
Learning with Non-Positive Kernels

Cheng Soon Ong

CHENG.ONG@ANU.EDU.AU

Computer Sciences Laboratory, RSISE, Australian National University, 0200 ACT, Australia

Xavier Mary

XAVIER.MARY@ENSAE.FR

ENSAE-CREST-LS, 3 avenue Pierre Larousse, 92240 Malakoff, France

Stéphane Canu

SCANU@INSA-ROUEN.FR

Laboratoire PSI FRE CNRS 2645 - INSA de Rouen, B.P. 08, 76131 Mont-Saint-Aignan Cedex, France

Alexander J. Smola

ALEX.SMOLA@ANU.EDU.AU

RSISE and NICTA Australia, Australian National University, 0200 ACT, Australia

Indefinite Kernels, Reproducing Kernel Kreĭn Space, Representer Theorem, Rademacher Average, Non-convex Optimization, Ill-posed Problems

Abstract

In this paper we show that many kernel methods can be adapted to deal with indefinite kernels, that is, kernels which are not positive semidefinite. They do not satisfy Mercer's condition and they induce associated functional spaces called Reproducing Kernel Kreĭn Spaces (RKKS), a generalization of Reproducing Kernel Hilbert Spaces (RKHS).

Machine learning in RKKS shares many "nice" properties of learning in RKHS, such as orthogonality and projection. However, since the kernels are indefinite, we can no longer minimize the loss, instead we stabilize it. We show a general representer theorem for constrained stabilization and prove generalization bounds by computing the Rademacher averages of the kernel class. We list several examples of indefinite kernels and investigate regularization methods to solve spline interpolation. Some preliminary experiments with indefinite kernels for spline smoothing are reported for truncated spectral factorization, Landweber-Fridman iterations, and MR-II.

Appearing in *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada, 2004. Copyright 2004 by the authors.

1. Why Non-Positive Kernels?

Almost all current research on kernel methods in machine learning focuses on functions $k(x, x')$ which are positive semidefinite. That is, it focuses on kernels which satisfy Mercer's condition and which consequently can be seen as scalar products in some Hilbert space. See (Vapnik, 1998; Schölkopf & Smola, 2002; Wahba, 1990) for details.

The purpose of this article is to point out that there is a much larger class of kernel functions available, which do not necessarily correspond to a RKHS but which nonetheless can be used for machine learning. Such kernels are known as *indefinite kernels*, as the scalar product matrix may contain a mix of positive and negative eigenvalues. There are several motivations for studying indefinite kernels:

- Testing Mercer's condition for a given kernel can be a challenging task which may well lie beyond the abilities of a practitioner.
- Sometimes functions which can be proven *not* to satisfy Mercer's condition may be of other interest. One such instance is the hyperbolic tangent kernel $k(x, x') = \tanh(\langle x, x' \rangle - 1)$ of Neural Networks, which is indefinite for any range of parameters or dimensions (Smola et al., 2000).
- There have been promising empirical reports on the use of indefinite kernels (Lin & Lin, 2003).
- In H^∞ control applications and discrimination the cost function can be formulated as the difference between two quadratic norms (Haasdonk,

2003; Hassibi et al., 1999), corresponding to an indefinite inner product.

- RKKS theory (concerning function spaces arising from indefinite kernels) has become a rather active area in interpolation and approximation theory.
- In recent work on learning the kernel, such as (Ong & Smola, 2003), the solution is a linear combination of positive semidefinite kernels. However, an arbitrary linear combination of positive kernels is not necessarily positive semidefinite (Mary, 2003). While the elements of the associated vector space of kernels can always be defined as the difference between two positive kernels, what is the functional space associated with such a kernel?

We will discuss the above issues using topological spaces similar to Hilbert spaces except for the fact that the inner product is no longer necessarily positive. Section 2 defines RKKS and some properties required in the subsequent derivations. We also give some examples of indefinite kernels and describe their spectrum. Section 3 extends Rademacher type generalization error bounds for learning using indefinite kernels. Section 4 shows that we can obtain a theorem similar to the representer theorem in RKHS. However, we note that there may be practical problems. Section 5 describes how we can perform approximation of the interpolation problem using the spectrum of the kernel also using iterative methods. It also shows preliminary results on spline regularization.

2. Reproducing Kernel Krein Spaces

Krein spaces are indefinite inner product spaces endowed with a Hilbertian topology, yet their inner product is no longer positive. Before we delve into definitions and state basic properties of Krein spaces, we give an example:

Example 1 (4 dimensional space-time)

Indefinite spaces were first introduced by Minkowski for the solution of problems in special relativity. There the inner product in space-time (x, y, z, t) is given by

$$\langle (x, y, z, t), (x', y', z', t') \rangle = xx' + yy' + zz' - tt'.$$

Clearly it is not positive. The vector $v = (1, 1, 1, \sqrt{3})$ belongs to the cone of so-called neutral vectors which satisfy $\langle v, v \rangle = 0$ (in coordinates $x^2 + y^2 + z^2 - t^2 = 0$). In special relativity this cone is also called the “light cone,” as it corresponds to the propagation of light from a point event.

2.1. Krein spaces

The above example shows that there are several differences between Krein spaces and Hilbert spaces. We now define Krein spaces formally. More detailed expositions can be found in (Bognár, 1974; Azizov & Iokhvidov, 1989). The key difference is the fact that the inner products are indefinite.

Definition 1 (Inner product) *Let \mathcal{K} be a vector space on the scalar field.¹ An inner product $\langle \cdot, \cdot \rangle_{\mathcal{K}}$ on \mathcal{K} is a bilinear form where for all $f, g, h \in \mathcal{K}$, $\alpha \in \mathbb{R}$:*

- $\langle f, g \rangle_{\mathcal{K}} = \langle g, f \rangle_{\mathcal{K}}$
- $\langle \alpha f + g, h \rangle_{\mathcal{K}} = \alpha \langle f, h \rangle_{\mathcal{K}} + \langle g, h \rangle_{\mathcal{K}}$
- $\langle f, g \rangle_{\mathcal{K}} = 0$ for all $g \in \mathcal{K}$ implies $\Rightarrow f = 0$

An inner product is said to be *positive* if for all $f \in \mathcal{K}$ we have $\langle f, f \rangle_{\mathcal{K}} \geq 0$. It is *negative* if for all $f \in \mathcal{K}$ $\langle f, f \rangle_{\mathcal{K}} \leq 0$. Otherwise it is called *indefinite*.

A vector space \mathcal{K} embedded with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{K}}$ is called an *inner product space*. Two vectors f, g of an inner product space are said to be *orthogonal* if $\langle f, g \rangle_{\mathcal{K}} = 0$. Given an inner product, we can define the associated space.

Definition 2 (Krein space) *An inner product space $(\mathcal{K}, \langle \cdot, \cdot \rangle_{\mathcal{K}})$ is a Krein space if there exist two Hilbert spaces $\mathcal{H}_+, \mathcal{H}_-$ spanning \mathcal{K} such that*

- All $f \in \mathcal{K}$ can be decomposed into $f = f_+ + f_-$, where $f_+ \in \mathcal{H}_+$ and $f_- \in \mathcal{H}_-$.
- $\forall f, g \in \mathcal{K}, \langle f, g \rangle_{\mathcal{K}} = \langle f_+, g_+ \rangle_{\mathcal{H}_+} - \langle f_-, g_- \rangle_{\mathcal{H}_-}$

This suggests that there is an “associated” Hilbert space, where the difference in scalar products is replaced by a sum:

Definition 3 (Associated Hilbert Space) *Let \mathcal{K} be a Krein space with decomposition into Hilbert spaces \mathcal{H}_+ and \mathcal{H}_- . Then we denote by $\overline{\mathcal{K}}$ the associated Hilbert space defined by*

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

Likewise we can introduce the symbol \ominus to indicate that

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

Note that $\overline{\mathcal{K}}$ is the smallest Hilbert space majorizing the Krein space \mathcal{K} and one defines the strong topology on \mathcal{K} as the Hilbertian topology of $\overline{\mathcal{K}}$. The topology does not depend on the decomposition chosen. Clearly $|\langle f, f \rangle_{\mathcal{K}}| \leq \|f\|_{\overline{\mathcal{K}}}^2$ for all $f \in \mathcal{K}$.

¹Like Hilbert spaces, Krein spaces can be defined on \mathbb{R} or \mathbb{C} . We use \mathbb{R} in this paper.

\mathcal{K} is said to be *Pontryagin* if it admits a decomposition with finite dimensional \mathcal{H}_- , and *Minkowski* if \mathcal{K} itself is finite dimensional. We will see how Pontryagin spaces arise naturally when dealing with conditionally positive definite kernels (see Section 2.4).

For estimation we need to introduce Krein spaces on functions. Let \mathcal{X} be the learning domain, and $\mathbb{R}^{\mathcal{X}}$ the set of functions from \mathcal{X} to \mathbb{R} . The evaluation functional tells us the value of a function at a certain point, and we shall see that the RKKS is a subset of $\mathbb{R}^{\mathcal{X}}$ where this functional is continuous.

Definition 4 (Evaluation functional)

$$T_x : \mathcal{K} \rightarrow \mathbb{R} \text{ where } f \mapsto T_x f = f(x).$$

Definition 5 (RKKS) A Krein space $(\mathcal{K}, \langle \cdot, \cdot \rangle_{\mathcal{K}})$ is a Reproducing Kernel Krein Space (Alpay, 2001, Chapter 7) if $\mathcal{K} \subset \mathbb{R}^{\mathcal{X}}$ and the evaluation functional is continuous on \mathcal{K} endowed with its strong topology (that is, via $\overline{\mathcal{K}}$).

2.2. From Krein spaces to Kernels

We prove an analog to the Moore-Aronszajn theorem (Wahba, 1990), which tells us that for every kernel there is an associated Krein space, and for every RKKS, there is a unique kernel.

Proposition 6 (Reproducing Kernel) Let \mathcal{K} be an RKKS with $\mathcal{K} = \mathcal{H}_+ \ominus \mathcal{H}_-$. Then

1. \mathcal{H}_+ and \mathcal{H}_- are RKHS (with kernels k_+ and k_-),
2. There is a unique symmetric $k(x, x')$ with $k(x, \cdot) \in \mathcal{K}$ such that for all $f \in \mathcal{K}$, $\langle f, k(x, \cdot) \rangle_{\mathcal{K}} = f(x)$,
3. $k = k_+ - k_-$.

Proof Since \mathcal{K} is a RKKS, the evaluation functional is continuous with respect to the strong topology. Hence the associated Hilbert Space $\overline{\mathcal{K}}$ is an RKHS. It follows that \mathcal{H}_+ and \mathcal{H}_- , as Hilbertian subspaces of an RKHS, are RKHS themselves with kernels k_+ and k_- respectively. Let $f = f_+ + f_-$. Then $T_x(f)$ is given by

$$\begin{aligned} T_x(f) &= T_x(f_+) + T_x(f_-) \\ &= \langle f_+, k_+(x, \cdot) \rangle_{\mathcal{H}_+} - \langle f_-, k_-(x, \cdot) \rangle_{\mathcal{H}_-} \\ &= \langle f, k_+(x, \cdot) - k_-(x, \cdot) \rangle_{\mathcal{K}}. \end{aligned}$$

In both lines we exploited the orthogonality of \mathcal{H}_+ with \mathcal{H}_- . Clearly $k := k_+ - k_-$ is symmetric. Moreover it is unique since the inner product is non-degenerate. ■

2.3. From Kernels to Krein spaces

Let k be a symmetric real valued function on \mathcal{X}^2 .

Proposition 7 The following are equivalent (Mary, 2003, Theorem 2.28):

- There exists (at least) one RKKS with kernel k .
- k admits a positive decomposition, that is there exists two positive kernels k_+ and k_- such that $k = k_+ - k_-$.
- k is dominated by some positive kernel p (that is, $p - k$ is a positive kernel).

There is no bijection but a surjection between the set of RKKS and the set of generalized kernels defined in the vector space generated out of the cone of positive kernels.

2.4. Examples and Spectral Properties

We collect several examples of indefinite kernels in Table 1 and plot a 2 dimensional example as well as 20 of the eigenvalues with the largest absolute value. We investigate the spectrum of radial kernels using the Hankel transform.

The Fourier transform allows one to find the eigenvalue decomposition of kernels of the form $k(x, x') = \kappa(x - x')$ by computing the Fourier transform of κ . For $x \in \mathbb{R}^n$ we have

$$F[f](\|\omega\|) = \|\omega\|^{-\nu} H_{\nu}[r^{\nu} \kappa(r)](\|\omega\|)$$

where $\nu = \frac{1}{2}n - 1$ and H_{ν} is the Hankel transform of order ν . Table 1 depicts the spectra of these kernels. Negative values in the Hankel transform correspond to \mathcal{H}_- , positive ones to \mathcal{H}_+ . Likewise the decomposition of $k(x, x') = \kappa(\langle x, x' \rangle)$ in terms of associated Legendre polynomials allows one to identify the positive and negative parts of the Krein space, as the Legendre polynomials commute with the rotation group.

One common class of translation invariant kernels which are not positive definite are so-called conditionally positive definite (cpd) kernels. A cpd kernel of order p leads to a positive semidefinite matrix in a subspace of coefficients orthogonal to polynomials of order up to $p - 1$. Moreover, in the subspace of $(p - 1)$ degree polynomials, the inner product is typically negative definite. This means that there is a space of polynomials of degree up to order $p - 1$ (which constitutes an up to $\binom{n+p-2}{p-1}$ -dimensional subspace) with negative inner product. In other words, we are dealing with a Pontryagin space.

The standard procedure to use such kernels is to project out the negative component, replace the lat-

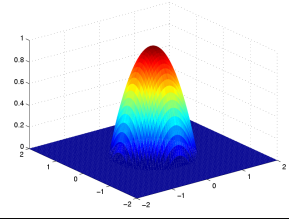
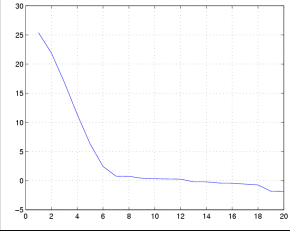
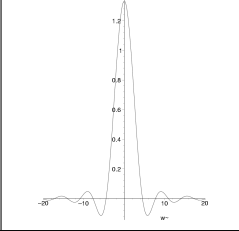
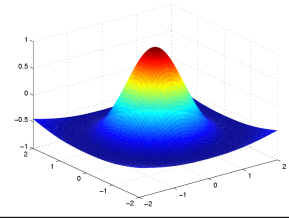
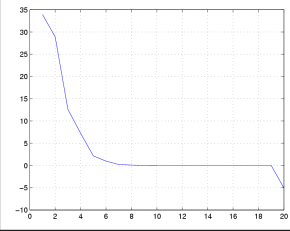
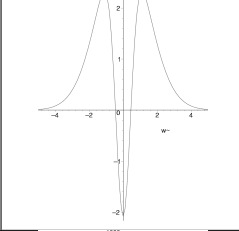
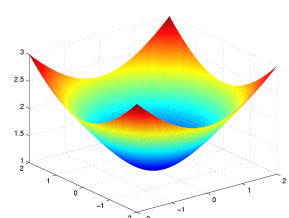
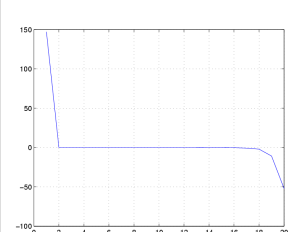
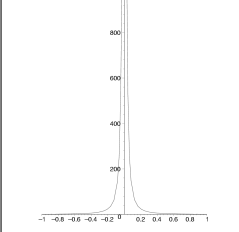
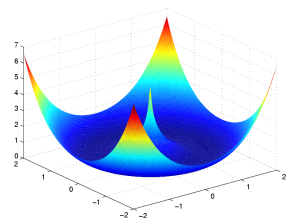
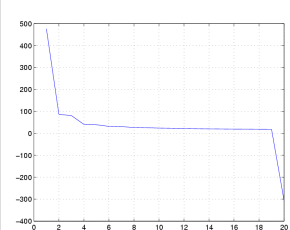
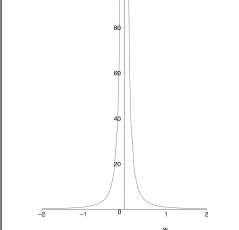
Kernel	2D kernel	20 main Eigenvalues	Fourier Transform
Epanechnikov kernel $\left(1 - \frac{\ s-t\ ^2}{\sigma}\right)^p, \text{ for } \frac{\ s-t\ ^2}{\sigma} \leq 1$			
Gaussian Combination $\exp\left(\frac{-\ s-t\ ^2}{\sigma_1}\right) + \exp\left(\frac{-\ s-t\ ^2}{\sigma_2}\right) - \exp\left(\frac{-\ s-t\ ^2}{\sigma_3}\right)$			
Multiquadric kernel $\sqrt{\frac{\ s-t\ ^2}{\sigma} + c^2}$			
Thin plate spline $\frac{\ s-t\ ^{2p}}{\sigma} \ln\left(\frac{\ s-t\ ^2}{\sigma}\right)$			

Table 1. Examples of indefinite kernels. Column 2 shows the 2D surface of the kernel with respect to the origin, column 3 shows plots of the 20 eigenvalues with largest magnitude of uniformly spaced data from the interval $[-2, 2]$, column 4 shows plots of the Fourier spectra.

ter by a suitably smoothed estimate in the polynomial subspace and treat the remaining subspace as any RKHS (Wahba, 1990). Using Krein spaces we can use these kernels directly, without the need to deal with the polynomial parts separately.

3. Generalization Bounds via Rademacher Average

An important issue regarding learning algorithms are their ability to generalize (to give relevant predictions). This property is obtained when the learning process considered shows a uniform convergence behavior. In (Mendelson, 2003) such a result is demonstrated in the case of RKHS through the control of the Rademacher average of the class of function considered. Here we

present an adaptation of this proof in the case of Krein spaces. We begin with setting the functional framework for the result.

Let k be a kernel defined on a set \mathcal{X} and choose a decomposition $k = k_+ - k_-$ where k_+ and k_- are both positive kernels. This given decomposition of the kernel can be associated with the RKHS $\bar{\mathcal{K}}$ defined by its positive kernel $\bar{k} = k_+ + k_-$ whose Hilbertian topology defines the strong topology of \mathcal{K} . We will then consider the set $\mathcal{B}_{\mathcal{K}}$ defined as follows:

$$\mathcal{B}_{\mathcal{K}} = \{f \in \mathcal{K} \mid \|f_+\|^2 + \|f_-\|^2 = \|f\|^2 \leq 1\}$$

Note that in a Krein space the norm of a function is the associated Hilbertian norm and usually $\|f\|^2 \neq \langle f, f \rangle_{\mathcal{K}}$ but always $\langle f, f \rangle_{\mathcal{K}} \leq \|f\|^2$.

The Rademacher average of a class of functions \mathcal{F}

with respect to a measure μ is defined as follows. Let $x_1, \dots, x_m \in \mathcal{X}$ be i.i.d random variables sampled according to μ . Let ε_i for $i = 1, \dots, m$ be Rademacher random variables, that is variables taking values $\{-1, +1\}$ with equal probability.

Definition 8 (Rademacher Average) *The Rademacher average, $R_m(\mathcal{F})$ of a set of functions \mathcal{F} (w.r.t. μ) is defined as*

$$R_m(\mathcal{F}) = \mathbb{E}_\mu \mathbb{E}_\varepsilon \frac{1}{\sqrt{m}} \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^m \varepsilon_i f(x_i) \right|$$

Using the Rademacher average as an estimate of the “size” of a function class, we can obtain generalization error bounds which are also called uniform convergence or sample complexity bounds (Mendelson, 2003, Corollary 3), that is for any $\varepsilon > 0$ and $\delta > 0$, there is an absolute constant C such that if $m \geq \frac{C}{\varepsilon^2} \max\{R_m^2(J(\mathcal{B}_\mathcal{K})), \log \frac{1}{\delta}\}$, then,

$$\Pr\left(\sup_{f \in \mathcal{B}_\mathcal{K}} \left| \frac{1}{m} \sum_{i=1}^m J(f(X_i)) - \mathbb{E}J(f) \right| \geq \varepsilon\right) \leq \delta,$$

where $J(f(x))$ denotes the quadratic loss defined as in (Mendelson, 2003). To get the expected result we have to show that the Rademacher average is bounded by a constant independent of the sample size m . To control the Rademacher average, we first give a lemma regarding the topology of Kreĭn spaces putting emphasis on both difference and close relationship with the Hilbertian case.

Lemma 9 *For all $g \in \mathcal{K}$:*

$$\sup_{f \in \mathcal{B}_\mathcal{K}} \langle f(\cdot), g(\cdot) \rangle_{\mathcal{K}} = \|g\|$$

Proof It is trivial if $g = 0$. $\forall g \in \mathcal{K}$, $g \neq 0$, let $h = g/\|g\|$. By construction $\|h\| = 1$.

$$\begin{aligned} \sup_{f \in \mathcal{B}_\mathcal{K}} \langle f(\cdot), g(\cdot) \rangle_{\mathcal{K}} &= \|g\| \sup_{f \in \mathcal{B}_\mathcal{K}} \langle f(\cdot), h(\cdot) \rangle_{\mathcal{K}} \\ &= \|g\| \sup_{f \in \mathcal{B}_\mathcal{K}} (\langle f_+, h_+ \rangle_{\mathcal{K}_+} - \langle f_-, h_- \rangle_{\mathcal{K}_-}) \\ &= \|g\| (\langle h_+, h_+ \rangle_{\mathcal{K}_+} + \langle h_-, h_- \rangle_{\mathcal{K}_-}) \\ &= \|g\| \end{aligned}$$

■

In the unit ball of a RKKS, the Rademacher average with respect to the probability measure μ behave the same way as the one of its associated RKHS.

Proposition 10 (Rademacher Average) *Let \overline{K} be the Gram matrix of kernel \overline{k} at points*

x_1, \dots, x_m , *If according to the measure μ on \mathcal{X} $x \mapsto \overline{k}(x, x) \in L^1(\mathcal{X}, \mu)$, then*

$$R_m(\mathcal{B}_\mathcal{K}) \leq M^{\frac{1}{2}}$$

with

$$M = \frac{1}{m} \mathbb{E}_\mu(\text{tr}(\overline{K})) = \int_{\mathcal{X}} \overline{k}(x, x) d\mu(x)$$

The proof works just as in the Hilbertian case (Mendelson, 2003, Theorem 16) with the application of lemma 9. As a second slight difference we choose to express the bound as a function of the $L^1(\mathcal{X}, \mu)$ norm of the kernel instead of going through its spectral representation. It is simpler since for instance, for the unnormalized gaussian kernel $k(x, y) = \exp(-(x - y)^2)$ on $\mathcal{X} = \mathbb{R}$ we have $M = 1$ regardless the measure μ considered. Since we are back to the Hilbertian context (Mendelson, 2003, Corollary 4) applies replacing Hilbert by Kreĭn, providing an uniform convergence result as expected.

4. Machine Learning in RKKS

In order to perform machine learning, we need to be able to optimize over a class of functions, and also to be able to prove that the solution exists and is unique. Instead of minimizing over a class of functionals as in a RKHS, we look for the stationary point. This is motivated by the fact that in a RKHS, minimization of the cost functional can be seen as a projection problem. The equivalent projection problem in RKKS gives us the stationary point of the cost functional.

4.1. Representer Theorem

The analysis of the learning problem in a RKKS gives similar representer theorems to the Hilbertian case (Schölkopf et al., 2001). The key difference is that the problem of minimizing a regularized risk functional becomes one of finding the stationary point of a similar functional. Moreover, the solution need not be unique any more. The proof technique, however, is rather similar. The main difference is that a) we deal with a constrained optimization problem directly and b) the Gateaux derivative has to vanish due to the nondegeneracy of the inner product. In the following, we define the training data $X := (x_1, \dots, x_m)$ drawn from the learning domain \mathcal{X} .

Theorem 11 *Let \mathcal{K} be an RKKS with kernel k . Denote by $L\{f, X\}$ a continuous convex loss functional depending on $f \in \mathcal{K}$ only via its evaluations $f(x_i)$ with $x_i \in X$, let $\Omega(\langle f, f \rangle)$ be a continuous stabilizer with*

strictly monotonic $\Omega : \mathbb{R} \rightarrow \mathbb{R}$ and let $C\{f, X\}$ be a continuous functional imposing a set of constraints on f , that is $C : \mathcal{K} \times \mathcal{X}^m \rightarrow \mathbb{R}^n$. Then if the optimization problem

$$\underset{f \in \mathcal{K}}{\text{stabilize}} L\{f, X\} + \Omega(\langle f, f \rangle_{\mathcal{K}}) \quad (1)$$

subject to $C\{f, X\} \leq d$

has a saddle point f^* , it admits the expansion

$$f^* = \sum_i \alpha_i k(x_i, \cdot) \text{ where } x_i \in X \text{ and } \alpha_i \in \mathbb{R}. \quad (2)$$

Proof The first order conditions for a solution of (1) imply that the Gateaux derivative of the Lagrangian

$$\mathcal{L}\{f, \lambda\} = L\{f, X\} + \Omega(\langle f, f \rangle_{\mathcal{K}}) + \lambda^\top (C\{f, X\} - d)$$

needs to vanish. By the nondegeneracy of the inner product, $\langle f, g \rangle_{\mathcal{K}} = 0$ for all $g \in \mathcal{K}$ implies $f = 0$.

Next observe that the functional subdifferential of $\mathcal{L}\{f, \lambda\}$ with respect to f satisfies (Rockafellar, 1996)

$$\begin{aligned} \partial_f \mathcal{L}\{f, \lambda\} = & \sum_{i=1}^m \partial_{f(x_i)} [L\{f, X\} + \lambda^\top C\{f, X\}] k(x_i, \cdot) \\ & + 2f \partial_{\langle f, f \rangle} \Omega(\langle f, f \rangle_{\mathcal{K}}). \end{aligned} \quad (3)$$

Here ∂ is understood to be the subdifferential with respect to the argument wherever the function is not differentiable, since C and Ω only depends on $f(x_i)$, the subdifferential always exists with respect to $[f(x_1), \dots, f(x_m)]^\top$. Since for stationarity the variational derivative needs to vanish, we have $0 \in \partial_f \mathcal{L}\{f, \lambda\}$ and consequently $f = \sum_i \alpha_i k(x_i, \cdot)$ for some $\alpha_i \in \partial_{f(x_i)} [L\{f, X\} + \lambda^\top C\{f, X\}]$. This proves the claim. ■

Theorem 12 (Semiparametric Extension) *The same result holds if the optimization is carried out over $f + g$, where $f \in \mathcal{K}$, and g is a parametric addition to f . Again f lies in the span of $k(x_i, \cdot)$.*

Proof [sketch only] In the Lagrange function the partial derivative with respect to f needs to vanish just as in (3). This is only possible if f is contained in the span of kernel functions on the data. ■

4.2. Application to general spline smoothing

We consider the general spline smoothing problem as presented in (Wahba, 1990), except we are considering

Kreĭn spaces. The general spline smoothing is defined as the function stabilizing (that is finding the stationary point) the following criterion:

$$J_m(f) = \frac{1}{m} \sum_{i=1}^m (y_i - f(x_i))^2 + \lambda \langle f, f \rangle_{\mathcal{K}}. \quad (4)$$

The form for the solution of equation (4) is given by the representer theorem, which says that the solution (if it exists) is the solution of the linear equation

$$(K + \lambda \mathbf{I})\alpha = y,$$

where $K_{ij} = k(x_i, x_j)$ is the Gram matrix.

The general spline smoothing problem can be viewed as applying Tikhonov regularization to the interpolation problem. However, since the matrix K is indefinite, it may have negative eigenvalues. For values of the regularization parameter λ which equal a negative eigenvalue of the Gram matrix K , $(K + \lambda \mathbf{I})$ is singular. Note that in the case where K is positive, this does not occur. Hence, solving the Tikhonov regularization problem directly may not be successful. Instead, we use the subspace expansion from Theorem 11 directly.

5. Algorithms for Kreĭn space Regularization

Tikhonov regularization restricts the solution of the interpolation error $\frac{1}{m} \sum_{i=1}^m (y_i - f(x_i))^2$ to a ball of radius $1/\lambda \langle f, f \rangle_{\mathcal{K}}$. Hence, it projects the solution of the equation onto the ball. To avoid the problems of singular $(K + \lambda \mathbf{I})$, the approach we take here is to set $\lambda = 0$, and to find an approximation to the solution in a small subspace of the possible solution space. That is, we are solving the following optimization problem,

$$\begin{aligned} \underset{f \in \mathcal{L}}{\text{stabilize}} & \frac{1}{m} \sum_{i=1}^m (y_i - f(x_i))^2 \\ \text{subject to} & f \in \mathcal{L} \subset \text{span}\{\alpha_i k(x_i, \cdot)\}. \end{aligned} \quad (5)$$

We describe several different ways of choosing the subspace \mathcal{L} . Defining $T : \mathcal{K} \rightarrow \mathbb{R}^m$ to be the evaluation functional (Definition 4), we can express the interpolation problem $f(x_i) = y_i$ given the training data $(x_1, y_1), \dots, (x_m, y_m) \in (\mathcal{X} \times \mathbb{R})^m$, as the linear system $Tf = y$, where $f \in \mathcal{K}$, and \mathcal{K} is a RKKS. Define $T^* : \mathbb{R}^m \rightarrow \mathcal{K}$ to be the adjoint operator of T such that $\langle Tf, y \rangle = \langle f, T^*y \rangle$. Note that since T operates on elements of a Kreĭn space, $TT^* = K$ is indefinite.

5.1. Truncated Spectral Factorization

We perform regularization by controlling the spectrum of K . We can obtain the eigenvalue decomposition of K , $Ku_i = \mu_i u_i$, where u_1, \dots, u_m are the

orthonormal eigenvectors of K , and μ_i are the associated nonzero eigenvalues (assume K is regular). Let $v_i = \frac{\text{sign}(\mu_i)}{\sqrt{|\mu_i|}} T^* u_i$, then v_i are the orthogonal eigenvectors for T^*T . The solution of $Tf = y$ (if it exists), is given by

$$f = \sum_{i=1}^m \frac{\langle y, u_i \rangle}{\mu_i} T^* u_i$$

Intuitively, we associate eigenvalues with large absolute values to the underlying function, and eigenvalues close to zero corresponds to signal noise. The Truncated Spectral Factorization (TSF) (Engl & K ugler, 2003) method can be obtained by setting all the eigenvalues of small magnitude to zero. This means that the solution is in the subspace

$$\mathcal{L} = \text{span}\{T^* u_i, |\mu_i| > \lambda\}$$

5.2. Iterative Methods

Iterative methods can be used to minimize the squared error $J(f) := \frac{1}{2} \|Tf - y\|^2$. Since $J(f)$ is convex, we can perform gradient descent. Since $\nabla_f J(f) = T^*Tf - T^*y$, we have the iterative definition $f_{k+1} = f_k - \lambda(T^*Tf - T^*y)$, which results in Landweber-Fridman (LF) iteration (Hanke & Hansen, 1993). The solution subspace in this case is the polynomial

$$\mathcal{L} = \text{span}\{(\mathbf{I} - \lambda T^*T)^k T^*y\} \text{ for } 1 \leq k \leq m.$$

A more efficient method which utilizes the Krylov subspaces is MR-II (Hanke, 1995). MR-II, which generalizes conjugate gradient methods to indefinite kernels, searches for the minimizer of $\|K\alpha - y\|$ within the Krylov subspace

$$\mathcal{L} = \text{span}\{Kr_0, K^2r_0, \dots, K^{k-2}r_0\},$$

where $r_0 = y - K\alpha_0$. The algorithm is shown in Figure 1. The convergence proof and regularization behavior can be found in (Hanke, 1995).

5.3. Illustration with Toy Problem

We apply TSF, LF, and MR-II to the spline approximation of $\text{sinc}(x)$ and $\cos(\exp(x))$. The experiments was performed using 100 random restarts. The results using a Gaussian combinations kernel are shown in Figure 2. The aim of these experiments is to show that we can solve the regression problem using iterative methods. The three methods perform equally well on the toy data, based on visually inspecting the approximation. TSF requires the explicit computation of the largest eigenvalues, and hence would not be

```

 $r_0 = y - Sx_0; r_1 = r_0; x_1 = x_0;$ 
 $v_{-1} = 0; v_0 = Sr_0; w_{-1} = 0; w_0 = Sv_0;$ 
 $\beta = \|w_0\|; v_0 = v_0/\beta; w_0 = w_0/\beta;$ 
 $k = 1;$ 
while (not stop) do
   $\varrho = \langle r_k, w_{k-1} \rangle; \alpha = \langle w_{k-1}, Sw_{k-1} \rangle;$ 
   $x_{k+1} = x_k + \varrho v_{k-1}; r_{k+1} = r_k + \varrho w_{k-1};$ 
   $v_k = w_{k-1} - \alpha v_{k-1} - \beta v_{k-2};$ 
   $w_k = Sw_{k-1} - \alpha w_{k-1} - \beta w_{k-2};$ 
   $\beta = \|w_k\|; v_k = v_k/\beta; w_k = w_k/\beta;$ 
   $k = k + 1;$ 
end while

```

Figure 1. Algorithm: MR-II. Note that there is only one matrix-vector product in each iteration. Since a matrix-vector product is $\mathcal{O}(m)$, the total number of operations is just $\mathcal{O}(km)$, where k is the number of iterations.

suitable for large problems. LF has been previously shown to have slow convergence (Hanke & Hansen, 1993), requiring a large number of iterations. MR-II has the benefits of being an iterative method and also has faster convergence. The results above required 30 iterations for LF, but only 8 for MR-II.

6. Conclusion

The aim of this paper is to introduce the concept of an indefinite kernel to the machine learning community. These kernels, which induce an RKKS, exhibit many of the properties of positive definite kernels. Several examples of indefinite kernels are given, along with their spectral properties. Due to the lack of positivity, we stabilize the loss functional instead of minimizing it. We have proved that stabilization provides us with a representer theorem, and also generalization error bounds via the Rademacher average. We discussed regularization with respect to optimizing in Kre in spaces, and illustrated the spline smoothing problem on toy datasets.

References

- Alpay, D. (2001). *The schur algorithm, reproducing kernel spaces and system theory*, vol. 5 of *SMF/AMS Texts and Monographs*. SMF.
- Azizov, T. Y., & Iokhvidov, I. S. (1989). *Linear operators in spaces with an indefinite metric*. John Wiley & Sons. Translated by E. R. Dawson.
- Bogn ar, J. (1974). *Indefinite inner product spaces*. Springer Verlag.

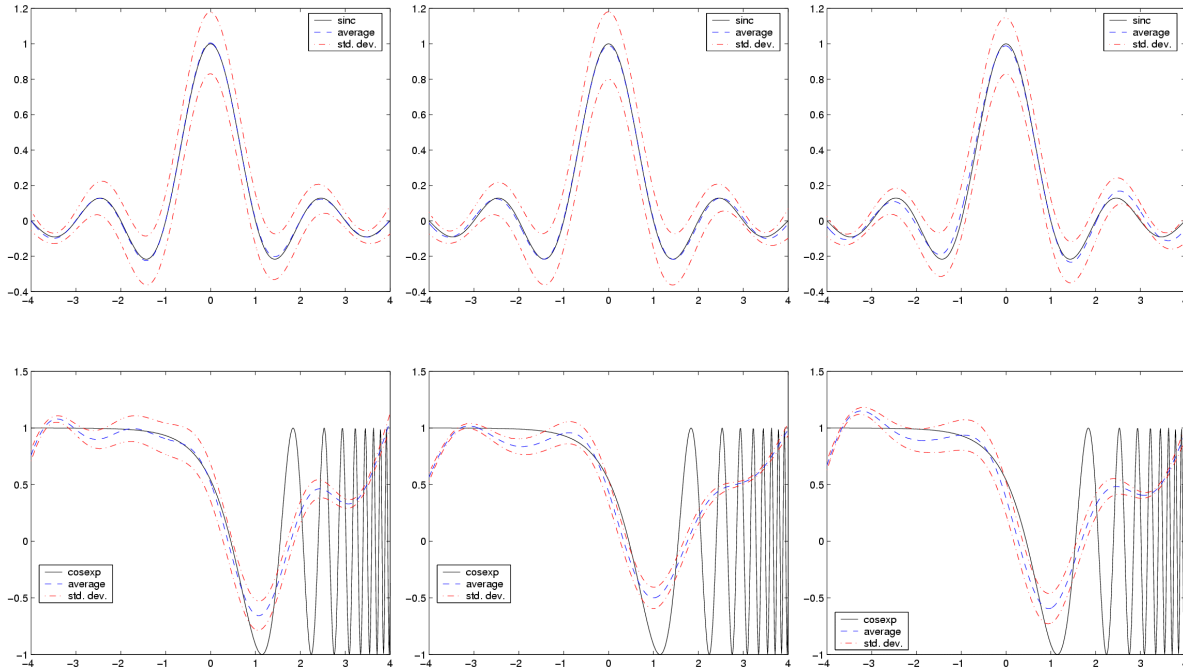


Figure 2. Mean and one standard deviation of 100 random experiments to estimate $\text{sinc}(x)$ (top row) and $\text{cos}(\exp(x))$ (bottom row) using the Gaussian combination with $\sigma_1 = 0.8, \sigma_2 = 1.2, \sigma_3 = 10$. The left column shows the results using TSF, the middle column using LF, and the right column using MR-II

Engl, H. W., & Kügler, P. (2003). Nonlinear inverse problems: Theoretical aspects and some industrial applications. *Inverse Problems: Computational Methods and Emerging Applications Tutorials*, UCLA.

Haasdonk, B. (2003). Feature space interpretation of SVMs with non positive definite kernels. Unpublished.

Hanke, M. (1995). *Conjugate gradient type methods for ill-posed problems*. Pitman Research Notes in Mathematics Series. Longman Scientific & Technical.

Hanke, M., & Hansen, P. (1993). Regularization methods for large-scale problems. *Surveys Math. Ind.*, 3, 253–315.

Hassibi, B., Sayed, A. H., & Kailath, T. (1999). *Indefinite-quadratic estimation and control: A unified approach to h^2 and h^∞ theories*. SIAM.

Lin, H.-T., & Lin, C.-J. (2003). A study on sigmoid kernels for svm and the training of non-psd kernels by smo-type methods. March.

Mary, X. (2003). *Hilbertian subspaces, subdualities*

and applications. Doctoral dissertation, Institut National des Sciences Appliquées Rouen.

Mendelson, S. (2003). A few notes on statistical learning theory. *Advanced Lectures in Machine Learning* (pp. 1–40). Springer Verlag.

Ong, C. S., & Smola, A. J. (2003). Machine learning with hyperkernels. *International Conference of Machine Learning* (pp. 568–575).

Rockafellar, R. T. (1996). *Convex analysis*. Princeton Univ. Pr. Reprint edition.

Schölkopf, B., Herbrich, R., Smola, A., & Williamson, R. (2001). A generalized representer theorem. *Proceedings of Computational Learning Theory*.

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels*. MIT Press.

Smola, A. J., Ovari, Z. L., & Williamson, R. C. (2000). Regularization with dot-product kernels. *NIPS* (pp. 308–314).

Vapnik, V. N. (1998). *Statistical learning theory*. John Wiley & Sons.

Wahba, G. (1990). *Spline models for observational data*. SIAM.