Scalable Learning in Reproducing Kernel Kreĭn Spaces

Dino Oglic [1] **Thomas Gärtner** [2]

Abstract

We provide the first mathematically complete derivation of the Nyström method for low-rank approximation of indefinite kernels and propose an efficient method for finding an approximate eigendecomposition of such kernel matrices. Building on this result, we devise highly scalable methods for learning in reproducing kernel Kreĭn spaces. The devised approaches provide a principled and theoretically well-founded means to tackle large scale learning problems with indefinite kernels. The main motivation for our work comes from problems with structured representations (e.g., graphs, strings, time-series), where it is relatively easy to devise a pairwise (dis)similarity function based on intuition and/or knowledge of domain experts. Such functions are typically not positive definite and it is often well beyond the expertise of practitioners to verify this condition. The effectiveness of the devised approaches is evaluated empirically using indefinite kernels defined on structured and vectorial data representations.

1. Introduction

In learning problems with structured data it is relatively easy to devise a pairwise similarity/dissimilarity function based on intuition/knowledge of domain experts. Such functions are typically not positive definite and it is often the case that verifying this condition is well beyond the expertise of practitioners. The learning problems with indefinite similarity/dissimilarity functions are typically modeled via Kreĭn spaces (e.g., see Ong et al., 2004; Loosli et al., 2016; Oglic and Gärtner, 2018), which are vector spaces with an indefinite bilinear form (Azizov and Iokhvidov, 1981; Iokhvidov et al., 1982). The computational and space complexities of these approaches are similar to those of the standard kernel methods that work with positive definite kernels (Schölkopf and Smola, 2001). The Nyström method (Nyström, 1930; Smola and Schölkopf, 2000; Williams and Seeger, 2001) is an effective approach for low-rank approximation of positive definite kernels that can scale kernel methods to problems with millions of instances (Schölkopf and Smola, 2001). We provide the first mathematically complete derivation that extends the Nyström method to low-rank approximation of indefinite kernels and propose an efficient method for finding an approximate eigendecomposition of such kernel matrices. To tackle the computational issues arising in large scale problems with indefinite kernels, we also devise several novel Nyström-based low-rank approaches tailored for scalable learning in reproducing kernel Kreĭn spaces.

We start by showing that the Nyström method can be used for low-rank approximations of indefinite kernel matrices and provide means for finding their approximate eigendecompositions (Section 2.2). We then devise two landmark sampling strategies based on state-of-the-art techniques (Gittens and Mahoney, 2016; Oglic and Gärtner, 2017) used in Nyström approximations of positive definite kernels (Section 2.3). Having described means for finding low-rank factorizations of indefinite kernel matrices, we formulate low-rank variants of two least squares methods (Tikhonov and Arsenin, 1977; Rifkin, 2002; Oglic and Gärtner, 2018) for learning in reproducing kernel Kreĭn spaces (Section 2.4). We also derive a novel low-rank variant of the support vector machine for scalable learning in reproducing kernel Kreĭn spaces (Section 2.5), inspired by Oglic and Gärtner (2018). Having introduced means for scalable learning in reproducing kernel Kreĭn spaces, we evaluate the effectiveness of these approaches and the Nyström low-rank approximations on datasets from standard machine learning repositories (Section 3). The empirical results demonstrate the effectiveness of the proposed approaches in: *i*) classification tasks and *ii*) problems of finding a low-rank approximation of an indefinite kernel matrix. The experiments are performed using 15 representative datasets and a variety of indefinite kernels. The paper concludes with a discussion where we contrast ours and other relevant approaches (Section 4).

2. Scalable Learning in Kreĭn Spaces

In this section, we first provide a novel derivation of the Nyström method that allows us to extend the approach to

[1]Department of Informatics, King's College London, UK [2]School of Computer Science, University of Nottingham, UK. Correspondence to: Dino Oglic <dino.oglic@uni-bonn.de>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

low-rank approximation of indefinite kernel matrices. Building on this result, we then provide means to scale Kreĭn kernel methods to datasets with millions of instances/pairwise (dis)similarities. More specifically, we devise low-rank variants of kernel ridge regression and support vector machines in reproducing kernel Kreĭn spaces, as well as a low-rank variant of the variance constrained ridge regression proposed in Oglic and Gärtner (2018). As the effectiveness of low-rank approximations based on the Nyström method critically depends on the selected landmarks, we also adapt two state-of-the-art landmark sampling strategies proposed for the approximation of positive definite kernels. In addition to this, we make a theoretical contribution (Proposition 3) relating the Kreĭn support vector machines (Loosli et al., 2016) to previous work on learning with the flip-spectrum kernel matrices (Graepel et al., 1998; Chen et al., 2009).

2.1. Reproducing Kernel Kreĭn Spaces

Let $\mathcal{K}$ be a vector space defined on the scalar field $\mathbb{R}$. A bilinear form on $\mathcal{K}$ is a function $\langle\cdot,\cdot\rangle_{\mathcal{K}} : \mathcal{K}\times\mathcal{K}\to\mathbb{R}$ such that, for all $f,g,h\in\mathcal{K}$ and scalars $\alpha,\beta\in\mathbb{R}$, it holds: *i)* $\langle \alpha f+\beta g, h\rangle_{\mathcal{K}} = \alpha\langle f,h\rangle_{\mathcal{K}} + \beta\langle g,h\rangle_{\mathcal{K}}$ and *ii)* $\langle f, \alpha g+\beta h\rangle_{\mathcal{K}} = \alpha\langle f,g\rangle_{\mathcal{K}} + \beta\langle f,h\rangle_{\mathcal{K}}$. The bilinear form is called non-degenerate if

$$(\forall f\in\mathcal{K}) : \Big((\forall g\in\mathcal{K})\ \langle f,g\rangle_{\mathcal{K}} = 0\Big) \implies f = 0.$$

The bilinear form $\langle\cdot,\cdot\rangle_{\mathcal{K}}$ is symmetric if, for all $f,g\in\mathcal{K}$, we have $\langle f,g\rangle_{\mathcal{K}} = \langle g,f\rangle_{\mathcal{K}}$. The form is called indefinite if there exists $f,g\in\mathcal{K}$ such that $\langle f,f\rangle_{\mathcal{K}} > 0$ and $\langle g,g\rangle_{\mathcal{K}} < 0$. On the other hand, if $\langle f,f\rangle_{\mathcal{K}}\geq 0$ for all $f\in\mathcal{K}$, then the form is called positive. A non-degenerate, symmetric, and positive bilinear form on $\mathcal{K}$ is called inner product. Any two elements $f,g\in\mathcal{K}$ that satisfy $\langle f,g\rangle_{\mathcal{K}} = 0$ are $\langle\cdot,\cdot\rangle_{\mathcal{K}}$-orthogonal. Similarly, any two subspaces $\mathcal{K}_1,\mathcal{K}_2\subset\mathcal{K}$ that satisfy $\langle f_1,f_2\rangle_{\mathcal{K}} = 0$ for all $f_1\in\mathcal{K}_1$ and $f_2\in\mathcal{K}_2$ are called $\langle\cdot,\cdot\rangle_{\mathcal{K}}$-orthogonal. Having reviewed bilinear forms, we are now ready to introduce the notion of a Kreĭn space.

Definition 1. *(Bognár, 1974; Azizov and 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 the bilinear form 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 $f_\pm, g_\pm\in\mathcal{H}_\pm$.

For a fixed 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 $\mathcal{K}$. For a Kreĭn space $\mathcal{K}$, the decomposition $\mathcal{K} = \mathcal{H}_+\oplus\mathcal{H}_-$ is not necessarily unique. Thus, a Kreĭn space can, in general, be associated with infinitely many Hilbert spaces. However, for any such Hilbert space $\mathcal{H}_{\mathcal{K}}$ the topology introduced on $\mathcal{K}$ via the norm $\|f\|_{\mathcal{H}_{\mathcal{K}}} = \sqrt{\langle f,f\rangle_{\mathcal{H}_{\mathcal{K}}}}$ is independent of the decomposition and the associated Hilbert space. More specifically, all 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. Having reviewed basic properties of Kreĭn spaces, we are now ready to introduce a notion of reproducing kernel Kreĭn space. For that, let $\mathcal{X}$ be an instance space and denote with $\mathbb{R}^{\mathcal{X}}$ the set of functions from $\mathcal{X}$ to $\mathbb{R}$. For a fixed element $x\in\mathcal{X}$, the map $\mathbb{T}_x\colon\mathbb{R}^{\mathcal{X}}\to\mathbb{R}$ that assigns a real number to each function $f\in\mathcal{R}^{\mathcal{X}}$ is called the evaluation functional at x, i.e., $\mathbb{T}_x(f) = f(x)$ for all $f\in\mathbb{R}^{\mathcal{X}}$.

Definition 2. *(Alpay, 1991; Ong et al., 2004) A Kreĭn space $(\mathcal{K},\langle\cdot,\cdot\rangle_{\mathcal{K}})$ is a reproducing kernel Kreĭn space if $\mathcal{K}\subset\mathbb{R}^{\mathcal{X}}$ and the evaluation functional is continuous on $\mathcal{K}$ with respect to the strong topology.*

The following theorem provides a characterization of reproducing kernel Kreĭn spaces.

Theorem 1. *(Schwartz, 1964; Alpay, 1991) Let $k\colon\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ be a real-valued symmetric function. Then, there is an associated reproducing kernel Kreĭn space if and only if $k = k_+ - k_-$, where k_+ and k_- are positive definite kernels. When the function k admits such a decomposition, one can choose k_+ and k_- such that the corresponding reproducing kernel Hilbert spaces are disjoint.*

2.2. Nyström Method for Reproducing Kreĭn Kernels

Let $\mathcal{X}$ be an instance space and $X = \{x_1,\ldots,x_n\}$ an independent sample from a Borel probability measure defined on $\mathcal{X}$. For a positive definite kernel h and a set of landmarks $Z = \{z_1,\ldots,z_m\}\subset\mathcal{X}$, the Nyström method (Nyström, 1930; Smola and Schölkopf, 2000; Williams and Seeger, 2001) first projects the evaluation functionals $h(x_i,\cdot)$ onto $\operatorname{span}(\{h(z_1,\cdot),\ldots,h(z_m,\cdot)\})$ and then approximates the kernel matrix H with entries $\{H_{ij} = h(x_i,x_j)\}_{i,j=1}^n$ by inner products between the projections of the corresponding evaluation functionals. The projections of the evaluation functionals $h(x_i,\cdot)$ are linear combinations of the landmarks and these coefficients are given by the following convex optimization problem

$$\alpha^* = \operatorname*{arg\,min}_{\alpha\in\mathbb{R}^{m\times n}} \sum_{i=1}^n \left\| h(x_i,\cdot) - \sum_{j=1}^m \alpha_{j,i} h(z_j,\cdot)\right\|_{\mathcal{H}}^2 . \quad (1)$$

While this approach works for positive definite kernels, it cannot be directly applied to reproducing Kreĭn kernels. In

particular, let $\mathcal{K}$ be a reproducing kernel Kreĭn space with an indefinite kernel $k\colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}$. The reproducing Kreĭn kernel k is defined by an indefinite bilinear form $\langle\cdot,\cdot\rangle_{\mathcal{K}}$ which does not induce a norm on $\mathcal{K}$ and for all $a,b\in\mathcal{K}$ the value of $\langle a-b,a-b\rangle_{\mathcal{K}}$ does not capture the distance. More specifically, as the bilinear form is indefinite then there exists an element $c\in\mathcal{K}$ such that $\langle c,c\rangle_{\mathcal{K}}<0$.

For an evaluation functional $k(x,\cdot)\in\mathcal{K}$ and a linear subspace $\mathcal{L}_Z\subset\mathcal{K}$ spanned by a set of evaluation functionals $\{k(z_1,\cdot),\ldots,k(z_m,\cdot)\}$, we define $\tilde{k}(x,\cdot)$ to be the orthogonal projection of $k(x,\cdot)$ onto the subspace $\mathcal{L}_Z$ if the evaluation functional admits a decomposition (Azizov and Iokhvidov, 1981; Iokhvidov et al., 1982)

$$k(x,\cdot)=\tilde{k}(x,\cdot)+k^{\perp}(x,\cdot)\ ,$$

where $\tilde{k}(x,\cdot)=\sum_{i=1}^{m}\alpha_{i,x}k(z_i,\cdot)$ with $\alpha_x\in\mathbb{R}^m$, and $\langle k^{\perp}(x,\cdot),\mathcal{L}_Z\rangle_{\mathcal{K}}=0$. For a landmark $z\in Z$, the inner product between the corresponding evaluation functional $k(z,\cdot)$ and $k(x,\cdot)$ then gives

$$k(x,z)=\langle k(x,\cdot),k(z,\cdot)\rangle_{\mathcal{K}}=\sum_{i=1}^{m}\alpha_{i,x}\tilde{k}(z_i,z)\ . \quad (2)$$

Denote with $K_{Z\times Z}$ the block in the kernel matrix K corresponding to landmarks Z and let $k_x=\operatorname{vec}(k(x,z_1),\ldots,k(x,z_m))$. From Eq. (2) it then follows that $k_x=K_{Z\times Z}\alpha_x$. Thus, in Kreĭn spaces an orthogonal projection exists if the matrix $K_{Z\times Z}$ is non-singular. If this condition is satisfied, then the projection is given by

$$\tilde{k}(x,\cdot)=\sum_{i=1}^{m}\alpha^{*}_{i,x}k(z_i,\cdot)\ \text{ with }\alpha^{*}_{x}=K^{-1}_{Z\times Z}k_x\in\mathbb{R}^m\ .$$

Having computed the projection of a point onto the span of the landmarks in a Kreĭn space, we now proceed to define the Nyström approximation of the corresponding indefinite kernel matrix. In this, we follow the approach for positive definite kernels (Schölkopf and Smola, 2001; Smola and Schölkopf, 2000) and approximate the Kreĭn kernel matrix K using the bilinear form on the span of the landmarks. More specifically, we have that

$$\tilde{k}(x_i,x_j)=\left\langle\sum_{p=1}^{m}\alpha^{*}_{p,i}k(z_p,\cdot),\sum_{q=1}^{m}\alpha^{*}_{q,j}k(z_q,\cdot)\right\rangle_{\mathcal{K}}=k_{x_i}^{\top}K^{-1}_{Z\times Z}k_{x_j}\ .$$

Thus, the low-rank approximation of the Kreĭn kernel matrix K is given by

$$\tilde{K}_{X|Z}=K_{X\times Z}K^{-1}_{Z\times Z}K_{Z\times Z}K^{-1}_{Z\times Z}K_{Z\times X}=K_{X\times Z}K^{-1}_{Z\times Z}K_{Z\times X}\ . \quad (3)$$

This approach for low-rank approximation of Kreĭn kernel matrices also provides a direct way for an out-of-sample extension in the non-transductive setting. In particular, for an out-of-sample instance $x\in\mathcal{X}$ we have that if holds

$$\tilde{k}_{x\times X}=\operatorname{vec}(\tilde{k}(x,x_1),\ldots,\tilde{k}(x,x_n))=K_{X\times Z}K^{-1}_{Z\times Z}k_x.$$

In applications to estimation problems (see Sections 2.4 and 2.5) an approximate low-rank *eigendecomposition* of the kernel matrix, also known as the *one-shot* variant of the Nyström method (Fowlkes et al., 2004), is sometimes preferred over the plain Nyström approximation described above. To derive such a factorization, we first observe that the low-rank approximation of the indefinite kernel matrix can be written as

$$\tilde{K}_{X|Z}=LSL^{\top}\ \text{ with }\ L=K_{X\times Z}U_{Z\times Z}\left|D_{Z\times Z}\right|^{-\frac{1}{2}}\ ,$$

and where $K_{Z\times Z}=U_{Z\times Z}D_{Z\times Z}U^{\top}_{Z\times Z}$ is an eigendecomposition of the block in the kernel matrix corresponding to landmarks and $S=\operatorname{sign}(D_{Z\times Z})$. Substituting a singular value decomposition of $L=A\Sigma B^{\top}$ into the latter equation (with orthonormal matrices $A\in\mathbb{R}^{n\times m}$ and $B\in\mathbb{R}^{m\times m}$), we deduce the following low-rank factorization

$$\tilde{K}_{X|Z}=A\,\Sigma B^{\top}SB\Sigma\,A^{\top}=AMA^{\top}\ ,$$

where $M=\Sigma B^{\top}SB\Sigma$ is a symmetric matrix with an eigendecomposition $M=P\Lambda P^{\top}$. Hence,

$$\tilde{K}_{X|Z}=(AP)\,\Lambda\,(AP)^{\top}\quad\text{with}\quad(AP)^{\top}AP=\mathbb{I}_m\ .$$

As the matrix $\tilde{U}=AP\in\mathbb{R}^{n\times m}$ contains m orthonormal column vectors and Λ is a diagonal matrix, we have then derived an approximate low-rank eigendecomposition of K.

2.3. Landmark Selection for the Nyström Method with Indefinite Kernels

The effectiveness of a low-rank approximation based on the Nyström method depends crucially on the choice of landmarks and an optimal choice is a difficult discrete optimization problem. The landmark selection for the Nyström method has been studied extensively in the context of approximation of positive definite matrices (e.g., see Drineas and Mahoney, 2005; Kumar et al., 2012; Gittens and Mahoney, 2016; Alaoui and Mahoney, 2015; Oglic and Gärtner, 2017). We follow this line of research and present two landmark selection strategies for indefinite Kreĭn kernels inspired by the state-of-the-art sampling schemes: (approximate) kernel K-means++ sampling (Oglic and Gärtner, 2017) and statistical leverage scores (Alaoui and Mahoney, 2015; Drineas et al., 2006; Drineas and Mahoney, 2005).

In both cases, we propose to first sample a small number of instances uniformly at random and create a sketch matrix $\tilde{K}=\tilde{U}\Lambda\tilde{U}^{\top}$ using the procedure described in Section 2.2. Then, using this sketch matrix we propose to

approximate: *i*) statistical leverage scores for all instances, and/or *ii*) squared distances between instances in the feature space of the factorization $\tilde{H} = \tilde{L}\tilde{L}^\top$ with $\tilde{L} = \tilde{U}\left|\Lambda\right|^{1/2}$. An approximate leverage score assigned to the i-th instance is given as the squared norm of the i-th row in the matrix $\tilde{U}$, that is $\ell\left(x_i\right) = \|\tilde{U}(i)\|^2$ with $1 \leq i \leq n$. As the two matrices $\tilde{H}$ and $\tilde{K}$ have identical eigenvectors, the approximate leverage scores obtained using the positive definite matrix $\tilde{H}$ capture the informative part of the eigenspace of the indefinite matrix $\tilde{K}$. This follows from the Eckart–Young–Mirsky theorem (Eckart and Young, 1936; Mirsky, 1960) which implies that the optimal low-rank approximation of an indefinite kernel matrix is given by a set of landmarks spanning the same subspace as that spanned by the eigenvectors corresponding to the top eigenvalues, sorted in descending order with respect to their absolute values. The landmark selection strategy based on the approximate leverage scores then works by taking a set of independent samples from

$$p_\ell\left(x\right) = {}^{\ell(x)}\!/_{\sum_{i=1}^{n}\ell(x_i)}\ .$$

For approximate kernel K-means++ landmark selection, we propose to perform K-means++ clustering (Arthur and Vassilvitskii, 2007) in the feature space defined by the factorization matrix $\tilde{L}$, that is each instance is represented with a row from this matrix. The approach works by first sampling an instance uniformly at random and setting it as the first landmark (i.e., the first cluster centroid). Following this, the next landmark/centroid is selected by sampling an instance with the probability proportional to its clustering contribution. More formally, assuming that the landmarks $\{z_1, z_2, \ldots, z_s\}$ have already been selected the $(s+1)$-st one is selected by taking a sample from the distribution

$$p_{s+1}^{++}\left(x\right) = {}^{\min_{1\leq i\leq s}\|x-z_i\|^2}\!/_{\sum_{i=1}^{n}\min_{1\leq j\leq s}\|x_i-z_j\|^2}\ .$$

2.4. Scaling Least Squares Methods for Indefinite Kernels using the Nyström Method

We present two regularized risk minimization problems with squared error loss function for scalable learning in reproducing kernel Kreĭn spaces. Our choice of the regularization term is motivated by the considerations in Oglic and Gärtner (2018), where the authors regularize with respect to a decomposition of the Kreĭn kernel into a direct sum of Hilbert spaces. We start with a Kreĭn least squares method (KREĬN LSM) which is a variant of kernel ridge regression, i.e.,

$$f^* = \operatorname*{arg\,min}_{f\in\mathcal{K}} \quad \frac{1}{n}\sum_{i=1}^{n}\Big(f\left(x_i\right) - y_i\Big)^2 + \lambda_+\left\|f_+\right\|_+^2 + \lambda_-\left\|f_-\right\|_-^2\ ,$$

where $f = f_+ \oplus f_- \in \mathcal{K}$, $\mathcal{K} = \mathcal{H}_+ \oplus \mathcal{H}_-$ with disjoint $\mathcal{H}_\pm$, $f_\pm \in \mathcal{H}_\pm$, and hyperparameters $\lambda_\pm \in \mathbb{R}^+$. This is a convex optimization problem for which the representer theorem holds (Oglic and Gärtner, 2018, Appendix A) and the optimal solution $f^* = \sum_{i=1}^{n}\alpha_i^* k\left(x_i, \cdot\right)$ with $\alpha^* \in \mathbb{R}^n$. Applying the reproducing property of the Kreĭn kernel and setting the gradient of the objective to zero, we derive

$$\alpha^* = \Big(H + n\Lambda_\pm\Big)^{-1} P y\ ,$$

where $K = UDU^\top$, $S = \operatorname{sign}\left(D\right)$, $H = UDSU^\top$, $P = USU^\top$, and $\Lambda_\pm = \lambda_+ S_+ + \lambda_-\left|S_-\right|$ with $S_\pm = {}^{(S\pm\mathbb{I})}\!/_2$.

An important difference compared to stabilization approaches (e.g., see Loosli et al., 2016) is that we are solving a regularized risk minimization problem for which a globally optimal solution can be found in polynomial time. Another difference is that stabilization approaches perform subspace descent while we are optimizing jointly over decomposition components of a Krein space. In the special case with $\lambda_+ = \lambda_-$, the approach outputs a hypothesis equivalent to that of a stabilization approach along the lines of Loosli et al. (2016). In particular, the matrix H is called the flip-spectrum transformation of K and $k_{x\times X}^\top P$ is the corresponding out-of-sample transformation. Learning with the flip-spectrum transformation of an indefinite kernel matrix was first considered in Graepel et al. (1998) and the corresponding out-of-sample transformation was first proposed in Chen et al. (2009). The following proposition (a proof is provided in Appendix A) establishes the equivalence between the least squares method with the flip-spectrum matrix in place of an indefinite kernel matrix and Kreĭn kernel ridge regression regularized with a single hyperparameter.

Proposition 2. *If the Kreĭn kernel ridge regression problem is regularized via the norm $\|\cdot\|_{\mathcal{H}_\mathcal{K}}$ with $\lambda = \lambda_+ = \lambda_-$, then the optimal hypothesis is equivalent to that obtained with kernel ridge regression and the flip-spectrum matrix in place of an indefinite Kreĭn kernel matrix.*

Having established this, we now proceed to formulate a Kreĭn regression problem with a low-rank approximation $\tilde{K}_{X|Z}$ in place of the indefinite kernel matrix K. More formally, after substituting the low-rank approximation into Kreĭn kernel ridge regression problem we transform it by

$$z = \left|D_{Z\times Z}\right|^{-1/2} U_{Z\times Z}^\top K_{Z\times X}\alpha = L_{X|Z}^\top\alpha\ ,$$
$$\Phi = K_{X\times Z}U_{Z\times Z}\left|D_{Z\times Z}\right|^{-1/2} S_{Z\times Z} = L_{X|Z}S_{Z\times Z}\ ,$$
$$\tilde{K}_{X|Z}\alpha = L_{X|Z}S_{Z\times Z}z = \Phi z \ \text{ and } \ \alpha^\top H_\pm\alpha = z_\pm^\top z_\pm\ ,$$

where $H_\pm = L_{X|Z}\left|S_{Z\times Z,\pm}\right|L_{X|Z}^\top$, $z_\pm = \left|S_{Z\times Z,\pm}\right|z$, and $S_{Z\times Z,\pm} = {}^{(S_{Z\times Z}\pm\mathbb{I})}\!/_2$. Hence, we can write a low-rank variant of the Kreĭn kernel ridge regression problem as

$$z^* = \operatorname*{arg\,min}_{z\in\mathbb{R}^m} \ \|\Phi z - y\|_2^2 + n\lambda_+\left\|z_+\right\|_2^2 + n\lambda_-\left\|z_-\right\|_2^2\ .$$

The problem is convex in z and the optimal solution satisfies

$$z^* = \left(\Phi^\top\Phi + n\Lambda_\pm\right)^{-1}\Phi^\top y\ .$$

An out-of-sample extension for this learning problem is

$$\tilde{f}^* (x) = k_x^\top U_{Z\times Z} \left|D_{Z\times Z}\right|^{-1/2} S_{Z\times Z} z^* \ .$$

Having introduced a low-rank variant of Kreĭn kernel ridge regression, we proceed to define a scalable variance constrained least squares method (KREĬN VC-LSM). This risk minimization problem is given by (Oglic and Gärtner, 2018)

$$\begin{aligned} \min_{f\in\mathcal{K}} \quad & \frac{1}{n}\sum_{i=1}^{n} \left(f\left(x_i\right) - y_i\right)^2 + \lambda_+ \left\|f_+\right\|_+^2 + \lambda_- \left\|f_-\right\|_-^2 \\ s.t. \quad & \sum_{i=1}^{n} f\left(x_i\right)^2 = r^2 \ , \end{aligned}$$

with hyperparameters $r \in \mathbb{R}$ and $\lambda_\pm \in \mathbb{R}^+$. To simplify our derivations (just as in Oglic and Gärtner, 2018), we have without loss of generality assumed that the kernel matrix K is centered. Then, the hard constraint fixes the variance of the predictor over training instances. Similar to Kreĭn kernel ridge regression, we can transform this problem into

$$\begin{aligned} z^* = \quad & \operatorname*{arg\,min}_{z\in\mathbb{R}^m} \ n\lambda_+ \left\|z_+\right\|^2 + n\lambda_- \left\|z_-\right\|^2 - 2z^\top \Phi^\top y \\ s.t. \quad & z^\top \Phi^\top \Phi z = r^2 \ . \end{aligned}$$

Now, performing a singular value decomposition of $\Phi = A\Delta B^\top$ and taking $\gamma = \Delta B^\top z$ we obtain

$$\begin{aligned} \gamma^* = \quad & \operatorname*{arg\,min}_{\gamma\in\mathbb{R}^m} \ n\gamma^\top \Delta^{-1} B^\top \Lambda_\pm B \Delta^{-1} \gamma - 2(A^\top y)^\top \gamma \\ s.t. \quad & \gamma^\top \gamma = r^2 \ . \end{aligned}$$

A globally optimal solution to this non-convex problem can be computed by following the procedures outlined in Gander et al. (1989) and Oglic and Gärtner (2018). The cost of computing the solution is $\mathcal{O}\left(m^3\right)$ and the cost for the low-rank transformation of the problem is $\mathcal{O}\left(m^3 + m^2 n\right)$. An out-of-sample extension can also be obtained by following the derivation for Kreĭn kernel ridge regression.

2.5. Scaling Support Vector Machines for Indefinite Kernels using the Nyström Method

In this section, we propose a low-rank support vector machine for scalable classification with indefinite kernels. Our regularization term is again motivated by the considerations in Oglic and Gärtner (2018) and that is one of the two main differences compared to Kreĭn support vector machine proposed in Loosli et al. (2016). The latter approach outputs a hypothesis which can equivalently be obtained using the standard support vector machine with the flip-spectrum kernel matrix combined with the corresponding out-of-sample transformation (introduced in Section 2.4). The second difference of our approach compared to Loosli et al. (2016) stems from the fact that in low-rank formulations one optimizes the primal of the problem, defined with the squared hinge loss instead of the plain hinge loss. In particular, the latter loss function is not differentiable and that can complicate the hyperparameter optimization. We note that the identical choice of loss function was used in other works for primal-based optimization of support vector machines (e.g., see Mangasarian, 2002; Keerthi and DeCoste, 2005).

We propose the following optimization problem as the Kreĭn squared hinge support vector machine (KREĬN SH-SVM)

$$\begin{aligned} f^* = \operatorname*{arg\,min}_{f\in\mathcal{K}} \quad & \frac{1}{n}\sum_{i=1}^{n} \max\left\{1 - y_i f\left(x_i\right), 0\right\}^2 + \\ & \lambda_+ \left\|f_+\right\|_+^2 + \lambda_- \left\|f_-\right\|_-^2 \ . \end{aligned}$$

Similar to Section 2.4, the representer theorem holds for this problem and applying the reproducing property of the Kreĭn kernel we can transform it to a matrix form. If we again substitute a low-rank approximation $\tilde{K}_{X|Z}$ in place of the Kreĭn kernel matrix K, we observe that

$$\begin{aligned} f\left(x_i\right) = \quad & \tilde{k}_{x_i}^\top \alpha = k_{x_i}^\top K_{Z\times Z}^{-1} K_{Z\times X} \alpha = \\ & k_{x_i}^\top U_{Z\times Z} \left|D_{Z\times Z}\right|^{-1/2} S_{Z\times Z} z = \Phi_i z \ , \end{aligned}$$

where Φ_i denotes the i-th row in the matrix Φ. The low-rank variant of the approach can then be written as

$$\begin{aligned} z^* = \operatorname*{arg\,min}_{z\in\mathbb{R}^m} \quad & \sum_{i=1}^{n} \max\left\{1 - y_i \Phi_i z, 0\right\}^2 + \\ & n\lambda_+ \left\|z_+\right\|_2^2 + n\lambda_- \left\|z_-\right\|_2^2 \ . \end{aligned}$$

The derivation of the solution follows that for the standard primal-based training of support vector machines with the only difference being that the diagonal matrix $\Lambda_\pm$ is used instead of the scalar hyperparameter controlling the hypothesis complexity (e.g., see Mangasarian, 2002; Keerthi and DeCoste, 2005). To automatically tune the hyperparameters, one can follow the procedure described in Chapelle et al. (2002) and use implicit derivation to compute the gradient of the optimal solution with respect to the hyperparameters.

We conclude with a discussion of a potential shortcoming inherent to the Kreĭn support vector machine (Loosli et al., 2016). As the following proposition shows (a proof can be found in Appendix A), that approach is equivalent to the standard support vector machine with the flip-spectrum matrix in place of an indefinite Kreĭn kernel matrix (Graepel et al., 1998), combined with the corresponding out-of-sample transformation (Chen et al., 2009).

Proposition 3. *The Kreĭn support vector machine (Loosli et al., 2016) is equivalent to the standard support vector machine with the flip-spectrum matrix in place of an indefinite Kreĭn kernel matrix (Graepel et al., 1998), combined with the out-of-sample transformation from Chen et al. (2009).*

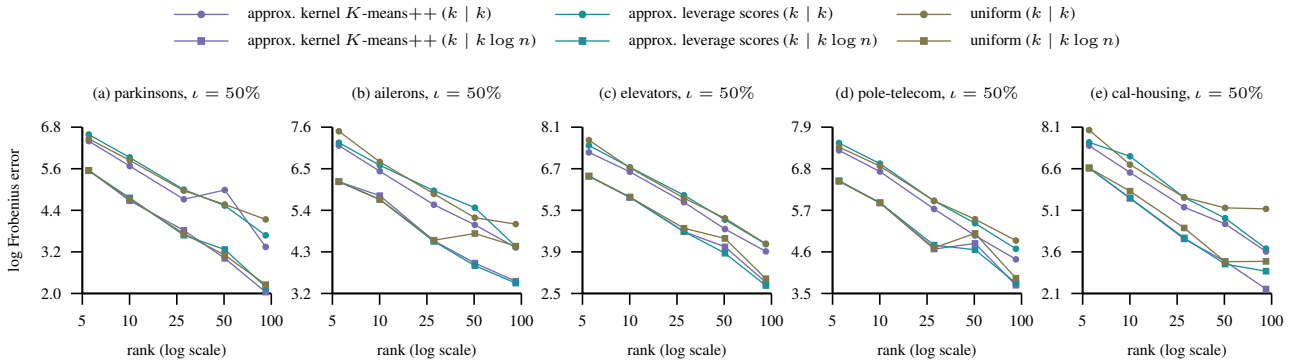

Figure 1: The figure shows the reduction in the approximation error for an indefinite kernel matrix defined as the difference between two Gaussian kernels, which comes as a result of the increase in the approximation rank. In the figure legend, we use $(k \mid l)$ to express the fact that a rank k approximation of the kernel matrix is computed using a set of l landmarks.

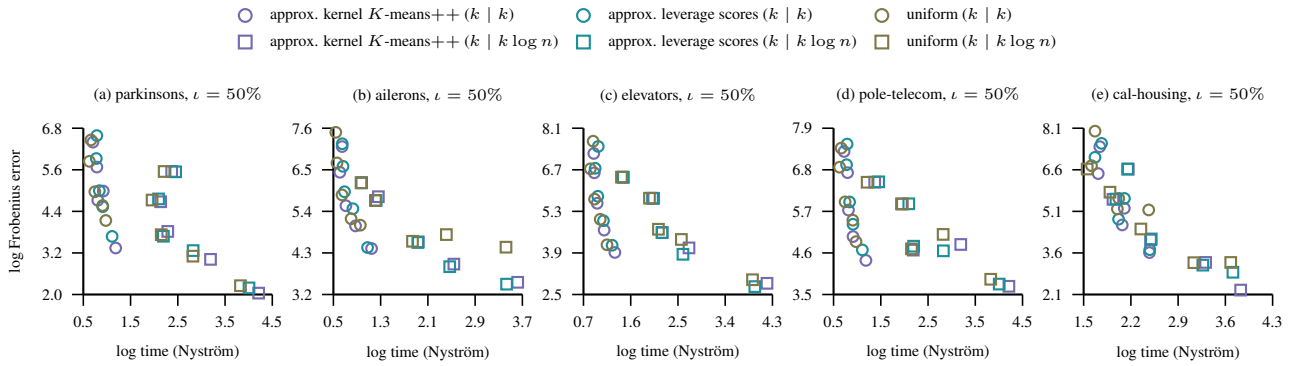

Figure 2: The figure shows the approximation errors of Nyström low-rank approximations (with different approximation ranks) as a function of time required to compute these approximations. In the figure legend, we again use $(k \mid l)$ to express the fact that a rank k approximation of the kernel matrix is computed using a set of l landmarks.

While at the first glance this result seems incremental, it makes an important contribution towards understanding the Kreĭn support vector machines (Loosli et al., 2016). In particular, the discussion of experiments in Loosli et al. (2016) differentiates between the Kreĭn support vector machines and the flip-spectrum approach. This happens despite the illustration indicating that they produce identical hypotheses in synthetic experiments (e.g., see Figures 3 and 4 in Loosli et al., 2016, and the discussion therein).

3. Experiments

In this section, we report the results of experiments aimed at demonstrating the effectiveness of: *i*) the Nyström method in low-rank approximations of indefinite kernel matrices, and *ii*) the described scalable Kreĭn approaches in classification tasks with pairwise (dis)similarity matrices.

In the first set of experiments, we take several datasets from UCI and LIACC repositories and define kernel matrices on them using the same indefinite kernels as previous work (Oglic and Gärtner, 2018, Appendix D). We use

$$0 \leq \iota = \textstyle\sum_{\{i:\ \lambda_i < 0\}} |\lambda_i| / \sum_i |\lambda_i| \leq 1$$

to quantify the level of indefiniteness of a kernel matrix. Prior to computation of kernel matrices, all the data matrices were normalized to have mean zero and unit variance across features. Following this, we have applied the Nyström method with landmark selection strategies presented in Section 2.3 to derive approximations with different ranks. We measure the effectiveness of a low-rank approximation with its error in the Frobenius norm. To quantify the effectiveness of the approximate eigendecomposition of the kernel matrix (i.e., the one-shot Nyström method) derived in Section 2.2, we have performed rank k approximations using sets of $k \log n$ landmarks. Figures 1 and 2 summarize the results obtained with an indefinite kernel defined by the difference between two Gaussian kernels. The reported error/time is the median error/time over 10 repetitions of the experiment. Figure 1 indicates a sharp (approximately exponential) decay in the approximation error as the rank

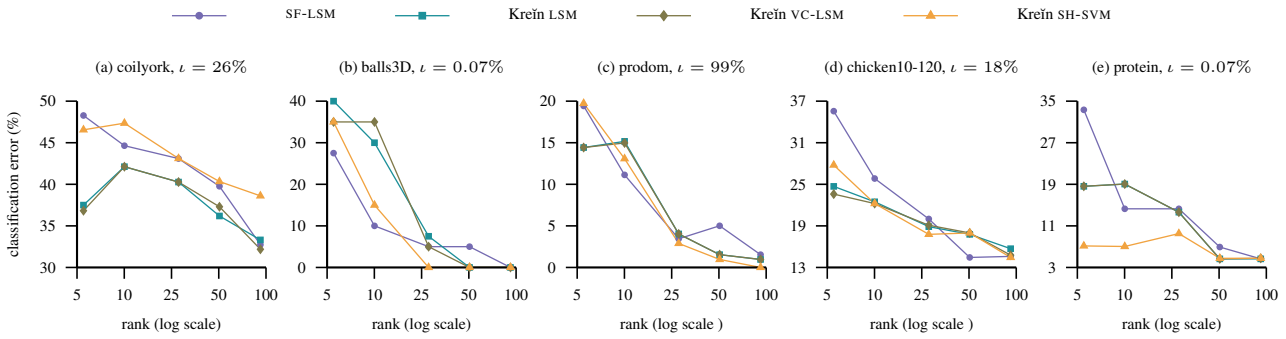

Figure 3: The figure shows the reduction in the classification error as the approximation rank of a Nyström low-rank approximation increases. The reported error is the median classification error obtained using 10-fold stratified cross-validation.

of the approximation increases. The devised approximate kernel K-means++ sampling strategy performs the best in terms of the accuracy in the experiments where rank k approximations are generated using k landmarks. The approximate leverage score strategy is quite competitive and in rank k approximations generated using $k \log n$ landmarks it performs as good or even better than the approximate kernel K-means++ sampling scheme. Oglic and Gärtner (2017) have evaluated these two state-of-the-art strategies on the low-rank approximation of positive definite kernels. In contrast to that work, we had to resort to an approximate kernel K-means++ sampling scheme because of the indefiniteness of the bilinear form defining a Kreĭn space. As a result of this, we can observe the lack of a gap between the curves describing the two sampling strategies, compared to the results reported in Oglic and Gärtner (2017) for positive definite kernels. Our hypothesis is that this is due to sub-optimal choices of landmarks that define sketch matrices. In our simulations, we have generated sketches by sampling the corresponding landmarks uniformly at random. In support of this hypothesis, rather large approximation errors for uniformly selected landmarks in approximation of other indefinite kernels can be observed (see Appendix C). Figure 2 reports the time required to generate a Nyström low-rank approximation and indicates that the considered sampling strategies amount to only a small fraction of the total time required to generate the low-rank approximation.

In the second set of experiments, we evaluate the effectiveness of the proposed least square methods and the support vector machine on classification tasks[1] with pairwise dissimilarity matrices (Pekalska and Duin, 2005; Duin and Pekalska, 2009). Following the instructions in Pekalska and Haasdonk (2009), the dissimilarity matrices are converted to similarities by applying the transformation characteristic to multi-dimensional scaling (e.g., see the negative double-centering transformation in Cox and Cox, 2000). In each simulation, we perform 10-fold stratified cross-validation and measure the effectiveness of an approach with the average/median percentage of misclassified examples. For multi-class problems, we only evaluate the effectiveness of a single binary *one-vs-all* classifier (just as in Oglic and Gärtner, 2018, Appendix C). Figure 3 shows the reduction in the classification error as the approximation rank increases. The reported error is the median error over 10-folds. Here, SF-LSM represents the baseline in which similarities are used as features and a linear ridge regression model is trained in that instance space (Chen et al., 2009; Alabdulmohsin et al., 2015). The figure indicates that the baseline is quite competitive, but overall the proposed low-rank variants perform very well across different datasets (additional plots are provided in Appendix C). Tables 1 provides the detailed results over all the datasets. In Appendix C (Table 2), we also compare the effectiveness of our low-rank approaches with respect to the relevant state-of-the-art methods which make no approximations and represent hypotheses via the span of kernel functions centered at training instances. The empirical results indicate a competitive performance of our low-rank approaches with only 100 landmarks across all the datasets and a variety of indefinite kernel functions.

4. Discussion

The Nyström method has recently been used for approximate eigendecomposition and low-rank approximation of indefinite kernel matrices (Gisbrecht and Schleif, 2015; Schleif and Tiňo, 2015; Schleif et al., 2016). To circumvent the fact that the original derivations of the approach are restricted to positive definite Mercer kernels (Smola and Schölkopf, 2000; Williams and Seeger, 2001), Gisbrecht and Schleif (2015) provide a derivation of the approach based on approximations of integral eigenfunctions arising in an eigendecomposition of an indefinite kernel. In particular, the authors of that work introduce an integral operator defined with an indefinite kernel and its empirical/sample-based approximation which asymptotically converges to the original (indefinite) integral operator. Based on this re-

[1] http://prtools.org/disdatasets/index.html

DATASET	DISSIMILARITY TYPE	RANK 100 APPROXIMATION							
		KREĬN VC-LSM		KREĬN LSM		KREĬN SH-SVM		SF-LSM	
coilyork	Graph matching	32.22	(±7.89)	**31.21**	(±5.28)	38.20	(±7.20)	35.33	(±10.09)
balls 3D	Shortest distance between balls	1.00	(±2.00)	0.50	(±1.50)	**0.00**	(±0.00)	0.50	(±1.50)
prodom	Structural alignment of proteins	0.92	(±0.46)	0.92	(±0.46)	**0.54**	(±0.47)	1.57	(±0.58)
chicken10	String edit distance	16.35	(±4.31)	15.69	(±4.97)	16.82	(±6.57)	**14.37**	(±4.02)
protein	Structural alignment of proteins	4.19	(±2.47)	**3.72**	(±2.76)	5.23	(±2.89)	5.15	(±3.91)
zongker	Deformable template matching	17.70	(±2.06)	17.75	(±2.23)	**15.30**	(±3.39)	17.05	(±2.36)
chicken25	String edit distance	19.29	(±4.64)	20.41	(±4.09)	25.77	(±4.68)	**18.17**	(±6.67)
pdish57	Hausdorff distance	3.40	(±0.39)	3.40	(±0.42)	**2.73**	(±0.62)	3.03	(±0.67)
pdism57	Hausdorff distance	0.38	(±0.26)	0.38	(±0.26)	**0.30**	(±0.29)	0.63	(±0.42)
woody50	Plant leaves' shape dissimilarity	30.84	(±5.25)	30.47	(±5.54)	38.42	(±7.13)	**26.41**	(±4.42)

Table 1: The table reports the results of our experiment on benchmark datasets for learning with indefinite kernels (Pekalska and Duin, 2005). The goal of the experiment is to evaluate the effectiveness of the state-of-the-art approaches for scalable learning in reproducing kernel Kreĭn spaces on classification tasks with pairwise dissimilarity matrices. We measure the effectiveness of an approach using the average classification error obtained using 10-fold stratified cross-validation (standard deviations are given in the brackets).

sult, Gisbrecht and Schleif (2015) provide a derivation of the Nyström method for indefinite kernels that treats the approximate equalities arising in the approximations of integral eigenfunctions as if they were exact. While such an assumption might hold for some datasets it fails to hold in the general case and this fact makes their extension of the Nyström method to indefinite kernels mathematically incomplete. Our derivation of the approach does not rely on such an assumption and, thus, provides a stronger result. Moreover, our proof is much simpler than the one in Gisbrecht and Schleif (2015) and provides a geometrical intuition for the approximation.

In addition to this, Gisbrecht and Schleif (2015); Schleif and Tiňo (2015); Schleif et al. (2016) proposed a method for finding an approximate low-rank eigendecomposition of an indefinite kernel matrix (for the sake of completeness, we review this approach in Appendix B). From the perspective of the exact number of floating point operations (FLOPs), the approach by Gisbrecht and Schleif (2015); Schleif and Tiňo (2015); Schleif et al. (2016) requires 7 matrix-to-matrix multiplications (each with the cost of m^2n FLOPs) and 2 eigendecompositions (each with the cost of m^3 FLOPs). Thus, in total their approach requires $7m^2n + 2m^3$ FLOPs to find an approximate low-rank eigendecomposition of an indefinite kernel matrix. In contrast to this, the approach proposed in Section 2.2 comes with a much better runtime complexity and requires at most $3m^2n + 3m^3$ FLOPs. To see a practical runtime benefit of our approach, take a problem of approximating the kernel matrix defined with $n = 10^6$ instances using $m = 10^3$ landmarks. Our method for approximate low-rank eigendecomposition requires 3×10^{12} less FLOPs than the approach proposed by Gisbrecht and Schleif (2015); Schleif and Tiňo (2015); Schleif et al. (2016).

Beside the considered low-rank approximations, it is possible to treat indefinite similarity functions as features and learn with linear models (Alabdulmohsin et al., 2015; Chen et al., 2009) or squared kernel matrices (Graepel et al., 1998). However, Balcan et al. (2008) have showed that learning with a positive definite kernel corresponding to a feature space where the target concept is separable by a linear hypothesis yields a larger margin compared to learning with a linear model in a feature space constructed using that kernel function. As a result, if a kernel is used to construct a feature representation the sample complexity of a linear model in that space might be higher compared to learning with a kernelized variant of regularized risk minimization.

The effectiveness of a particular landmark selection strategy is a problem studied separately from the derivation of the Nyström method and we, therefore, do not focus on that problem in this work. However, clustering and leverage score sampling have been proposed and validated in earlier publications and are state-of-the-art for low-rank approximation of positive definite kernels (Kumar et al., 2012; Alaoui and Mahoney, 2015; Gittens and Mahoney, 2016; Oglic and Gärtner, 2017). As the flip-spectrum matrix shares the eigenspace with the indefinite kernel matrix, the convergence results on the effectiveness of landmark selection strategies for Nyström low-rank approximation of positive definite kernels apply to indefinite kernels (e.g., see Section 2.3 or Eckart and Young, 1936; Mirsky, 1960). In particular, bounds for the leverage score sampling strategy applied to the flip-spectrum matrix carry over to our derivation of the Nyström method for indefinite kernels.

We conclude with a reference to Schleif et al. (2018) and Loosli et al. (2016), where an issue concerning the sparsity of a solution returned by the Kreĭn support vector machine has been raised. We hypothesize that our approach can overcome this limitation by either controlling the approximation rank or penalizing the low-rank objective with the ℓ_1-norm of the linear model. We leave the theoretical study and evaluation of such an approach for future work.

Acknowledgment: We are grateful for access to the University of Nottingham High Performance Computing Facility. Dino Oglic was supported in part by EPSRC grant EP/R012067/1.

References

Alabdulmohsin, I., Gao, X., and Zhang, X. Z. (2015). Support vector machines with indefinite kernels. In *Proceedings of the Sixth Asian Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR.

Alaoui, A. E. and Mahoney, M. W. (2015). Fast randomized kernel ridge regression with statistical guarantees. In *Advances in Neural Information Processing Systems 28*.

Alpay, D. (1991). Some remarks on reproducing kernel Kreĭn spaces. *Rocky Mountain Journal of Mathematics*.

Arthur, D. and Vassilvitskii, S. (2007). K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*.

Azizov, T. Y. and Iokhvidov, I. S. (1981). Linear operators in spaces with indefinite metric and their applications. *Journal of Soviet Mathematics*.

Balcan, M.-F., Blum, A., and Srebro, N. (2008). A theory of learning with similarity functions. *Machine Learning*.

Bognár, J. (1974). *Indefinite inner product spaces*. Ergebnisse der Mathematik und ihrer Grenzgebiete. Springer.

Chapelle, O., Vapnik, V., Bousquet, O., and Mukherjee, S. (2002). Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1-3):131–159.

Chen, Y., Garcia, E. K., Gupta, M. R., Rahimi, A., and Cazzanti, L. (2009). Similarity-based classification: Concepts and algorithms. *Journal of Machine Learning Research*.

Cox, T. F. and Cox, M. A. A. (2000). *Multidimensional Scaling*. Chapman and Hall/CRC, 2nd edition.

Drineas, P., Kannan, R., and Mahoney, M. W. (2006). Fast Monte Carlo algorithms for matrices II: Computing a low-rank approximation to a matrix. *SIAM Journal on Computing*.

Drineas, P. and Mahoney, M. W. (2005). On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*.

Duin, R. P. and Pekalska, E. (2009). Datasets and tools for dissimilarity analysis in pattern recognition. *Beyond Features: Similarity-Based Pattern Analysis and Recognition*.

Eckart, C. and Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218.

Fowlkes, C., Belongie, S., Chung, F., and Malik, J. (2004). Spectral grouping using the Nyström method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):214–225.

Gander, W., Golub, G. H., and von Matt, U. (1989). A constrained eigenvalue problem. *Linear Algebra and its Applications*.

Gisbrecht, A. and Schleif, F.-M. (2015). Metric and non-metric proximity transformations at linear costs. *Neurocomputing*, 167:643–657.

Gittens, A. and Mahoney, M. W. (2016). Revisiting the Nyström method for improved large-scale machine learning. *Journal Machine Learning Research*.

Graepel, T., Herbrich, R., Bollmann-Sdorra, P., and Obermayer, K. (1998). Classification on pairwise proximity data. In *Advances in Neural Information Processing Systems 11*.

Iokhvidov, I. S., Kreĭn, M. G., and Langer, H. (1982). *Introduction to the spectral theory of operators in spaces with an indefinite metric*. Berlin: Akademie-Verlag.

Keerthi, S. S. and DeCoste, D. (2005). A modified finite Newton method for fast solution of large scale linear SVMs. *Journal of Machine Learning Research*, 6:341–361.

Kumar, S., Mohri, M., and Talwalkar, A. (2012). Sampling methods for the Nyström method. *Journal of Machine Learning Research*.

Langer, H. (1962). Zur Spektraltheorie J-selbstadjungierter Operatoren. *Mathematische Annalen*.

Loosli, G., Canu, S., and Ong, C. S. (2016). Learning SVM in Kreĭn spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Mangasarian, O. L. (2002). A finite Newton method for classification. *Optimization Methods and Software*, 17:913–929.

Mirsky, L. (1960). Symmetric gauge functions and unitarily invariant norms. *Quaterly Journal of Mathematics, Oxford II. Series*, 11:50–59.

Nyström, E. J. (1930). Über die praktische Auflösung von Integralgleichungen mit Anwendungen auf Randwertaufgaben. *Acta Mathematica*.

Oglic, D. (2018). *Constructive Approximation and Learning by Greedy Algorithms*. PhD thesis, University of Bonn, Germany.

Oglic, D. and Gärtner, T. (2017). Nyström method with kernel K-means++ samples as landmarks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*. PMLR.

Oglic, D. and Gärtner, T. (2018). Learning in reproducing kernel Kreĭn spaces. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*. PMLR.

Ong, C. S., Mary, X., Canu, S., and Smola, A. J. (2004). Learning with non-positive kernels. In *Proceedings of the Twenty-First International Conference on Machine Learning*.

Pekalska, E. and Duin, R. P. W. (2005). *The Dissimilarity Representation for Pattern Recognition: Foundations And Applications (Machine Perception and Artificial Intelligence)*. World Scientific Publishing Co., Inc.

Pekalska, E. and Haasdonk, B. (2009). Kernel discriminant analysis with positive definite and indefinite kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(6).

Rifkin, R. M. (2002). *Everything Old is New Again: A Fresh Look at Historical Approaches in Machine Learning*. PhD thesis.

Schleif, F., Gisbrecht, A., and Tiňo, P. (2016). Probabilistic classifiers with low rank indefinite kernels. *arXiv preprint arXiv:1604.02264*.

Schleif, F.-M., Raab, C., and Tino, P. (2018). Sparsification of indefinite learning models. In *Structural, Syntactic, and Statistical Pattern Recognition*, pages 173–183. Springer.

Schleif, F.-M. and Tiňo, P. (2015). Indefinite proximity learning: A review. *Neural Computation*, 27(10):2039–2096.

Schleif, F.-M. and Tino, P. (2017). Indefinite core vector machine. *Pattern Recognition*, 71:187–195.

Schölkopf, B. and Smola, A. J. (2001). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT Press.

Schwartz, L. (1964). Sous-espaces hilbertiens d'espaces vectoriels toplogiques et noyaux associés (noyaux reproduisants). *Journal d'Analyse Mathematique*.

Smola, A. J. and Schölkopf, B. (2000). Sparse greedy matrix approximation for machine learning. In *Proceedings of the 17th International Conference on Machine Learning*.

Tikhonov, A. N. and Arsenin, V. Y. (1977). *Solutions of ill-posed problems*. W. H. Winston, Washington D. C.

Williams, C. K. I. and Seeger, M. (2001). Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems 13*.

A. Proofs

Proposition 2. *If the Kreĭn kernel ridge regression problem is regularized via the norm $\|\cdot\|_{\mathcal{H}_\mathcal{K}}$ with $\lambda = \lambda_+ = \lambda_-$, then the optimal hypothesis is equivalent to that obtained with kernel ridge regression and the flip-spectrum matrix in place of an indefinite Kreĭn kernel matrix.*

Proof. The optimal hypothesis over training data satisfies

$$\begin{aligned} K\alpha^* = \; & UDU^\top U \left(DS + n\lambda\mathbb{I}\right)^{-1} U^\top USU^\top y = \\ & UDS \left(DS + n\lambda\mathbb{I}\right)^{-1} U^\top y = \\ & H \left(H + n\lambda\mathbb{I}\right)^{-1} y \; . \end{aligned}$$

Thus, if we only regularize with $\|f\|_{\mathcal{H}_\mathcal{K}}$ then the Kreĭn kernel ridge regression problem is equivalent to that with the flip-spectrum transformation combined with the corresponding out-of-sample extension. More formally, if $\alpha_H^* = \left(H + n\lambda\mathbb{I}\right)^{-1} y$ denotes the optimal solution of the kernel ridge regression with the flip-spectrum matrix H in place of the indefinite kernel matrix K then the predictions at out-of-sample test instances are given by

$$f\left(x\right) = k_x^\top P \alpha_H^* \; .$$

□

Proposition 3. *The Kreĭn support vector machine (Loosli et al., 2016) is equivalent to the standard support vector machine with the flip-spectrum matrix in place of an indefinite Kreĭn kernel matrix (Graepel et al., 1998), combined with the out-of-sample transformation from Chen et al. (2009).*

Proof. The optimal hypothesis over training instances is

$$K\alpha^* = KP\alpha_H^* = UDU^\top USU^\top \alpha_H^* = H\alpha_H^* \; ,$$

where α_H^* is the optimal solution for the support vector machine problem with the flip-spectrum matrix H in place of the indefinite Kreĭn kernel matrix K. Thus, this variant of Kreĭn support vector machine is equivalent to learning with the flip-spectrum transformation of an indefinite kernel matrix. An out-of-sample extension for a test instance x is

$$f\left(x\right) = k_x^\top \alpha^* = k_x^\top P \alpha_H^* \; .$$

□

B. Discussion Addendum

We provide here a brief review of the approach by Gisbrecht and Schleif (2015); Schleif and Tiňo (2015); Schleif et al. (2016) for approximate eigendecomposition of an indefinite matrix [2]. The approach is motivated by the observation that an indefinite symmetric matrix and its square have identical eigenvectors. For this reason, the authors first form the squared low-rank Kreĭn kernel matrix

$$\tilde{K}^2 = K_{X\times Z} K_{Z\times Z}^{-1} K_{Z\times X} K_{X\times Z} K_{Z\times Z}^{-1} K_{Z\times X} \; .$$

The matrix $A = K_{Z\times Z}^{-1} K_{Z\times X} K_{X\times Z} K_{Z\times Z}^{-1}$ is positive definite because it can be written as $LL^\top$ (e.g., taking $L = K_{Z\times Z}^{-1} K_{Z\times X}$). Thus, all the eigenvalues in an eigendecomposition of $A = V\Gamma V^\top$ are non-negative and we can set $A = LL^\top$ with $L = V\Gamma^{\frac{1}{2}}$. From here it then follows that the matrix $\tilde{K}^2$ can be factored as $\tilde{K}^2 = BB^\top$ with $B = K_{X\times Z} V\Gamma^{\frac{1}{2}}$. Following this, Gisbrecht and Schleif (2015); Schleif and Tiňo (2015); Schleif et al. (2016) mimic the standard procedure for the derivation of approximate eigenvectors and eigenvalues characteristic to the Nyström method for positive definite kernels (Williams and Seeger, 2001; Fowlkes et al., 2004; Drineas and Mahoney, 2005; Drineas et al., 2006). In particular, they first decompose the positive definite matrix $B^\top B = Q\Delta Q^\top$ and then compute the approximate eigenvectors of $\tilde{K}^2$ as $\tilde{U} = BQ\Delta^{-\frac{1}{2}}$. Now, to obtain an approximate eigendecomposition of the Kreĭn kernel matrix the authors use these eigenvectors in combination with the posited form of the low-rank approximation $\tilde{K} = K_{X\times Z} K_{Z\times Z}^{-1} K_{Z\times X}$ and compute the approximate eigenvalues as $\tilde{D} = \tilde{U}^\top \tilde{K} \tilde{U}$. As the diagonal matrix Δ contains the eigenvalues of $\tilde{K}^2$ this step retrieves the signed eigenvalues of $\tilde{K}$. The *one-shot* Nyström approximation of the kernel matrix is then given as (Gisbrecht and Schleif, 2015; Schleif and Tiňo, 2015; Schleif et al., 2016)

$$\begin{aligned} & K_{X|Z}^{\text{SGT}^*} = \tilde{U}\tilde{D}\tilde{U}^\top = \\ & K_{X\times Z} V\Gamma^{\frac{1}{2}} Q\Delta^{-\frac{1}{2}} \; \tilde{D} \; \Delta^{-\frac{1}{2}} Q^\top \Gamma^{\frac{1}{2}} V^\top K_{Z\times X} \; . \end{aligned}$$

[2] https://www.techfak.uni-bielefeld.de/~fschleif/eigenvalue_corrections_demos.tgz, accessed in May 2018

C. Additional Experiments

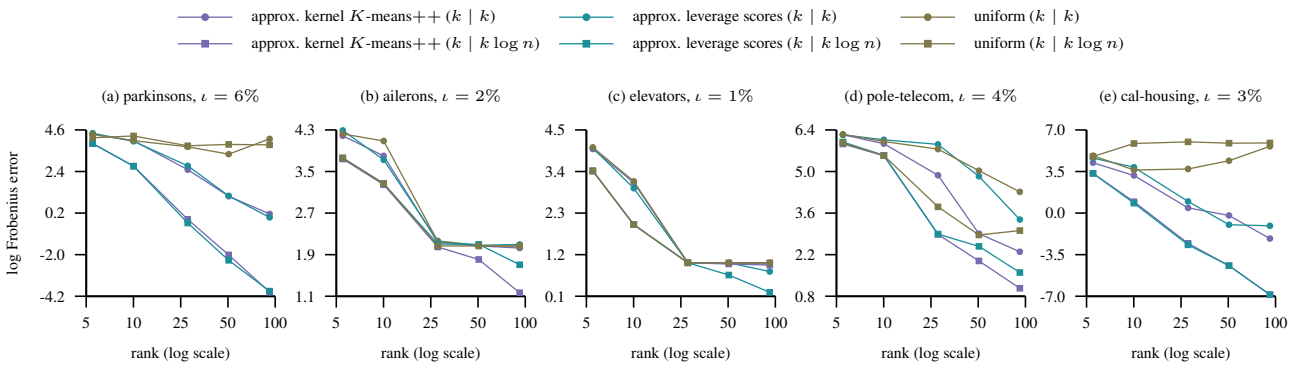

Figure 4: The figure shows the reduction in the approximation error for an indefinite kernel matrix obtained using the SIGMOID kernel (Oglic and Gärtner, 2018, Appendix D), which comes as a result of the increase in the approximation rank. In the figure legend, we use $(k \mid l)$ to express the fact that a rank k approximation of the kernel matrix is computed using a set of l landmarks.

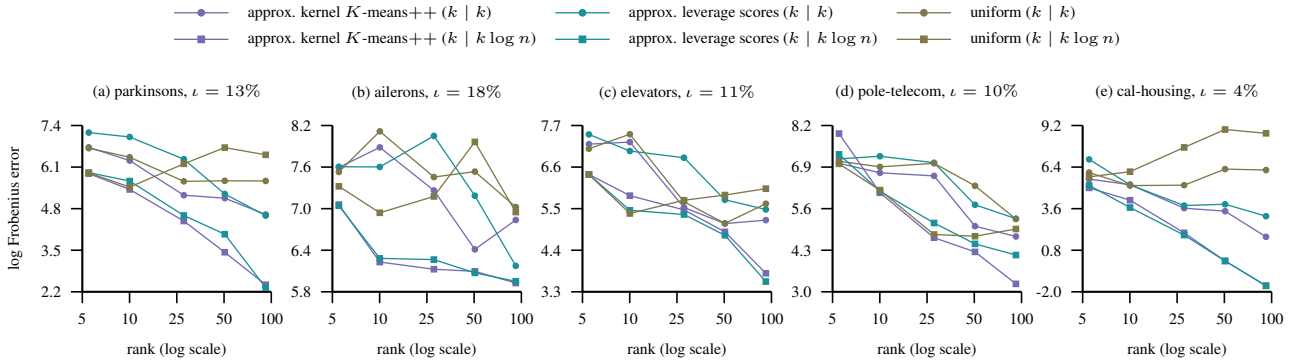

Figure 5: The figure shows the reduction in the approximation error for an indefinite kernel matrix obtained using the RL-SIGMOID kernel (Oglic and Gärtner, 2018, Appendix D), which comes as a result of the increase in the approximation rank. In the figure legend, we use $(k \mid l)$ to express the fact that a rank k approximation of the kernel matrix is computed using a set of l landmarks.

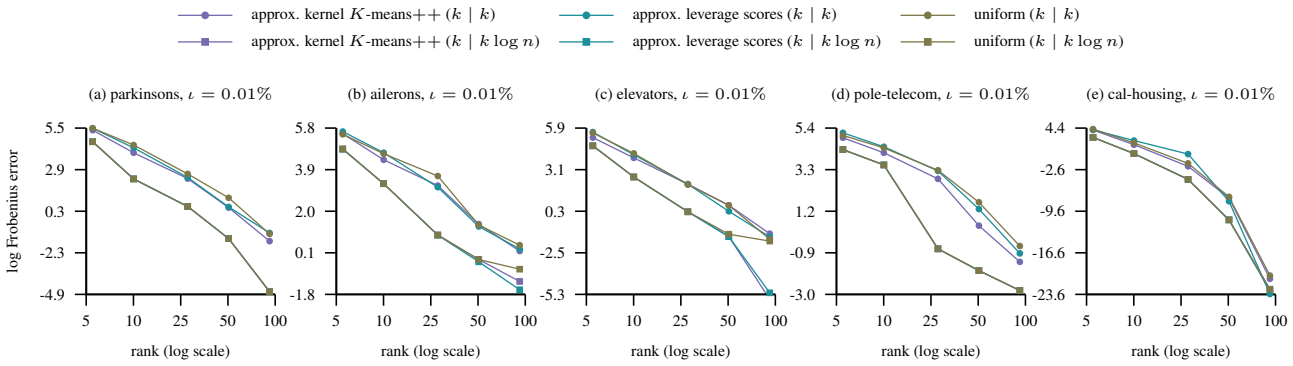

Figure 6: The figure shows the reduction in the approximation error for an indefinite kernel matrix obtained using the EPANECHNIKOV kernel (Oglic and Gärtner, 2018, Appendix D), which comes as a result of the increase in the approximation rank. In the figure legend, we use $(k \mid l)$ to express the fact that a rank k approximation of the kernel matrix is computed using a set of l landmarks.

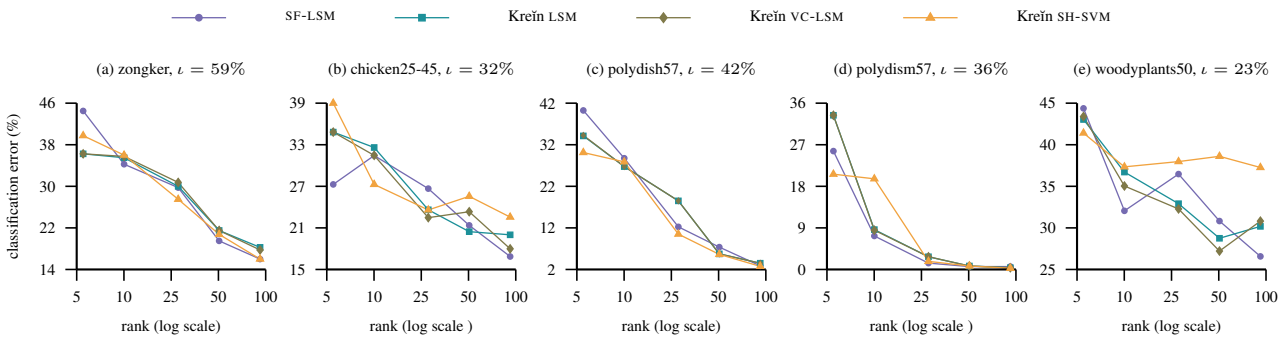

Figure 7: The figure shows the reduction in the classification error as the approximation rank of a Nyström low-rank approximation increases. The reported error is the median classification error obtained using 10-fold stratified cross-validation.

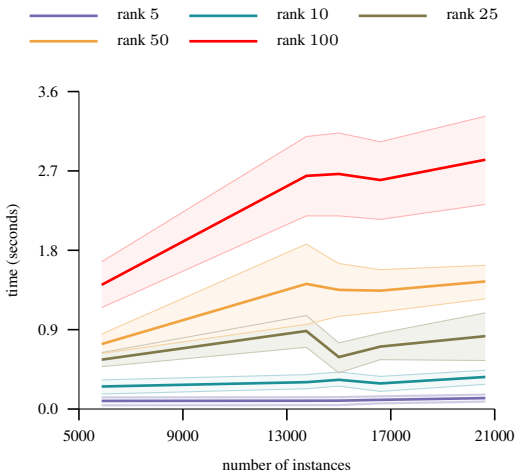

Figure 8: The figure depicts the computational cost for Nyström approximations of different ranks as a function of the number of instances in a dataset. The reported time is the average time required to compute the Nyström approximation of a given rank (averaged over 10 repetitions of the experiment). The confidence interval for a *cost-curve* is computed by subtracting/adding the corresponding standard deviations from the average computational costs. The landmarks are selected using the approximate kernel K-means++ sampling strategy proposed in Section 2.3. For all the considered approximation ranks, the *cost-curves* indicate that the approach scales (approximately) linearly with respect to the dataset size (the slope of a *cost-curve* depends on the approximation rank).

Dataset	Rank 100 Approximation (Primal Optimization)								Whole Kreĭn Space (Dual Optimization)					
	Kreĭn VC-LSM		Kreĭn LSM		Kreĭn SH-SVM		SF-LSM		K-SVM		SF-LSM		VC-LSM	
coilyork	32.22	(±7.89)	31.21	(±5.28)	38.20	(±7.20)	35.33	(±10.09)	32.91	(±8.06)	26.03	(±5.60)	22.56	(±7.66)
balls 3D	1.00	(±2.00)	0.50	(±1.50)	0.00	(±0.00)	0.50	(±1.50)	0.00	(±0.00)	0.00	(±0.00)	0.00	(±0.00)
prodom	0.92	(±0.46)	0.92	(±0.46)	0.54	(±0.47)	1.57	(±0.58)	0.00	(±0.00)	0.04	(±0.11)	0.00	(±0.00)
chicken10	16.35	(±4.31)	15.69	(±4.97)	16.82	(±6.57)	14.37	(±4.02)	30.95	(±7.81)	11.91	(±3.56)	5.62	(±2.55)
protein	4.19	(±2.47)	3.72	(±2.76)	5.23	(±2.89)	5.15	(±3.91)	5.17	(±3.34)	2.83	(±3.15)	0.00	(±0.00)
zongker	17.70	(±2.06)	17.75	(±2.23)	15.30	(±3.39)	17.05	(±2.36)	16.00	(±1.41)	5.60	(±1.20)	0.95	(±1.68)
chicken25	19.29	(±4.64)	20.41	(±4.09)	25.77	(±4.68)	18.17	(±6.67)	17.72	(±6.57)	16.38	(±5.14)	4.73	(±3.29)
pdish57	3.40	(±0.39)	3.40	(±0.42)	2.73	(±0.62)	3.03	(±0.67)	0.42	(±0.25)	0.20	(±0.19)	0.35	(±0.37)
pdism57	0.38	(±0.26)	0.38	(±0.26)	0.30	(±0.29)	0.63	(±0.42)	0.13	(±0.23)	0.15	(±0.17)	0.11	(±0.18)
woody50	30.84	(±5.25)	30.47	(±5.54)	38.42	(±7.13)	26.41	(±4.42)	37.04	(±5.07)	22.89	(±4.07)	2.53	(±2.66)

K-SVM denotes the Kreĭn support vector machine (Loosli et al., 2016).

SF-LSM denotes a variant of the least squares method with similarities as features (Graepel et al., 1998; Chen et al., 2009; Alabdulmohsin et al., 2015).

VC-LSM denotes the variance constrained least squares method (Oglic and Gärtner, 2018).

Table 2: The table reports the results of our experiments on benchmark datasets for learning with indefinite kernels (Pekalska and Duin, 2005). The effectiveness of an approach is measured using the average classification error obtained via 10-fold stratified cross-validation. In contrast to the experiments over the whole Kreĭn space (Oglic and Gärtner, 2018), the hyper-parameter optimization for low-rank approaches did not involve any random restarts (which could further improve the reported results for considered Kreĭn methods).