Scalable Learning in Reproducing Kernel Kreın 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 Kreın 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 Kreın 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 Kreın 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 Kreın spaces (Section 2.4). We also derive a novel low-rank variant of the support vector machine for scalable learning in reproducing kernel Kreın spaces (Section 2.5), inspired by Oglic and Gärtner (2018). Having introduced means for scalable learning in reproducing kernel Kreın 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 these approaches and the Nyström low-rank approximations for learning in reproducing kernel Kreın spaces (Section 2.5), inspired by Oglic and Gärtner (2018). Having introduced means for scalable learning in reproducing kernel Kreın 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 these approaches and the Nyström low-rank approximations on datasets from standard machine learning repositories (Section 3). The empirical results demonstrate 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 these approaches and the Nyström low-rank approximations on datasets from standard machine learning repositories (Section 3). The empirical results demonstrate 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 these approaches and the Nyström low-rank approximations on datasets from standard machine learning repositories (Section 3).

2. Scalable Learning in Kreın Spaces
In this section, we first provide a novel derivation of the Nyström method that allows us to extend the approach to...
low-rank approximation of indefinite kernel matrices. Building on this result, we then provide means to scale Krein 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 Krein 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 Krein 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 Krein Spaces

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

\[
\forall f \in K : \left( \forall g \in K \right) \langle f, g \rangle_K = 0 \implies f = 0.
\]

The bilinear form \( \langle \cdot, \cdot \rangle_K \) is symmetric if, for all \( f, g \in K \), we have \( \langle f, g \rangle_K = \langle g, f \rangle_K \). The form is called indefinite if there exists \( f, g \in K \) such that \( \langle f, f \rangle_K > 0 \) and \( \langle g, g \rangle_K < 0 \). On the other hand, if \( \langle f, f \rangle_K \geq 0 \) for all \( f \in K \), then the form is called positive. A non-degenerate, symmetric, and positive bilinear form on \( K \) is called an inner product. Any two elements \( f, g \in K \) that satisfy \( \langle f, g \rangle_K = 0 \) are called \( \langle \cdot, \cdot \rangle_K \)-orthogonal. Similarly, any two subspaces \( K_1, K_2 \subset K \) that satisfy \( \langle f_1, f_2 \rangle_K = 0 \) for all \( f_1 \in K_1 \) and \( f_2 \in K_2 \) are called \( \langle \cdot, \cdot \rangle_K \)-orthogonal. Having reviewed bilinear forms, we are now ready to introduce the notion of a Krein space.

**Definition 1.** (Bognár, 1974; Azizov and Iokhvidov, 1981) The vector space \( K \) with a bilinear form \( \langle \cdot, \cdot \rangle_K \) is called a Krein space if it admits a decomposition into a direct sum \( K = H_+ \oplus H_- \) of \( \langle \cdot, \cdot \rangle_K \)-orthogonal Hilbert spaces \( H_{\pm} \) such that the bilinear form can be written as

\[
\langle f, g \rangle_K = \langle f_+, g_+ \rangle_{H_+} - \langle f_-, g_- \rangle_{H_-},
\]

where \( H_{\pm} \) are endowed with inner products \( \langle \cdot, \cdot \rangle_{H_{\pm}} \), \( f = f_+ \oplus f_- \), \( g = g_+ \oplus g_- \), and \( f_\pm, g_\pm \in H_{\pm} \).

For a fixed decomposition \( K = H_+ \oplus H_- \), the Hilbert space \( H_K = H_+ \oplus H_- \) endowed with inner product

\[
\langle f, g \rangle_{H_K} = \langle f_+, g_+ \rangle_{H_+} + \langle f_-, g_- \rangle_{H_-}, \quad (f_\pm, g_\pm \in H_{\pm})
\]

can be associated with \( K \). For a Krein space \( K \), the decomposition \( K = H_+ \oplus H_- \) is not necessarily unique. Thus, a Krein space can, in general, be associated with infinitely many Hilbert spaces. However, for any such Hilbert space \( H_K \) the topology introduced on \( K \) via the norm \( \| f \|_{H_K} = \sqrt{\langle f, f \rangle_{H_K}} \) is independent of the decomposition and the associated Hilbert space. More specifically, all norms \( \| \cdot \|_{H_K} \) generated by different decompositions of \( K \) into direct sums of Hilbert spaces are topologically equivalent (Langer, 1962). The topology on \( K \) defined by the norm of an associated Hilbert space is called the strong topology. Having reviewed basic properties of Krein spaces, we are now ready to introduce a notion of reproducing kernel Krein space. For that, let \( \mathcal{X} \) be an instance space and denote with \( \mathbb{R}^m \) the set of functions from \( \mathcal{X} \) to \( \mathbb{R} \). For a fixed element \( x \in \mathcal{X} \), the map \( T_x : \mathbb{R}^m \rightarrow \mathbb{R} \) that assigns a real number to each function \( f \in \mathbb{R}^m \) is called the evaluation functional at \( x \), i.e., \( T_x(f) = f(x) \) for all \( f \in \mathbb{R}^m \).

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

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

**Theorem 1.** (Schwartz, 1964; Alpay, 1991) Let \( k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R} \) be a real-valued symmetric function. Then, there is an associated reproducing kernel Krein 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 Krein 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 span \( \{h(z_1, \cdot), \ldots, h(z_m, \cdot)\} \) and then approximates the kernel matrix \( H \) with entries \( H_{ij} = h(x_i, x_j) \sum_{j=1}^m \alpha_j h(z_{ij}, \cdot) \) 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

\[
\alpha^* = \arg \min_{\alpha \in \mathbb{R}^{m \times n}} \sum_{i=1}^n \left\| h(x_i, \cdot) - \sum_{j=1}^m \alpha_j h(z_j, \cdot) \right\|^2_{H_K}.
\]

While this approach works for positive definite kernels, it cannot be directly applied to reproducing Krein kernels. In
particular, let $\mathcal{K}$ be a reproducing kernel Krein space with an indefinite kernel $k: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The reproducing Krein 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 $(a - b, a - b)_{\mathcal{K}}$ 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 $\{ k(z_l, \cdot), \ldots, k(z_m, \cdot) \}$, 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)

$$k(x, \cdot) = \tilde{k}(x, \cdot) + k_1(x, \cdot),$$

where $\tilde{k}(x, \cdot) = \sum_{i=1}^{m} \alpha_i z_i$ with $\alpha_i \in \mathbb{R}^m$, and $\langle k_1(x, \cdot), \mathcal{L}_Z \rangle = 0$. For a landmark $z \in Z$, the inner product between the corresponding evaluation functional $k(z, \cdot)$ and $k(x, \cdot)$ then gives

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

Denote with $K_{Z \times Z}$ the block in the kernel matrix $K$ corresponding to landmarks $Z$ and let $k_x = \text{vec}(k(x, z_1), \ldots, k(x, z_m))$. From Eq. (2) it then follows that $k_x = K_{Z \times Z} z$. Thus, in Krein 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

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

Having computed the projection of a point onto the span of the landmarks in a Krein 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 Krein kernel matrix $K$ using the bilinear form on the span of the landmarks. More specifically, we have that

$$\tilde{k}(x_i, x_j) = \sum_{p=1}^{m} \alpha_{i,p}^* k(z_p, \cdot) \sum_{q=1}^{m} \alpha_{j,q}^* k(z_q, \cdot) = K_{Z \times Z}^{-1} k_{x_i, x_j}.$$

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

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

This approach for low-rank approximation of Krein 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 holds

$$\tilde{k}_x(x) = \text{vec}(\tilde{k}(x, x_1), \ldots, \tilde{k}(x, x_n)) = K_{X \times Z} K_{Z \times Z}^{-1} k_x.$$
approximate: i) statistical leverage scores for all instances, and/or ii) squared distances between instances in the feature space of the factorization \( \hat{H} = \hat{L}\hat{L}^\top \) with \( \hat{L} = \hat{U} |\Lambda|^{1/2} \). An approximate leverage score assigned to the \( i \)-th instance is given as the squared norm of the \( i \)-th row in the matrix \( \hat{U} \), that is \( \ell (x_i) = \| \hat{U} (i) \|^2 \) with \( 1 \leq i \leq n \). As the two matrices \( H \) and \( \hat{K} \) have identical eigenvectors, the approximate leverage scores obtained using the positive definite matrix \( H \) capture the informative part of the eigenspace of the indefinite matrix \( \hat{K} \). 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
\[
p_L^+ (x) = \min_{1 \leq i \leq n} \| x-z_i \|^2 / \sum_{i=1}^n \min_{1 \leq j \leq n} \| x_i - z_j \|^2 .
\]

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 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 Krejn kernel into a direct sum of Hilbert spaces. We start with a Krejn least squares method (KREIN LSM) which is a variant of kernel ridge regression, i.e.,
\[
f^* = \arg \min_{f \in \mathcal{K}} \frac{1}{n} \sum_{i=1}^n \left( f(x_i) - y_i \right)^2 + \lambda_+ \| f_+ \|^2_2 + \lambda_- \| f_- \|^2_2 ,
\]
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 \( f^* = \sum_{i=1}^n \alpha_i^* k(x_i, \cdot) \) with \( \alpha^* \in \mathbb{R}^n \). Applying the reproducing property of the Krejn kernel and setting the gradient of the objective to zero, we derive
\[
\alpha^* = \left( H + n\Lambda_\pm \right)^{-1} P y ,
\]
where \( K = U D U^\top, S = \text{sign} (D), H = U S D U^\top, P = U S U^\top, \) and \( \Lambda_\pm = \lambda_+ S_+ + \lambda_- |S_-| \) with \( S_\pm = (S \pm 1)/2 \).

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 Krejn 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 \( k_{\pm} \times \times X \) 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 transformation of \( H \) and the flip-spectrum matrix in place of an indefinite kernel matrix and Krejn kernel ridge regression regularized with a single hyperparameter.

**Proposition 2.** If the Krejn kernel ridge regression problem is regularized via the norm \( \| \cdot \|_{\mathcal{H}_+} \) 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 Krejn kernel matrix.

Having established this, we now proceed to formulate a Krejn regression problem with a low-rank approximation \( K_{\times \times Z} \) in place of the indefinite kernel matrix \( K \). More formally, after substituting the low-rank approximation into Krejn kernel ridge regression problem we transform it by
\[
z = |D_{Z \times Z}|^{-1/2} U_{Z \times Z}^\top K_{Z \times Z} x \alpha = L_{X \times Z} x \alpha ,
\]
\[
\Phi = K_{X \times Z} U_{Z \times Z} |D_{Z \times Z}|^{-1/2} S_{Z \times Z} = L_{X \times Z} S_{Z \times Z} ,
\]
\[
K_{X \times Z} x \alpha = L_{X \times Z} S_{Z \times Z} = \Phi z \quad \text{and} \quad \alpha^\top H_{\pm} \alpha = z_{\pm}^2 ,
\]
where \( H_{\pm} = L_{X \times Z} S_{Z \times Z, \pm} = L_{X \times Z} |S_{Z \times Z, \pm}| z \), and \( S_{Z \times Z, \pm} = (S_{Z \times Z, \pm} \pm 1)/2 \). Hence, we can write a low-rank variant of the Krejn kernel ridge regression problem as
\[
z^* = \arg \min_{z \in \mathbb{R}^m} \| \Phi z - y \|^2_2 + n \lambda_+ \| z_+ \|^2_2 + n \lambda_- \| z_- \|^2_2 .
\]

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

\[ \hat{f}^*(x) = k_x^T U_{Z \times Z} |D_{Z \times Z}|^{-1/2} S_{Z \times Z} z^* . \]

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

\[
\begin{align*}
\min_{f \in \mathcal{K}} & \quad \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2 + \lambda_+ \|f_+\|_2^2 + \lambda_- \|f_-\|_2^2 \\
\text{s.t.} & \quad \sum_{i=1}^n f(x_i)^2 = r^2 ,
\end{align*}
\]

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 Kreın kernel ridge regression, we can transform this problem into

\[
z^* = \arg\min_{z \in \mathbb{R}^m} n \lambda_+ \|z_+\|^2 + n \lambda_- \|z_-\|^2 - 2 z^T \Phi^T y
\]

\[
s.t. \quad z^T \Phi z = r^2 .
\]

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

\[
\gamma^* = \arg\min_{\gamma \in \mathbb{R}^m} n \gamma^T \Delta^{-1} B^T \Lambda_\pm B \Delta^{-1} \gamma - 2 (A^T y)^T \gamma
\]

\[
s.t. \quad \gamma^T \gamma = r^2 .
\]

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 \( O(m^3) \) and the cost for the low-rank transformation of the problem is \( O(m^3 + m^2 n) \). An out-of-sample extension can also be obtained by following the derivation for Kreın 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 Kreın 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 Kreın squared hinge support vector machine (KREIN SH-SVM)

\[
f^* = \arg\min_{f \in \mathcal{K}} \frac{1}{n} \sum_{i=1}^n \max \{1 - y_i f(x_i), 0\}^2 + \lambda_+ \|f_+\|_2^2 + \lambda_- \|f_-\|_2^2 .
\]

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

\[
f(x_i) = \tilde{k}_i^T \alpha = \tilde{k}_i^T K_{X \times Z}^{-1} K_{Z \times X} \alpha = k_i^T U_{Z \times Z} |D_{Z \times Z}|^{-1/2} S_{Z \times Z} z = \Phi_i z ,
\]

where \( \Phi_i \) denotes the \( i \)-th row in the matrix \( \Phi \). The low-rank variant of the approach can then be written as

\[
z^* = \arg\min_{z \in \mathbb{R}^m} \sum_{i=1}^n \max \{1 - y_i \Phi_i z, 0\}^2 + n \lambda_+ \|z_+\|^2 + n \lambda_- \|z_-\|^2 .
\]

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 Kreın 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 Kreın kernel matrix (Graepel et al., 1998), combined with the corresponding out-of-sample transformation (Chen et al., 2009).

**Proposition 3.** The Kreın 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 Kreın kernel matrix (Graepel et al., 1998), combined with the out-of-sample transformation from Chen et al. (2009).
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 LIAACC repositories and define kernel matrices on them using the same indefinite kernels as previous work (Oglic and Gärtner, 2018, Appendix D). We use

$$0 \leq \iota = \sum_{i : x_i \in \Omega} |\lambda_i|/\sum |\lambda_i| \leq 1$$

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 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 $(k \mid \ell)$ 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 $(k \mid \ell)$ 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).
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 Krein 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 squares method 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 Haasdink (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 Tiño, 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-
scalable, 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 \(7m^2n + 2m^3\) 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 \(3 \times 10^{12}\) 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 Kreın 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 \(\ell_1\)-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.

---

**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 Kreın 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).
References


### A. Proofs

**Proposition 2.** If the Kreın kernel ridge regression problem is regularized via the norm $\| \cdot \|_{\mathcal{H}_C}$ 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 Kreın kernel matrix.

**Proof.** The optimal hypothesis over training data satisfies

$$K \alpha^* = U D U^T (D S + n \lambda \mathbb{I})^{-1} U^T S U^T y = U D S (D S + n \lambda \mathbb{I})^{-1} U^T y = H (H + n \lambda \mathbb{I})^{-1} y.$$  

Thus, if we only regularize with $\| f \|_{\mathcal{H}_C}$ then the Kreın kernel ridge regression problem is equivalent to that with the flip-spectrum transformation combined with the corresponding out-of-sample extension. More formally, if $\alpha_H^* = (H + n \lambda \mathbb{I})^{-1} y$ 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

$$f(x) = k_x^T P \alpha_H^*.$$

**Proposition 3.** The Kreın 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 Kreın 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

$$K \alpha^* = K P \alpha_H^* = U D U^T U S U^T \alpha_H^* = H \alpha_H^*,$$

where $\alpha_H^*$ is the optimal solution for the support vector machine problem with the flip-spectrum matrix $H$ in place of the indefinite Kreın kernel matrix $K$. Thus, this variant of Kreın 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

$$f(x) = k_x^T \alpha^* = k_x^T P \alpha_H^*.$$

### B. Discussion Addendum

We provide here a brief review of the approach by Gisbrecht and Schleif (2015); Schleif and Tiño (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 Kreın kernel matrix

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

The matrix $A = K_{Z \times Z}^{-1} K_{X \times X} K_{X \times Z} K_{Z \times X}$ is positive definite because it can be written as $L L^T$ (e.g., taking $L = K_{Z \times Z}^{-1} K_{X \times X}$). Thus, all the eigenvalues in an eigendecomposition of $A = V T V^T$ are non-negative and we can set $A = L L^T$ with $L = V T^{\frac{1}{2}}$. From here it then follows that the matrix $K^2$ can be factored as $K^2 = B B^T$ with $B = K_{X \times Z} V T^{\frac{1}{2}}$. Following this, Gisbrecht and Schleif (2015); Schleif and Tiño (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 $B^T B = Q \Delta Q^T$ and then compute the approximate eigenvectors of $K^2$ as $\hat{U} = B Q \Delta^{-\frac{1}{2}}$. Now, to obtain an approximate eigendecomposition of the Kreın kernel matrix the authors use these eigenvectors in combination with the posited form of the low-rank approximation $K = K_{X \times Z} K_{Z \times X}^{-1} K_{X \times Z}$. Compute the approximate eigenvalues as $\hat{D} = \hat{U} \hat{D} \hat{U}^T$. As the diagonal matrix $\Delta$ contains the eigenvalues of $\hat{K}^2$ this step retrieves the signed eigenvalues of $K$. The one-shot Nyström approximation of the kernel matrix is then given as (Gisbrecht and Schleif, 2015; Schleif and Tiño, 2015; Schleif et al., 2016)

$$K^s = \hat{U} D \hat{U}^T = K_{X \times Z} V T^{\frac{1}{2}} Q \Delta^{-\frac{1}{2}} \hat{D} \Delta^{-\frac{1}{2}} Q^T V T \cdot K_{Z \times X} \cdot 2\text{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.
Scalable Learning in Reproducing Kernel Krein Spaces

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).

<table>
<thead>
<tr>
<th>DATASET</th>
<th>KREIN VC-LSM</th>
<th>KREIN LSM</th>
<th>KREIN SH-SVM</th>
<th>SF-LSM</th>
<th>K-SVM</th>
<th>SF-LSM</th>
<th>VC-LSM</th>
</tr>
</thead>
<tbody>
<tr>
<td>coilyork</td>
<td>32.22±7.89</td>
<td>31.21±5.28</td>
<td>38.20±7.20</td>
<td>35.53±10.09</td>
<td>32.91±8.00</td>
<td>26.83±5.60</td>
<td>22.56±7.60</td>
</tr>
<tr>
<td>balls3D</td>
<td>1.00±2.00</td>
<td>0.50±1.50</td>
<td>0.00±0.00</td>
<td>0.50±1.50</td>
<td>0.00±0.00</td>
<td>0.00±0.00</td>
<td>0.00±0.00</td>
</tr>
<tr>
<td>prodom</td>
<td>0.92±0.46</td>
<td>0.92±0.46</td>
<td>0.54±0.47</td>
<td>1.57±0.56</td>
<td>0.00±0.00</td>
<td>0.04±0.11</td>
<td>0.00±0.00</td>
</tr>
<tr>
<td>chicken10</td>
<td>16.35±4.31</td>
<td>15.69±4.97</td>
<td>16.82±6.57</td>
<td>14.37±4.02</td>
<td>30.95±7.81</td>
<td>11.91±3.56</td>
<td>5.62±2.55</td>
</tr>
<tr>
<td>protein</td>
<td>4.19±2.47</td>
<td>3.72±2.76</td>
<td>5.23±2.89</td>
<td>5.15±3.91</td>
<td>5.17±3.34</td>
<td>2.83±3.15</td>
<td>0.00±0.00</td>
</tr>
<tr>
<td>zongker</td>
<td>17.70±2.06</td>
<td>17.75±2.23</td>
<td>15.30±3.39</td>
<td>17.05±2.36</td>
<td>16.00±1.41</td>
<td>5.60±1.20</td>
<td>0.95±1.68</td>
</tr>
<tr>
<td>pdish57</td>
<td>3.40±0.39</td>
<td>3.40±0.42</td>
<td>2.73±0.62</td>
<td>3.03±0.67</td>
<td>0.42±0.25</td>
<td>0.20±0.19</td>
<td>0.35±0.37</td>
</tr>
<tr>
<td>pdism57</td>
<td>0.38±0.26</td>
<td>0.38±0.26</td>
<td>0.30±0.29</td>
<td>0.63±0.42</td>
<td>0.13±0.23</td>
<td>0.15±0.17</td>
<td>0.11±0.18</td>
</tr>
<tr>
<td>woody50</td>
<td>30.84±5.25</td>
<td>30.47±5.54</td>
<td>38.42±7.13</td>
<td>26.41±4.42</td>
<td>37.04±5.07</td>
<td>22.89±4.07</td>
<td>2.53±2.60</td>
</tr>
</tbody>
</table>

K-SVM denotes the Krein support vector machine (Looshi 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 Krein 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 Krein methods).