density is $\psi^*(\omega) \propto p_I(\omega)\sqrt{B_q(\omega)B_k(\omega)}$. Using $p_I(\omega) \propto \exp(-\frac{1}{2}\|\omega\|^2)$ and $B_q = B_k$ yields

$$\begin{aligned}\psi^*(\omega) &\propto \exp\Big(-\tfrac{1}{2}\|\omega\|^2 + \sum_{i=1}^{d}\beta_i(\omega_i')^2\Big)\\ &= \exp\Big(-\tfrac{1}{2}\sum_{i=1}^{d}(1-2\beta_i)\,(\omega_i')^2\Big),\end{aligned}$$

where we used $\|\omega\|^2 = \|\omega'\|^2$. Under the technical assumption that $1 - 2\beta_i > 0$ for all i, or equivalently, $\lambda_i < \frac{1}{2}$ (so that ψ^* is integrable), this is the unnormalized density of a centered Gaussian with covariance

$$\Sigma^* = U\ \mathrm{diag}\Big(\tfrac{1}{1-2\beta_1}, \ldots, \tfrac{1}{1-2\beta_d}\Big)U^\top,$$

which proves that $\psi^* = \mathcal{N}(0, \Sigma^*)$ and establishes the first part of the theorem.

From this expression, we can see that Σ^* clearly shares the eigenvectors of Λ. Moreover, each diagonal entry $1/(1-2\beta_i)$ is a scalar function of λ_i, so all eigenvalues of Σ^* are equal if and only if all λ_i are equal, i.e., if and only if $\Lambda \propto I_d$. This proves item (1).

For item (2), recall from Equation 6 that

$$V(\psi) := \mathbb{E}_{q,k\sim\mathcal{D}}\big[\mathrm{Var}_\omega[\hat{\kappa}_\psi(q,k)]\big] = \frac{1}{m}\int_{\mathbb{R}^d}\frac{f(\omega)^2}{\psi(\omega)}\,d\omega,$$

where $f(\omega) = p_I(\omega)\sqrt{B_q(\omega)B_k(\omega)}$. The Cauchy–Schwarz inequality in 7 implies $V(\psi) \geq (\int f)^2/m$ with equality if and only if ψ is proportional to f, i.e., $\psi = \psi^*$ after normalization. In particular, $V(\psi^*) \leq V(p_I)$, and $V(\psi^*) = V(p_I)$ can hold only if p_I is proportional to f.

In the Gaussian setting above, $B_q(\omega)$ is non-constant whenever $\Lambda \neq 0$, so $f(\omega) \propto p_I(\omega)$ can occur only in the degenerate case $\Lambda = 0$. Consequently, for any non-degenerate covariance $\Lambda \neq 0$ we have $V(\psi^*) < V(p_I)$, proving item (2). □

B Data-aware Positive Random Feature Kernel: Derivations

This appendix provides the derivations of the results on DARKFormer's PRF kernel, presented in Section 4. We first derive the expression for the exponential kernel with learned covariance given in Equation 3, which forms the core of our data-aware random feature kernel. We then prove Proposition 4.1, which establishes the equivalence between DARKFormer's covariance learning and importance sampling—a critical insight that allows our method to implicitly implement importance sampling.

Derivation of Equation 3. Here we derive the expression for DARKFormer's data-aware kernel as presented in Equation 3, showing how a learned covariance matrix Σ can produce an unbiased approximation of the exponential kernel $\exp(q^\top\Sigma k)$ as discussed in Section 4.1. Let $\Sigma \in \mathbb{R}^{d\times d}$ be positive–semidefinite and write $\Sigma = M^\top M$ with $M \in \mathbb{R}^{r\times d}$ ($r \leq d$). We aim to show, for any $\mathbf{q}, \mathbf{k} \in \mathbb{R}^d$,

$$\exp\big(\mathbf{q}^\top\Sigma\,\mathbf{k}\big) = \mathbb{E}_{\tilde{\omega}\sim\mathcal{N}(0,\Sigma)}\Big[\exp\big(\tilde{\omega}^\top\mathbf{q} - \tfrac{1}{2}\,\mathbf{q}^\top\Sigma\mathbf{q}\big)\ \exp\big(\tilde{\omega}^\top\mathbf{k} - \tfrac{1}{2}\,\mathbf{k}^\top\Sigma\mathbf{k}\big)\Big]. \tag{8}$$

Let $\tilde{q} := Mq,\ \tilde{k} := Mk \in \mathbb{R}^r$. Applying Lemma 2.1 in $\mathbb{R}^r$ with $w \sim \mathcal{N}(0, I_r)$ gives

$$\exp(\tilde{q}^\top\tilde{k}) = \mathbb{E}_{w\sim\mathcal{N}(0,I_r)}\left[\exp\left(w^\top\tilde{q} - \frac{1}{2}\|\tilde{q}\|^2\right)\ \exp\left(w^\top\tilde{k} - \frac{1}{2}\|\tilde{k}\|^2\right)\right]. \tag{9}$$

Replacing w with $\tilde{\omega} := M^\top w$ in the expectation, the expectation will be taken with respect to $\tilde{\omega} \sim \mathcal{N}(0, \Sigma)$. Moreover $w^\top\tilde{q} = w^\top Mq = (M^\top w)^\top q = \tilde{\omega}^\top q$ and similarly $w^\top\tilde{k} = \tilde{\omega}^\top k$, while $\|\tilde{q}\|^2 = q^\top\Sigma q$ and $\|\tilde{k}\|^2 = k^\top\Sigma k$. Substituting these into Equation 9 yields Equation 3.

Proof of Proposition 4.1. Here we prove Proposition 4.1, which establishes that sampling projection vectors from a non-isotropic distribution $p_\Sigma(\omega)$ is equivalent to sampling from the isotropic distribution $p_I(\omega)$ and applying importance weights. For any integrable f,

$$\mathbb{E}_{\omega\sim p_\Sigma}[f(\omega)] = \int f(\omega)\, p_\Sigma(\omega) d\omega = \int f(\omega)\, w_\Sigma(\omega)\, p_I(\omega) d\omega = \mathbb{E}_{\omega\sim p_I}[w_\Sigma(\omega) f(\omega)].$$

Choosing $f(\omega) = \phi_\Sigma(q,\omega)\,\phi_\Sigma(k,\omega)$ gives

$$\mathbb{E}_{\omega\sim p_\Sigma}\big[\phi_\Sigma(q,\omega)\phi_\Sigma(k,\omega)\big] = \mathbb{E}_{\omega\sim p_I}\big[w_\Sigma(\omega)\phi_\Sigma(q,\omega)\phi_\Sigma(k,\omega)\big],$$

and hence, for the empirical average with iid sampling we have $\mathbb{E}_{\omega_{1:m}\sim p_\Sigma}\left[\hat{\kappa}^1_\Sigma(q,k)\right] = \mathbb{E}_{\omega_{1:m}\sim p_I}\left[\hat{\kappa}^{w_\Sigma}_\Sigma(q,k)\right]$. □

C Attention with Mahalanobis Geometry

In Section 4 we introduce a data-aware kernel of the form $\exp(q^\top \Sigma k)$. This appendix provides additional details on this kernel. In particular, we explain how the Σ-inner product measures similarity through Mahalanobis distance and show how choosing Σ as an inverse covariance whitens the representation space. This formalizes why replacing $q^\top k$ by $q^\top \Sigma k$ corrects both scale and correlation mismatch under anisotropic representations.

Attention with Mahalanobis Geometry. For $\Sigma \succ 0$, define the Mahalanobis norm $\|x\|^2_\Sigma := x^\top \Sigma x$ and the induced distance $\|x-y\|^2_\Sigma = (x-y)^\top \Sigma (x-y)$. Note that $x^\top \Sigma y = \frac{1}{2}\big(\|x\|^2_\Sigma + \|y\|^2_\Sigma - \|x-y\|^2_\Sigma\big)$. Consequently,

$$\exp(x^\top \Sigma y) = \exp\Big(\tfrac{1}{2}\|x\|^2_\Sigma\Big)\, \exp\Big(\tfrac{1}{2}\|y\|^2_\Sigma\Big)\, \exp\Big(-\tfrac{1}{2}\|x-y\|^2_\Sigma\Big).$$

Hence, replacing $q^\top k$ by $q^\top \Sigma k$ corresponds to measuring similarity in a Mahalanobis geometry. In particular, $\exp(q^\top \Sigma k)$ is, up to scaling, a Gaussian kernel in the Mahalanobis distance $\|q-k\|_\Sigma$. This directly addresses issues that arise under anisotropic representations, since Σ can downweight high-variance directions and incorporate correlations across coordinates.

Proposition C.1. *Given $\Sigma \succ 0$, let M be the matrix satisfying $\Sigma = M^\top M$ and suppose $r = d$. Then for all $q, k \in \mathbb{R}^d$ we have $q^\top \Sigma k = (Mq)^\top (Mk)$ and $\|q-k\|^2_\Sigma = \|Mq - Mk\|^2$. If the representation covariance is $\Lambda = \mathrm{Cov}(q) = \mathrm{Cov}(k) \succ 0$ and we choose $\Sigma = \Lambda^{-1}$ with $M = \Lambda^{-1/2}$, then $\mathrm{Cov}(Mq) = I_d$ and $\|q-k\|^2_{\Lambda^{-1}} = \sum_{i=1}^d \delta_i^2/\lambda_i$, where $\Lambda = U\mathrm{diag}(\lambda_1,\dots,\lambda_d)U^\top$ and $\delta = U^\top(q-k)$.*

Proof. For any $q, k \in \mathbb{R}^d$ we have $q^\top \Sigma k = q^\top M^\top M k = (Mq)^\top (Mk)$. Similarly,

$$\|q-k\|^2_\Sigma = (q-k)^\top \Sigma (q-k) = (q-k)^\top M^\top M (q-k) = \|M(q-k)\|^2 = \|Mq - Mk\|^2.$$

Now assume $\Lambda = \mathrm{Cov}(q) = \mathrm{Cov}(k) \succ 0$ and choose $\Sigma = \Lambda^{-1}$ with $M = \Lambda^{-1/2}$. Then

$$\mathrm{Cov}(Mq) = M\,\mathrm{Cov}(q)\,M^\top = \Lambda^{-1/2}\,\Lambda\,\Lambda^{-1/2} = I_d.$$

Now write $\Lambda = U\mathrm{diag}(\lambda_1,\dots,\lambda_d)U^\top$ with U orthogonal and $\lambda_i > 0$. Then $\Lambda^{-1} = U\mathrm{diag}(\lambda_1^{-1},\dots,\lambda_d^{-1})U^\top$, and with $\delta := U^\top(q-k)$, we obtain

$$\|q-k\|^2_{\Lambda^{-1}} = (q-k)^\top \Lambda^{-1}(q-k) = \delta^\top \mathrm{diag}(\lambda_1^{-1},\dots,\lambda_d^{-1})\delta = \sum_{i=1}^d \delta_i^2/\lambda_i,$$

as claimed. □

The identity $\|q-k\|^2_{\Lambda^{-1}} = \sum_i (\delta_i/\sqrt{\lambda_i})^2$ shows that the Mahalanobis distance measures displacements in units of principal standard deviations. Thus, directions with large variance are naturally downweighted and correlations are removed by whitening. In the attention kernel, this corresponds to applying the standard dot-product softmax kernel to the re-embedded coordinates $\tilde{q} = Mq$ and $\tilde{k} = Mk$. This motivates learning Σ as a tractable way to adapt the attention geometry to anisotropic inputs, rather than relying on the transformer to reshape representations toward isotropy.

D RELATED WORK

Random Feature Kernels. Kernel methods have long been valued for their ability to model complex relationships. However, their computational cost has been a significant limitation. Using random basis approximations (Rahimi & Recht, 2008; 2007), Random Fourier Features approximate shift-invariant kernels by mapping inputs into a randomized feature space while achieving linear complexity. Subsequent work has characterized approximation error and improved sample efficiency via structured and orthogonal feature constructions or data-dependent sampling viewpoints such as leverage scores and kernel quadrature (Sutherland & Schneider, 2015; Le et al., 2013; Yu et al., 2016; Choromanski et al., 2017; Alaoui & Mahoney, 2015; Bach, 2017). These data-dependent views suggest that when the input distribution is anisotropic, Monte Carlo error is often dominated by a subset of directions, motivating sampling schemes that align with the geometry of the inputs rather than treating all directions equally. These developments have enabled large-scale kernel learning in practical domains such as speech recognition (May et al., 2019).

Efficient Transformers. The quadratic complexity of the attention mechanism in transformers has been a limiting factor. Extending random feature kernels to attention, Performer (Choromanski et al., 2021) uses positive (orthogonal) random features (Choromanski et al., 2017) to obtain an asymptotically unbiased softmax kernel estimator with linear complexity, while related linear attention variants adopt alternative normalizations (Katharopoulos et al., 2020). Other approximation methods achieve efficiency through low-rank decompositions, sketching-based methods, or sparsity (Wang et al., 2020; Xiong et al., 2021; Chen et al., 2021; Lee et al., 2024; Kacham et al., 2024; Kitaev et al., 2020). Recent work has also studied the approximation gap of random-feature attention and proposed variance-reduction mechanisms such as control variates (Zheng et al., 2023). Meanwhile, complementary approaches accelerate attention by optimizing memory access and parallelism (Dao et al., 2022; Dao, 2023; Shah et al., 2024; Li et al., 2023).

Filling the Gap. Despite this progress, random-feature attention mechanisms typically sample projections from an isotropic distribution, which can exhibit high Monte Carlo variance when query–key representations are anisotropic, a common phenomenon in pretrained Transformers (Ethayarajh, 2019; Godey et al., 2024). More broadly, data-dependent sampling can substantially reduce Monte Carlo error, but applying it to PRF-based attention is non-trivial because the optimal sampling distribution is input-dependent and costly to compute. DARKFormer fills this gap by learning the sampling geometry through a learned covariance for positive random features, connecting to data-dependent kernel sampling (Alaoui & Mahoney, 2015; Bach, 2017) and enabling importance-sampling behavior without explicit per-sample importance weights (Sernau et al., 2024).