Learning in Reproducing Kernel Kreĭn Spaces

Dino Oglic [1 2] **Thomas Gärtner** [1]

Abstract

We formulate a novel regularized risk minimization problem for learning in reproducing kernel Kreĭn spaces and show that the strong representer theorem applies to it. As a result of the latter, the learning problem can be expressed as the minimization of a quadratic form over a hypersphere of constant radius. We present an algorithm that can find a globally optimal solution to this non-convex optimization problem in time cubic in the number of instances. Moreover, we derive the gradient of the solution with respect to its hyperparameters and, in this way, provide means for efficient hyperparameter tuning. The approach comes with a generalization bound expressed in terms of the Rademacher complexity of the corresponding hypothesis space. The major advantage over standard kernel methods is the ability to learn with various domain specific similarity measures for which positive definiteness does not hold or is difficult to establish. The approach is evaluated empirically using indefinite kernels defined on structured as well as vectorial data. The empirical results demonstrate a superior performance of our approach over the state-of-the-art baselines.

1. Introduction

We build on the work by Ong et al. (2004) and formulate a novel regularized risk minimization problem for learning in reproducing kernel Kreĭn spaces (reviewed in Section 2). The proposed risk minimization problem is of interest to several applications of machine learning (Laub & Müller, 2004) where the instance space can be accessed only implicitly, through a kernel function that outputs a real-value for a pair of instances. Typically, for a given set of instances the kernel matrix does not exhibit properties required by standard machine learning algorithms such as positive definiteness or metricity. A common practice in dealing with such data is to map the indefinite kernel matrix to a positive definite one using a spectrum transformation. This conversion can cause information loss and affect our ability to model a functional dependence of interest. In particular, Laub & Müller (2004) have used three real-world datasets to demonstrate that for symmetric kernel functions corresponding to indefinite kernel matrices, the negative parts of their spectra contain useful information which gets discarded by some of the standard procedures that learn by first transforming the indefinite kernel matrix to a positive definite one.

We show that the strong representer theorem applies to the proposed risk minimization problem and utilize this theoretical result to express the learning problem as the minimization of a quadratic form over a hypersphere of constant radius (Section 3.1). The optimization problem is, in general, neither convex nor concave and it can have exponentially many local optima (with respect to the representation size). Despite this, a globally optimal solution to this problem can be found in time cubic in the number of training examples. The algorithm for solving this nonconvex problem relies on the work by Forsythe & Golub (1965) and Gander et al. (1989), who were first to consider the optimization of a quadratic form over a hypersphere of constant radius. The proposed risk minimization problem is consistent and comes with a generalization bound expressed in terms of the Rademacher complexity of the corresponding hypothesis space, which is a subset of a reproducing kernel Kreĭn space of functions (Section 3.2). In Section 3.3, we derive the gradient of an optimal solution to the risk minimization problem with respect to the hyperparameters of the model (e.g., the regularization parameters, hypersphere radius, and/or kernel-specific parameters). The derived solution gradient allows one to tune the hyperparameters of the model using an off-the-shelf optimization algorithm (e.g., L-BFGS-B minimization procedure, available in most numerical packages). In Section 4, we place our work in the context of relevant existing approaches for learning in reproducing kernel Kreĭn spaces. The effectiveness of the approach is evaluated empirically using indefinite kernels defined on structured and vectorial data. The results show a superior performance of our approach over the state-of-the-art baselines and indicate that on some problems indefinite kernels can be more effective than the positive definite ones.

[1] School of Computer Science, University of Nottingham, UK [2] Institut für Informatik III, Universität Bonn, Germany. Correspondence to: Dino Oglic <dino.oglic@uni-bonn.de>.

Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

2. Reproducing Kernel Kreĭn Spaces

This section provides a brief overview of reproducing kernel Kreĭn spaces. The review follows closely the study by Azizov & Iokhvidov (1981) and the work by Ong et al. (2004). For a more extensive introduction, we refer to works by Bognár (1974) and Iokhvidov et al. (1982).

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) $\langle f, \alpha g+\beta h\rangle_{\mathcal{K}} = \alpha\langle f,g\rangle_{\mathcal{K}} + \beta\langle f,h\rangle_{\mathcal{K}}$.

For $f\in\mathcal{K}$, if $\langle f,g\rangle_{\mathcal{K}}=0$ for all $g\in\mathcal{K}$ implies that $f=0$, then the form is non-degenerate. 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. *(Azizov & Iokhvidov, 1981; Bognár, 1974) 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 $f_\pm,g_\pm\in\mathcal{H}_\pm$.

Thus, a Kreĭn space is defined with a non-degenerate, symmetric, and indefinite bilinear form. 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 the norms $\|\cdot\|_{\mathcal{H}_{\mathcal{K}}}$ generated by different decompositions of $\mathcal{K}$ into direct sums of Hilbert spaces are topologically equivalent (Langer, 1962). The topology on $\mathcal{K}$ defined by the norm of an associated Hilbert space is called the strong topology on $\mathcal{K}$. Henceforth, notions of convergence and continuity on a Kreĭn space are defined with respect to the strong topology. As the strong topology of a Kreĭn space is a Hilbert space topology, the Riesz representation theorem holds. More formally, for a continuous linear functional $\mathcal{L}$ on a Kreĭn space $\mathcal{K}$ there exists a unique $g\in\mathcal{K}$ such that the functional $\mathcal{L}$, for all $f\in\mathcal{K}$, can be written as $\mathcal{L}f=\langle f,g\rangle_{\mathcal{K}}$.

Having reviewed basic properties of Kreĭn spaces, we are now ready to introduce the notion of a 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 $\mathbb{R}$.

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. *(Alpay, 1991; Schwartz, 1964) 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.*

In contrast to reproducing kernel Hilbert spaces, there is no bijection between reproducing kernel Kreĭn spaces and indefinite reproducing kernels. Moreover, it is important to note that not every symmetric kernel function admits a representation as a difference between two positive definite kernels. A symmetric function that does not admit such a representation has been constructed by Schwartz (1964) and it can also be found in Alpay (Theorem 2.2, 1991). On finite discrete spaces, however, any symmetric kernel function admits a Kreĭn decomposition.

3. Regularized Risk Minimization in Reproducing Kernel Kreĭn Spaces

Building on the work by Ong et al. (2004), we first propose a novel regularized risk minimization problem for learning in reproducing kernel Kreĭn spaces and then show that the strong representer theorem applies to it (Section 3.1). The main difference compared to previous stabilization approaches due to Ong et al. (2004) is in the way the optimization problem accounts for the complexity of hypotheses. As a result of our representer theorem, the proposed regularized risk minimization problem defined over a reproducing kernel Kreĭn space can be transformed into a non-convex optimization problem over a Euclidean space. Following this, we build on the work by Gander et al. (1989) and show how to find a globally optimal solution to the transformed non-convex optimization problem. Having provided means for finding an optimal solution to the learning problem, we present a sample complexity bound (Ong et al., 2004) which shows that learning in a reproducing kernel Kreĭn space is consistent (Section 3.2). The section concludes with a procedure for the optimization of hyperparameters arising in our regularized risk minimization problem (Section 3.3).

3.1. Optimization Problem

We retain the notation from Section 2 and assume that a sample $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^n$ has been drawn independently from a Borel probability measure ρ defined on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, with $\mathcal{Y} \subset \mathbb{R}$. For an approximation of the target function $f_\rho(x) = \int y \, \mathrm{d}\rho(y \mid x)$, we measure the goodness of fit with the expected squared error in ρ, i.e., $\mathcal{E}_\rho(f) = \int (f(x) - y)^2 \, \mathrm{d}\rho$. The empirical counterpart of the error, defined over a sample $\mathbf{z} \in \mathcal{Z}^n$ is denoted with $\mathcal{E}_\mathbf{z}(f) = \frac{1}{n}\sum_{i=1}^n (f(x_i) - y_i)^2$.

Early attempts at defining a regularized risk minimization problem for learning in reproducing kernel Kreĭn spaces are based on the stabilization approach by Ong et al. (2004). We start with an instance of that approach where the stabilization is replaced with minimization over a reproducing kernel Kreĭn space. More formally, we refer to the following risk minimization problem over a reproducing kernel Kreĭn space as the OMCS-KREĬN problem (Ong et al., 2004)

$$\begin{aligned} \min_{f \in \mathcal{K}} \quad & \frac{1}{n}\sum_{i=1}^n (f(x_i) - y_i)^2 + \lambda \langle f, f \rangle_\mathcal{K} \\ s.t. \quad & \frac{1}{n}\sum_{i=1}^n \left(f(x_i) - \frac{1}{n}\sum_{j=1}^n f(x_j) \right)^2 = r^2 \ . \end{aligned} \tag{1}$$

The empirical squared error depends on $f \in \mathcal{K}$ only through its evaluations $f(x_i)$, with $1 \leq i \leq n$. Moreover, the squared error loss function is convex and, thus, satisfies the requirement on the loss function from the representer theorem for stabilization (Theorem 11, Ong et al., 2004). In Eq. (1), we choose the linear identity function as the stabilizer and constrain the solution space by matching the variance of the estimator f to an a priori specified hyperparameter. Thus, the OMCS-KREĬN problem satisfies the conditions from the representer theorem for stabilization (Ong et al., 2004) and any saddle point of the optimization problem in Eq. (1) admits the expansion as $f^* = \sum_{i=1}^n \alpha_i k(x_i, \cdot)$ with $\alpha_i \in \mathbb{R}$. This allows us to express the optimization problem from Eq. (1) in terms of the parameters $\alpha \in \mathbb{R}^n$. To simplify our derivations, we can without loss of generality assume that the kernel matrix K is centered, where $K_{ij} = k(x_i, x_j)$ for $1 \leq i, j \leq n$. Then, substituting $f = \sum_{i=1}^n \alpha_i k(x_i, \cdot)$ and using the reproducing property of the Kreĭn kernel k we can rewrite the optimization problem from Eq. (1) as

$$\begin{aligned} \min_{\alpha \in \mathbb{R}^n} \quad & \|K\alpha - y\|_2^2 + n\lambda^2 \ \alpha^\top K \alpha \\ s.t. \quad & \alpha^\top K^2 \alpha = nr^2 \ . \end{aligned} \tag{2}$$

The OMCS-KREĬN regularized risk minimization problem is non-convex and can have exponentially many local optima. Despite this, we subsequently show how to find a globally optimal solution to this problem in time cubic in the size of the kernel expansion. However, our empirical evaluation of the approach (presented in Section 5) demonstrates that it fails to generalize to unseen instances. As $\langle f, f \rangle_\mathcal{K} = \|f_+\|_{\mathcal{H}_+}^2 - \|f_-\|_{\mathcal{H}_-}^2$ does not define a norm, we suspect that the regularization term does not capture the complexity of hypotheses from the reproducing kernel Kreĭn space $\mathcal{K}$. To address this, we propose to penalize the complexity of hypotheses via decomposition components $\mathcal{H}_\pm$ and/or the strong topology on $\mathcal{K}$. More formally, we propose the following regularized risk minimization problem for learning in reproducing kernel Kreĭn spaces and henceforth refer to it as the KREĬN problem

$$\begin{aligned} \min_{f \in \mathcal{K}} \quad & \frac{1}{n}\sum_{i=1}^n (f(x_i) - y_i)^2 + \lambda_+ \|f_+\|_{\mathcal{H}_+}^2 + \lambda_- \|f_-\|_{\mathcal{H}_-}^2 \\ s.t. \quad & \frac{1}{n}\sum_{i=1}^n \left(f(x_i) - \frac{1}{n}\sum_{j=1}^n f(x_j) \right)^2 = r^2 \ . \end{aligned} \tag{3}$$

Having introduced our regularized risk minimization problem, we show that the following strong representer theorem applies to it (a proof is provided in Appendix A).

Theorem 2. *Let $f^* \in \mathcal{K}$ be an optimal solution to the* KREĬN *optimization problem from Eq. (3). Then, f^* admits the expansion $f^* = \sum_{i=1}^n \alpha_i k(x_i, \cdot)$ with $\alpha_i \in \mathbb{R}$.*

The representer theorem allows us to express the regularized risk minimization problem as an optimization problem over a Euclidean space. In particular, substituting $f = \sum_{i=1}^n \alpha_i k(x_i, \cdot)$ into Eq. (3) we deduce

$$\begin{aligned} \min_{\alpha \in \mathbb{R}^n} \quad & \|K\alpha - y\|_2^2 + n\alpha^\top \left(\lambda_+^2 \ K_+ + \lambda_-^2 \ K_- \right) \alpha \\ s.t. \quad & \alpha^\top K^2 \alpha = nr^2 \ , \end{aligned} \tag{4}$$

where $K_\pm$ are kernel matrices corresponding to disjoint reproducing kernel Hilbert spaces given by positive definite kernels $k_\pm$, $k = k_+ - k_-$, and $K = K_+ - K_-$.

The optimization problems in Eq. (2) and (4) are minimizing quadratic forms over hyperellipsoids with radius r and center at the origin. As such, the problems are non-convex even in the cases when the regularization term is defined with a positive definite matrix. Despite this, it is possible to find a globally optimal solution to such a problem using a method proposed by Gander et al. (1989). To simplify our presentation, we focus on our regularized risk minimization problem from Eq. (4) and note that the derivation for the OMCS-Kreĭn problem follows along these lines. First, we provide a proposition which is crucial for finding a globally optimal solution to the problem in Eq. (4). To this end, let us derive the Lagrangian of that optimization problem as

$$\begin{aligned}\mathcal{L}\left(\alpha,\mu\right) = \quad & \alpha^\top\left(\lambda_+^2\ K_+ + \lambda_-^2\ K_-\right)\alpha - 2y^\top K\alpha - \\ & \mu\left(\alpha^\top K^2\alpha - r^2\right),\end{aligned}$$

and denote with $\Theta\left(\alpha\right)$ the optimization objective in problem (4). If we now set the derivative of the Lagrangian to zero, we obtain the following two stationary constraints

$$\begin{aligned}&\left(\lambda_+^2\ K_+ + \lambda_-^2\ K_-\right)\alpha = Ky + \mu K^2\alpha \\ &\alpha^\top K^2\alpha = r^2\ .\end{aligned} \tag{5}$$

Having introduced all the relevant terms, we are now ready to characterize a globally optimal solution to problem (4).

Proposition 3. *(Forsythe & Golub, 1965; Gander et al., 1989) The optimization objective $\Theta\left(\alpha\right)$ attains its minimal value at the tuple $\left(\alpha^*,\mu^*\right)$ satisfying the stationary constraints (5) with the smallest value of μ. Analogously, the maximal value of $\Theta\left(\alpha\right)$ is attained at the tuple with the largest value of the Lagrange multiplier μ.*

Hence, instead of the original optimization problem (4) we can solve the system with two stationary equations (5) and minimal μ. Gander et al. (1989) propose two methods for solving such problems. In the first approach, the problem is reduced to a quadratic eigenvalue problem and afterwards transformed into a linear eigenvalue problem. In the second approach, the problem is reduced to solving a one-dimensional secular equation. The first approach is more elegant, as it allows us to compute the solution in a closed form. More specifically, the solution to problem (4) is given by (Gander et al., 1989)

$$\alpha^* = \left(\lambda_+^2\ P_+ - \lambda_-^2\ P_- - \mu^* K\right)^{-1} y\ , \tag{6}$$

where μ^* is the smallest real eigenvalue of the matrix

$$\left[\begin{array}{cc} \lambda_+^2\ K_+^\dagger + \lambda_-^2\ K_-^\dagger & -\mathbb{I} \\ {}^{-yy^\top}\!/_{r^2} & \lambda_+^2\ K_+^\dagger + \lambda_-^2\ K_-^\dagger \end{array}\right],$$

$P_\pm = V\mathbb{I}_\pm V^\top$, $K = V\Sigma V^\top$ is an eigendecomposition of K, and $\mathbb{I}_\pm$ are diagonal matrices with ones at places corresponding to positive/negative eigenvalues of K.

Despite its elegance, the approach requires us to: *i*) invert/decompose a positive definite matrix, and *ii*) decompose a non-symmetric block matrix of dimension $2n$, which is not a numerically stable task for every such matrix. Furthermore, the computed solution α^* highly depends on the precision up to which the optimal μ is computed and for an imprecise value the solution might not be on the correct hyperellipsoid at all (e.g., see Gander et al., 1989).

For this reason, we rely on the secular approach in the computation of the optimal solution. Gander et al. (1989) proposed an efficient algorithm for the computation of the optimal Lagrange multiplier to machine precision. For the sake of completeness (and brevity), we review this approach in Appendix B and in the remainder of the section describe how to derive the secular equation required to compute the optimal multiplier. First, we perform the eigendecomposition of the symmetric and indefinite kernel matrix $K = V\Sigma V^\top$. From this eigendecomposition, we derive the decompositions of matrices $K_\pm = V\Sigma_\pm V^\top$, where Σ_+/Σ_- are diagonal matrices with the absolute values of the positive/negative eigenvalues of K at their respective diagonals, padded with zeros. The decomposition of K allows us to transform the stationary constraints from Eq. (5) as

$$V\left(\lambda_+^2\ \Sigma_+^\dagger + \lambda_-^2\ \Sigma_-^\dagger\right)V^\top u = y + \mu u\ ,$$

where $u = K\alpha$, $u^\top u = r^2$, and $\Sigma_\pm^\dagger$ denote the pseudo-inverses of the diagonal matrices $\Sigma_\pm$. Then, this resulting equation is multiplied with the orthogonal matrix $V^\top$ from the left and transformed into

$$\left(\lambda_+^2\ \Sigma_+^\dagger + \lambda_-^2\ \Sigma_-^\dagger\right)\hat{u} = \hat{y} + \mu\hat{u}\ ,$$

with $\hat{y} = V^\top y$ and $\hat{u} = V^\top u$. From here, we deduce

$$\hat{u}_i(\mu) = {}^{\sigma_i\hat{y}_i}\!/_{(\lambda_{\mathrm{sign}(\sigma_i)}^2 - \mu\sigma_i)} \quad (i = 1, 2, ..., n)\ ,$$

and substitute the computed vector $\hat{u}\left(\mu\right) \in \mathbb{R}^n$ into the second stationary constraint to form the secular equation

$$g(\mu) = {}^{\sigma_i\hat{y}_i}\!/_{(\lambda_{\mathrm{sign}(\sigma_i)}^2 - \mu\sigma_i)} - r^2 = 0\ . \tag{7}$$

The optimal value of the parameter μ is the smallest root of this non-linear secular equation and the optimal solution to problem (4) is given by $u^* = V\hat{u}(\mu^*)$. Moreover, the interval at which the root lies is known (Gander et al., 1989). In particular, the quadratic term from Eq. (4) is a positive definite matrix and $\mu^* \in \left(-\infty, {}^{\lambda_+^2}\!/_{\sigma_+}\right)$, where σ_+ is the largest eigenvalue of the matrix $|K|$. On the other hand, the quadratic term from Eq. (2) is an indefinite matrix and $\mu^* \in \left(-\infty, {}^{\lambda_-^2}\!/_{\sigma_-}\right)$, where σ_- is the largest negative eigenvalue of the matrix K. The condition on the interval of the optimal Lagrange multiplier implies that the matrix defining the optimal solution u^* is positive semidefinite. Thus, the proposed regularized risk minimization problem is well-posed if $\mu^* \neq {}^{\lambda_\pm^2}\!/_{\sigma_\pm}$. The computational complexity of both approaches (secular and eigenvalue) is $O(n^3)$.

3.2. Generalization Bound

In this section, we present a generalization bound for learning in a reproducing kernel Kreĭn space using the proposed regularized risk minimization problem. The key to such bounds over Kreĭn spaces is to be able to quantify the complexity of a hypothesis space. In the considered case, this refers to a hypothesis space corresponding to problem (3).

We observe that as $n \to \infty$, for zero-mean hypotheses, the hard constraint in problem (3) converges to $\|\cdot\|_\rho^2$, where $\|\cdot\|_\rho$ denotes the norm in the space of square integrable functions defined on $\mathcal{X}$ in the measure ρ. Thus, we can under this assumption define our hypothesis space as

$$\mathcal{F} = \left\{ f \in \mathcal{K} \mid \|f\|_{\mathcal{H}_\mathcal{K}} \leq R \,\wedge\, \|f\|_\rho = r \right\} .$$

Let us now, similar to Ong et al. (2004), define a ball in the reproducing kernel Kreĭn space via the strong topology as

$$\mathcal{B}_\mathcal{K} = \left\{ f \in \mathcal{K} \mid \|f\|_{\mathcal{H}_\mathcal{K}} \leq R \right\} .$$

Then, we have $\mathcal{F} \subset \mathcal{B}_\mathcal{K}$ and $\mathcal{R}_n(\mathcal{F}) \leq \mathcal{R}_n(\mathcal{B}_\mathcal{K})$, where $\mathcal{R}_n(\mathcal{F})$ denotes the Rademacher complexity of $\mathcal{F}$ (defined subsequently). On the other hand, we can bound the Rademacher complexity of the hypothesis space $\mathcal{B}_\mathcal{K}$ using a result by Ong et al. (Lemma 9, 2004). In particular, we have that (Ong et al., 2004)

$$\begin{aligned}\mathcal{R}_n(\mathcal{B}_\mathcal{K}) &= \mathbb{E}_{\nu,\sigma}\left[\sup_{f\in\mathcal{B}_\mathcal{K}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sigma_i f(x_i) \right| \mid x_1, \ldots, x_n \right] \\ &\leq \frac{R}{\sqrt{n}} \sqrt{\int h_\mathcal{K}(x,x)\, \mathrm{d}\nu(x)} \,,\end{aligned}$$

where σ_i are Rademacher random variables taking values in $\{-1, 1\}$, $h_\mathcal{K}$ is the reproducing kernel corresponding to $\mathcal{H}_\mathcal{K}$, and ν is a measure on $\mathcal{X}$. Having provided a bound on the Rademacher complexity of our hypothesis space, we can now use a result by Mendelson (Corollary 2.24, 2003) to give a generalization bound for learning in a reproducing kernel Kreĭn space using the proposed variant of regularized risk minimization. The proof of the following sample complexity bound mimics that for the reproducing kernel Hilbert spaces and can be found in Ong et al. (2004).

Theorem 4. *(Mendelson, 2003; Ong et al., 2004) Let $h_\mathcal{K}$ be the reproducing kernel of a Hilbert space associated to a reproducing kernel Kreĭn space $\mathcal{K}$. For all $0 < \epsilon, \delta < 1$ there exists $N \in \Omega\left(\frac{1}{\epsilon^2} \max\left\{\mathcal{R}_n^2(J(\mathcal{B}_\mathcal{K})), \log\frac{1}{\delta}\right\}\right)$ such that for any $n \geq N$ it holds*

$$\mathbb{P}\left(\sup_{f\in\mathcal{B}_\mathcal{K}} \left| \mathcal{E}_\mathbf{z}(f) - \mathcal{E}_\rho(f) \right| \geq \epsilon \right) \leq \delta \,,$$

where J denotes the squared error loss function.

3.3. Optimization of Hyperparameters

We now show how to improve the inductive bias (Baxter, 2000) of our approach by automatically tuning the hyperparameters while performing inner cross-validation. In this process, we split the training data into training and validation folds and select a validation function that will be optimized with respect to the hyperparameter vector. The optimization can be performed with an off-the-shelf algorithm (e.g., L-BFGS-B solver) as long as we are able to compute the hyperparameter gradient of the validation function.

Denote the training and validation examples with F and $F^\perp$, respectively. Then, the validation function corresponding to the squared error loss function is given by

$$\Xi(F, f) = \frac{1}{|F^\perp|} \sum_{(x,y)\in F^\perp} \left(f(x) - y\right)^2 \,,$$

where $f = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot)$ is a hypothesis from the reproducing kernel Kreĭn space defined by training examples in F. Now, denote the hyperparameter vector with θ consisting of scalars $\lambda_\pm$ and r that control the capacity of the hypothesis and a vector η parameterizing the kernel function. Then, the gradient of this validation function is given by

$$\begin{aligned}\nabla\Xi(F, f) = \; & {}^2/_{|F^\perp|} \sum_{(x,y)\in F^\perp} \left(K_x^\top \alpha - y\right) \\ & \cdot \left(\left({}^{\partial K_x}/_{\partial\theta}\right)^\top \alpha + K_x^\top \, {}^{\partial\alpha}/_{\partial\theta}\right) .\end{aligned} \tag{8}$$

A globally optimal solution to our regularized risk minimization problem is given in a closed form in Eq. (6). From that solution, we can derive the gradient of α with respect to the hyperparameters. More specifically, we have

$$\begin{aligned}\tau^\top \frac{\partial\alpha}{\partial\theta} = \; & -t^\top P_+ \alpha \,\frac{\partial}{\partial\theta}(\lambda_+)^2 + t^\top P_- \alpha \,\frac{\partial}{\partial\theta}(\lambda_-)^2 + \\ & t^\top u \,\frac{\partial\mu^*}{\partial\theta} + \mu^* t^\top \,\frac{\partial K}{\partial\theta}\alpha \,,\end{aligned}$$

with $\tau = {}^2/_{|F^\perp|} \sum_{(x,y)\in F^\perp} \left(K_x^\top \alpha - y\right) K_x$, $St = \tau$, and $S = V\left(\lambda_+^2 \mathbb{I}_+ - \lambda_-^2 \mathbb{I}_- - \mu^* \Sigma\right) V^\top$ that can be computed from the eigendecomposition of K. Thus, t is the solution of a linear system which can be solved in time quadratic in the number of instances using the eigendecomposition of S.

Before we give the gradients of the hyperparameters, we need to find the derivative of the optimal Lagrange multiplier μ^*. In order to do this, we substitute the expression for u^* into the second stationary constraint from Eq. (5) to obtain

$$y^\top \left(\lambda_+^2 \, K_+^{-1} + \lambda_-^2 \, K_-^{-1} - \mu^* \mathbb{I}\right)^{-2} y = r^2.$$

To find the derivative of μ^* with respect to θ we need to implicitly derive the latter equation. In particular, taking the derivative of both sides with respect to θ we deduce

$$\begin{aligned}\frac{\partial}{\partial\theta}\left(\frac{r^2}{2}\right) = \; & -q^\top P_+ u \,\frac{\partial}{\partial\theta}\left(\lambda_+^2\right) + q^\top P_- u \,\frac{\partial}{\partial\theta}\left(\lambda_-^2\right) + \\ & u^\top \frac{\partial K}{\partial\theta}\alpha + \mu^* q^\top K \frac{\partial K}{\partial\theta}\alpha + q^\top K u \,\frac{\partial\mu^*}{\partial\theta} \,,\end{aligned}$$

where q is the solution of the linear system $Sq = u$ which can again be solved in quadratic time using the eigendecomposition of S. From the latter equation, we derive the gradient of the optimal Lagrange multiplier μ^* with respect to the individual hyperparameters (a detailed derivation is provided in Appendix E). If we now substitute the derived gradients of the optimal multiplier μ^* into $\frac{\partial \alpha}{\partial \theta}$, we obtain

$$\begin{aligned}
\tau^\top \frac{\partial \alpha}{\partial r} &= r \frac{t^\top u}{q^\top K u} \ , \\
\tau^\top \frac{\partial \alpha}{\partial \eta} &= \frac{t^\top u}{q^\top K u} \left(-u - \mu^* \ K q\right)^\top \frac{\partial K}{\partial \eta} \alpha + \mu^* \ t^\top \frac{\partial K}{\partial \eta} \alpha \ , \\
\tau^\top \frac{\partial \alpha}{\partial \lambda_\pm} &= 2\lambda_\pm \left(\pm \frac{t^\top u}{q^\top K u} \cdot q^\top P_\pm u \mp t^\top P_\pm \alpha \right) \ .
\end{aligned}$$

Now, the gradient of the validation function $\Xi \left(F, f\right)$ can be derived by substituting the individual gradients into Eq. (8).

4. Related Work

From the perspective of practitioners, the main advantage of the proposed regularized risk minimization problem over standard kernel methods is the fact that a kernel function does not need to be positive definite. It is often well beyond the ability of practitioners to verify this condition and many intuitive/interpretable similarity functions are not positive definite. Previous approaches for dealing with indefiniteness of kernel matrices can be divided into three classes: *i*) transformations of the kernel spectrum, *ii*) stabilization instead of minimization of a risk functional, and *iii*) learning with evaluation functionals as features.

The first class of approaches aims at converting an indefinite kernel function, which defines a reproducing kernel Kreĭn space, to a positive definite one. Perhaps the simplest such approach is to *clip* the spectrum of the kernel matrix, i.e., set the negative eigenvalues to zero (Wu et al., 2005). This corresponds to projecting an indefinite kernel matrix to the cone of positive definite matrices. The approach can be motivated by problems in which negative spectrum amounts to noise, rather than useful information. Another approach from this class, considers shifting the spectrum of the kernel matrix by adding the absolute value of the smallest eigenvalue to the diagonal of the kernel matrix (Roth et al., 2003; Zhang et al., 2006). While spectrum clip changes the kernel matrix, spectrum *shift* modifies only its diagonal entries. Some approaches consider mapping of an indefinite kernel matrix to its *square* which is positive definite (Chen et al., 2009; Graepel et al., 1998). Another popular transformation *flips* the spectrum by taking the absolute value of the eigenvalues (Graepel et al., 1998; Loosli et al., 2016). This transformation is equivalent to learning in an associated Hilbert space corresponding to a decomposition of a Kreĭn kernel. We conclude our brief review of spectral transformations with the work by Ong et al. (2004), which regularizes the risk minimization by setting to zero the eigenvalues with the absolute value below an a priori specified threshold. The hypothesis is then obtained by solving the linear system given by the minimization of the expected squared error.

In the second class of approaches, the minimization of a regularized risk functional is replaced with its stabilization. The stabilization of a risk functional, first proposed by Ong et al. (2004), can intuitively be interpreted as settling with a *good* stationary point of the regularized risk minimization. Early approaches from this class involved optimization of support vector machines while ignoring the non-convexity of the optimization problem (Lin & Lin, 2003). Recently, Loosli et al. (2016) have proposed a support vector machine for learning in Kreĭn spaces that performs stabilization by finding a hypothesis in a reproducing kernel Kreĭn space $\left(\mathcal{H}_+ \oplus \mathcal{H}_-, \langle \cdot, \cdot \rangle_{\mathcal{K}}\right)$ that solves the corresponding primal optimization problem by minimizing over $\mathcal{H}_+$ and maximizing over $\mathcal{H}_-$. As the authors of that work show, this amounts to solving the dual optimization problem over the associated reproducing kernel Hilbert space $\left(\mathcal{H}_+ \oplus \mathcal{H}_-, \langle \cdot, \cdot \rangle_{\mathcal{H}_{\mathcal{K}}}\right)$. The approach is related to considerations by Graepel et al. (1998), where the eigenvalues of an indefinite kernel matrix are replaced with their absolute values. Another support vector machine approach for learning in Kreĭn spaces was proposed by Luss & d'Aspremont (2009). A key idea in that work is to first find a positive definite matrix that approximates well the indefinite one and then learn a support vector machine predictor with that positive definite matrix as the kernel matrix. Thus, the approach can be seen as a sophisticated transformation of spectrum, where an indefinite matrix is mapped to a positive definite one using training examples. Chen & Ye (2008) have provided a fast algorithm for this variant of support vector machines in Kreĭn spaces.

The third class of approaches first embeds instances into a feature space defined by kernel values between them and a fixed number of landmarks from the instance space. Following this, a linear model is used in the constructed feature space to learn a target concept. Chen et al. (2009) have considered such an approach for learning with symmetric similarity/kernel functions, providing a detailed empirical study and a generalization bound. Recently, Alabdulmohsin et al. (2015) have reported promising empirical results using support vector machines with ℓ_1-norm regularization and indefinite kernels as features. Balcan et al. (2008) have studied generalization properties of learning with kernel/similarity functions as features. Their theoretical results demonstrate 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

Table 1. This table presents the results of synthetic experiments in which the proposed approach (i.e., the KREĬN method) is compared to frequently used transformations of the spectrum of indefinite kernel matrices on regression tasks. We measure the effectiveness of a baseline/method using the average root mean squared error, computed after performing 10 fold outer cross-validation.

$K - K^{-1}$	min	max	# pos	# neg	ι	KREĬN	FLIP	CLIP	SHIFT	SQUARE	OMCS-KREĬN
GAUSS ($n = 100$)	-272.08	91.03	1	99	0.99	8.99 (±1.96)	9.11 (±2.00)	23.14 (±9.45)	22.20 (±4.95)	9.73 (±1.43)	9.28 (±2.23)
GAUSS ($n = 500$)	-70.75	378.65	51	449	0.98	10.76 (±1.54)	10.80 (±1.38)	19.87 (±5.54)	15.84 (±1.86)	11.01 (±1.28)	29.89 (±16.51)
GAUSS ($n = 1000$)	-520.19	886.49	55	945	0.99	9.70 (±0.67)	9.77 (±0.68)	14.13 (±1.49)	14.10 (±1.71)	11.31 (±4.62)	50.24 (±18.11)
SIGMOID ($n = 100$)	-19.73	21.68	44	56	0.45	12.21 (±1.67)	12.30 (±1.93)	14.75 (±2.94)	16.16 (±3.03)	13.05 (±2.01)	27.59 (±24.29)
SIGMOID ($n = 500$)	-43.87	129.87	179	321	0.47	7.94 (±1.07)	8.06 (±0.96)	9.78 (±1.02)	10.90 (±1.27)	9.40 (±0.80)	22.51 (±13.80)
SIGMOID ($n = 1000$)	-385.50	300.26	375	625	0.48	6.63 (±0.34)	6.62 (±0.33)	13.21 (±1.13)	13.56 (±1.20)	7.09 (±0.46)	16.58 (±3.87)
*SIGMOID ($n = 100$)	-5.86×10^5	5.32×10^5	2	98	0.56	5.70 (±0.60)	5.67 (±0.51)	12.97 (±2.20)	14.00 (±2.34)	20.62 (±5.81)	18.56 (±4.29)
*SIGMOID ($n = 500$)	-22.61	1.66×10^6	400	100	0.01	9.04 (±1.32)	8.06 (±0.92)	8.11 (±0.88)	15.08 (±14.27)	38.04 (±26.12)	14.59 (±12.30)

model in that space might be higher compared to learning with a kernelized variant of regularized risk minimization.

An important aspect of learning with indefinite kernels is the consistent treatment of training and test instances known as the *out-of-sample extension*. While this problem does not occur in transductive setting, where the kernel matrix can be constructed using both training and test samples, it affects a number of approaches based on spectral transformations. In particular, Chen et al. (2009) have constructed a linear operator to deal with training and test samples consistently in the case of spectrum clip. In addition to this, the authors of that work have provided an out-of-sample extension for spectrum flip without a theoretical result guaranteeing its consistency. Contrary to some of the previous empirical studies, these out-of-sample extensions are used in our experiments to transform test samples. For other transformations, such as spectrum shift, the described regularization by Ong et al. (2004), and/or matrix inversion no linear transformation exists to consistently deal with training and test samples. In these cases, it is possible to use a heuristic proposed by Wu et al. (2005), that can also be found in Chen et al. (2009). The heuristic first applies the spectral transformation to a kernel matrix comprised of training instances and a test sample and then uses the transformed part of the kernel matrix corresponding to the test sample to define its kernel expansion. In our experiments, we use this heuristic for shift and square transformations of the kernel spectrum.

5. Experiments

The presented optimization procedure can compute a globally optimal solution to the regularized risk minimization problem defined by either a positive definite (e.g., regularization via decomposition components $\mathcal{H}_{\pm}$) or an indefinite regularization/quadratic term (e.g., regularization via $\langle \cdot, \cdot \rangle_{\mathcal{K}}$). In the first set of experiments, we exploit this to gain an insight into the effectiveness of learning in reproducing kernel Kreĭn spaces using: *i*) our approach that regularizes via decomposition components (KREĬN), *ii*) an approach that regularizes via the strong topology (FLIP), *iii*) a variant of the stabilization approach (OMCS-KREĬN) motivated by Ong et al. (2004), and *iv*) approaches relying on spectral transformations of the kernel matrix (CLIP, SHIFT, SQUARE). In the first case, we find a globally optimal solution to the problem from Eq. (4) and in others we solve the problem from Eq. (2), defined with an indefinite kernel matrix or a spectral transformation in place the matrix K. Having established that the regularization via decomposition components of a reproducing kernel Kreĭn space is effective, we perform a series of experiments on real-world datasets with different structured representations (i.e., strings, graphs, shapes). More specifically, we evaluate the effectiveness of our approach with respect to the state-of-the-art baselines for learning in reproducing kernel Kreĭn spaces: *i*) Kreĭn support vector machine (Loosli et al., 2016), and *ii*) linear ridge regression with similarities as features (Alabdulmohsin et al., 2015; Chen et al., 2009). In addition to this, we perform a series of experiments with variants of standard indefinite kernels on vectorial data (described in Appendix D) and demonstrate that on some problems indefinite kernels can be more effective than the positive definite ones. A detailed description of the experimental setup can be found in Appendix C. To quantify the indefiniteness of a kernel matrix, we use the following measure (Alabdulmohsin et al., 2015)

$$\iota = \sum_{\{i\colon \lambda_i < 0\}} |\lambda_i| / \sum_i |\lambda_i| \quad \text{with} \quad 0 \leq \iota \leq 1 .$$

Table 2. This table presents the results of experiments with indefinite kernels derived from dissimilarity matrices defined on structured data. The effectiveness of an approach is measured using the average percentage of misclassified examples, computed after performing 10 fold stratified cross-validation.

DATASET	DISSIM SOURCE	KREĬN (%)	K-SVM (%)	LRR-SF (%)
coilyork	Graph matching	**22.56** (±7.66)	32.91 (±8.06)	26.03 (±5.60)
balls 3D	Ball-to-ball distances	0.00 (±0.00)	0.00 (±0.00)	0.00 (±0.00)
prodom	Protein alignment	0.00 (±0.00)	0.00 (±0.00)	0.04 (±0.11)
chicken10	String edit distance	**5.62** (±2.55)	30.95 (±7.81)	11.91 (±3.56)
protein	Protein alignment	0.00 (±0.00)	5.17 (±3.34)	2.83 (±3.15)
zongker	Template matching	**0.95** (±1.68)	16.00 (±1.41)	5.60 (±1.20)
chicken25	String edit distance	**4.73** (±3.29)	17.72 (±6.57)	16.38 (±5.14)
pdish57	Hausdorff distance	0.35 (±0.37)	0.42 (±0.25)	0.20 (±0.19)
pdism57	Hausdorff distance	0.11 (±0.18)	0.13 (±0.23)	0.15 (±0.17)
woody50	Shape dissimilarity	**2.53** (±2.66)	37.04 (±5.07)	22.89 (±4.07)

In Table 1, we present the results of our synthetic experiments designed to evaluate the effectiveness of regularization via decomposition components $\mathcal{H}_{\pm}$ and/or the strong topology of a Kreĭn space. In these experiments, we first

Table 3. This table presents the results of experiments on real-world datasets in which the proposed risk minimization problem is used to evaluate the effectiveness of indefinite kernels on classification and regression tasks. For classification tasks, we measure the effectiveness of a kernel using the average percentage of misclassified examples, computed after performing 10 fold cross-validation. The effectiveness on regression tasks is measured using the root mean squared error, which is also computed after performing 10 fold cross-validation.

DATASET	SIGMOID	ι	RL-SIGMOID	ι	DELTA-GAUSS	ι	EPANECHNIKOV	ι	GAUSS	ι	RL-GAUSS	ι
ⓒ HABERMAN ($n = 306$)	29.99 (±5.76)	0.11	30.79 (±5.85)	0.21	30.31 (±9.63)	0.60	32.29 (±7.28)	0.02	30.61 (±8.70)	0.00	29.69 (±8.00)	0.00
ⓒ IONOSPHERE ($n = 351$)	9.35 (±4.26)	0.34	7.96 (±5.22)	0.40	6.29 (±4.92)	0.60	7.45 (±4.67)	0.02	6.29 (±4.92)	0.00	8.25 (±3.42)	0.00
ⓒ BREASTCANCER ($n = 683$)	2.63 (±1.71)	0.29	3.25 (±2.91)	0.26	2.93 (±1.86)	0.40	3.36 (±2.79)	0.03	3.21 (±2.68)	0.00	2.63 (±1.94)	0.00
ⓒ AUSTRALIAN ($n = 690$)	14.32 (±4.89)	0.14	13.89 (±4.02)	0.38	14.18 (±4.80)	0.60	13.76 (±4.80)	0.01	14.18 (±4.58)	0.00	13.74 (±4.16)	0.00
ⓒ DIABETES ($n = 768$)	27.08 (±4.61)	0.20	26.30 (±5.31)	0.32	26.30 (±3.84)	0.30	25.65 (±4.81)	0.03	26.17 (±4.71)	0.00	24.74 (±5.08)	0.00
ⓡ YACHT ($n = 308$)	2.12 (±1.73)	0.11	**1.63** (±1.24)	0.35	2.80 (±2.44)	0.70	5.75 (±1.85)	0.04	5.09 (±2.33)	0.00	3.47 (±1.93)	0.00
ⓡ PM10 ($n = 500$)	16.05 (±2.81)	0.16	16.06 (±2.38)	0.37	15.72 (±2.83)	0.50	15.63 (±2.34)	0.04	15.54 (±2.67)	0.00	15.78 (±2.15)	0.00
ⓡ WAGE ($n = 534$)	10.01 (±2.17)	0.17	9.95 (±2.13)	0.36	9.85 (±2.11)	0.20	10.02 (±1.99)	0.01	9.87 (±2.12)	0.00	9.92 (±2.13)	0.00
ⓡ AIRFOIL ($n = 1503$)	8.31 (±1.62)	0.07	7.54 (±1.03)	0.25	**6.12** (±0.53)	0.49	8.84 (±0.77)	0.04	8.54 (±1.89)	0.00	9.21 (±1.74)	0.00

sample hyperparameters of a kernel matrix and then define an indefinite matrix as the difference between the sampled kernel matrix and its inverse. Having selected the kernel matrix, we pick a Kreĭn hypothesis by sampling coefficients of the kernel expansion from the standard normal distribution and dividing them with the square root of the size of the expansion. Note that the target function is, thus, unlikely to be contained in the span of the training data only. After sampling the hypothesis, we perturb it with a noise vector sampled from the standard normal distribution with zero mean and scale that corresponds to 5% of the hypothesis range. The empirical results show that clipping and shifting of the kernel spectrum can result in a significant performance degradation compared to regularization via decomposition components and/or the strong topology of a reproducing kernel Kreĭn space (KREĬN and FLIP). While the flip spectrum transformation has been considered in previous work (Chen et al., 2009; Graepel et al., 1998; Loosli et al., 2016), the results reported here are obtained using a novel regularized risk minimization problem with different generalization properties compared to the previous approaches. Overall, our KREĬN approach is the best performing method across the considered problems characterized by different spectrum structure/decay. For kernel matrices, which in the absolute value have a large positive and a large negative eigenvalue, squaring of the spectrum can result in a performance degradation. Another important insight from the synthetic experiments is that for fixed hyperparameters the OMCS-KREĬN approach results in a hypothesis with large norm over the negative part of the spectrum. The hyperparameter optimization on a validation set penalizes over-fitting on the training data by pushing the radius r to zero or *'encourages fitting'* of the validation data without capacity control by pushing λ to zero. As a result of this, the approach fails to generalize to unseen examples.

In Table 2, we present the results of our experiments on classification tasks using a set of benchmark datasets for learning with indefinite kernels (Duin & Pekalska, 2009). The set consists of matrices with pairwise dissimilarities between instances and the corresponding labels. In our simulations, we follow the guidelines from Pekalska & Haasdonk (2009) and use the negative double-centering transformation characteristic to multidimensional scaling (Cox & Cox, 2000) to map dissimilarity matrices to kernel matrices expressing the pairwise similarities between instances. In the table header, K-SVM refers to Kreĭn support vector machine and LRR-SF to linear ridge regression with similarities as features. For each dataset, we have performed the Welch t-test (Welch, 1947) with $p = 0.05$ and marked the statistically significantly better results in bold (standard deviations are provided in the brackets). The results show that our approach which regularizes via decomposition components of a Kreĭn space performs statistically significantly better than the two competing approaches on the considered datasets.

Table 3 presents the results of our empirical evaluation on real-world vectorial datasets using the proposed approach. The goal of the experiment is to show that indefinite kernels define an important class of kernel functions. All the kernels used in this experiment are described in Appendix D, together with the corresponding hyperparameters. The results show that on the YACHT and AIRFOIL datasets, the error obtained with RL-SIGMOID and DELTA-GAUSS kernels is statistically significantly better than the one obtained with positive definite kernels (the Welch t-test with $p = 0.05$).

Conclusion

We have proposed a novel regularized risk minimization problem for learning in reproducing kernel Kreĭn spaces and showed that the strong representer theorem applies to it. The approach is consistent and guaranteed to find an optimal solution in time cubic in the number of training examples. Moreover, we have provided means for efficient hyperparameter tuning by deriving the gradient of the solution with respect to its hyperparameters. Our empirical results demonstrate the effectiveness of regularizing via decomposition components of a reproducing kernel Kreĭn space compared to learning with different spectrum transformations, as well as the state-of-the-art competing approaches. The results obtained on real-world vectorial datasets show that on some problems variants of the well-known indefinite kernels can outperform the frequently used positive definite ones.

Acknowledgments: We are grateful for access to the University of Nottingham High Performance Computing Facility. The authors would also like to thank Michael Kamp for valuable discussions and comments.

References

Alabdulmohsin, I., Gao, X., and Zhang, X. Z. Support vector machines with indefinite kernels. In Phung, D. and Li, H. (eds.), *Proceedings of the Sixth Asian Conference on Machine Learning*, volume 39 of *Proceedings of Machine Learning Research*, pp. 32–47. PMLR, 2015.

Alpay, D. Some remarks on reproducing kernel Kreĭn spaces. *Rocky Mountain Journal of Mathematics*, 21(4):1189–1205, 1991.

Azizov, T. Y. and Iokhvidov, I. S. Linear operators in spaces with an indefinite metric and their applications. *Journal of Soviet Mathematics*, 15(4):438–490, 1981.

Balcan, M.-F., Blum, A., and Srebro, N. A theory of learning with similarity functions. *Machine Learning*, 72(1):89–112, 2008.

Baxter, J. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12(1):149–198, 2000.

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

Bunch, J. R., Nielsen, C. P., and Sorensen, D. Rank-one modification of the symmetric eigenproblem. *Numerische Mathematik*, 31(1):31–48, 1978.

Chen, J. and Ye, J. Training SVM with indefinite kernels. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 136–143. ACM, 2008.

Chen, Y., Garcia, E. K., Gupta, M. R., Rahimi, A., and Cazzanti, L. Similarity-based classification: Concepts and algorithms. *Journal of Machine Learning Research*, 10:747–776, 2009.

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

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

Forsythe, G. E. and Golub, G. H. On the stationary values of a second-degree polynomial on the unit sphere. *Journal of the Society for Industrial and Applied Mathematics*, 13(4):1050–1068, 1965.

Gander, W., Golub, G. H., and von Matt, U. A constrained eigenvalue problem. *Linear Algebra and its Applications*, 114-115: 815–839, 1989.

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

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

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

Laub, J. and Müller, K.-R. Feature discovery in non-metric pairwise data. *Journal of Machine Learning*, 5:801–818, 2004.

Lin, H.-T. and Lin, C.-J. A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods. Technical report, National Taiwan University, 2003.

Loosli, G., Canu, S., and Ong, C. S. Learning SVM in Kreĭn spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(6):1204–1216, 2016.

Luss, R. and d'Aspremont, A. Support vector machine classification with indefinite kernels. *Mathematical Programming Computation*, 1(2):97–118, 2009.

Mendelson, S. *A Few Notes on Statistical Learning Theory*, pp. 1–40. Advanced Lectures on Machine Learning: Machine Learning Summer School 2002 Canberra, Australia. Springer Berlin Heidelberg, 2003.

Oglic, D., Paurat, D., and Gärtner, T. Interactive knowledge-based kernel PCA. In Calders, T., Esposito, F., Hüllermeier, E., and Meo, R. (eds.), *Machine Learning and Knowledge Discovery in Databases*, pp. 501–516. Springer Berlin Heidelberg, 2014.

Ong, C. S., Mary, X., Canu, S., and Smola, A. J. Learning with non-positive kernels. In *Proceedings of the Twenty-First International Conference on Machine Learning*, volume 69 of *ACM International Conference Proceeding Series*. ACM, 2004.

Pekalska, E. and Haasdonk, B. Kernel discriminant analysis with positive definite and indefinite kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(6):1017–1032, 2009.

Roth, V., Laub, J., Kawanabe, M., and Buhmann, J. M. Optimal cluster preserving embedding of nonmetric proximity data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25:1540–1551, 2003.

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

Schwartz, L. Sous-espaces hilbertiens d'espaces vectoriels toplogiques et noyaux associés (noyaux reproduisants). *Journal d'Analyse Mathematique*, 13:115–256, 1964.

Welch, B. L. The generalization of student's problem when several different population variances are involved. *Biometrika*, 34(1/2), 1947.

Wu, G., Chang, E. Y., and Zhang, Z. An analysis of transformation on non-positive semidefinite similarity matrix for kernel machines. Technical report, Department of Electrical and Computer Engineering, University of California, Santa Barbara, 2005.

Zhang, H., Berg, A. C., Maire, M., and Malik, J. SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2126–2136. IEEE Computer Society, 2006.

A. Proofs

Theorem 2. *Let $f^* \in \mathcal{K}$ be an optimal solution to the* KREĬN *optimization problem from Eq. (3). Then, f^* admits the expansion $f^* = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot)$ with $\alpha_i \in \mathbb{R}$.*

Proof. From Definition 1 it follows that a function $f \in \mathcal{K}$ admits a decomposition $f = f_+ \oplus f_-$ with $f_\pm \in \mathcal{H}_\pm$. Denote with $\mathcal{H}_\pm(X) = \text{span}(\{k_\pm(x, \cdot) \in \mathcal{H}_\pm \mid x \in X\})$ the two spans of the evaluation functionals centered at data instances from $X = \{x_1, \ldots, x_n\}$. Let $f_\pm = u_\pm + v_\pm$ such that $u_\pm \in \mathcal{H}_\pm(X)$ and $v_\pm \perp \mathcal{H}_\pm(X)$. Then, for all instances $x \in X$ it holds that

$$\langle v_\pm, k_\pm(x, \cdot)\rangle_{\mathcal{H}_\pm} = 0 \; .$$

Thus, for all $x \in X$ a Kreĭn hypothesis $f \in \mathcal{K}$ evaluated at x is independent of $v_\pm$. More specifically, we have that

$$\begin{aligned} & f(x) = \langle f_+ \oplus f_-, k(x, \cdot)\rangle_{\mathcal{K}} = \\ & \langle u_+ + v_+, k_+(x, \cdot)\rangle_{\mathcal{H}_+} - \langle u_- + v_-, k_-(x, \cdot)\rangle_{\mathcal{H}_-} = \\ & u_+(x) - u_-(x) = u(x) \; . \end{aligned}$$

From here it follows that we can express the optimization problem from Eq. (3) as

$$\begin{aligned} \min_{f \in \mathcal{K}} \quad & \frac{1}{n}\sum_{i=1}^{n}\Big(u(x_i) - y_i\Big)^2 + \lambda_+\left(\|u_+\|^2_{\mathcal{H}_+} + \|v_+\|^2_{\mathcal{H}_+}\right) \\ & + \lambda_-\left(\|u_-\|^2_{\mathcal{H}_-} + \|v_-\|^2_{\mathcal{H}_-}\right) \\ s.t. \quad & \frac{1}{n}\sum_{i=1}^{n}\Big(u(x_i) - \frac{1}{n}\sum_{j=1}^{n} u(x_j)\Big)^2 = r^2 \; . \end{aligned}$$

As the hard constraint is independent of $v_\pm$, this optimization problem attains the minimal value at $v_\pm = 0$. Hence, an optimal solution to the optimization problem from Eq. (3) admits the expansion $f^* = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot)$ with $\alpha_i \in \mathbb{R}$ and $1 \leq i \leq n$. □

For the sake of completeness, we provide a proof of Proposition 3 which can also be found in Oglic et al. (2014). While the result itself has not been explicitly formulated by Forsythe & Golub (1965) and Gander et al. (1989), it follows directly from the considerations in these two papers.

Proposition 3. *(Forsythe & Golub, 1965; Gander et al., 1989) The optimization objective $\Theta(\alpha)$ attains its minimal value at the tuple (α^*, μ^*) satisfying the stationary constraints (5) with the smallest value of μ. Analogously, the maximal value of $\Theta(\alpha)$ is attained at the tuple with the largest value of the Lagrange multiplier μ.*

Proof. Denote with $C = \lambda_+^2 K_+^{-1} + \lambda_-^2 K_-^{-1}$ and let $u = K\alpha$. Then, the two stationary constraints from Eq. (5) can be written as

$$\begin{aligned} & Cu = y + \mu u \\ & u^\top u = r^2 \; . \end{aligned}$$

Here, C is a symmetric matrix as the sum of two symmetric and positive definite matrices.

Let (α_1, μ_1) and (α_2, μ_2) be two tuples satisfying the stationary constraints from Eq. (5) with $\mu_1 \geq \mu_2$. Then, substituting $u_1 = K\alpha_1$ and $u_2 = K\alpha_2$ into the first stationary constraint, we have that

$$Cu_1 = \mu_1 u_1 + y \; , \tag{9}$$

$$Cu_2 = \mu_2 u_2 + y \; . \tag{10}$$

Substracting (10) from (9) we deduce

$$Cu_1 - Cu_2 = \mu_1 u_1 - \mu_2 u_2 \; . \tag{11}$$

Multiplying Eq. (11) first with $u_1^\top$ and then with $u_2^\top$ and adding the resulting two equations (having in mind that the matrix C is symmetric) we derive

$$u_1^\top C u_1 - u_2^\top C u_2 = (\mu_1 - \mu_2)(r^2 + u_1^\top u_2) \; . \tag{12}$$

On the other hand, combining the Cauchy-Schwarz inequality with the second stationary constraint we obtain that

$$u_1^\top u_2 \leq \|u_1\|\|u_2\| = r^2 \; . \tag{13}$$

Now, combining the results obtained in (12) and (13) with the initial assumption $\mu_1 \geq \mu_2$,

$$u_1^\top C u_1 - u_2^\top C u_2 \leq 2r^2(\mu_1 - \mu_2) \; . \tag{14}$$

Finally, subtracting the optimization objectives for the two tuples and using (9) and (10) multiplied by $u_1^\top$ and $u_2^\top$, respectively, we derive

$$\begin{aligned} & \Theta(\alpha_1) - \Theta(\alpha_2) = u_1 C u_1 - u_2 C u_2 - 2y^\top(u_1 - u_2) = \\ & 2r^2(\mu_1 - \mu_2) - (u_1^\top C u_1 - u_2^\top C u_2) \geq 0 \; , \end{aligned}$$

where the last inequality follows from (14). □

B. Secular Root Finder

In this appendix, we review an effective iterative method (Gander et al., 1989) for finding the smallest/largest root of the secular equation introduced in Section 3.1. An obvious choice for the root finder is the Newton method and, yet, it is not well suited for the problem. The tangent at certain points in the interval of interest crosses the x-axis outside that interval leading to incorrect solution or division by zero. An efficient root finder, then, must overcome these issues and converge very quickly. The main idea behind an efficient iterative root finder is to first approximate the secular equation with a quadratic surrogate and then update the current root estimate with the root of the surrogate function.

As the smallest root $\mu_* \in \left(-\infty, \lambda_{\pm}^{2}/\sigma_{\pm}\right)$, the secular equation has a quadratic surrogate for only one interval endpoint (Gander et al., 1989), i.e.,

$$h(\mu) = \frac{p}{(q-\mu)^2} - r^2 .$$

In order to determine the coefficients of the quadratic surrogate at step t, the secular equation and its derivative are matched to the corresponding surrogate approximations at the candidate root. In other words, the following constraints are enforced on the surrogate function

$$h(\mu_t) = g(\mu_t) \quad \wedge \quad h'(\mu_t) = g'(\mu_t) .$$

From the derivative constraint it follows that

$$g'\left(\mu_t\right) = 2\frac{g\left(\mu_t\right) + r^2}{q - \mu_t} \quad \Longrightarrow \quad q = \mu_t + 2\frac{g\left(\mu_t\right) + r^2}{g'\left(\mu_t\right)} .$$

Now, combining the computed coefficient q with the constraint on the surrogate value at μ_t we deduce

$$p = 4\frac{\left(g\left(\mu_t\right) + r^2\right)^3}{g'\left(\mu_t\right)^2} .$$

Having computed the coefficients p and q, the next secular root candidate is given by

$$\begin{aligned} &\frac{p}{\left(q - \mu_{t+1}\right)^2} - r^2 = 0 \quad \Longrightarrow \\ &\mu_{t+1} = q - \frac{\sqrt{p}}{r} \quad \Longrightarrow \\ &\mu_{t+1} = \mu_t + 2\frac{g\left(\mu_t\right) + r^2}{g'\left(\mu_t\right)} \left(1 - \frac{\sqrt{g\left(\mu_t\right) + r^2}}{r}\right) . \end{aligned}$$

For an initial solution $\mu_* < \mu_0 < \lambda_{\pm}^{2}/\sigma_{\pm}$ the convergence is monotonic (Bunch et al., 1978), i.e., for all $t > 0$ we have that $\mu_* < \mu_{t+1} < \mu_t$.

C. Experimental Setup

In all the experiments, we have performed 10 fold outer cross-validation to evaluate the effectiveness of the considered baselines (with stratified fold splitting on classification tasks). To tune the hyperparameters of the approach ($\lambda_{\pm}$, r, and η where applicable) we have performed 5 fold (stratified) inner cross-validation. For one such split, the training is performed on the batch of $(k-1)$ training folds and the hyperparameters are optimized on the remaining validation fold. Each inner cross-validation fold is used exactly once as a validation fold and we refer to the hyperparameter gradients computed on these folds as *fold gradients*. Having computed the fold gradients for all inner cross-validation splits, the ultimate hyperparameter gradient is their average. In the inner cross-validation, we have used the derived hyperparameter gradients (Section 3.3) with the L-BFGS-B implementation from the *scipy* package. The hyperparameter optimization is performed with 10 random restarts such that for each initial solution the minimization procedure makes at most 20 iterations of L-BFGS-B minimization and then continues with the best hyperparameter vector for at most 200 iterations. In all our simulations, we have used identical initialization procedures for hyperparameter optimization (described in Appendix E). For real-world vectorial datasets, we have normalized the instances so that the data matrix has zero mean and unit variance. To facilitate the comparison between different problems, the labels in regression tasks were normalized so that their range is equal to one. In classification tasks, the input labels $\{-1, 1\}$ were set to

$$y_+ = \sqrt{\frac{n_-}{n_+}} \qquad \text{and} \qquad y_- = -\sqrt{\frac{n_+}{n_-}} ,$$

where $n_{\pm}$ denote the number of positive/negative class labels present in a given sample of labeled examples.

In the experiments with structured data[1], we have used the negative double-centering transformation (Cox & Cox, 2000; Pekalska & Haasdonk, 2009) to convert dissimilarity matrices to indefinite kernel/similarity matrices. More formally, for a symmetric dissimilarity matrix D the indefinite kernel/similarity matrix is given by

$$K = -\frac{1}{2} HD \odot DH ,$$

where $H = \mathbb{I}_n - \frac{1}{n}\mathbf{e}\mathbf{e}^{\top}$, $\mathbf{e}$ is the vector of all ones, and $\odot$ denotes the elementwise multiplication of two matrices. Some of the classification tasks on structured data are multi-class problems and for them we only evaluate the effectiveness of one-vs-all classifier for the class with the label one.

D. Kernels

GAUSS

$$k(x, x') = \exp\left(-\frac{\left\|x - x'\right\|^2}{2\eta^2}\right) ,$$

where $\eta \in \mathbb{R}$

RL-GAUSS

$$k(x, x') = \exp\left(-\left(x - x'\right)^{\top} D \left(x - x'\right)\right) ,$$

where $x, x', \eta \in \mathbb{R}^d$ and $D = \operatorname{diag}\left(\eta^{-2}\right)$

[1] The datasets are available at http://prtools.org/disdatasets/index.html.

SIGMOID

$$k(x, x') = \tanh\left(\frac{-0.5 + x^\top x'}{\eta^2}\right) \ ,$$

where $\eta \in \mathbb{R}$

RL-SIGMOID

$$k(x, x') = \tanh\left(x^\top D x'\right) \ ,$$

where $x, x', \eta \in \mathbb{R}^d$ and $D = \operatorname{diag}\left(\eta^{-2}\right)$

DELTA-GAUSS

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\eta_1^2}\right) - \exp\left(-\frac{\|x - x'\|^2}{2\eta_2^2}\right) \ ,$$

where $\eta_1, \eta_2 \in \mathbb{R}$

EPANECHNIKOV

$$k(x, x') = \max\left(0, 1 - (x - x')^\top D \left(x - x'\right)\right)^2 \ ,$$

where $x, x', \eta \in \mathbb{R}^d$ and $D = \operatorname{diag}\left(\eta^{-2}\right)$

E. Hyperparameter Optimization

E.1. Initialization Schemes

The outputs are normalized so that their mean is equal to zero and their range is equal to one.

RADIUS

- $\text{var} \leftarrow \frac{1}{n}\sum_{i=1}^{n} y_i^2$
- $\delta \leftarrow \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i^2 - \text{var}\right)^2}$
- $r_{\min} \leftarrow \max\left\{10^{-4}, \text{var} - 2.5\delta\right\}$
- $r_{\max} \leftarrow \text{var} + 2.5\delta$
- $u \sim \mathcal{U}\left[0, 1\right]$ and $r \leftarrow \sqrt{\left(r_{\max} - r_{\min}\right) u + r_{\min}}$

REGULARIZATION PARAMETERS $\lambda_\pm$

- $\mathcal{S} \leftarrow \text{logspace}\left(-1, 2, 5\right)$
- $u \sim \mathcal{U}\left\{1, 2, 3, 4, 5\right\}$ and $\lambda_\pm \leftarrow \sqrt{\mathcal{S}[u]/n}$

BANDWIDTH

i) GAUSS

- $p \leftarrow \text{sq\_pairwise\_distances}\left(X\right)$
- $u \sim \mathcal{U}\left[0, 1\right]$ and $c \leftarrow 0.4(u + 1)$
- $\eta \leftarrow {}^{\text{median}(p)}/_{c}$

ii) RL-GAUSS

- $M \leftarrow \text{col\_max}\left(X\right)$ and $m \leftarrow \text{col\_min}\left(X\right)$
- $u \sim \mathcal{U}\left[0, 1\right]^d$ and $c \leftarrow 0.4(u + 1)$
- $\eta \leftarrow \sqrt{d} \cdot c \otimes \left(M - m\right)$

iii) SIGMOID

- $u \sim \mathcal{U}\left[0, 1\right]$ and $c \leftarrow 0.4(u + 1)$
- $\eta \leftarrow \sqrt{{}^{\max\{\text{row\_norm}(X)\}}/_{c}}$

iv) RL-SIGMOID

- $u \sim \mathcal{U}\left[0, 1\right]^d$ and $c \leftarrow 0.4(u + 1)$
- $\eta \leftarrow \sqrt{\text{col\_max}\left(\text{abs}\left(X\right)\right) \otimes c^{-1}}$

v) DELTA-GAUSS

- $u \sim \mathcal{U}\left[0, 1\right]$ and $c \leftarrow 2u - 1$
- $\eta \sim$ GAUSS
- $\eta_1 \leftarrow (1 - c)\eta$ and $\eta_2 \leftarrow (1 + c)\eta$

vi) EPANECHNIKOV

- $\hat{\eta} \sim$ RL-SIGMOID
- $\eta \leftarrow \hat{\eta}^2$

E.2. Derivation of Hyperparameter Gradients

$$K = V \Sigma V^\top \quad \wedge \quad S = \lambda_+^2 \ P_+ - \lambda_-^2 \ P_- - \mu^* K$$
$$S = V \left(\lambda_+^2 \mathbb{I}_+ - \lambda_-^2 \mathbb{I}_- - \mu^* \Sigma\right) V^\top$$

$$\alpha = \left(\lambda_+^2 \ P_+ - \lambda_-^2 \ P_- - \mu^* K\right)^{-1} y = S^{-1} y$$
$$u = K\alpha = K S^{-1} y$$

$$\Xi\left(F, f\right) = \quad \frac{1}{|F^\perp|} \sum_{(x,y) \in F^\perp} \left(f\left(x\right) - y\right)^2 =$$
$$\frac{1}{|F^\perp|} \sum_{(x,y) \in F^\perp} \left(K_x^\top \alpha - y\right)^2$$

$$\nabla \Xi\left(F, f\right) = {}^{2}/_{|F^\perp|} \sum_{(x,y) \in F^\perp} \left(K_x^\top \alpha - y\right) \cdot$$
$$\left(\left({}^{\partial K_x}/_{\partial \theta}\right)^\top \alpha + K_x^\top \, {}^{\partial \alpha}/_{\partial \theta}\right)$$

$$\tau = {}^{2}/_{|F^\perp|} \sum_{(x,y) \in F^\perp} \left(K_x^\top \alpha - y\right) K_x \quad \wedge \quad St = \tau$$

$$\begin{aligned}
\tau^\top \frac{\partial \alpha}{\partial \theta} = \quad & \tau^\top S^{-1} \left(-\frac{\partial}{\partial \theta} \left(\lambda_+\right)^2 P_+ + \frac{\partial}{\partial \theta} \left(\lambda_-\right)^2 P_- \right) \alpha + \\
& \tau^\top S^{-1} \left(\frac{\partial \mu^*}{\partial \theta} K + \mu^* \frac{\partial K}{\partial \theta} \right) \alpha = \\
& - t^\top P_+ \alpha \, \frac{\partial}{\partial \theta} \left(\lambda_+\right)^2 + t^\top P_- \alpha \, \frac{\partial}{\partial \theta} \left(\lambda_-\right)^2 + \\
& t^\top u \, \frac{\partial \mu^*}{\partial \theta} + \mu^* t^\top \, \frac{\partial K}{\partial \theta} \alpha
\end{aligned}$$

$$y^\top \left(\lambda_+^2 \, K_+^{-1} + \lambda_-^2 \, K_-^{-1} - \mu^* \mathbb{I} \right)^{-2} y = r^2$$

$$Sq = u$$

$$\begin{aligned}
\frac{\partial}{\partial \theta} \left(\frac{r^2}{2} \right) = \quad & - q^\top P_+ u \, \frac{\partial}{\partial \theta} \left(\lambda_+^2 \right) + q^\top P_- u \, \frac{\partial}{\partial \theta} \left(\lambda_-^2 \right) + \\
& u^\top \frac{\partial K}{\partial \theta} \alpha + \mu^* q^\top K \frac{\partial K}{\partial \theta} \alpha + q^\top K u \, \frac{\partial \mu^*}{\partial \theta}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \mu^*}{\partial r} = \quad & \frac{r}{q^\top K u} \\
\frac{\partial \mu^*}{\partial \eta} = \quad & - \frac{1}{q^\top K u} \left(u + \mu^* K q \right)^\top \frac{\partial K}{\partial \eta} \alpha \\
\frac{\partial \mu^*}{\partial \lambda_\pm} = \quad & \pm \, 2 \lambda_\pm \frac{q^\top P_\pm u}{q^\top K u}
\end{aligned}$$

$$\begin{aligned}
\tau^\top \frac{\partial \alpha}{\partial r} = \quad & r \frac{t^\top u}{q^\top K u} \\
\tau^\top \frac{\partial \alpha}{\partial \eta} = \quad & - \frac{t^\top u}{q^\top K u} \left(u + \mu^* \, K q \right)^\top \frac{\partial K}{\partial \eta} \alpha + \mu^* \, t^\top \frac{\partial K}{\partial \eta} \alpha \\
\tau^\top \frac{\partial \alpha}{\partial \lambda_\pm} = \quad & 2 \lambda_\pm \left(\pm \frac{t^\top u}{q^\top K u} \cdot q^\top P_\pm u \mp t^\top P_\pm \alpha \right)
\end{aligned}$$