Scalable Learning in Reproducing Kernel Kreĭn Spaces

Dino Oglic [1]   Thomas Gärtner [2]

[1] Department of Informatics, King's College London
[2] School of Computer Science, University of Nottingham

In Proceedings of the 36th International Conference on Machine Learning (ICML 2019)

In learning problems with structured data (e.g., time-series, strings, graphs), it is relatively easy to devise a pairwise (dis)similarity function based on intuition of a domain expert

To find an **optimal hypothesis** with standard kernel methods **positive definiteness** of the kernel/similarity function needs to be established

A large number of **pairwise (dis)similarity functions** devised by experts are **indefinite**
(e.g., edit distances for strings and graphs, dynamic time-warping algorithm, Wasserstein and Haussdorf distances)

Goal

Scalable kernel methods for learning with any notion of (dis)similarity between instances.

Kreĭn Space (Bognár, 1974; Azizov & Iokhvidov, 1981)

The vector space $\mathcal{K}$ with a bilinear form $\langle \cdot,\cdot \rangle_{\mathcal{K}}$ is called Kreĭn space if it admits a decomposition into a direct sum $\mathcal{K} = \mathcal{H}_+ \oplus \mathcal{H}_-$ of $\langle \cdot,\cdot \rangle_{\mathcal{K}}$-orthogonal Hilbert spaces $\mathcal{H}_\pm$ such that $\langle \cdot,\cdot \rangle_{\mathcal{K}}$ can be written as

$$\langle f,g \rangle_{\mathcal{K}} = \langle f_+,g_+ \rangle_{\mathcal{H}_+} - \langle f_-,g_- \rangle_{\mathcal{H}_-} ,$$

where $\mathcal{H}_\pm$ are endowed with inner products $\langle \cdot,\cdot \rangle_{\mathcal{H}_\pm}$, $f = f_+ \oplus f_-$, $g = g_+ \oplus g_-$, and $f_\pm, g_\pm \in \mathcal{H}_\pm$.

Associated Hilbert Space

For a decomposition $\mathcal{K} = \mathcal{H}_+ \oplus \mathcal{H}_-$, the Hilbert space $\mathcal{H}_\mathcal{K} = \mathcal{H}_+ \oplus \mathcal{H}_-$ endowed with inner product

$$\langle f, g \rangle_{\mathcal{H}_\mathcal{K}} = \langle f_+, g_+ \rangle_{\mathcal{H}_+} + \langle f_-, g_- \rangle_{\mathcal{H}_-} \quad (f_\pm, g_\pm \in \mathcal{H}_\pm)$$

can be associated with $\mathcal{K}$.

All the norms $\|\cdot\|_{\mathcal{H}_\mathcal{K}}$ generated by different decompositions of $\mathcal{K}$ into direct sums of Hilbert spaces are **topologically equivalent** (Langer, 1962)

The topology on $\mathcal{K}$ defined by the **norm of an associated Hilbert space** is called the **strong topology** on $\mathcal{K}$

$\exists f \in \mathcal{K}\colon \langle f, f \rangle_\mathcal{K} < 0 \Longrightarrow \langle f, f \rangle_\mathcal{K} = \|f_+\|^2_{\mathcal{H}_+} - \|f_-\|^2_{\mathcal{H}_-}$ does not induce a norm on a reproducing kernel Kreĭn space $\mathcal{K}$

The complexity of hypotheses can be penalized via decomposition components $\mathcal{H}_\pm$ and the strong topology

Scalability !

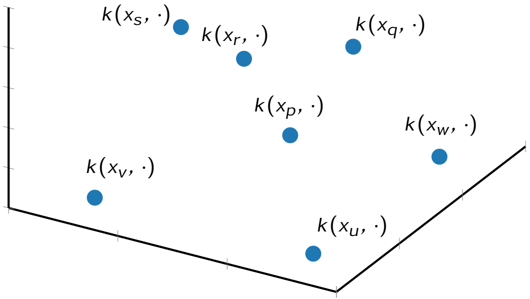
Computational and space complexities are often **quadratic in the number of instances** and in several approaches the computational complexity is cubic.

Nyström Method for Indefinite Kernels
Overview

$\mathcal{X}$ is an instance space

$X = \{x_1, \ldots, x_n\}$ is an independent sample from a probability measure defined on $\mathcal{X}$

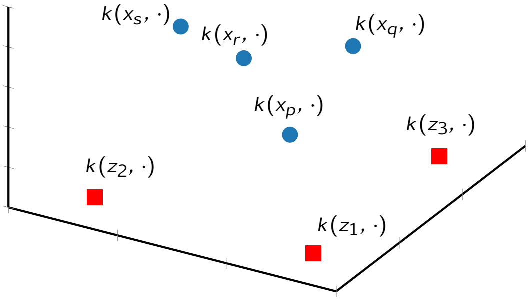
$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a reproducing Kreĭn kernel with $k(x, x') = \langle k(x, \cdot), k(x', \cdot) \rangle_{\mathcal{K}}$



Nyström Method for Indefinite Kernels

Landmarks

$Z = \{z_1, \ldots, z_m\}$ is a set of landmarks (not necessarily a subset of X)

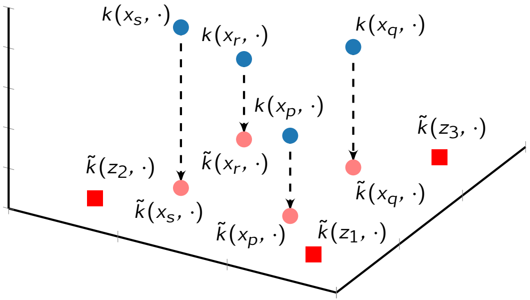


Nyström Method for Indefinite Kernels

Projections onto $\mathcal{L}_Z = \operatorname{span}(\{k(z_1,\cdot),\cdots,k(z_m,\cdot)\})$

For a given set of landmarks Z, the Nyström method approximates the kernel matrix K with a low-rank matrix $\tilde{K}$ given by $\tilde{K}_{ij} = \tilde{k}\left(x_i, x_j\right) = \left\langle \tilde{k}(x_i,\cdot), \tilde{k}\left(x_j,\cdot\right)\right\rangle_{\mathcal{K}}$

$$k(x,\cdot) = \tilde{k}(x,\cdot) + k^{\perp}(x,\cdot) \quad \text{with} \quad \tilde{k}(x,\cdot) = \sum_{i=1}^{m} \alpha_{i,x} k(z_i,\cdot) \quad \wedge \quad \left\langle k^{\perp}(x,\cdot), \mathcal{L}_Z \right\rangle_{\mathcal{K}} = 0$$



$$\tilde{K} = K_{n,m} K_{m,m}^{-1} K_{m,n} = \tilde{U}_m \tilde{\Lambda}_m \tilde{U}_m^{\top} \quad \text{with} \quad \tilde{U}_m^{\top} \tilde{U}_m = \mathbb{I}_m$$

Contributions

First **mathematically complete derivation** of the Nyström method for indefinite kernels

An approach for efficient **low-rank eigendecomposition of indefinite kernel matrices**

Two effective **landmark selection strategies** for the Nyström method with **indefinite kernels**

Nyström-based **scalable least squares methods** for learning in reproducing kernel Kreĭn spaces

Nyström-based **scalable support vector machine** for learning in reproducing kernel Kreĭn spaces

Effective **regularization via decomposition components** $\mathcal{H}_\pm$ and the strong topology

PYTHON package for learning in reproducing kernel Kreĭn spaces
(in preparation, early version available upon request)