Random Gegenbauer Features for Scalable Kernel Methods

Insu Han [*1] Amir Zandieh [*2] Haim Avron [3]

Abstract

We propose efficient random features for approximating a new and rich class of kernel functions that we refer to as *Generalized Zonal Kernels (GZK)*. Our proposed GZK family, generalizes the zonal kernels (i.e., dot-product kernels on the unit sphere) by introducing *radial factors* in the Gegenbauer series expansion of these kernel functions. The GZK class of kernels includes a wide range of ubiquitous kernel functions such as the entirety of dot-product kernels as well as the Gaussian and the recently introduced Neural Tangent kernels. Interestingly, by exploiting the reproducing property of the *Gegenbauer (Zonal) Harmonics*, we can construct efficient random features for the GZK family based on randomly oriented Gegenbauer harmonics. We prove subspace embedding guarantees for our Gegenbauer features which ensures that our features can be used for approximately solving learning problems such as kernel k-means clustering, kernel ridge regression, etc. Empirical results show that our proposed features outperform recent kernel approximation methods.

1. Introduction

Kernel methods are an important family of learning algorithms, which are applicable for a wide range of tasks, e.g. regression (Saunders et al., 1998), clustering (Dhillon et al., 2004), graph learning (Vishwanathan et al., 2010), nonparametric modeling (Rasmussen, 2004) as well as wide deep neural networks analysis (Jacot et al., 2018; Lee et al., 2019). However, unfortunately, they suffer from scalability issues, often due to the fact that applying the aforementioned methods requires operating on the kernel matrix (Gram matrix) of the data, whose size scales quadratically in the number of training samples. For example, solving kernel ridge regression generally requires a prohibitively large quadratic memory and at least quadratic runtime. To alleviate this issue, there has been a long line of efforts on efficiently approximating kernel matrices by low-rank factors (Williams & Seeger, 2001; Rahimi & Recht, 2009; Avron et al., 2014; Alaoui & Mahoney, 2015; Musco & Musco, 2017; Avron et al., 2017b; Zandieh et al., 2020; Ahle et al., 2020; Woodruff & Zandieh, 2020). Most relevant to this work is the widely used *random features* approach, originally proposed by Rahimi & Recht (2009).

In this work, we propose efficient random features for approximating a new and rich class of kernel functions that we refer to as *Generalized Zonal Kernels (GZK)* (see Definition 3). Our proposed class of kernels extends the zonal kernels (i.e., dot-product kernels restricted to the unit sphere) to entire $\mathbb{R}^d$ space, and includes a wide range of ubiquitous kernels, e.g. the entire family of dot-product kernels, the Gaussian kernel, and the recently introduced Neural Tangent kernels (Jacot et al., 2018). We start by considering the series expansion of zonal kernel functions in terms of the Gegenbauer polynomials, which are central in our analysis. Then we generalize these kernels by allowing *radial factors* in the Gegenbauer expansion. We construct the GZK family of kernels in Section 3.2. We design efficient random features for this class of kernels by exploiting various properties of Gegenbauer polynomials and using leverage scores sampling techniques (Li et al., 2013).

Specifically, for a given GZK function and its corresponding kernel matrix $\boldsymbol{K} \in \mathbb{R}^{n \times n}$, we seeks to find a low-rank matrix that can serve as a proxy to the kernel matrix $\boldsymbol{K}$. We present an algorithm that for given $\varepsilon, \lambda > 0$, computes a matrix $\boldsymbol{Z} \in \mathbb{R}^{m \times n}$ such that $\boldsymbol{Z}^\top \boldsymbol{Z}$ is an (ε, λ)-*spectral approximation* to the GZK kernel matrix $\boldsymbol{K}$, meaning that

$$\frac{\boldsymbol{K} + \lambda \boldsymbol{I}}{1+\varepsilon} \preceq \boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I} \preceq \frac{\boldsymbol{K} + \lambda \boldsymbol{I}}{1-\varepsilon}. \quad (1)$$

The spectral approximation guarantee can be directly used to obtain statistical guarantees for downstream kernel-based learning applications, such as bounds on the empirical risk of kernel ridge regression (Avron et al., 2017b).

1.1. Overview of Our Contributions

The Gegenbauer polynomials are a family of orthogonal polynomials that include Chebyshev and Legendre poly-

[*] Equal contribution [1] Yale University [2] Max-Planck-Institut für Informatik [3] Tel Aviv University. Correspondence to: Insu Han <insu.han@yale.edu>, Amir Zandieh <azandieh@mpi-inf.mpg.de>, Haim Avron <haimav@tauex.tau.ac.il>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

nomials and are widely employed in approximation theory (Gautschi, 2004). The Gegenbauer polynomials naturally provide positive definite dot-product kernels on the unit sphere known as the *Gegenbauer (Zonal) Harmonics*. In this work, we first define a rich class of kernels based on Gegenbauer harmonics and then present efficient random features for this new family of kernels by using the fact that Gegenbauer harmonics induce a natural feature map on themselves because of their reproducing property (see Lemma 1 for details). To the best of our knowledge, this is the first work on random features of orthogonal polynomials with provable guarantees. We analyze our proposed random features and prove that they spectrally approximate the exact kernel matrix. Our contributions are listed as follows,

- We extend the zonal kernels from unit sphere to entire $\mathbb{R}^d$ by adding radial components to the Gegenbauer series expansion of such kernels in Definition 3. Then we derive the Mercer decomposition of this class of kernels based on Gegenbauer polynomials in Lemma 5.

- We show that our newly proposed class of kernels is rich and contains all dot-product kernels Lemma 4, as well as Gaussian and Neural Tangent kernels Appendix C.

- We propose efficient random features for our proposed class of kernels in Definition 8 and prove both spectral approximation and projection-cost preserving guarantees for our proposed features in Theorem 9 and Theorem 10. These properties ensure that our random features can be used for downstream learning tasks such as kernel regression, kernel k-means, and principal/canonical component analysis, see Appendix A.

- We apply our main spectral approximation results on dot-product and Gaussian kernels and show our method gives improved random features for these types of kernels in Theorem 11 and Theorem 12.

- Our empirical results verify that the proposed method outperforms previous approaches for accelerating kernel ridge regression and kernel k-means methods.

1.2. Related Work

A popular line of work on kernel approximation is based on the random Fourier features method (Rahimi & Recht, 2009), which works well for shift-invariant kernels and with some modifications can embed the Gaussian kernel near optimally in constant dimension (Avron et al., 2017b). Other random feature constructions have been suggested for a variety of kernels, e.g., arc-cosine kernels (Cho & Saul, 2009), polynomial kernels (Pennington et al., 2015), and Neural Tangent kernels (Zandieh et al., 2021).

For the polynomial kernel, sketching methods have been developed extensively (Avron et al., 2014; Pham & Pagh, 2013; Woodruff & Zandieh, 2020; Song et al., 2021). For example, Ahle et al. (2020) proposed a subspace embedding for high-degree Polynomial kernels as well as the Gaussian kernel. However, approximating non-polynomial kernels using these tools require sketching the Taylor expansion of the kernel which can perform somewhat poorly due to slow convergence rate of Taylor series. On the other hand, we focus on Gegenbauer series that generally converge faster (Fox & Parker, 1968; Mason & Handscomb, 2002).

Another popular kernel approximation approach is the Nyström method (Williams & Seeger, 2001; Yang et al., 2012). While the recursive Nyström sampling of Musco & Musco (2017) can embed kernel matrices using near optimal number of landmarks, this method is inherently data dependent, so unlike our data oblivious random features, it cannot provide one-round distributed protocols and/or single-pass streaming algorithms.

2. Preliminaries

Notations. Throughout the paper, all logarithms are natural $\log$ functions unless we specify the base. We denote by $\mathbb{S}^{d-1}$ the unit sphere in d dimension. We use $|\mathbb{S}^{d-1}| = \frac{2\pi^{d/2}}{\Gamma(d/2)}$ to denote the surface area of the unit sphere $\mathbb{S}^{d-1}$ and $\mathcal{U}(\mathbb{S}^{d-1})$ to denote the uniform probability distribution on $\mathbb{S}^{d-1}$. We use $\mathbb{1}_{\{\mathcal{E}\}}$ as an indicator function for event $\mathcal{E}$. All matrices are in boldface, e.g., $\boldsymbol{K}$, and we let $\boldsymbol{I}_n$ be the $n \times n$ identity matrix and sometimes omit the subscript. For any function $\kappa(\cdot)$ and any integer i we denote the i^{th} derivative of κ with $\kappa^{(i)}(t)$ or $\frac{d^i}{dt^i}\kappa(t)$. We use $\|\cdot\|$ and $\|\cdot\|_{\mathrm{op}}$ to denote the ℓ_2-norm of vectors and the operator norm of matrices, respectively. The *statistical dimension* of a positive semidefinite matrix $\boldsymbol{K}$ and parameter $\lambda \geq 0$ is defined as $s_\lambda := \mathrm{Tr}\big(\boldsymbol{K}(\boldsymbol{K}+\lambda\boldsymbol{I})^{-1}\big)$.

2.1. Gegenbauer Polynomials

The Gegenbauer polynomial (a.k.a. *ultraspherical polynomial*) of degree $\ell \geq 0$ in dimension $d \geq 2$ is given by

$$P_d^\ell(t) := \sum_{j=0}^{\lfloor \ell/2 \rfloor} c_j \cdot t^{\ell-2j} \cdot (1-t^2)^j, \tag{2}$$

where $c_0 = 1$ and $c_{j+1} = -\frac{(\ell-2j)(\ell-2j-1)}{2(j+1)(d-1+2j)} c_j$ for $j = 0, 1, \dots \lfloor \ell/2 \rfloor - 1$. This class of polynomials includes Chebyshev polynomials of the first kind when $d = 2$ and Legendre polynomials when $d = 3$. Furthermore, when $d = \infty$, these polynomials reduce to monomials i.e., $P_\infty^\ell(t) = t^\ell$. They also fall into the important class of Jacobi polynomials.

Gegenbauer polynomials satisfy an orthogonality property on interval $[-1,1]$ with respect to measure $(1-t^2)^{\frac{d-3}{2}}$:

$$\int_{-1}^{1} P_d^{\ell}(t) P_d^{\ell'}(t)(1-t^2)^{\frac{d-3}{2}}\, dt = \frac{\left|\mathbb{S}^{d-1}\right| \cdot \mathbb{1}_{\{\ell=\ell'\}}}{\alpha_{\ell,d} \cdot \left|\mathbb{S}^{d-2}\right|}, \quad (3)$$

where $\alpha_{\ell,d}$ is the dimensionality of the space of *spherical harmonics* of order ℓ in dimension d defined as $\alpha_{0,d} := 1, \alpha_{1,d} := d$ and for $\ell \geq 2$

$$\alpha_{\ell,d} := \binom{d+\ell-1}{\ell} - \binom{d+\ell-3}{\ell-2}. \quad (4)$$

The following alternative expression for $P_d^{\ell}(t)$, proved in (Morimoto, 1998), is known as Rodrigues' formula,

$$P_d^{\ell}(t) = \frac{(-1)^{\ell}\Gamma\left(\frac{d-1}{2}\right)}{2^{\ell}(1-t^2)^{\frac{d-3}{2}}\Gamma\left(\ell+\frac{d-1}{2}\right)} \frac{d^{\ell}\left(1-t^2\right)^{\ell+\frac{d-3}{2}}}{dt^{\ell}} \quad (5)$$

for any $d \geq 3$.

2.2. Hilbert Space of Function in $L^2\left(\mathbb{S}^{d-1}, \mathbb{R}^s\right)$

For any integer $s \geq 1$ and any vector-valued functions $f, g \in L^2(\mathbb{S}^{d-1}, \mathbb{R}^s)$ meaning that $f, g : \mathbb{S}^{d-1} \to \mathbb{R}^s$, we define the inner product of these maps as follows,

$$\langle f, g\rangle_{L^2(\mathbb{S}^{d-1},\mathbb{R}^s)} := \mathbb{E}_{w\sim\mathcal{U}(\mathbb{S}^{d-1})}\left[\langle f(w), g(w)\rangle\right]. \quad (6)$$

With this inner product, $L^2\left(\mathbb{S}^{d-1}, \mathbb{R}^s\right)$ is a *Hilbert space*, with norm $\|f\|_{L^2(\mathbb{S}^{d-1},\mathbb{R}^s)} = \sqrt{\langle f, f\rangle_{L^2(\mathbb{S}^{d-1},\mathbb{R}^s)}}$. Furthermore, we shorten the notation for the space of square-integrable functions $L^2(\mathbb{S}^{d-1}, \mathbb{R})$ to $L^2(\mathbb{S}^{d-1})$.

2.3. Gegenbauer (Zonal) Harmonics

The Gegenbauer polynomials naturally provide positive definite dot-product kernels on the unit sphere $\mathbb{S}^{d-1}$ known as Gegenbauer (zonal) harmonics. In fact, Schoenberg (1942) proved that a dot-product kernel $k(x, y) = \kappa(\langle x, y\rangle)$ is positive definite if and only if $\kappa(t) = \sum_{\ell=0}^{\infty} c_{\ell} P_d^{\ell}(t)$ with all $c_{\ell} \geq 0$ (see Theorem 3 therein).

Particularly the following reproducing property of Gegenbauer harmonics is useful which follows from the Funk–Hecke formula (See (Atkinson & Han, 2012)).

Lemma 1 (Reproducing property of Gegenbauer Harmonics)**.** *Let $P_d^{\ell}(\cdot)$ be the Gengenbauer polynomial of degree ℓ in dimension d. For any $x, y \in \mathbb{S}^{d-1}$:*

$$P_d^{\ell}(\langle x, y\rangle) = \alpha_{\ell,d} \cdot \mathbb{E}_{w\sim\mathcal{U}(\mathbb{S}^{d-1})}\left[P_d^{\ell}\left(\langle x, w\rangle\right) P_d^{\ell}\left(\langle y, w\rangle\right)\right],$$

Furthermore, for any $\ell' \neq \ell$:

$$\mathbb{E}_{w\sim\mathcal{U}(\mathbb{S}^{d-1})}\left[P_d^{\ell}\left(\langle x, w\rangle\right) \cdot P_d^{\ell'}\left(\langle y, w\rangle\right)\right] = 0.$$

3. Generalized Zonal Kernels (GZK)

In this section, we introduce our proposed class of Generalized Zonal Kernels (GZK). We start by deriving a practical Mercer decomposition of zonal kernels, i.e., dot-product kernels on the unit sphere, and then extend it to a large class of kernel functions – Generalized Zonal Kernels.

3.1. Warm-up: Mercer Decomposition of Zonal Kernels

A function $k : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to \mathbb{R}$ is called *zonal kernel* if it can be represented by $k(x, y) = \kappa(\langle x, y\rangle)$ for some scalar function $\kappa : [-1, 1] \to \mathbb{R}$. Note that zonal kernels are rotation invariant, i.e., $k(x, y) = k(\boldsymbol{R}x, \boldsymbol{R}y)$ for any rotation matrix $\boldsymbol{R} \in \mathbb{R}^{d\times d}$. Due to this property, zonal kernels have been used in various geo-science applications including climate change simulation (Sanderson et al., 2010), Ozone prediction (Su et al., 2020) and mantle convection (Bercovici, 2003).

Assuming that the Gegenbauer series expansion of the function $\kappa(\cdot)$ is $\kappa(t) = \sum_{\ell=0}^{\infty} c_{\ell} P_d^{\ell}(t)$, by orthogonality property in Eq. (3), the Gegenbauer coefficients c_{ℓ} can be computed as

$$c_{\ell} = \alpha_{\ell,d} \cdot \frac{\left|\mathbb{S}^{d-2}\right|}{\left|\mathbb{S}^{d-1}\right|} \cdot \int_{-1}^{1} \kappa(t) P_d^{\ell}(t)(1-t^2)^{\frac{d-3}{2}} dt. \quad (7)$$

So, we have

$$k(x, y) = \kappa(\langle x, y\rangle) = \sum_{\ell=0}^{\infty} c_{\ell} \cdot P_d^{\ell}(\langle x, y\rangle). \quad (8)$$

It is known that polynomial approximation with Chebyshev series (i.e., $d = 2$) generally has faster convergence rate compared to Taylor series (i.e., $d = \infty$) (Fox & Parker, 1968; Mason & Handscomb, 2002). We empirically verify that the Gegenbauer series (i.e., $2 < d < \infty$) interpolates between Taylor and Chebyshev series in Section 6.1.

Throughout this work, we assume that $\kappa(\cdot)$ is an analytic function so that the corresponding Gegenbauer series expansion exists and converges. With Eq. (8) in-hand and applying Lemma 1 we obtain a Mercer decomposition for zonal kernels.

Lemma 2 (Feature map for zonal kernels)**.** *Suppose $\kappa : [-1, 1] \to \mathbb{R}$ is analytic and let $\{c_{\ell}\}_{\ell=0}^{\infty}$ be the coefficients of its Gegenbauer series expansion in dimension $d \geq 2$. For $x, w \in \mathbb{S}^{d-1}$, define the real-valued function $\phi_x \in L^2(\mathbb{S}^{d-1})$ as*

$$\phi_x(w) := \sum_{\ell=0}^{\infty} \sqrt{c_{\ell} \cdot \alpha_{\ell,d}} \cdot P_d^{\ell}(\langle x, w\rangle). \quad (9)$$

Then, for all $x, y \in \mathbb{S}^{d-1}$, it holds that

$$\mathbb{E}_{w\sim\mathcal{U}(\mathbb{S}^{d-1})}\left[\phi_x(w) \cdot \phi_y(w)\right] = \kappa(\langle x, y\rangle). \quad (10)$$

The proof of Lemma 2 can be found in Appendix D.1.

3.2. Extension to Dot-product Kernels and Beyond

In this section, we generalize the zonal kernel functions from $\mathbb{S}^{d-1}$ to entire $\mathbb{R}^d$ by letting the kernel function be factorizable into angular and radial parts.

Definition 3 (Generalized zonal kernels)**.** *For an integer $s \geq 1$ and a sequence of vector-valued functions $h_\ell : \mathbb{R} \to \mathbb{R}^s$ for $\ell = 0, 1, \dots$, we define the* generalized zonal kernel (GZK) *of order s as*

$$k(x, y) := \sum_{\ell=0}^{\infty} \langle h_\ell(\|x\|), h_\ell(\|y\|) \rangle P_d^\ell \left(\frac{\langle x, y \rangle}{\|x\| \|y\|} \right). \quad (11)$$

We remark that for any series of real-valued vector functions $h_\ell : \mathbb{R} \to \mathbb{R}^s$, Eq. (11) defines a valid positive definite kernel (we give the Mercer decomposition of the GZK function in Lemma 4). While we defined the GZK functions for finite order s, the definition can be extended to include $s = +\infty$ by letting $h_\ell(t)$ be a map to the square-summable sequences (a.k.a. l^2-sequence-space[1]) and letting the term $\langle h_\ell(\|x\|), h_\ell(\|y\|) \rangle$ in Eq. (11) be the standard l^2-inner-product of sequences $h_\ell(\|x\|), h_\ell(\|y\|)$.

The class of GZK in Definition 3 includes a wide range of familiar kernel functions such as all dot-product kernels, the Gaussian and Neural Tangent Kernels. In the following lemma we show that dot-products kernels are GZK.

Lemma 4 (Dot-product kernels are GZKs)**.** *For any $x, y \in \mathbb{R}^d$, any integer $d \geq 3$, and any dot-product kernel $k(x, y) = \kappa(\langle x, y \rangle)$ with analytic $\kappa(\cdot)$, the eigenfunction expansion of $k(x, y)$ can be written as,*

$$k(x, y) := \sum_{\ell=0}^{\infty} \left(\sum_{i=0}^{\infty} \widetilde{h}_{\ell,i}(\|x\|) \widetilde{h}_{\ell,i}(\|y\|) \right) P_d^\ell \left(\frac{\langle x, y \rangle}{\|x\| \|y\|} \right),$$

where $\widetilde{h}_{\ell,i}(\cdot)$ are real-valued monomials defined as follows for integers $\ell, i \geq 0$ and any $t \in \mathbb{R}$:

$$\widetilde{h}_{\ell,i}(t) := \sqrt{\frac{\alpha_{\ell,d}}{2^\ell} \frac{\Gamma(\frac{d}{2}) \ \kappa^{(\ell+2i)}(0)}{\sqrt{\pi}(2i)!} \frac{\Gamma(i + \frac{1}{2})}{\Gamma(i + \ell + \frac{d}{2})}} \cdot t^{\ell+2i}. \quad (12)$$

The proof of Lemma 4 is provided in Appendix B. The proof starts by expressing the monomials in Taylor series expansion of $\kappa(\langle x, y \rangle)$ in the Gegenbauer basis, i.e., $\langle x, y \rangle^j = (\|x\| \|y\|)^j \cdot \langle \frac{x}{\|x\|}, \frac{y}{\|y\|} \rangle^j = (\|x\| \|y\|)^j \cdot \sum_{\ell=0}^{j} c_\ell P_d^\ell (\frac{\langle x, y \rangle}{\|x\| \|y\|})$. The coefficients c_ℓ can be computed using Eq. (7) along with the Rodrigues' formula in Eq. (5). Lemma 4 shows that any dot-product kernel $\kappa(\cdot)$ is indeed a GZK of order s if its derivatives $\kappa^{(2i)}(t)$ at $t = 0$ for $i \geq s$ are zeros. If the derivatives of $\kappa(t)$ do not vanish at $t = 0$ then the kernel can be a GZK of potentially infinite order $s = +\infty$ with $h_\ell(t) = [\widetilde{h}_{\ell,i}(t)]_{i=0}^\infty$, where $\widetilde{h}_{\ell,i}(\cdot)$ are defined as per Eq. (12). In Section 5 we show that very often $\widetilde{h}_{\ell,i}(\cdot)$ rapidly decay with respect to i, thus dot-product kernels can be tightly approximated by GZKs with small finite order s. Furthermore, when inputs are on the unit sphere, i.e., $\|x\| = 1$, the radial functions $\widetilde{h}_{\ell,i}(\|x\|)$ turn out to be constant so a dot-product kernel on the sphere (a.k.a a zonal kernel as per Eq. (8)) is a GZK of order $s = 1$.

Now we present a feature map for the GZK which will be the basis of our efficient random features.

Lemma 5 (Feature map for GZK)**.** *Consider a GZK $k(\cdot, \cdot)$ with real-valued functions $h_\ell : \mathbb{R} \to \mathbb{R}^s$ for $\ell = 0, 1, \dots$ as in Definition 3. For any $x \in \mathbb{R}^d, w \in \mathbb{S}^{d-1}$, define the function $\phi_x \in L^2(\mathbb{S}^{d-1}, \mathbb{R}^s)$ as*

$$\phi_x(w) := \sum_{\ell=0}^{\infty} \sqrt{\alpha_{\ell,d}} \ h_\ell(\|x\|) \ P_d^\ell \left(\frac{\langle x, w \rangle}{\|x\|} \right). \quad (13)$$

Then, for any $x, y \in \mathbb{R}^d$, it holds that

$$\mathbb{E}_{w \sim \mathcal{U}(\mathbb{S}^{d-1})} [\langle \phi_x(w), \phi_y(w) \rangle] = k(x, y).$$

The proof of Lemma 5 is given in Appendix D.2. For this feature map to be well-defined we require the series in Eq. (13) to be convergent for every $x \in \mathbb{R}^d$ in our dataset.

Remark. Several works have attempted to extend simple zonal kernels from $\mathbb{S}^{d-1}$ to $\mathbb{R}^d$ (Smola et al., 2001; Cho & Saul, 2010; Scetbon & Harchaoui, 2021). These papers focus on the eigensystem of the dot-product kernels based on the spherical harmonics. However, it is intractable to compute spherical harmonics in general (Minh et al., 2006) which renders the above-mentioned eigendecomposition results mainly existential and non-practical. On the other hand, we propose a computationally practical Mercer decomposition of the GZK (and a fortiori dot-product kernels) in Lemma 5, which unlike (Smola et al., 2001) does not rely on spherical harmonics and will lead to efficient kernel approximations.

4. Spectral Approximation of GZK

In this section, we propose efficient random features for GZK kernels based on our feature map in Eq. (13) and then analyze their approximation guarantee. We first introduce the following notations that are essential in our analysis.

Consider a dataset $\boldsymbol{X} = [x_1, , \dots, x_n] \in \mathbb{R}^{d \times n}$ and a GZK $k(\cdot, \cdot)$ as per Definition 3 and let the n-by-n kernel matrix $\boldsymbol{K}$ be defined as $[\boldsymbol{K}]_{i,j} := k(x_i, x_j)$. Let ϕ_{x_j} be the feature map defined in Eq. (13) for all $j \in [n]$. For $v \in \mathbb{R}^n$, we define an operator $\boldsymbol{\Phi} : \mathbb{R}^n \to L^2 \left(\mathbb{S}^{d-1}, \mathbb{R}^s \right)$ (a.k.a. quasi-

[1] l^2 space not be confused with the index ℓ in functions $h_\ell(\cdot)$

matrix) as follows,

$$\mathbf{\Phi} \cdot v := \sum_{j=1}^{n} v_j \cdot \phi_{x_j}. \tag{14}$$

The adjoint of this operator $\mathbf{\Phi}^* : L^2\left(\mathbb{S}^{d-1}, \mathbb{R}^s\right) \to \mathbb{R}^n$ is the following for $f \in L^2\left(\mathbb{S}^{d-1}, \mathbb{R}^s\right)$ and $j \in [n]$,

$$[\mathbf{\Phi}^* f]_j = \langle \phi_{x_j}, f \rangle_{L^2(\mathbb{S}^{d-1}, \mathbb{R}^s)}, \tag{15}$$

where the inner product above is defined as per Eq. (6). With this definition, it follows from Lemma 5 that

$$\mathbf{\Phi}^* \mathbf{\Phi} = \boldsymbol{K}.$$

Our approach for spectrally approximating $\boldsymbol{K}$ is sampling the "rows" of the quasi-matrix $\mathbf{\Phi}$ with probabilities proportional to their ridge leverage scores (Li et al., 2013). The ridge leverage scores of $\mathbf{\Phi}$ are defined as follows,

Definition 6 (Ridge leverage scores of $\mathbf{\Phi}$)**.** *Let $\mathbf{\Phi} : \mathbb{R}^n \to L^2(\mathbb{S}^{d-1}, \mathbb{R}^s)$ be the operator defined in Eq. (14). Also, for every $w \in \mathbb{S}^{d-1}$, define $\Phi_w \in \mathbb{R}^{n \times s}$ as,*

$$\Phi_w := \left[\phi_{x_1}(w), \phi_{x_2}(w), \ldots \phi_{x_n}(w)\right]^\top. \tag{16}$$

For any $\lambda > 0$, the row leverage scores of $\mathbf{\Phi}$ are defined as,

$$\tau_\lambda(w) := \operatorname{Tr}\left(\Phi_w^\top \cdot (\boldsymbol{K} + \lambda \boldsymbol{I})^{-1} \cdot \Phi_w\right). \tag{17}$$

An important quantity for the spectral approximation to $\boldsymbol{K}$ is the *average* of the ridge leverage scores with respect to the uniform distribution on $\mathbb{S}^{d-1}$ which can be shown to be equal to the *statistical dimension* of kernel matrix $\boldsymbol{K}$:

$$\mathbb{E}_{w \sim \mathcal{U}(\mathbb{S}^{d-1})}[\tau_\lambda(w)] = \operatorname{Tr}(\boldsymbol{K}(\boldsymbol{K} + \lambda \boldsymbol{I})^{-1}) = s_\lambda. \tag{18}$$

Remark. Our definition of leverage scores is slightly non-standard and different from the prior works such as (Avron et al., 2017a; 2019) because it is not normalized with the distribution of $w \sim \mathcal{U}(\mathbb{S}^{d-1})$. The difference stems from the definition of inner product in $L^2(\mathbb{S}^{d-1}, \mathbb{R}^s)$ space in Eq. (6).

4.1. Random Features Based on the Leverage Scores

In this section, we propose our random features according to the leverage scores of $\mathbf{\Phi}$, and show that they are able to spectrally approximate $\boldsymbol{K}$. However, computing the leverage scores exactly is expensive in general and even if we could it is not necessarily easy to sample from them efficiently. So, we focus on approximating the leverage scores of the GZK with a distribution which is easy to sample from. Specifically, we find a $\widehat{\tau}_\lambda(\cdot)$ such that $\widehat{\tau}_\lambda(w) \ge \tau_\lambda(w)$ for all $w \in \mathbb{S}^{d-1}$. For any GZK and its corresponding feature operator defined in Eq. (14), we have the following upper bound,

Lemma 7 (Upper bound on leverage scores of GZK)**.** *For any dataset $\boldsymbol{X} = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d \times n}$, let $\mathbf{\Phi}$ be the feature operator for the order s GZK on $\boldsymbol{X}$ defined in Eq. (14). For any $\lambda > 0$ and $w \in \mathbb{S}^{d-1}$, the ridge leverage scores of $\mathbf{\Phi}$ defined in Definition 6 are uniformly upper bounded by*

$$\tau_\lambda(w) \le \sum_{\ell=0}^{\infty} \alpha_{\ell,d} \min\left\{ \frac{\pi^2(\ell+1)^2}{6\lambda} \sum_{j \in [n]} \left\| h_\ell(\|x_j\|) \right\|^2, s \right\}.$$

Proof Sketch. To find a proper upper bound on the ridge leverage function, we first show that it can be expressed as the sum of a collection of regularized least-squares problems, i.e., $\tau_\lambda = \sum_{i=1}^s \tau_i^*$ for

$$\tau_i^* := \min_{g_i \in L^2(\mathbb{S}^{d-1}, \mathbb{R}^s)} \|g_i\|^2_{L^2(\mathbb{S}^{d-1}, \mathbb{R}^s)} + \lambda^{-1} \left\| \mathbf{\Phi}^* g_i - \Phi_w^i \right\|_2^2,$$

where $\Phi_w^i \in \mathbb{R}^n$ is the i^{th} column of matrix Φ_w defined in Eq. (16). Intuitively, the function $g_i(\sigma) = \sum_{\ell=q}^{\infty} \alpha_{\ell,d} \cdot P_d^\ell\left(\langle \sigma, w \rangle\right) \cdot e_i$, where e_i is the i^{th} standard basis vector in $\mathbb{R}^s$, can zero out the second term in the above objective function, by Lemma 1, while making the first term infinite $\|g_i\|^2_{L^2(\mathbb{S}^{d-1}, \mathbb{R}^s)} = \sum_{\ell=q}^{\infty} \alpha_{\ell,d}$. On the other hand, for $g_i = 0$, the first term in the objective function will be zero while the second term will be as large as $\lambda^{-1} \left\| \Phi_w^i \right\|_2^2$.

To find a balance between these two extremes, we make the heavy radial components in the second term small, i.e., ℓ's such that $\left\| h_\ell(\|x_j\|) \right\|^2$ is large, and ignore the small components to keep the norm of g_i as small as possible. Specifically, we choose the following feasible solution that is nearly optimal for the above least-squares problem

$$\widehat{g}_i(\sigma) = \left(\sum_{\ell=q}^{\infty} \alpha_{\ell,d} \mathbb{1}_{\left\{\sum_j \|h_\ell(\|x_j\|)\|^2 \ge \lambda s\right\}} P_d^\ell\left(\langle \sigma, w \rangle\right) \right) \cdot e_i.$$

Plugging this to the minimization problem gives the lemma. The full proof is in Appendix E. □

We will show in Section 5 that the bound in Lemma 7 is typically small for all practically important kernels because the radial components $h_\ell(\cdot)$ rapidly decay as ℓ increases. Inspired by this uniform bound on leverage score, we propose the following random features for the GZK by uniformly sampling the rows of the feature operator $\mathbf{\Phi}$ in Eq. (14).

Definition 8 (Random features for Generalized Zonal Kernels)**.** *For any GZK as per Definition 3 and dataset $\boldsymbol{X} \in \mathbb{R}^{d \times n}$, sample m i.i.d. points $w_1, \ldots, w_m \sim \mathcal{U}(\mathbb{S}^{d-1})$ and let $\Phi_{w_1}, \ldots, \Phi_{w_m} \in \mathbb{R}^{n \times s}$ be defined as per Eq. (16), then define the features matrix $\boldsymbol{Z} \in \mathbb{R}^{(m \cdot s) \times n}$ as:*

$$\boldsymbol{Z} := \frac{1}{\sqrt{m}} \cdot \left[\Phi_{w_1}, \ldots, \Phi_{w_m}\right]^\top. \tag{19}$$

These random features are *unbiased*, i.e., $\mathbb{E}\left[\boldsymbol{Z}^\top \boldsymbol{Z}\right] = \boldsymbol{K}$.

4.2. Main Theorems

We now formally prove that for the class of GZKs, the random features in Definition 8 yield a spectral approximation to the kernel matrix $\boldsymbol{K}$ with enough number of features.

Theorem 9 (Spectral approximation of GZK)**.** *For any dataset $\boldsymbol{X} = [x_1, x_2, \dots x_n] \in \mathbb{R}^{d\times n}$, let $\boldsymbol{K}$ be the corresponding GZK kernel matrix (Definition 3). For any $0 < \lambda \le \|\boldsymbol{K}\|_{\mathrm{op}}$, let $\boldsymbol{Z} \in \mathbb{R}^{(m\cdot s)\times n}$ be the random features matrix defined in Definition 8. Also let s_λ be the statistical dimension of $\boldsymbol{K}$. For any $\varepsilon, \delta > 0$, if $m \ge \frac{8}{3\varepsilon^2} \log \frac{16 s_\lambda}{\delta} \cdot \sum_{\ell=0}^{\infty} \alpha_{\ell,d} \min\left\{ \frac{\pi^2(\ell+1)^2}{6\lambda} \sum_{j\in[n]} \|h_\ell(\|x_j\|)\|^2, s \right\}$, then with probability of at least $1-\delta$,*

$$\frac{\boldsymbol{K}+\lambda\boldsymbol{I}}{1+\varepsilon} \preceq \boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I} \preceq \frac{\boldsymbol{K}+\lambda\boldsymbol{I}}{1-\varepsilon}. \tag{20}$$

We provide the proof of Theorem 9 in Appendix F. The proof follows the standard approach studied in (Avron et al., 2017b). By Lemma 7, there exists a bound $U \ge \tau_\lambda(w)$ for all $w \in \mathbb{S}^{d-1}$. This gives upper bounds of both the operator norm and the second moment of our kernel estimator. Applying a matrix concentration inequality (e.g., Corollary 7.3.3 in Tropp (2015)) with those bounds gives the result.

In addition to the basic spectral approximation guarantee of Theorem 9, we also prove that our random features method is able to produce *projection-cost preserving* samples.

Theorem 10 (Projection cost preserving GZK approximation)**.** *Let $\boldsymbol{K}$ be the GZK kernel matrix as in Theorem 9 with eigenvalues $\lambda_1 \ge \dots \ge \lambda_n$. For any positive integer r, let $\lambda := \frac{1}{r}\sum_{i=r+1}^{n} \lambda_i$ and let s_λ be the statistical dimension of $\boldsymbol{K}$. For any $\varepsilon, \delta > 0$, if $\boldsymbol{Z} \in \mathbb{R}^{(m\cdot s)\times n}$ is the random features matrix defined in Definition 8 with $m \ge \frac{8}{3\varepsilon^2}\log\frac{16 s_\lambda}{\delta} \cdot \sum_{\ell=0}^{\infty} \alpha_{\ell,d} \min\left\{ \frac{\pi^2(\ell+1)^2}{6\lambda} \sum_{j\in[n]} \|h_\ell(\|x_j\|)\|^2, s \right\}$, with probability at least $1-\delta$, the following holds for all rank-r orthonormal projections $\boldsymbol{P}$:*

$$1-\varepsilon \le \frac{\mathrm{Tr}\left(\boldsymbol{Z}^\top \boldsymbol{Z} - \boldsymbol{P}\boldsymbol{Z}^\top \boldsymbol{Z}\boldsymbol{P}\right)}{\mathrm{Tr}(\boldsymbol{K} - \boldsymbol{P}\boldsymbol{K}\boldsymbol{P})} \le 1+\varepsilon. \tag{21}$$

We prove Theorem 10 in Appendix G. This property ensures that it is possible to extract a near optimal low-rank approximation to the kernel matrix from our random features, thus they can be used for learning tasks such kernel k-means, principal component analysis (PCA) and Gaussian processes. We provide how the projection-cost preserving cost can be applied to these tasks in Appendix A.

5. Application to Popular Kernels

So far we have showed GZKs can be spectrally approximated using the random features we designed in Definition 8. We have also showed in Lemma 4 and Appendix C that all dot-product kernels as well as Gaussian and Neural Tangent Kernels are in the rich family of GZKs. Thus, our random features can be used to get a good spectral approximation for these kernels. In this section we answer the question of efficiency of our random features.

Note that Theorem 9 bounds the number of required features by $\sum_{\ell=0}^{\infty} \alpha_{\ell,d} \min\left\{ \frac{\pi^2(\ell+1)^2}{6\lambda} \sum_{j\in[n]} \|h_\ell(\|x_j\|)\|^2, s \right\}$. We show that for dot-product and Gaussian kernels and datasets with bounded radius, the radial components $\sum_{j\in[n]} \|h_\ell(\|x_j\|)\|^2$ decay very fast as ℓ increases and effectively only the terms with degree $\ell \lesssim \log\frac{n}{\lambda}$ matter. This way, we get simple bounds on the number of required features for these kernels and also show that the features given in Definition 8 are efficiently computable.

5.1. Dot-product Kernels

We proved in Lemma 4 that any dot-product kernel $k(x,y) = \kappa(\langle x,y\rangle)$ with analytic $\kappa(\cdot)$ is a GZK, thus can be spectrally approximated by Theorem 9. To bound the number of required random features, we need to know how fast the monomials $\widetilde{h}_{\ell,i}(\cdot)$ in Eq. (12) decay as a function of ℓ. To bound the decay of $\widetilde{h}_{\ell,i}(\cdot)$, we first need to quantify the growth rate of the derivatives of $\kappa(\cdot)$. We assume that derivatives of $\kappa(\cdot)$ at zero can be characterized by the following exponential growth.

Assumption 1. *For a dot-product kernel $\kappa(\cdot)$ suppose that there exist some constants $C_\kappa \ge 0$ and $\beta_\kappa \ge 1$ such that for any integer $\ell > d$, $\kappa^{(\ell)}(0) \le C_\kappa \cdot \beta_\kappa^\ell$.*

Schoenberg (1942) showed that for any dot-product kernel we have $\kappa^{(\ell)}(0) \ge 0$ for all ℓ. Assumption 1 is commonly observed in popular kernel functions. For example, the exponential kernel $\kappa(\langle x,y\rangle) = e^{\langle x,y\rangle}$ satisfies Assumption 1 with $C_\kappa = \beta_\kappa = 1$.

Now, for kernel $\kappa(\langle x,y\rangle)$ and positive integers s, q we let $k_{q,s}(x,y)$ be the order s GZK as per Definition 3 whose corresponding radial functions $h_\ell : \mathbb{R} \to \mathbb{R}^s$ are defined as follows for $i \in [s]$ and $\ell \le q$

$$[h_\ell(t)]_i = \sqrt{\frac{\alpha_{\ell,d}}{2^\ell} \frac{\Gamma(\frac{d}{2})\, \kappa^{(\ell+2i)}(0)}{\sqrt{\pi}(2i)!} \frac{\Gamma(i+\frac{1}{2})}{\Gamma(i+\ell+\frac{d}{2})}} \cdot t^{\ell+2i} \tag{22}$$

and $h_\ell(t) := 0$ for any $\ell > q$. We show that under Assumption 1, the GZK $k_{q,s}(x,y)$ tightly approximates $\kappa(\langle x,y\rangle)$ for reasonably small values of q and s, thus we can approximate $\kappa(\langle x,y\rangle)$ by invoking Theorem 9 on $k_{q,s}(x,y)$. Specifically, we prove the following theorem,

Theorem 11. *Suppose Assumption 1 holds for a dot-product kernel $\kappa(\langle x,y\rangle)$. Given $\boldsymbol{X} = [x_1,\dots,x_n] \in \mathbb{R}^{d\times n}$, assume that $\max_{j\in[n]} \|x_j\| \le r$. Let $\boldsymbol{K}$ be the kernel matrix corresponding to $\kappa(\cdot)$ and $\boldsymbol{X}$. For any $0 < \lambda \le \|\boldsymbol{K}\|_{\mathrm{op}}$ and $\varepsilon, \delta > 0$ let s_λ be the statistical dimension of $\boldsymbol{K}$ and define*

$q = \max\left\{d, 3.7r^2\beta_\kappa, r^2\beta_\kappa + \frac{d}{2}\log\frac{3r^2\beta_\kappa}{d} + \log\frac{C_\kappa n}{\varepsilon\lambda}\right\}$. *There exists a randomized algorithm that can output* $\boldsymbol{Z} \in \mathbb{R}^{m\times n}$ *with* $m = \frac{25q^2}{3\varepsilon^2}\cdot\binom{q+d-1}{q}\cdot\log\frac{16s_\lambda}{\delta}$, *such that with probability at least* $1-\delta$, $\boldsymbol{Z}^\top\boldsymbol{Z}$ *is an* (ε,λ)*-spectral approximation to* $\boldsymbol{K}$ *as per Eq. (1). Furthermore,* $\boldsymbol{Z}$ *can be computed in time* $\mathcal{O}((m/q)\cdot\mathrm{nnz}(\boldsymbol{X}))$.

In Appendix H we provide more formal statement and proof.

5.2. Gaussian Kernel

The Gaussian kernel $g(x,y) = e^{-\|x-y\|_2^2/2}$ is a GZK as shown in Lemma 15. Therefore, we can spectrally approximate it on datasets with bounded ℓ_2 radius efficiently.

In particular, we first approximate $g(x,y)$ by a low-degree GZK and then invoke Theorem 9 on the resulting low-degree kernel. More precisely, for positive integers s,q we let $g_{q,s}(x,y)$ be the order-s GZK as per Definition 3 whose corresponding radial functions $h_\ell : \mathbb{R}\to\mathbb{R}^s$ are defined as follows for $i\in[s]$ and $\ell \le q$

$$[h_\ell(t)]_i = \sqrt{\frac{\alpha_{\ell,d}}{2^\ell}\frac{\Gamma(\frac{d}{2})}{\sqrt{\pi}(2i)!}\frac{\Gamma(i+\frac{1}{2})}{\Gamma(i+\ell+\frac{d}{2})}}\cdot t^{\ell+2i}e^{-\frac{t^2}{2}} \quad (23)$$

and $h_\ell(t) := 0$ for any $\ell > q$. We show that $g_{q,s}(x,y)$ tightly approximates $g(x,y)$ for reasonably small values of q and s, thus we can approximate the Gaussian kernel matrix by invoking Theorem 9 on $g_{q,s}(x,y)$. Specifically, we prove,

Theorem 12. *Given* $\boldsymbol{X} = [x_1,\dots,x_n] \in \mathbb{R}^{d\times n}$ *for* $d\ge 3$, *assume that* $\max_{j\in[n]}\|x_j\| \le r$. *Let* $\boldsymbol{K}\in\mathbb{R}^{n\times n}$ *be the corresponding Gaussian kernel matrix* $[\boldsymbol{K}]_{i,j} = e^{-\|x_i-x_j\|_2^2/2}$. *For any* $0 < \lambda \le \|\boldsymbol{K}\|_{\mathrm{op}}$ *and* $\varepsilon,\delta > 0$, *let* s_λ *denote the statistical dimension of* $\boldsymbol{K}$ *and define* $q = \max\left\{3.7r^2, \frac{d}{2}\log\frac{2.8(r^2+\log\frac{n}{\varepsilon\lambda}+d)}{d} + \log\frac{n}{\varepsilon\lambda}\right\}$. *There exists an algorithm that can output a feature matrix* $\boldsymbol{Z}\in\mathbb{R}^{m\times n}$ *with* $m = \frac{25q^2}{3\varepsilon^2}\binom{q+d-1}{q}\log\left(\frac{16s_\lambda}{\delta}\right)$, *such that with probability at least* $1-\delta$, $\boldsymbol{Z}^\top\boldsymbol{Z}$ *is an* (ε,λ)*-spectral approximation to* $\boldsymbol{K}$ *as per Eq. (1). Furthermore,* $\boldsymbol{Z}$ *can be computed in time* $\mathcal{O}((m/q)\cdot\mathrm{nnz}(\boldsymbol{X}))$.

The proof of Theorem 12 is provided in Appendix I. We remark that for any constant $\varepsilon = \Theta(1)$, dimension $d = o\left(\log\frac{n}{\lambda}\right)$ and radius $r = \mathcal{O}\left(\sqrt{\log\frac{n}{\lambda}}\right)$ our number of random features for spectrally approximating the Gaussian kernel matrix is sub-polynomial in n/λ. More precisely,

$$m = \mathcal{O}\left(\frac{\left(\frac{3d}{2}+\log\frac{n}{\lambda}\right)^d + (3.7r^2+d)^d}{(d-1)!}\right)$$
$$= \mathcal{O}\left(\frac{\left(2\log\frac{n}{\lambda}\right)^d + (1.93r)^{2d}}{(d-1)!}\right) = (n/\lambda)^{o(1)}.$$

This result improves upon prior works in a number of interesting ways. First, note that the only prior random features that can spectrally approximate the Gaussian kernel and is independent of the maximum norm of the input dataset is the random Fourier features (Rahimi & Recht, 2009). Indeed, Avron et al. (2017b) showed that spectral approximation can be achieved using random Fourier features. However, they also proved that the number of Fourier features should be at least $\Omega(n/\lambda)$, which is significantly larger than our number of features for any $d = o\left(\log\frac{n}{\lambda}\right)$.

All other prior results on spectral approximation of the Gaussian kernel with features dimension that scales sub-linearly in n/λ, bear a dependence on the radius of the dataset, like our method. The modified Fourier features (Avron et al., 2017b) assumes that the ℓ_∞-norm of all data points are bounded by some $r > 0$ and constructs random features that spectrally approximate the Gaussian kernel matrix using

$$\mathcal{O}\left(\frac{(248r)^d\cdot(\log(n/\lambda))^{d/2} + (200\log(n/\lambda))^{2d}}{\Gamma(d/2+1)}\right)$$

features. This is strictly larger than our number of features, by a large margin, for any radius $r = \mathcal{O}\left(\sqrt{\log\frac{n}{\lambda}}\right)$ and any dimension d.

Additionally, there has been a line of work based on approximating the Gaussian kernel by low degree polynomials through Taylor expansion and then sketching the resulting polynomial. Ahle et al. (2020) proposed a sketching method that runs in time $\mathcal{O}\left(r^{12}\cdot(s_\lambda\cdot n + \mathrm{nnz}(\boldsymbol{X}))\cdot\mathrm{poly}(\log(n/\lambda))\right)$. Additionally, Woodruff & Zandieh (2020) improved the result of Ahle et al. (2020) for high dimensional sparse datasets by combining sketching with adaptive sampling techniques. Their result runs in time $\mathcal{O}\left(r^{15}\cdot s_\lambda^2\cdot n + r^5\cdot\mathrm{nnz}(\boldsymbol{X})\cdot\mathrm{poly}(\log(n/\lambda))\right)$. Because of the large exponent of the radius r, both of these bounds can easily become worse than our result for datasets with large radius in small constant dimensions $d = \mathcal{O}(1)$. Table 1 summarizes our result and all prior methods for approximating Gaussian kernel.

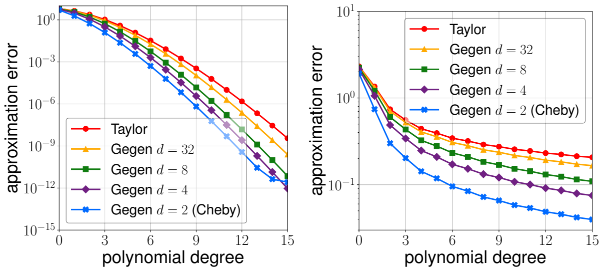
6. Experiments

6.1. Function Approximation via Gegenbauer Series

We first study function approximation of the Gegenbauer series for Gaussian and Neural Tangent Kernel of two-layer ReLU networks. They correspond to function $\kappa(x) = \exp(2x)$ and $a_1(a_1(x)) + (a_1(x) + xa_0(x))\cdot a_0(a_1(x))$ for $x\in[-1,1]$ where $a_0(x) := 1 - \frac{\mathrm{acos}(x)}{\pi}$ and $a_1(x) := \frac{\sqrt{1-x^2}+x(\pi-\mathrm{acos}(x))}{\pi}$. We approximate these functions by Taylor, Chebyshev and Gegenbauer series with degree up to 15 and compute approximation errors by $\max_{x\in[-1,1]}|\kappa(x) - \widetilde{\kappa}(x)|$ where $\widetilde{\kappa}$ is the polynomial ap-

Table 1. Comparison of Gaussian kernel approximation algorithms in terms of feature dimension and runtime for (ε, λ)-spectral guarantee. The norm of dataset is bounded by r. We omit $(\log n)^{\mathcal{O}(1)}$ dependency for clarity and consider constant ε. We assume that $\max_i \|x_i\| \le r$.

Algorithm	Feature Dimension (m)	Runtime
Fourier (Rahimi & Recht, 2009)	$\frac{n}{\lambda}$	$m \cdot \mathrm{nnz}(\boldsymbol{X})$
Modified Fourier (Avron et al., 2017b)	$(248r)^d (\log \frac{n}{\lambda})^{\frac{d}{2}} + (200 \log \frac{n}{\lambda})^{2d}$	$m \cdot \mathrm{nnz}(\boldsymbol{X})$
Nyström (Musco & Musco, 2017)	s_λ	$nm^2 + m \cdot \mathrm{nnz}(\boldsymbol{X})$
PolySketch (Ahle et al., 2020)	$r^{10} \cdot s_\lambda$	$r^{12}(ns_\lambda + \mathrm{nnz}(\boldsymbol{X}))$
Adaptive Sketch (Woodruff & Zandieh, 2020)	s_λ	$r^{15} s_\lambda^2 n + r^5 \mathrm{nnz}(\boldsymbol{X})$
Gegenbauer (This work)	$\frac{\left(2 \log \frac{n}{\lambda}\right)^d + (1.93r)^{2d}}{(d-1)!}$	$m \cdot \mathrm{nnz}(\boldsymbol{X})$



(a) Gaussian

(b) Neural Tangent

Figure 1. Kernel function approximation error of Taylor expansion and Gegenbauer expansion with $d \in \{2, 4, 8, 32\}$. The case of $d = 2$ is equivalent to the Chebyshev series expansion.

proximation. For the Gegenbauer, the dimension d varies in $\{2, 4, 8, 32\}$. Note that Taylor and Chebyshev are equivalent to Gegenbauer with $d = \infty$ and 2, respectively. Fig. 1 shows that Gegenbauer series with a proper choice of d provide better function approximators than the Taylor expansion. This can lead to performance improvement of the proposed random features, beyond Taylor series based kernel approximations, e.g., random Maclaurin (Kar & Karnick, 2012) and polynomial sketch (Ahle et al., 2020).

6.2. Kernel Ridge Regression

Next we approximate kernel ridge regression on problems from 4 real-world datasets, e.g., Earth Elevation, CO_2 , Climate and Protein. We consider the kernel ridge regression for predicting the outputs (e.g., earth elevation) with the Gaussian kernel. More details and additional results with the neural tangent kernel (NTK) can be found in Appendix J.1.

We also benchmark various Gaussian kernel approximations including Nyström (Musco & Musco, 2017), Random Fourier Features (Rahimi & Recht, 2009) and that equipped with Hadamard transform (known as FastFood) (Le et al., 2013), Random Maclaurin Features (Kar & Karnick, 2012) and PolySketch (Ahle et al., 2020). We choose the feature dimension $m = 1{,}024$ for all methods and datasets. Table 4 summarizes the results. We observe that our proposed features (Gegenbauer) achieves the best both for CO_2 and climate datasets, and the second best for elevation. But, for Protein dataset whose dimension is larger than others, we verify that others show better performance. This follows from Theorem 12 our methods requires large number of features when d is large. Although the Nyström method also performs well in practice, its runtime becomes much slower than ours.

Table 2. Results of kernel ridge regression with Gaussian kernel.

	Elevation		CO_2		Climate		Protein	
n	64,800		146,040		223,656		45,730	
Domain	$\mathbb{S}^2$		$[\mathbb{S}^2, \mathbb{R}]$		$[\mathbb{S}^2, \mathbb{R}]$		$\mathbb{R}^9$	
Metric	MSE	Time	MSE	Time	MSE	Time	MSE	Time
Nystrom	**1.14**	3.81	0.533	8.17	3.14	12.0	**18.9**	2.85
Fourier	1.30	2.10	0.548	4.73	3.15	6.93	19.8	1.66
FastFood	1.35	7.79	0.551	17.3	3.16	26.3	19.8	4.94
Maclaurin	1.90	1.07	0.593	2.38	3.18	3.55	25.9	1.05
PolySketch	1.56	7.65	0.590	16.4	3.15	23.5	26.9	4.96
Gegenbauer	1.15	1.71	**0.532**	3.49	**3.13**	5.41	21.0	9.72

Table 3. k-means clustering objective with the Gaussian kernel.

	Abalone	Pendigits	Mushroom	Magic	Statlog	Connect-4
n	4,177	7,494	8,124	19,020	43,500	67,557
d	8	16	21	10	9	42
Nyström	0.38	0.42	0.71	0.64	0.23	**0.61**
Fourier	0.38	0.43	0.72	0.66	0.24	0.81
FastFood	0.43	0.46	0.74	0.67	0.24	0.83
Maclaurin	0.43	0.46	0.72	0.73	0.23	0.90
PolySketch	**0.35**	0.45	**0.67**	0.66	**0.21**	0.82
Gegenbauer	**0.35**	**0.40**	0.71	**0.59**	**0.21**	0.78

6.3. Kernel k-means Clustering

We apply the proposed random features to kernel k-means clustering under 6 UCI classification datasets. We choose the Gaussian kernel and explore various approximating algorithms as described above where feature dimension is set to $m = 512$. We evaluate the average summation of squared distance to the nearest cluster centers. Formally, given data points $x_1, \dots, x_n$, let ϕ_i be some feature map of x_i and denote $\mu_i = \frac{1}{|C_i|} \sum_{x_j \in C_i} \phi_{x_i}$ be the centroid of the vectors in C_i after mapping to kernel space. The goal of kernel k-means is to choose partitions $\{C_1, \dots, C_k\}$ which minimize the following objective: $\sum_{i=1}^{k}\sum_{x_j \in C_i} \|\phi_{x_j} - \mu_i\|_2^2$. Table 3 reports the result of k-means clustering. We observe that our random Gegenbauer features shows promising performances except Mushroom and Connect-4 datasets, which have a higher input dimension. More details are in Appendix J.2.

7. Conclusion

We studied a new class of kernels expressed by Gegenbauer polynomials that covers a wide range of ubiquitous kernels. The proposed random features can spectrally approximate kernel matrices, making it useful for scalable kernel methods. One limitation is that it can tightly approximate when the inputs are in a low-dimensional space. We believe this can be solved when it combines with additional dimensionality reductions (e.g., JL-transform) and open the question for application of high-dimensional inputs for future work.

Acknowledgements

Haim Avron was partially supported by the Israel Science Foundation (grant no. 1272/17) and by the US-Israel Binational Science Foundation (grant no. 2017698). Amir Zandieh was supported by the Swiss NSF grant No. P2ELP2_195140. Insu Han was supported by TATA DATA Analysis (grant no. 105676).

References

Ahle, T. D., Kapralov, M., Knudsen, J. B., Pagh, R., Velingker, A., Woodruff, D. P., and Zandieh, A. Oblivious sketching of high-degree polynomial kernels. In *Symposium on Discrete Algorithms (SODA)*, 2020.

Alaoui, A. E. and Mahoney, M. W. Fast randomized kernel methods with statistical guarantees. In *Neural Information Processing Systems (NeurIPS)*, 2015.

Arthur, D. and Vassilvitskii, S. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006.

Atkinson, K. and Han, W. *Spherical harmonics and approximations on the unit sphere: an introduction*. Springer Science & Business Media, 2012.

Avron, H., Nguyen, H., and Woodruff, D. Subspace embeddings for the polynomial kernel. In *Neural Information Processing Systems (NeurIPS)*, 2014.

Avron, H., Clarkson, K. L., and Woodruff, D. P. Faster Kernel Ridge Regression Using Sketching and Preconditioning. In *SIAM Journal on Matrix Analysis and Applications (SIMAX)*, 2017a.

Avron, H., Kapralov, M., Musco, C., Musco, C., Velingker, A., and Zandieh, A. Random Fourier features for kernel ridge regression: Approximation bounds and statistical guarantees. In *International Conference on Machine Learning (ICML)*, 2017b.

Avron, H., Kapralov, M., Musco, C., Musco, C., Velingker, A., and Zandieh, A. A universal sampling method for reconstructing signals with simple fourier transforms. In *Symposium on the Theory of Computing (STOC)*, 2019.

Bach, F. Sharp analysis of low-rank kernel matrix approximations. In *Conference on Learning Theory (COLT)*, 2013.

Bercovici, D. The generation of plate tectonics from mantle convection. *Earth and Planetary Science Letters*, 2003.

Cho, Y. and Saul, L. Kernel methods for deep learning. In *Neural Information Processing Systems (NeurIPS)*, 2009.

Cho, Y. and Saul, L. K. Large-margin classification in infinite neural networks. *Neural computation*, 2010.

Cohen, M. B., Musco, C., and Musco, C. Input sparsity time low-rank approximation via ridge leverage score sampling. In *Symposium on Discrete Algorithms (SODA)*, 2017.

Dhillon, I. S., Guan, Y., and Kulis, B. Kernel k-means: spectral clustering and normalized cuts. In *Conference on Knowledge Discovery and Data Mining (KDD)*, 2004.

Fox, L. and Parker, I. B. Chebyshev polynomials in numerical analysis. Technical report, 1968.

Gautschi, W. *Orthogonal polynomials: computation and approximation*. OUP Oxford, 2004.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Neural Information Processing Systems (NeurIPS)*, 2018.

Kar, P. and Karnick, H. Random feature maps for dot product kernels. In *Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.

Le, Q., Sarlós, T., Smola, A., et al. Fastfood-approximating kernel expansions in loglinear time. In *International Conference on Machine Learning (ICML)*, 2013.

Lee, J., Xiao, L., Schoenholz, S., Bahri, Y., Novak, R., Sohl-Dickstein, J., and Pennington, J. Wide neural networks of any depth evolve as linear models under gradient descent. In *Neural Information Processing Systems (NeurIPS)*, 2019.

Li, C., Jegelka, S., and Sra, S. Fast dpp sampling for nystrom with application to kernel methods. In *International Conference on Machine Learning (ICML)*, 2016.

Li, M., Miller, G. L., and Peng, R. Iterative row sampling. In *Foundations of Computer Science (FOCS)*, 2013.

Mason, J. C. and Handscomb, D. C. *Chebyshev polynomials*. CRC press, 2002.

Minh, H. Q., Niyogi, P., and Yao, Y. Mercer's theorem, feature maps, and smoothing. In *Conference on Learning Theory (COLT)*, 2006.

Morimoto, M. *Analytic functionals on the sphere*. American Mathematical Soc., 1998.

Musco, C. and Musco, C. Recursive Sampling for the Nyström Method. In *Neural Information Processing Systems (NeurIPS)*, 2017.

Ogawa, H. An operator pseudo-inversion lemma. *SIAM Journal on Applied Mathematics*, 1988.

Paul, S. and Drineas, P. Feature selection for ridge regression with provable guarantees. *Neural computation*, 2016.

Pennington, J., Yu, F. X. X., and Kumar, S. Spherical random features for polynomial kernels. In *Neural Information Processing Systems (NeurIPS)*, 2015.

Pham, N. and Pagh, R. Fast and scalable polynomial kernels via explicit feature maps. In *Conference on Knowledge Discovery and Data Mining (KDD)*, 2013.

Rahimi, A. and Recht, B. Random Features for Large-Scale Kernel Machines. In *Neural Information Processing Systems (NeurIPS)*, 2009.

Rasmussen, C. E. Gaussian processes in machine learning. In *Advanced lectures on machine learning*. Springer, 2004.

Sanderson, B. M., Shell, K. M., and Ingram, W. Climate feedbacks determined using radiative kernels in a multi-thousand member ensemble of AOGCMs. *Climate dynamics*, 2010.

Saunders, C., Gammerman, A., and Vovk, V. Ridge regression learning algorithm in dual variables. In *International Conference on Machine Learning (ICML)*, 1998.

Scetbon, M. and Harchaoui, Z. A Spectral Analysis of Dot-product Kernels. In *Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

Schoenberg, I. Positive definite functions on spheres. *Duke Math. J*, 1942.

Smola, A. J., Ovari, Z. L., Williamson, R. C., et al. Regularization with dot-product kernels. *Neural Information Processing Systems (NeurIPS)*, 2001.

Song, Z., Woodruff, D., Yu, Z., and Zhang, L. Fast sketching of polynomial kernels of polynomial degree. In *International Conference on Machine Learning (ICML)*, 2021.

Su, X., An, J., Zhang, Y., Zhu, P., and Zhu, B. Prediction of ozone hourly concentrations by support vector machine and kernel extreme learning machine using wavelet transformation and partial least squares methods. *Atmospheric Pollution Research*, 2020.

Tropp, J. A. An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571*, 2015.

Vishwanathan, S. V. N., Schraudolph, N. N., Kondor, R., and Borgwardt, K. M. Graph kernels. *Journal of Machine Learning Research (JMLR)*, 2010.

Williams, C. and Seeger, M. Using the Nyström method to speed up kernel machines. In *Neural Information Processing Systems (NeurIPS)*, 2001.

Woodruff, D. P. and Zandieh, A. Near Input Sparsity Time Kernel Embeddings via Adaptive Sampling. In *International Conference on Machine Learning (ICML)*, 2020.

Yang, T., Li, Y.-F., Mahdavi, M., Jin, R., and Zhou, Z.-H. Nyström method vs random fourier features: A theoretical and empirical comparison. In *Neural Information Processing Systems (NeurIPS)*, 2012.

Zandieh, A., Nouri, N., Velingker, A., Kapralov, M., and Razenshteyn, I. Scaling up kernel ridge regression via locality sensitive hashing. In *Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

Zandieh, A., Han, I., Avron, H., Shoham, N., Kim, C., and Shin, J. Scaling Neural Tangent Kernels via Sketching and Random Features. In *Neural Information Processing Systems (NeurIPS)*, 2021.

A. Applications to Learning Tasks

In this section, we prove that our general kernel approximation guarantees from Theorem 9 and Theorem 10 are sufficient for many downstream learning tasks without sacrificing accuracy or statistical performance of our random features.

A.1. Kernel Ridge Regression

One way to analyze the quality of approximate kernel ridge regression (KRR) estimator is by bounding the excess risk compared to the exact KRR estimator. We consider a fixed design setting which has been particularly popular in analysis of KRR (Bach, 2013; Alaoui & Mahoney, 2015; Li et al., 2016; Paul & Drineas, 2016; Musco & Musco, 2017; Avron et al., 2017b; Zandieh et al., 2020). In this setting, we assume that our observed labels y_i represent some underlying true labels $f^*(x_i)$ perturbed with Gaussian noise with variance σ^2. More specifically, we assume y_i satisfies

$$y_i = f^*(x_i) + \nu_i$$

for some $f^* : \mathbb{R}^d \to \mathbb{R}$. Then, the empirical risk of an estimator f is defined as

$$\mathcal{R}(f) \coloneqq \mathbb{E}_{\{v_i\}_{i=1}^n} \left[\frac{1}{n} \sum_{i=1}^{n} |f(x_i) - f^*(x_i)|^2 \right] \tag{24}$$

Given this definition of risk, our Theorem 9 along with (Avron et al., 2017b, Lemma 2) immediately gives the following bound on the risk of approximate KRR using our feature matrix $\mathbf{\Phi}$,

Lemma 13 (Kernel ridge regression risk bound)**.** *Given that preconditions of Theorem 9 hold, let f be the exact KRR estimator using kernel $\boldsymbol{K} + \lambda \boldsymbol{I}$ and $\tilde{f}$ be the approximate estimator obtained using the approximate kernel $\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I}$. If $\|\boldsymbol{K}\|_{\text{op}} \geq 1$ and $\boldsymbol{Z}^\top \boldsymbol{Z}$ is an (ε, λ)-spectral approximation to $\boldsymbol{K}$ for some $0 \leq \varepsilon < 1$ as per (1) then*

$$\mathcal{R}(\tilde{f}) \leq \frac{\mathcal{R}(f)}{1-\varepsilon} + \frac{\varepsilon}{1+\varepsilon} \cdot \frac{\text{rank}(\boldsymbol{Z})}{n} \cdot \sigma^2.$$

A.2. Kernel k-means Clustering

Kernel k-means clustering aims at partitioning the data-points $x_1, \cdots, x_n \in \mathbb{R}^d$, into k cluster sets, $\{C_1, \ldots, C_k\}$ such that the sum of squares of *kernel distances* of data-points from their associated cluster center is minimized. Specifically, for our generalized zonal kernel function (Definition 3), if we let $\mu_i = \frac{1}{|C_i|} \sum_{x_j \in C_i} \phi_{x_i}$ be the centroid of the vectors in C_i after mapping to kernel space using the feature map ϕ_x defined in Lemma 5, then the goal of kernel k-means is to choose partitions $\{C_1, \ldots, C_k\}$ which minimize the following objective:

$$\sum_{i=1}^{k} \sum_{x_j \in C_i} \|\phi_{x_j} - \mu_i\|^2_{L^2(S^{d-1}, \mathbb{R}^s)}.$$

This optimization problem can be rewritten as a constrained low-rank approximation problem (Musco & Musco, 2017). In particular, for any clustering $\{C_1, \ldots, C_k\}$ we can define a rank-k orthonormal matrix $\boldsymbol{C} \in \mathbb{R}^{n \times k}$, called the cluster indicator matrix, as $\boldsymbol{C}_{j,i} \coloneqq \frac{1}{|C_i|} \cdot \mathbb{1}_{\{x_j \in C_i\}}$ for every $i \in [k]$ and $j \in [n]$. Note that with this definition we have $\boldsymbol{C}^\top \boldsymbol{C} = \boldsymbol{I}_k$, so $\boldsymbol{C}\boldsymbol{C}^\top$ is a rank k projection matrix. Therefore, if we let $\boldsymbol{K} \in \mathbb{R}^{n \times n}$ be the GZK kernel matrix, the kernel k-means cost function is equivalent to

$$\sum_{i=1}^{k} \sum_{x_j \in C_i} \|\phi_{x_j} - \mu_i\|^2_{L^2(S^{d-1}, \mathbb{R}^s)} \coloneqq \text{Tr}\left(\boldsymbol{K} - \boldsymbol{C}\boldsymbol{C}^\top \boldsymbol{K} \boldsymbol{C}\boldsymbol{C}^\top\right).$$

Thus we can approximately solve this problem by using our random features $\boldsymbol{Z}$ constructed in Definition 8 and solving the following problem:

$$\min_{\text{cluster indicator } \boldsymbol{C}} \|\boldsymbol{Z} - \boldsymbol{Z}\boldsymbol{C}\boldsymbol{C}^\top\|_F^2.$$

Specifically, using our Theorem 10 along with (Musco & Musco, 2017, Theorem 16) we have the following approximation bound,

Lemma 14. *Given that preconditions of Theorem 10 hold, if we let $\widetilde{\boldsymbol{C}} \in \mathbb{R}^{n\times k}$ be an approximately optimal cluster indicator matrix for the following k-means problem,*

$$\|\boldsymbol{Z} - \boldsymbol{Z}\widetilde{\boldsymbol{C}}\widetilde{\boldsymbol{C}}^\top\|_F^2 \le (1+\gamma) \min_{\textit{cluster indicator } \boldsymbol{C}} \|\boldsymbol{Z} - \boldsymbol{Z}\boldsymbol{C}\boldsymbol{C}^\top\|_F^2,$$

for some $\gamma \ge 0$, then we have the following,

$$\|\boldsymbol{Z} - \boldsymbol{Z}\widetilde{\boldsymbol{C}}\widetilde{\boldsymbol{C}}^\top\|_F^2 \le (1+\gamma)(1+\varepsilon) \min_{\textit{cluster indicator } \boldsymbol{C}} \operatorname{Tr}\left(\boldsymbol{K} - \boldsymbol{C}\boldsymbol{C}^\top\boldsymbol{K}\boldsymbol{C}\boldsymbol{C}^\top\right).$$

B. Class of GZKs Contains All Dot-product Kernels

In this section we prove Lemma 4, which implies that the class of GZK given in Definition 3 includes all dot-product kernels.

Lemma 4 (Dot-product kernels are GZKs)**.** *For any $x, y \in \mathbb{R}^d$, any integer $d \ge 3$, and any dot-product kernel $k(x,y) = \kappa(\langle x, y\rangle)$ with analytic $\kappa(\cdot)$, the eigenfunction expansion of $k(x,y)$ can be written as,*

$$k(x,y) := \sum_{\ell=0}^{\infty} \left(\sum_{i=0}^{\infty} \widetilde{h}_{\ell,i}(\|x\|)\widetilde{h}_{\ell,i}(\|y\|) \right) P_d^\ell\left(\frac{\langle x, y\rangle}{\|x\|\|y\|} \right),$$

where $\widetilde{h}_{\ell,i}(\cdot)$ are real-valued monomials defined as follows for integers $\ell, i \ge 0$ and any $t \in \mathbb{R}$:

$$\widetilde{h}_{\ell,i}(t) := \sqrt{\frac{\alpha_{\ell,d}}{2^\ell} \frac{\Gamma(\frac{d}{2})\ \kappa^{(\ell+2i)}(0)}{\sqrt{\pi}(2i)!} \frac{\Gamma(i+\frac{1}{2})}{\Gamma(i+\ell+\frac{d}{2})}} \cdot t^{\ell+2i}. \tag{12}$$

Proof of Lemma 4. We begin with the Taylor series expansion of the function $\kappa(\cdot)$ around zero. Because $\kappa(\cdot)$ is analytic, the series expansion exists and converges to κ. So we have,

$$\kappa(\langle x,y\rangle) := \sum_{j=0}^{\infty} \frac{\kappa^{(j)}(0)}{j!} \cdot \langle x,y\rangle^j = \sum_{j=0}^{\infty} \frac{\kappa^{(j)}(0)}{j!} \cdot \|x\|^j \cdot \|y\|^j \cdot \left(\frac{\langle x,y\rangle}{\|x\|\cdot\|y\|} \right)^j. \tag{25}$$

Now we write the degree-j monomial t^j for any integer $j \ge 0$, in the basis of d-dimensional Gegenbauer polynomials, $P_d^0(t), P_d^1(t), P_d^2(t), \dots P_d^j(t)$. More precisely, by Eq. (7), we find $t^j := \sum_{\ell=0}^{j} \mu_\ell^j \cdot P_d^\ell(t)$ where

$$\mu_\ell^j = \alpha_{\ell,d} \cdot \frac{|\mathbb{S}^{d-2}|}{|\mathbb{S}^{d-1}|} \int_{-1}^{1} t^j \cdot P_d^\ell(t) \cdot (1-t^2)^{\frac{d-3}{2}} dt. \tag{26}$$

By using the Rodrigues' formula in Eq. (5), we can compute the Gegenbauer coefficients of t^j as follows,

$$\mu_\ell^j = \alpha_{\ell,d} \cdot \frac{(-1)^\ell}{2^\ell} \cdot \frac{|\mathbb{S}^{d-2}|}{|\mathbb{S}^{d-1}|} \cdot \frac{\Gamma\left(\frac{d-1}{2}\right)}{\Gamma\left(\ell + \frac{d-1}{2}\right)} \int_{-1}^{1} t^j \cdot \frac{d^\ell}{dt^\ell}\left(1-t^2\right)^{\ell+\frac{d-3}{2}} dt. \tag{27}$$

By multiple applications of integration by parts we can compute the integral in Eq. (27) as follows,

$$\begin{aligned}
\int_{-1}^{1} t^j \cdot \frac{d^\ell\left(1-t^2\right)^{\ell+\frac{d-3}{2}}}{dt^\ell} dt &= t^j \cdot \frac{d^{\ell-1}(1-t^2)^{\ell+\frac{d-3}{2}}}{dt^{\ell-1}} \Bigg|_{-1}^{1} - j\int_{-1}^{1} t^{j-1} \cdot \frac{d^{\ell-1}(1-t^2)^{\ell+\frac{d-3}{2}}}{dt^{\ell-1}} dt \\
&= -j\int_{-1}^{1} t^{j-1} \cdot \frac{d^{\ell-1}(1-t^2)^{\ell+\frac{d-3}{2}}}{dt^{\ell-1}} dt \\
&= -jt^{j-1} \cdot \frac{d^{\ell-2}(1-t^2)^{\ell+\frac{d-3}{2}}}{dt^{\ell-2}} \Bigg|_{-1}^{1} + j(j-1)\int_{-1}^{1} t^{j-2} \cdot \frac{d^{\ell-2}(1-t^2)^{\ell+\frac{d-3}{2}}}{dt^{\ell-2}} dt \\
&= (-1)^2 \cdot j(j-1)\int_{-1}^{1} t^{j-2} \cdot \frac{d^{\ell-2}(1-t^2)^{\ell+\frac{d-3}{2}}}{dt^{\ell-2}} dt \\
&\ \ \vdots \\
&= (-1)^\ell \cdot \frac{j!}{(j-\ell)!} \int_{-1}^{1} t^{j-\ell} \cdot (1-t^2)^{\ell+\frac{d-3}{2}} dt.
\end{aligned} \tag{28}$$

Now note that the above integral is zero if $j-\ell$ is an odd integer. So, we focus on the cases where $j-\ell$ is an even integer. By a change of variables to $u=t^2$ we have,

$$\int_{-1}^{1} t^{j-\ell}\cdot(1-t^2)^{\ell+\frac{d-3}{2}}dt=\int_{0}^{1} u^{\frac{j-\ell-1}{2}}\cdot(1-u)^{\ell+\frac{d-3}{2}}du=\frac{\Gamma(\frac{j-\ell+1}{2})\cdot\Gamma(\ell+\frac{d-1}{2})}{\Gamma(\frac{j+\ell+d}{2})}.$$

By combining the above with Eq. (28) and Eq. (27) and using the fact that $\frac{|\mathbb{S}^{d-2}|}{|\mathbb{S}^{d-1}|}=\frac{\Gamma(\frac{d}{2})}{\sqrt{\pi}\cdot\Gamma(\frac{d-1}{2})}$, we find the following

$$\mu_\ell^j=\begin{cases}\frac{\alpha_{\ell,d}}{2^\ell}\cdot\frac{\Gamma(\frac{d}{2})\cdot j!}{\sqrt{\pi}\cdot(j-\ell)!}\cdot\frac{\Gamma(\frac{j-\ell+1}{2})}{\Gamma(\frac{j+\ell+d}{2})} & \text{if } j-\ell \text{ is even}\\ 0 & \text{if } j-\ell \text{ is odd}\end{cases} \tag{29}$$

Now if we plug the monomial expansion $t^j:=\sum_{\ell=0}^{j}\mu_\ell^j\cdot P_d^\ell(t)$ into Eq. (25), using the fact that $\mu_\ell^j=0$ for any odd $j-\ell$, we find that

$$\begin{aligned}\kappa(\langle x,y\rangle)&=\sum_{j=0}^{\infty}\frac{\kappa^{(j)}(0)}{j!}\cdot\|x\|^j\cdot\|y\|^j\cdot\sum_{\ell=0}^{j}\mu_\ell^j\cdot P_d^\ell\left(\frac{\langle x,y\rangle}{\|x\|\cdot\|y\|}\right)\\&=\sum_{\ell=0}^{\infty}\left(\sum_{j=\ell}^{\infty}\mu_\ell^j\cdot\frac{\kappa^{(j)}(0)}{j!}\cdot\|x\|^j\cdot\|y\|^j\right)\cdot P_d^\ell\left(\frac{\langle x,y\rangle}{\|x\|\cdot\|y\|}\right)\\&=\sum_{\ell=0}^{\infty}\left(\sum_{i=0}^{\infty}\mu_\ell^{\ell+2i}\cdot\frac{\kappa^{(\ell+2i)}(0)}{(\ell+2i)!}\cdot\|x\|^{\ell+2i}\cdot\|y\|^{\ell+2i}\right)\cdot P_d^\ell\left(\frac{\langle x,y\rangle}{\|x\|\cdot\|y\|}\right)\\&=\sum_{\ell=0}^{\infty}\left(\sum_{i=0}^{\infty}\widetilde{h}_{\ell,i}(\|x\|)\cdot\widetilde{h}_{\ell,i}(\|y\|)\right)\cdot P_d^\ell\left(\frac{\langle x,y\rangle}{\|x\|\cdot\|y\|}\right),\end{aligned}$$

where the functions $\widetilde{h}_{\ell,i}(\cdot)$ are defined as

$$\widetilde{h}_{\ell,i}(t):=\sqrt{\mu_\ell^{\ell+2i}\cdot\frac{\kappa^{(\ell+2i)}(0)}{(\ell+2i)!}}\cdot t^{\ell+2i}.$$

Note that since $\kappa(\cdot)$ is a valid positive semi-definite kernel function, it's derivatives $\kappa^{(\ell+2i)}(0)$ are all non-negative (Schoenberg, 1942), thus the above function is real-valued. Now by Eq. (29), the function $h_{i,\ell}(t)$ defined above satisfies

$$\widetilde{h}_{\ell,i}(t)=\sqrt{\frac{\alpha_{\ell,d}}{2^\ell}\cdot\frac{\Gamma(\frac{d}{2})\cdot\kappa^{(\ell+2i)}(0)}{\sqrt{\pi}\cdot(2i)!}\cdot\frac{\Gamma(i+\frac{1}{2})}{\Gamma(i+\ell+\frac{d}{2})}}\cdot t^{\ell+2i}.$$

This completes the proof of Lemma 4. □

C. Gaussian and Neural Tangent Kernels are GZK

In this section we show that the Gaussian and Neural Tangent Kernels are contained in the class of GZKs.

Lemma 15 (Gaussian kernel is a GZK). *For any $x,y\in\mathbb{R}^d$, any integer $d\geq 3$, the eigenfunction expansion of the Gaussian kernel can be written as,*

$$e^{-\|x-y\|_2^2/2}:=\sum_{\ell=0}^{\infty}\left(\sum_{i=0}^{\infty}\widetilde{h}_{\ell,i}(\|x\|)\widetilde{h}_{\ell,i}(\|y\|)\right)P_d^\ell\left(\frac{\langle x,y\rangle}{\|x\|\|y\|}\right),$$

where $\widetilde{h}_{\ell,i}(\cdot)$ are real-valued monomials defined as follows for integers $\ell,i\geq 0$ and any $t\in\mathbb{R}$:

$$\widetilde{h}_{\ell,i}(t):=\sqrt{\frac{\alpha_{\ell,d}}{2^\ell}\frac{\Gamma(\frac{d}{2})}{\sqrt{\pi}(2i)!}\frac{\Gamma(i+\frac{1}{2})}{\Gamma(i+\ell+\frac{d}{2})}}\cdot t^{\ell+2i}\cdot e^{-t^2/2}.$$

Proof of Lemma 15. First note that for the Gaussian kernel we can write, $k(x,y) = e^{-\|x-y\|^2/2} = e^{-\|x\|^2/2}e^{-\|y\|^2/2}e^{\langle x,y\rangle}$. Applying Lemma 4 to the exponential kernel function $e^{\langle x,y\rangle}$, we have

$$e^{\langle x,y\rangle} := \sum_{\ell=0}^{\infty}\left(\sum_{i=0}^{\infty}\widetilde{h}_{\ell,i}^{\exp}(\|x\|)\widetilde{h}_{\ell,i}^{\exp}(\|y\|)\right)P_d^{\ell}\left(\frac{\langle x,y\rangle}{\|x\|\|y\|}\right), \tag{30}$$

where

$$\widetilde{h}_{\ell,i}^{\exp}(t) = \sqrt{\frac{\alpha_{\ell,d}}{2^{\ell}}\cdot\frac{\Gamma(\frac{d}{2})}{\sqrt{\pi}(2i)!}\cdot\frac{\Gamma(i+\frac{1}{2})}{\Gamma(i+\ell+\frac{d}{2})}}\cdot t^{\ell+2i}. \tag{31}$$

The reason for the above is because all derivatives of the exponential function are equal to 1 at the origin. So, using the above we have,

$$e^{-\frac{\|x-y\|_2^2}{2}} = \sum_{\ell=0}^{\infty}\left(\sum_{i=0}^{\infty}e^{-\frac{\|x\|^2}{2}}\widetilde{h}_{\ell,i}^{\exp}(\|x\|)\cdot e^{-\frac{\|y\|^2}{2}}\widetilde{h}_{\ell,i}^{\exp}(\|y\|)\right)\cdot P_d^{\ell}\left(\frac{\langle x,y\rangle}{\|x\|\cdot\|y\|}\right). \tag{32}$$

This shows that the Gaussian kernel can be represented in the form of

$$e^{-\frac{\|x-y\|_2^2}{2}} = \sum_{\ell=0}^{\infty}\left(\sum_{i=0}^{\infty}\widetilde{h}_{\ell,i}(\|x\|)\widetilde{h}_{\ell,i}(\|y\|)\right)P_d^{\ell}\left(\frac{\langle x,y\rangle}{\|x\|\|y\|}\right),$$

with $\widetilde{h}_{\ell,i}(t) = e^{-t^2/2}\cdot\widetilde{h}_{\ell,i}^{\exp}(t)$. □

Lemma 15 shows that the Gaussian kernel is a GZK as per Definition 3 with

$$h_{\ell}(t) = \left[\sqrt{\frac{\alpha_{\ell,d}}{2^{\ell}}\frac{\Gamma(\frac{d}{2})}{\sqrt{\pi}(2i)!}\frac{\Gamma(i+\frac{1}{2})}{\Gamma(i+\ell+\frac{d}{2})}}\cdot t^{\ell+2i}\cdot e^{-t^2/2}\right]_{i=0}^{\infty}.$$

Next, we show that the Neural Tangent Kernel (NTK) of an infinitely wide network with ReLU activation is a GZK. It was shown in (Zandieh et al., 2021, Definition 1) that the depth-L NTK with ReLU activation has the following normalized dot-product form,

$$\Theta_{\texttt{ntk}}^{(L)}(x,y) := \|x\|\|y\|\cdot K_{\texttt{relu}}^{(L)}\left(\frac{\langle x,y\rangle}{\|x\|_{\|}y\|}\right),\quad \text{for any } x,y\in\mathbb{R}^d, \tag{33}$$

where $K_{\texttt{relu}}^{(L)}:[-1,1]\to\mathbb{R}$ is some smooth univariate function that can be computed using a recursive relation. We show that this kernel is indeed a GZK.

Lemma 16 (Neural Tangent Kernel is a GZK)**.** *For any $x,y\in\mathbb{R}^d$, any integers $d\geq 3$ and $L\geq 1$, the eigenfunction expansion of the depth-L NTK defined in (Zandieh et al., 2021, Definition 1) can be written as,*

$$\Theta_{\texttt{ntk}}^{(L)}(x,y) := \sum_{\ell=0}^{\infty}\widetilde{h}_{\ell}(\|x\|)\widetilde{h}_{\ell}(\|y\|)\cdot P_d^{\ell}\left(\frac{\langle x,y\rangle}{\|x\|\|y\|}\right),$$

where $\widetilde{h}_{\ell}(\cdot)$ are linear univariate functions defined as follows for integer $\ell\geq 0$ and any $t\in\mathbb{R}$:

$$\widetilde{h}_{\ell}(t) := \sqrt{\alpha_{\ell,d}\cdot\frac{|\mathbb{S}^{d-2}|}{|\mathbb{S}^{d-1}|}\cdot\int_{-1}^{1}K_{\texttt{relu}}^{(L)}(\tau)P_d^{\ell}(\tau)(1-\tau^2)^{\frac{d-3}{2}}d\tau}\cdot t,$$

where $K_{\texttt{relu}}^{(L)}:[-1,1]\to\mathbb{R}$ is the univariate function defined as per (Zandieh et al., 2021, Definition 1).

Proof of Lemma 16. We start by finding the Gegenbauer series expansion of $K_{\texttt{relu}}^{(L)}(t)$ using Eq. (8) and Eq. (7):

$$K_{\texttt{relu}}^{(L)}(t) = \sum_{\ell=0}^{\infty}c_{\ell}\cdot P_d^{\ell}(t), \tag{34}$$

where the Gegenbauer coefficients c_ℓ, can be computed as follows,

$$c_\ell = \alpha_{\ell,d} \cdot \frac{|\mathbb{S}^{d-2}|}{|\mathbb{S}^{d-1}|} \cdot \int_{-1}^{1} K_{\mathtt{relu}}^{(L)}(t) P_d^\ell(t)(1-t^2)^{\frac{d-3}{2}} dt.$$

Therefore, using Eq. (33) we have,

$$\Theta_{\mathtt{ntk}}^{(L)}(x,y) := \sum_{\ell=0}^{\infty} c_\ell \cdot \|x\|_2 \|y\|_2 \cdot P_d^\ell\left(\frac{\langle x, y\rangle}{\|x\|_2 \|y\|_2}\right).$$

Therefore the lemma follows. □

D. Mercer Decomposition of GZK

In this section we prove the lemmas about the Mercer decomposition of Zonal and Generalized Zonal kernels.

D.1. Proof of Lemma 2

Lemma 2 (Feature map for zonal kernels)**.** *Suppose $\kappa : [-1,1] \to \mathbb{R}$ is analytic and let $\{c_\ell\}_{\ell=0}^{\infty}$ be the coefficients of its Gegenbauer series expansion in dimension $d \geq 2$. For $x, w \in \mathbb{S}^{d-1}$, define the real-valued function $\phi_x \in L^2(\mathbb{S}^{d-1})$ as*

$$\phi_x(w) := \sum_{\ell=0}^{\infty} \sqrt{c_\ell \cdot \alpha_{\ell,d}} \cdot P_d^\ell(\langle x, w\rangle). \tag{9}$$

Then, for all $x, y \in \mathbb{S}^{d-1}$, it holds that

$$\mathbb{E}_{w \sim \mathcal{U}(\mathbb{S}^{d-1})}\left[\phi_x(w) \cdot \phi_y(w)\right] = \kappa(\langle x, y\rangle). \tag{10}$$

Proof of Lemma 2. We observe that

$$\begin{aligned}
\mathbb{E}_{w \sim \mathcal{U}(\mathbb{S}^{d-1})}[\phi_x(w) \cdot \phi_y(w)] &= \mathbb{E}_w\left[\sum_{\ell,\ell'=0}^{\infty} \sqrt{c_\ell c_{\ell'} \alpha_{\ell,d} \alpha_{\ell',d}} \cdot P_d^\ell(\langle x, w\rangle) \cdot P_d^{\ell'}(\langle y, w\rangle)\right] \\
&= \sum_{\ell,\ell'=0}^{\infty} \sqrt{c_\ell c_{\ell'} \alpha_{\ell,d} \alpha_{\ell',d}} \cdot \mathbb{E}_w\left[P_d^\ell(\langle x, w\rangle) \cdot P_d^{\ell'}(\langle y, w\rangle)\right] \\
&= \sum_{\ell,\ell'=0}^{\infty} \sqrt{c_\ell c_{\ell'} \alpha_{\ell,d} \alpha_{\ell',d}} \cdot \frac{P_d^\ell(\langle x, y\rangle)}{\alpha_{\ell,d}} \cdot \mathbb{1}_{\{\ell=\ell'\}} \\
&= \sum_{\ell=0}^{\infty} c_\ell P_d^\ell(\langle x, y\rangle) = \kappa(\langle x, y\rangle).
\end{aligned}$$

where the third equality comes from Lemma 1. This completes the proof of Lemma 2. □

D.2. Proof of Lemma 5

In this section we prove that Lemma 5 gives a Mercer decomposition of the GZK.

Lemma 5 (Feature map for GZK)**.** *Consider a GZK $k(\cdot,\cdot)$ with real-valued functions $h_\ell : \mathbb{R} \to \mathbb{R}^s$ for $\ell = 0, 1, \dots$ as in Definition 3. For any $x \in \mathbb{R}^d, w \in \mathbb{S}^{d-1}$, define the function $\phi_x \in L^2(\mathbb{S}^{d-1}, \mathbb{R}^s)$ as*

$$\phi_x(w) := \sum_{\ell=0}^{\infty} \sqrt{\alpha_{\ell,d}}\, h_\ell(\|x\|)\, P_d^\ell\left(\frac{\langle x, w\rangle}{\|x\|}\right). \tag{13}$$

Then, for any $x, y \in \mathbb{R}^d$, it holds that

$$\mathbb{E}_{w \sim \mathcal{U}(\mathbb{S}^{d-1})}\left[\langle \phi_x(w), \phi_y(w)\rangle\right] = k(x, y).$$

Proof of Lemma 5. By Eq. (13),

$$
\begin{aligned}
\mathbb{E}_w\left[\langle \phi_x(w), \phi_y(w)\rangle\right] &= \mathbb{E}_w\left[\langle \phi_x(w), \phi_y(w)\rangle\right]\\
&= \mathbb{E}_w\left[\left\langle \sum_{\ell=0}^{\infty} \sqrt{\alpha_{\ell,d}} h_\ell(\|x\|) P_d^\ell\left(\frac{\langle x, w\rangle}{\|x\|}\right), \sum_{\ell'=0}^{\infty} \sqrt{\alpha_{\ell',d}} h_{\ell'}(\|y\|) P_d^{\ell'}\left(\frac{\langle y, w\rangle}{\|y\|}\right)\right\rangle\right]\\
&= \sum_{\ell=0}^{\infty}\sum_{\ell'=0}^{\infty} \sqrt{\alpha_{\ell,d}\cdot\alpha_{\ell',d}} \cdot \langle h_\ell(\|x\|), h_{\ell'}(\|y\|)\rangle \cdot \mathbb{E}_w\left[P_d^\ell\left(\frac{\langle x, w\rangle}{\|x\|}\right) P_d^{\ell'}\left(\frac{\langle y, w\rangle}{\|y\|}\right)\right]\\
&= \sum_{\ell=0}^{\infty}\sum_{\ell'=0}^{\infty} \sqrt{\alpha_{\ell,d}\cdot\alpha_{\ell',d}} \cdot \langle h_\ell(\|x\|), h_{\ell'}(\|y\|)\rangle \cdot \frac{1}{\alpha_{\ell,d}} \cdot P_d^\ell\left(\frac{\langle x, y\rangle}{\|x\|\|y\|}\right) \cdot \mathbb{1}_{\{\ell=\ell'\}}\\
&= \sum_{\ell=0}^{\infty} \langle h_\ell(\|x\|), h_\ell(\|y\|)\rangle \cdot P_d^\ell\left(\frac{\langle x, y\rangle}{\|x\|\|y\|}\right),
\end{aligned}
$$

where the second last line above follows from Lemma 1. This completes the proof of Lemma 5. □

E. Leverage Scores of the GZK Feature Operator Φ

In this section we prove the uniform upper bound on ridge leverage scores of the GZK feature operator $\mathbf{\Phi}$ defined in Eq. (14) as well as some other useful properties of the leverage scores. We start by calculating the *average* of the ridge leverage scores defined in Definition 6, a.k.a. *statistical dimension* of the kernel matrix,

$$
\begin{aligned}
s_\lambda &:= \mathbb{E}_{w\sim\mathcal{U}(\mathbb{S}^{d-1})}\left[\tau_\lambda(w)\right]\\
&= \mathbb{E}_{w\sim\mathcal{U}(\mathbb{S}^{d-1})}\left[\operatorname{Tr}\left(\Phi_w^\top \cdot (\mathbf{\Phi}^*\mathbf{\Phi} + \lambda \boldsymbol{I})^{-1} \cdot \Phi_w\right)\right]\\
&= \operatorname{Tr}\left((\mathbf{\Phi}^*\mathbf{\Phi} + \lambda \boldsymbol{I})^{-1} \cdot \mathbb{E}_{w\sim\mathcal{U}(\mathbb{S}^{d-1})}\left[\Phi_w\Phi_w^\top\right]\right)\\
&= \operatorname{Tr}\left((\mathbf{\Phi}^*\mathbf{\Phi} + \lambda \boldsymbol{I})^{-1} \cdot \mathbf{\Phi}^*\mathbf{\Phi}\right)\\
&= \operatorname{Tr}\left((\boldsymbol{K} + \lambda \boldsymbol{I})^{-1} \cdot \boldsymbol{K}\right).
\end{aligned}
$$

Next, we use the fact that the ridge leverage scores can be characterized in terms of a least-squares minimization problem, which is crucial for approximately computing the leverage scores distribution. This fact was previously exploited in (Avron et al., 2017b).

Lemma 17 (Minimization characterization of ridge leverage scores)**.** *For any $\lambda > 0$, let $\mathbf{\Phi}$ be the operator defined in Eq. (14), and its leverage score $\tau_\lambda(\cdot)$ be defined as in Definition 6. If we let Φ_w^i denote the i^{th} column of the matrix $\Phi_w \in \mathbb{R}^{n\times s}$ defined in Eq. (16) for any $i \in [s]$, the following holds,*

$$
\tau_\lambda(w) = \sum_{i\in[s]} \left(\min_{g_i\in L^2(\mathbb{S}^{d-1},\mathbb{R}^s)} \|g_i\|^2_{L^2(\mathbb{S}^{d-1},\mathbb{R}^s)} + \lambda^{-1}\cdot\left\|\mathbf{\Phi}^* g_i - \Phi_w^i\right\|_2^2\right) \quad \textit{for } w \in \mathbb{S}^{d-1}. \tag{35}
$$

We remark that this lemma is in fact a modification and generalization of Lemma 11 of (Avron et al., 2017b). We prove this lemma here for the sake of completeness.

Proof of Lemma 17. For any $i \in [s]$ let g_i^* denote the least-squares solution to the i^{th} summand in right hand side of Eq. (35). The optimal solution g_i^* can be obtained from the normal equation as follows,

$$
g_i^* = \left(\mathbf{\Phi}\mathbf{\Phi}^* + \lambda \boldsymbol{I}_{L^2(\mathbb{S}^{d-1},\mathbb{R}^s)}\right)^{-1} \cdot \mathbf{\Phi}\cdot\Phi_w^i = \mathbf{\Phi}\cdot(\mathbf{\Phi}^*\mathbf{\Phi} + \lambda\boldsymbol{I}_n)^{-1}\cdot\Phi_w^i = \mathbf{\Phi}\cdot(\boldsymbol{K}+\lambda\boldsymbol{I}_n)^{-1}\cdot\Phi_w^i,
$$

where the second equality above follows from the matrix inversion lemma for operators (Ogawa, 1988). We now have,

$$\begin{aligned}\|g_i^*\|^2_{L^2(\mathbb{S}^{d-1},\mathbb{R}^s)} &= \left\langle \boldsymbol{\Phi} \cdot (\boldsymbol{K}+\lambda \boldsymbol{I}_n)^{-1} \cdot \Phi_w^i, \boldsymbol{\Phi} \cdot (\boldsymbol{K}+\lambda \boldsymbol{I}_n)^{-1} \cdot \Phi_w^i \right\rangle_{L^2(\mathbb{S}^{d-1},\mathbb{R}^s)} \\ &= \left\langle \Phi_w^i, \left(\boldsymbol{\Phi} \cdot (\boldsymbol{K}+\lambda \boldsymbol{I}_n)^{-1}\right)^* \cdot \boldsymbol{\Phi} \cdot (\boldsymbol{K}+\lambda \boldsymbol{I}_n)^{-1} \cdot \Phi_w^i \right\rangle \\ &= \left\langle \Phi_w^i, (\boldsymbol{K}+\lambda \boldsymbol{I}_n)^{-1} \cdot \boldsymbol{K} \cdot (\boldsymbol{K}+\lambda \boldsymbol{I}_n)^{-1} \cdot \Phi_w^i \right\rangle \\ &= \Phi_w^{i\top} \cdot (\boldsymbol{K}+\lambda \boldsymbol{I})^{-1} \cdot \Phi_w^i - \lambda \cdot \Phi_w^{i\top} \cdot (\boldsymbol{K}+\lambda \boldsymbol{I})^{-2} \cdot \Phi_w^i.\end{aligned}$$

We also have,

$$\begin{aligned}\left\|\boldsymbol{\Phi}^* g_i^* - \Phi_w^i\right\|_2^2 &= \left\|\boldsymbol{\Phi}^*\boldsymbol{\Phi} \cdot (\boldsymbol{K}+\lambda \boldsymbol{I}_n)^{-1} \cdot \Phi_w^i - \Phi_w^i\right\|_2^2 \\ &= \left\|-\lambda \left(\boldsymbol{K}+\lambda \boldsymbol{I}_n\right)^{-1} \cdot \Phi_w^i\right\|_2^2 \\ &= \lambda^2 \cdot \Phi_w^{i\top} \cdot (\boldsymbol{K}+\lambda \boldsymbol{I}_n)^{-2} \cdot \Phi_w^i.\end{aligned}$$

Now by combining these equalities we have,

$$\|g_i^*\|^2_{L^2(\mathbb{S}^{d-1},\mathbb{R}^s)} + \lambda^{-1} \cdot \left\|\boldsymbol{\Phi}^* g_i^* - \Phi_w^i\right\|_2^2 = \Phi_w^{i\top} \cdot (\boldsymbol{K}+\lambda \boldsymbol{I})^{-1} \cdot \Phi_w^i.$$

Now summing the above over all $i \in [s]$ gives the lemma,

$$\begin{aligned}\sum_{i\in[s]} \|g_i^*\|^2_{L^2(\mathbb{S}^{d-1},\mathbb{R}^s)} + \lambda^{-1} \cdot \left\|\boldsymbol{\Phi}^* g_i^* - \Phi_w^i\right\|_2^2 &= \sum_{i\in[s]} \Phi_w^{i\top} \cdot (\boldsymbol{K}+\lambda \boldsymbol{I})^{-1} \cdot \Phi_w^i \\ &= \operatorname{Tr}\left(\Phi_w^{\top} \cdot (\boldsymbol{K}+\lambda \boldsymbol{I})^{-1} \cdot \Phi_w\right) \\ &:= \tau_\lambda(w).\end{aligned}$$

□

Now using the minimization characterization of the leverage score we can prove a uniform upper bound for any GZK and its corresponding feature as follows,

Lemma 7 (Upper bound on leverage scores of GZK)**.** *For any dataset $\boldsymbol{X} = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d\times n}$, let $\boldsymbol{\Phi}$ be the feature operator for the order s GZK on $\boldsymbol{X}$ defined in Eq. (14). For any $\lambda > 0$ and $w \in \mathbb{S}^{d-1}$, the ridge leverage scores of $\boldsymbol{\Phi}$ defined in Definition 6 are uniformly upper bounded by*

$$\tau_\lambda(w) \le \sum_{\ell=0}^{\infty} \alpha_{\ell,d} \min\left\{ \frac{\pi^2(\ell+1)^2}{6\lambda} \sum_{j\in[n]} \|h_\ell(\|x_j\|)\|^2, s \right\}.$$

Proof of Lemma 7. We prove the lemma using min-characterization of ridge leverage scores. Let $\mu := \frac{6\lambda s}{\pi^2 n}$ and define the data-dependent quantities R_ℓ as follows:

$$R_\ell := \frac{(\ell+1)^2}{n} \cdot \sum_{j\in[n]} \|h_\ell(\|x_j\|)\|^2, \qquad \text{for } \ell = 0, 1, 2, \ldots$$

Now, for any $i \in [s]$, let us define the function $g_w^i \in L^2(\mathbb{S}^{d-1}, \mathbb{R}^s)$ as,

$$g_w^i(\sigma) := \left(\sum_{\ell=0}^{\infty} \alpha_{\ell,d} \cdot \mathbb{1}_{\{R_\ell \ge \mu\}} \cdot P_d^\ell\left(\langle \sigma, w\rangle\right)\right) \cdot e_i,$$

where $e_i \in \mathbb{R}^s$ is the standard basis vector along the i^{th} coordinate. For this function we have,

$$\begin{aligned}
\|g_w^i\|^2_{L^2(\mathbb{S}^{d-1},\mathbb{R}^s)} &= \left\| \sum_{\ell=0}^{\infty} \alpha_{\ell,d} \cdot \mathbb{1}_{\{R_\ell \ge \mu\}} \cdot P_d^\ell\left(\langle \cdot, w\rangle\right) \right\|^2_{L^2(\mathbb{S}^{d-1})} \\
&= \sum_{\ell=0}^{\infty} \sum_{\ell'=0}^{\infty} \alpha_{\ell,d}\alpha_{\ell',d} \cdot \mathbb{1}_{\{R_\ell \ge \mu\}} \cdot \mathbb{1}_{\{R_{\ell'} \ge \mu\}} \cdot \mathbb{E}_{\sigma \sim \mathcal{U}(\mathbb{S}^{d-1})} \left[P_d^\ell\left(\langle \sigma, w\rangle\right) \cdot P_d^{\ell'}\left(\langle \sigma, w\rangle\right) \right] \\
&= \sum_{\ell=0}^{\infty} \alpha_{\ell,d} \cdot \mathbb{1}_{\{R_\ell \ge \mu\}} \cdot P_d^\ell\left(\langle w, w\rangle\right) = \sum_{\ell=0}^{\infty} \alpha_{\ell,d} \cdot \mathbb{1}_{\{R_\ell \ge \mu\}},
\end{aligned}$$

where the second line above follows from the definition of norm in the Hilbert space $L^2(\mathbb{S}^{d-1}, \mathbb{R})$ and the third line follows from Lemma 1 together with the fact that $P_d^\ell\left(\langle w, w\rangle\right) = P_d^\ell(1) = 1$. Thus, by summing the above over all $i \in [s]$ we get the following,

$$\sum_{i \in [s]} \|g_w^i\|^2_{L^2(\mathbb{S}^{d-1},\mathbb{R}^s)} = s \cdot \sum_{\ell=0}^{\infty} \alpha_{\ell,d} \cdot \mathbb{1}_{\{R_\ell \ge \mu\}} \tag{36}$$

Furthermore, for any $j \in [n]$ we have,

$$\begin{aligned}
[\mathbf{\Phi}^* g_w^i]_j &= \langle \phi_{x_j}, g_w^i \rangle_{L^2(\mathbb{S}^{d-1},\mathbb{R}^s)} \\
&= \left\langle \sum_{\ell=0}^{\infty} \sqrt{\alpha_{\ell,d}} \cdot \left[h_\ell(\|x_j\|)\right]_i \cdot P_d^\ell\left(\frac{\langle x_j, \cdot\rangle}{\|x_j\|}\right), \sum_{\ell=0}^{\infty} \alpha_{\ell,d} \cdot \mathbb{1}_{\{R_\ell \ge \mu\}} \cdot P_d^\ell\left(\langle \cdot, w\rangle\right) \right\rangle_{L^2(\mathbb{S}^{d-1})} \\
&= \sum_{\ell=0}^{\infty} \sqrt{\alpha_{\ell,d}} \cdot \left[h_\ell(\|x_j\|)\right]_i \cdot \mathbb{1}_{\{R_\ell \ge \mu\}} \cdot P_d^\ell\left(\frac{\langle x_j, w\rangle}{\|x_j\|}\right),
\end{aligned}$$

where the third line above follows from Lemma 1. Using the above equality along with definition of Φ_w in Eq. (16) and noting that Φ_w^i is the i^{th} column of this matrix, we can write,

$$\begin{aligned}
\left\|\mathbf{\Phi}^* g_w^i - \Phi_w^i\right\|_2^2 &= \sum_{j=1}^{n} \left| [\mathbf{\Phi}^* g_w^i]_j - [\Phi_w]_{j,i} \right|^2 \\
&= \sum_{j=1}^{n} \left| \langle \phi_{x_j}, g_w^i \rangle_{L^2(\mathbb{S}^{d-1},\mathbb{R}^s)} - \left[\phi_{x_j}(w)\right]_i \right|^2 \\
&= \sum_{j=1}^{n} \left| \sum_{\ell=0}^{\infty} \sqrt{\alpha_{\ell,d}} \cdot \left[h_\ell(\|x_j\|)\right]_i \cdot \mathbb{1}_{\{R_\ell < \mu\}} \cdot P_d^\ell\left(\frac{\langle x_j, w\rangle}{\|x_j\|}\right) \right|^2 \\
&\le \sum_{j=1}^{n} \left(\sum_{\ell=0}^{\infty} \sqrt{\alpha_{\ell,d}} \cdot \left| \left[h_\ell(\|x_j\|)\right]_i \right| \cdot \mathbb{1}_{\{R_\ell < \mu\}} \right)^2 \\
&= \sum_{j=1}^{n} \left(\sum_{\ell=0}^{\infty} \sqrt{\alpha_{\ell,d}} \cdot \sqrt{R_\ell} \cdot \mathbb{1}_{\{R_\ell < \mu\}} \cdot \frac{\left| \left[h_\ell(\|x_j\|)\right](i) \right|}{\sqrt{R_\ell}} \right)^2 \\
&\le \sum_{j=1}^{n} \left(\sum_{\ell=0}^{\infty} \alpha_{\ell,d} \cdot R_\ell \cdot \mathbb{1}_{\{R_\ell < \mu\}} \right) \cdot \left(\sum_{\ell=0}^{\infty} \frac{\left| \left[h_\ell(\|x_j\|)\right](i) \right|^2 \cdot \mathbb{1}_{\{0 < R_\ell < \mu\}}}{R_\ell} \right) \\
&= \left(\sum_{\ell=0}^{\infty} \alpha_{\ell,d} \cdot R_\ell \cdot \mathbb{1}_{\{R_\ell < \mu\}} \right) \cdot \sum_{\ell=0}^{\infty} \frac{\sum_{j=1}^{n} \left| \left[h_\ell(\|x_j\|)\right](i) \right|^2 \cdot \mathbb{1}_{\{0 < R_\ell < \mu\}}}{R_\ell},
\end{aligned}$$

where the first inequality above follows from the fact that $\left|P_d^\ell(t)\right| \le 1$ for $t \in [-1, 1]$ (See Equation (2.116) in (Atkinson & Han, 2012)) and the second inequality comes from Cauchy–Schwarz inequality. Therefore, if we sum the above over all

$i \in [s]$ we find the following inequlity,

$$\sum_{i \in [s]} \left\| \mathbf{\Phi}^* g_w^i - \Phi_w^i \right\|_2^2 \le \left(\sum_{\ell=0}^{\infty} \alpha_{\ell,d} \cdot R_\ell \cdot \mathbb{1}_{\{R_\ell < \mu\}} \right) \cdot \sum_{\ell=0}^{\infty} \frac{\sum_{j=1}^n \left\| h_\ell(\|x_j\|) \right\|^2 \cdot \mathbb{1}_{\{0 < R_\ell < \mu\}}}{R_\ell}$$
$$\le \frac{\pi^2 n}{6} \cdot \sum_{\ell=0}^{\infty} \alpha_{\ell,d} \cdot R_\ell \cdot \mathbb{1}_{\{R_\ell < \mu\}},$$

where the last line above follows from the definition of R_ℓ. Therefore, by combining the above with the norm of g_w^i's in Eq. (36), we find that,

$$\sum_{i \in [s]} \|g_w^i\|^2_{L^2(\mathbb{S}^{d-1}, \mathbb{R}^s)} + \lambda^{-1} \cdot \left\| \mathbf{\Phi}^* g_w^i - \Phi_w^i \right\|_2^2 \le s \cdot \sum_{\ell=0}^{\infty} \alpha_{\ell,d} \cdot \mathbb{1}_{\{R_\ell \ge \mu\}} + \frac{\pi^2 n}{6\lambda} \cdot \sum_{\ell=0}^{\infty} \alpha_{\ell,d} \cdot R_\ell \cdot \mathbb{1}_{\{R_\ell < \mu\}}$$
$$\le \sum_{\ell=0}^{\infty} \alpha_{\ell,d} \cdot \left(s \cdot \mathbb{1}_{\{R_\ell \ge \mu\}} + s\mu^{-1} R_\ell \cdot \mathbb{1}_{\{R_\ell < \mu\}} \right)$$
$$\le \sum_{\ell=0}^{\infty} \alpha_{\ell,d} \cdot \min\left\{ s\mu^{-1} R_\ell, s \right\}.$$

Plugging in the values of R_ℓ proves the lemma, because by Lemma 17, $\tau_\lambda(w) \le \sum_{i \in [s]} \|g_w^i\|^2_{L^2(\mathbb{S}^{d-1}, \mathbb{R}^s)} + \lambda^{-1} \cdot \left\| \mathbf{\Phi}^* g_w^i - \Phi_w^i \right\|_2^2$ for any $w \in \mathbb{S}^{d-1}$. □

F. Spectral Approximation to GZK Kernel Matrix

We will use the following version of the matrix Bernstein inequality to show spectral guarantees for our leverage scores sampling method.

Lemma 18 (Restatement of Corollary 7.3.3 of (Tropp, 2015))**.** *Let $\boldsymbol{B}$ be a fixed $n \times n$ matrix. Construct an $n \times n$ matrix $\boldsymbol{R}$ that, almost surely, satisfies,*

$$\mathbb{E}[\boldsymbol{R}] = \boldsymbol{B} \quad \textit{and} \quad \|\boldsymbol{R}\|_{\text{op}} \le L.$$

Let $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ be semi-definite upper bounds for the expected squares,

$$\mathbb{E}[\boldsymbol{R}\boldsymbol{R}^*] \preceq \boldsymbol{M}_1, \quad \textit{and} \ \mathbb{E}[\boldsymbol{R}^*\boldsymbol{R}] \preceq \boldsymbol{M}_2$$

Define the quantities $M = \max\{\|\boldsymbol{M}_1\|_{\text{op}}, \|\boldsymbol{M}_2\|_{\text{op}}\}$. Form the matrix sampling estimator,

$$\bar{\boldsymbol{R}} = \frac{1}{m} \sum_{j=1}^m \boldsymbol{R}_j,$$

where each $\boldsymbol{R}_j$ is an independent copy of $\boldsymbol{R}$. Then,

$$\Pr\left[\|\bar{\boldsymbol{R}} - \boldsymbol{B}\|_{\text{op}} \ge \varepsilon \right] \le 4 \cdot \frac{\text{Tr}(\boldsymbol{M}_1 + \boldsymbol{M}_2)}{M} \cdot \exp\left(\frac{-m\varepsilon^2/2}{M + 2L\varepsilon/3} \right).$$

Now we can prove Theorem 9. Our proof is a generalized version of Lemma 6 in (Avron et al., 2017b). We prove this theorem here for the sake of completeness.

Theorem 9 (Spectral approximation of GZK)**.** *For any dataset $\boldsymbol{X} = [x_1, x_2, \dots x_n] \in \mathbb{R}^{d \times n}$, let $\boldsymbol{K}$ be the corresponding GZK kernel matrix (Definition 3). For any $0 < \lambda \le \|\boldsymbol{K}\|_{\text{op}}$, let $\boldsymbol{Z} \in \mathbb{R}^{(m \cdot s) \times n}$ be the random features matrix defined in Definition 8. Also let s_λ be the statistical dimension of $\boldsymbol{K}$. For any $\varepsilon, \delta > 0$, if $m \ge \frac{8}{3\varepsilon^2} \log \frac{16 s_\lambda}{\delta} \cdot \sum_{\ell=0}^{\infty} \alpha_{\ell,d} \min\left\{ \frac{\pi^2(\ell+1)^2}{6\lambda} \sum_{j \in [n]} \left\| h_\ell(\|x_j\|) \right\|^2, s \right\}$, then with probability of at least $1 - \delta$,*

$$\frac{\boldsymbol{K} + \lambda \boldsymbol{I}}{1 + \varepsilon} \preceq \boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I} \preceq \frac{\boldsymbol{K} + \lambda \boldsymbol{I}}{1 - \varepsilon}. \tag{20}$$

Proof of Theorem 9. Let $\boldsymbol{K} + \lambda \boldsymbol{I} = \boldsymbol{V}^\top \boldsymbol{\Sigma}^2 \boldsymbol{V}$ be the singular value decomposition of the kernel matrix $\boldsymbol{K} + \lambda \boldsymbol{I}$. It is sufficient to show that,

$$\Pr\left[\left\| \boldsymbol{\Sigma}^{-1}\boldsymbol{V} \cdot \boldsymbol{Z}^\top \boldsymbol{Z} \cdot \boldsymbol{V}^\top \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{V} \cdot \boldsymbol{K} \cdot \boldsymbol{V}^\top \boldsymbol{\Sigma}^{-1} \right\|_{\mathrm{op}} \le \varepsilon \right] \ge 1 - \delta.$$

Now note that from definition of our random features matrix $\boldsymbol{Z}$ in Definition 8 we have,

$$\boldsymbol{\Sigma}^{-1}\boldsymbol{V} \cdot \boldsymbol{Z}^\top \boldsymbol{Z} \cdot \boldsymbol{V}^\top \boldsymbol{\Sigma}^{-1} = \frac{1}{m} \cdot \sum_{j=1}^{m} \boldsymbol{\Sigma}^{-1}\boldsymbol{V} \cdot \Phi_{w_j} \Phi_{w_j}^\top \cdot \boldsymbol{V}^\top \boldsymbol{\Sigma}^{-1}.$$

Thus, because in Definition 8, w_j's are sampled independently from each other from the distribution $\mathcal{U}(\mathbb{S}^{d-1})$, we can invoke Lemma 18 with the following arguments,

$$\boldsymbol{B} := \boldsymbol{\Sigma}^{-1}\boldsymbol{V} \cdot \boldsymbol{K} \cdot \boldsymbol{V}^\top \boldsymbol{\Sigma}^{-1}, \quad \text{and} \quad \boldsymbol{R}_j := \boldsymbol{\Sigma}^{-1}\boldsymbol{V} \cdot \Phi_{w_j} \Phi_{w_j}^\top \cdot \boldsymbol{V}^\top \boldsymbol{\Sigma}^{-1}.$$

Now we verify that the preconditions of Lemma 18 holds. First note that $\mathbb{E}[\boldsymbol{R}_j] = \boldsymbol{\Sigma}^{-1}\boldsymbol{V} \cdot \mathbb{E}_{w_j \sim \mathcal{U}(\mathbb{S}^{d-1})}[\Phi_{w_j} \Phi_{w_j}^\top] \cdot \boldsymbol{V}^\top \boldsymbol{\Sigma}^{-1} = \boldsymbol{B}$. Now we need to bound the operator norm of $\boldsymbol{R}_j$ and the stable rank $\mathbb{E}[\boldsymbol{R}_j^2]$. Using the cyclic property of trace, we can upper bound the operator norm $\|R_j\|_{\mathrm{op}}$ as follows,

$$\begin{aligned}
\|\boldsymbol{R}_j\|_{\mathrm{op}} &\le \operatorname{Tr}(\boldsymbol{R}_j) \\
&= \operatorname{Tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{V} \cdot \Phi_{w_j} \Phi_{w_j}^\top \cdot \boldsymbol{V}^\top \boldsymbol{\Sigma}^{-1} \right) \\
&= \operatorname{Tr}\left(\Phi_{w_j}^\top \cdot \boldsymbol{V}^\top \boldsymbol{\Sigma}^{-2} \boldsymbol{V} \cdot \Phi_{w_j} \right) \\
&= \operatorname{Tr}\left(\Phi_{w_j}^\top \cdot (\boldsymbol{K} + \lambda \boldsymbol{I})^{-1} \cdot \Phi_{w_j} \right) \\
&= \tau_\lambda(w_j),
\end{aligned}$$

where the last line above follows from Definition 6. This implies the following for any j,

$$\|\boldsymbol{R}_j\|_{\mathrm{op}} \le \max_{w \in \mathbb{S}^{d-1}} \tau_\lambda(w) := L. \tag{37}$$

We also have,

$$\begin{aligned}
\boldsymbol{R}_j^2 &= \boldsymbol{\Sigma}^{-1}\boldsymbol{V} \cdot \Phi_{w_j} \Phi_{w_j}^\top \cdot \boldsymbol{V}^\top \boldsymbol{\Sigma}^{-1} \cdot \boldsymbol{\Sigma}^{-1}\boldsymbol{V} \cdot \Phi_{w_j} \Phi_{w_j}^\top \cdot \boldsymbol{V}^\top \boldsymbol{\Sigma}^{-1} \\
&= \boldsymbol{\Sigma}^{-1}\boldsymbol{V} \cdot \Phi_{\sigma_j} \Phi_{w_j}^\top \cdot (\boldsymbol{K} + \lambda \boldsymbol{I})^{-1} \cdot \Phi_{w_j} \Phi_{w_j}^\top \cdot \boldsymbol{V}^\top \boldsymbol{\Sigma}^{-1} \\
&\preceq \operatorname{Tr}\left(\Phi_{w_j}^\top \cdot (\boldsymbol{K} + \lambda \boldsymbol{I})^{-1} \cdot \Phi_{w_j} \right) \cdot \boldsymbol{\Sigma}^{-1}\boldsymbol{V} \cdot \Phi_{w_j} \Phi_{w_j}^\top \cdot \boldsymbol{V}^\top \boldsymbol{\Sigma}^{-1} \\
&= \tau_\lambda(w_j) \cdot \boldsymbol{\Sigma}^{-1}\boldsymbol{V} \cdot \Phi_{w_j} \Phi_{w_j}^\top \cdot \boldsymbol{V}^\top \boldsymbol{\Sigma}^{-1}.
\end{aligned}$$

Now if we let $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_n$ be the eigenvalues of the kernel matrix $\boldsymbol{K}$ we find that the following holds for any j,

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{R}_j^2\right] &\preceq \left(\max_{w \in \mathbb{S}^{d-1}} \tau_\lambda(w) \right) \cdot \mathbb{E}_{w_j \sim \mathcal{U}(\mathbb{S}^{d-1})} \left[\boldsymbol{\Sigma}^{-1}\boldsymbol{V} \cdot \Phi_{w_j} \Phi_{w_j}^\top \cdot \boldsymbol{V}^\top \boldsymbol{\Sigma}^{-1} \right] \\
&= L \cdot \boldsymbol{\Sigma}^{-1}\boldsymbol{V} \cdot \boldsymbol{K} \cdot \boldsymbol{V}^\top \boldsymbol{\Sigma}^{-1} \\
&= L \cdot \left(\boldsymbol{I} - \lambda \boldsymbol{\Sigma}^{-2} \right) \\
&= L \cdot \operatorname{Diag}\left(\frac{\lambda_1}{\lambda_1 + \lambda}, \frac{\lambda_2}{\lambda_2 + \lambda}, \ldots \frac{\lambda_n}{\lambda_n + \lambda} \right) := \boldsymbol{D},
\end{aligned} \tag{38}$$

where L is the operator norm upper bound defined in Eq. (37). Now by invoking Lemma 7, we have the following upper bound,

$$L = \max_{w \in \mathbb{S}^{d-1}} \tau_\lambda(w) \le \sum_{\ell=0}^{\infty} \alpha_{\ell,d} \min\left\{ \frac{\pi^2(\ell+1)^2}{6\lambda} \sum_{j \in [n]} \|h_\ell(\|x_j\|)\|^2, s \right\}.$$

Therefore, by Lemma 18,

$$\Pr\left[\left\|\frac{1}{m}\sum_{j=1}^{m}\boldsymbol{R}_j - \boldsymbol{\Sigma}^{-1}\boldsymbol{V}\cdot\boldsymbol{K}\cdot\boldsymbol{V}^\top\boldsymbol{\Sigma}^{-1}\right\|_{\mathrm{op}} \geq \varepsilon\right] \leq 8\cdot\frac{\mathrm{Tr}(\boldsymbol{D})}{\|\boldsymbol{D}\|_{\mathrm{op}}}\cdot\exp\left(\frac{-m\varepsilon^2/2}{\|\boldsymbol{D}\|_{\mathrm{op}}+2L\varepsilon/3}\right)$$
$$\leq 8\cdot\frac{s_\lambda}{\lambda_1/(\lambda_1+\lambda)}\cdot\exp\left(\frac{-m\varepsilon^2/2}{L+2L\varepsilon/3}\right)$$
$$\leq \delta,$$

where the last line above is due to the fact that $\lambda_1 = \|\boldsymbol{K}\|_{\mathrm{op}} \geq \lambda$ along with the value of m. □

G. Projection Cost Preserving Samples for GZK

In this section we show that our random features results in an approximate kernel matrix that satisfies the projection-cost preservation condition. This property ensures that it is possible to extract a near optimal low-rank approximation from the random features. The proof of our result is based on (Cohen et al., 2017) which showed that unbiased leverage score sampling is sufficient for achieving this guarantee in discrete matrices. We extend this proof to the GZK quasi-matrix $\boldsymbol{\Phi}$.

Theorem 10 (Projection cost preserving GZK approximation)**.** *Let $\boldsymbol{K}$ be the GZK kernel matrix as in Theorem 9 with eigenvalues $\lambda_1 \geq \ldots \geq \lambda_n$. For any positive integer r, let $\lambda := \frac{1}{r}\sum_{i=r+1}^{n}\lambda_i$ and let s_λ be the statistical dimension of $\boldsymbol{K}$. For any $\varepsilon, \delta > 0$, if $\boldsymbol{Z} \in \mathbb{R}^{(m\cdot s)\times n}$ is the random features matrix defined in Definition 8 with $m \geq \frac{8}{3\varepsilon^2}\log\frac{16 s_\lambda}{\delta}\cdot\sum_{\ell=0}^{\infty}\alpha_{\ell,d}\min\left\{\frac{\pi^2(\ell+1)^2}{6\lambda}\sum_{j\in[n]}\|h_\ell(\|x_j\|)\|^2, s\right\}$, with probability at least $1-\delta$, the following holds for all rank-r orthonormal projections $\boldsymbol{P}$:*

$$1-\varepsilon \leq \frac{\mathrm{Tr}\left(\boldsymbol{Z}^\top\boldsymbol{Z} - \boldsymbol{P}\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{P}\right)}{\mathrm{Tr}(\boldsymbol{K}-\boldsymbol{P}\boldsymbol{K}\boldsymbol{P})} \leq 1+\varepsilon. \tag{21}$$

Proof of Theorem 10. The proof is nearly identical to the proof of Theorem 6 in (Cohen et al., 2017) which proves that unbiased leverage score sampling results in projection-cost preserving samples in discrete matrices. We adopt the proof of Theorem 6 of (Cohen et al., 2017) to our continuous operator $\boldsymbol{\Phi}$. First, for ease of notation let $\boldsymbol{Y} := \boldsymbol{I} - \boldsymbol{P}$. Now, note that we have the following,

$$\mathrm{Tr}(\boldsymbol{K}-\boldsymbol{P}\boldsymbol{K}\boldsymbol{P}) = \mathrm{Tr}(\boldsymbol{Y}\boldsymbol{K}\boldsymbol{Y}),$$
$$\mathrm{Tr}\left(\boldsymbol{Z}^\top\boldsymbol{Z}-\boldsymbol{P}\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{P}\right) = \mathrm{Tr}\left(\boldsymbol{Y}\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{Y}\right).$$

So it is enough to show that,

$$\frac{\mathrm{Tr}(\boldsymbol{Y}\boldsymbol{K}\boldsymbol{Y})}{1+\varepsilon} \leq \mathrm{Tr}\left(\boldsymbol{Y}\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{Y}\right) \leq \frac{\mathrm{Tr}(\boldsymbol{Y}\boldsymbol{K}\boldsymbol{Y})}{1-\varepsilon}.$$

Let t be the index of the smallest eigenvalue of $\boldsymbol{K}$ such that $\lambda_t \geq \frac{1}{r}\sum_{i=r+1}^{n}\lambda_i = \lambda$. Let $\boldsymbol{Q}_t$ denote the projection onto the eigenspace of matrix $\boldsymbol{K}$ corresponding to $\lambda_1, \lambda_2, \ldots, \lambda_t$. Also let $\boldsymbol{Q}_{\setminus t} := \boldsymbol{I} - \boldsymbol{Q}_t$. We split,

$$\mathrm{Tr}(\boldsymbol{Y}\boldsymbol{K}\boldsymbol{Y}) = \mathrm{Tr}(\boldsymbol{Y}\boldsymbol{Q}_t\boldsymbol{K}\boldsymbol{Q}_t\boldsymbol{Y}) + \mathrm{Tr}(\boldsymbol{Y}\boldsymbol{Q}_{\setminus t}\boldsymbol{K}\boldsymbol{Q}_{\setminus t}\boldsymbol{Y}) \tag{39}$$

Additionally, we split:

$$\mathrm{Tr}(\boldsymbol{Y}\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{Y}) = \mathrm{Tr}(\boldsymbol{Y}\boldsymbol{Q}_t\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{Q}_t\boldsymbol{Y}) + \mathrm{Tr}(\boldsymbol{Y}\boldsymbol{Q}_{\setminus t}\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{Q}_{\setminus t}\boldsymbol{Y}) + 2\mathrm{Tr}(\boldsymbol{Y}\boldsymbol{Q}_t\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{Q}_{\setminus t}\boldsymbol{Y}). \tag{40}$$

Head Terms. We first bound the term $\mathrm{Tr}(\boldsymbol{Y}\boldsymbol{Q}_t\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{Q}_t\boldsymbol{Y}) - \mathrm{Tr}(\boldsymbol{Y}\boldsymbol{Q}_t\boldsymbol{K}\boldsymbol{Q}_t\boldsymbol{Y})$. First note that by Eq. (20), for any vector $v \in \mathbb{R}^n$ we have,

$$(1-\varepsilon)v^\top\boldsymbol{Q}_t\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{Q}_t v - \varepsilon\lambda\|\boldsymbol{Q}_t v\|_2^2 \leq v^\top\boldsymbol{Q}_t\boldsymbol{K}\boldsymbol{Q}_t v \leq (1+\varepsilon)v^\top\boldsymbol{Q}_t\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{Q}_t v + \varepsilon\lambda\|\boldsymbol{Q}_t v\|_2^2.$$

By definition of t, $\boldsymbol{Q}_t v$ is orthogonal to all eigenvectors of $\boldsymbol{K}$ except those with eigenvalue greater than or equal to λ. Thus,

$$v^\top\boldsymbol{Q}_t\boldsymbol{K}\boldsymbol{Q}_t v \geq \lambda\|\boldsymbol{Q}_t v\|_2^2.$$

This inequality combines with the previous equality to give,

$$\frac{1-\varepsilon}{1+\varepsilon} \cdot v^\top \boldsymbol{Q}_t \boldsymbol{Z}^\top \boldsymbol{Z} \boldsymbol{Q}_t v \le v^\top \boldsymbol{Q}_t \boldsymbol{K} \boldsymbol{Q}_t v \le \frac{1+\varepsilon}{1-\varepsilon} \cdot v^\top \boldsymbol{Q}_t \boldsymbol{Z}^\top \boldsymbol{Z} \boldsymbol{Q}_t v,$$

for all $v \in \mathbb{R}^n$. This implies that,

$$\frac{1-\varepsilon}{1+\varepsilon} \cdot \boldsymbol{Q}_t \boldsymbol{Z}^\top \boldsymbol{Z} \boldsymbol{Q}_t \preceq \boldsymbol{Q}_t \boldsymbol{K} \boldsymbol{Q}_t \preceq \frac{1+\varepsilon}{1-\varepsilon} \cdot \boldsymbol{Q}_t \boldsymbol{Z}^\top \boldsymbol{Z} \boldsymbol{Q}_t. \tag{41}$$

Using the above we conclude that,

$$(1-3\varepsilon) \cdot \mathrm{Tr}(\boldsymbol{Y}\boldsymbol{Q}_t \boldsymbol{Z}^\top \boldsymbol{Z} \boldsymbol{Q}_t \boldsymbol{Y}) \le \mathrm{Tr}(\boldsymbol{Y}\boldsymbol{Q}_t \boldsymbol{K} \boldsymbol{Q}_t \boldsymbol{Y}) \le (1+3\varepsilon) \cdot \mathrm{Tr}(\boldsymbol{Y}\boldsymbol{Q}_t \boldsymbol{Z}^\top \boldsymbol{Z} \boldsymbol{Q}_t \boldsymbol{Y}).$$

Tail Terms. For the lower singular vectors of $\boldsymbol{K}$, Theorem 9 does not give a multiplicative bound, so we do things a bit differently. Specifically, we start by writing:

$$\begin{aligned}
\mathrm{Tr}(\boldsymbol{Y}\boldsymbol{Q}_{\setminus t}\boldsymbol{K}\boldsymbol{Q}_{\setminus t}\boldsymbol{Y}) &= \mathrm{Tr}(\boldsymbol{Q}_{\setminus t}\boldsymbol{K}\boldsymbol{Q}_{\setminus t}) - \mathrm{Tr}(\boldsymbol{P}\boldsymbol{Q}_{\setminus t}\boldsymbol{K}\boldsymbol{Q}_{\setminus t}\boldsymbol{P}),\\
\mathrm{Tr}(\boldsymbol{Y}\boldsymbol{Q}_{\setminus t}\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{Q}_{\setminus t}\boldsymbol{Y}) &= \mathrm{Tr}(\boldsymbol{Q}_{\setminus t}\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{Q}_{\setminus t}) - \mathrm{Tr}(\boldsymbol{P}\boldsymbol{Q}_{\setminus t}\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{Q}_{\setminus t}\boldsymbol{P})
\end{aligned}$$

We handle $\mathrm{Tr}(\boldsymbol{Q}_{\setminus t}\boldsymbol{K}\boldsymbol{Q}_{\setminus t})$ and $\mathrm{Tr}(\boldsymbol{Q}_{\setminus t}\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{Q}_{\setminus t})$ first. Since $\boldsymbol{Z}$ is constructed via an unbiased sampling of $\boldsymbol{\Phi}$ rows, $\mathbb{E}[\boldsymbol{Q}_{\setminus t}\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{Q}_{\setminus t}] = \boldsymbol{Q}_{\setminus t}\boldsymbol{K}\boldsymbol{Q}_{\setminus t}$ and a scalar-version Chernoff bound is sufficient for showing that this value concentrates around its expectation. We have the following bound:

$$\left|\mathrm{Tr}(\boldsymbol{Q}_{\setminus t}\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{Q}_{\setminus t}) - \mathrm{Tr}(\boldsymbol{Q}_{\setminus t}\boldsymbol{K}\boldsymbol{Q}_{\setminus t})\right| \le \varepsilon r \lambda.$$

Note that the above inequality does not depend on the choice of projection $\boldsymbol{P}$, so it holds simultaneously for all $\boldsymbol{P}$. We do not provide more details on why the above inequality holds but it follows fairly straightforwardly from scalar Chernoff bound. For example, one can find a detailed proof in Lemma 20 of (Cohen et al., 2017).

Next, we compare $\mathrm{Tr}(\boldsymbol{P}\boldsymbol{Q}_{\setminus t}\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{Q}_{\setminus t}\boldsymbol{P})$ to $\mathrm{Tr}(\boldsymbol{P}\boldsymbol{Q}_{\setminus t}\boldsymbol{K}\boldsymbol{Q}_{\setminus t}\boldsymbol{P})$. We first claim that:

$$\boldsymbol{Q}_{\setminus t}\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{Q}_{\setminus t} - 3\varepsilon\lambda\boldsymbol{I} \preceq \boldsymbol{Q}_{\setminus t}\boldsymbol{K}\boldsymbol{Q}_{\setminus t} \preceq \boldsymbol{Q}_{\setminus t}\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{Q}_{\setminus t} + 3\varepsilon\lambda\boldsymbol{I}. \tag{42}$$

The argument is similar to the one for Eq. (41). Now, since $\boldsymbol{P}$ is a rank r projection matrix this inequality implies that,

$$\mathrm{Tr}(\boldsymbol{P}\boldsymbol{Q}_{\setminus t}\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{Q}_{\setminus t}\boldsymbol{P}) - 3\varepsilon r\lambda \le \mathrm{Tr}(\boldsymbol{P}\boldsymbol{Q}_{\setminus t}\boldsymbol{K}\boldsymbol{Q}_{\setminus t}\boldsymbol{P}) \le \mathrm{Tr}(\boldsymbol{P}\boldsymbol{Q}_{\setminus t}\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{Q}_{\setminus t}\boldsymbol{P}) + 3\varepsilon r\lambda$$

which combines with the previous bound to give the final bound:

$$\left|\mathrm{Tr}(\boldsymbol{Y}\boldsymbol{Q}_{\setminus t}\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{Q}_{\setminus t}\boldsymbol{Y}) - \mathrm{Tr}(\boldsymbol{Y}\boldsymbol{Q}_{\setminus t}\boldsymbol{K}\boldsymbol{Q}_{\setminus t}\boldsymbol{Y})\right| \le 4\varepsilon r\lambda.$$

Cross Term. Finally, we handle the cross term $2\mathrm{Tr}(\boldsymbol{Y}\boldsymbol{Q}_m\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{Q}_{\setminus t}\boldsymbol{Y})$. We just need to show that it is small. To do so, we rewrite:

$$\mathrm{Tr}(\boldsymbol{Y}\boldsymbol{Q}_t\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{Q}_{\setminus t}\boldsymbol{Y}) = \mathrm{Tr}(\boldsymbol{Y}\boldsymbol{K}\boldsymbol{K}^\dagger\boldsymbol{Q}_t\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{Q}_{\setminus t}), \tag{43}$$

which holds since the columns of $\boldsymbol{Q}_t\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{Q}_{\setminus t}$ fall in the span of $\boldsymbol{K}$'s columns and the trailing $\boldsymbol{Y}$ gets eliminated by cyclic property of the trace. Now let us define the semi-inner product of matrices $\langle \boldsymbol{M}, \boldsymbol{N}\rangle := \mathrm{Tr}(\boldsymbol{M}\boldsymbol{K}^\dagger\boldsymbol{N}^\top)$. Thus, by Cauchy-Schwarz inequality, if we let $\boldsymbol{K} = \boldsymbol{U}\boldsymbol{\Sigma}^2\boldsymbol{U}^\top$ be the singular value decomposition of $\boldsymbol{K}$, we have,

$$\begin{aligned}
\mathrm{Tr}(\boldsymbol{Y}\boldsymbol{K}\boldsymbol{K}^\dagger\boldsymbol{Q}_m\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{Q}_{\setminus t}) &\le \sqrt{\mathrm{Tr}(\boldsymbol{Y}\boldsymbol{K}\boldsymbol{K}^\dagger\boldsymbol{K}\boldsymbol{Y}) \cdot \mathrm{Tr}(\boldsymbol{Q}_{\setminus t}\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{Q}_t\boldsymbol{K}^\dagger\boldsymbol{Q}_t\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{Q}_{\setminus t})}\\
&= \sqrt{\mathrm{Tr}(\boldsymbol{Y}\boldsymbol{K}\boldsymbol{Y}) \cdot \mathrm{Tr}(\boldsymbol{Q}_{\setminus t}\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{U}_t\boldsymbol{\Sigma}_t^{-2}\boldsymbol{U}_t^\top\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{Q}_{\setminus t})}\\
&= \sqrt{\mathrm{Tr}(\boldsymbol{Y}\boldsymbol{K}\boldsymbol{Y}) \cdot \|\boldsymbol{\Sigma}_t^{-1}\boldsymbol{U}_t^\top\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{Q}_{\setminus t}\|_F^2}.
\end{aligned} \tag{44}$$

To bound the second term, we write,

$$\left\|\boldsymbol{\Sigma}_t^{-1}\boldsymbol{U}_t^\top \boldsymbol{Z}^\top \boldsymbol{Z}\boldsymbol{Q}_{\setminus t}\right\|_F^2 = \sum_{i=1}^{t} \lambda_i^{-1} \cdot \|\boldsymbol{Q}_{\setminus t}\boldsymbol{Z}^\top \boldsymbol{Z} u_i\|_2^2,$$

where u_i is the i^{th} column of $\boldsymbol{U}$. Now we show that the summand is small for every $i \in [m]$. Let vector q_i be defined as $q_i := \frac{\boldsymbol{Q}_{\setminus t}\boldsymbol{Z}^\top \boldsymbol{Z} u_i}{\|\boldsymbol{Q}_{\setminus t}\boldsymbol{Z}^\top \boldsymbol{Z} u_i\|_2}$. Then we have,

$$\left\|\boldsymbol{\Sigma}_t^{-1}\boldsymbol{U}_t^\top \boldsymbol{Z}^\top \boldsymbol{Z}\boldsymbol{Q}_{\setminus t}\right\|_F^2 = \sum_{i=1}^{t} \lambda_i^{-1} \cdot \left(q_i^\top \boldsymbol{Z}^\top \boldsymbol{Z} u_i\right)^2. \tag{45}$$

Now, let us define the vector $v := \frac{u_i}{\sqrt{\lambda_i}} + \frac{q_i}{\sqrt{\lambda}}$. Using Eq. (20) we can write,

$$(1-\varepsilon) v^\top \boldsymbol{Z}^\top \boldsymbol{Z} v - \varepsilon\lambda \|v\|_2^2 \le v^\top \boldsymbol{K} v.$$

This expands out to,

$$\begin{aligned}
\frac{1-\varepsilon}{\lambda_i} u_i^\top \boldsymbol{Z}^\top \boldsymbol{Z} u_i + \frac{1-\varepsilon}{\lambda} q_i^\top \boldsymbol{Z}^\top \boldsymbol{Z} q_i + 2\frac{1-\varepsilon}{\sqrt{\lambda_i \cdot \lambda}} u_i^\top \boldsymbol{Z}^\top \boldsymbol{Z} q_i &\le \varepsilon\left(\frac{\lambda}{\lambda_i} + 1\right) + v^\top \boldsymbol{K} v \\
&\le 2\varepsilon + \frac{u_i^\top \boldsymbol{K} u_i}{\lambda_i} + \frac{q_i^\top \boldsymbol{K} q_i}{\lambda} \\
&= 2\varepsilon + 1 + \frac{q_i^\top \boldsymbol{K} q_i}{\lambda},
\end{aligned} \tag{46}$$

where the first inequality above follows because $u_i^\top q_i = \frac{u_i^\top \boldsymbol{Q}_{\setminus t}\boldsymbol{Z}^\top \boldsymbol{Z} u_i}{\|\boldsymbol{Q}_{\setminus t}\boldsymbol{Z}^\top \boldsymbol{Z} u_i\|_2} = 0$ for every $i \in [t]$. The second inequality above also follows because $u_i^\top \boldsymbol{K} q_i = \lambda_i \cdot u_i^\top q_i = 0$. Now note that, $u_i = \boldsymbol{Q}_t u_i$ for every $i \in [t]$, thus, by Eq. (41), $u_i^\top \boldsymbol{Z}^\top \boldsymbol{Z} u_i \ge \frac{1-\varepsilon}{1+\varepsilon} \cdot u_i^\top \boldsymbol{K} u_i = \frac{1-\varepsilon}{1+\varepsilon} \cdot \lambda_i$. Furthermore, using the fact that $p_i = \boldsymbol{Q}_{\setminus t} q_i$ along with Eq. (42), we have $q_i^\top \boldsymbol{Z}^\top \boldsymbol{Z} q_i \ge q_i^\top \boldsymbol{K} q_i - 3\varepsilon r\lambda \|q_i\|_2^2 = q_i^\top \boldsymbol{K} q_i - 3\varepsilon r\lambda$. Plugging these inequalities into Eq. (46) gives,

$$2\frac{1-\varepsilon}{\sqrt{\lambda_i \cdot \lambda}} u_i^\top \boldsymbol{Z}^\top \boldsymbol{Z} q_i \le 9\varepsilon + \varepsilon \cdot \frac{q_i^\top \boldsymbol{K} q_i}{\lambda} \le 10\varepsilon,$$

where the second inequality follows because q_i lies in the column span of $\boldsymbol{Q}_{\setminus t}$, thus $q_i^\top \boldsymbol{K} q_i \le \lambda_{t+1} \le \lambda$. Therefore,

$$(u_i^\top \boldsymbol{Z}^\top \boldsymbol{Z} q_i)^2 \le 26\varepsilon^2 \cdot \lambda \cdot \lambda_i.$$

Plugging into Eq. (45) gives:

$$\left\|\boldsymbol{\Sigma}_t^{-1}\boldsymbol{U}_t^\top \boldsymbol{Z}^\top \boldsymbol{Z}\boldsymbol{Q}_{\setminus t}\right\|_F^2 \le \sum_{i=1}^{t} 26\varepsilon^2\lambda \le 52\varepsilon^2 r\lambda,$$

where for the second inequality we used the fact that $t \le 2r$. Returning to Eq. (44) gives,

$$\mathrm{Tr}(\boldsymbol{Y}\boldsymbol{K}\boldsymbol{K}^\dagger \boldsymbol{Q}_t \boldsymbol{Z}^\top \boldsymbol{Z}\boldsymbol{Q}_{\setminus t}) \le 8\varepsilon \cdot \sqrt{r\lambda \cdot \mathrm{Tr}(\boldsymbol{Y}\boldsymbol{K}\boldsymbol{Y})} \le 8\varepsilon \cdot \mathrm{Tr}(\boldsymbol{Y}\boldsymbol{K}\boldsymbol{Y}),$$

where the second inequality follows from the fact that $r\lambda = \sum_{i=r+1}^{n} \lambda_i \le \mathrm{Tr}(\boldsymbol{Y}\boldsymbol{K}\boldsymbol{Y})$.

Final Bound. Finally by combining the bounds we obtained for **Head Terms**, **Tail Terms**, and **Cross Term** and applying the fact that $r\lambda = \sum_{i=r+1}^{n} \lambda_i \le \mathrm{Tr}(\boldsymbol{Y}\boldsymbol{K}\boldsymbol{Y})$, we find that,

$$\left|\mathrm{Tr}(\boldsymbol{Y}\boldsymbol{Z}^\top \boldsymbol{Z}\boldsymbol{Y}) - \mathrm{Tr}(\boldsymbol{Y}\boldsymbol{K}\boldsymbol{Y})\right| \le 4\varepsilon \mathrm{Tr}(\boldsymbol{Y}\boldsymbol{Q}_t \boldsymbol{K}\boldsymbol{Q}_t \boldsymbol{Y}) + 4\varepsilon r\lambda + 8\varepsilon \mathrm{Tr}(\boldsymbol{Y}\boldsymbol{K}\boldsymbol{Y}) \le 16\varepsilon \mathrm{Tr}(\boldsymbol{Y}\boldsymbol{K}\boldsymbol{Y}).$$

The proof of Theorem 10 follows by substituting $\varepsilon/16$ in place of ε in all the bounds above. □

H. Spectral Approximation of Dot-product Kernels

In this section we first provide formal statement of Theorem 11 and prove it.

Theorem (Formal statement of Theorem 11)**.** *Suppose Assumption 1 holds for a dot-product kernel $\kappa(\langle x, y\rangle)$. Given $\boldsymbol{X} = [x_1, \ldots, x_n] \in \mathbb{R}^{d \times n}$ for $d \geq 3$, assume that $\max_{j \in [n]} \|x_j\| \leq r$. Let $\boldsymbol{K}$ be the kernel matrix corresponding to $\kappa(\cdot)$ and $\boldsymbol{X}$. For any $0 < \lambda \leq \|\boldsymbol{K}\|_{\mathrm{op}}$ and $\varepsilon, \delta > 0$ let s_λ be the statistical dimension of $\boldsymbol{K}$. Also let $\boldsymbol{Z}$ be the proposed random features in Eq. (19) with $q = \max\left\{d, 3.7 r^2 \beta_\kappa, r^2 \beta_\kappa + \frac{d}{2} \log \frac{3 r^2 \beta_\kappa}{d} + \log \frac{C_\kappa n}{\varepsilon \lambda}\right\}$, $s = \max\left\{\frac{d}{2}, 3.7 r^2 \beta_\kappa, \frac{r^2 \beta_\kappa}{4} + \frac{1}{2} \log \frac{C_\kappa n}{\varepsilon \lambda}\right\}$ and $m = \frac{5 q^2}{4 \varepsilon^2} \cdot \binom{q+d-1}{q} \cdot \log \frac{16 s_\lambda}{\delta}$. Then, with probability at least $1 - \delta$, $\boldsymbol{Z}^\top \boldsymbol{Z}$ is an (ε, λ)-spectral approximation to $\boldsymbol{K}$ as per Eq. (1). Furthermore, $\boldsymbol{Z}$ can be computed in time $\mathcal{O}((ms/q) \cdot \mathrm{nnz}(\boldsymbol{X}))$.*

Proof. We first show that the low-degree GZK $k_{q,s}(x, y)$ corresponding to the radial functions $h_\ell(\cdot)$ defined in Eq. (22), tightly approximates the kernel $\kappa(\langle x, y\rangle)$ on every pair of points x, y in our dataset for $q = \max\left\{d, 3.7 r^2 \beta_\kappa, r^2 \beta_\kappa + \frac{d}{2} \log \frac{3 r^2 \beta_\kappa}{d} + \log \frac{C_\kappa n}{\varepsilon \lambda}\right\}$ and $s = \max\left\{\frac{d}{2}, 3.7 r^2 \beta_\kappa, \frac{r^2 \beta_\kappa}{4} + \frac{1}{2} \log \frac{C_\kappa n}{\varepsilon \lambda}\right\}$. By Lemma 4 and triangle inequality we have,

$$|k_{q,s}(x, y) - \kappa(\langle x, y\rangle)| \leq \left| \sum_{\ell > q} \left(\sum_{i=0}^{\infty} \widetilde{h}_{\ell,i}(\|x\|) \cdot \widetilde{h}_{\ell,i}(\|y\|) \right) \cdot P_d^\ell \left(\frac{\langle x, y\rangle}{\|x\| \cdot \|y\|} \right) \right| \tag{47}$$

$$+ \left| \sum_{\ell=0}^{q} \left(\sum_{i \geq s} \widetilde{h}_{\ell,i}(\|x\|) \cdot \widetilde{h}_{\ell,i}(\|y\|) \right) \cdot P_d^\ell \left(\frac{\langle x, y\rangle}{\|x\| \cdot \|y\|} \right) \right|, \tag{48}$$

where $\widetilde{h}_{\ell,i}(\cdot)$ is defined as per Eq. (12). We bound the terms in Eq. (47) and Eq. (48) separately. We first show that the coefficients of the monomials $\widetilde{h}_{\ell,i}(\cdot)$ in Eq. (12) decay exponentially as a function of i and ℓ. Since $\kappa(\langle x, y\rangle)$ is a valid kernel function, the derivative $\kappa^{(\ell+2i)}(0)$ must be non-negative for any ℓ and i (Schoenberg, 1942). Using the fact that $\alpha_{\ell,d} \leq \frac{(\ell+d-1)!}{\ell!(d-1)!}$ along with Assumption 1, we find the following bound for any $t \geq 0$,

$$0 \leq \widetilde{h}_{\ell,i}(t) \leq \sqrt{\frac{C_\kappa \cdot \beta_\kappa^{\ell+2i} \cdot \Gamma(\frac{d}{2})}{\sqrt{\pi} \cdot (d-1)!} \cdot \frac{(\ell+d-1)!}{2^\ell \cdot \ell! \cdot (2i)!} \cdot \frac{\Gamma(i+\frac{1}{2})}{\Gamma(i+\ell+\frac{d}{2})}} \cdot t^{\ell+2i}. \tag{49}$$

Now, using Eq. (49), we can bound the term in Eq. (47) as follows,

$$\begin{aligned}
\left| \sum_{\ell > q} \left(\sum_{i=0}^{\infty} \widetilde{h}_{\ell,i}(\|x\|) \widetilde{h}_{\ell,i}(\|y\|) \right) \cdot P_d^\ell \left(\frac{\langle x, y\rangle}{\|x\| \|y\|} \right) \right| &\leq \sum_{\ell > q} \left(\sum_{i=0}^{\infty} \widetilde{h}_{\ell,i}(\|x\|) \cdot \widetilde{h}_{\ell,i}(\|y\|) \right) \\
&\leq \sum_{\ell > q} \sum_{i=0}^{\infty} \frac{C_\kappa \cdot \beta_\kappa^{\ell+2i} \cdot \Gamma(\frac{d}{2})}{\sqrt{\pi} \cdot (d-1)!} \cdot \frac{(\ell+d-1)!}{2^\ell \cdot \ell! \cdot (2i)!} \cdot \frac{\Gamma(i+\frac{1}{2})}{\Gamma(i+\ell+\frac{d}{2})} \cdot r^{2\ell+4i} \\
&= \frac{C_\kappa \cdot \Gamma(\frac{d}{2})}{\sqrt{\pi} \cdot (d-1)!} \sum_{\ell > q} \frac{(\ell+d-1)!}{2^\ell \cdot \ell!} \cdot \sum_{i=0}^{\infty} \frac{\Gamma(i+\frac{1}{2})}{\Gamma(i+\ell+\frac{d}{2})} \cdot \frac{(r^2 \beta_\kappa)^{\ell+2i}}{(2i)!} \\
&\leq \frac{C_\kappa \cdot \Gamma(\frac{d}{2}) \cdot e^{r^2 \beta_\kappa}}{4 \cdot (d-1)!} \sum_{\ell > q} \frac{(\ell+d-1)!}{2^\ell \cdot \ell!} \cdot \frac{(r^2 \beta_\kappa)^\ell}{\Gamma(\ell+\frac{d}{2})}
\end{aligned}$$

where in the last line above we used the fact that $\frac{\Gamma(i+\frac{1}{2})}{\Gamma(i+\ell+\frac{d}{2})}$ is a decreasing function of i and the sum $\sum_{i=0}^{\infty} \frac{(r^2 \beta_\kappa)^{\ell+2i}}{(2i)!} =$

$\cosh(r^2\beta_\kappa) \le 0.57 e^{r^2\beta_\kappa}$. Now we can further upper bound the above as follows

$$\begin{aligned}
\left| \sum_{\ell > q} \left(\sum_{i=0}^{\infty} \widetilde{h}_{\ell,i}(\|x\|) \widetilde{h}_{\ell,i}(\|y\|) \right) \cdot P_d^\ell \left(\frac{\langle x, y \rangle}{\|x\|\|y\|} \right) \right| &\le \frac{C_\kappa \cdot \Gamma(\frac{d}{2}) \cdot e^{r^2\beta_\kappa}}{4 \cdot (d-1)!} \sum_{\ell > q} \frac{(\ell + d - 1)!}{2^\ell \cdot \ell!} \cdot \frac{(r^2\beta_\kappa)^\ell}{\Gamma(\ell + \frac{d}{2})} \\
&\le \frac{C_\kappa \cdot \Gamma(\frac{d}{2}) \cdot e^{r^2\beta_\kappa}}{4 \cdot (d-1)!} \cdot \sum_{\ell > q} \frac{1}{\ell^{\ell - \frac{d}{2}}} \cdot \left(\frac{e \cdot r^2\beta_\kappa}{2} \right)^\ell \cdot \left(1 + \frac{d-1}{\ell} \right)^{\frac{d}{2}} \\
&\le \frac{C_\kappa \cdot \Gamma(\frac{d}{2}) \cdot 2^{\frac{d}{2}} \cdot e^{r^2\beta_\kappa}}{5 \cdot (d-1)!} \cdot \sum_{\ell > q} \frac{1}{\ell^{\ell - \frac{d}{2}}} \cdot \left(\frac{e \cdot r^2\beta_\kappa}{2} \right)^\ell \\
&\le \frac{C_\kappa \cdot e^{r^2\beta_\kappa}}{20 (d/2)^{d/2}} \cdot \sum_{\ell > q} \frac{1}{\ell^{\ell - \frac{d}{2}}} \cdot \left(\frac{e \cdot r^2\beta_\kappa}{2} \right)^\ell \\
&\le \frac{C_\kappa \cdot e^{r^2\beta_\kappa}}{20} \cdot \left(\frac{e \cdot r^2\beta_\kappa}{d} \right)^{d/2} \sum_{\ell > q} \left(\frac{e \cdot r^2\beta_\kappa}{2\ell} \right)^{\ell - \frac{d}{2}} \\
&\le \frac{\varepsilon\lambda}{20n}. \qquad (50)
\end{aligned}$$

Similarly we upper bound the term in Eq. (48)

$$\begin{aligned}
\left| \sum_{\ell=0}^{q} \left(\sum_{i \ge s} \widetilde{h}_{\ell,i}(\|x\|) \widetilde{h}_{\ell,i}(\|y\|) \right) \cdot P_d^\ell \left(\frac{\langle x, y \rangle}{\|x\|\|y\|} \right) \right| &\le \sum_{\ell=0}^{q} \left(\sum_{i \ge s} \widetilde{h}_{\ell,i}(\|x\|) \cdot \widetilde{h}_{\ell,i}(\|y\|) \right) \\
&\le \sum_{\ell=0}^{q} \sum_{i \ge s} \frac{C_\kappa \cdot \beta_\kappa^{\ell+2i} \cdot \Gamma(\frac{d}{2})}{\sqrt{\pi} \cdot (d-1)!} \cdot \frac{(\ell + d - 1)!}{2^\ell \cdot \ell! \cdot (2i)!} \cdot \frac{\Gamma(i + \frac{1}{2})}{\Gamma(i + \ell + \frac{d}{2})} \cdot r^{2\ell + 4i} \\
&\le \frac{C_\kappa \cdot \Gamma(\frac{d}{2})}{\sqrt{\pi} \cdot (d-1)!} \sum_{i=s}^{\infty} \frac{\Gamma(i + \frac{1}{2}) \cdot (r^2\beta_\kappa)^{2i}}{(2i)!} \sum_{\ell=0}^{q} \frac{(\ell + d - 1)! \cdot (r^2\beta_\kappa)^\ell}{2^\ell \cdot \ell! \cdot \Gamma(i + \ell + \frac{d}{2})} \\
&\le \frac{C_\kappa \cdot \Gamma(\frac{d}{2})}{5 \cdot (d-1)!} \sum_{i=s}^{\infty} \frac{\Gamma(i + \frac{1}{2}) \cdot (r^2\beta_\kappa)^{2i}}{(2i)!} \cdot \frac{(d-1)! \cdot e^{\frac{r^2\beta_\kappa}{2}}}{\Gamma(i + \frac{d}{2})} \\
&= \frac{C_\kappa \cdot \Gamma(\frac{d}{2}) \cdot e^{\frac{r^2\beta_\kappa}{2}}}{5} \sum_{i=s}^{\infty} \frac{\Gamma(i + \frac{1}{2}) \cdot (r^2\beta_\kappa)^{2i}}{(2i)! \cdot \Gamma(i + \frac{d}{2})} \\
&\le \frac{C_\kappa \cdot e^{\frac{r^2\beta_\kappa}{2}}}{20} \sum_{i=s}^{\infty} \left(\frac{e \cdot r^2\beta_\kappa}{2i} \right)^{2i} \\
&\le \frac{\varepsilon\lambda}{20n}. \qquad (51)
\end{aligned}$$

Thus, by combining Eq. (50) and Eq. (51), we find that for every pair of points $x, y \in \mathbb{R}^d$ with $\|x\|, \|y\| \le r$ the following holds,

$$|k_{q,s}(x, y) - \kappa(\langle x, y \rangle)| \le \frac{\varepsilon\lambda}{10n}.$$

Therefore, if we let $\widetilde{\boldsymbol{K}} \in \mathbb{R}^{n \times n}$ be the kernel matrix corresponding to kernel function $k_{s,q}(\cdot)$ and dataset $\boldsymbol{X}$, then we have the following,

$$\left\| \widetilde{\boldsymbol{K}} - \boldsymbol{K} \right\|_F \le \frac{\varepsilon\lambda}{10}.$$

Now we let $\boldsymbol{Z} \in \mathbb{R}^{(m \cdot s) \times n}$ be the random features matrix as in Definition 8 corresponding to the kernel function $k_{q,s}(x, y)$.

The bound on the number of features given in Theorem 9 for the kernel function $k_{q,s}(x,y)$ is upper bounded by,

$$\begin{aligned}\sum_{\ell=0}^{q} \alpha_{\ell,d} \min\left\{ \frac{\pi^2(\ell+1)^2}{6\lambda} \sum_{j\in[n]} \|h_\ell(\|x_j\|)\|_2^2 \,,\ s \right\} &\le \sum_{\ell=0}^{q} \alpha_{\ell,d} \cdot s = s\left(\binom{q+d-1}{q} + \binom{q+d-2}{q-1} \right)\\ &\le s \cdot \binom{q+d-1}{q} \cdot \left(1 + \frac{q}{q+d-1}\right)\\ &= 2s \cdot \binom{q+d-1}{q}.\end{aligned}$$

Thus by plugging this bound into Theorem 9 we get that,

$$(1-8\varepsilon/10)\cdot(\widetilde{\boldsymbol{K}}+\lambda\boldsymbol{I}) \preceq \boldsymbol{Z}^\top\boldsymbol{Z}+\lambda\boldsymbol{I} \preceq (1+8\varepsilon/10)\cdot(\widetilde{\boldsymbol{K}}+\lambda\boldsymbol{I}).$$

The fact that $\left\|\widetilde{\boldsymbol{K}}-\boldsymbol{K}\right\|_F \le \frac{\varepsilon\lambda}{10}$ gives the lemma.

Runtime. The runtime of computing the features in Definition 8 is equal to the time to compute $\boldsymbol{X}^\top w_j$ for all $j\in[m]$ along with the time to evaluate the polynomials $P_d^\ell(t)$ at mn different values of t for all $\ell\in[q]$. These operations can be done in total time $\mathcal{O}\left(m\cdot \mathrm{nnz}\left(\boldsymbol{X}\right)\right) = \mathcal{O}\left((ms/q)\cdot \mathrm{nnz}\left(\boldsymbol{X}\right)\right)$. Note that, to compute these random features we also need to evaluate the derivatives of function $\kappa(t)$ at zero (up to order q), however this is just a one time computation and does not need to be repeated for each data-point, thus, we can assume that this time would not depend on n or m or d and is negligible compared to $\mathcal{O}\left((ms/q)\cdot \mathrm{nnz}\left(\boldsymbol{X}\right)\right)$.

□

I. Spectral Approximation to Gaussian Kernel

In this section we prove Theorem 12.

Theorem 12. *Given $\boldsymbol{X} = [x_1,\ldots,x_n]\in\mathbb{R}^{d\times n}$ for $d\ge 3$, assume that $\max_{j\in[n]}\|x_j\|\le r$. Let $\boldsymbol{K}\in\mathbb{R}^{n\times n}$ be the corresponding Gaussian kernel matrix $[\boldsymbol{K}]_{i,j} = e^{-\|x_i-x_j\|_2^2/2}$. For any $0<\lambda\le\|\boldsymbol{K}\|_{\mathrm{op}}$ and $\varepsilon,\delta>0$, let s_λ denote the statistical dimension of $\boldsymbol{K}$ and define $q = \max\left\{3.7r^2, \frac{d}{2}\log\frac{2.8(r^2+\log\frac{n}{\varepsilon\lambda}+d)}{d} + \log\frac{n}{\varepsilon\lambda}\right\}$. There exists an algorithm that can output a feature matrix $\boldsymbol{Z}\in\mathbb{R}^{m\times n}$ with $m = \frac{25q^2}{3\varepsilon^2}\binom{q+d-1}{q}\log\left(\frac{16s_\lambda}{\delta}\right)$, such that with probability at least $1-\delta$, $\boldsymbol{Z}^\top\boldsymbol{Z}$ is an (ε,λ)-spectral approximation to $\boldsymbol{K}$ as per Eq. (1). Furthermore, $\boldsymbol{Z}$ can be computed in time $\mathcal{O}((m/q)\cdot\mathrm{nnz}\left(\boldsymbol{X}\right))$.*

Proof. We first show that the low-degree GZK $g_{q,s}(x,y)$ corresponding to the radial functions $h_\ell(\cdot)$ defined in Eq. (23), tightly approximates the Gaussian kernel $g(x,y)$ on every pair of points x,y in our dataset for $q = \max\left\{3.7r^2, \frac{d}{2}\log\frac{2.8(r^2+\log\frac{n}{\varepsilon\lambda}+d)}{d} + \log\frac{n}{\varepsilon\lambda}\right\}$ and $s = \max\left\{\frac{d}{2}, 3.7r^2, \frac{1}{2}\log\frac{n}{\varepsilon\lambda}\right\}$. By Lemma 15 and triangle inequality we have the following,

$$|g_{q,s}(x,y) - g(x,y)| \le \left| \sum_{\ell>q} \left(\sum_{i=0}^{\infty} \widetilde{h}_{\ell,i}(\|x\|)\cdot\widetilde{h}_{\ell,i}(\|y\|) \right)\cdot P_d^\ell\left(\frac{\langle x,y\rangle}{\|x\|\cdot\|y\|}\right) \right| \tag{52}$$

$$+ \left| \sum_{\ell=0}^{q} \left(\sum_{i\ge s} \widetilde{h}_{\ell,i}(\|x\|)\cdot\widetilde{h}_{\ell,i}(\|y\|) \right)\cdot P_d^\ell\left(\frac{\langle x,y\rangle}{\|x\|\cdot\|y\|}\right) \right|, \tag{53}$$

where $\widetilde{h}_{\ell,i}(\cdot)$ is defined as in the statement of Lemma 15. We bound the terms in Eq. (52) and Eq. (53) separately. By

Cauchy–Schwarz inequality and the fact that $\widetilde{h}_{\ell,i}(\|x\|)$ and $\widetilde{h}_{\ell,i}(\|y\|)$ are non-negative, we can bound Eq. (52) as follows,

$$\begin{aligned}\left|\sum_{\ell>q}\left(\sum_{i=0}^{\infty}\widetilde{h}_{\ell,i}(\|x\|)\widetilde{h}_{\ell,i}(\|y\|)\right)P_d^{\ell}\left(\frac{\langle x,y\rangle}{\|x\|\|y\|}\right)\right| &\le \left|\sum_{\ell>q}\left(\sum_{i=0}^{\infty}\widetilde{h}_{\ell,i}(\|x\|)\cdot\widetilde{h}_{\ell,i}(\|y\|)\right)\right|\\ &\le \sqrt{\sum_{\ell>q}\sum_{i=0}^{\infty}|\widetilde{h}_{\ell,i}(\|x\|)|^2\cdot\sum_{\ell>q}\sum_{i=0}^{\infty}|\widetilde{h}_{\ell,i}(\|y\|)|^2}.\end{aligned}$$

Now we can bound the term $\sum_{\ell>q}\sum_{i=0}^{\infty}|\widetilde{h}_{\ell,i}(\|x\|)|^2$, using the definition of $\widetilde{h}_{\ell,i}(\cdot)$, as follows,

$$\begin{aligned}\sum_{\ell>q}\sum_{i=0}^{\infty}|\widetilde{h}_{\ell,i}(\|x\|)|^2 &\le \frac{\Gamma(\frac{d}{2})}{\sqrt{\pi}\cdot(d-1)!}\cdot\sum_{\ell>q}\frac{(\ell+d-1)!}{2^{\ell}\cdot\ell!}\cdot\sum_{i=0}^{\infty}\frac{\Gamma(i+\frac{1}{2})}{\Gamma(i+\ell+\frac{d}{2})}\cdot\frac{\|x\|^{2\ell+4i}e^{-\|x\|^2}}{(2i)!}\\ &\le \frac{\Gamma(\frac{d}{2})}{4\cdot(d-1)!}\cdot\sum_{\ell>q}\frac{(\ell+d-1)!}{2^{\ell}\cdot\ell!}\cdot\frac{\|x\|^{2\ell}}{\Gamma(\ell+\frac{d}{2})}\\ &\le \frac{1}{4}\sum_{\ell>q}\left(\frac{e\cdot\ell}{d}\right)^{\frac{d}{2}}\cdot\frac{\|x\|^{2\ell}}{2^{\ell}\cdot\ell!}\\ &\le \frac{1}{4}\sum_{\ell>q}\left(\frac{e\cdot\ell}{d}\right)^{\frac{d}{2}}\cdot\left(\frac{e\cdot r^2}{2\ell}\right)^{\ell}\\ &\le \frac{\varepsilon\lambda}{20n}.\end{aligned}$$

Similarly, we can show $\sum_{\ell>q}\sum_{i=0}^{\infty}|\widetilde{h}_{\ell,i}(\|y\|)|^2\le\frac{\varepsilon\lambda}{20n}$, thus, the term in Eq. (52) is bounded by,

$$\left|\sum_{\ell>q}\left(\sum_{i=0}^{\infty}\widetilde{h}_{\ell,i}(\|x\|)\cdot\widetilde{h}_{\ell,i}(\|y\|)\right)\cdot P_d^{\ell}\left(\frac{\langle x,y\rangle}{\|x\|\cdot\|y\|}\right)\right|\le\frac{\varepsilon\lambda}{20n}. \tag{54}$$

Now we upper bound the term in Eq. (53) using Cauchy–Schwarz inequality as follows,

$$\begin{aligned}\left|\sum_{\ell=0}^{q}\left(\sum_{i\ge s}\widetilde{h}_{\ell,i}(\|x\|)\cdot\widetilde{h}_{\ell,i}(\|y\|)\right)P_d^{\ell}\left(\frac{\langle x,y\rangle}{\|x\|\|y\|}\right)\right| &\le \left|\sum_{\ell=0}^{q}\left(\sum_{i\ge s}\widetilde{h}_{\ell,i}(\|x\|)\cdot\widetilde{h}_{\ell,i}(\|y\|)\right)\right|\\ &\le \sqrt{\sum_{\ell=0}^{q}\sum_{i\ge s}|\widetilde{h}_{\ell,i}(\|x\|)|^2\cdot\sum_{\ell=0}^{q}\sum_{i\ge s}|\widetilde{h}_{\ell,i}(\|y\|)|^2}.\end{aligned}$$

Now we can bound the term $\sum_{\ell=0}^{q}\sum_{i\ge s}|\widetilde{h}_{\ell,i}(\|x\|)|^2$, using the definition of $\widetilde{h}_{\ell,i}(\cdot)$, as follows,

$$\begin{aligned}\sum_{\ell=0}^{q}\sum_{i\ge s}|\widetilde{h}_{\ell,i}(\|x\|)|^2 &\le \frac{\Gamma(\frac{d}{2})}{\sqrt{\pi}\cdot(d-1)!}\cdot\sum_{i\ge s}\frac{\Gamma(i+\frac{1}{2})}{(2i)!}\sum_{\ell=0}^{q}\frac{(\ell+d-1)!}{2^{\ell}\cdot\ell!}\cdot\frac{\|x\|^{2\ell+4i}e^{-\|x\|^2}}{\Gamma(i+\ell+\frac{d}{2})}\\ &\le \frac{\Gamma(\frac{d}{2})}{\sqrt{\pi}\cdot(d-1)!}\cdot\sum_{i\ge s}\frac{\Gamma(i+\frac{1}{2})}{(2i)!}\cdot\frac{(d-1)!}{\Gamma(i+\frac{d}{2})}\sum_{\ell=0}^{q}\frac{\|x\|^{2\ell+4i}e^{-\|x\|^2}}{2^{\ell}\cdot\ell!}\\ &\le \frac{\Gamma(\frac{d}{2})}{\sqrt{\pi}}\cdot\sum_{i\ge s}\frac{\Gamma(i+\frac{1}{2})}{\Gamma(i+\frac{d}{2})\cdot(2i)!}\cdot\|x\|^{4i}e^{-\|x\|^2/2}\\ &\le \frac{e^{-\|x\|^2/2}}{5}\sum_{i\ge s}\left(\frac{e\cdot\|x\|^2}{2i}\right)^{2i}\\ &\le \frac{\varepsilon\lambda}{20n}.\end{aligned}$$

Similarly, we can show $\sum_{\ell=0}^{q} \sum_{i \ge s} |\widetilde{h}_{\ell,i}(\|y\|)|^2 \le \frac{\varepsilon\lambda}{20n}$, thus, the term in Eq. (53) is bounded by,

$$\left| \sum_{\ell=0}^{q} \left(\sum_{i \ge s} \widetilde{h}_{\ell,i}(\|x\|) \cdot \widetilde{h}_{\ell,i}(\|y\|) \right) \cdot P_d^{\ell}\left(\frac{\langle x, y \rangle}{\|x\| \cdot \|y\|} \right) \right| \le \frac{\varepsilon\lambda}{20n}. \tag{55}$$

Thus by combining Eq. (54) and Eq. (55), we find that for every pair of points $x, y \in \mathbb{R}^d$ with $\|x\|, \|y\| \le r$ the following holds,

$$|g_{q,s}(x,y) - g(x,y)| \le \frac{\varepsilon\lambda}{10n}.$$

Therefore, if we let $\widetilde{\boldsymbol{K}} \in \mathbb{R}^{n \times n}$ be the kernel matrix corresponding to kernel function $g_{s,q}(\cdot)$ and dataset $\boldsymbol{X}$, then,

$$\left\| \widetilde{\boldsymbol{K}} - \boldsymbol{K} \right\|_F \le \frac{\varepsilon\lambda}{10}.$$

Now we let $\boldsymbol{Z} \in \mathbb{R}^{(m \cdot s) \times n}$ be the random features matrix as in Definition 8 corresponding to the kernel function $g_{q,s}(x,y)$. The bound on the number of features given in Theorem 9 for the kernel function $g_{q,s}(x,y)$ is upper bounded by,

$$\sum_{\ell=0}^{q} \alpha_{\ell,d} \min \left\{ \frac{\pi^2(\ell+1)^2}{6\lambda} \sum_{j \in [n]} \|h_\ell(\|x_j\|)\|^2, s \right\} \le 1.1s \cdot \binom{q+d-1}{q}.$$

Thus by plugging this bound into Theorem 9 we get that,

$$(1 - 8\varepsilon/10) \cdot (\widetilde{\boldsymbol{K}} + \lambda \boldsymbol{I}) \preceq \boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I} \preceq (1 + 8\varepsilon/10) \cdot (\widetilde{\boldsymbol{K}} + \lambda \boldsymbol{I}).$$

Using the fact that $\left\| \widetilde{\boldsymbol{K}} - \boldsymbol{K} \right\|_F \le \frac{\varepsilon\lambda}{10}$ gives the lemma.

Runtime. The runtime of computing the features in Definition 8 is equal to the time to compute $\boldsymbol{X}^\top \sigma_j$ for all $j \in [m]$ along with the time to evaluate the polynomials $P_d^{\ell}(t)$ at mn different values of t for all $\ell \in [q]$. These operations can be done in total time $\mathcal{O}\left(m \cdot \mathrm{nnz}\left(\boldsymbol{X}\right)\right) = \mathcal{O}\left((ms/q) \cdot \mathrm{nnz}\left(\boldsymbol{X}\right)\right)$

□

J. Experimental Details

J.1. Details on Kernel Ridge Regression

For kernel ridge regression, we use 4 real-world datasets, e.g., Earth Elevation[2], CO_2 [3], Climate[4] and Protein[5]. For Elevation, CO_2 , Climate datasets, each data point is represented by a (latitude, longitude) pair. We convert the location values into the 3D-Cartesian coordinates (i.e., $\mathbb{S}^2$). In addition, both CO_2 and climate datasets contain 12 different temporal values. and append the temporal one if they exist. For Protein dataset, each data point is given by 10-dimensional features. We consider the first 9 features as training data and the final feature as label. We also normalize those features so that each feature has zero mean and 1 standard deviation. For all datasets, we randomly split 90% training and 10% testing, and find the ridge parameter via the 2-fold cross-validation on the training set. For all kernel approximation methods, we set the final feature dimension to $m = 1{,}024$.

Results of NTK. We conduct additional experiments on kernel ridge regression (Section 6.2) with Neural Tangent Kernel (NTK) of a 2-layer fully-connect neural network as the kernel function. Since the Random Fourier Features and Fastfood can be only applied for the Gaussian kernel, we only report results of Nyström, Random Maclaurin, PolySketch and ours. Similar to Table 4, our random Gegenbauer features achieves the best MSE aside from the Nyström. However, the runtime of Nyström is up to 29 times slower than our method.

[2] https://github.com/fatiando/rockhound
[3] https://db.cger.nies.go.jp/dataset/ODIAC/
[4] http://berkeleyearth.lbl.gov/
[5] https://archive.ics.uci.edu/

Table 4. Results of kernel ridge regression with NTK.

	Elevation		CO_2		Climate	
Metric	MSE	Time	MSE	Time	MSE	Time
Nyström	0.61	43.2	0.43	95.3	3.12	152
Maclaurin	2.88	0.91	0.63	2.09	3.32	3.27
PolySketch	2.88	7.64	0.63	16.1	3.32	24.6
Gegenbauer	0.94	1.59	0.51	3.45	3.13	5.25

J.2. Details on Kernel k-means Clustering

For kernel k-means clustering, we use 6 UCI classification datasets[6]. We normalize the inputs by this l_2 norms so that they are on the unit sphere. In addition, we use the k-means clustering algorithm from an open-source scikit-learn[7] package (`sklearn.cluster.KMeans`) where initial seed points are chosen by k-mean++ initialization (Arthur & Vassilvitskii, 2006). The number of clusters is set to the number of classes of each dataset and the number of features are commonly set to $m = 512$.

J.3. Numerical Stability

The Gegenbauer polynomials $P_d^\ell(\cdot)$ can be computed using their recursion:

$$P_d^{\ell+1}(t) = \frac{2\ell + d - 4}{\ell + d - 3} P_d^\ell(t) - \frac{\ell - 1}{\ell + d - 3} P_d^{\ell-1}(t)$$

where $P_d^1(t) = t, P_d^0(t) = 1, \ell \geq 1$ and we do not see any numerical issues for computing P_d^ℓ. However, coefficients c_ℓ of Gegenbauer series involve in integration of some scalar functions in Eq. (8). We use `scipy.integrate.quad` to estimate a definite integration with Gaussian quadrature. We verify that coefficients for $d \geq 32$ and $\ell \geq 16$ are numerically unstable. To avoid this, we suggest to set the maximum degree to 15, and this provides a sufficiently accurate polynomial approximation.

[6] http://persoal.citius.usc.es/manuel.fernandez.delgado/papers/jmlr/data.tar.gz

[7] https://scikit-learn.org/