IJCNN 2021

Towards Unbiased Random Features with Lower Variance For Stationary Indefinite Kernels

Qin Luo, Kun Fang, Jie Yang, Xiaolin Huang

Institute of Image Processing and Pattern Recognition
Shanghai Jiao Tong University

INNS - International Neural Network Society

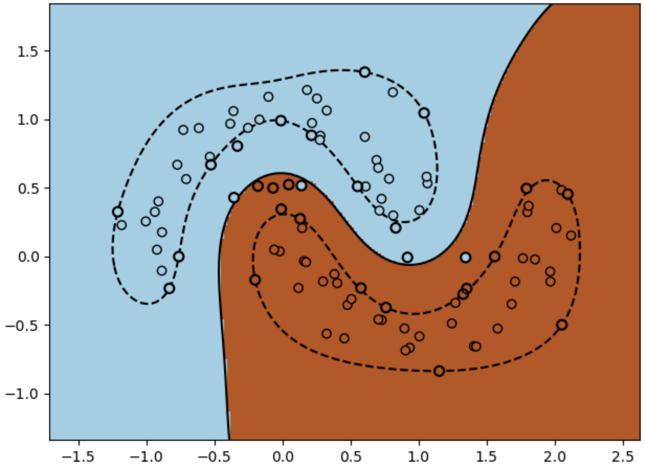
IEEE - Computational Intelligence Society

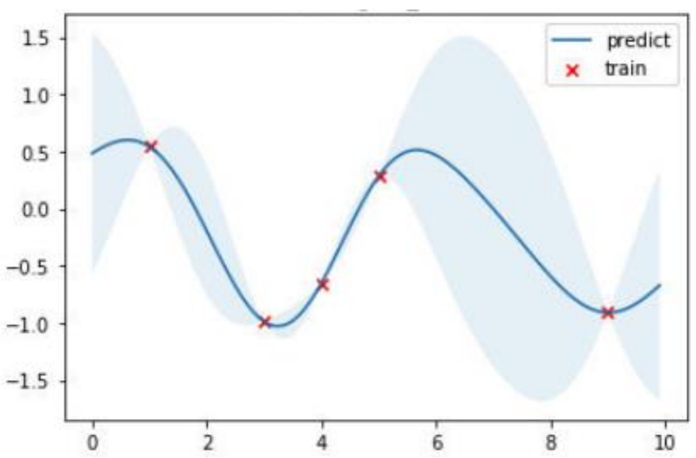
Contents

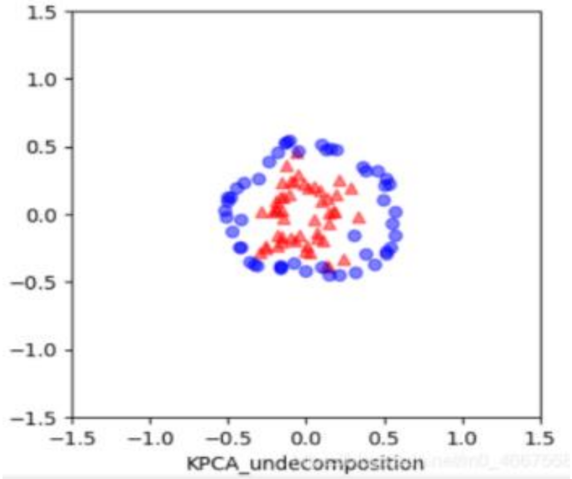
Contents

Background

- Kernel methods are extensively used in classification [1], regression [2] and dimension reduction [3].
- Kernel methods scale poorly to large datasets because of $\mathcal{O}(N^3)$ time complexity and $\mathcal{O}(N^2)$ space complexity
- Random Fourier Features reduce the computation cost and storage space to $\mathcal{O}(Ns^2)$ and $\mathcal{O}(Ns)$ $(s \ll N)$







[1] Schölkopf B, Smola A J, Bach F. Learning with kernels: support vector machines, regularization, optimization, and beyond[M]. MIT press, 2002.
[2] Wilson A, Adams R. Gaussian process kernels for pattern discovery and extrapolation[C]//International conference on machine learning. PMLR, 2013: 1067-1075.
[3] Schölkopf B, Smola A, Müller K R. Kernel principal component analysis[C]//International conference on artificial neural networks. Springer, Berlin, Heidelberg, 1997: 583-588.

Background

- Random Fourier Features are restricted to the kernels:
  1) **shift-invariant (stationary)**

$$k(\boldsymbol{x}, \boldsymbol{y}) = k(\boldsymbol{x} - \boldsymbol{y})$$

  2) **positive definite (PD)**

$$\boldsymbol{\alpha}^T \boldsymbol{K} \boldsymbol{\alpha} \geq 0 \text{ for all } \boldsymbol{\alpha} \neq \mathbf{0} \text{ and } K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

- Not satisfy the requirement
  1) **non-stationary kernels**: polynomial kernel, neural tangent kernel (NTK)
     When data is restricted on the sphere → stationary but indefinite kernel
  2) **non-PD kernel**: linear combination of Gaussian kernel (Delta-Gaussian kernel), TL1-kernel

(Bohner's Theorem) A continuous and **stationary** function $k\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is **positive definite** if and only if it can be represented as

$$k(\boldsymbol{x} - \boldsymbol{y}) = \int_{\mathbb{R}^d} \exp\left(i\boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{y})\right) p(d\boldsymbol{w}) = \mathbb{E}_{\boldsymbol{w} \sim p(\boldsymbol{w})}\left[\exp\left(\mathrm{i}\boldsymbol{w}^{\boldsymbol{T}}(\boldsymbol{x} - \boldsymbol{y})\right)\right]$$

where p(**w**) is the positive finite measure over **w**, i is imaginary unit.

Background

Related Work

Methods	Kernel Types	Unbiasedness	Variance
Random Maclaurin (RM) [1]	Polynomial	√	$\mathcal{O}(\left(\frac{32RL}{\epsilon}\right)^{2d}\exp\left(-\frac{D\epsilon^2}{8C_\Omega^2}\right))$
Tensor Sketch (TS) [2]	Polynomial	√	$\mathcal{O}(\exp\left(-\frac{t\epsilon^2}{2R^{4p}}\right))$
Spherical Random Features (SRF) [3]	Stationary Indefinite	×	________________
Double Variation Random Features (DIGMM) [4]	Stationary Indefinite	×	________________
Generalized Random Fourier Features (GRFF) [5]	Stationary Indefinite	√	$\mathcal{O}(\left(\frac{2\sigma R}{\epsilon}\right)^{2d}\exp\left(-\frac{s\epsilon^2}{32(d+2)}\right))$

[1] Kar P, Karnick H. Random feature maps for dot product kernels[C]//Artificial intelligence and statistics. PMLR, 2012: 583-591.
[2] Pham N, Pagh R. Fast and scalable polynomial kernels via explicit feature maps[C]//Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. 2013: 239-247.
[3] Pennington J, Felix X Y, Kumar S. Spherical Random Features for Polynomial Kernels[C]//NIPS. 2015.
[4] Liu F, Huang X, Shi L, et al. A double-variational bayesian framework in random fourier features for indefinite kernels[J]. IEEE transactions on neural networks and learning systems, 2019, 31(8): 2965-2979.
[5] Liu F, Huang X, Chen Y, et al. Fast Learning in Reproducing Kernel Krein Spaces via Signed Measures[C]//International Conference on Artificial Intelligence and Statistics. PMLR, 2021: 388-396.

Contents

- Background
- **Motivation**
- Methods
- Experiments
- Conclusion

Motivation

Objective: Unbiased random Fourier approximation with lower variance for stationary indefinite kernels

Contribution:

- Unbiased approximation and lower the variance utilizing orthogonal sampling
- Theoretical analysis of the unbiasedness and variance reduction
- Experimental validation of the approximation error and classification or regression performance compared with the existing approximation method

Contents

Methods

Preliminaries

(Signed Measure) Let Ω be some set, and $\mathcal{A}$ be a σ-algebra of subsets on Ω. A signed measure is a function $\mu: \mathcal{A} \to [-\infty, +\infty)$ or $(-\infty, +\infty]$ satisfying σ-additivity.

(Jordan Decomposition) Let μ be a signed measure defined on the σ-algebra $\mathcal{A}$. There exists two nonnegative measures μ_+ and μ_- (one of the measure) such that $\mu = \mu_+ - \mu_-$. The total mass is defined as: $||\mu|| = ||\mu_+|| + ||\mu_-||$

Methods

$p(\boldsymbol{w})$ is not a probability measure, viewed as a signed measure.

$$k(\boldsymbol{x}-\boldsymbol{y}) = k(\boldsymbol{z}) = \int_{\mathbb{R}^d} \exp(i\boldsymbol{w}^T\boldsymbol{z})\, p(d\boldsymbol{w}) = \int_{\mathbb{R}^d} \exp(i\boldsymbol{w}^T\boldsymbol{z})\, p_+(d\boldsymbol{w}) - \int_{\mathbb{R}^d} \exp(i\boldsymbol{w}^T\boldsymbol{z})\, p_-(d\boldsymbol{w})$$

$$= ||p_+||\mathbb{E}_{\boldsymbol{w}\sim\widetilde{p_+}(\boldsymbol{w})}\left(\exp\left(\mathrm{i}\mathbf{w}^{\mathrm{T}}\mathbf{z}\right)\right) - ||p_-||\mathbb{E}_{\boldsymbol{w}\sim\widetilde{p_-}(\boldsymbol{w})}\left(\exp\left(\mathrm{i}\mathbf{w}^{\mathrm{T}}\mathbf{z}\right)\right)$$

where $\widetilde{p_+} = \frac{p_+(\boldsymbol{w})}{||p_+||}$ and $\widetilde{p_-} = \frac{p_-(\boldsymbol{w})}{||p_-||}$

Define $\phi(\boldsymbol{x}) = \frac{1}{\sqrt{s}}[\psi_1(\boldsymbol{x}), \psi_2(\boldsymbol{x}), \psi_3(\boldsymbol{x}), \dots, \psi_s(\boldsymbol{x})]^T$ with $\psi_i(\boldsymbol{x})$:

$$\psi_i(\boldsymbol{x}) = \left[\sqrt{||p_+||}\cos\left(\boldsymbol{w}_i^T\boldsymbol{x}\right), \sqrt{||p_+||}\sin\left(\boldsymbol{w}_i^T\boldsymbol{x}\right), \mathrm{i}\sqrt{||p_-||}\cos\left(\boldsymbol{v}_i^T\boldsymbol{x}\right), \mathrm{i}\sqrt{||p_-||}\sin\left(\boldsymbol{v}_i^T\boldsymbol{x}\right)\right]^T$$

→ Generalized Random Fourier Features

$$k(\boldsymbol{x}-\boldsymbol{y}) \approx \frac{1}{s}\sum_{i=1}^{s} < \psi_i(\boldsymbol{x}), \psi_i(\boldsymbol{y}) > = \phi(\boldsymbol{x})^T\phi(\boldsymbol{y})$$

Methods

Unbiased Approximation

$$\mathbb{E}\left(K_{GRFF}(\boldsymbol{z})\right) = k(\boldsymbol{z})$$

Variance determines the whole approximation error. Orthogonal sampling could reduce the variance.

QR decomposition

1) Amplitude sampling $\left|\left|\boldsymbol{w}_i\right|\right|_2 \sim \widetilde{p_+(\boldsymbol{w})}$ $\left|\left|\boldsymbol{v}_i\right|\right|_2 \sim \widetilde{p_-(\boldsymbol{w})}$

2) Orthogonal direction $\boldsymbol{a}_j \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{2m})$ $\boldsymbol{b}_{\boldsymbol{j}} \sim \boldsymbol{\mathcal{N}}(\boldsymbol{0}, \boldsymbol{I}_{\boldsymbol{2m}})$ $\boldsymbol{M} = [\boldsymbol{a}_{\boldsymbol{1}}, \dots, \boldsymbol{a}_{\boldsymbol{m}}, \boldsymbol{b}_{\boldsymbol{1}}, \dots, \boldsymbol{b}_{\boldsymbol{m}}]$

$M^{orth} = QR(M)$

3) Composition $\boldsymbol{w}_i = \left|\left|\boldsymbol{w}_i\right|\right|_2 M_i^{orthn}$ $\boldsymbol{v}_i = \left|\left|\boldsymbol{v}_i\right|\right|_2 M_{s+i}^{orthn}$

Methods

Variance Reduction

Orthogonality on W_{pos}

Mutual orthogonality on W_{pos} and W_{neg}

$$Var\left(K_{GORF}(\boldsymbol{z})\right) - Var\left(K_{GRFF}(\boldsymbol{z})\right) = \boxed{||p_+||^2 G_{\widetilde{k_+}}(\boldsymbol{z})} + \boxed{||p_-||^2 G_{\widetilde{k_-}}(\boldsymbol{z})} + \boxed{H(\boldsymbol{z})}$$

Orthogonality on W_{neg}

Theorem 5 [1] For a PD radial kernel k on $\mathbb{R}^d$ with Fourier measure $p(\boldsymbol{w})$ and $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, writing $\boldsymbol{z} = \boldsymbol{x} - \boldsymbol{y}$, we have:

$$G_k(\boldsymbol{z}) = Var\left(K_{ORF}(\boldsymbol{z})\right) - Var\left(K_{RFF}(\boldsymbol{z})\right) = \frac{s-1}{s}\mathbb{E}_{R_1}\left[\frac{J_{\frac{d}{2}-1}(R_1||\boldsymbol{z}||)\Gamma\left(\frac{d}{2}\right)}{\left(\frac{R_1||\boldsymbol{z}||}{2}\right)^{\frac{d}{2}-1}}\right]^2 - \frac{s-1}{s}\mathbb{E}_{R_1,R_2}\left[\frac{J_{\frac{d}{2}-1}\left(\sqrt{R_1^2+R_2^2}||\boldsymbol{z}||\right)\Gamma\left(\frac{d}{2}\right)}{\left(\frac{\sqrt{R_1^2+R_2^2}||\boldsymbol{z}||}{2}\right)^{\frac{d}{2}-1}}\right]^2$$

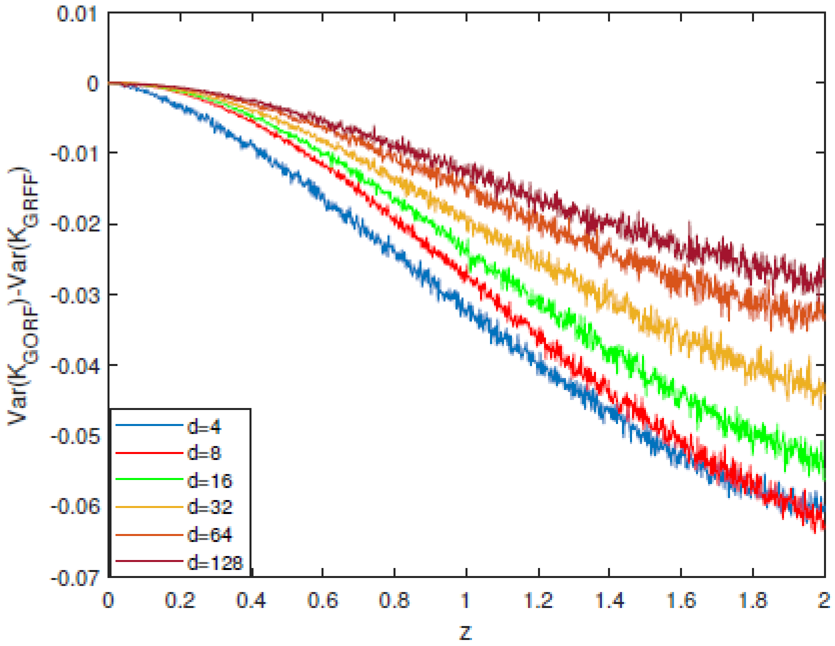
where $R_1, R_2 \sim p(\boldsymbol{w})$, and J_α is the Bessel function of the first kind of degree α.

$$H(\boldsymbol{z}) = 2||p_+||||p_-||[\mathbb{E}(a_1)\mathbb{E}(b_1) - \mathbb{E}(a_1 b_1)], \qquad a_1 = \cos(\boldsymbol{w}_1^T\boldsymbol{z}), \qquad b_1 = \cos(\boldsymbol{v}_1^T\boldsymbol{z})$$

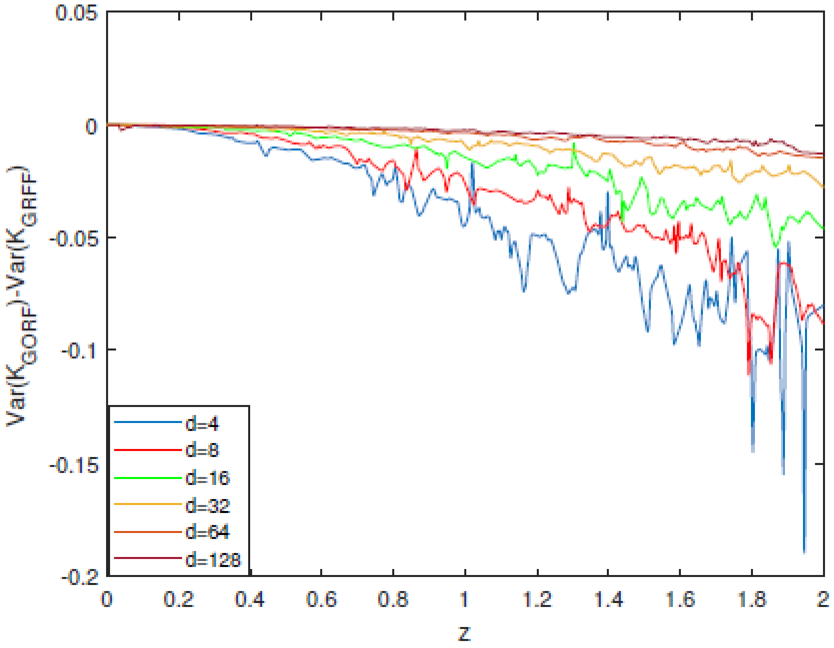
[1] Choromanski K, Rowland M, Sarlos T, et al. The geometry of random features[C]//International Conference on Artificial Intelligence and Statistics. PMLR, 2018: 1-9.

Methods

Variance Reduction



(a) polynomial kernel on the unit sphere ($a = 3, m = 1$)



(b) delta-gaussian kernel ($a_1 = 1, a_2 = -1, \sigma_1 = 1, \sigma_2 = 10$)

Contents

Experiments

Setup

1) Kernels

Polynomial kernel

$$k(\boldsymbol{x},\boldsymbol{y}) = \alpha(q + <\boldsymbol{x},\boldsymbol{y}>)^m = \left(1 - \frac{||\boldsymbol{x}-\boldsymbol{y}||^2}{a^2}\right)^m$$

where $q = a^2/2 - 1$, $\alpha = (2/\alpha^2)^m$

Delta-Gaussian kernel

$$k(\boldsymbol{x},\boldsymbol{y}) = \sum_{i=1}^{m} \alpha_i e^{-\frac{||\boldsymbol{x}-\boldsymbol{y}||^2}{2\sigma_i^2}}$$

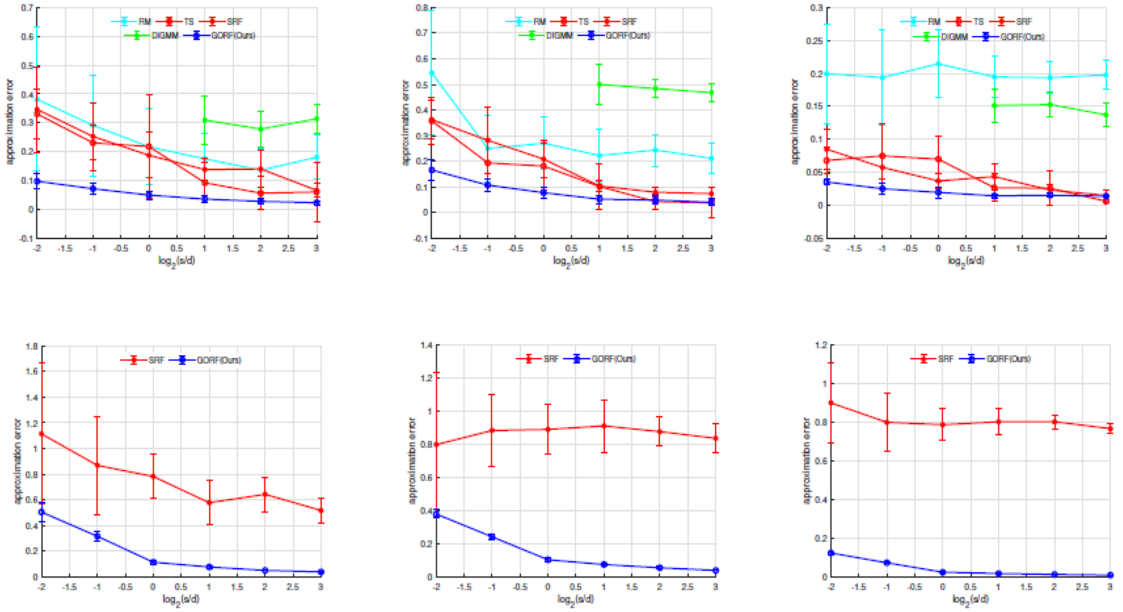
2) Datasets

Datasets	**d**	**training**	**testing**
letter	16	12000	6000
ijcnn1	22	49990	91701
usps	256	7291	2007

housing: d=13, training=405, testing=101

Experiments

Approximation error



(a) *letter* (b) *ijcnn1* (c) *usps*

Fig. 2. Comparisons of various algorithms for kernel approximation in terms of approximation error across two typical stationary indefinite kernels and three datasets with different dimensions. Top: polynomial kernel on the unit sphere. Below: delta-gaussian kernel

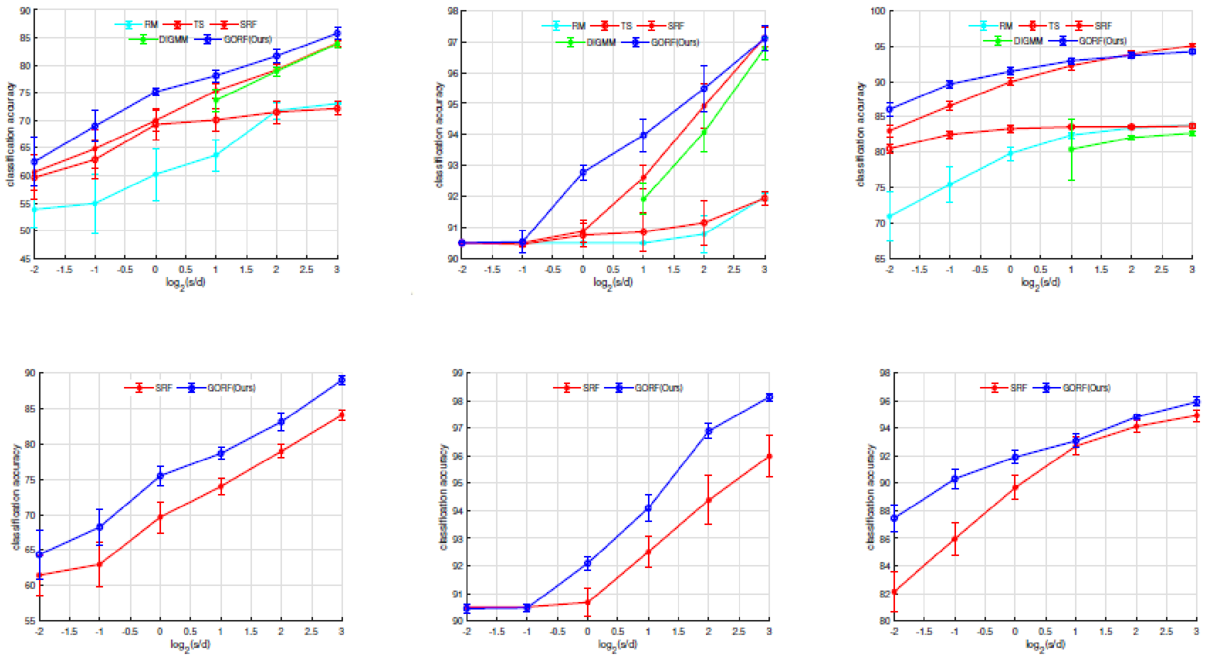
TABLE II
COMPARISON RESULTS BETWEEN APPLYING ORTHOGONAL SAMPLING AND I.I.D SAMPLING ON STATIONARY INDEFINITE KERNELS IN TERMS OF APPROXIMATION ERROR (MEAN±STD.). THE LOWEST ERROR IS HIGHLIGHTED IN **BOLDFACE**

Kernel	DataSet	Method	s=1/2d	s=d	s=2d	s=8d
polynomial	*letter*	*GRFF*	0.0859±0.0309	0.0547±0.0078	0.0469±0.0109	0.0261±0.0059
		GORF	**0.0716±0.0175**	**0.0495±0.0139**	**0.0360±0.0110**	**0.0231±0.0078**
	ijcnn1	*GRFF*	0.1159±0.0158	0.0907±0.0194	0.0794±0.0142	0.0433±0.0059
		GORF	**0.1072±0.0228**	**0.0775±0.0204**	**0.0487±0.0155**	**0.0397±0.0135**
	usps	*GRFF*	0.0270±0.0056	0.0213±0.0063	0.0160±0.0029	0.0137±0.0020
		GORF	**0.0251±0.0078**	**0.0194±0.0087**	**0.0143±0.0030**	**0.0137±0.0016**
delta-gaussian	*letter*	*GRFF*	0.3918± 0.0428	0.2736±0.0345	0.1887±0.0201	0.1017±0.0088
		GORF	**0.3154±0.0424**	**0.1133±0.0181**	**0.0760±0.0090**	**0.0376±0.0039**
	ijcnn1	*GRFF*	0.2924±0.0188	0.2171±0.0222	0.1504±0.0134	0.0757±0.0081
		GORF	**0.2415±0.0190**	**0.1026±0.0129**	**0.0739±0.0065**	**0.0383±0.0022**
	usps	*GRFF*	0.1005±0.0061	0.0690±0.0050	0.0500±0.0024	0.0253±0.0023
		GORF	**0.0724±0.0049**	**0.0235±0.0009**	**0.0166±0.0008**	**0.0083±0.0003**

Unbiased estimation + Lower variance = Lower approximation error

Experiments

SVM classification problem



(a) *letter* (b) *ijcnn1* (c) *usps*

Fig. 3. Comparisons of various algorithms for SVM classification task in terms of accuracy across two typical stationary indefinite kernels and three datasets with different dimensions. Top: polynomial kernel on the unit sphere. Below: delta-gaussian kernel

SVR regression problem

TABLE III
REGRESSION ERROR FOR KERNEL APPROXIMATION METHODS ON POLYNOMIAL KERNEL AND *HOUSING* DATASET (RMSE: MEAN±STD). THE LOWEST ERROR IS HIGHLIGHTED IN BOLDFACE

Methods	s=2d	s=4d	s=8d
RM	7.153±1.772	5.436±0.917	4.491±0.008
TS	5.414±0.879	4.772±0.177	4.657±0.316
SRF	4.391±0.368	3.906±0.219	3.555±0.130
DIGMM	4.897±0.368	4.130±0.324	4.000±0.475
GORF(OURS)	**4.079±0.233**	**3.817±0.204**	**3.472±0.137**

TABLE IV
REGRESSION ERROR FOR KERNEL APPROXIMATION METHODS ON DELTA-GAUSSIAN KERNEL AND *HOUSING* DATASET (RMSE: MEAN±STD). THE LOWEST ERROR IS HIGHLIGHTED IN BOLDFACE

Methods	s=2d	s=4d	s=8d
SRF	5.432±0.729	3.845±0.379	3.321±0.274
GORF(OURS)	**3.739±0.360**	**3.474±0.330**	**3.164±0.452**

Contents

- Background
- Motivation
- Methods
- Experiments
- Conclusion

Conclusion

1) Propose an unbiased random feature approximation with lower variance for stationary indefinite kernels

2) Verify the unbiasedness and numerically calculate the reduced variance.

3) Experimentally demonstrate the approximation error and performance in classification and regression task compared with other methods.