Memory Efficient Kernel Approximation for Non-Stationary and Indefinite Kernels

FH·W-S

Simon Heilig [1] Maximilian Münch [1,2] Frank-Michael Schleif [1]

[1]University of Applied Sciences Würzburg-Schweinfurt [2]University of Groningen

university of groningen

Take-Home Message

Large scale machine learning:

- Runtime as well as memory issues arise from quadratic matrices which are central to many kernel models.
- Kernel approximation is of high relevance in the age of large scale data.

Memory efficient kernel approximation (MEKA) from [4]:

- achieves very low approximation error, but
- results in non-positive definite approximations, and
- is restricted to shift-invariant kernels

Challenges:

- **To what extend does the MEKA approximation introduce indefiniteness?**
- **How to extend the class of kernels used in MEKA, in particular to indefinite ones?**
- **How to correct the approximation while maintaining the memory efficiency?**

Solution:

- **Spherical normalization and indefinite Nyström to extend the range of kernel functions**
- **Lanczos-Iteration based spectrum shift**

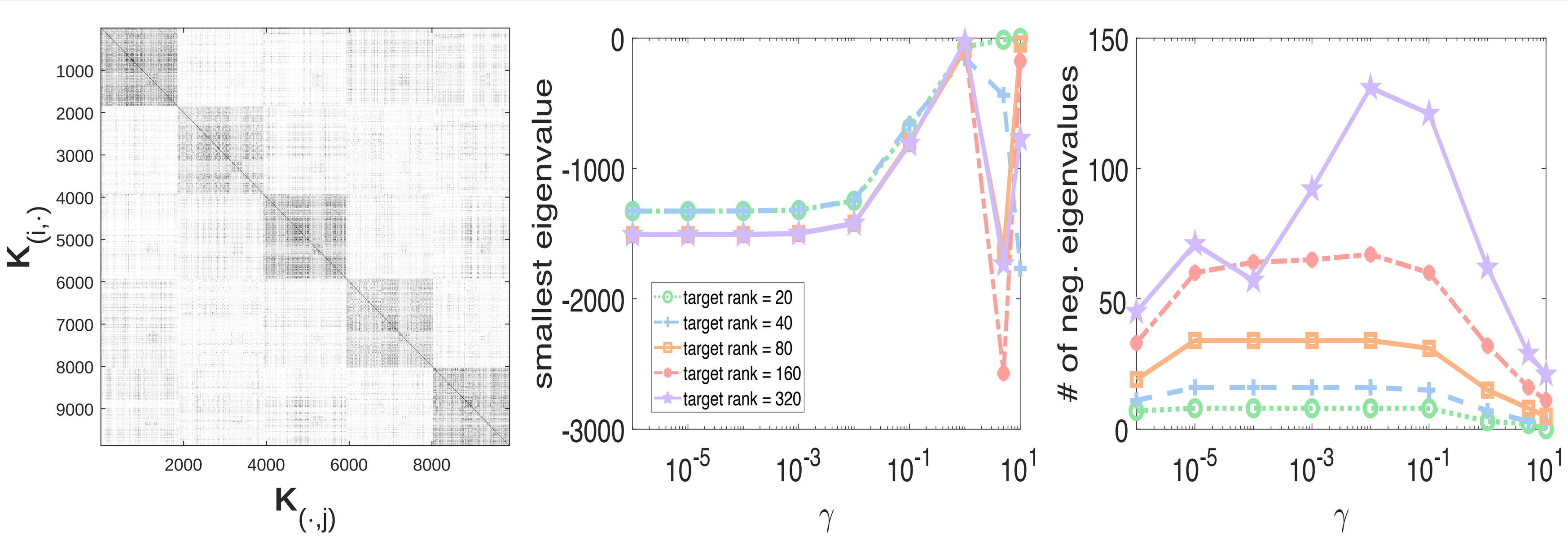
Introduction

Initial situation:

Kernel matrices are the central object studied kernelized models such as support vector machine (SVM), kernel principal component analysis or Gaussian Process models. Common approaches are:

- Random Fourier Features explicitly approximate the kernel function
- Nyström methods work on arbitrary symmetric matrices and result in $O(n\hat{k})$ memory [2]
- MEKA proposed by [4] minimizes the memory requirement to $O(nk + (ck)^2)$

Analysis of MEKA



Key steps of MEKA:

1. Approximate c-means clustering in the input space
2. Nyström approximation of *diagonal* blocks
3. Least-Squares approximation of *off-diagonal* blocks

Observations:

- Substantial negative eigenspace is present for $10^{-6} \leq \gamma \leq 1$
- When the matrix is close to full rank, it results in an increasing approximation error
- Direct correlation between the target rank of the approximation and the number of negative eigenvalues

Extending the classes of kernels

Non-Stationary kernels

Normalizing the data to the unit sphere $\mathcal{S}^{d-1}$ implies that: $\|\mathbf{x}-\mathbf{y}\|_2^2 = 2 - 2\langle \mathbf{x}, \mathbf{y}\rangle_2$, but the associated single variable function $k(\mathbf{x}-\mathbf{y})$ can be non-psd, as shown for the polynomial kernel by [3].

Indefinite kernels

In the light of indefinite kernel functions all steps of MEKA are applicable in a bounded error, since it has been shown that Nyström is also bounded in such cases [2].

Handling indefinite kernels

Sources of indefiniteness

- MEKA approximation of psd kernels
- Spherical normalization for non-stationary kernels [3]
- Domain specific similarity measures, e.g. protein sequence alignment or local learning (TL1 kernel) [1]

Lanczos-Iteration based spectrum shift

Correcting approximated matrix efficiently by $\tilde{\mathbf{K}} = \mathbf{QLQ}^T + \lambda_{shift}\mathbf{I}$, iff. $\lambda_{shift} \geq |\lambda_{min}|$. Where λ_{shift} is obtained by:

$$\lambda_{shift} = \min_{\mathbf{x} \neq 0} \frac{\langle \mathbf{x}, \, \mathbf{Ax}\rangle}{\langle \mathbf{x}, \, \mathbf{x}\rangle},$$

while requiring only a matrix-times-vector multiplication, which is proportional to $O(nk + (ck)^2)$ due to the decomposition of MEKA.
Derived error bound for corrected approximation:

$$\begin{aligned}\|\mathbf{K} - (\tilde{\mathbf{K}} + \lambda_{shift}\mathbf{I})\|_F \leq & \|\mathbf{K}^+ - \mathbf{K}_k^+\|_F + \left(\frac{64k}{l}\right)^{\frac{1}{4}} n\mathbf{K}_{max}^+ (1+\theta)^{\frac{1}{2}} + 2\|\Delta_+\|_F \\ & + \|\mathbf{K}^- - \mathbf{K}_k^-\|_F + \left(\frac{64k}{l}\right)^{\frac{1}{4}} n\mathbf{K}_{max}^- (1+\theta)^{\frac{1}{2}} + 2\|\Delta_-\|_F \\ & + \sqrt{n}|\lambda_{shift}|\end{aligned}$$

Experimental validation

SVM classification accuracy ($\pm$ std.), where **n.c.** refers to *not converged*.

	RBF Kernel		Sph. Poly. Kernel	
Dataset	MEKA	L-MEKA	MEKA	L-MEKA
artificial	85.48 ± 11.93	89.23 ± 1.42	**n.c.**	82.49 ± 1.03
cpusmall	**n.c.**	86.47 ± 1.32	**n.c.**	77.27 ± 1.76
pendigit	21.04 ± 12.88	87.79 ± 2.54	39.23 ± 32.40	98.01 ± 0.56

Contact Information

Simon Heilig
University of Applied Sciences Würzburg-Schweinfurt
Email: simon99.heilig@gmail.com
Overview about indefinite learning at:
http://promos-science.blogspot.com/
QR-Code for supplementary details and full paper

References

[1] Maximilian Münch, Christoph Raab, Michael Biehl, and Frank-Michael Schleif. Data-Driven Supervised Learning for Life Science Data. *Frontiers in Appl. Math. and Stat.*, 6:56, 2020.

[2] Dino Oglic and Thomas Gärtner. Scalable learning in reproducing kernel krein spaces. In *Int. Conf. on Mach. Learning*, pages 4912–4921. PMLR, 2019.

[3] Jeffrey Pennington, Felix X Yu, and Sanjiv Kumar. Spherical random features for polynomial kernels. In *Adv. in NIPS*, pages 1837–1845. MIT Press, 2015.

[4] Si Si, Cho-Jui Hsieh, and Inderjit Dhillon. Memory efficient kernel approximation. *The Journal of Machine Learning Research*, 18(1):682–713, 2017.

Supported by the Bavarian HighTech agenda and the Würzburg Center for Artificial Intelligence and Robotics (CAIRO)