Random Gegenbauer Features for Scalable Kernel Methods

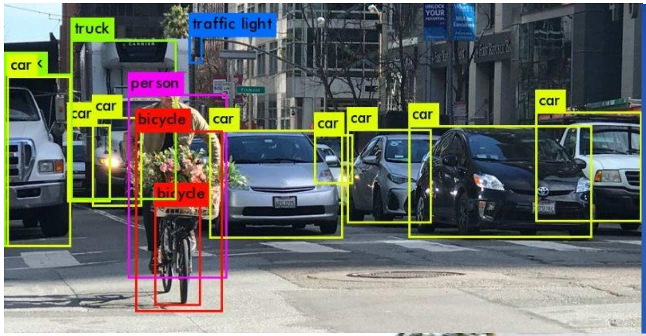
Insu Han◐, Amir Zandieh▣, Haim Avron◈

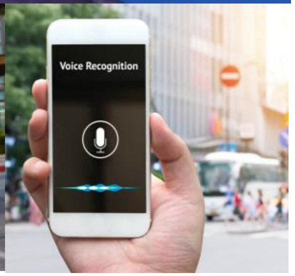
◐Yale University
▣Max-Planck-Institute
◈Tel Aviv University

ICML 2022

Kernel Methods

- Widely used in kernel-based learning, statistics and control
- Classical machine learning tool with real-world applications





Kernel Regression

- **Kernel:** a similarity function over pairs of data points in raw representation
  - Mercer decomposition: for every kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and $x, x' \in \mathbb{R}^d$

$$k(x, x') = \langle \phi(x), \phi(x') \rangle$$

  - ϕ is called a feature map

Kernel Regression

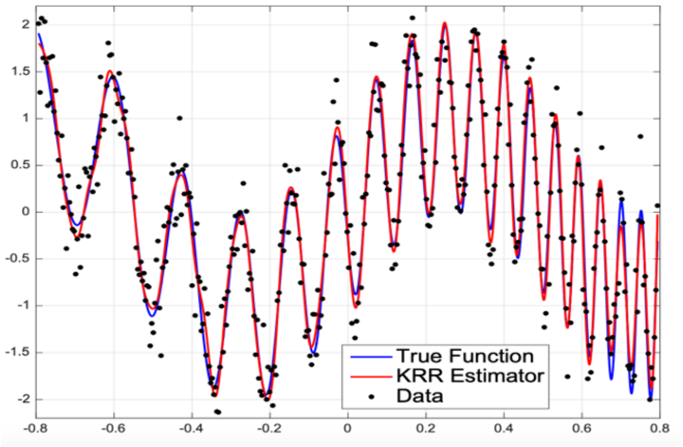
- **Kernel:** a similarity function over pairs of data points in raw representation
  - Mercer decomposition: for every kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and $x, x' \in \mathbb{R}^d$

$$k(x, x') = \langle \phi(x), \phi(x') \rangle$$

  - ϕ is called a feature map

- **Kernel ridge regression:**

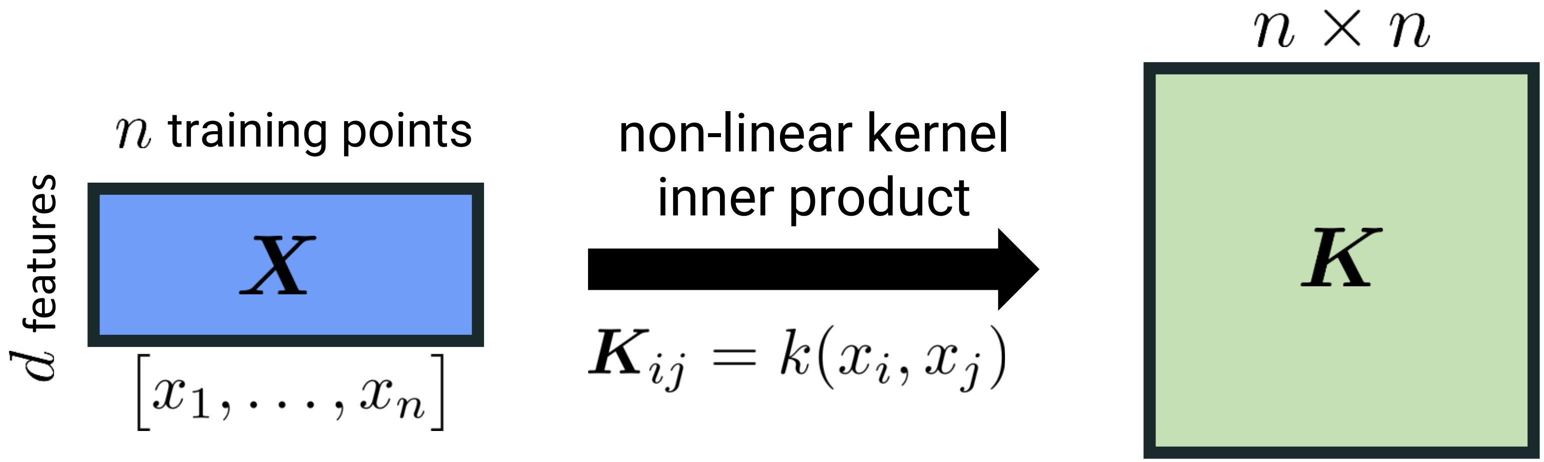
$$w^* = \underset{w}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (y_i - \phi(x_i)^\top w)^2 + \lambda \|w\|^2$$



  - Simple yet powerful tool for learning non-linear relationships between data points

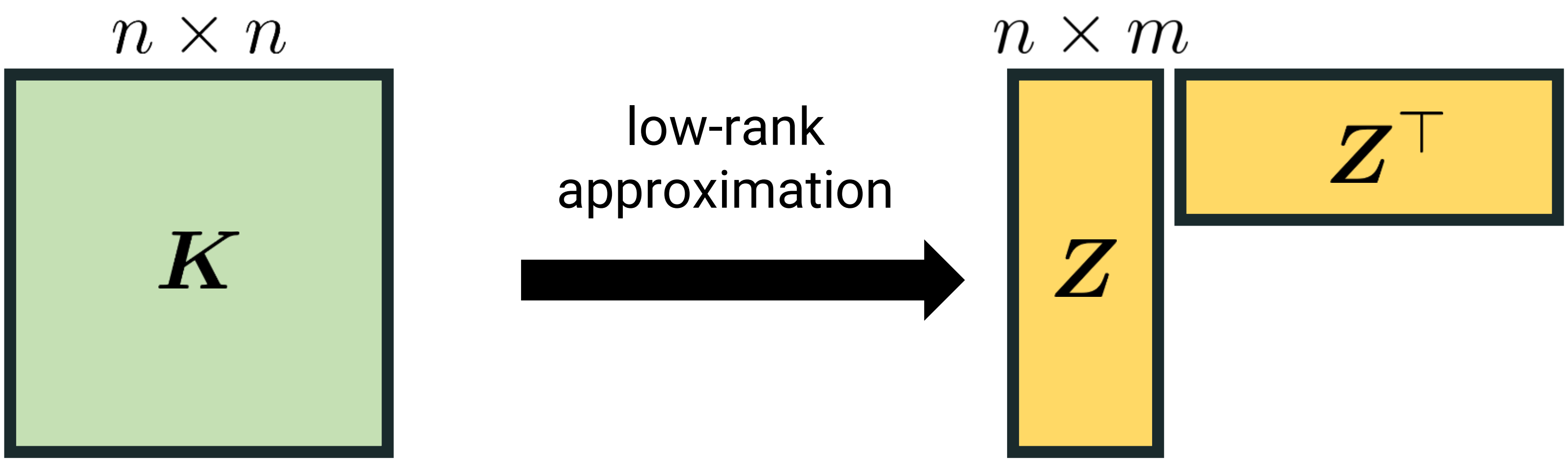
Scalability of Kernel Methods

- **Kernel methods are expensive**



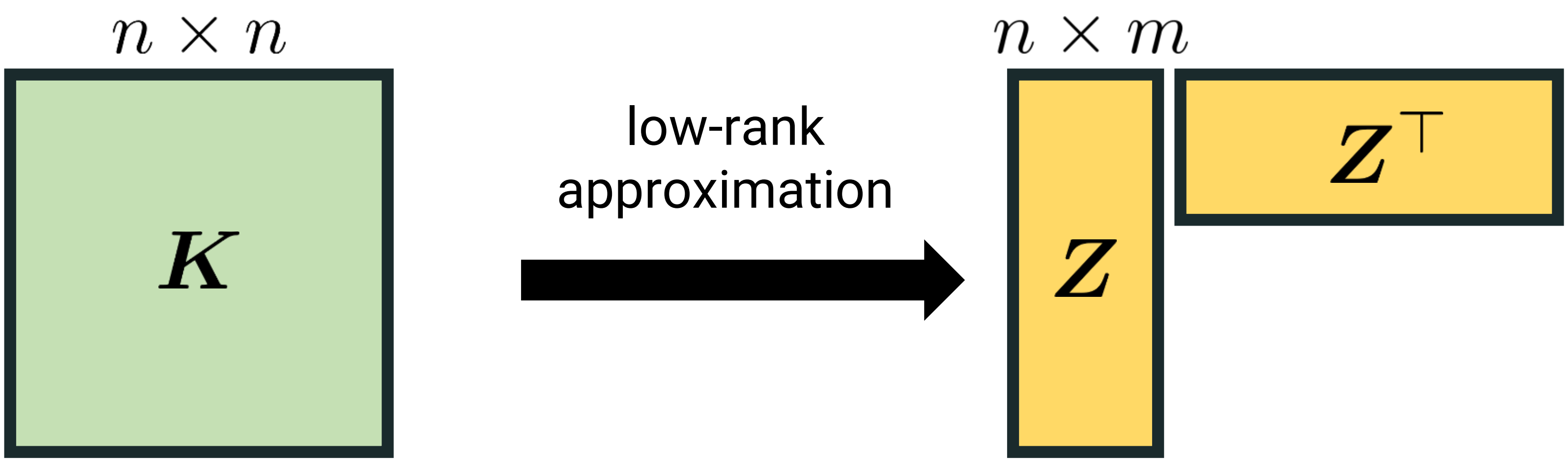
- Computing all kernel entries take $\Omega(n \cdot \mathrm{nnz}(\boldsymbol{X}) + n^2)$ time
- Even writing it down takes $\Omega(n^2 d)$ time and $\Omega(n^2)$ memory
- A single iteration of a linear system solver takes $\Omega(n^2)$ time
- For $n = 100{,}000$, $\boldsymbol{K}$ has 10 billion entries. Take 80 GB of storage!

Classical Solution: Dimensionality Reduction



- Storing $\boldsymbol{Z}$ uses $\mathcal{O}(nm)$ space and computing $\boldsymbol{Z}^\top \boldsymbol{Z} x$ takes $\mathcal{O}(nm)$ time
- Orthogonalization, eigen-decomposition and pseudo-inversion of $\boldsymbol{Z}\boldsymbol{Z}^\top$ all take just $\mathcal{O}(nm^2)$ time

Classical Solution: **Dimensionality Reduction**



- Storing $\boldsymbol{Z}$ uses $\mathcal{O}(nm)$ space and computing $\boldsymbol{Z}^\top \boldsymbol{Z} x$ takes $\mathcal{O}(nm)$ time
- Orthogonalization, eigen-decomposition and pseudo-inversion of $\boldsymbol{Z}\boldsymbol{Z}^\top$ all take just $\mathcal{O}(nm^2)$ time
- **Our approach:**
  - a low-rank approximation based on series expansion of Gegenbauer polynomials and their reproducing property

Overview of Our Contributions

- Extend the zonal kernels from $\mathbb{S}^{d-1}$ to $\mathbb{R}^d$ (that contains dot-product, Gaussian, Neural Tangent kernels) and derive the Mercer decomposition based on Gegenbauer polynomials

- Introduce random feature approach and provide spectral approximation (for kernel ridge regression) and projection-cost preserving approximation (for kernel k-mean clustering) guarantees

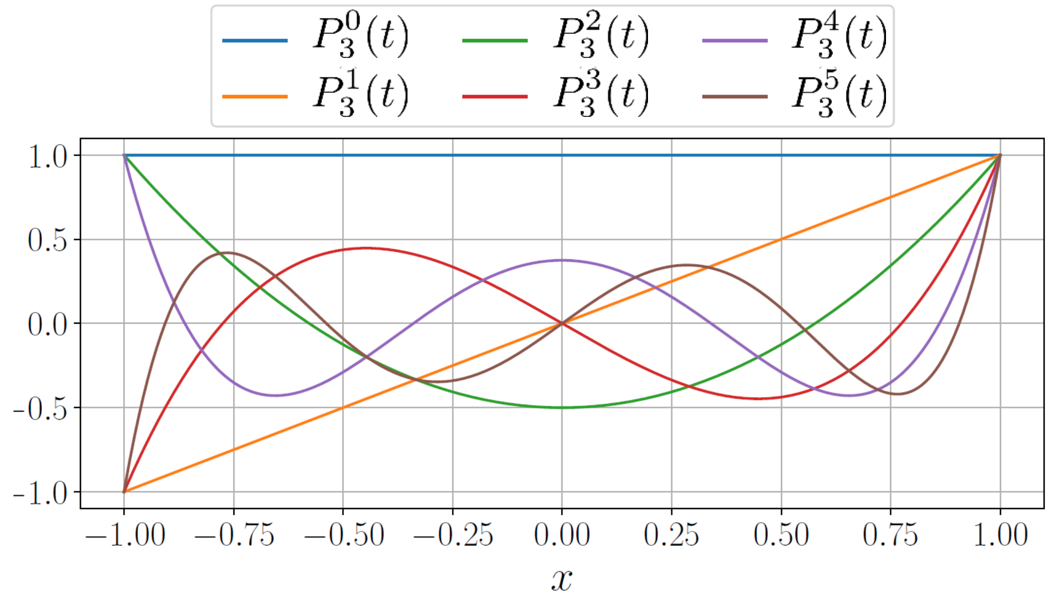
- Achieve the best sample complexity for spectrally approximating **Gaussian** kernel compared to the prior known methods when input dimension is small

Gegenbauer Harmonics

- **Gegenbauer polynomials** $\{P_d^\ell(\cdot)\}_{\ell\geq 0}$: a family of orthogonal polynomials

$$\int_{-1}^{1} P_d^\ell(t) P_d^m(t)(1-t^2)^{\frac{d-3}{2}}\, dt = \frac{\left|\mathbb{S}^{d-1}\right|}{\alpha_{\ell,d}\left|\mathbb{S}^{d-2}\right|} \cdot \mathbb{1}_{\{\ell=m\}}$$

  - d : dimension parameter, $\alpha_{\ell,d} = \binom{d+\ell-1}{\ell} - \binom{d+\ell-3}{\ell-2}$
  - $\left|\mathbb{S}^{d-1}\right|$: surface area of $\mathbb{S}^{d-1}$



Gegenbauer Harmonics

- **Reproducing property:** for any $x, y \in \mathbb{S}^{d-1}$

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

  - $\alpha_{\ell,d} = \binom{d+\ell-1}{\ell} - \binom{d+\ell-3}{\ell-2}$ (dimension of spherical harmonics)
  - $\mathcal{U}(\mathbb{S}^{d-1})$: uniform distribution over $\mathbb{S}^{d-1}$

Gegenbauer Harmonics

- **Reproducing property:** for any $x, y \in \mathbb{S}^{d-1}$

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

  - $\alpha_{\ell,d} = \binom{d+\ell-1}{\ell} - \binom{d+\ell-3}{\ell-2}$ (dimension of spherical harmonics)
  - $\mathcal{U}(\mathbb{S}^{d-1})$: uniform distribution over $\mathbb{S}^{d-1}$

- Gegenbauer polynomial kernel has a feature map as

$$\phi_x(w) = \sqrt{\alpha_{\ell,d}} \cdot P_d^\ell(\langle x, w\rangle)$$

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

Gegenbauer Harmonics

- **Reproducing property:** for any $x, y \in \mathbb{S}^{d-1}$

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

  - $\alpha_{\ell,d} = \binom{d+\ell-1}{\ell} - \binom{d+\ell-3}{\ell-2}$ (dimension of spherical harmonics)
  - $\mathcal{U}(\mathbb{S}^{d-1})$: uniform distribution over $\mathbb{S}^{d-1}$

- Gegenbauer polynomials can span all positive definite dot-product kernels on $\mathbb{S}^{d-1}$:

Theorem [Schoenberg, 1941]**.** Consider a function $\kappa(t) = \sum_{\ell=0}^{\infty} b_\ell P_\ell^d(t)$ and a kernel function $k(x, y) = \kappa(\langle x, y\rangle)$. The kernel k defined on $\mathbb{S}^{d-1}$ is positive definite if and only if $b_\ell \geq 0$.

Gegenbauer Harmonics

- **Reproducing property:** for any $x, y \in \mathbb{S}^{d-1}$

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

- **Zonal kernel:** $k : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to \mathbb{R}$ if $k(x, y) = \kappa(\langle x, y\rangle)$ for some $\kappa : \mathbb{R} \to \mathbb{R}$ ($\Leftrightarrow$ Dot-product kernels with a restriction of inputs)

Gegenbauer Harmonics

- **Reproducing property:** for any $x, y \in \mathbb{S}^{d-1}$

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

- **Zonal kernel:** $k : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to \mathbb{R}$ if $k(x, y) = \kappa(\langle x, y\rangle)$ for some $\kappa : \mathbb{R} \to \mathbb{R}$ (⇔ Dot-product kernels with a restriction of inputs)
  - Suppose the series expansion with Gegenbauer polynomials:

$$\kappa(t) = \sum_{\ell=0}^{\infty} c_\ell \cdot P_d^\ell(t)$$

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

(*κ with $\int_{-1}^{1} |\kappa(t)|^2 \left(1-t^2\right)^{\frac{d-3}{2}} dt < \infty$ has the unique series expansion)

Gegenbauer Harmonics

- **Reproducing property:** for any $x, y \in \mathbb{S}^{d-1}$

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

- **Zonal kernel:** $k : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to \mathbb{R}$ if $k(x,y) = \kappa(\langle x, y\rangle)$ for some $\kappa : \mathbb{R} \to \mathbb{R}$ (⇔ Dot-product kernels with a restriction of inputs)
  - Suppose the series expansion with Gegenbauer polynomials:

$$\kappa(t) = \sum_{\ell=0}^{\infty} c_\ell \cdot P_d^\ell(t)$$

  - The feature map of zonal kernels:

For positive semidefinite kernel, $c_\ell \geq 0$

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

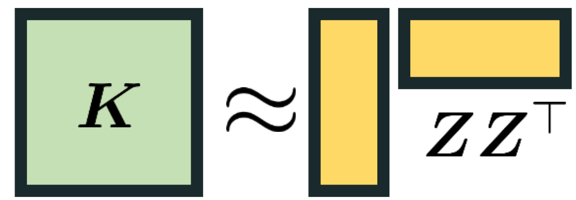
such that $\mathbb{E}_{w\sim\mathcal{U}(\mathbb{S}^{d-1})}\left[\phi_x(w) \cdot \phi_y(w)\right] = \kappa(\langle x, y\rangle)$

Gegenbauer Harmonics

- **Feature map of zonal kernel:** for $\kappa(t) = \sum_{\ell=0}^{\infty} c_\ell \cdot P_d^\ell(t)$ and $x, y \in \mathbb{S}^{d-1}$

$$\phi_x(w) = \sum_{\ell=0}^{\infty} \sqrt{c_\ell \alpha_{\ell,d}} P_d^\ell(\langle x, w\rangle) \quad \Rightarrow \quad \mathop{\mathbb{E}}_{w \sim \mathcal{U}(\mathbb{S}^{d-1})} [\phi_x(w)\phi_y(w)] = \kappa(\langle x, y\rangle)$$

- **Goal:** design a low-rank kernel approximation
  - Given $x_1, \dots, x_n \in \mathbb{S}^{d-1}$, draw i.i.d. $w_1, \dots, w_m \sim \mathcal{U}(\mathbb{S}^{d-1})$ and compute



$$[\boldsymbol{K}]_{ij} = \kappa(\langle x_i, x_j\rangle)$$

$$\boldsymbol{Z} = \begin{bmatrix} \phi_{x_1}(w_1) & \cdots & \phi_{x_1}(w_m) \\ \vdots & \ddots & \vdots \\ \phi_{x_n}(w_1) & \cdots & \phi_{x_n}(w_m) \end{bmatrix} \quad \Rightarrow \quad \mathbb{E}\left[\boldsymbol{Z}\boldsymbol{Z}^\top\right] = \boldsymbol{K}$$

Gegenbauer Harmonics

- **Feature map of zonal kernel:** for $\kappa(t) = \sum_{\ell=0}^{\infty} c_\ell \cdot P_d^\ell(t)$ and $x, y \in \mathbb{S}^{d-1}$

$$\phi_x(w) = \sum_{\ell=0}^{\infty} \sqrt{c_\ell \alpha_{\ell,d}} P_d^\ell(\langle x, w \rangle) \quad \Rightarrow \quad \mathop{\mathbb{E}}_{w \sim \mathcal{U}(\mathbb{S}^{d-1})} [\phi_x(w) \phi_y(w)] = \kappa(\langle x, y \rangle)$$

- **Goal:** design a low-rank kernel approximation
  - Given $x_1, \dots, x_n \in \mathbb{S}^{d-1}$, draw i.i.d. $w_1, \dots, w_m \sim \mathcal{U}(\mathbb{S}^{d-1})$ and compute

$$[\boldsymbol{K}]_{ij} = \kappa(\langle x_i, x_j \rangle)$$

$$\boldsymbol{Z} = \begin{bmatrix} \phi_{x_1}(w_1) & \cdots & \phi_{x_1}(w_m) \\ \vdots & \ddots & \vdots \\ \phi_{x_n}(w_1) & \cdots & \phi_{x_n}(w_m) \end{bmatrix} \quad \Rightarrow \quad \mathbb{E}\left[\boldsymbol{Z}\boldsymbol{Z}^\top\right] = \boldsymbol{K}$$

- **Challenge:** Can we extend zonal kernel functions to $\mathbb{R}^d$?

Generalized Zonal Kernel

- **Generalized zonal kernel:** for any $x, y \in \mathbb{R}^d$ and $h_\ell : \mathbb{R} \to \mathbb{R}^s$ for $\ell = 0, 1, \ldots$

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

Generalized Zonal Kernel

- **Generalized zonal kernel:** for any $x, y \in \mathbb{R}^d$ and $h_\ell : \mathbb{R} \to \mathbb{R}^s$ for $\ell = 0, 1, \ldots$

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

  - This includes all dot-product kernels, i.e., $k(x, y) = \kappa(\langle x, y \rangle), x, y \in \mathbb{R}^d$

Lemma. For any $x, y \in \mathbb{R}^d$ and an analytic function $\kappa : \mathbb{R} \to \mathbb{R}$, define

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

Then,

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

Generalized Zonal Kernel

- **Generalized zonal kernel:** for any $x, y \in \mathbb{R}^d$ and $h_\ell : \mathbb{R} \to \mathbb{R}^s$ for $\ell = 0, 1, \dots$

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

- **Feature map of generalized zonal kernel:**

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

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

  - When $\|x\| = 1$, this falls into the feature map of zonal kernel

Generalized Zonal Kernel

- **Generalized zonal kernel:** for any $x, y \in \mathbb{R}^d$ and $h_\ell : \mathbb{R} \to \mathbb{R}^s$ for $\ell = 0, 1, \dots$

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

- **Random features of generalized zonal kernel:**
  - Given $x_1, \dots, x_n \in \mathbb{R}^d$, draw i.i.d. $w_1, \dots, w_m \sim \mathcal{U}(\mathbb{S}^{d-1})$ and compute

$$\boldsymbol{Z} = \begin{bmatrix} \phi_{x_1}(w_1) & \cdots & \phi_{x_1}(w_m) \\ \vdots & \ddots & \vdots \\ \phi_{x_n}(w_1) & \cdots & \phi_{x_n}(w_m) \end{bmatrix} \Rightarrow \begin{array}{c} \mathbb{E}\left[\boldsymbol{Z}\boldsymbol{Z}^\top\right] = \boldsymbol{K} \\ [\boldsymbol{K}]_{ij} = k(x_i, x_j) \end{array}$$

Generalized Zonal Kernel

- **Generalized zonal kernel:** for any $x, y \in \mathbb{R}^d$ and $h_\ell : \mathbb{R} \to \mathbb{R}^s$ for $\ell = 0, 1, \dots$

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

- **Random features of generalized zonal kernel:**
  - Given $x_1, \dots, x_n \in \mathbb{R}^d$, draw i.i.d. $w_1, \dots, w_m \sim \mathcal{U}(\mathbb{S}^{d-1})$ and compute

$$\boldsymbol{Z} = \begin{bmatrix} \phi_{x_1}(w_1) & \cdots & \phi_{x_1}(w_m) \\ \vdots & \ddots & \vdots \\ \phi_{x_n}(w_1) & \cdots & \phi_{x_n}(w_m) \end{bmatrix} \quad \Rightarrow \quad \begin{array}{c} \mathbb{E}\left[\boldsymbol{Z}\boldsymbol{Z}^\top \right] = \boldsymbol{K} \\ [\boldsymbol{K}]_{ij} = k(x_i, x_j) \end{array}$$

- **Goal:** how many random vectors are needed? (lower bound on m)

Generalized Zonal Kernel

- **Spectral approximation of GZK:**

Theorem. For any $0 < \lambda < \|\boldsymbol{K}\|_{\mathrm{op}}$, let $s_\lambda := \mathrm{Tr}(\boldsymbol{K}(\boldsymbol{K} + \lambda \boldsymbol{I})^{-1}))$. For any $\delta, \varepsilon > 0$, if

$$m \geq \frac{8}{3\varepsilon^2} \log \frac{16 s_\lambda}{\delta} \cdot \sum_{\ell=0}^{\infty} \alpha_{\ell,d} \min \left\{ \frac{\pi^2(\ell+1)^2}{6\lambda} \sum_{j \in [n]} \|h_\ell(\|x_j\|)\|^2, s \right\}$$

Then,

$$\Pr\left[(1-\varepsilon)(\boldsymbol{Z}\boldsymbol{Z}^\top + \lambda \boldsymbol{I}) \preceq \boldsymbol{K} + \lambda \boldsymbol{I} \preceq (1+\varepsilon)(\boldsymbol{Z}\boldsymbol{Z}^\top + \lambda \boldsymbol{I})\right] \geq 1 - \delta$$

  - Spectral approximation can directly guarantee empirical risk bound of kernel ridge regression

Generalized Zonal Kernel

- **Projection-cost preserving approximation of GZK:**

Theorem. For any $0 < \lambda < \|\boldsymbol{K}\|_{\text{op}}$, let $s_\lambda := \mathrm{Tr}(\boldsymbol{K}(\boldsymbol{K} + \lambda \boldsymbol{I})^{-1}))$. For any positive integer r, let $\lambda := \frac{1}{r}\sum_{i=r+1}^{n} \lambda_i$ where $\lambda_1 \geq \cdots \geq \lambda_n$ are eigenvalues of $\boldsymbol{K}$. For all rank-r orthonormal projection matrices $\boldsymbol{P} \in \mathbb{R}^{n \times n}$ and for any $\delta, \varepsilon > 0$ if

$$m \geq \frac{8}{3\varepsilon^2} \log \frac{16 s_\lambda}{\delta} \cdot \sum_{\ell=0}^{\infty} \alpha_{\ell,d} \min \left\{ \frac{\pi^2(\ell+1)^2}{6\lambda} \sum_{j \in [n]} \left\| h_\ell(\|x_j\|) \right\|^2, s \right\}$$

Then,

$$\Pr\left[(1-\varepsilon) \leq \frac{\mathrm{Tr}(\boldsymbol{Z}\boldsymbol{Z}^\top - \boldsymbol{P}\boldsymbol{Z}\boldsymbol{Z}^\top \boldsymbol{P})}{\mathrm{Tr}(\boldsymbol{K} - \boldsymbol{P}\boldsymbol{K}\boldsymbol{P})} \leq (1+\varepsilon) \right] \geq 1 - \delta$$

- Projection cost preserving approximation can be used for kernel k-means clustering, principal component analysis (PCA)

Gaussian Kernel

- **Spectral approximation of the Gaussian kernels:**

Theorem. Given $x_1, \ldots, x_n \in \mathbb{R}^d$, assume that $\max_{i\in[n]} \|x_i\| \le r$. Let $\boldsymbol{K} \in \mathbb{R}^{n\times n}$ and $[\boldsymbol{K}]_{ij} = \exp(-\|x_i - x_j\|^2/2)$. For $0 < \lambda < \|\boldsymbol{K}\|_{\mathrm{op}}$, let $s_\lambda := \mathtt{tr}(\boldsymbol{K}(\boldsymbol{K} + \lambda \boldsymbol{I})^{-1}))$ and for any $\delta, \varepsilon > 0$, if

$$m = \Omega\left(\frac{\left(2\log\frac{n}{\lambda}\right)^d + (1.93r)^{2d}}{(d-1)!}\right)$$

Then,

$$\Pr\left[(1-\varepsilon)(\boldsymbol{Z}\boldsymbol{Z}^\top + \lambda\boldsymbol{I}) \preceq \boldsymbol{K} + \lambda\boldsymbol{I} \preceq (1+\varepsilon)(\boldsymbol{Z}\boldsymbol{Z}^\top + \lambda\boldsymbol{I})\right] \ge 1-\delta$$

Gaussian Kernel

- **Comparison to prior results**

Method	Feature dimension (m)	Runtime
Gegenbauer features (Our work)	$\frac{\left(2\log\frac{n}{\lambda}\right)^d + (1.93r)^{2d}}{(d-1)!}$	$m \cdot \mathrm{nnz}(\boldsymbol{X})$
Fourier features (RR'07)	$\frac{n}{\lambda}$	$m \cdot \mathrm{nnz}(\boldsymbol{X})$
Modified Fourier features (AKMMVZ'17)	$(248r)^d\left(\log\frac{n}{\lambda}\right)^{\frac{d}{2}} + \left(200\log\frac{n}{\lambda}\right)^d$	$m \cdot \mathrm{nnz}(\boldsymbol{X})$
PolySketch (AKKPVWZ'20)	$r^{10}s_\lambda$	$r^{12}(s_\lambda n + \mathrm{nnz}(\boldsymbol{X}))$
Adaptive Sketch (WZ'20)	s_λ	$r^{15}s_\lambda n + r^5\mathrm{nnz}(\boldsymbol{X})$

$d = o\left(\log\frac{n}{\lambda}\right)$

Gaussian Kernel

- **Comparison to prior results**

Method	Feature dimension (m)	Runtime
Gegenbauer features (Our work)	$\frac{\left(2\log\frac{n}{\lambda}\right)^d + (1.93r)^{2d}}{(d-1)!}$	$m \cdot \mathrm{nnz}(\boldsymbol{X})$
Fourier features (RR'07)	$\frac{n}{\lambda}$	$m \cdot \mathrm{nnz}(\boldsymbol{X})$
Modified Fourier features (AKMMVZ'17)	$(248r)^d\left(\log\frac{n}{\lambda}\right)^{\frac{d}{2}} + \left(200\log\frac{n}{\lambda}\right)^d$	$m \cdot \mathrm{nnz}(\boldsymbol{X})$
PolySketch (AKKPVWZ'20)	r $r = o\left(\sqrt{\log\frac{n}{\lambda}}\right)$	$r^{12}(s_\lambda n + \mathrm{nnz}(\boldsymbol{X}))$
Adaptive Sketch (WZ'20)	s_λ	$r^{15}s_\lambda n + r^5\mathrm{nnz}(\boldsymbol{X})$

Gaussian Kernel

- **Comparison to prior results**

Method	Feature dimension (m)	Runtime
Gegenbauer features (Our work)	$\frac{\left(2\log\frac{n}{\lambda}\right)^d + (1.93r)^{2d}}{(d-1)!}$	$m \cdot \mathrm{nnz}(\boldsymbol{X})$
Fourier features (RR'07)	$\frac{n}{\lambda}$	$m \cdot \mathrm{nnz}(\boldsymbol{X})$
Modified Fourier features (AKMMVZ'17)	$(248r)^d\left(\log\frac{n}{\lambda}\right)^{\frac{d}{2}} + \left(200\log\frac{n}{\lambda}\right)^d$	$m \cdot \mathrm{nnz}(\boldsymbol{X})$
PolySketch (AKKPVWZ'20)	$r^{10}s_\lambda$	$r^{12}(s_\lambda n + \mathrm{nnz}(\boldsymbol{X}))$
Adaptive Sketch (WZ'20)	s_λ	$r^{15}s_\lambda n + r^5\mathrm{nnz}(\boldsymbol{X})$

$\boxed{d = \mathcal{O}(1)}$

Gaussian Kernel

- **Comparison to prior results**

Method	Feature dimension (m)	Runtime
Gegenbauer features (Our work)	$\frac{\left(2\log\frac{n}{\lambda}\right)^d + (1.93r)^{2d}}{(d-1)!}$	$m \cdot \mathrm{nnz}(\boldsymbol{X})$
Fourier features (RR'07)	$\frac{n}{\lambda}$	$m \cdot \mathrm{nnz}(\boldsymbol{X})$
Modified Fourier features (AKMMVZ'17)	$(248r)^d \left(\log\frac{n}{\lambda}\right)^{\frac{d}{2}} + \left(200\log\frac{n}{\lambda}\right)^d$	$m \cdot \mathrm{nnz}(\boldsymbol{X})$
PolySketch (AKKPVWZ'20)	$r^{10} s_\lambda$	$r^{12}(s_\lambda n + \mathrm{nnz}(\boldsymbol{X}))$
Adaptive Sketch (WZ'20)	s_λ	$r^{15} s_\lambda n + r^5 \mathrm{nnz}(\boldsymbol{X})$

$d = \mathcal{O}(1)$

Experiments

- **Kernel ridge regression with Gaussian kernel**
  - Random Gegenbauer features achieve the best MSE except "Elevation" and "Protein" datasets
  - For "Protein" dataset (larger d), Nystrom method is the best

	Elevation		CO_2		Climate		Protein	
n	64,800		146,040		223,656		45,730	
d	3		4		4		9	
	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
Ours	1.15	1.71	**0.532**	3.49	**3.13**	5.41	21.0	9.72

MSE of kernel ridge regression and runtime for kernel approximation

Experiments

- **Kernel k-means clustering with Gaussian kernel**
  - Random Gegenbauer features show the promising performance except "Mushroom" and "Connect-4" datasets which have a higher input dimension

	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
Ours	**0.35**	**0.40**	0.71	**0.59**	**0.21**	0.78

The average sum of squared distance to the nearest cluster centers

Conclusion

- **Summary:**
  - We study a new class of kernels expressed by Gegenbauer polynomials that covers a wide range of ubiquitous kernels
  - We analyze that our 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
- **Future work:**
  - **Our limitation can be resolved by combining with additional dimensionality reductions (e.g., JL-transform)**