Spectral Analysis of Symmetric and Anti-Symmetric Pairwise Kernels

MARKUS VILJANEN

PRO GRADU-TUTKIELMA

TURUN YLIOPISTO

INFORMAATIOTEKNOLOGIAN LAITOS

TIETOJENKÄSITTELYTIEDE

MARRASKUU 2015

Contents

1 Prelude

1.1 Thesis topic

This thesis presents my small contribution to the topic of symmetric and anti-symmetric kernels in the context of a comprehensive theoretical framework developed by other researchers. The mathematical theory of kernels defining a function space saw its beginnings in the mid-20th centry (Aronszajn, Theory of Reproducing Kernels, 1948), but the adaptation of kernels into machine learning gained significant traction only in the last two decades (Steinwart & Christmann, 2008). Mathematics formulating and investigating theoretical properties of this learning process is constantly evolving, and major learning theoretic results this thesis builds on are closer to only one decade old (Bauer;Pereverzev;& Rosasco, 2007).

The title of this thesis is inaccessible to a layman, but it can be described in a more general way as: "Suppose that we have a particular algorithm which is able to form a model of a quantifiable relationship after seeing only examples of it. If this algorithm is adapted to learn relations between things, and we know that those relations are symmetric or anti-symmetric, can we force it to learn only those relations and then guarantee that it learns better?" Given this question, a vague intuitive answer is a resounding yes. But mathematically formulating what this question exactly means, incorporating the solution into the algorithm and then answering the question is not trivial.

1.2 Thesis contribution

The research presented in the chapters up to the final one has been the work of other researchers, the aim was only to create a compendium of results required to understand the fundamental result of this thesis. This may have proven to be an ill-advised journey; the topic is very much more complicated than the author estimated. My small but central contribution in question is the result on the effective dimension of those kernels. After existence of this result, Prof. Pahikkala's productivity greatly outpaced mine to produce rest of the research draft (Pahikkala, Viljanen, Airola, & Waegeman, 2015).

Choosing this problem was my own decision from multiple options of Prof. Pahikkala's, and it was presented as it is stated. Is $D_\lambda(SKS) \leq D_\lambda(K)$? I did not even understand the question at first. Reviewing linear algebra and gaining significant intuition to the environment around the problem took approximately two months. Forming a big picture from Pahikkala's papers was a central part of this process. The actual solution used the intuition obtained to transform the problem into another formulation to which solution seemed geometrically obvious. A particular name existed for this formulation, principal submatrix, and a literature search arrived at Poincaré's result (Poincaré, 1890) from a different field which completed the final step. This 3-page original proof is contained in Appendix A.1.

An infinite-dimensional version of Poincaré's result was quite hard to track down, understanding the literature took almost a month and two weeks of following multiple references were required to arrive at a textbook which mentioned a similar result (Stenger & Alexander, 1972), but with formulation through complement names! Highly ironically the first legible statement was stated in (Aronszajn, Rayleigh-Ritz and A. Weinstein methods for approximation of eigenvalues: I. Operations in a Hilbert space, 1948), a paper by the originator of the monumental work on kernels (Aronszajn, Theory of Reproducing Kernels, 1948) published the very same year.

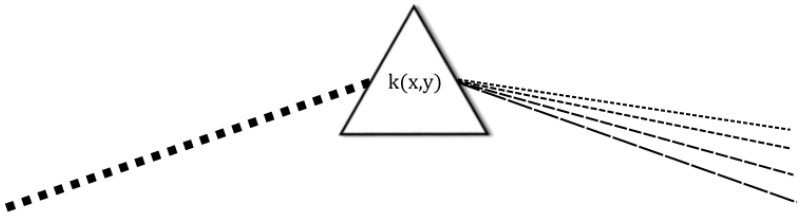
1.3 Research skills

From an academic perspective, the process of writing this thesis certainly cultivated research skills. The central piece of work was solving the problem, and this was the first time I was exposed to a difficult problem without any context. I did not know if the problem was solvable with the knowledge I had; if, when and how to acquire more knowledge. Not visible in this thesis are the attempts that did not go anywhere, or reviews of knowledge that were not fruitful for solving the problem. Those alone might total the length this thesis. Even though in hindsight much of that labour was wasted and naïve, developing these skills was a success because in the future I know how to be wiser about approaching similar problems.

Yet the proof of this result was a minor consumption of my time. Most of the time was spent orienting around the vast machine learning literature which was or only seemed to be related to the result. The task was not made easy by the fact that each

paper seemed to use a slightly different formulation and the required material was scattered among multiple papers. This is perhaps even more important research skill the process taught me; to plan and execute research despite vagueness and open-endedness, orient oneself around a vast body of proximally related literature and perform a sufficiently thorough yet efficient review.

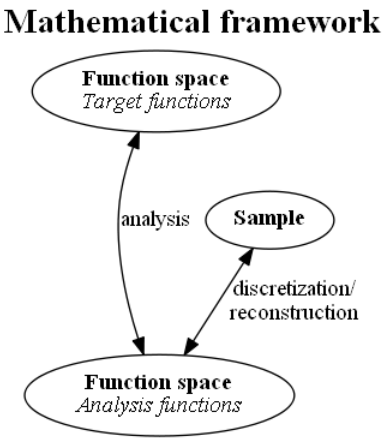
There are many more research skills that I have in my stupor only recently realized. Much more optimal approach would have been to plan a smooth progression small goals instead of directly attempting sink-or-swim on a difficult problem. Collaboration was an essential part of research: wherever it seemed like there might be results about a problem I consulted the related literature. In the future I should also learn to better collaborate with other researchers, since orienting around the vastness of the research context may be possible only after a decade of experience.



2 Machine Learning

2.1 Conceptual framework

The goal of supervised machine learning can be stated in a very simple way. We are provided data as `(input,output)` pairs:



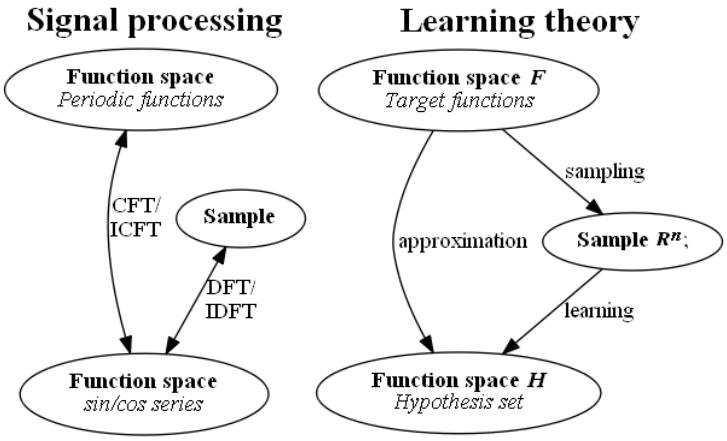
INPUT	OUTPUT
(21.39, 15.52)	11.80
(18.17, 21.55)	0.59
(20.38, 16.29)	8.48
…	…

The goal is to form a model, such that for new `input`, the model gives an `output` which is close to the true unknown answer. Formally, we have:

- Input space X and output space Y
- Unknown target function $f: X \to Y$
- Data set $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$ where $y_i = f(x_i)$ and $x_i \in X, y_i \in Y$.
- A learning algorithm $\mathcal{L}$ which returns a function $g = \mathcal{L}(D)$ such that $g \approx f$.
- The learning algorithm $\mathcal{L}$ also often explicitly encompasses:
  - A hypothesis set H of functions $X \to Y$ from which the function is selected
  - An error measure $E(D, g)$ which defines the best approximation hypothesis.
  - An efficient algorithm to find $g \in H$ minimizing $E(D,g)$.

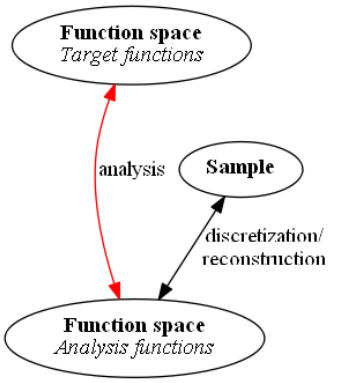
Machine learning is closely related to other fields, sharing mathematical foundations and methods in particular with learning theory, approximation theory, inverse problems, signal processing and statistics. The terminology is different however, and often the shared commonality is not obvious from the methods which make field-specific assumptions. For example, in signal processing, the Discrete Fourier transform (DFT) is used to transform a function into a discrete set of (point, value)-

pairs with a well defined inverse transform (IDFT). Given a sample satisfying certain requirements, this means we fully reconstruct an underlying periodic function from the sample. In machine learning, the goal is not to discretize and reconstruct known functions, and the lack of assumptions we can make about the target means full reconstruction is impossible. But both function-sample-transforms can be seen as instances of sampling.



This thesis aims to use the mathematical formalizations on the most general level possible, and then use their well-known special cases for practical applications. This allows a cross-disciplinary approach with increased concinnity at the cost of concreteness. The learning theory literature which encompasses RLS kernel methods and its learning bounds is often written in a highly general level as well, making varying special assumptions as needed.

2.2 Approximation



Linear combination of functions belonging to a certain function space can be used in the analysis of properties of a known target function or derive an expression for a function uniquely defined through some property. Function representation and approximation using a set of functions is a very important topic in mathematics, such as the use of power series in analysis, and has close connections to learning. Representation expresses the full function using the set of analysis functions, whereas approximation is an estimate with an error term of certain magnitude, possibly bounded locally. An accessible and application oriented resource is (Moon, 1999) whereas theoretical discussion is found for example in

(Kreyszig, 1989). One of the central mathematical tools used in approximation is the theory of abstract vector spaces and projections.

Projection Theorem (Kreyszig, 1989)

Let X be a Hilbert space and $X' \subseteq X$ a closed subspace of X. For any vector $\overline{x} \in X$, there exists a unique vector $\overline{x}_0 \in X'$ closest to $\overline{x}$: $\|\overline{x} - \overline{x}_0\| \leq \|\overline{x} - \overline{x}_1\|$ for all $\overline{x}_1 \in X'$. The point $\overline{x}_0$ is a minimizer of $\|\overline{x} - \overline{x}_0\|$ if and only if $\overline{x} - \overline{x}_0 \perp X'$.

2.2.1 Approximation with a finite set of vectors

Let X be a real Hilbert space and $X' = \text{span}\{p_1, \dots, p_m\}$ be a subspace spanned by linearly independent vectors $p_1, \dots, p_m$. Given a vector $\overline{x} \in X$, the approximation goal is to find vector $\hat{x} \in X'$ which approximates $\overline{x}$:

$$\overline{x} = \hat{x} + \overline{e} = \sum_{i=1}^{m} c_i \overline{p}_i + \overline{e}$$

The best approximation can be defined as the vector which minimizes the norm of the error:

$$\|\overline{x} - \hat{x}\| = \|\overline{e}\|$$

Using the projection theorem, the error is minimized when the error $\overline{e} = \overline{x} - \hat{x}$ is orthogonal to X'. This is the case if and only if the error is orthogonal to each vector that spans X':

$$\left\langle \overline{x} - \sum_{i=1}^{m} c_i \overline{p}_i , \overline{p}_j \right\rangle = 0, j = 1, \dots, m$$

These conditions may be written as a matrix

$$\begin{pmatrix} \langle \overline{p}_1, \overline{p}_1 \rangle & \cdots & \langle \overline{p}_m, \overline{p}_1 \rangle \\ \vdots & \ddots & \vdots \\ \langle \overline{p}_1, \overline{p}_m \rangle & \cdots & \langle \overline{p}_m, \overline{p}_m \rangle \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_m \end{pmatrix} = \begin{pmatrix} \langle \overline{x}, \overline{p}_1 \rangle \\ \vdots \\ \langle \overline{x}, \overline{p}_m \rangle \end{pmatrix}$$

Define the Gramian matrix $G_{ij} = \left\langle \overline{p}_i, \overline{p}_j \right\rangle$ (Moon, 1999) and the expression can be written:

$$G\overline{c} = \overline{p}$$

The cofficients solved:

$$\overline{c} = G^{-1}\overline{p}$$

The Gramian is:

- Symmetric: follows from the symmetry of the inner product $\left\langle \overline{p}_i, \overline{p}_j \right\rangle = \left\langle \overline{p}_j, \overline{p}_i \right\rangle$.
- Positive semidefinite, futhermore positive definite iff $p_1, \dots, p_m$ are linearly independent:$\overline{y}^* G \overline{y} = \sum_{i=1}^{m} \sum_{j=1}^{m} y_i y_j \left\langle \overline{p}_i, \overline{p}_j \right\rangle = \sum_{i=1}^{m} \sum_{j=1}^{m} \left\langle y_i \overline{p}_i, y_j \overline{p}_j \right\rangle = \left\langle \sum_{i=1}^{m} y_i \overline{p}_i , \sum_{j=1}^{m} y_j \overline{p}_j \right\rangle = \left\| \sum_{i=1}^{m} y_i \overline{p}_i \right\|^2 \geq 0.$

In the special case that the vectors $p_1, \dots, p_m$ are orthogonal, we do not have to solve a linear system, because the coefficients are simply $c_j = \left\langle \overline{x}, \overline{p}_j \right\rangle / \langle \overline{p}_i, \overline{p}_i \rangle$. Futhermore, orthogonal vectors of unit length are called orthonormal and $c_i = \left\langle \overline{x}, \overline{p}_j \right\rangle$. Because of this simplification, in many approximation tasks the approximating set is defined so that the vectors are orthonormal.

*2.2.2 **Approximation with infinite set of basis vectors***

Given a set of orthonormal vectors $\{u_1, \dots u_n\}$, we may associate with any $x \in X$ the series

$$\hat{x} = \sum_{i=1}^{n} \langle \overline{x}, u_i \rangle u_i$$

Then from the considerations before, let $\overline{e} = \overline{x} - \hat{x}$ with the orthogonality $\overline{e} \perp \hat{x}$ and

$$\overline{x} = \hat{x} + \overline{e}$$

From the orthogonality property:

$$\|\overline{x}\|^2 = \|\hat{x}\|^2 + \|\overline{e}\|^2 \Longrightarrow \|\overline{x}\| \geq \|\hat{x}\| \geq 0$$

For $n = 1,2, \dots$, the sequence $\|\hat{x}\|^2 = \sum_{i=1}^{n} |\langle \overline{x}, u_i \rangle|^2$ is a monotonically increasing bounded sequence, which thus converges. This leads to the following.

Lemma: Bessel's inequality

Let (u_i) be a sequence of orthonormal vectors, then for any $x \in X$:

$$\sum_{i=1}^{\infty} |\langle x, u_i \rangle|^2 \leq \|x\|^2$$

Theorem: Converge of coefficients (Kreyszig, 1989)

The series $s_n = \sum_{i=1}^{n} \alpha_i u_i$ of partial sums of vectors converges in the norm of a Hilbert space X, i.e:

$$\|s_n \to s\| \text{ as } n \to \infty$$

if and only if the series $\alpha_n = \sum_{i=1}^{n} \alpha_i^2$ converges.

Proof

Because of orthonormality, $\|s_m - s_n\| = \|\alpha_m u_m + \cdots + \alpha_{n+1} u_{n+1}\| = \alpha_m^2 + \cdots + \alpha_{n+1}^2$, i.e. the series s_n is Cauchy iff α_n is Cauchy. Both X and $\mathbb{R}$ are complete, in which series is Cauchy iff it converges.

In particular, this combined with Bessel's inequality implies that for any $\overline{x} \in X$, the series $\hat{x} = \sum_{i=1}^{\infty} \langle \overline{x}, u_i \rangle u_i$ converges to an element in X. The natural question is when $\hat{x} = \overline{x}$?

An essential condition is the totality of (u_i), which can be shown to be equivalent to multiple conditions, of which Parseval's equality is common (Moon, 1999). The existence of countable total set (u_i) can be shown to be equivalent to seperability of X (Kreyszig, 1989).

Definition

An orthonormal set $U = \{u_1, u_2 \dots\}$ in a Hilbert space X is total if its span is dense in X:

$$\overline{span(U)} = X$$

Lemma: Parseval's equality (Kreyszig, 1989).

An orthonormal set $U = \{u_1, u_2 \dots\}$ is total if and only if Bessel's inequality is an equality:

$$\forall \overline{x} \in X : \sum_{i=1}^{\infty} \langle \overline{x}, u_i \rangle^2 = \|\overline{x}\|^2$$

Furthermore, then $\overline{x} = \sum_{i=1}^{\infty} \langle \overline{x}, u_i \rangle u_i$.

2.2.3 *Example: Linear regression as approximation in* $\mathbb{R}^n$

Given a set of vectors $\{\overline{x}_1, \overline{x}_2, \dots \overline{x}_n\}$ with $\overline{x}_i \in \mathbb{R}^d$ and outputs $\{y_1, \dots, y_n\}$, the goal of linear regression is to model an unknown linear relationship

$$x_{i1}w_1 + \cdots + x_{id}w_d = y_i \text{ for } i = 1 \dots n$$

The best approximation is defined as the minimizer of the squared error $\sum_{i=1}^{n}\left(\sum_j x_{ij}w_j - y_i\right)^2$, which may be written in matrix form by stacking vectors as rows of $X = (\overline{x}_1, \dots \overline{x}_n)^T$:

$$E(\overline{w}) = \|X\overline{w} - \overline{y}\|_2^2$$

But $\|.\|_2^2$ is simply the Euclidean norm in the space $\mathbb{R}^n$ and denoting the columns $X = \left(\overline{x}'_1, \dots, \overline{x}'_n\right)$ we have the approximator $X\overline{w} = \sum_{i=1}^{n}\overline{x}'_i w_i$. By the projection theorem the minimizer is the orthogonal projection in $\mathbb{R}^n$ into the column space of X, therefore for the minimizer:

$$\begin{pmatrix} \overline{x}'^T_1\overline{x}'_1 & \cdots & \overline{x}'^T_n\overline{x}'_1 \\ \vdots & \ddots & \vdots \\ \overline{x}'^T_1\overline{x}_n & \cdots & \overline{x}'^T_n\overline{x}'_n \end{pmatrix}\begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} = \begin{pmatrix} \overline{x}'^T_1\overline{y} \\ \vdots \\ \overline{x}'^T_n\overline{y} \end{pmatrix}$$

Which may be written more succinctly in matrix form, since in this case $G = X^TX$:

$$X^TX\overline{w} = X^T\overline{y}$$

If the columns are linearly independent this is can be solved directly.

2.2.4 *Example: Approximation in* $L^2[a,b]$

Let $L^2[0,1]$ be the Hilbert space obtained by taking the completion of the inner product space of continuous real valued functions on the interval [0,1] with the inner product

$$\langle f, g\rangle = \int_0^1 f(t)g(t)dt$$

Suppose we want to find the best $m-1$ degree polynomial approximation of a continuous function $f(t)$ in the sense that we minimize the norm induced by the inner product $\|f\| = \sqrt{\langle f, f\rangle}$:

$$\int_0^1 \left(f(t) - p(t)\right)^2 dt$$

Take $1, t, \dots, t^{m-1}$ as the approximation vectors so that $p(t) = c_0 + c_1t + c_2t^2 + \cdots + c_{m-1}t^{m-1}$. The minimum norm approximation problem becomes:

$$\begin{bmatrix} \langle 1,t\rangle & \cdots & \langle 1,t^{m-1}\rangle \\ \vdots & \ddots & \vdots \\ \langle t^{m-1},1\rangle & \cdots & \langle t^{m-1},t^{m-1}\rangle \end{bmatrix}\begin{bmatrix} c_0 \\ \vdots \\ c_m \end{bmatrix} = \begin{bmatrix} \langle f,1\rangle \\ \vdots \\ \langle f,t^{m-1}\rangle \end{bmatrix}$$

Where:

$$\langle t^i, t^j\rangle = \int_0^1 t^{i+j}dt = \frac{1}{i+j+1}$$

The Gramian is the well known ill-conditioned Hilbert matrix:

$$H = \begin{bmatrix} 1 & \cdots & \frac{1}{m} \\ \vdots & \ddots & \vdots \\ \frac{1}{m} & \cdots & \frac{1}{2m} \end{bmatrix}$$

Suppose we solved the coefficients (c_i) of the above problem $H\overline{c} = \overline{f}$ through taking the inverse $\overline{c} = H^{-1}\overline{f}$. The solution then is:

$$p(t) = c_0 + c_1 t + c_2 t^2 + \cdots + c_{m-1}t^{m-1}$$

In practice we might use the Legendre polynomials, which are obtained by applying the Gram-Schmidt algorithm to $\{1, t, t^2 \ldots\}$ and also form a total orthonormal set on $L^2[0,1]$ (Kreyszig, 1989).

2.2.5 ***Example: Fourier series***

Any (complex) function periodic on $[0,2\pi)$ may be represented using the series (Moon, 1999):

$$f(t) = \sum_{n=-\infty}^{\infty} \langle x, p_n\rangle p_n(t)$$

Where $p_n(t)$ are the following orthonormal Fourier basis functions:

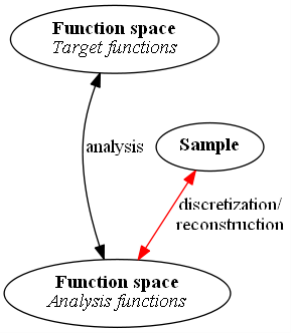
$$p_n(t) = \frac{1}{\sqrt{2\pi}}e^{nti}$$

The coefficients can be found simply by:

$$\langle x, p_n\rangle = \frac{1}{\sqrt{2\pi}}\int_0^{2\pi} f(t)e^{-nti}$$

The representation is the Fourier transform of $f(t)$.

2.3 Sampling



Sampling a function is an important mathematical concept, but the context is very different in machine learning compared to those often occurring in theoretical discussion or applications such as signal processing. This setting has been recently used in different forms in the machine learning literature, see for example the discussion in (De Vito;Rosasco;Caponetto;De Giovannini;& Odone, 2005). Connection to learning theory and clear formalization were recently given in (Smale & Zhou, Shannon Sampling and Function reconstruction, 2004) and (Smale & Zhou, Shannon sampling II: Connections to learning theory, 2005). Below we present their formalization and relate it to our goal.

2.3.1 Discretization of a function

Let $\mathcal{H}$ be a space of functions of the form $f: X \to \mathbb{R}$. In discretization, the goal is to define $(x_1, .., x_n) \in X^n$ such that the mapping $\mathcal{F}_{\overline{x}}: f \mapsto \big(f(x_1), \dots, f(x_n)\big)$ is a **bijection** $\mathcal{F}_{\overline{x}}: \mathcal{H} \to \mathbb{R}^n$. This is unlike sampling in machine learning, where the sample is not purposefully chosen and it often is impossible to fully reconstruct the function from the sample, in other words the inverse $\mathcal{F}_{\overline{x}}^{-1}$ does not exist. If the inverse does exists, the map $\mathcal{F}_{\overline{x}}$ is said to discretize the function space. The two following examples of discretization are widely used, see for example (Smale & Zhou, Shannon sampling II: Connections to learning theory, 2005).

Example 1

Consider a space of polynomials $\mathcal{H}_d$ up to degree d: $p(x) = a_0 + a_1 x + \cdots + a_d x^d$. Given any sample of $d+1$ points $\overline{x} = (x_1, .., x_{d+1})$ its value may be computed:

$$\begin{pmatrix} f(x_1) \\ \vdots \\ f(x_{d+1}) \end{pmatrix} = \begin{pmatrix} 1 & \dots & x_1^d \\ \vdots & \ddots & \vdots \\ 1 & \dots & x_{d+1}^d \end{pmatrix} \begin{pmatrix} a_0 \\ \vdots \\ a_d \end{pmatrix}$$

More succinctly, we have a linear transformation $L_{\overline{x}}: \mathbb{R}^{d+1} \to \mathbb{R}^{d+1}$ known as the Vandermonde matrix:

$$f(\overline{x}) = L_{\overline{x}} \overline{a}$$

When the x_i are distinct $L_{\overline{x}}: \overline{a} \mapsto \left(f(x_i)\right)_{i=1}^{d+1}$ is a bijection, since the determinant is nonzero. Similarly the map $\sum_{i=0}^{d} a_i x^i \to \overline{a}$ from coefficients to the polynomial is a bijection, so their composition is a discretizing bijection $\mathcal{F}_{\overline{x}}: \mathcal{H}_d \to \mathbb{R}^{d+1}$.

Example 2 (Whittaker-Shannon-Nyquist Sampling Theorem)

Let $\phi(x) = \frac{\sin(\pi x)}{\pi x}$ and $\phi_t(x) = \phi(x - t)$. If a function $f \in L^2(\mathbb{R})$ has its Fourier transform supported in $[-\pi, \pi]$, then

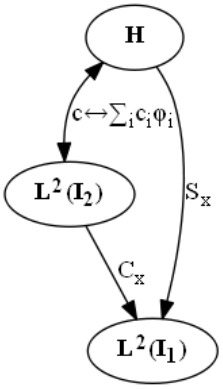
$$f = \sum_{t \in \mathbb{Z}} f(t)\phi_t$$

Functions $\left\{\sum_{t\in\mathbb{Z}} a_t\, \phi_t : \overline{a} \in \mathbb{R}^{\mathbb{Z}}\right\}$ are the representation space, and as a consequence of the theorem the bijection $\mathcal{F}_{\overline{x}}: f \mapsto \left(f(t)\right)_{t\in\mathbb{Z}}$ exists.

2.3.2 ***Reconstruction of a function***

Following the formulation of (Smale & Zhou, Shannon Sampling and Function reconstruction, 2004), we present the sampling framework for machine learning as follows. Let $f: \mathbb{R}^d \to \mathbb{R}$ be function and for $x_i \in \mathbb{R}^d$ denote the finite or infinite set where the function has been sampled $\overline{x} = \{x_1, x_2, \dots\} = \{x_i\}_{i\in I_1}$. Assuming countability the index set I_1 is $\{1, \dots n\}$ or $\mathbb{N}$. Define the sample space $\ell^2(I_1)$ ($= \mathbb{R}^n$ or ℓ^2), and denote the values of f at $\overline{x}$ as $\overline{y} = (y_i)_{i\in I_1} \in \ell^2(I_1)$.

Likewise, let $\mathcal{H}$ be a Hilbert space of continuous functions and let $\{\phi_i\}_{i\in I_2}$ be a finite or countably infinite dimensional orthonormal basis of $\mathcal{H}$. Define the coefficient space $\ell^2(I_2)$ as the set of square summable sequences $(a_i)_{i\in I_2}$ with the inner product $\langle a, b\rangle = \sum_{i\in I_2} a_i b_i$. Then the mapping $\sum_{i\in I_2} c_i\phi_i \mapsto \overline{c}$ between $\mathcal{H} \to \ell^2(I_2)$ is an isomorphism. The space of sequences is defined so that instead of $f \in \mathcal{H}$ we may equivalently consider the coefficient sequence $(a_i)_{i\in I_2}$ of the series $f = \sum_{i\in I_2} a_i\phi_i$:



The sampling operator $S_{\overline{x}}$ is a map from the function space $\mathcal{H} \to \ell^2(I_1)$:

$$S_{\overline{x}}(f) = \left(f(x_i)\right)_{i\in I_1}$$

The sampling operator from the sequence space $\mathcal{C}_{\overline{x}}: \ell^2(I_2) \to \ell^2(I_1)$ is an infinite matrix:

$$\mathcal{C}_{\overline{x}} = \left(\phi_j(x_i)\right)_{i\in I_1, j\in I_2}$$

This is well defined because it denotes the linear operator

$$(\mathcal{C}_{\overline{x}}\overline{c})_i = \sum_{j \in I_2} c_j \phi_j(x_i)$$

Considering either $\mathcal{C}_{\overline{x}}$ or $S_{\overline{x}}$ as the sampling operator, there are two fundamental problems with its inverse. Since the sample was provided and not purposefully chosen, the operator may be neither injective nor surjective. If the inverse is not a bijection, we need to clarify how we define function reconstruction.

a) Surjectivity: If the function space $\mathcal{H}$ is not expressive enough on a large enough sample $\overline{x}$, for values $(y_i)_{i \in I_1}$ there may not be a function f such that $S_{\overline{x}} f = \overline{y}$.
b) Injectivity: If the function space $\mathcal{H}$ is expressive enough or the sample $\overline{x}$ too small, there may very well be two functions f_1, f_2 such that $S_{\overline{x}} f_1 = S_{\overline{x}} f_2$.

Because it may be impossible to use an inverse operator, instead of full function reconstruction one needs to formulate a well-defined minimization problem. Sometimes injectivity or surjectivity may be assumed, or the existence problem of the inverse solved by:

a) Defining an approximation task, such as minimizing an error $\mathcal{E}_{\overline{x},\overline{y}}(f)$.
b) Defining a property which makes the solution $f \in \mathcal{H}$ unique, such as penalized or minimum norm.

Aside from injectivity and surjectivity, there may be an additional problem with the stability of the solution. It will be shown that in the minimum norm or unsubstantially penalized solution reconstructed from a sample of an unknown function with little noise added produces a very different solution. In learning this expressed as generalization ability, which this solution does not have: The learned function $\hat{f}$ approximates the target f in the known sample $\overline{x}$ very closely: $S_{\overline{x}} \hat{f} \cong S_{\overline{x}} f$. But given a new sample $\overline{x}'$, the values $S_{\overline{x}'} \hat{f}$ it produces can be wildly different from $S_{\overline{x}'} f$.

Clearly new tools are needed to determine a suitable hypothesis set $\mathcal{H}$ and the rule which picks an approximating function $\hat{f} \in \mathcal{H} \approx f$ based on the sample $S_{\overline{x}} f$. Aside from injectivity, surjectivity and generalization, these tools need to incorporate considerations of the computational efficiency as well. Because there was nothing sampling operator specific in the formulation, for full generality it may be studied as an abstract linear inverse problem.

2.4 Learning

2.4.1 Learning as an inverse problem

Definition: Hadamard's criteria

Consider an abstract linear operator $T: \mathcal{X} \to \mathcal{Y}$ equation of the form

$$Tx = y$$

Hadamard's definition of a well-posed inverse problem, with equivalent conditions for the special case of a linear operator in parenthesis, is (Engl;Hanke;& Neubauer, 1996):

a) For all admissible data, a solution exists ($\mathcal{R}(T) = \mathcal{Y}$)
b) For all admissible data, a solution is unique ($\mathcal{N}(T) = \{0\}$)
c) The solution depends continuously on the data (T^{-1} is continuous/bounded)

Supervised learning is an algorithm which solves a well-posed inverse problem. Since $S_{\overline{x}}: \mathcal{H} \to \ell^2(I_1)$ does not necessarily satisfy these conditions, a new problem formulation is needed.

2.4.2 Least squares and regularized least squares

In this thesis we focus on one particular learning algorithm, called **regularized least squares (RLS)**, which is widely used, effective and furthermore has very useful analytical and statistical properties (Cucker & Smale, 2001). It solves this problem as follows:

a) The goal is to minimize the least squares error,:

$$\mathcal{E}_{\overline{x},\overline{y}}(f) = \sum_{i \in I_1} (f(x_i) - y_i)^2$$

b) Define the optimal solution as the solution with minimum norm $\|\|_{\mathcal{H}}$.
c) Possibly regularize (penalize) the norm to achieve desirable stability properties.

For example in (Smale & Zhou, Shannon Sampling and Function reconstruction, 2004), it is assumed that $\overline{x}$ provides rich data with respect to $\mathcal{H}$ in the sense that

$\inf_{v\in\ell^2(I_2)} \frac{\|\mathcal{C}_{\overline{x}}v\|_{\ell^2(I_1)}}{\|v\|_{\ell^2(I_2)}} > 0$, which is equivalent to $\mathcal{C}_{\overline{x}}^T\mathcal{C}_{\overline{x}}$ having bounded inverse which means that the map is injective. The solution is then easily obtained:

Theorem: Least Squares solution (Smale & Zhou, Shannon Sampling and Function reconstruction, 2004)

The problem is to reconstruct f^* from $\overline{y} \in \ell^2(I_1)$. Consider the minimization problem

$$\underset{f\in\mathcal{H}}{\operatorname{argmin}} \sum_{i\in I_1} (f(x_i) - y_i)^2$$

If $\overline{x}$ provides rich data, the solution is

$$f_{\overline{x}} = \sum_{i\in I_2} c_i \phi_i$$

Where $\overline{c} = L\overline{y}$ with $L = \left(\mathcal{C}_{\overline{x}}^T\mathcal{C}_{\overline{x}}\right)^{-1}\mathcal{C}_{\overline{x}}^T$, which is the Moore-Penrose inverse of $\mathcal{C}_{\overline{x}}$.

Proof

This is a quadratic form that may be expressed:

$$Q(\overline{c}) \coloneqq \left\|\mathcal{C}_{\overline{x},\overline{t}}\overline{c} - \overline{y}\right\|_{\ell^2(I_2)}^2 = \left\langle \mathcal{C}_{\overline{x}}^T\mathcal{C}_{\overline{x}}\overline{c}, \overline{c}\right\rangle_{\ell^2(I_2)} - 2\langle \mathcal{C}_{\overline{x}}\overline{c}, \overline{y}\rangle_{\ell^2(I_1)} + \langle \overline{y}, \overline{y}\rangle_{\ell^2(I_1)}$$

Take the functional derivative and equate $\frac{\delta}{\delta\overline{c}}Q(\overline{c}) = 0$ to find the minimizer:

$$\mathcal{C}_{\overline{x}}^T\mathcal{C}_{\overline{x}}\overline{c} = \mathcal{C}_{\overline{x}}^T\overline{y}$$
$$\overline{c} = \left(\mathcal{C}_{\overline{x}}^T\mathcal{C}_{\overline{x}}\right)^{-1}\mathcal{C}_{\overline{x}}^T\overline{y}$$

If the function space is a Hilbert space, one possible choice for the property which removes the requirement for the injectivity assumption and makes the solution unique is minimum norm or penalized norm solution. To use a penalized norm, define a new error, for $\lambda > 0$:

$$\mathcal{E}_{\overline{x},\overline{y}}(f) = \underset{f\in\mathcal{H}}{\operatorname{argmin}} \sum_{x\in I_1} (f(x_i) - y_i)^2 + \lambda\|f\|_H$$

Theorem: Regularized Least Squares solution (Smale & Zhou, Shannon Sampling and Function reconstruction, 2004)

The problem is to reconstruct f^* from $\overline{y} \in \ell^2(I_1)$. Consider the minimization problem

$$\underset{f \in \mathcal{H}}{\operatorname{argmin}} \sum_{x \in I_1} (f(x_i) - y_i)^2 + \lambda \|f\|_H$$

For an orthonormal basis $\{\phi_i\}_{i \in I_2}$ of $\mathcal{H}$, the solution is

$$f_{\overline{x}} = \sum_{i \in I_2} c_i \phi_i$$

Where $\overline{c} = L\overline{y}$ with $L = \left(\mathcal{C}_{\overline{x}}^T \mathcal{C}_{\overline{x}} + \lambda I\right)^{-1} \mathcal{C}_{\overline{x}}^T$, which is the Moore-Penrose inverse of $\mathcal{C}_{\overline{x}}$.

Proof

This is a quadratic form that may be expressed:

$$\begin{aligned} Q(\overline{c}) &\coloneqq \|\mathcal{C}_{\overline{x}}\overline{c} - \overline{y}\|_{\ell^2(I_1)}^2 + \lambda \|\overline{c}\|_{\ell^2(I_2)}^2 \\ &= \left\langle \mathcal{C}_{\overline{x}}^T \mathcal{C}_{\overline{x}} \overline{c}, \overline{c} \right\rangle_{\ell^2(I_2)} - 2\langle \mathcal{C}_{\overline{x}}\overline{c}, \overline{y} \rangle_{\ell^2(I_1)} + \langle \overline{y}, \overline{y} \rangle_{\ell^2(I_1)} + \lambda \langle \overline{c}, \overline{c} \rangle_{\ell^2(I_2)} \end{aligned}$$

Take the functional derivative and equate $\frac{\delta}{\delta \overline{c}} Q(\overline{c}) = 0$ to find the minimizer:

$$(\mathcal{C}_{\overline{x}}^T \mathcal{C}_{\overline{x}} + \lambda I)\overline{c} = \mathcal{C}_{\overline{x}}^T \overline{y}$$

$$\overline{c} = \left(\mathcal{C}_{\overline{x}}^T \mathcal{C}_{\overline{x}} + \lambda I\right)^{-1} \mathcal{C}_{\overline{x}}^T \overline{y}$$

2.5 Hypotheses spaces

In this section we consider different hypothesis spaces $\mathcal{H}$ from which the sampling operator maps to the sample space $S_{\overline{x}}: \mathcal{H} \to \ell^2(I_1)$. As before, denote $\mathcal{H}$ a Hilbert space of continuous functions and $\{\phi_i\}_{i \in I_2}$ an orthonormal basis of $\mathcal{H}$, so that $\overline{span\left(\{\phi_i\}_{i \in I_2}\right)}$ forms an analysis space. Because of the isomorphism, we may then equivalently operate on the coefficient sequence $(a_i)_{i \in I_2} \in \ell^2(I_2)$ of the series $f = \sum_{i \in I_2} a_i \phi_i$. Given a sample $\overline{y} \in \ell^2(I_1)$, to solve the problem one forms the sampling operator $\mathcal{C}_{\overline{x}}: \ell^2(I_2) \to \ell^2(I_1)$:

$$\mathcal{C}_{\overline{x}} = \left(\phi_j(x_i)\right)_{x \in I_1, j \in I_2}$$

The minimum norm closest $\|\|_2$ approximation solution is given by the pseudoinverse (see chapter 4).:

$$\overline{c}^{\dagger} = \mathcal{C}_{\overline{x}}^{\dagger} \overline{y}$$

Regularized solution is essential if $\mathcal{H}$ is expressive or sample noisy, Tikhonov regularization gives:

$$\overline{c}_\lambda^\dagger = \left(\mathcal{C}_{\overline{x}}^T \mathcal{C}_{\overline{x}} + \lambda I\right)^{-1} \mathcal{C}_{\overline{x}}^T \overline{y}$$

Solving for the coefficients $\overline{c}$, the solution is then given by the isomorphism between $\ell^2(I_2)$ and $\mathcal{H}$:

$$\overline{c} \mapsto \sum_{I \in I_2} c_i \phi_i$$

2.5.1 ***Hypothesis spaces***

Linear hypotheses

Previous examples used functions $f: \mathbb{R} \to \mathbb{R}$ with 1-dimensional domain. In general, the target function may be of the form $f: \mathbb{R}^d \to \mathbb{R}$. Consider now the linear hypothesis space $\mathcal{H}_d$ of functions $\mathbb{R}^d \to \mathbb{R}$:

$$\mathcal{H}_d = \left\{ f(\overline{x}) = \overline{w}^T \overline{x} \mid \overline{w} \in \mathbb{R}^d \right\}$$

Since by definition for any function $f(\overline{x}) \in \mathcal{H}_d$:

$$f(\overline{x}) = w_1 x_1 + \cdots + w_d x_d$$

A basis for this function space is the set of functions $\{\phi_i\}_{i=1}^d$ where $\phi_i(\overline{x}) = \overline{x}_i$, i.e the set $\{x_1, \dots, x_d\}$, with the coefficient space $\overline{w} \in \mathbb{R}^d$. For samples $\{\overline{x}_1, \dots, \overline{x}_n\}$, the sampling matrix is:

$$\mathcal{C}_{\overline{x}} = \begin{pmatrix} (\overline{x}_1)_1 & \cdots & (\overline{x}_1)_d \\ \vdots & \ddots & \vdots \\ (\overline{x}_n)_1 & \cdots & (\overline{x}_n)_d \end{pmatrix}$$

Given a set of samples $\{\overline{x}_1, \dots, \overline{x}_n\}$ and values $\{y_1, \dots, y_n\}$, the solution is obtained with previous tools.

Polynomial hypotheses

Linear hypotheses are very simple. The polynomial hypothesis up to degree m may be generalized from $\mathbb{R} \to \mathbb{R}$ to $\mathbb{R}^d \to \mathbb{R}$. Denote the polynomial hypothesis space $\mathcal{H}_d^m$ of m–degree polynomials $\mathbb{R}^d \to \mathbb{R}$. Then we may define the following set of integer vectors whose entries sum to less than or equal to m:

$$\Omega = \{\overline{n} \in \mathbb{N}^d \mid \sum n_i \leq m\}$$

The space may then be expressed concisely:

$$\mathcal{H}_d = \left\{ \sum_{\sigma \in \Omega} c_\sigma x_1^{\sigma_1} \dots x_d^{\sigma_d} \mid \overline{c} \in \mathbb{R}^{|\Omega|} \right\}$$

In other words, every $f \in \mathcal{H}_d$ may be written:

$$f(\overline{x}) = c_1 x_1^m + c_2 x_1^{m-1} + \cdots + c_{m+1} x_1^{m-1} x_2 + \cdots + c_{|\Omega|} x_d^m$$

Every function is a linear function of the monomials $\phi_\sigma(\overline{x}) = x_1^{\sigma_1} \dots x_d^{\sigma_d}$ where the sum of the exponents is equal to or less than m. Therefore, the monomials $\left\{ x_1^{\sigma_1} \dots x_d^{\sigma_d} \right\}_{\sigma \in \Omega}$ form a basis of $\mathcal{H}_d$ with the coefficient space $\overline{c} \in \mathbb{R}^{|\Omega|}$. For samples $\{\overline{x}_1, \dots, \overline{x}_n\}$, the sampling matrix is:

$$\mathcal{C}_{\overline{x}} = \begin{pmatrix} \phi_{\sigma_1}(\overline{x}_1) & \cdots & \phi_{\sigma_\Omega}(\overline{x}_1) \\ \vdots & \ddots & \vdots \\ \phi_{\sigma_1}(\overline{x}_n) & \cdots & \phi_{\sigma_\Omega}(\overline{x}_n) \end{pmatrix}$$

Given a set of samples $\{\overline{x}_1, \dots, \overline{x}_n\}$ and values $\{y_1, \dots, y_n\}$, the solution is obtain with previous tools. However, the number of these monomials grows rapidly as a function of the degree m. The $n \times |\Omega|$ matrix $\mathcal{C}_{\overline{x}}$ almost certainly has multiple solutions, a high degree of instability and the inverse lends itself to growing computational complexity.

Abstract feature space hypotheses

A tool often used in machine learning is the feature transform, let $\phi(\overline{x}) = \overline{u}$ be an arbitrary mapping $\phi: \mathbb{R}^d \to \mathbb{R}^z$ to a high dimensional space. Define the hypothesis space $\mathcal{H}_d^z$ as the set of linear hypothesis $\mathbb{R}^z \to \mathbb{R}$, which may correspond to a highly complicated nonlinear hypothesis $\mathbb{R}^d \to \mathbb{R}$:

$$\mathcal{H}_d^z = \left\{ \overline{w}^T \phi(\overline{x}) \mid \overline{w} \in \mathbb{R}^z \right\}$$

An obvious basis for this function space is the set of functions $\{\phi(\overline{x})_i\}_{i=1}^z$, consisting of each entry of the feature vector $\phi(\overline{x})_i$ with the coefficient space $(w_i)_{i=1}^z \in \mathbb{R}^z$. For samples $\{\overline{x}_1, \dots, \overline{x}_n\}$, the sampling matrix can be written as the feature vectors as rows:

$$\mathcal{C}_{\overline{x}} = \begin{pmatrix} \phi(\overline{x}_1)^T \\ \vdots \\ \phi(\overline{x}_n)^T \end{pmatrix}$$

Depending on the dimensionality of the feature transform $\phi: \mathbb{R}^d \to \mathbb{R}^z$, the $n \times z$ matrix $\mathcal{C}_{\overline{x}}$ may be very large which means the problem possibly has multiple solutions, instability and growing computational complexity.

2.5.2 ***Kernel from a feature space***

The problem with the feature transform is that the dimension of $\mathbb{R}^z$ may be large, thus explicitly performing this transform for each point and solving the linear system could make the approach computationally prohibitive. Deeper investigation into this setting reveals a trick one can utilize to overcome computational problems.

Let $\phi: \mathbb{R}^d \to \mathbb{R}^z$ be an arbitrary feature transform. Define $\mathcal{H}_d^z$ as the set of linear hypothesis in $\mathbb{R}^z \to \mathbb{R}$:

$$\mathcal{H}_d^z = \{\langle \overline{w}, \phi(\overline{x}) \rangle \mid \overline{w} \in \mathbb{R}^z\}$$

Where we have replaced the Euclidean inner product $\overline{w}^T \phi(\overline{x})$ by the general form. Suppose we have a set of basis functions for $f \in \mathcal{H}_d^z$:

$$f = c_1 \langle \phi(\overline{x}), \overline{w}_1 \rangle + \cdots + c_z \langle \phi(\overline{x}), \overline{w}_z \rangle$$

The approximation problem is to solve, denoting the sampling matrix by Φ^T:

$$\begin{pmatrix} \hat{f}(x_1) \\ \vdots \\ \hat{f}(x_n) \end{pmatrix} = \begin{pmatrix} \langle \phi(\overline{x}_1), \overline{w}_1 \rangle & \cdots & \langle \phi(\overline{x}_1), \overline{w}_z \rangle \\ \vdots & \ddots & \vdots \\ \langle \phi(\overline{x}_n), \overline{w}_1 \rangle & \cdots & \langle \phi(\overline{x}_n), \overline{w}_z \rangle \end{pmatrix} \begin{pmatrix} c_1 \\ \vdots \\ c_z \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_z \end{pmatrix}$$

In the previous setting we used the obvious choice $\overline{w}_i = \overline{e}_i$, but then Φ^T is an $n \times z$ matrix and the problem is to solve $\Phi^T \overline{c} = \overline{y}$. Instead, we make the claim that we can select $\overline{w}_i = \phi(\overline{x}_i)$. Then we have an $n \times n$ matrix G_ϕ and the problem becomes:

$$G_\phi \overline{c} = \overline{y}$$

Theorem (adapted from (Moon, 1999)):

Let $\overline{\phi}_1, \dots, \overline{\phi}_n \in \mathbb{R}^z$ be linearly independent and let $V = \text{span}\{\overline{\phi}_1, \dots, \overline{\phi}_n\}$. Then $\|\overline{w}\| \in \mathbb{R}^z$ with minimum norm satisfying:

$$\begin{gathered} \langle \overline{w}, \overline{\phi}_1 \rangle = y_1 \\ \vdots \\ \langle \overline{w}, \overline{\phi}_n \rangle = y_n \end{gathered}$$

belongs to V: $\overline{w} = \sum_{i=1}^n w_i \overline{\phi}_i$ with the coefficients $\overline{w}$ satisfying:

$$\begin{pmatrix} \langle \overline{\phi}_1, \overline{\phi}_1 \rangle & \cdots & \langle \overline{\phi}_n, \overline{\phi}_1 \rangle \\ \vdots & \ddots & \vdots \\ \langle \overline{\phi}_1, \overline{\phi}_n \rangle & \cdots & \langle \overline{\phi}_n, \overline{\phi}_n \rangle \end{pmatrix} \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

Proof:

Suppose $\overline{w}_0$ is a solution. Then so is any $\overline{w}_0 + \overline{n}$, where $\overline{n} \perp \overline{\phi}_1, \dots, \overline{\phi}_n$, i.e $\overline{n} \in V^{\perp}$. In fact any solution $\overline{w}$ can be written $\overline{w} = \overline{w}_0 + \overline{n}$ with $\overline{n} \in V^{\perp}$ ($\overline{w}_1$ also a solution $\Longrightarrow \Phi^T(\overline{w}_1 - \overline{w}_0) = 0 \Longrightarrow \overline{w}_1 - \overline{w}_0 \in V^{\perp}$)).

The problem is to find $\underset{\overline{n} \in V^{\perp}}{\operatorname{argmin}} \|\overline{w}_0 + \overline{n}\|$. This is equivalent to $\underset{\overline{n} \in V^{\perp}}{\operatorname{argmin}} \|\overline{n} - (-\overline{w}_0)\|$, finding the minimum norm approximation of $-\overline{w}_0$ in the closed subspace $V^{\perp}$. The projection theorem states that the error of optimal solution $\overline{n}^*$ is orthogonal to $V^{\perp}$: $\overline{n}^* - (-w_0) \in (V^{\perp})^{\perp} = V$. This proves that the optimal solution $\overline{w}^* = \overline{w}_0 + \overline{n}^* \in V$. Denote $\overline{w} = \sum_{i=1}^{n} w_i \overline{\phi}_i$ and take the inner products with $\overline{\phi}_1, \dots, \overline{\phi}_n$ to obtain the co-efficient formulation.

One implication of this is that instead of the potentially massive $n \times z$ system, we can solve $n \times n$ system G_{ϕ}. Furthermore, since the solution only depends on the dot products between feature transforms, even the feature transformation does not have to be performed if we have a **kernel** function k such that

$$k(\overline{x}_i, \overline{x}_j) = \langle \phi(\overline{x}_i), \phi(\overline{x}_j) \rangle$$

Utilizing $k(\overline{x}_i, \overline{x}_j)$ as a black box means that we have an algorithm which may have computationally better qualities but is equivalent to one with the feature transform. The underlying feature transform is implicit. The problem them is expressed through the **kernel matrix**:

$$\begin{pmatrix} k(\overline{x}_1, \overline{x}_1) & \cdots & k(\overline{x}_n, \overline{x}_1) \\ \vdots & \ddots & \vdots \\ k(\overline{x}_1, \overline{x}_n) & \cdots & k(\overline{x}_n, \overline{x}_n) \end{pmatrix} \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

With the solution

$$f(\overline{x}) = w_1 k(\overline{x}, \overline{x}_1) + \cdots + w_n k(\overline{x}, \overline{x}_n)$$

2.5.3 ***Kernel from a function space***

Suppose we have a Hilbert space $\mathcal{H}$ of functions $f: X \to \mathbb{R}$ with a total orthonormal set $\{\phi_i\}_{i\in I_2}$ forming the analysis space, with the evaluation functional $\delta_x: f \mapsto f(x)$ continuous. Then for every $f \in \mathcal{H}$ there exists a representation $f = \sum_{i\in I_2} a_i\phi_i$ and a corresponding sequence $(a_i)_{i\in I_2} \in \ell^2(I_2)$.

Denoting the space of functionals $\mathcal{H}'$, by the Riesz representer theorem (Kreyszig, 1989, s. 188), for every bounded functional $\mathcal{g} \in \mathcal{H}'$ there exists a representer $g \in \mathcal{H}$ such that: $\mathcal{g}(f) = \langle f, g\rangle_{\mathcal{H}}\ \forall f \in \mathcal{H}$. Denote this map $\mathcal{R}: \mathcal{H}' \to \mathcal{H}$. Since evaluation functional was assumed to be continuous, for every $x \in X$ there exists a function in $\mathcal{H}$ representing evaluation at x given by $\mathcal{R}: \delta_x \mapsto \mathcal{R}\delta_x$. In fact, it can be explicitly written:

$$\mathcal{R}\delta_x = \sum_{i\in I_2} \langle \mathcal{R}\delta_x, \phi_i\rangle_{\mathcal{H}}\phi_i = \sum_{i\in I_2} \phi_i(x)\,\phi_i$$

Futhermore:

$$\mathcal{R}\delta_x(x) = \sum_{i\in I_2} \phi_i(x)^2 \Longrightarrow \big(\phi_i(x)\big) \in \ell^2(I_2)$$

Lemma: $\mathrm{span}\{\mathcal{R}\delta_x\}_{x\in X}$ is dense in $\mathcal{H}$.

Proof: Suppose not. This is equivalent to $\mathrm{span}\left\{\big(\phi_i(x)\big)_{i\in I_2}\right\}_{x\in X}$ is not dense in $\ell^2(I_2)$. Then exists $\overline{a} \neq 0 \in \mathrm{span}\big\{\big(\phi_i(x)\big)\big\}_{x\in X}^{\perp}$ implying $\sum_{i\in I} a\phi_i(x) = 0\ \forall x \in X$, contradicting that $\{\phi_i\}_{i\in I_2}$ is a basis.

Therefore any $f \in \mathcal{H}$ belongs to the closure of $\{\mathcal{R}\delta_x\}_{x\in X}$, and can be written in terms of the representers of the evaluation functionals. However, consider the following which motivates the mapping.

Claim: Given data $\big((x_1, y_1), \dots, (x_m, y_m)\big)$, the best approximation minimum norm solution belongs to $H_{\overline{x}} := \mathrm{span}\big\{\mathcal{R}\delta_{x_1}, \dots, \mathcal{R}\delta_{x_m}\big\}$.

Proof: Let $H = H_{\overline{x}} \oplus H_{\overline{x}}^{\perp}$ and for $f \in H$ factorize accordingly $f = f_{\overline{x}} + f_{\overline{x}}^{\perp}$. Then

$$f(x_i) = \left\langle f, \mathcal{R}\delta_{x_i}\right\rangle_H = \left\langle f_{\overline{x}} + f_{\overline{x}}^{\perp}, \mathcal{R}\delta_{x_i}\right\rangle_H = \left\langle f_{\overline{x}}, \mathcal{R}\delta_{x_i}\right\rangle_H = f_{\overline{x}}(x_i)$$

$$\|f\|_H = \left\|f_{\overline{x}} + f_{\overline{x}}^{\perp}\right\|_H = \|f_{\overline{x}}\|_H + \left\|f_{\overline{x}}^{\perp}\right\|_H \Longrightarrow \|f_{\overline{x}}\|_H \leq \|f\|_H$$

This result applied to kernels is known as the representer theorem (Mohri;Rostamizadeh;& Ameet, 2012), because it is both norm penalty and dataset loss invariant.

Because $f_{\overline{x}}$ evaluates to f at $\overline{x} = \{x_1, \dots, x_n\}$ and has smaller or equal norm, for minimum or penalized norm approximation problems at these points we only need to consider $f_{\overline{x}} \in H_{\overline{x}}$. A basis of $H_{\overline{x}}$ or linear combination $f_{\overline{x}} = \sum_{i=1}^{n} a_i \mathcal{R}\delta_{x_i}$ is sufficient for approximation. Because $\{\mathcal{R}\delta_{x_1}, \dots, \mathcal{R}\delta_{x_n}\}$ span the subspace by definition, but are not necessarily linearly independent, it is possible that multiple coefficient vectors $(a_i)_{i=1}^{n} \leftrightarrow \sum_{i=1}^{n} a_i \mathcal{R}\delta_{x_i}$ denote the same function. The Gram matrix $(G_{\overline{a}})_{i,j} = \left\langle \mathcal{R}\delta_{x_i}, \mathcal{R}\delta_{x_j} \right\rangle_H$ is then positive semidefinite and multiple vectors in which denote the same solution in $H_{\overline{x}}$ minimize the norm.

The problem is to find a minimum norm ($\|\overline{a}\| = \overline{a}^T G_{\overline{a}} \overline{a}$) solution to:

$$\begin{pmatrix} \mathcal{R}\delta_{x_1}(x_1) & \cdots & \mathcal{R}\delta_{x_n}(x_1) \\ \vdots & \ddots & \vdots \\ \mathcal{R}\delta_{x_1}(x_n) & \cdots & \mathcal{R}\delta_{x_n}(x_n) \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

Having obtained $\overline{a}$ the solution is $f(x) = \sum_{i=1}^{n} a_i \mathcal{R}\delta_{x_i}(x)$.

Now assume we define the following symmetric positive definite function, also called a **kernel**:

$$k(x, y) = \sum_{i=1}^{n} \phi_i(x)\phi_i(y)$$

Rewrite the previous subspace problem using this newly defined function using the fact that $k(x_i, x_j) = \mathcal{R}\delta_{x_i}(x_j) = \left\langle \mathcal{R}\delta_{x_i}, \mathcal{R}\delta_{x_j} \right\rangle_H$. Define the matrix $K_{i,j} = k(x_i, x_j)$. The problem is to find a minimum norm ($\|\overline{a}\| = \overline{a}^T K \overline{a}$) or solution to:

$$\begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

Having obtained $(a_i)_{i=1}^{n}$ solution is $f(x) = \sum_{i=1}^{n} a_i k(x, x_i)$.

Example

Define the function $k(x, y) = (1 + x \cdot y)^d$. Expanding, it can be written:

$$k(x,y) = 1 \cdot 1 + \sqrt{\binom{d}{1}}x \cdot \sqrt{\binom{d}{1}}y + \sqrt{\binom{d}{2}}x^2 \cdot \sqrt{\binom{d}{2}}y^2 + \cdots + x^d \cdot y^d$$

Denote $\phi_0(x) = 1, \phi_1(x) = \sqrt{\binom{d}{1}}x,\ \phi_2(x) = \sqrt{\binom{d}{2}}x^2,\ \ldots,\ \phi_d(x) = x^d$, which are clearly linearly independent. Define a Hilbert space $\mathcal{H} = \text{span}\{\phi_0, \phi_1, \phi_2, \ldots, \phi_d\}$ with the norm $\|\sum_{i=0}^n a_i\phi_i\|_X^2 = \sum_{i=0}^n a_i^2$ so that $\{\phi_i\}_{i=0}^d$ is an orthonormal basis. Then the minimum norm best approximation problem can be solved in two different ways. The regularized solutions may be obtained, for example, minimizing the Tikhonov functional as before.

Using the $n \times d$ sampling matrix: Take minimum $\|\overline{a}\|_X = \overline{a}^T\overline{a}$ norm closest approximation

$$\begin{pmatrix} \phi_0(x_1) & \cdots & \phi_d(x_1) \\ \vdots & \ddots & \vdots \\ \phi_0(x_n) & \cdots & \phi_d(x_n) \end{pmatrix}\begin{pmatrix} a_0 \\ \vdots \\ a_d \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

The solution is $f(x) = \sum_{i=0}^d a_i\phi_i(x)$.

Using the $n \times n$ kernel matrix. Take minimum $\|\overline{\alpha}\|_K = \overline{\alpha}^T K\overline{\alpha}$ norm closest approximation

$$\begin{pmatrix} k(x_1,x_1) & \cdots & k(x_1,x_n) \\ \vdots & \ddots & \vdots \\ k(x_n,x_1) & \cdots & k(x_n,x_n) \end{pmatrix}\begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

The solution is $f(x) = \sum_{i=\text{i}}^n \alpha_i k(x, x_i)$.

In this case the kernel is fast to compute $k(x,y) = (1 + x \cdot y)^d$ yet in most of the practical scenarios $n > d$. However, generalizing this kernel to the polynomial case $k\colon \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$:

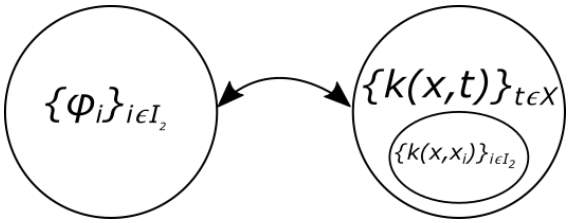
$$k(\overline{x}, \overline{y}) = (1 + \overline{x} \cdot \overline{y})^d$$

The kernel is still fast to compute, but we now have $\binom{m+d}{d}$ basis functions, rapidly increasing the computational complexity of the ordinary solution. However, if $k(\overline{x}, \overline{y})$ is computed in constant time c, the kernel matrix may still be computed in $O(cn^2)$ time and the solution then obtained in $O(n^3)$ time.

However, one should note that aside from defining the implicit possibly complicated function space, a kernel defines a norm as well for this space. For example, compared to the traditional setting of using $\{1, x, x^2, \ldots, x^d\}$ as a basis of $\mathcal{H}$ with the norm

$\|a_0 + a_1x + a_2x^2 + \cdots + a_dx^d\| = \sum_{i=0}^{d} a_i^2$, this kernel weighted the norm by $\|a_0 + a_1x + +a_2x^2 + \cdots + a_dx^d\|_{\mathcal{H}} = a_0^2 + \sqrt{\binom{d}{1}}a_1^2 + \sqrt{\binom{d}{2}}a_2^2 + \cdots + a_d^2$.

2.5.4 ***Reverse direction***



We proved that if we had a Hilbert space H of functions $f: X \to \mathbb{R}$ with orthonormal basis $(\phi_i)_{i\in I}$, we can create an equivalent approximation problem by defining a symmetric positive definite function $k(x,y) = \sum_{i\in I} \phi_i(x)\phi_i(y)$ which spans the space $H_K = \text{span}\{k(x,t)\}_{t\in X}$ equipped with the norm $\left\langle \sum_{i=0}^{d} a_i k(x,t_i)\,, \sum_{i=0}^{d} b_i k\left(x,t_j\right)\right\rangle_{H_K} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i b_j k\left(t_i,t_j\right)$. Furthermore, given a sample $\{x_1, \dots, x_n\}$ this space simplifies seeking the minimum norm solution into $H_{\overline{x}} = \text{span}\{k(x,t)\}_{t\in\{x_1,\dots,x_n\}}$.

Suppose we are given an arbitrary kernel, a positive-definite symmetric function $k(x,y)$. Which space and orthonormal basis does substituting it in the approximation problem, thus using the closure of the space $H_K = \text{span}\{k(x,t)\}_{t\in X}$ with $\langle , \rangle_{H_K}$, to obtain a solution correspond to? If we know a decomposition $k(x,y) = \sum_{i\in I} \phi_i(x)\phi_i(y)$ such that $\{\phi_i(x)\}_{i\in I}$ is a basis of H_K, is unknown, we may either

- Construct the closure and use H_K as the analysis space.
- Apply Gram-Schmidt in H_K to obtain a sequence of orthonormal vectors.
- Apply Mercer's theorem to obtain a decomposition with this property.

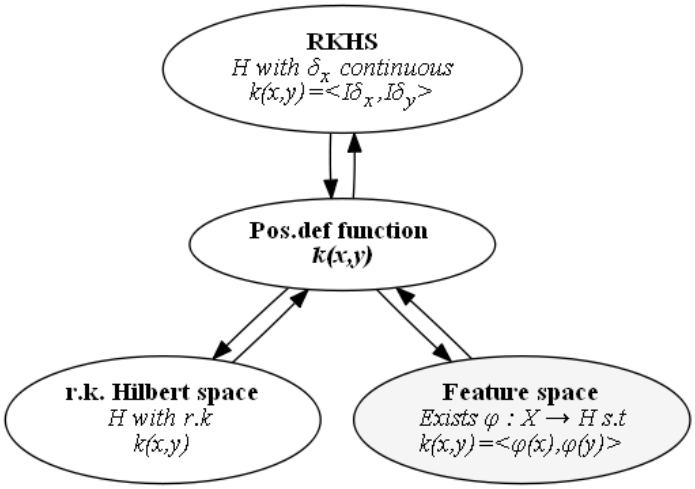
This means using a given kernel $k(x,y)$ to solve an approximation problem corresponds to implicitly using a possibly complicated, even infinite, function space. An orthonormal basis for the space can be obtained from Mercer's theorem, for example. This is advantageous compared to solving using the full basis for the function space and the sampling operator if the kernel is easy to compute and data dimensionality is high compared to the amount of data. In applications, kernels are chosen to utilize function spaces, or alternatively feature spaces, with well-known properties applicable to the task at hand and the space is taken to be implicit. The next chapter studies this based on an arbitrary kernel $k(x,y)$.

3 Reproducing Kernel Hilbert Space

3.1 Spaces associated to a kernel

To each kernel k, there corresponds an unique space of functions $\mathcal{H}$ for which k is a reproducing kernel. Often the existence of this space is taken to be implicit, and the algorithms are expressed solely in terms of the kernel function. The solution then consists of linear combination of functions obtained by fixing one argument of the kernel: $k(x, x_i): X \to \mathbb{R}$. Yet, the space of functions still exists behind the scenes. An explicit formulation is particularly useful in our case since this allows the learning problem to be formulated as a linear operator from the kernel Hilbert space to the sample Hilbert space. Another characterization often used in machine learning is given by providing an explicit feature mapping ϕ to a feature space $\mathcal{H}_0$ and the kernel as the inner product in this feature space. This space may not consist of functions, but there is an unique corresponding space of functions $\mathcal{H}$ which the learning problem in $\mathcal{H}_0$ corresponds to.

In the literature, there are several different characterizations of the space of functions $\mathcal{H}$ associated to a kernel k. These characterizations are all equivalent, and are closely related to the feature space formulation. The space $\mathcal{H}$ has several interesting properties, some following most naturally from a given characterization. These properties are used in theoretical and practical considerations, and are therefore worth illuminating. In the following we demonstrate the equivalence of 3 characterizations:



In the diagram, arrows denote implications we are going to prove. As can be seen, having shown these implications, we have equivalence between the concepts. Graying of the feature space denotes a caveat; the feature space is not unique as shall be seen. One can think of the node as 'all the feature spaces' with the inner product equivalent to k.

(Aronszajn, Theory of Reproducing Kernels, 1948) originated much of the theory of reproducing kernels, and is the most often referred resource. An introduction to reproducing kernels with emphasis on machine learning is (Steinwart & Christmann, 2008), to which this section is mostly based on.

Definitions

In the following, let $\mathcal{H}$ be a Hilbert space of functions $f: X \to \mathbb{R}$. Denote the inner product $\langle f, g \rangle_{\mathcal{H}}$ and the corresponding norm $\|f\|_{\mathcal{H}}$.

For a fixed $x \in X$, the map $\delta_x: \mathcal{H} \to \mathbb{R}$ defined by $\delta_x: f \mapsto f(x)$ is called the evaluation functional at x.

Definition (RKHS)

$\mathcal{H}$ is a Reproducing Kernel Hilbert Space (RKHS) if δ_x is continuous $\forall x \in X$.

Definition (reproducing kernel)

A function $k: X \times X \to \mathbb{R}$ is called a reproducing kernel of $\mathcal{H}$ if it satisfies:

$$\forall x \in X, \qquad k(.,x) \in \mathcal{H}$$

$$\forall x \in X, \forall f \in \mathcal{H}, \qquad \langle f, k(.,x) \rangle_{\mathcal{H}} = f(x)$$

Definition (Feature space)

The function $k: X \times X \to \mathbb{R}$ is a kernel if there exists a real Hilbert space $\mathcal{F}$ and a map $\phi: X \to \mathcal{F}$ such that

$$\forall x, y \in X, k(x,y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{F}}$$

The map ϕ is called a feature map and $\mathcal{F}$ a feature space. Note that in the definition, we do not require $\mathcal{F}$ to be a space of functions; hence a different symbol.

Definition (positive definite function)

A symmetric function $h: X \times X \to \mathbb{R}$ is positive definite if

$$\forall n \geq 1, \overline{a} \in \mathbb{R}^n, \forall \overline{x} \in X^n, \sum_{i=1}^{n} \sum_{j=1}^{n} a_i, a_j h(x_i, x_j) \geq 0$$

Equivalently, we may require the matrix $H_{ij} \coloneqq h(x_i, x_j)$ to be positive semidefinite for all $n \in \mathbb{N}, x_1, \dots x_n$.

3.2 Equivalence of spaces

Theorem

Given a symmetric positive definite function $k: X \times X \to \mathbb{R}$, define the following space of functions (Steinwart & Christmann, 2008):

$$\mathcal{H}_{pre} \coloneqq \left\{ \sum_{i=1}^{n} \alpha_i k(., x_i) \ : \alpha_1, \dots, \alpha_n \in \mathbb{R}, x_1, \dots, x_n \in X \right\}$$

Given $f \coloneqq \sum_{i=1}^{n} \alpha_i k(., x_i)$ and $g \coloneqq \sum_{j=1}^{m} \beta_j k(., x_j')$, define:

$$\langle f, g \rangle \coloneqq \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(x_i, x_j')$$

Then $\mathcal{H}_{pre}$ is an inner product space. Denote $\mathcal{H}$ the completion, which is then a Hilbert space with a reproducing kernel k.

Futhermore the reproducing kernel Hilbert space is unique. If k is the reproducing kernel of $\mathcal{F}$, then an isometric isomorphism exists between $\mathcal{F}$ and $\mathcal{H}$.

Proof

$\langle ., . \rangle$ is a well-defined inner product on $\mathcal{H}_{pre}$:

- Well-defined: $\langle f, g \rangle = \sum_{j=1}^{m} \beta_j f(x_j') = \sum_{i=1}^{n} \alpha_i g(x_i)$, implying independent of representation of f and g.
- Bilinear, symmetric and positive: follows directly with k assumed symmetric positive definite.
- Separates points: Schwarz inequality implies $|\langle f, g \rangle| = \langle f, f \rangle \langle g, g \rangle \ \forall f, g \in \mathcal{H}_{pre}$. Then $\|f\| = 0$ implies $|f(x)| = |\sum_{i=1}^{n} \alpha_i k(x, x_i)|^2 = |\langle f, k(., x) \rangle|^2 \leq \langle f, f \rangle \langle k(., x), k(., x) \rangle = \|f\|^2 k(x, x) = 0$.

Then k is the reproducing kernel of $\mathcal{H}_{pre}$:

- $k(., x) \in \mathcal{H}_{pre} \ \forall x \in X$

- $\langle f, k(.,x)\rangle = f(x)\ \forall f \in \mathcal{H}_{pre}$

$\mathcal{H}_{pre}$ is a well-defined inner product space with the reproducing kernel k, but it may not be a Hilbert space, i.e. complete.

Completion is defined as an extension of a space such that the new space is complete and contains a dense subspace isometric with the original space. There exists a standard completion of any inner product space $\overline{\mathcal{H}_{pre}} = \mathcal{H}$ (Kreyszig, 1989, ss. 41,69,139). Given a space $\mathcal{H}_{pre}$ with inner product $\langle .,. \rangle$ and induced norm $\|.\|$, completion is defined:

$$\mathcal{H} = \{[(f_n)] : (f_n)\ is\ Cauchy\ in\ \mathcal{H}_{pre}\}$$

Where $[(f_n)]$ is the equivalence class of (f_n) defined relative to the equivalence $(f_n) \sim (g_n)$ if

$$\lim_{n\to\infty} \|f_n - g_n\| = 0$$

To make the space a vector space, addition and scalar multiplication are defined:

$$[(f_n)] + [(g_n)] = [(f_n + g_n)]$$
$$\alpha[(f_n)] = [(\alpha f_n)]$$

For any representatives $(f_n) \in [(f_n)]$ and $(g_n) \in [(g_n)]$ an inner product on $\mathcal{H}$ is defined:

$$\langle [(f_n)], [(g_n)] \rangle_H = \lim_{n\to\infty} \langle f_n, g_n \rangle$$

It can be shown that this construction satisfies the properties required from completion. The inner product and vector space operations are well defined and indepedent of the representers. The mapping $T: \mathcal{H}_{pre} \to \mathcal{H}$ defined by $f \mapsto [(f)]$ is an isometric mapping onto a dense subspace of $\mathcal{H}$. Futhermore, $\mathcal{H}$ is complete and unique except for isometries.

However, the space $\mathcal{H}$ does not consist of functions. When a space of functions is completed such that the completion consists of functions as well and preservers certain continuity properties, we speak of **functional completion** (Aronszajn, Theory of Reproducing Kernels, 1948). For each Cauchy sequence (f_n) in $\mathcal{H}_{pre}$:

$$|f_m(x) - f_n(x)| = |\langle f_m, k(.,x)\rangle - \langle f_n, k(.,x)\rangle| = |\langle f_m - f_n, k(.,x)\rangle|$$
$$\leq \|f_m - f_n\| k(x,x)^{\frac{1}{2}}$$

Since $\left(f_n(x)\right)$ is Cauchy in $\mathbb{R}$ which is complete, the sequence converges. Substituting two equivalent sequences in the inequality we see $(f_n)\sim(g_n) \Longrightarrow |f_n(x) - g_n(x)| \to 0$. Therefore it it possible to associate to each equivalence class $[(f_n)]$ the function:

$$f(x) = \lim_{n\to\infty} f_n(x)$$

Consider the sequence $\left(k(.,x)\right)$ associated to $k(.,x)$. Existence of the reproducing kernel is easily verified:

$$f(x) = \lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} \langle f_n, k(.,x)\rangle = \left\langle [(f_n)]\,, \left[\left(k(.,x)\right)\right] \right\rangle_{\mathcal{F}}$$

Given an arbitrary $\mathcal{F}$ with a reproducing kernel k, first property $k(.,x) \in \mathcal{F}\ \forall x$ implies $\mathcal{H}_{pre} \subset \mathcal{F}$. The reproducing property implies $\forall f \in \left(\mathcal{H}_{pre}\right)^{\perp}: f(x) = \langle f, k(.,x)\rangle = 0 \Longrightarrow \left(\mathcal{H}_{pre}\right)^{\perp} = \{0\}$ which implies $\mathcal{H}_{pre}$ is dense in $\mathcal{F}$. Because the completion is unique up to isometric isomorphism (Kreyszig, 1989), we have $\mathcal{H} = \mathcal{F}$.

Finally, every space has a unique reproducing kernel, since if k_1 and k_2 are reproducing kernels of $\mathcal{H}$:

$$\begin{aligned}\|k_1(.,x) - k_2(.,x)\|^2 &= \langle k_1(.,x) - k_2(.,x), k_1(.,x) - k_2(.,x)\rangle \\ &= k_1(x,x) - k_2(x,x) - \left(k_1(x,x) - k_2(x,x)\right) = 0\end{aligned}$$

It follows $\forall x:\ k_1(y,x) = k_2(y,x) \Longrightarrow k_1 = k_2$.

This shows that to every symmetric positive definite function $k(x,y)$ corresponds a unique Hilbert space of functions for which k is a reproducing kernel. For a space with reproducing kernel k, the kernel is unique and symmetric positive semidefinite.

Theorem

$\mathcal{H}$ is an RKHS if and only if it has a reproducing kernel. Futhermore, the reproducing kernel is given by (Steinwart & Christmann, 2008):

$$k(x,y) = \langle \mathcal{R}\delta_Y, \mathcal{R}\delta_x\rangle$$

Where $\mathcal{R}: \mathcal{H}' \to \mathcal{H}$ maps a bounded functional to the representer in $\mathcal{H}$, given by Riesz's theorem (Kreyszig, 1989, s. 188).

Proof

k is the reproducing kernel of $\mathcal{H}$:

- $k(x,y) = \langle \mathcal{R}\delta_y, \mathcal{R}\delta_x \rangle = \delta_x(\mathcal{R}\delta_y) = \mathcal{R}\delta_y(x) \Longrightarrow k(.,y) = \mathcal{R}\delta_y \in \mathcal{H}\ \forall y$
- $f(x) = \delta_x(f) = \langle f, \mathcal{R}\delta_x \rangle = \langle f, k(.,x) \rangle$

If $\mathcal{H}$ has a reproducing kernel k, δ_x is continous:

$$|\delta_x(f)| = |f(x)| = |\langle f, k(.,x) \rangle| \leq \|f\| k(x,x)^{\frac{1}{2}}$$

RKHS formulation is often used to highlight an important convergence property (Aronszajn, Theory of Reproducing Kernels, 1948). Let (f_n) be a sequence converging to f: $\|f_n - f\| \to 0$ as $n \to \infty$. Then continuity of δ_x means norm convergence implies pointwise convergence:

$$\lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} \delta_x(f_n) = \delta_x(f) = f(x)$$

Theorem

$k: X \times X \to \mathbb{R}$ is a symmetric positive definite function if and only if it is a kernel.

Proof

Let $\mathcal{H}$ be a Hilbert space for which the symmetric positive definite function k is a reproducing kernel. Define the feature mapping $\Phi: X \to \mathcal{H}$, also called the *canonical feature map* (Steinwart & Christmann, 2008):

$$\Phi(x) = k(.,x)$$

Then k realizes the inner product on $\mathcal{H}$:

$$\langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}} = \langle \Phi(y), \Phi(x) \rangle_{\mathcal{H}} = \langle k(.,y), k(.,x) \rangle_{\mathcal{H}} = k(x,y)$$

Suppose $k(x,y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}}$. Then k is symmetric positive definite:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) = \left\langle \sum_{i=1}^{n} \alpha_i \Phi(x_i), \sum_{i=1}^{n} \alpha_i \Phi(x_j) \right\rangle_{\mathcal{H}} \geq 0$$

Note that the theorem implies existence of a feature map, not uniqueness. There are several feature spaces for which kernel realizes an inner product. For a trivial example, define the feature maps $\phi_1: x \mapsto x \in \mathbb{R}$ and $\phi_2: x \mapsto \left(\frac{x}{\sqrt{2}}, \frac{x}{\sqrt{2}}\right) \in \mathbb{R}^2$. Then

$k(x,y) = xy = \langle \phi_1(x), \phi_1(y) \rangle_{\mathbb{R}} = \langle \phi_2(x), \phi_2(y) \rangle_{\mathbb{R}^2}$. However as seen above, the function space for which k is a reproducing kernel is unique, there exists a feature map which realizes it as a feature space. Futhermore, the following fact implies that there exists a metric surjection from any feature space into it, which means it can be seen as the *smallest* feature space in this sense.

Theorem (Steinwart & Christmann, 2008)

Let $k: X \times X \to \mathbb{R}$ be a kernel with feature space $\mathcal{H}_0$ and feature map $\Phi_0: X \to \mathcal{H}_0$. Then the RKHS may also be expressed through the arbitrary feature space:

$$\mathcal{H} = \{f: X \to \mathbb{R} : \exists w \in \mathcal{H}_0 \; f(x) = \langle w, \Phi(x) \rangle_{\mathcal{H}_0}\}$$

equipped with the norm:

$$\|f\|_{\mathcal{H}} \coloneqq \inf\{\|w\|_{\mathcal{H}_0} : w \in \mathcal{H}_0 \;\; f(x) = \langle w, \Phi(x) \rangle_{\mathcal{H}_0}\}$$

Moreover, a metric surjection $V: \mathcal{H}_0 \to \mathcal{H}$ may be defined by

$$(Vw)(x) = \langle w, \Phi(x) \rangle_{\mathcal{H}_0} \; \forall w \in \mathcal{H}_0$$

This theorem is particularly useful for considering the effect on the RKHS of modifications of symmetric positive definite functions by compositions through operations on the feature space, compared to the complicated proofs in (Aronszajn, Theory of Reproducing Kernels, 1948).

3.3 Kernel and space modifications

Consider kernels K_i with RKHSs $\mathcal{H}_i$. If we have a composition K^* consisting of K_i which can be readily verified to be positive definite, we know that there exists an RKHS $\mathcal{H}^*$ associated to it. Likewise, if an operation such as restriction modifies kernel K with RKHS $\mathcal{H}$ such that K remains symmetric positive semidefinite there is a RKHS $O(\mathcal{H})$ associated to the modified kernel $O(\mathcal{H})$. Can these RKHS be expressed in terms of the original $\mathcal{H}$? (Aronszajn, 1948) systematically developed the following results, with the proofs expressed much more simply using the theorem in the previous section:

Sum of Kernels

If K_1, K_2 are reproducing kernels of $\mathcal{H}_1, \mathcal{H}_2$, then $K(x,y) = K_1(x,y) + K_2(x,y)$ is the reproducing kernel of $\mathcal{H} = \{f_1 + f_2 \mid f_1 \in \mathcal{H}_1, f_2 \in \mathcal{H}_2\}$ with the norm defined as $\|f\|_{\mathcal{H}}^2 = \min_{f_1 \in \mathcal{H}_1, f_2 \in \mathcal{H}_2} \|f_1\|_{\mathcal{H}_1}^2 + \|f_2\|_{\mathcal{H}_2}^2$.

Proof

Let $K_1(x,y) = \langle \phi_1(x), \phi_1(y) \rangle_{\mathcal{H}_1}, K_2(x,y) = \langle \phi_2(x), \phi_2(y) \rangle_{\mathcal{H}_2}$. Define the map $\phi(x) = (\phi_1(x) \quad \phi_2(x))$. Then $\phi: X \to \mathcal{H}_1 \oplus \mathcal{H}_2$ is a feature map with kernel $K(x,y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}} = \langle \phi_1(x), \phi_1(y) \rangle_{\mathcal{H}_1} + \langle \phi_2(x), \phi_2(y) \rangle_{\mathcal{H}_2} = K_1(x,y) + K_2(x,y)$. Since $\langle w, \phi(y) \rangle_{\mathcal{H}} = \langle w_1, \phi_1(y) \rangle_{\mathcal{H}_1} + \langle w_2, \phi_2(y) \rangle_{\mathcal{H}_2}$ substituting this to the theorem gives the RKHS result.

Product of Kernels

If K_1, K_2 are reproducing kernels of $\mathcal{H}_1, \mathcal{H}_2$, then $K(x_1, x_2, y_1, y_2) = K_1(x_1, y_1) K_2(x_2, y_2)$ is the reproducing kernel of $\mathcal{H} = \mathcal{H}_1 \otimes \mathcal{H}_2$.

Proof

Let $K_1(x,y) = \langle \phi_1(x), \phi_1(y) \rangle_{\mathcal{H}_1}, K_2(x,y) = \langle \phi_2(x), \phi_2(y) \rangle_{\mathcal{H}_2}$. Define the map $\phi(x) = \phi_1(x) \otimes \phi_2(x)$. Then $\phi: X \to \mathcal{H}_1 \otimes \mathcal{H}_2$ is a feature map with kernel $K(x,y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}} = \langle \phi_1(x) \otimes \phi_2(x), \phi_1(y) \otimes \phi_2(y) \rangle_{\mathcal{H}} = \langle \phi_1(x), \phi_1(y) \rangle_{\mathcal{H}_1} \langle \phi_2(x), \phi_2(y) \rangle_{\mathcal{H}_2} = K_1(x,y) K_2(x,y)$. Substituting $\langle w, \phi(y) \rangle_{\mathcal{H}} = \langle w_1, \phi_1(y) \rangle_{\mathcal{H}_1} \langle w_2, \phi_2(y) \rangle_{\mathcal{H}_2}$ to the theorem gives the RKHS result.

Taylor series of a Kernel (Steinwart & Christmann, 2008)

Denote $\dot{B}_{\mathbb{R}}$ and $\dot{B}_{\mathbb{R}^d}$ the unit balls in $\mathbb{R}$ and $\mathbb{R}^d$ and let $r \in [0, \infty)$. Assume f is analytic with Taylor series $f(x) = \sum_{i=0}^{\infty} a_i z^i$ where $z \in r\dot{B}_{\mathbb{R}}$ and $a_i \geq 0$. Then

$$k(x,y) = f(\langle x, y \rangle) = \sum_{i=0}^{\infty} a_i \langle x, y \rangle^i \text{ for } x, y \in \sqrt{r}\dot{B}_{\mathbb{R}^d}$$

Is a kernel, called **taylor type** kernel.

Restriction of Kernels

If K is the reproducing kernel of $\mathcal{H}$ of functions $f: X \to \mathbb{R}$, then $K_{|X'}$ defined K restricted to $X' \subset X$ is the reproducing kernel of $\mathcal{H}_0$ of functions $f_{|X'}$ where $f \in \mathcal{H}$ with the norm $\|f\|_{\mathcal{H}_0}^2 = \min_{g \in \mathcal{H}: g_{|X'} = f} \|g\|_{\mathcal{H}}^2$.

Proof

Let $K(x,y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$. Define the map $\phi_0(x) = \phi(Px)$ where $P: X' \to X$ is inclusion. Then $\phi_0: X' \to \mathcal{H}$ is a feature map with kernel $K_{|X'}(x,y) = \langle \phi_0(x), \phi_0(y) \rangle_{\mathcal{H}_0} = \langle \phi(Px), \phi(Py) \rangle_{\mathcal{H}} = K(Px, Py)$. Substituting $\langle w_0, \phi_0(y) \rangle_{\mathcal{H}_0} = \langle w, \phi(Py) \rangle_{\mathcal{H}}$ to the theorem gives the RKHS result.

These simple results can be used to construct many commonly used kernels on $\mathbb{R}^d$, for example the following are often used:

- Linear kernel: $k(x,y) = \langle x, y \rangle$
- Polynomial kernel: $k(x,y) = (\langle x, y \rangle + c)^m$ where $c \geq 0$.
- Exponential kernel: $k(x,y) = \exp(2\sigma x \cdot y)$
- Gaussian RBF kernel: $k(x,y) = \exp(-\sigma \|x - x'\|^2)$

3.4 Prior knowledge in Kernels

Multiple methods for incorporating prior knowledge into the kernel learning process have been considered in the literature. Prior knowledge is classified into two main categories in (Lauer & Bloch, 2007): invariance in the input space and knowledge on the data. By invariance they refer to the constraint that the output should not change under transformations or permutations of samples; for example transition or rotation should not change the letter in digit recognition. They define knowledge on the data as manipulating the data set: one may incorporate additional knowledge on unlabelled data such as training set imbalance or add quality relative weighting on the labelled sample.

Formally, invariance in the input space can be expressed as a constraint in terms of a parametrized transformation $T_\theta: x \mapsto T_\theta x$ of the input pattern (Lauer & Bloch, 2007):

$$f(x) = f(T_\theta x) \; \forall x, \theta$$

Often only local invariance is enforced through the tangent at $\theta = 0$:

$$\frac{\delta}{\delta \theta} |_{\theta=0} f(T_\theta x)$$

Some methods constrain the domain directly without specifying a transformation:

$$f(x) = y_{\mathcal{P}} \; \forall x \in \mathcal{P} \text{ where } \mathcal{P} \subseteq \text{X is constrained}$$

These constraints may be enforced in different places of the learning setting. (Schölkopf & Smola, 2001) classify the methods into three types based on the component they affect:

1. Sample methods modify data by generating or modifying samples
2. Kernel methods modify the kernel function
3. Optimization methods modify the minimization process or problem formulation

A popular sample method is virtual examples (Decoste & Bernhard, 2002). This method simply generates new samples from the training data which satisfy the invariance and have the same label. For a singular non-parametric transformation new example is are added for each vector in the training set $(x_i, y_i) \mapsto (Tx_i, y_i),\ i = 1 \dots N$. If the transformation is continuous, the transformation may be discretized and samples added for $T_\theta x_i, y_i, \theta = 1, \dots M$. This method is used in neural networks as hints and in SVMs as virtual SVs. For enforcing symmetry in the training data it is particularly simple, because we simply need to add

$$\big((v, v'), y_i\big) \mapsto ((v', v), y_i)$$

It can be shown that for permutation invariant kernels and non-zero regularization parameter it results in enforcing symmetry in the solution as well.

The method used in our study is to modify the kernel to incorporate prior knowledge. One example of such knowledge is invariance of kernel k under given transition T:

$$k(x, z) = k(Tx, z)$$

The review (Lauer & Bloch, 2007) mentions at least four methods which utilize the same technique: jittering kernels, tangent distance (TD), Haar-integration (HI) and Kernels between sets. In our case, the transition is not arbitrary like in these methods and the simplicity allows additional considerations. Because these methods are useful illustrations for their generality, they are briefly reviewed.

3.5 Examples of prior knowledge in Kernels

Jittering kernels (Decoste & Bernhard, 2002)

The jittered variations (translated, rotated, etc.) of samples are incorporated to the kernel itself. The new kernel $k^J(x_i, x_j)$ is defined in terms of the old kernel as follows

- Select a jittered variation $x_q \in \mathcal{L}(x_i)$ of a sample x_i closest to x_j by minimizing the distance $\|x_q - x_j\|_{\mathcal{H}} = \sqrt{k_{qq} - 2k_{qj} + k_{jj}}$ in the feature space $\mathcal{H}$.
- Define a new kernel by $k_{ij}^J = k_{qj}$

Tangent distance (TD) (Haasdonk & Keysers, 2002)

Let P_x be the set of all samples generated by transformation T from a sample x. For sample x_1 and x_2 tangent distance uses approximation techniques to define the kernel through the distance between P_{x_1} and P_{x_2}. Often linear approximation of the transformation T is used, which may be extended by using the first terms in the Taylor expansion of T. Express Tx as the combination of n_L local and simple parametrized linear transformations L_{α_k}, $\alpha_k \in \left[\alpha_k^{\min}, \alpha_k^{\max}\right]$ based on tangent vectors ℓ_{α_k}:

$$Tx \approx x + \sum_{k=1}^{n_L} \alpha_k L_{\alpha_k}(x)$$

TD is then defined as the minimal Euclidean distance between all approximated transformations:

$$\rho(P_x, P_z) = \left(x - z + \sum_{k=1}^{n_L} \left(\alpha_k \ell_{\alpha_k}(x) - \beta_k \ell_{\beta_k}(z) \right) \right)^2, \ \alpha_k \in \left[\alpha_k^{\min}, \alpha_k^{\max}\right], \ \beta_k \in \left[\beta_k^{\min}, \beta_k^{\max}\right]$$

The input space tangent distance is then combined with a kernel which uses the Euclidean distance in the input space, for example the Gaussian kernel $k(x, y) = \exp\left(-\frac{\rho(P_x, P_z)^2}{2\sigma^2}\right)$.

Haar-integration (Schulz-Mirbach, 1994)

For a standard kernel k_0 and a transformation group $\mathcal{T}$, an average of the standard kernel output is computed over all pairwise combinations of the transformed examples:

$$k(x, z) = \int_{\mathcal{T}} \int_{\mathcal{T}} k_0(Tx, T'z) dT dT'$$

This is equivalent to computing the inner product between averages $\overline{\Phi(x)}$ and $\overline{\Phi(z)}$ of the transformed examples:

$$k(x,z) = \int_{\mathcal{T}} \int_{\mathcal{T}} \langle \Phi(Tx), \Phi(T'z) \rangle dT dT' = \left\langle \int_{\mathcal{T}} \Phi(Tx) dT \, , \int_{\mathcal{T}} \Phi(T'z) dT' \right\rangle$$

Kernels between sets (Wolf & Shashua, 2003)

Kernels between sets aim to define kernels invariant to a certain kind of permutation of the input. For example, if one considers kernels over sets of vectors, and represents these as matrices $A = [\Phi(a_1), \ldots., \Phi(a_n)], B = [\Phi(b_1), \ldots., \Phi(b_n)]$, one may define a kernel by the principal angles θ_i in the feature space:

$$k(A,B) = \prod_{i=1}^{n} \cos(\theta_i)$$

This kernel does not modify an existing kernel, but it demonstrates a kernel with an invariance property with respect to the permutations of the columns of the matrices.

4 Theory of Linear Inverse Problems

In the regularized least squares formulation, the sampling operator $S_{\overline{x}}: \mathcal{H} \to \ell^2(I_1)$ is a map from a normed space $\|\|_{\mathcal{H}}$ to a normed space $\|\|_2$. In practise the sample space is often $\mathbb{R}^n$, which is finite dimensional. We saw that the inverse may not exist, so an optimal solution was defined through these norms:

$$f^* = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \|S_{\overline{x}} f - \overline{y}\|_2 \text{ of minimum/penalized } \|f\|_{\mathcal{H}}$$

Unlike most other machine learning algorithms (Schölkopf & Smola, 2001), using optimization techniques is not necessary because this formulation allows derivation of the solution analytically. Analytical treatment allows further insights into the algorithm. In the general, consider an abstract linear map $T: \mathcal{X} \to \mathcal{Y}$ between Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$. Under certain conditions, it is then possible to define an "inverse" operator where T is not bijective. These operators are needed when kernel approximation is later considered from learning theoretic perspective.

4.1 Linear operators

Linear operator is a map $T: \mathcal{X} \to \mathcal{Y}$ between vector spaces over the same field K such that $\forall x, y \in X, \alpha \in \mathrm{K}$:

$$T(x + y) = Tx + Ty$$
$$T(\alpha x) = \alpha T x$$

Bounded operator is a linear operator $T: \mathcal{X} \to \mathcal{Y}$ between normed spaces such that

$$\exists c \in \mathbb{R} : \ \forall x \in \mathcal{X} \ \|Tx\|_{\mathcal{Y}} \leq c\|x\|_{\mathcal{X}}$$

Compact operator (Kreyszig, 1989) is a linear operator $T: \mathcal{X} \to \mathcal{Y}$ between normed spaces such that

$$M \subset \mathcal{X} \text{ bounded} \Longrightarrow \overline{T(M)} \text{ compact}$$

A foundational result of linear algebra is that any linear operator $T: \mathcal{X} \to \mathcal{Y}$ between **finite-dimensional** vector spaces $\dim(\mathcal{X}) = m, \dim(\mathcal{Y}) = n$ with respect given bases may be represented by an $n \times m$ **matrix**:

$$T = \begin{pmatrix} t_{11} & \cdots & t_{1m} \\ \vdots & \ddots & \vdots \\ t_{n1} & \cdots & t_{nm} \end{pmatrix}$$

Combining several straightforward results in (Kreyszig, 1989), we obtain the following hierarchy:

$$\text{Finite-dimensional operator (matrix)} \subset \text{Compact operator} \subset \text{Bounded operator} \subset \text{Linear operator}$$

This hiearchy is useful because the increasingly specifized cases allow increasingly powerful results.

Finally, the following linear operator is very useful for considering the spaces associated to a linear operator.

Definition (Kreyszig, 1989)

Let $T: \mathcal{X} \to H_2$ be a bounded Linear operator, where $\mathcal{X}$ and $\mathcal{Y}$ are Hilbert spaces. Then the Hilbert-**adjoint operator** T^* of T is the operator $T^*: \mathcal{Y} \to \mathcal{X}$ such that for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$:

$$\langle Tx, y\rangle_{\mathcal{Y}} = \langle x, T^*y\rangle_{\mathcal{X}}$$

4.2 Fundamental subspaces of a linear operator

Define the following subspaces, known as range and nullspace, for a linear operator $A: \mathcal{X} \to \mathcal{Y}$:

$$\mathcal{R}(A) = \{y : y = Ax,\ x \in \mathcal{X}\}$$
$$\mathcal{N}(A) = \{x : Ax = 0, x \in \mathcal{X}\}$$

Then there are four subspaces associated to the operator $A: \mathcal{X} \to \mathcal{Y}$ and its adjoint $A^*: \mathcal{Y} \to \mathcal{X}$:

$$\mathcal{R}(A) \subset \mathcal{Y}$$
$$\mathcal{N}(A) \subset \mathcal{X}$$
$$\mathcal{R}(A^*) \subset \mathcal{X}$$
$$\mathcal{N}(A^*) \subset \mathcal{Y}$$

With the following property:

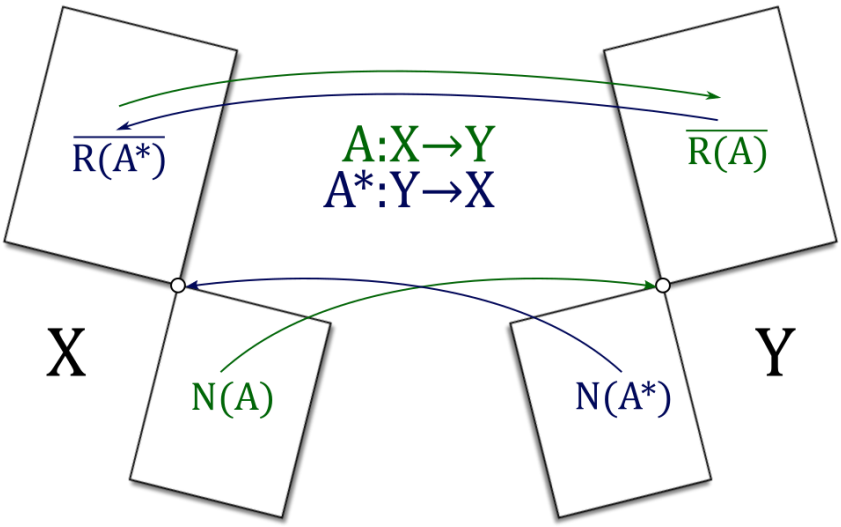
$$\mathcal{R}(A)^\perp = \mathcal{N}(A^*)$$
$$\mathcal{R}(A^*)^\perp = \mathcal{N}(A)$$

Since $\mathcal{N}(A)$ and $\mathcal{N}(A^*)$ are closed we have the factorization $\mathcal{X} = \mathcal{N}(A) \oplus \mathcal{N}(A)^\perp$ and $\mathcal{Y} = \mathcal{N}(A^*) \oplus \mathcal{N}(A^*)^\perp$. Furthermore, if $\mathcal{R}(A)$ is closed $(\mathcal{R}(A)^\perp)^\perp = \mathcal{R}(A)$, otherwise $(\mathcal{R}(A)^\perp)^\perp = \overline{\mathcal{R}(A)}$ (Kreyszig, 1989).

Therefore we get the following factorization, where closure can be omitted if the range is closed:

$$\mathcal{X} = \overline{\mathcal{R}(A^*)} \oplus \mathcal{N}(A)$$
$$\mathcal{Y} = \overline{\mathcal{R}(A)} \oplus \mathcal{N}(A^*)$$

Any vector $\overline{x} \in \mathcal{X}$ can in this case be factored into $\overline{\mathcal{R}(A^*)}$ and $\mathcal{N}(A)$ component and the action of $A\overline{x}$ expressed as a bijection $A_{|\overline{\mathcal{R}(A^*)}}$ and null mapping $A_{|\mathcal{N}(A)}$ (Moon, 1999):



$$\overline{x} = \overline{x}_r + \overline{x}_n$$
$$A\overline{x} = A\overline{x}_r + A\overline{x}_n \text{ with } A\overline{x}_r \in \mathcal{R}(A), A\overline{x}_n = 0$$

Subspaces associated to a matrix.

For a matrix $A: \mathbb{R}^n \to \mathbb{R}^m$ since $\dim\big(\mathcal{R}(A)\big) = \dim\big(\mathcal{R}(A^*)\big)$ this means we have the following rank results:

$$\dim\big(\mathcal{R}(A)\big) = r$$
$$\dim\big(\mathcal{R}(A^*)\big) = r$$
$$\dim\big(\mathcal{N}(A)\big) = n - r$$
$$\dim\big(\mathcal{N}(A^*)\big) = m - r$$

Where $\mathcal{R}(A)$ is called column space, $\mathcal{N}(A)$ the nullspace and $\mathcal{R}(A^*)$ the row space. The spaces are equipped with the inner product $\langle \overline{x}, \overline{y} \rangle = \overline{x}^T \overline{y}$ implying for the adjoint $A^* = A^T$ and the terminology follows. (Kreyszig, 1989).

4.3 Spectral Theorems

The following factorizations are defined for linear operators $T: \mathcal{H} \to \mathcal{H}$ which are self-adjoint $T^* = T$, widely referred to as *spectral theorems*. They are presented in the order of abstraction.

Spectral Theorem for Bounded Operators (Kreyszig, 1989)

Let $\mathcal{H}$ be a complex Hilbert space and $T: \mathcal{H} \to \mathcal{H}$ be a bounded self-adjoint operator. Then T has the spectral representation

$$T = \int_m^M \lambda dE_\lambda$$

where $\{E_\lambda\}$ is the spectral family associated with T. The integral is to be understood in the sense of uniform operator convergence.

This theorem is useful to illustrate the most general case, but is not used in this thesis. For details, see the details and extensive derivation in (Kreyszig, 1989, ss. 459-522).

Spectral Theorem for Compact Operators (Kreyszig, 1989)

Let $\mathcal{H}$ be a Hilbert space and $T: \mathcal{H} \to \mathcal{H}$ be a compact self-adjoint operator. Then there is a countable orthonormal set $\{e_j\}_{j \in J}$ in $\mathcal{H}$ and $\{\lambda_j\}_{j \in J}$ $|\lambda_1| \geq |\lambda_2| \geq \cdots > 0$ converging to zero such that

$$Tf = \sum_{j \in J} \lambda_j \langle f, e_j \rangle_{\mathcal{H}} e_j$$

Diagonalization of Symmetric Matrices (Moon, 1999)

Every $m \times m$ real matrix $A: \mathbb{R}^m \to \mathbb{R}^m$ such that $A^T = A$ is called symmetric (self-adjoint in $\mathbb{R}^m$) and can be factored

$$A = U \Sigma U^T$$

where $U \in R^{m \times m}$ is orthogonal and $\Sigma \in \mathbb{R}^{m \times m}$ has the form $\Sigma = diag(\lambda_1, \lambda_2, \dots, \lambda_n)$. U contains the eigenvectors of A as columns and (λ_i) the corresponding eigenvalues.

4.4 Singular Value Decompositions

4.4.1 SVD for compact operators

Compact operators are an important special case, since specific integral operators are compact, and they are used in connection with kernels.

Theorem: Compact integral operators (Engl;Hanke;& Neubauer, 1996)

If $\Omega \subseteq \mathbb{R}^n$ is compact and Jordan measurable and the kernel k is continuous on $\{(s,t) \in \Omega \times \Omega \mid s \neq t\}$ and for all $s \neq t \in \Omega$, $|k(s,t)| \leq \frac{M}{|s-t|^{n-\epsilon}}$ with $M > 0$, $\epsilon > 0$, then

$$K: \mathcal{L}^2(\Omega) \to \mathcal{L}^2(\Omega)$$
$$x \mapsto (Kx)(s) := \int_\Omega k(s,t)x(t)dt$$

is compact.

The following two results are analogous between compact operators and finite-dimensional operators.

Theorem: Singular values of a compact operator (Engl;Hanke;& Neubauer, 1996)

With a singular system $(\{\sigma_i\}, \{v_i\}, \{u_i\})$ of orthonormal eigenvectors v_i of K^*K, orthonormal eigenvectors u_i of KK^* and nonzero eigenvalues σ_i^2 of K^*K/KK^* ordered by multiplicity, a compact operator K may be exressed:

$$Kv_i = \sigma_i u_i$$
$$K^*u_i = \sigma_i v_i$$
$$Kx = \sum_{i=1}^{\infty} \sigma_i \langle x, v_i \rangle u_i$$
$$K^*x = \sum_{i=1}^{\infty} \sigma_i \langle x, u_i \rangle v_i$$
$$K^*Kx = \sum_{i=1}^{\infty} \sigma_i^2 \langle x, v_i \rangle v_i$$
$$KK^*x = \sum_{i=1}^{\infty} \sigma_i^2 \langle x, u_i \rangle u_i$$

where these infinite series converge, they are called the singular value expansions.

K has finitely many singular values iff it has a finite-dimensional range. For an integral operator K with square integrable kernel k, this is the case iff k is degenerate (Engl;Hanke;& Neubauer, 1996):

$$k(s,t) = \sum_{i=1}^{n} \phi_i(s)\psi_i(t)$$

If there are infinitely many singular values, they accumulate at 0 (Kreyszig, 1989):

$$\lim_{i\to\infty} \sigma_i = 0$$

As a simple application of Banach's open mapping theorem in (Engl;Hanke;& Neubauer, 1996), the range of K is closed iff it is finite-dimensional. Therefore $\mathcal{R}(K)$ of non-degenerate K is non-closed.

*4.4.2 **SVD for matrices***

Every matrix $A \in \mathbb{R}^{m\times n}$ can be factored as

$$A = U\Sigma V^T$$

where $U \in \mathbb{R}^{m\times m}, V \in \mathbb{R}^{n\times n}$ are orthogonal and $\Sigma \in \mathbb{R}^{m\times n}$ has the form $\Sigma = diag(\sigma_1, \sigma_2, \dots, \sigma_p)$ where $p = \min(m,n)$. The diagonal values of Σ are called singular values of A and are usually ordered $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$.

We see that V,U contains the eigenvectors of A^TA, AA^T as columns, respectively. Σ contains the square roots $\sqrt{\lambda_i}$ of the eigenvalues of A^TA/AA^T.

$$A^TA = V\Sigma^T\Sigma V^T = V\Sigma_n V^T$$
$$AA^T = U\Sigma^T\Sigma U^T = U\Sigma_m U^T$$

This way the SVD can also be used to diagonalize the symmetric positive semidefinite matrix A^TA or AA^T.

It is often convinient to break the SVD into two parts. Split the singular values into non-zero and zero parts:

$\Sigma = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix}$ where $\Sigma_1 = diag(\sigma_1, \dots, \sigma_r) \in \mathbb{R}^{r\times r}$ and $\Sigma_1 = diag(0, \dots, 0) \in \mathbb{R}^{(m-r)\times(n-r)}$

Then the SVD can be written (Moon, 1999):

$$A = \begin{pmatrix} U_1 & U_2 \end{pmatrix} \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix} \begin{pmatrix} V_1^T \\ V_2^T \end{pmatrix} = U_1\Sigma_1 V_1^T = \sum_{i=1}^{r} \sigma_i \overline{u}_i \overline{v}_i^T$$

From this factorization, it is easy to determine the four fundamental subspaces of A:

- $\mathcal{R}(A) = span(U_1)$
- $\mathcal{N}(A) = span(V_2)$
- $\mathcal{R}(A^T) = span(V_1)$
- $\mathcal{N}(A^T) = span(U_2)$

Relating this back to our diagram, for the $m \times n$ matrix A: $\mathbb{R}^n \to \mathbb{R}^m$ the SVD factorizes into:

- U decomposes $\mathbb{R}^m = \mathcal{R}(A) \oplus \mathcal{N}(A^T) = span(U_1) \oplus span(U_2)$
- V decomposes $\mathbb{R}^n = \mathcal{R}(A^T) \oplus \mathcal{N}(A) = span(V_1) \oplus span(V_2)$

4.5 Pseudoinverse of linear operators

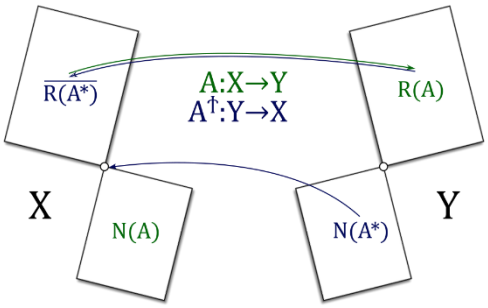
Definition: Best-approximate solution (Engl;Hanke;& Neubauer, 1996)

Let $T: \mathcal{X} \to \mathcal{Y}$ be a bounded linear operator where $\mathcal{X}$ and $\mathcal{Y}$ are Hilbert spaces .

- $x \in \mathcal{X}$ is called the least-squares solution of $Tx = y$ if $\|Tx - y\| = inf\{\|Tz - y\| \mid z \in \mathcal{X}\}$
- $x \in \mathcal{X}$ is called best-approximate solution of $Tx = y$ if x is a least-squares solution of $Tx = y$ and $\|x\| = inf\{\|z\| \mid z \text{ is a least squares solution of } Tx = y\}$

Bounded operator pseudoinverse (Engl;Hanke;& Neubauer, 1996)

Let $\overline{T} \coloneqq T|_{\mathcal{N}(T)^\perp}: \mathcal{N}(T)^\perp \to \mathcal{R}(T)$. The Moore-Penrose inverse $T^\dagger$ of $T \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ is defined as the unique linear extension of $\overline{T}^{-1}$ to $\mathcal{D}(T^\dagger) = \mathcal{R}(T) \oplus \mathcal{R}(T)^\perp$ with $\mathcal{N}(T^\dagger) = \mathcal{R}(T)^\perp$.



Theorem (Engl;Hanke;& Neubauer, 1996)

Let $y \in \mathcal{D}(T^\dagger)$. Then $Tx = y$ has a unique best-approximative solution given by

$$x^\dagger = T^\dagger y$$

where the set of all least-squares solutions is $x^\dagger + \mathcal{N}(T)$. $x \in \mathcal{X}$ is a least-squares solution of $Tx = y$ iff

$$T^*Tx = T^*y$$

The reason for assuming boundedness or $y \in \mathcal{D}(T^\dagger)$ is explained by two basic facts (Kreyszig, 1989):

- A linear operator $T: \mathcal{D}(T) \to Y$ is bounded iff it is continuous.
- If $\mathcal{R}(T)$ is closed, then $\mathcal{D}(T^\dagger) = \mathcal{R}(T) + \mathcal{R}(T)^\perp = \mathcal{Y}$.

And the following proposition (Engl;Hanke;& Neubauer, 1996):

- $T^\dagger$ is bounded (continuous) if and only if $\mathcal{R}(T)$ is closed.

This implies that $\mathcal{R}(T)$ being closed is equivalent to the universal existence and continuous dependence of the solution. This is the case for example when T is finite dimensional. In general this is not the case. For example when a compact operator K is not finite dimensional and thus the range is not closed, the pseudoinverse $K^\dagger$ is a densely defined unbounded linear operator. In this case the best-approximate solution of the inverse problem $Kx = y$ does not depend continuously on the right hand side and the problem is ill-posed. This implies means that further investigation is needed to stabilize the solution.

4.5.1 ***Compact operator pseudoinverse***

For a compact operator, following condition implies the existence and spectral representation of the solution (Engl;Hanke;& Neubauer, 1996):

$$y \in \mathcal{D}(K^\dagger) \Leftrightarrow \sum_{i=1}^{\infty} \frac{|\langle y, u_i \rangle|^2}{\sigma_i^2} < \infty$$

$$\text{For } y \in \mathcal{D}(K^\dagger) : K^\dagger y = \sum_{i=1}^{\infty} \frac{\langle y, u_i \rangle}{\sigma_i} v_i$$

This expression also shows how the pseudoinverse is instable. Small errors in the data are amplified by large constants $\frac{1}{\sigma_i}$. Suppose for example we have added an error term $y_\delta = y + \delta u_i$. Then $\|y - y_\delta\| = \delta$ but $\|Ky - Ky_\delta\| = \left\| \frac{\langle \delta u_i, u_i \rangle}{\sigma_i} v_n \right\| = \frac{\delta}{\sigma_i} \to \infty$ as $i \to \infty$. Therefore if we assume that we have not observed y but a noisy sample y_δ close to it, $K^\dagger y_\sigma \approx K^\dagger y$ is not a good solution. We seek to replace $K^\dagger$ by a new operator $K_\alpha^\dagger$ which is continuous and stable against noise $K_\alpha^\dagger y_\sigma \approx K^\dagger y$. This property is called regularization and the regularization parameter α determining the operator should be noise dependent so that as $\delta \to 0$ the solutions converge $K_\alpha^\dagger y_\sigma \to K^\dagger y$.

4.5.2 ***Matrix pseudoinverse***

Assume matrix $A \in \mathbb{R}^{m\times n}$ has the SVD:

$$A = U\Sigma V^T$$

where $U \in \mathbb{R}^{m\times m}, V \in \mathbb{R}^{n\times n}$ are orthogonal and $\Sigma \in \mathbb{R}^{m\times n}$.

Then for the associated to the singular values $\Sigma = diag(\sigma_1, \sigma_2, \dots, \sigma_p)$ define the matrix:

$$\Sigma^\dagger = diag(\sigma_1', \sigma_2', \dots, \sigma_p')$$
$$\sigma_i' = \begin{matrix} 1/\sigma_i \text{ if } \sigma_i \neq 0 \\ 0 \text{ otherwise} \end{matrix}$$

Based on the previous SVD factorization, it is the easily seen that the pseudoinverse can be written:

$$A^\dagger = V\Sigma^\dagger U^T$$

4.6 Regularization of the pseudoinverse

Define the following notation for a compact operator $K^*Kx = \sum_{i=1}^{\infty} \sigma_i^2 \langle x, v_i\rangle v_i$:

$$f(K^*K) \coloneqq \sum_{i=1}^{\infty} f(\sigma_n^2)\langle ., v_n\rangle v_n$$

Then a regularized inverse is an operator which replaces the singular values $\sigma_i \hookrightarrow g_\alpha(\sigma_i)$ based on a function g_α (Engl;Hanke;& Neubauer, 1996):

$$x_\alpha^\dagger \coloneqq K_\alpha^\dagger y = \sum_{i=1}^{\infty} g_\alpha(\sigma_i)\langle y, u_i\rangle v_i$$

where $g_\alpha: [0, \|T\|^2] \to \mathbb{R}$ is a piecewise continuous function function such that

$$\forall \alpha > 0\ \exists C > 0 : |\lambda g_\alpha(\lambda)| \leq C$$
$$\lim_{\alpha\to 0} g_\alpha(\lambda) = \frac{1}{\lambda}$$

Then the following convergence result can be shown

$$\lim_{\alpha\to 0} g_\alpha(T^*T)T^*y = \begin{cases} T^\dagger y \text{ when } y \in \mathcal{D}(T^\dagger) \\ \infty \text{ when } y \notin \mathcal{D}(T^\dagger) \end{cases}$$

A recent application to machine learning is in (De Vito;Rosasco;& Verri, Spectral Methods for Regularization in Learning Theory, 2005) where different such functions are investigated in connection to algorithms.

Tikhonov regularization is a regularization method which defines the following g_α fulfilling the assumptions:

$$g_\alpha(\lambda) = \frac{1}{\lambda + \alpha}$$

Using the spectral theorem, then the regularized estimator can then be written through either T^*T or TT^*, and the solution expressed through the modified spectrum:

$$T_\alpha^\dagger y = \sum_{i=1}^{\infty} \frac{\sigma_n}{\sigma_n^2 + \alpha} \langle y, u_i \rangle v_i$$

$$T_\alpha^\dagger y = (T^*T + \alpha I)^{-1} T^* y = T^*(TT^* + \alpha I)^{-1}\, y$$

The Tikhonov regularized solution $x_\alpha^\dagger = T_\alpha^\dagger y$ is the unique minimizer of the Tikhonov functional (Engl;Hanke;& Neubauer, 1996) we saw previously:

$$f_\alpha(x) = \|Tx - y\|^2 + \alpha \|x\|^2$$

When the singular values are large compared to α, the factors $\frac{\sigma_n}{\sigma_n^2+\alpha}$ are close to $\frac{1}{\sigma_n}$. But as $n \to \infty$ the factors $\frac{\sigma_n}{\sigma_n^2+\alpha} \to 0$ and they stay bounded; in particular when $\sigma_i < \alpha$ the factors are less than 1. This propagates the error $\langle y, u_i \rangle$ much less than the large factors $\frac{1}{\sigma_n}$, and therefore Tikhonov regularization stabilizes the solution as a function of the regularization strength α.

5 Theory of Kernel Approximation

In this section we present the mathematical formulation of learning theory in connection to kernel approximation. Operator theoretic results relevant to the approximation and our learning bound are also briefly discussed.

5.1 Random variable formulation

Define the learning scenario as follows. Assume the input space X is a closed and bounded subset of $\mathbb{R}^n$. Let the output space Y be a closed interval $[-M, M] \subseteq \mathbb{R}$. The sample space is defined $Z = X \times Y$. This is a common practical setting in machine learning (De Vito;Rosasco;& Verri, Spectral Methods for Regularization in Learning Theory, 2005), but it could be further generalized by assuming X to be a manifold and $Y = \mathbb{R}^k$ (Cucker & Smale, 2001).

With the sample space we have associated an unknown probability distribution expressing the target:

$$\rho(x, y) = \rho(y|x)\rho_X(x)$$

This is formulation encompasses the scenario where target function f^* has a random noise term σ:

$$\rho(x, y) = f^*(x) + \sigma(x, y)$$

The training set $\overline{z}$ is then drawn i.i.d from Z according to ρ:

$$\overline{z} = \{(x_1, y_1), \dots, (x_n, y_n)\}$$

To study the learning setting theoretically, define the **empirical error** $\mathcal{E}_{\overline{z}}: \mathcal{H} \to \mathbb{R}$ and **population error** $\mathcal{E}: \mathcal{H} \to \mathbb{R}$:

$$\mathcal{E}_{\overline{z}}(f) = \frac{1}{m}\sum_{i=1}^{m}(f(x_i) - y_i)^2$$

$$\mathcal{E}(f) = \int_{X\times Y}(f(x) - y)^2 d\rho(x, y)$$

We are interested in a function f which minimizes the error $\mathcal{E}(f)$ over the space of square integrable functions $L^2(X, p_X)$. For this purpose, define the regression function as

$$f_\rho(x) = \int_Y y d\rho(y|x)$$

Lemma (Cucker & Smale, 2001)

The error for $f \in L^2(X, p_X)$ can be decomposed:

$$\mathcal{E}(f) = \int_X \Big(f(x) - f_\rho(x)\Big)^2 d\rho_X(x) + \mathcal{E}(f_\rho)$$

Proof

$$\begin{aligned}
\mathcal{E}(f) &= \int_Z \big(f(x) - f_\rho(x) + f_\rho(x) - y\big)^2 d\rho(x,y) \\
&= \int_X \Big(f(x) - f_\rho(x)\Big)^2 d\rho_X(x) + \int_Z \big(f_\rho(x) - y\big)^2 d\rho(x,y) \\
&+ 2\int_Z (f(x) - f_\rho(x))(f_\rho(x) - y) d\rho(x,y) \\
&= \int_X \Big(f(x) - f_\rho(x)\Big)^2 + \mathcal{E}(f_\rho)
\end{aligned}$$

The error $\mathcal{E}(f)$ of function f decomposes into two terms. The second term is the error associated with the regression function f_ρ, is independent of f, and the first term is the expected squared difference between f and f_ρ. This also shows that the regression function achieves the minimum error and that in considering the minimizer of $\mathcal{E}(f)$, the term $\mathcal{E}(f_\rho)$ may be omitted since it is independent of f.

However, since the probability distribution $\rho(x, y)$ is unknown, we cannot use it to define the regression function, which is thus also unknown. We would like to approximate the regression function as closely as possible, given the sample $\overline{z}$ that we do know and the hypothesis space $\mathcal{H}$ we have. The goal is thus to find a learning algorithm $\mathcal{L}$:

$$\mathcal{L}: \overline{z} \mapsto f_{\overline{z}} \text{ such that } f_{\overline{z}} \text{ approximates } f_\rho$$

where the subscript highlights the fact that the function is deduced from the sample $\overline{z}$.

The learning algorithm implicitly or explicitly is based on a family of functions $\mathcal{H}$ used for learning. In general it may be that $f_\rho \notin \mathcal{H}$. Given a fixed hypothesis set, in that case our goal is to learn $f_{\mathcal{H}}$ defined as the closest approximation of f_ρ in $\mathcal{H}$. Since

this is unavailable, we have also define a hypothesis $f_{\overline{z}}$ which minimizes the empirical error:

$$f_{\mathcal{H}} = \underset{f \in \mathcal{H}}{\operatorname{argmin}}\, \mathcal{E}(f)$$

$$f_{\overline{z}} = \underset{f \in \mathcal{H}}{\operatorname{argmin}}\, \mathcal{E}_{\overline{z}}(f)$$

Define the error of a function $f \in H$ as:

$$\mathcal{E}_{\mathcal{H}}(f) = \mathcal{E}(f) - \mathcal{E}(f_{\mathcal{H}})$$

This allows the following decomposition for a function $f_{\overline{z}}$ (Cucker & Smale, 2001)

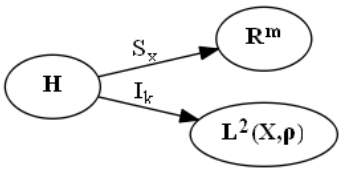
$$\mathcal{E}(f_{\overline{z}}) = \mathcal{E}_{\mathcal{H}}(f_{\overline{z}}) + \mathcal{E}(f_{\mathcal{H}})$$

The first term $\mathcal{E}_{\mathcal{H}}(f_{\overline{z}})$ is the error incurred due to using $f_{\overline{z}}$ instead of the best available function in $\mathcal{H}$, called the *sampling error*, and the second term is the error incurred due to using $\mathcal{H}$ to approximate f_{ρ}, called the *approximation error*. Theoretical considerations often analyse these errors separately.

Algorithms that aim to minimize a population error $\mathcal{E}(f)$ based on the empirical error $\mathcal{E}_{\overline{z}}(f)$, with a possibly added regularization functional also known as complexity penalty, are called **empirical risk minimization** algorithms. The least squares algorithm is a special case of such an algorithm in the sense that the loss is not only convex but can be defined through a norm, which defines a convex loss by the triangle inequality. We now analyse the minimizers and their properties by contrasting them based on this fact.

5.2 Abstract minimizers

Define $\|x\|_m = \frac{1}{m}\|x\|_{\mathbb{R}^m}$, a norm on the sample output space $y_i \in \mathbb{R}^m$, and $\|f\|_{\rho} = \left(\int_{X\times Y} f(x)^2 d\rho(x)\right)^{\frac{1}{2}}$ where $\rho(x)$ is a shorthand for the marginal measure $\rho_X(x) = \int_Y \rho(x,y)$, which defines a norm on the space of square integrable functions $L^2(X,\rho)$. Define the sampling operator $S_{\overline{x}}: \mathcal{H} \to \mathbb{R}^m$ as before and the inclusion $I_k: \mathcal{H} \to L^2(X,\rho)$ as the map of f to its equivalence class in $L^2(X,\rho)$:



$$S_{\overline{x}}f = \big(f(x_1), \dots, f(x_m)\big)^T$$

$$I_k f = [f]$$

The inclusion operator I_k is associated with a number of technical considerations:

- If ρ is degenerate, $I_k f$ is an equivalence class of functions and evaluation $I_k f(x)$ is not be well-defined (Steinwart & Christmann, 2008). For simplicity some theoretical literature assumes that ρ is non-degenrate, to avoid considerations involving zero-measure sets. Then the inclusion is injective, because $I_k f = 0$ in implies that $f(x) \neq 0$ is possible in some open set of ρ-measure zero, which must be empty (ρ nondegenerate). This implies if k is continous $f(x) = 0$, otherwise identically zero for ρ almost all $x \in X$. Thus $(I_k f)(x) = \langle f, K_x \rangle_H = f(x)$ is well-defined. In this case I_k merely changes the norm from $\|f\|_H$ to $\|f\|_{L^2}$.
- The range of inclusion operator needs to be a subset of $L^2(X, \rho)$. When the kernel is measurable and bounded $I_k(\mathcal{H}) \subset L^2(X, \rho)$. Since $k(x, y)^2 \leq k(x, x)k(y, y)$, sufficient condition for boundedness is that the kernel is continuous and X is compact, because then the kernel is bounded: $\sup_{x \in X} \sqrt{k(x, x)} = \sup_{x \in X} \|K_x\|_H < \infty$
- The range $I_k(\mathcal{H})$ is in some theoretical considerations required to be dense in $L^2(X, \rho)$ so that $\inf_{f \in \mathcal{H}} \mathcal{E}(I_k f) = \inf_{f \in L^2(X,v)} \mathcal{E}(f) = \mathcal{E}(f_\rho)$. However, if this is not the case, the optimal solution is the projection of f_ρ into the closure of $I_k(\mathcal{H})$, often denoted by Pf_ρ (De Vito;Rosasco;& Verri, Spectral Methods for Regularization in Learning Theory, 2005).

Based on our discussion of subspaces associated to a linear operator and basic properties (Kreyszig, 1989) of the adjoint, I_k^*, which to be seen has an informative form for kernels, we have on injectivity/surjectivity:

- $\mathcal{R}(I_k)$ is dense in $L^2(X, \rho) \Leftrightarrow \mathcal{N}(I_k^*) = \{0\}$
- $\mathcal{N}(I_k) = \{0\} \Leftrightarrow \mathcal{R}(I_k^*)$ is dense in $\mathcal{H}$

Then the errors may be written through the norms:

$$\mathcal{E}_{\overline{z}}(f) = \|S_{\overline{x}}f - \overline{y}\|_m^2$$

$$\mathcal{E}(f) = \|I_k f - y\|_\rho$$

The augmented empirical and population error add the norm penalty $\lambda > 0$:

$$\mathcal{E}_{\overline{z}}^{\lambda}(f) = \|S_{\overline{x}}f - \overline{y}\|_m^2 + \lambda\|f\|_H^2$$

$$\mathcal{E}^{\lambda}(f) = \|I_k f - y\|_\rho + \lambda\|f\|_H^2$$

Notice that even though we seek solution in $\mathcal{H}$, the error is defined through the norm in $L^2(X, \rho)$, so we need to combine the function with inclusion I_k for this to be defined. Using the norm formulation and the two linear operators $S_{\overline{x}}$ and I_k between Hilbert spaces, the minimization problems become linear inverse problems. As was seen in the general inverse case, for an expressive hypothesis space regularization is necessary to enforce stability since generalization is of central importance in learning. The solution to the regularized empirical and population least squares error is defined:

$$f_{\overline{z}}^{\lambda} = \underset{f\in\mathcal{H}}{\operatorname{argmin}}\|S_{\overline{x}}f - \overline{y}\|_n^2 + \lambda\|f\|_H^2$$

$$f^{\lambda} = \underset{f\in\mathcal{H}}{\operatorname{argmin}}\left\|I_k f - f_\rho\right\|_\rho + \lambda\|f\|_H^2$$

The unregularized solutions to the empirical and population cases can be defined as the pseudoinverse solutions to the linear inverse problems (Cucker & Smale, 2001):

$$S_{\overline{x}}f = \overline{y}$$

$$I_K f = f_p$$

The regularized solutions can be obtained for example through the functional derivative (Smale & Zhou, 2007) or modifying the singular values in the pseudoinverse series expansion if the operator is compact. Because adding λI ($\lambda > 0$) to a self-adjoint operator makes the eigenvalues strictly positive, these have unique solutions expressible through the approximator operator (De Vito;Rosasco;Caponetto;De Giovannini;& Odone, 2005):

$$f_{\overline{z}}^{\lambda} = \left(S_{\overline{x}}^* S_{\overline{x}} + \lambda I\right)^{-1} S_{\overline{x}}^* \overline{y}$$

$$f^{\lambda} = (I_K^* I_K + \lambda I)^{-1} I_K^* f_p$$

It was previously stated that as $\lambda \to 0$ these operators converge to the pseudoinverse.

5.3 Kernel minimizers

Assume that $\mathcal{H}$ is seperable and for the kernel k:

$$\|k\|_\rho = \left(\int_X k(x,x)\, d\rho(x) \right)^{\frac{1}{2}} < \infty$$

This requirement is satisfied if $k: X \times X \to \mathbb{R}$ is continuous on a compact set X.

It is immediately verified that $I_K: \mathcal{H} \to L^2(X,\rho)$ is continuous (bounded) (Steinwart & Christmann, 2008):

$$\int_X |f(x)|^2\, d\rho(x) = \int_X |\langle f, k(.,x)\rangle|^2\, d\rho(x) \leq \|f\|_H^2 \int_X k(x,x)\, d\rho(x)$$

This implies that the adjoint I_K^* exists and is likewise bounded since $\|I_K^*\| = \|I_K\|$ (Kreyszig, 1989).

The adjoint $S_{\overline{x}}^*: \mathbb{R}^n \to \mathcal{H}$ is given by (Rosasco;Belkin;& De Vito, 2010):

$$S_{\overline{x}}^* \overline{c} = \frac{1}{m} \sum_{i=1}^{m} c_i K_{x_i}$$

This follows trivially from the definition $\langle S_{\overline{x}} f, \overline{c}\rangle_{\mathbb{R}^m} = \frac{1}{m}\sum_{i=1}^m \langle f, K_{x_i}\rangle_H c_i = \left\langle f, \frac{1}{m}\sum_{i=1}^m c_i K_{x_i} \right\rangle_H$.

The adjoint $I_k^*: L^2(X,\rho) \to \mathcal{H}$ is given by (Rosasco;Belkin;& De Vito, 2010):

$$(I_k^* f)(x) = \int_X k(x,y) f(y) d\rho(y)$$

This follows from the definition $\langle i^* f, g\rangle_H = \langle f, ig\rangle_{L^2}$. Setting $g = k(.,x)$, we have $(i^* f)(x) = \langle f, ik(.,x)\rangle_{L^2} = \int_X k(x,y) f(y) d\rho(x)$.

Therefore we may define the following **empirical operators**, the **kernel matrix** $K: \mathbb{R}^m \to \mathbb{R}^m$ and the **covariance operator** $C: \mathcal{H} \to \mathcal{H}$:

$$K = S_{\overline{x}} S_{\overline{x}}^*$$

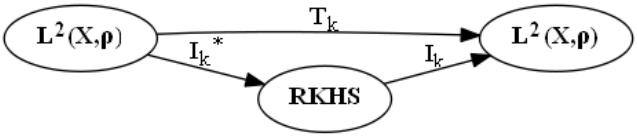
$$C = S_{\overline{x}}^* S_{\overline{x}}$$

The following expressions are derived for reproducing kernel Hilbert spaces:

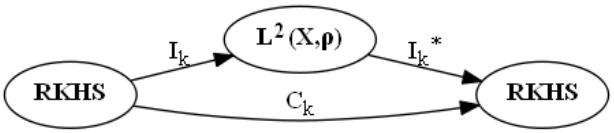
$$K\overline{c} = S_{\overline{x}}S_{\overline{x}}^{*}\overline{c} = \begin{pmatrix} \langle \frac{1}{m}\sum_{i=1}^{m} c_i K_{x_i}, K_{x_1} \rangle \\ \cdots \\ \langle \frac{1}{m}\sum_{i=1}^{m} c_i K_{x_i}, K_{x_n} \rangle \end{pmatrix} \overline{c} = \frac{1}{m}\begin{pmatrix} k(x_1,x_1) & \cdots & k(x_n,x_1) \\ \vdots & \ddots & \vdots \\ k(x_1,x_n) & \cdots & k(x_n,x_n) \end{pmatrix} \overline{c}$$

$$Cf = S_{\overline{x}}^{*}S_{\overline{x}}f = S_{\overline{x}}^{*}\begin{pmatrix} f(x_1) \\ \cdots \\ f(x_n) \end{pmatrix} = \frac{1}{m}\sum_{i=1}^{m} f(x_i)K_{x_i}$$

Also define following **population operators,** the **kernel operator** $T_k: L^2(X,\rho) \to L^2(X,\rho)$ and the **covariance operator** $C_k: \mathcal{H} \to \mathcal{H}$ (Hein & Olivier, 2004):



$$T_k = I_k I_k^{*}$$



$$C_k = I_k^{*} I_k$$

Using $(I_k f)(x) = \langle f, k(.,x) \rangle_H$ and the fact that the integral commutes with the inner product:

$$T_k f = I_k I_k^{*} f = \left\langle \int_X K_y f(y) d\rho_X(y), K_x \right\rangle_H = \int_X k(x,y) f(y) d\rho(y)$$

$$C_k f = I_k^{*} I_k f = \int_X K_y (I_k f)(y) d\rho_X(y) = \int_X K_y \langle f, K_y \rangle_H d\rho(y)$$

The first integral converges in $L^2(X,\rho)$ norm and the second in $\mathcal{H}$ norm.

The population operators closely connected to their empirical variants, for example the following descriptions are used in the literature:

	Empirical	**Population**
Kernel	$\frac{1}{m} S_{\overline{x}} S_{\overline{x}}^{*}$	$I_k I_k^{*}$

Covariance	$\frac{1}{m} S_{\overline{x}}^* S_{\overline{x}}$	$I_k^* I_k$

Therefore the unique solution $f_{\overline{z}}^{\lambda} = \left(\frac{1}{m} S_{\overline{x}}^* S_{\overline{x}} + \lambda I\right)^{-1} \frac{1}{m} S_{\overline{x}}^* \overline{y} = \frac{1}{m} S_{\overline{x}}^* \left(\frac{1}{m} S_{\overline{x}} S_{\overline{x}}^* + \lambda I\right)^{-1} \overline{y}$ in $\mathcal{H}$ minimizing the regularized least squares error can be written through either the kernel matrix or the covariance operator. Using the kernel matrix, we obtain the following finite dimensional problem:

$$f_{\overline{z}}^{\lambda}(x) = \sum_{i=1}^{m} a_i \, K(x, x_i) \text{ where } \overline{a} = \left(\frac{1}{m} K + \lambda I\right)^{-1} \frac{1}{m} \overline{y}$$

Using the previous result that $(I_k^* I_k + \lambda I)^{-1} I_k^* = I_k^* (I_k I_k^* + \lambda I)^{-1}$, the population error minimizer can be written through either T_k or C_k and optionally combined with the inclusion I_k, the resulting function considered in either $\mathcal{H}$ or $L^2(X, \rho)$. The population error minimizer as an element in $L^2(X, \rho)$ is given by:

$$f^{\lambda} = (T_k + \lambda I)^{-1} T_k f_{\rho}$$

The specific form used in the literature is quite variable, contrast for example (Smale & Zhou, 2007) (Rosasco;Belkin;& De Vito, 2010) (De Vito;Rosasco;Caponetto;De Giovannini;& Odone, 2005).

5.4 Spectral and singular value decompositions

The following stronger results, contingent on $\|A\| \leq \|A\|_{\mathrm{HS}} \leq \|A\|_{\mathrm{TC}}$ (Rosasco;Belkin;& De Vito, 2010), may be established under the assumptions. See the appendix (Steinwart & Christmann, 2008, ss. 507-529) which has additional straightforward results founded on these operator norms, closely connected to the spectral theorem in (Kreyszig, 1989). These results allow intuitive interpretations for the operators.

Lemma: The inclusion I_k and adjoint I_k^* are Hilbert-Schmidt and thus compact (Steinwart & Christmann, 2008). For ONB $(e_i)_{i \geq 1}$ of $\mathcal{H}$:

$$\|I_k^*\|_{\mathrm{HS}} = \sum_{i=1}^{\infty} \|I_k^* e_i\|_{\rho}^2 = \int_X \sum_{i=1}^{\infty} |I_k^* e_i(x)|^2 \, d\rho(x) = \int_X \sum_{i=1}^{\infty} e_i(x)^2 \, d\rho(x) = \|k\|_{\rho} < \infty$$

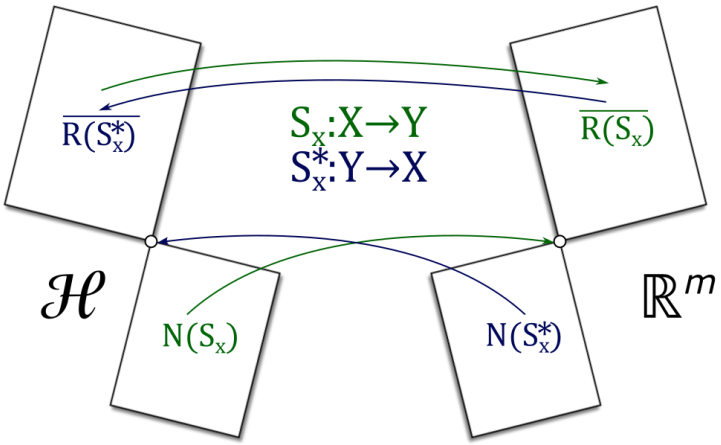
For I_k^*, which implies the same for I_k since $\|A\|_{\mathrm{HS}} = \|A^*\|_{\mathrm{HS}}$.

This leads us to the following theorem, where compactness is essential to be able to apply several results in functional analysis.

Theorem: $T_k = I_k I_k^*$ and $C_k = I_k^* I_k$ are compact, positive, self-adjoint and nuclear operators with $\|T_k\|_{\mathrm{nuc}} = \|C_k\|_{\mathrm{nuc}} = \|I_k^*\|_{\mathrm{HS}}$.

Proof: Compactness of I_k^* implies compactness of $I_k I_k^*$ and $\|I_k^*\|_{\mathrm{HS}} < \infty \Longrightarrow \|I_k I_k^*\|_{\mathrm{nuc}} = \|I_k^*\|_{\mathrm{HS}}$ (Steinwart & Christmann, 2008). Case for C_k goes analogously.

5.4.1 ***Empirical decomposition***



Factor $\mathcal{H} = \mathcal{N}(S_{\overline{x}}) \otimes \overline{\mathcal{R}(S_{\overline{x}}^*)}$ and $\mathbb{R}^m = \mathcal{N}(S_{\overline{x}}^*) \oplus \overline{\mathcal{R}(S_{\overline{x}})}$.

Then we have the following factorizations for operators $S_{\overline{x}}: \mathcal{H} \to \mathbb{R}^m, S_{\overline{x}}^*: \mathbb{R}^m \to \mathcal{H}, S_{\overline{x}} S_{\overline{x}}^*: \mathbb{R}^m \to \mathbb{R}^m, S_{\overline{x}}^* S_{\overline{x}}: \mathcal{H} \to \mathcal{H}$:

$S_{\overline{x}} f = \sqrt{\lambda_i}\langle f, \hat{v}_i\rangle x_i$	$Kx \coloneqq S_{\overline{x}} S_{\overline{x}}^* x = \lambda_i \langle x, x_i\rangle x_i$
$S_{\overline{x}}^* x = \sqrt{\lambda_i}\langle x, x_i\rangle \hat{v}_i$	$Cf \coloneqq S_{\overline{x}}^* S_{\overline{x}} f = \lambda_i \langle f, \hat{v}_i\rangle \hat{v}_i$

Where $(\{x_i\}_{i\geq 1}, \{\hat{v}_i\}_{i\geq 1}, \{\lambda_i\}_{i\geq 1})$ is a singular system of orthonormal eigenvectors of K, orthonormal eigenvectors of C and nonzero eigenvalues of K/C. The sets $\{x_i\}_{i\geq 1}, \{\hat{v}_i\}_{i\geq 1}$ are also a basis for $\mathcal{R}(S_{\overline{x}}), \mathcal{R}(S_{\overline{x}}^*)$, respectively.

The following kernel specific explicit expressions can be written:

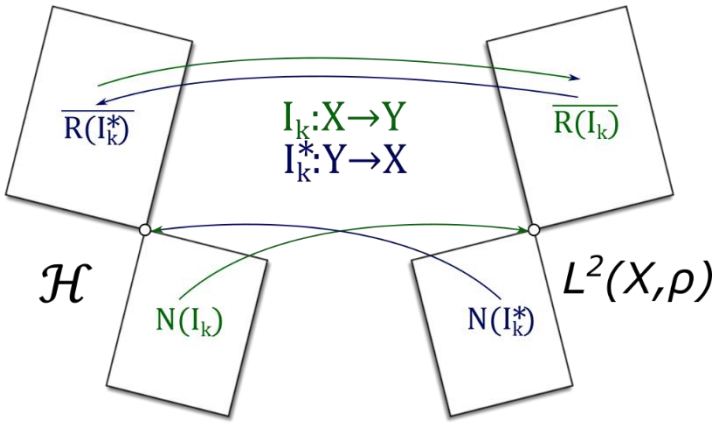
$$S_{\overline{x}} \hat{v}_i = \sqrt{\lambda_i} x_i \Longrightarrow x_i = \frac{1}{\sqrt{\lambda_i}} \big(\hat{v}_i(x_1), \dots, \hat{v}_i(x_m)\big)^T \in \mathbb{R}^m$$

$$S_{\vec{x}}^{*} x_i = \sqrt{\lambda_i}\hat{v}_i \Longrightarrow \hat{v}_i = \frac{1}{\sqrt{\lambda_i} m} \sum_{i=1}^{m} x_i K_{x_i} \in \mathcal{H}$$

The latter expression is sometimes called **kernel PCA** or **KPCA**, because it can be seen as an infinite dimensional analog to the PCA of a finite dimensional feature space. Given a data point x and i'th eigevector x_i of K, the projection to i'th component may be computed for each $\hat{v}_i$:

$$\hat{v}_i(x) = \frac{1}{\sqrt{\lambda_i} m} \sum_{i=1}^{m} x_i k(x, x_i)$$

*5.4.2 **Population decomposition***



Factor $\mathcal{H} = \mathcal{N}(I_k) \otimes \overline{\mathcal{R}(I_k^*)}$ and $L^2(X, \rho) = \mathcal{N}(I_k^*) \oplus \overline{\mathcal{R}(I_k)}$.

Then we have the following factorizations for operators $I_k : \mathcal{H} \to L^2(X, \rho), I_k^* : L^2(X, \rho) \to \mathcal{H}, I_k I_k^* : L^2(X, \rho) \to v, I_k^* I_k : \mathcal{H} \to \mathcal{H}$:

$I_k f = \sqrt{\mu_i}\langle f, v_i\rangle u_i$	$T_k g \coloneqq I_k I_k^* g = \mu_i \langle g, u_i\rangle u_i$
$I_k^* g = \sqrt{\mu_i}\langle g, u_i\rangle v_i$	$C_k f \coloneqq I_k^* I_k f = \mu_i \langle f, v_i\rangle v_i$

Where $(\{u_i\}_{i\geq 1}, \{v_i\}_{i\geq 1}, \{\mu_i\}_{i\geq 1})$ is a singular system of orthonormal eigenvectors of T_k, orthonormal eigenvectors of C_k and nonzero eigenvalues of T_k / C_k. The sets $\{u_i\}_{i\geq 1}, \{v_i\}_{i\geq 1}$ are also a basis for $\mathcal{R}(I_k), \mathcal{R}(I_k^*)$, respectively.

The following kernel specific explicit expressions can be written:

$$I_k v_i = \sqrt{\mu_i} u_i \Longrightarrow u_i(y) = \frac{1}{\sqrt{\mu_i}} v_i(x) \text{ for } \rho\text{-almost all } x \in X$$

$$I_k^* u_i = \sqrt{\mu_i} v_i \Longrightarrow v_i(x) = \frac{1}{\sqrt{\mu_i}} \int_{y \in X} k(x,y) u(y) \rho(y) \in \mathcal{H}$$

5.4.3 *Decompositions and associated projections*

Consider the following operators:

$$\overline{y}_{emp} = S_{\overline{x}} S_{\overline{x},\lambda}^{\dagger} \overline{y} = S_{\overline{x}} S_{\overline{x}}^* \left(S_{\overline{x}} S_{\overline{x}}^* + \lambda I \right)^{-1} \overline{y} = K(K + \lambda I)^{-1} \overline{y}$$

$$f_{emp}^H = S_{\overline{x},\lambda}^{\dagger} S_{\overline{x}} f = \left(S_{\overline{x}}^* S_{\overline{x}} + \lambda I \right)^{-1} S_{\overline{x}}^* S_{\overline{x}} f = (C + \lambda I)^{-1} C f$$

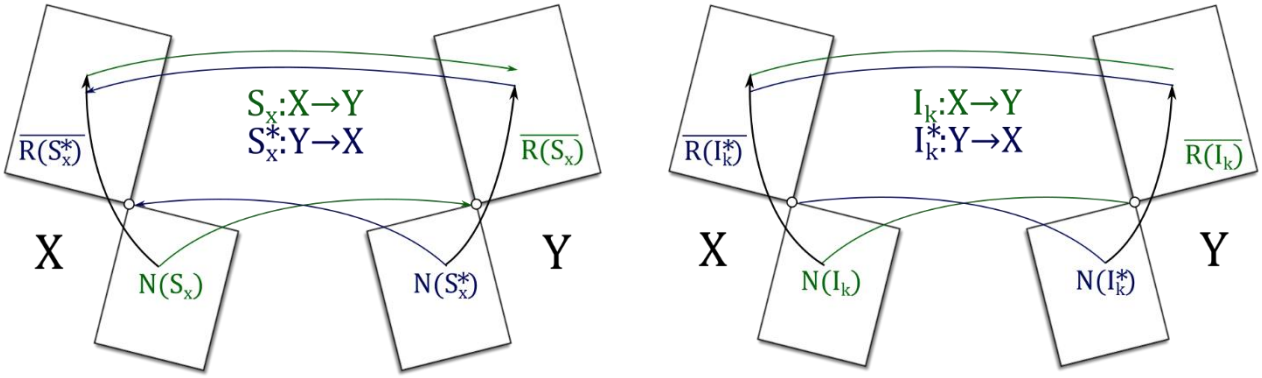
$$g_{pop} = I_k I_{k,\lambda}^{\dagger} g = I_k I_k^* (I_k I_k^* + \lambda I)^{-1} g = T_k (T_k + \lambda I)^{-1} g$$

$$f_{pop}^H = I_{k,\lambda}^{\dagger} I_k f = (I_k^* I_k + \lambda I)^{-1} I_k^* I_k f = (C_k + \lambda I)^{-1} C_k f$$

These can be interpreted as follows:

- Given data $\overline{y}$, predictions $\overline{y}_{emp}$ given by a function learned using $\mathcal{H}$.
- Given function $f \in \mathcal{H}$, function f_{emp}^H learned from $f(\overline{x})$ using $\mathcal{H}$.
- Given function $g \in L^2(X, \rho)$, function in $g_{pop} \in L^2(X, \rho)$ learned using $\mathcal{H}$.
- Given function $f \in \mathcal{H}$, function f_{pop}^H learned from $I_k f \in L^2(X, \rho)$ using $\mathcal{H}$.

Consider the spectral representations of the operators and the following projections:



$K(K + \lambda I)^{-1} \overline{y} = \sum_{i \geq 1} \frac{\lambda_i}{\lambda_i + \lambda} \langle x, x_i \rangle_{\mathbb{R}^m} x_i$	$P_{\overline{\mathcal{R}(S_{\overline{x}})}} \overline{y} = \sum_{i \geq 1 : \lambda_i > 0} \langle x, x_i \rangle_{\mathbb{R}^m} x_i$
$(C + \lambda I)^{-1} C f = \sum_{i \geq 1} \frac{\lambda_i}{\lambda_i + \lambda} \langle f, \hat{v}_i \rangle_H \hat{v}_i$	$P_{\overline{\mathcal{R}(S_{\overline{x}}^*)}} f = \sum_{i \geq 1 : \lambda_i > 0} \langle f, \hat{v}_i \rangle_H \hat{v}_i$

$T_k(T_k+\lambda I)^{-1}g = \sum_{i\geq 1}\frac{\mu_i}{\mu_i+\lambda}\langle g,u_i\rangle u_i$	$P_{\overline{\mathcal{R}(I_k)}}g = \sum_{i\geq 1:\mu_i>0}\langle g,u_i\rangle u_i$
$(C_k+\lambda I)^{-1}C_k f = \sum_{i\geq 1}\frac{\mu_i}{\mu_i+\lambda}\langle f,v_i\rangle v_i$	$P_{\overline{\mathcal{R}(I_k^*)}}f = \sum_{i\geq 1:\mu_i>0}\langle f,v_i\rangle_H v_i$

As $\lambda \to 0$, the regularized pseudoinverse converges to the pseudoinverse, and these operators converge to the corresponding projections. They can be interpreted as ´filtered projections´, in the sense that the term λ scales back the projection coefficients based on the regularized pseudoinverse filtering.

5.5 Connection between $\mathcal{H}$ and L^2 and Mercer's theorem

5.5.1 ***RKHS isometrically embedded in L^2***

Consider a measure μ on $L^2(X,\mu)$. Define a function $f\colon \mathrm{supp}(\mu) \to \mathbb{R}$ associated to each equivalence class $[(f)]$ in $L^2(X,\mu)$, which is then well defined:

$$f \coloneqq g_{|\mathrm{supp}(\mu)} \text{ for any representer } g \in [(f)]$$

Let $\mathrm{supp}(\mu) = X$, i.e. $\mu(X') > 0$ for every open subset $X' \subset X$. It was previously stated for a continous kernel $k\colon X \times X \to \mathbb{R}$ the inclusion $I_k\colon \mathcal{H} \to L^2(X,\mu)$ is then injective, and if X is compact the functions in RKHS are bounded so inclusion is well defined. Consider the decomposition from before:

$$I_k e_i = e_i = \sqrt{\lambda_i}\tilde{e}_i \in L^2(X,\mu)$$

Where $\{\tilde{e}_i\}_{i\geq 1} \subset L^2(X,\mu)$ is an ONB of $\mathcal{N}(I_k^*)^\perp$ and $\{e_i\}_{i\geq 1} \subset \mathcal{H}$ is an ONB of $\mathcal{N}(I_k)^\perp = \mathcal{H}$, $\{\lambda_i\}_{i\geq 0}$ are the eigenvalues of $C_k = I_k I_k^*$ / the nonzero eigenvalues of $T_k = I_k^* I_k$.

Inclusion maps to the same function by definition $e_i(x) = \sqrt{\lambda_i}\tilde{e}_i(x)\ \forall x \in X$, but changes the norms: if $f = \sum_{i\geq 1}\alpha_i e_i$ then $I_k(f) = \sum_{i\geq 1}\sqrt{\lambda_i}\alpha_i\tilde{e}_i$, implying $\|f\|_H = \sum_{i\geq 1}\alpha_i^2$ and $\|I_k(f)\|_{L^2} = \sum_{i\geq 1}\lambda_i\alpha_i^2$. Since inclusion is injective, I_k is a bijection between $\mathcal{H}$ and $\mathcal{N}(I_k^*)^\perp$. We may therefore consider the 'reverse inclusion':

$$I_k^{-1}\tilde{e}_i = \tilde{e}_i = \frac{1}{\sqrt{\lambda_i}}e_i \in \mathcal{H}$$

Now consider the following map $\Im$ between the bases of $\mathcal{N}(I_k^*)^\perp$ and $\mathcal{H}$:

$$\Im\tilde{e}_i \coloneqq e_i = \sqrt{\lambda_i}\tilde{e}_i$$

This maps to a different function, but retains norms: if $f = \sum_{i\geq 1}\beta_i\tilde{e}_i$ then $\Im(f) = \sum_{i\geq 1}\sqrt{\lambda_i}\beta_i\tilde{e}_i = \sum_{i\geq 1}\beta_i e_i$, implying $\|f\|_{L^2} = \sum_{i\geq 1}\beta_i^2$ and $\|\Im(f)\|_H = \sum_{i\geq 1}\beta_i^2$. It is clearly a bijection between $\mathcal{N}(I_k^*)^\perp$ and $\mathcal{H}$, inducing an isometric isomorphism between the spaces.

Since T_k is a positive self-adjoint compact operator, the square root $T_k = T_k^{\frac{1}{2}} \circ T_k^{\frac{1}{2}}$ is defined and given by:

$$T_k^{\frac{1}{2}} f = \sum_{i\geq 1}\sqrt{\lambda_i}\,\langle f, \tilde{e}_i\rangle_{L^2}\tilde{e}_i$$

Considering $T_k^{\frac{1}{2}} f$ as an element of $\mathcal{H}$, we see that it coincides with $\Im$ and induces an isometric isomorphism between $\mathcal{R}(I_k) = \mathcal{N}(I_k^*)^\perp$ and $\mathcal{H}$.

*5.5.2 **Mercer representation***

Recall that $k: X \times X \to \mathbb{R}$ was assumed continuous, X compact and $\text{supp}(\mu) = X$.

We know from before that if $\{e_i\}_{i\geq 1}$ is an ONB of $\mathcal{H}$, then

$$k(x,y) = \sum_{i\geq 1} e_i(x)e_i(y)$$

Where the convergence is uniform and absolute. However, since $e_i(x) = \sqrt{\lambda_i}\tilde{e}_i(x)$, we can express $k(x,y)$ in terms of the eigenfunctions $\{\tilde{e}_i\}_{i\geq 1}$ associated to nonzero eigenvalues of the operator $T_k = I_k^* I_k$:

$$k(x,y) = \sum_{i\geq 1}\lambda_i\tilde{e}_i(x)\tilde{e}_i(y)$$

This result is known as **Mercer's theorem** in the literature. It can be used to derive an expression for k, and an ONB representation of the RKHS, having obtained the eigenfunctions of the positive self-adjoint compact operator $T_k: L^2(X,\mu) \to L^2(X,\mu)$.

Since $k(x,y) = \langle\sqrt{\lambda_i}\tilde{e}_i(x), \sqrt{\lambda_i}\tilde{e}_i(y)\rangle_{l^2}$, Mercer's theorem also gives a feature map $\phi: X \to l^2(J)$ related to a spectral decomposition (Steinwart & Christmann, 2008):

$$\phi: x \mapsto \left(\sqrt{\lambda_i}\tilde{e}_i(x)\right)_{i\geq 1}$$

Which is well defined because $\sum_{i\geq 1}\left|\sqrt{\lambda_i}e_i(x)\right|^2 = k(x,x) < \infty$

*5.5.3 **ONB construction of a RKHS***

Under the assumptions of Mercer's theorem, an alternative construction exists for the RKHS in terms of eigenfunctions $\{\tilde{e}_i\}_{i\geq 1} \subset L^2(X,\mu)$ associated to nonzero eigenvalues $\{\lambda_i\}_{i\geq 1}$ of T_k. Define (Steinwart & Christmann, 2008):

$$\mathcal{H} = \left\{ f = \sum_{j\in J} \sqrt{\lambda_j} a_j \tilde{e}_j : \{a_j\} \in l^2(J) \right\}$$

With an inner product

$$\langle \sum_{j\in J} \sqrt{\lambda_j} a_j \tilde{e}_j , \sum_{j\in J} \sqrt{\lambda_j} b_j \tilde{e}_j \rangle = \sum_{j\in J} a_j b_j$$

Then $\mathcal{H}$ is the RKHS of k and $\left\{\sqrt{\lambda_i}\tilde{e}_i\right\}_{i\geq 1}$ clearly an ONB of $\mathcal{H}$.

Proof

The functions are well-defined since they converge for all x, from Hölder's inequality (Kreyszig, 1989, s. 14):

$$\sum_{j\in J}\left|\sqrt{\lambda_j}a_j\tilde{e}_j\right| \leq \left[\sum_{j\in J}|a_j|^2\right]^{\frac{1}{2}}\left[\sum_{j\in J}\left|\sqrt{\lambda_i}\tilde{e}_j(x)\right|^2\right]^{\frac{1}{2}} = \|\{a_j\}\|_{l^2}\sqrt{k(x,x)}$$

Then k obtained from Mercer's theorem is the reproducing kernel of $\mathcal{H}$:

- $k(.,x) \in \mathcal{H}$, since $k(.,x) = \sum_{j\in J}\left(\lambda_j\tilde{e}_j(x)\right)\tilde{e}_j \Longrightarrow \sum_{j\in J}\left|\sqrt{\lambda_i}\tilde{e}_j(x)\right|^2 = \sum_{j\in J}\lambda_j\tilde{e}_j(x)^2 = k(x,x) < \infty$
- k has the reproducing property since $f = \sum_{j\in J} a_j\tilde{e}_j \Longrightarrow \langle f, k(.,x)\rangle = \langle \sum_{j\in J} a_j\tilde{e}_j , \sum_{j\in J}\left(\lambda_j\tilde{e}_j(x)\right)\tilde{e}_j\rangle = \sum_{j\in J} a_j\tilde{e}_j(x) = f(x)$

*5.5.4 **Connection between the measure and the embedding***

Note that the measure μ on $L^2(X,\mu)$ was **arbitrary**, with the relatively strong assumption $\text{supp}(\mu) = X$. Since the kernel and the RKHS is independent of the measure, for all such measures this construction results in the same RKHS. Futhermore, if the kernel k is such that $\{e_i\}_{i\geq 1}$ is a basis of $L^2(X,\mu)$, the previous map

is an isometric isomorphism between the spaces $\mathcal{H} \leftrightarrow L^2(X,\mu)$, instead of an embedding to a subspace.

Mercer's theorem gives useful intuition into not only to the functions which make up the RKHS, but also the norm used for approximation. Suppose for example that μ is a nondegenerate probability measure. Then by definition each basis function $\tilde{e}_j \in L^2(X,\mu)$ limits an area size of the unit square. However, for a function $\sum_{j\in J}\sqrt{\lambda_j}a_j\tilde{e}_j$ in the RKHS, the functions are scaled to zero by multiplicative factors λ_j, which may decrease quite rapidly. While in principle a kernel method may use 'an infinite dimensional feature space', this guides the learning process in practise to arrive at a function determined by the amount of data and it's capacity to overwrite these factors. This observation is closely related to the effective dimension to be presented later.

Suppose this condition does not hold, $\mathrm{supp}(\mu) \neq X$. Then for any two representers $g_1, g_2 \in [(f)]$, in the set $X_0 \coloneqq X \backslash X_\mu \neq \emptyset$ it is possible that $g_1(x) \neq g_2(x)$. This motivated the association to the function restricted to $\mathrm{supp}(\mu)$:

$$g_{|\mathrm{supp}(\mu)} \text{ for any representer } g \in [(f)]$$

For such measure, for each equivalence class in $L^2(X,\mu)$ and the associated function $g_{|\mathrm{supp}(\mu)}$, $\mathcal{H}$ contains the extensions of functions $g_{|\mathrm{supp}(\mu)}$ to X. We obtain the following hierarchy

	k universal	k ¬universal
$\mathrm{supp}(\mu) = X$	$\mathcal{H} = L^2(X,\mu)$	$\mathcal{H} \subset L^2(X,\mu)$
$\mathrm{supp}(\mu) \subset X$	$\mathcal{H}$ extension $L^2(X,\mu)_{\|X_\mu}$	$\mathcal{H}$ extension $\subset L^2(X,\mu)_{\|X_\mu}$

5.6 Convergence

The empirical operators are closely connected to the population operators. To utilize the law of large numbers in a precise way, one can use inequalities to bound the difference between the empirical mean and the expectation of random variables.

Theorem: Hoeffding inequality (Mohri;Rostamizadeh;& Ameet, 2012)

If $X_1, \ldots, X_n$ are independent real random variables such that $|X_i| \leq M$ and $\overline{X} = \frac{1}{n}\sum_i X_i$, then

$$\mathrm{P}\left(\left|\overline{X} - \mathrm{E}[X]\right| \geq \epsilon\right) \leq 2e^{-\frac{n\epsilon^2}{2M^2}}$$

We can reformulate this is as the following bound (Abu-Mostafa;Magdon-Ismail;& Lin, 2012). With probability greater or equal to $1 - \delta$:

$$\left|\overline{X} - \mathrm{E}[X]\right| \leq \sqrt{\frac{2M^2}{n}\log\left(\frac{2}{\delta}\right)}$$

The same inequality extends to Hilbert spaces (Rosasco;Belkin;& De Vito, 2010). If $\xi_1, \ldots, \xi_n$ are random variables in a separable real Hilbert space H such that $\|\xi_i\|_H \leq M$, then with probability greater or equal to $1 - \delta$:

$$\left\|\overline{X} - \mathrm{E}[X]\right\|_H \leq \sqrt{\frac{2M^2}{n}\log\left(\frac{2}{\delta}\right)}$$

5.6.1 ***Operator and spectral converge***

To obtain insight about the relation of the sample operators and population operators, consider the Hilbert space random variable $\xi(f) = f(x)K_x \in \mathcal{H}$ on (Z, ρ). We obtain for the empirical mean of $\xi_1, \ldots, \xi_m \sim \xi$ and the population mean (Smale & Zhou, 2007):

$$\frac{1}{m}\sum_{i=1}^{m}\xi_i(f) = \frac{1}{m}\sum_{i=1}^{m} f(x_i)K_{x_i} = \frac{1}{m}S_{\overline{x}}^{*}S_{\overline{x}}f$$

$$E[\xi(f)] = \int_X K_x f(x) d\rho_X(x) = C_k f$$

This shows that $C = \frac{1}{m}S_{\overline{x}}^{*}S_{\overline{x}}$ is an unbiased estimate of C_K, the probability of deviation decreasing with increasing number of samples as a consequence of the law of large numbers. In fact, the difference can be explicitly bounded.

Theorem (De Vito;Rosasco;Caponetto;De Giovannini;& Odone, 2005)

With probability greater or equal to $1 - \delta$:

$$\|C - C_k\|_{HS} \leq \sqrt{\frac{2\|k\|_\rho^{\,2}}{n}\log\left(\frac{2}{\delta}\right)}$$

Proof

$\xi_i(f) = \langle f, K_{x_i} \rangle K_{x_i}$ is an operator $\mathcal{H} \to \mathcal{H}$ and $\|\xi\|_{HS}^2 = \|K_x\|^4 \leq \|k\|_\rho^2$ implies $\|\xi_i\|_{HS} \leq \|k\|_\rho$. Apply Hoeffding inequality to obtain the result.

The relation between $K: \mathbb{R}^m \to \mathbb{R}^m$ and $T_k: L^2(X, \rho) \to L^2(X, \rho)$ cannot be stated through operator norm, because they do not operate on the same space. However, a simple spectral converge result can be show utilizing the following theorem refered in (Rosasco;Belkin;& De Vito, 2010).

Theorem

Let $\mathcal{H}$ be a seperable Hilbert space with A, B self-adjoint compact operators. Let $\{\mu_i\}_{i \geq 1}$ be an enumeration of eigenvalues of $B - A$. Then there exists extended enumerations $\{\alpha_i\}_{i \geq 1}$ and $\{\beta_i\}_{i \geq 1}$of eigenvalues of A and B such that for any nonnegative convex function ϕ with $\phi(0) = 0$:

$$\sum_{i \geq 1} \phi(\beta_i - \alpha_i) \leq \sum_{i \geq 1} \phi(\mu_i)$$

Theorem (Rosasco;Belkin;& De Vito, 2010)

If $\{\lambda_i\}_{i \geq 1}$and $\{\mu_i\}_{i \geq 1}$ are the eigenvalues of C and C_k, they are also the eigenvalues of K and T_k, respectively, and for nonnegative convex function ϕ:

$$\sum_{i \geq 1} \phi(\mu_i - \lambda_i)^2 \leq \|C - C_k\|_{HS}^2$$

Proof

Apply the above theorem for $\phi(x) = x^2, A = C, B = C_k$.

Because $\|C - C_k\|_{HS}^2$ was bounded previously, we have bound the difference of the spectrum of K and T_k in the ℓ^2 norm.

Finally, equivalent converge results and bounds may be obtained for the operators related to the projections. For example, consider the following definitions, called the **empirical effective dimension** and the **effective dimension**:

$$D_\lambda(K) \coloneqq \text{trace}[K(K + \lambda I)^{-1}] = \sum_{i \geq 1} \frac{\lambda_i}{\lambda_i + \lambda}$$

$$D_\lambda(T_k) \coloneqq \text{trace}[T_k(T_k + \lambda I)^{-1}] = \sum_{i \geq 1} \frac{\mu_i}{\mu_i + \lambda}$$

Note that $D_\lambda(T_k)$, depending on the unknown measure, is a theoretical quantity, while $D_\lambda(K)$ may be computed in practise. Similar to the previous operators, converge implies we can take the empirical version as an approximation of the population version. The following result which establishes the convergence of their spectrum was proven in (Caponnetto;Rosasco;De Vito;& Verri, 2005):

For $0 < \eta < 1, \Delta > 0$ and $m \geq \left(\frac{4\|k\|_\rho}{\lambda\Delta}\left(1 + \frac{\|k\|_\rho}{\lambda}\right)\log\left(\frac{2}{\eta}\right)\right)^2$, with probability $1 - 2\eta$:

$$|D_\lambda(T_k) - D_\lambda(K)| \leq \Delta$$

5.6.2 ***Solution convergence***

Recall the error for a hypothesis $f_{\overline{z}} = \underset{f \in \mathcal{H}}{\operatorname{argmin}}\, \mathcal{E}_{\overline{z}}(f)$ learned from data was defined:

$$\mathcal{E}(f_{\overline{z}}) = \mathcal{E}_{\mathcal{H}}(f_{\overline{z}}) + \mathcal{E}(f_{\mathcal{H}})$$

Where $f_{\mathcal{H}} = \underset{f \in \mathcal{H}}{\operatorname{argmin}}\, \mathcal{E}(f)$ and the **sampling error** $\mathcal{E}_{\mathcal{H}}(f_{\overline{z}})$ was defined:

$$\mathcal{E}_{\mathcal{H}}(f_{\overline{z}}) = \mathcal{E}(f_{\overline{z}}) - \mathcal{E}(f_{\mathcal{H}})$$

One of the most important questions in learning is whether the function $f_{\overline{z}}$ minimizing the empirical risk approaches to the function $f_{\mathcal{H}}$ minimizing the population risk in the sense that the sampling error $\mathcal{E}_{\mathcal{H}}(f_{\overline{z}}) \to 0$ as $n \to \infty$. This is called **consistency** of the algorithm.

Consistency depends on the hypothesis set, and a modern research is still refining the necessary and sufficient conditions for the consistency of empirical risk minimization. Necessary and sufficient conditions include $\mathcal{H}$ being a Glivenko-Cantelli class of functions or having a finite V_γ dimension, and a sufficient condition for consistency is compactness of $\mathcal{H}$ (Poggio & Smale, 2003).

Since $f_{\overline{z}}$ is a random variable, we can the express consistency requirement as a **learning bound**:

$$P[\mathcal{E}(f_{\overline{z}}) - \mathcal{E}(f_{\mathcal{H}})] \leq \eta(\epsilon, n)\ \forall \epsilon > 0, n \in \mathbb{N}$$

where $\eta(\epsilon, n)$ does not depend on ρ and $\eta(\epsilon, n) \to 0$ as $n \to \infty$.

However, to obtain such uniform estimates for the convergence, additional restrictions need to specified on the measure ρ. The problem is that even though there always exists a solution to the empirical minimization problem which is guaranteened to converge to the target function, the converge may be arbitrarily slow given an arbitrary probability measure ρ (Bauer;Pereverzev;& Rosasco, 2007).

5.6.3 ***Learning bounds for kernel RLS***

Next we specialize this for the RLS algorithm where the hypothesis space is an RKHS. Recalling the factorization $\mathcal{E}(f) = \left\|f - f_\rho\right\|_\rho + \mathcal{E}(f_\rho)$, the error of a hypothesis may be studied through the norm:

$$\mathcal{E}(f_{\overline{z}}) - \mathcal{E}(f_\rho) = \left\|f - f_\rho\right\|_\rho = \|f_{\overline{z}} - f_{\mathcal{H}}\|_\rho - \left\|f_{\mathcal{H}} - f_\rho\right\|_\rho$$

where $\|f_{\overline{z}} - f_{\mathcal{H}}\|_\rho$ corresponds to the sampling error and $\left\|f_{\mathcal{H}} - f_\rho\right\|_\rho$ to the approximation error. Alternatively, since the converge in $\mathcal{H}$ with respect to the RKHS norm $\|.\|_H$ implies pointwise converge, this norm is used in some analysis (Smale & Zhou, 2007).

However, the above analysis needs caution. In general, given a RKHS $\mathcal{H}$ the approximation error may be nonzero $\inf_{f\in\mathcal{H}} \mathcal{E}(f) > \mathcal{E}(f_\rho)$ and existence of the optimal approximator $f_{\mathcal{H}}$ is not even ensured. Denote $P: L^2(X,\rho) \to L^2(X,\rho)$ the projection into $\overline{\mathcal{R}(I_K)}$. If $\mathcal{H}$ is dense in $L^2(X,\rho)$ then $\inf_{f\in\mathcal{H}} \mathcal{E}(f) = \mathcal{E}(f_\rho)$ and we previously saw that necessary and sufficient condition for the existence and uniqueness of the best approximation minimum norm solution is $Pf_\rho \in \mathcal{R}(I_K)$. The solution then was then defined through the pseudoinverse $f_{\mathcal{H}}^\dagger = I_K^\dagger f_\rho$. In this case we can replace the analysis of $\left\|f - f_\rho\right\|_\rho$ by $\left\|f - f_{\mathcal{H}}^\dagger\right\|_\rho$ (Bauer;Pereverzev;& Rosasco, 2007). However, because the pseudoinverse is not continuous it does not result in a consistent algorithm.

Consider the regularized pseudoinverse and the corresponding minimization problem using a RKHS $\mathcal{H}$:

$$f_{\overline{z}}^\lambda = \operatorname*{argmin}_{f\in\mathcal{H}} \|S_{\overline{x}} f - \overline{y}\|_n^2 + \lambda\|f\|_H^2$$

$$f^\lambda = \operatorname*{argmin}_{f\in\mathcal{H}} \left\|I_k f - f_\rho\right\|_\rho + \lambda\|f\|_H^2$$

This algorithm has implicit sufficient requirements on $\mathcal{H}$ for consistency, based on the regularization and assumptions on the kernel, such as measurability, boundedness and seperability of $\mathcal{H}$ (Bauer;Pereverzev;& Rosasco, 2007). The conditions on ρ required for bounds are often more complicated (Caponetto & De Vito, Optimal Rates for the Regularized Least Squares Algorithm, 2007).

Different regularizion schemes can be analyzed by specifying general conditions on regularization which are sufficient for the consistency of learning. Tikhonov regularization, the regularized least squares formulation, satisfies these conditions and therefore results in a consistent algorithm (Bauer;Pereverzev;& Rosasco, 2007) (De Vito;Rosasco;& Verri, Spectral Methods for Regularization in Learning Theory, 2005).

In (Cucker & Smale, 2001) it is shown that for a fixed $\lambda > 0$, regularized least squares RKHS formulation specifies a compact hypothesis set $\overline{I_K(B_R)}$ where B_R is a ball of radius R in $\mathcal{H}$ and compactness is in $C(X)$. They also gave the following simple consistency bound derived through the Hoeffding inequality and covering numbers as a measure of an arbitrary hypothesis set $\mathcal{H}$:

Theorem

Suppose $|f(x) - y| \leq M$ for all $f \in \mathcal{H}$ for almost all $z \in Z$. Then

$$P_{z \in Z^m}[\mathcal{E}_{\mathcal{H}}(f_{\overline{z}}) \leq \epsilon] \leq 1 - \delta$$

where $\delta = Cov\#\left(\mathcal{H}, \frac{\epsilon}{24M}\right) e\text{^}(-\frac{m\epsilon}{288M^2})$.

However, not only does the regularization parameter imply a nonzero approximation error $\left\|f_\lambda - f_\rho\right\|_\rho$, we should furthermore consider a regularization parameter λ variable in the number of samples to obtain converge to the best approximator of f_ρ in the whole RKHS $\mathcal{H}$. Both in practise and in theory, this implies an intricate balancing between the sampling and approximation errors (Poggio & Smale, 2003).

The most general learning bounds considering both the sampling error and approximation error with possibly variable λ can be very complicated. Consider the bound, where with probability $1 - \delta$:

$$\mathcal{E}(f_{\overline{z}}) - \mathcal{E}(f_{\mathcal{H}}) \leq B(\delta, \mathcal{H}, n)$$

In these bounds, the analysis focuses especially on specifying and improving the constant in the bound and the convergence rate as a function of the number of samples m. The constant and the convergence rate are both variable, depending on the algorithm, the hypothesis set and restrictions on ρ. For an example of RLS algorithm

with assumptions resulting in $O\left(\frac{1}{\sqrt{m}}\right)$ rate, see (Caponetto, Optimal Rates for Regularization Operators in Learning Theory, 2006).

The tightness of the bound depends on the properties of the hypothesis space, which has been previously quantified with VC dimension (Poggio & Smale, 2003), Rademacher averages (Steinwart & Christmann, 2008), Covering numbers (Cucker & Smale, 2001), and recently the effective dimension (Zhang, 2005) . Recall we defined the effective dimension, a property of the RLS estimator using kernel $k(x, y)$, as follows:

$$D_\lambda(T_k) = \sum_i \frac{\lambda_i}{\lambda_i + \lambda}$$

where λ_i are the eigenvalues of the integral operator T_k associated to kernel k and λ is the regularization parameter, obtained as the trace of the regularized population error minimizer. Effective dimension is widely used in learning bounds in the kernel methods literature, because it allows both a simple interpretation, small constant factors and optimal convergence rates (Caponetto & De Vito, Optimal Rates for the Regularized Least Squares Algorithm, 2007). In our investigation, the simplicity is very useful for comparing different hypothesis spaces and considering modifications of a hypothesis space obtained by composition or constraints.

For a general effective dimension based learning bound, see (Caponetto & De Vito, Optimal Rates for the Regularized Least Squares Algorithm, 2007), recent advancements are discussed in (Hsu;Kakade;& Zhang, 2010)

6 Learning relations

6.1 Introduction to relations

Relation is a relationship between objects which may be considered for each pair of objects $v, v' \in \mathcal{V}$. Assume the relation is quantifiable as a function $Q(v, v')$ over the Cartesian product $(v, v') \in \mathcal{V} \times \mathcal{V}$. Then in machine learning we observe samples from the underlying relation as pairs of objects coupled with an output $\mathcal{D} = \{((v, v')_i, y_i)\}_{i=1}^n$ where $y_i = Q(v, v')$. The goal is to model the underlying relation $Q(v, v')$ based on these samples. In (Waegeman, ym., 2012) quantifiable relations are classified based on output

- $Q: \mathcal{V}^2 \to \{0,1\}$ is called a crisp relation
- $Q: \mathcal{V}^2 \to [0,1]$ graded relation, which may extended to $\mathbb{R}$ by scaling
- $Q: \mathcal{V}^2 \to \mathbb{N}$ as an injective map is an ordinal relation, where only the order is considered

Another way to classify relations is by the properties they have. In particular (Waegeman, ym., 2012) considers the following types of relations

- $Q(v, v') = Q(v', v)$ is a symmetric relation.
- $Q(v, v') = -Q(v', v)$ is an anti-symmetric relation.
- $Q(v, v') = 1 - Q(v', v)$ for Q in $[0,1]$ is a reciprocal relation.

Since is possible to transform between anti-symmetric and reciprocal relation through a scaling operation, often one of them is taken as the focus of a paper, as in (Pahikkala;Waegeman;Tsivtsivadze;Salakoski;& De Baets, 2010).

In a kernel based relation learning algorithm we have a kernel $k: \mathcal{V}^2 \times \mathcal{V}^2 \to \mathbb{R}$ defined between pairs $v, \overline{v} \in \mathcal{V}^2$. A simple notational convenience is to write this kernel as a four argument **pairwise kernel**:

$$k\left((v, v'), (\overline{v}, \overline{v}')\right) = k(v, v', \overline{v}, \overline{v}')$$

Our discussion of an abstract kernel $k: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ encompass this scenario with $\mathcal{X} = \mathcal{V}^2$. However, pairwise kernels allow special considerations based on the properties of the relation or the objects being predicted. In the simplest case we are pro-

vided a kernel for the pairs or a feature mapping for a pair, which is simply the standard kernel setting. However, we may impose extra constraints defined only for relations, or build up the feature mapping for the pair from the feature mappings for the objects. Examples or relation specific constraints are the aforementioned symmetry/anti-symmetry, and a widely used technique for building a relation-predictor from object features is the Kronecker kernel. These considerations are incorporated in (Pahikkala;Waegeman;Tsivtsivadze;Salakoski;& De Baets, 2010).

6.2 Symmetric and anti-symmetric kernel

In this section we specify restrictions on a pairwise kernel with the goal of restricting the learning to only symmetric and anti-symmetric relations.

Let $\Psi(v, v')$ be a feature mapping on $\mathcal{V}^2$. Consider the symmetric and anti-symmetric parts

$$\Psi(v, v') = \frac{1}{2}\big(\Psi(v, v') + \Psi(v', v)\big) + \frac{1}{2}\big(\Psi(v, v') - \Psi(v', v)\big)$$

Take as new features the projections into the symmetric and anti-symmetric parts

$$\Phi_{\mathrm{S}}(v, v') = \frac{1}{2}\big(\Psi(v, v') + \Psi(v', v)\big)$$

$$\Phi_{\mathrm{A}}(v, v') = \frac{1}{2}\big(\Psi(v, v') - \Psi(v', v)\big)$$

Given a kernel $K^{\Psi} = \langle \Psi(v, v'), \Psi(\overline{v}, \overline{v}') \rangle$, define as new kernels the **symmetric kernel** K_S^{Φ} and the **anti-symmetric kernel** K_A^{Φ} through these features:

$$\begin{aligned}
K_S^{\Phi}(v, v', \overline{v}, \overline{v}') &= \langle \Phi_{\mathrm{S}}(v, v'), \Phi_{\mathrm{S}}(\overline{v}, \overline{v}') \rangle \\
&= \frac{1}{4}\big\{\langle \Psi(v, v'), \Psi(\overline{v}, \overline{v}') \rangle + \big\langle \Psi(v, v'), \Psi\big(\overline{v}', \overline{v}\big) \big\rangle + \langle \Psi(v', v), \Psi(\overline{v}, \overline{v}') \rangle \\
&\quad + \big\langle \Psi(v', v), \Psi\big(\overline{v}', \overline{v}\big) \big\rangle \big\} \\
&= \frac{1}{4}\big\{K^{\Psi}(v, v', \overline{v}, \overline{v}') + K^{\Psi}\big(v, v', \overline{v}', \overline{v}\big) + K^{\Psi}\big(v', v, \overline{v}, \overline{v}'\big) \\
&\quad + K^{\Psi}\big(v', v, \overline{v}', \overline{v}\big)\big\} \\
K_A^{\Phi}(v, v', \overline{v}, \overline{v}') &= \langle \Phi_{\mathrm{A}}(v, v'), \Phi_{\mathrm{A}}(\overline{v}, \overline{v}') \rangle \\
&= \frac{1}{4}\big\{\langle \Psi(v, v'), \Psi(\overline{v}, \overline{v}') \rangle - \big\langle \Psi(v, v'), \Psi\big(\overline{v}', \overline{v}\big) \big\rangle - \langle \Psi(v', v), \Psi(\overline{v}, \overline{v}') \rangle \\
&\quad + \big\langle \Psi(v', v), \Psi\big(\overline{v}', \overline{v}\big) \big\rangle \big\} \\
&= \frac{1}{4}\big\{K^{\Psi}(v, v', \overline{v}, \overline{v}') - K^{\Psi}\big(v, v', \overline{v}', \overline{v}\big) - K^{\Psi}\big(v', v, \overline{v}, \overline{v}'\big) \\
&\quad + K^{\Psi}\big(v', v, \overline{v}', \overline{v}\big)\big\}
\end{aligned}$$

Next consider their sum, called the **permutation invariant** kernel:

$$K_{PI}(v, v', \overline{v}, \overline{v}') = K_S(v, v', \overline{v}, \overline{v}') + K_A(v, v', \overline{v}, \overline{v}')$$
$$= \frac{1}{2}\left(K^{\Psi}(v, v', \overline{v}, \overline{v}') + K^{\Psi}(v', v, \overline{v}', \overline{v})\right)$$

Even though the above kernels were defined using the feature map, this was only an intermediate step. The final kernel depended only on the given pairwise kernel K^{Ψ}. Given an kernel $K: \mathcal{V}^2 \times \mathcal{V}^2 \to \mathbb{R}$, these may be considered as operator projections $\Gamma_{\mathrm{S}}(K) = K_S$, $\Gamma_{\mathrm{A}}(K) = K_A$ and $\Gamma_{\mathrm{PI}}(K) = K_{PI}$. In this paper these specific kernels are referred to as the symmetric, anti-symmetric and permutation invariant kernels of a given kernel K. Given an arbitrary kernel, we may define these properties as follows (Pahikkala, Viljanen, Airola, & Waegeman, 2015):

- A kernel $K_S(v, v', \overline{v}, \overline{v}')$ is a symmetric pairwise kernel if $K_S(v, v', \overline{v}, \overline{v}') = K_S(v', v, \overline{v}, \overline{v}')$
- A kernel $K_A(v, v', \overline{v}, \overline{v}')$ is a anti-symmetric pairwise kernel if $K_A(v, v', \overline{v}, \overline{v}') = -K_A(v', v, \overline{v}, \overline{v}')$
- A kernel $K_{PI}(v, v', \overline{v}, \overline{v}')$ is permutatation invariant if $K_{PI}(v, v', \overline{v}, \overline{v}') = K_{PI}(v', v, \overline{v}', \overline{v})$

6.3 Examples of pairwise kernels

The anti-symmetric kernel was originally introduced in (Pahikkala;Waegeman;Tsivtsivadze;Salakoski;& De Baets, 2010) where the authors contrasted two methods for learning pairwise relations: Kronecker kernel and transitive pairwise kernel. Using the previous definition, anti-symmetric Kronecker kernel was then formed from a given Kronecker kernel to be able to learn intransitive pairwise relations. Both kernels can be viewed in a unified theoretical framework as anti-symmetrizations of different kernels or feature mappings. Their resulting forms have shared properties which motivate our theoretical considerations of the symmetrisation/anti-symmetrization of a general abstract kernel. Our main learning theoretic result about the effects of symmetrisation/anti-symmetrisation process therefore applies to both kernels.

6.3.1 ***Transitive pairwise kernel***

Given a feature representation $\phi(v)$ or a kernel K^{Φ} for the objects $v \in \mathcal{V}$, define a feature mapping for the relation:

$$\Psi(v, v') = \phi(v)$$

The kernel for the feature mapping is then defined as the object kernel:

$$K^{\Psi}(v, v', \overline{v}, \overline{v}') = K^{\Phi}(v, \overline{v})$$

Consider the effect of anti-symmetrization on this kernel:

$$K_A^{\Psi}(v, v', \overline{v}, \overline{v}') = \frac{1}{4}\left\{K^{\Phi}(v, \overline{v}) - K^{\Phi}\left(v, \overline{v}'\right) - K^{\Phi}(v', \overline{v}) + K^{\Phi}\left(v', \overline{v}'\right)\right\}$$

The predictor becomes:

$$\mathrm{h}(x, x') = \langle w, \Psi(x, x') - \Psi(x', x)\rangle = \langle w, \phi(v)\rangle - \langle w, \phi(v')\rangle = f(v) - f(v')$$

This particular form for the predictor has consequences for the type of relation one can learn. In (Pahikkala;Waegeman;Tsivtsivadze;Salakoski;& De Baets, 2010) the authors recollected the following definitions and the following lemma:

Definition

A function $Q: X^2 \to [0,1]$ is called a reciprocal relation if

$$\forall (x, x') \in X^2 \; Q(x, x') + Q(x', x) = 1$$

Definition

A reciprocal relation $Q: X^2 \to [0,1]$ is called strongly ranking representable if there exists a function $f: X \to \mathbb{R}$ and a cumulative distribution function $G: \mathbb{R} \to [0,1]$ with $G(0) = \frac{1}{2}$:

$$g\left(f(x), f(x')\right) = G\left(f(x) - f(x')\right)$$

Definition

A reciprocal relation $Q: X^2 \to [0,1]$ is called strongly stochastically transitive if

$$\forall \left(x_i, x_j, x_k\right) \in X^3 \; \left(Q\left(x_i, x_j\right) \geq \frac{1}{2} \wedge Q\left(x_j, x_k\right) \geq \frac{1}{2}\right) \Longrightarrow Q(x_i, x_k) \geq \max\left(Q\left(x_i, x_j\right), Q\left(x_j, x_k\right)\right)$$

Lemma

If $Q: X^2 \rightarrow [0,1]$ is strongly ranking representable, then it is strongly stochastically transitive.

This means that this practicular kernel can only learn features which have this property, which can be a positive or negative consequence depending on the prior knowledge assumptions one should make. This was the authors motivation for presenting the following kernel.

6.3.2 ***Intransitive pairwise kernel***

Given a feature mapping $\phi(x)$ for the object, define a feature mapping for the relation

$$\Psi(x, x') = \phi(x) \otimes \phi(x')$$

where $\otimes$ denotes the Kronecker product. The kernel then becomes (Waegeman, ym., 2012):

$$\begin{aligned} K^{\Psi}\left(x_i, x_i', x_j, x_j'\right) &= \left\langle \phi(x_i) \otimes \phi(x_i'), \phi\left(x_j\right) \otimes \phi(x_j') \right\rangle \\ &= \left\langle \phi(x_i), \phi(x_j) \right\rangle \otimes \left\langle \phi(x_i'), \phi\left(x_j'\right) \right\rangle = K^{\phi}(x_i, x_j) K^{\phi}(x_i', x_j') \end{aligned}$$

Consider the effect of anti-symmetrization on this kernel:

$$K_A^{\Psi}\left(x_i, x_i', x_j, x_j'\right) = \frac{1}{2}\left(K^{\phi}\left(x_i, x_j\right) K^{\phi}\left(x_i', x_j'\right) - K^{\phi}\left(x_i', x_j\right) K^{\phi}\left(x_i, x_j'\right)\right)$$

The symmetric version was introduced originally in (Ben-Hur & Stafford Noble, 2005):

$$K_S^{\Psi}\left(x_i, x_i', x_j, x_j'\right) = \frac{1}{2}\left(K^{\phi}\left(x_i, x_j\right) K^{\phi}\left(x_i', x_j'\right) + K^{\phi}\left(x_i', x_j\right) K^{\phi}\left(x_i, x_j'\right)\right)$$

The authors (Waegeman, ym., 2012) proved that the Kronecker kernel K^{Ψ} can be used to learn arbitrary pairwise preference relation, given that the feature representation ϕ is powerful enough (K^{ϕ} universal on $\mathcal{V}$). Futhermore, they showed that K_A^{Ψ} can be used to learn arbitrary anti-symmetric pairwise preference relations. The expressibility of the hypothesis space induced by this kernel is considerable.

6.4 Approximation properties of symmetric and anti-symmetric kernels

Recalling the representer theorem, the learned function has the following expression:

$$h(v, v') = \langle w, \Phi(v, v') \rangle = \sum_{i=1}^{n} a_i K^{\Phi}(v, v', \overline{v}, \overline{v}')$$

Observing the expression of the predictor, this implies that

$$h(v', v) = h(v, v') \text{ if } K^{\Phi} \text{ is symmetric}$$
$$h(v', v) = -h(v, v') \text{ if } K^{\Phi} \text{ is anti-symmetric}$$

It is trivial to verify that for the symmetrized kernel K_S^{Φ} and anti-symmetrized kernel K_A^{Φ}

$$K_S^{\Phi}(v, v', \overline{v}, \overline{v}') = K_S^{\Phi}(v', v, \overline{v}, \overline{v}')$$

$$K_A^{\Phi}(v, v', \overline{v}, \overline{v}') = -K_A^{\Phi}(v', v, \overline{v}, \overline{v}')$$

These simple results imply the following corollary.

Corollary

The symmetric/anti-symmetric kernel can learn **only** symmetric/anti-symmetric relations.

The next important question is: what relations can they learn? This of course depends on the capabilities of the original kernel; the following definitions and approximation result give justification for their broad use.

Definition: universal kernel (Steinwart & Christmann, 2008)

A continuous kernel K on a compact metric space $\mathcal{X}$ is called universal if the RKHS induced by K is dense in the space $\mathcal{C}(\mathcal{X})$, of all continous functions $f: \mathcal{X} \to \mathbb{R}$. That is, for every function $f \in \mathcal{C}(\mathcal{X})$ and every $\epsilon > 0$, there exists a set of input points $\{x_i\}_{i=1}^{m} \in \mathcal{X}$ and real numbers $\{\alpha_i\}_{i=1}^{m}$, with $m \in \mathbb{N}$, such that

$$\max_{x \in V} \left\{ \left| f(x) - \sum_{i=1}^{m} \alpha_i K(x_i, x) \right| \right\} \leq \epsilon$$

Accordingly, the hypothesis space induced by the kernel K can approximate any function in $\mathcal{C}(\mathcal{X})$ arbitrarily well, and hence it is said to have the universal approximating property.

Definition: subset kernel

Let K be a continous kernel K on a compact metric space $\mathcal{X}$ and let $\mathcal{F} \subseteq \mathcal{C}(\mathcal{X})$. Denote $\mathcal{F} \subseteq \mathcal{H}(K)$ if the RKHS induced by K is dense in the subset $\mathcal{F}$. That is, for every function $f \in \mathcal{F}$:

$$\forall \epsilon > 0 \; \exists \{x_i\}_{i=1}^m \subseteq \mathcal{X}, \{\alpha_i\}_{i=1}^m \subseteq \mathbb{R}: \max_{x \in \mathcal{X}} \left\{ \left| f(x) - \sum_{i=1}^m \alpha_i K(x_i, x) \right| \right\}$$

This means that hypothesis space induced by kernel K can approximate any function in $\mathcal{F}$ arbitrary well.

Theorem (Pahikkala, Viljanen, Airola, & Waegeman, 2015)

Let $\mathcal{F}$ be an arbitrary set of continous functions and let

$$\mathcal{S} = \{t | r \in \mathcal{F}, t(v, v') = r(v, v') + r(v', v)\}$$
$$\mathcal{A} = \{t | r \in \mathcal{F}, t(v, v') = r(v, v') - r(v', v)\}$$

Be the set of symmetric and anti-symmetric functions determined by $\mathcal{F}$.

Let K be a kernel on $\mathcal{F}^2$ and K^S and K^A be the corresponding symmetric and anti-symmetric kernels. Then

$$\mathcal{F} \subseteq \mathcal{H}(K) \Longrightarrow \mathcal{S} \subseteq \mathcal{H}(K^S), \mathcal{A} \subseteq \mathcal{H}(K^A)$$

Proof

For a rather long and technical proof, see (Pahikkala, Viljanen, Airola, & Waegeman, 2015).

6.5 Kernel matrices of symmetric and anti-symmetric kernels

Next we will provide the kernel matrix and integral operator forms of the predictors for the symmetrized and anti-symmetrized kernels. These kernel matrices are considered for machine learning in (Pahikkala;Airola;Stock;De Baets;& Waegeman, 2013) and they are more widely used in conjunction with Kronecker type matrix products. For an extensive discussion of their properties and use in linear algebra see the book by (Abadir & Magnus, 2005).

Definition

The **vectorizing operator** is a linear operator stacking the columns of $s \times s$ matrix into an s^2 column vector:

$$\mathrm{vec}(M) = \left(M_{1,1}\, M_{2,1} \ldots M_{s,1}\, M_{1,2} \ldots M_{s,s}\right)^T$$

Definition

Let A be a real $m \times n$ matrix. Define matrix P_{mn} $(mn \times mn)$, called the **commutation matrix**, such that

$$P_{mn}vec(A) = vec(A^T)\, P_{mn}$$

When $m = n$, we write $P_{mn} = P_n$. Often the subscript is clear from the context and it is omitted entirely.

Definition

Define the **symmetrizer** S_n $(n^2 \times n^2)$ and **skew-symmetrizer** A_n $(n^2 \times n^2)$:

$$S_n vec(A) = vec\left(\frac{1}{2}(A + A^T)\right)$$

$$A_n vec(A) = vec\left(\frac{1}{2}(A - A^T)\right)$$

Lemma

It is easily verified that

$$S_n = \frac{1}{2}(I_{n^2} + P_n)$$

$$A_n = \frac{1}{2}(I_{n^2} - P_n)$$

The following properties also follow straightforwardly:

$$P^T = P$$

$$PP = I$$

$$SS = S \text{ and } AA = A$$

$$S^T = S \text{ and } A^T = A$$

$$SA = 0$$

These abstract properties are equivalent to the following characterization: P is a permutation matrix and S, A are orthogonal projection matrices which are orthogonal to each other.

Lemma (Abadir & Magnus, 2005)

For $M, N \in \mathbb{R}^{s \times t}$, the permutation matrix has a special application as the commuter of the Kronecker product:

$$P^{s^2}(M \otimes N) = (N \otimes M)P^{t^2}$$

Lemma (Abadir & Magnus, 2005)

The Kronecker product has the following property:

$$(A \otimes B)(C \otimes D) = (AC) \otimes (AB)$$

Consider an abstract pairwise kernel matrix $K^{\Phi}(e_i, e_j) \coloneqq K^{\Phi}(v_x, v_y, v_s, v_t)$ indexed as follows:

K^{Φ}		v_1			…	v_n		
		v_1	…	v_n		v_1	…	v_n
v_1	v_1							
	…							
	v_n							
…					$K^{\Phi}(v_x, v_y, v_s, v_t)$			
v_n	v_1							
	…							
	v_n							

Then we can address it by tensor products of the unit basis vectors:

$$K^{\Phi}(v_x, v_y, v_s, v_t) = (e_x^T \otimes e_y^T) K^{\Phi} (e_s \otimes e_t)$$

Theorem

Given a kernel matrix K^{Φ} indexed as above, the symmetric and anti-symmetric matrices may be expressed:

$$K_S^{\Phi} = SK^{\Phi}S$$
$$K_A^{\Phi} = AK^{\Phi}A$$
$$K_{PI}^{\Phi} = \frac{1}{2}(SK^{\Phi}S + AK^{\Phi}A)$$

Proof

For full generality let K^{Φ} be a $ms \times nt$ matrix

$$A_n K^{\Phi} A_n(v, v', \overline{v}, \overline{v}') = (e_v^T \otimes e_{v'}^T) A_n K^{\Phi} A_n (e_{\overline{v}} \otimes e_{\overline{v}'})$$

$$= \frac{1}{4}(e_v^T \otimes e_{v'}^T)(I_{n^2} - K_{ms}) K^{\Phi} (I_{n^2} - K_{nt})(e_{\overline{v}} \otimes e_{\overline{v}'})$$

$$= \frac{1}{4}\{(e_v^T \otimes e_{v'}^T) K^{\Phi}(e_{\overline{v}} \otimes e_{\overline{v}'}) - (e_v^T \otimes e_{v'}^T) K^{\Phi} K_{nt}(e_{\overline{v}} \otimes e_{\overline{v}'})$$
$$- (e_v^T \otimes e_{v'}^T) K_{ms} K^{\Phi}(e_{\overline{v}} \otimes e_{\overline{v}'}) + (e_v^T \otimes e_{v'}^T) K_{ms} K^{\Phi} K_{nt}(e_{\overline{v}} \otimes e_{\overline{v}'})\}$$

$$= \frac{1}{4}\{(e_v^T \otimes e_{v'}^T) K^{\Phi}(e_{\overline{v}} \otimes e_{\overline{v}'}) - (e_v^T \otimes e_{v'}^T) K^{\Phi}(e_{\overline{v}'} \otimes e_{\overline{v}})$$
$$- (e_{v'}^T \otimes e_v^T) K^{\Phi}(e_{\overline{v}} \otimes e_{\overline{v}'}) + (e_{v'}^T \otimes e_v^T) K^{\Phi}(e_{\overline{v}'} \otimes e_{\overline{v}})\}$$

$$= \frac{1}{4}\{K^{\Phi}(v, v', \overline{v}, \overline{v}') - K^{\Phi}(v, v', \overline{v}', \overline{v}) - K^{\Phi}(v', v, \overline{v}, \overline{v}') + K^{\Phi}(v', v, \overline{v}', \overline{v})\}$$

The case for $S_n K^{\Phi} S_n$ is equivalent. The permutation invariant kernel follows directly from the definition.

Note that even though we used the Kronecker product to define an indexing scheme, this is unrelated to the Kronecker kernel. The kernel indexed as before is arbitrary, and the indexing was used to define the symmetric and anti-symmetric kernel matrices using the symmetrizer/anti-symmetrizer matrices applied to the original kernel matrix.

Eigensystem of the permutation invariant kernel

It is easily verified that the matrices $K_S^{\Phi}, K_A^{\Phi}, K_{PI}^{\Phi}$ commute. This implies that they share eigenvectors. Using diagonalization of self-adjoint matrices, they can be written:

$$K_S^{\Phi} = V \Lambda^S V^*$$
$$K_A^{\Phi} = V \Lambda^A V^*$$
$$K_{PI}^{\Phi} = V \Lambda^{PI} V^*$$

Lemma

In the Kronecker kernel case $K^{\Phi} = K^{\phi} \otimes K^{\phi}$, where $K^{\phi}(v, v')$ contains the node evaluations.

Proof

$$K^{\phi} \otimes K^{\phi}(v, v', \overline{v}, \overline{v}') = (e_v^T \otimes e_{v'}^T)(K^{\phi} \otimes K^{\phi})(e_{\overline{v}} \otimes e_{\overline{v}'})$$
$$= (e_v^T K^{\phi}) \otimes (e_{v'}^T K^{\phi})(e_{\overline{v}} \otimes e_{\overline{v}'}) = (e_v^T K^{\phi} e_{\overline{v}}) \otimes (e_{v'}^T K^{\phi} e_{\overline{v}'})$$
$$= (e_v^T K^{\phi} e_{\overline{v}})(e_{v'}^T K^{\phi} e_{\overline{v}'}) = K^{\phi}(v, \overline{v}) K^{\phi}(v', \overline{v}')$$

Theorem

If K^{Φ} is a Kronecker kernel:

$$K_S^{\Phi} = SK^{\Phi}$$
$$K_A^{\Phi} = AK^{\Phi}$$

Proof

Because $P_n\left(K^{\phi} \otimes K^{\phi}\right) = \left(K^{\phi} \otimes K^{\phi}\right)P_n$ and $P_n^2 = I$, we have:

$A\left(K^{\phi} \otimes K^{\phi}\right)A = \frac{1}{4}\left(K^{\phi} \otimes K^{\phi} - \left(K^{\phi} \otimes K^{\phi}\right)P_n - P_n\left(K^{\phi} \otimes K^{\phi}\right) + P_n\left(K^{\phi} \otimes K^{\phi}\right)P_n\right) = \frac{1}{2}\left(K^{\phi} \otimes K^{\phi} - P_n\left(K^{\phi} \otimes K^{\phi}\right)\right) = A\left(K^{\phi} \otimes K^{\phi}\right)$

Same applies for the symmetric matrix.

Theorem

If K^{Φ} is a Kronecker kernel, it is equivalent to its symmetric and anti-symmetric part, i.e. it is permutation invariant kernel:

$$K^{\Phi} = K_{PI}^{\Phi} = K_S^{\Phi} + K_A^{\Phi}$$

This makes the Kronecker kernel and it's spectrum relative to the symmetric/anti-symmetric parts particularly simple.

Proof

This follows from the same property $P_n\left(K^{\phi} \otimes K^{\phi}\right) = \left(K^{\phi} \otimes K^{\phi}\right)P_n$ combined with $P_n^2 = I$.

$$K_S^{\Phi} + K_A^{\Phi} = \frac{1}{2}(K^{\Phi} + PK^{\Phi}P) = K^{\Phi}$$

Above we showed that if the kernel is permutation invariant, such as the Kronecker kernel is, the eigensystem is particularly simple. If kernel is not permutation invariant, there does not seem to be a straightforward statement to make about the eigensystem. This motivates the separate abstract consideration for the kernel not assumed to be permutation invariant.

6.6 Integral operators of symmetric and anti-symmetric kernels

This section uses the following definitions and theorem for the spectral analysis of the permutation invariant kernel.

Definition (Majorization)

Let $\overline{r} = (r_i)_{i=1}^{\infty}$ and $\overline{s} = (s_i)_{i=1}^{\infty}$ be sequences decreasing monotonically to 0. Then $\overline{s}$ majorizes $\overline{r}$, denoted $\overline{r} \prec \overline{s}$, if

$$\sum_{i=1}^{m} r_i \leq \sum_{i=1}^{m} s_i \ \forall m \in \mathbb{N} \text{ and } \sum_{i=1}^{\infty} r_i = \sum_{i=1}^{\infty} s_i$$

For two trace class operators T_1 and T_2 on a Hilbert space with sequences of eigenvalues $\overline{\lambda_1}$ and $\overline{\lambda_2}$, denote

$$T_1 \prec T_2 \text{ if } \overline{\lambda_1} \prec \overline{\lambda_2}$$

Definition (Doubly-stochastic operation)

Let $\mathcal{T}(\mathcal{L})$ be the Banach space of all trace class operators on a Hilbert space $\mathcal{L}$, i.e. space of operators complete with respect to norm $\|T\|_1 = trace(T) < \infty$. The operation $\Gamma: \mathcal{T}(\mathcal{L}) \to \mathcal{T}(\mathcal{L})$ is doubly stochastic if it is trace-preserving ($trace(T) = trace(\Gamma(T))$), unital ($I = \Gamma(I)$) and there exists a sequence $\{E_i\}_{i=1}^{\infty}$ of compact operators such that $\Gamma(T) = \sum_{i=1}^{\infty} E_i T E_i^*$.

Theorem (Uhlmann's theorem for infinite dimensional Hilbert spaces)

If T_1 and T_2 are trace-class operators on a Hilbert space, then $T_1 \prec T_2$ iff there exists a doubly-stochastic operation Γ such that $T_2 = \Gamma(T_1)$.

Theorem (Integral operators) (Pahikkala, Viljanen, Airola, & Waegeman, 2015)

Let K be an arbitrary pairwise kernel and K_{PI}, K_S, K_A be its permutation invariant, symmetric and anti-symmetric forms. Let $T_K, T_{K_{PI}}, T_{K_S}, T_{K_A}$ be the integral operators of the corresponding kernels. Then

$$T_{K_P} = P^{\mu *} T_K P^{\mu}$$
$$T_{K_S} = S^{\mu *} T_K S^{\mu}$$
$$T_{K_A} = A^{\mu *} T_K A^{\mu}$$
$$T_{K_{PI}} = \frac{1}{2}(T_K + P^{\mu *} T_K P^{\mu}) = S^{\mu *} T_K S^{\mu} + A^{\mu *} T_K A^{\mu}$$

Where $P^{\mu}: L^2(\mathcal{P}^2, \mu) \to L^2(\mathcal{P}^2, \mu)$ is a permutation operator with respect to measure μ:

$$P^{\mu}: h(\overline{v}, \overline{v}') \mapsto \frac{\mu(\overline{v}', \overline{v})}{\mu(\overline{v}, \overline{v}')} h(\overline{v}', \overline{v})$$

$$P^{\mu*}: h(\overline{v}, \overline{v}') \mapsto \frac{\mu(\overline{v}, \overline{v}')}{\mu(\overline{v}', \overline{v})} h(\overline{v}', \overline{v})$$

And the symmetrizer and anti-symmetrizer with respect to the measure μ can be written:

$$S^{\mu} = \frac{1}{2}(I + P^{\mu})$$
$$A^{\mu} = \frac{1}{2}(I - P^{\mu})$$

Proof

For the long proof see (Pahikkala, Viljanen, Airola, & Waegeman, 2015).

An important special case is when the measure is symmetric

$$\mu(\overline{v}, \overline{v}') = \mu(\overline{v}', \overline{v}) \; \forall (\overline{v}, \overline{v}') \in \mathcal{P}^2$$

Theorem (Eigensystem analysis of symmetric μ) (Pahikkala, Viljanen, Airola, & Waegeman, 2015)

If μ is symmetric, the set of operators $\{S, A, T_{K_S}, T_{K_A}, T_{K_{PI}}\}$ commutes, which means they share an eigensystem:

$$T_{K_S} = V \Lambda^S V^*$$
$$T_{K_A} = V \Lambda^A V^*$$
$$T_{K_{PI}} = V \Lambda^{PI} V^*$$

In addition $T_{K_{PI}} = S T_K S + A T_K A$ and properties of S, A split the eigenvalues into two distinct blocks corresponding to symmetric and anti-symmetric functions

$$\Lambda^{PI} = \Lambda^S + \Lambda^A \text{ and } \Lambda^S \Lambda^A = 0$$

Finally, the sequence of eigenvalues of T_K majorizes the eigenvalues of $T_{K_{PI}}$:

$$T_{K_{PI}} \prec T_K$$

Proof

With symmetric μ, S and A are self-adjoint and hence orthogonal projections. Futhermore, they are orthonal to each other:

$$SA = AS = 0$$

We can define a trace preserving and unital operation $\Gamma: \mathcal{T}\left(L^2(\mathcal{P}^2, \mu)\right) \to \mathcal{T}(L^2(\mathcal{P}^2, \mu))$ with the operators $\left\{\frac{1}{\sqrt{2}} I, \frac{1}{\sqrt{2}} P\right\}$:

$$\Gamma: T_K \mapsto \frac{1}{2}(T_K + PT_KP)$$

Majorization then follows from Uhlmann's theorem.

6.7 Effective dimension of symmetric and anti-symmetric kernels

This consideration is split into two effective dimension considerations, both of which are used in the literature. The first is defined as the trace of the finite dimensional approximator based on the kernel matrix and the second as the trace of the trace-class integral approximator based on T_K.

Theorem (finite-dimensional general case)

The effective dimension of the symmetric and anti-symmetric kernel matrices are smaller:

$$D_\lambda(SKS) \leq D_\lambda(K)$$
$$D_\lambda(AKA) \leq D_\lambda(K)$$

Proof:

See **Appendix A**.

Theorem (infinite-dimensional general case) (Pahikkala, Viljanen, Airola, & Waegeman, 2015)

$$D_\lambda\left(T_{K_S}\right) \leq D_\lambda(T_K)$$
$$D_\lambda\left(T_{K_A}\right) \leq D_\lambda(T_K)$$

Proof

Since A^μ is a projection matrix, $A^\mu\left(L^2(\mathcal{P}^2,\mu)\right) \subseteq L^2(\mathcal{P}^2,\mu)$, the operator $A^{\mu^*}T_KA^\mu$ constrains the action of the integral operator T_K onto the range of A^μ, which is a subspace of $L^2(\mathcal{P}^2,\mu)$. The eigenfunctions ϕ_i associated with nonzero eigenvalues λ_i of T_{K_A} belong to this subspace, and are seen to satisfy:

$$T_K\phi_i - \lambda_i\phi_i = p \text{ with } p \perp A^\mu\left(L^2(\mathcal{P}^2,\mu)\right)$$

Since $A^\mu\left(L^2(\mathcal{P}^2,\mu)\right) \subseteq L^2(\mathcal{P}^2,\mu)$, we can use a well known theorem (see e.g. (Aronszajn, Rayleigh-Ritz and A. Weinstein methods for approximation of eigenvalues: I. Operations in a Hilbert space, 1948) and references therein) to obtain:

$$\lambda_i^{K^A} \leq \lambda_i^K \text{ for } i = 1,2,\ldots$$

Same considerations apply to S^{μ} so the case with T_{K_S} goes analogously: $\lambda_i^{K^S} \leq \lambda_i^K$ for $i = 1,2, \ldots$

Each term in the effective dimension is a monotonically increasing function of λ_i, from term-wise comparison the result follows.

Theorem (Permutation-invariance with symmetric μ)

If the kernel is permutation invariant, for the kernel matrix and the integral operator with symmetric measure μ, we previously showed $\Lambda^{PI} = \Lambda^S + \Lambda^A$ and $\Lambda^S \Lambda^A = 0$. Then we have a particularly simple result for the effective dimension

$$D_\lambda(T_K) = D_\lambda\left(T_{K_S}\right) + D_\lambda\left(T_{K_A}\right)$$

Theorem (Pahikkala, Viljanen, Airola, & Waegeman, 2015)

If the measure is symmetric

$$D_\lambda(T_K) \leq D_\lambda\left(T_{K_{PI}}\right)$$

Proof

A following result was recently proven by Mari et al. (2014).

Let $\overline{r} = (r_i)_{i=1}^{\infty}$ and $\overline{s} = (s_i)_{i=1}^{\infty}$ be sequences decreasing monotonically to 0 with $\sum_{i=1}^{\infty} r_i = \sum_{i=1}^{\infty} s_i = 1$. Then

$$\overline{r} \prec \overline{s} \Leftrightarrow \sum_{i=1}^{m} \rho(r_i) \geq \sum_{i=1}^{m} \rho(s_i)$$

For all real non-negative stricly concave functions ρ defined on the segment [0,1]. Since $\rho(r) = \frac{r}{r+\lambda}$ satisfies these properties, with scaling of the eigenvalues we have that

$$T_2 \prec T_1 \Longrightarrow D_\lambda(T_1) \leq D_\lambda(T_2)$$

And the result follows from previous majorization theorem.

Summarizing the results in chapters:

- $D_\lambda(K) = D_\lambda(SKS) + D_\lambda(AKA)$ if the kernel is permutation invariant.
- $D_\lambda(T_K) = D_\lambda\left(T_{K_S}\right) + D_\lambda\left(T_{K_A}\right)$ if the kernel is permutation invariant and μ is symmetric.

- $D_\lambda(T_K) \leq D_\lambda\left(T_{K_{PI}}\right)$ if μ is symmetric.
- $D_\lambda(SKS) \leq D_\lambda(K)$ and $D_\lambda(AKA) \leq D_\lambda(K)$ in any case.

References

Abadir, K.;& Magnus, J. (2005). *Matrix Algebra.* Cambridge University Press.

Abu-Mostafa, Y.;Magdon-Ismail, M.;& Lin, H.-T. (2012). *Learning from Data - A Short Course.* AMLBook.com.

Aronszajn. (1948). Rayleigh-Ritz and A. Weinstein methods for approximation of eigenvalues: I. Operations in a Hilbert space. *Proceedings of the National Academy of Sciences*, 474-480.

Aronszajn. (1948). Theory of Reproducing Kernels. *AMS*, 337-404.

Bauer, F.;Pereverzev, S.;& Rosasco, L. (2007). On regularization algorithms in learning theory. *Journal of complexity*.

Ben-Hur, A.;& Stafford Noble, W. (2005). Kernel methods for predicting protein–protein. *Bioinformatics*, i38–i46.

Caponetto, A. (2006). *Optimal Rates for Regularization Operators in Learning Theory.* MIT Computer Science and Artificial Intelligence Laboratory.

Caponetto, A.;& De Vito, E. (2007). Optimal Rates for the Regularized Least Squares Algorithm. *Foundations of Computational Mathematics*, 331-368.

Caponnetto, A.;Rosasco, L.;De Vito, E.;& Verri, A. (2005). *Empirical Effective Dimension and Optimal Rates for Regularized Least Squares Algorithm.* MIT CSAIL.

Cucker, F.;& Smale, S. (2001). On the Mathematical Foundations of Learning. *Bulletin of the American Mathematical Society*, 1-49.

De Vito, E.;Rosasco, L.;& Verri, A. (2005). *Spectral Methods for Regularization in Learning Theory.* Università di Genova.

De Vito, E.;Rosasco, L.;Caponetto, A.;De Giovannini, U.;& Odone, F. (2005). Learning from Examples as an Inverse Problem. *Journal of Machine Learning Research*, 883-904.

De Vito, E.;Rosasco, L.;Caponnetto, A.;Piana, M.;& Verri, A. (2004). Some Properties of Regularized Kernel Methods. *Journal of Machine Learning Research*, 1363-1390.

Decoste, D.;& Bernhard, S. (2002). Training Invariant Support Vector Machines. *Machine Learning*, 161-190.

Engl, H.;Hanke, M.;& Neubauer, A. (1996). *Regularization of Inverse Problems.* Kluwer Academic Publishers.

Haasdonk, B.;& Keysers, D. (2002). Tangent distance kernels for Support Vector Machines. *Proceedings of the 16th International Conference on Pattern Recognition*, (ss. 864-868).

Hein, M.;& Olivier, B. (2004). *Kernels, Associated Structures and Generalizations.* Max Planck Institute for Biological Cybernetics.

Hsu, D.;Kakade, S.;& Zhang, T. (2010). Random Design Analysis of Ridge Regression.

Kreyszig, E. (1989). *Introductory Functional Analysis with Applications.* John Wiley & Sons.

Lauer, F.;& Bloch, G. (2007). Incorporating prior knowledge in support vector machines for classification - A review. *Neurocomputing*, 1578-1594.

Mohri, M.;Rostamizadeh, A.;& Ameet, T. (2012). *Foundations of Machine Learning.* The MIT Press.

Moon, T. (1999). *Mathematical Methods and Algorithms for Signal Processing.* Prentice Hall.

Pahikkala, T., Viljanen, M., Airola, A., & Waegeman, W. (2015). *Spectral Analysis of Symmetric and Anti-Symmetric Pairwise Kernels.* arXiv. Retrieved from http://arxiv.org/abs/1506.05950

Pahikkala, T.;Airola, A.;Stock, M.;De Baets, B.;& Waegeman, W. (2013). Efficient Regularized Least-Squares Algorithms for Conditional Ranking on Relational Data. *Machine Learning*, 321-356.

Pahikkala, T.;Waegeman, W.;Tsivtsivadze, E.;Salakoski, T.;& De Baets, B. (2010). Learning intransitive reciprocal relations with kernel methods. *European Journal of Operational Research*, 676–685.

Poggio, T.;& Smale, S. (2003). The Mathematics of Learning: Dealing with Data. *Notices of the AMS*.

Poincaré, H. (1890). Sur les Equations aux Dérivées Partielles de la Physique Mathématique. *American Journal of Mathematics*, 211-294.

Rosasco, L.;Belkin, M.;& De Vito, E. (2010). On Learning with Integral Operators. *Journal of Machine Learning Research*, 905-934.

Schulz-Mirbach. (1994). Constructing invariant features by averaging techniques. *Proceedings of the 12th International Conference on Pattern Recognition*, (ss. 387–390). Jerusalem, Israel.

Schölkopf, B.;& Smola, A. (2001). *Learning with Kernels.* The MIT Press.

Smale, S.;& Zhou, D.-X. (2004). Shannon Sampling and Function reconstruction. *Bulletin of the American Mathematical Society*, 279-305.

Smale, S.;& Zhou, D.-X. (2005). Shannon sampling II: Connections to learning theory. *Applied and Computational Harmonic Analysis*, 285-302.

Smale, S.;& Zhou, D.-X. (2007). Learning Theory Estimates via Integral Operators and Their Approximations. *Constructive Approximation*, 153-172.

Steinwart, I.;& Christmann, A. (2008). *Support Vector Machines.* Springer.

Stenger, W.;& Alexander, W. (1972). *Methods of intermediate problems for Eigenvalues : theory and ramifications.* Academic Press.

Waegeman, W.;Pahikkala, T.;Airola, A.;Salakoski, T.;Stock, M.;& Baets, B. D. (2012). A kernel-based framework for learning graded relations from data. *IEEE Transactions on Fuzzy Systems*, 1090 - 1101.

Wolf, L.;& Shashua, A. (2003). Learning over sets using kernel principal angles. *The Journal of Machine Learning Research*, 913-931.

Zhang, T. (2005). Learning Bounds for Kernel Regression Using Effective Data Dimensionality. *Neural Computation*, 2077-2098.

Appendix A: Effective dimension

A.1 Proof of $D_\lambda(SKS) \leq D_\lambda(K)$

Geometrical interpretation of S and A

P is a symmetric matrix with $P^2 = I$ and thus it has eigenvalues -1 and 1. This means P has a spectral decomposition:

$$P = P_1 - P_{-1}$$

Because P_1 and P_{-1} are projection matrices to the eigenspaces of a symmetric matrix P we additionally have that

$$I = P_1 + P_{-1} \text{ and } P_1 P_{-1} = 0$$

Through this we arrive at a natural interpretation of S and A:

$$S \coloneqq \frac{1}{2}(I + P) = P_1$$
$$A \coloneqq \frac{1}{2}(I - P) = P_{-1}$$

This means S and A are orthogonal projections into orthogonal subspaces of V that span the space. These subspaces are spanned by eigenvectors of P associated with eigenvalues 1 and -1, respectively.

Denoting $V_S = im(S)$ and $V_A = im(A)$, we have $\ker(S) = V_A$, $\ker(A) = V_S$ and $V_S^\perp = V_A$.

This implies $V = V_S \oplus V_A$.

Eigenvectors of SKS

Next a simple result about the eigenvectors of SKS.

Write any vector as a sum of its components in each subspace: $x = x_s + x_a$ where $x_s \in V_S, x_a \in V_A$. Denote $x' = Kx$ and write $x' = x_s' + x_a'$.

Then $SKSx = SKx_s = S(x_s' + x_a') = x_s'$. We have $x_S' = \mu x \Leftrightarrow x_s' = \mu(x_s + x_a)$.

If $\mu \neq 0$, then $x_S' = \mu x_s$ and $x_a = 0$.

If $\mu = 0$, then $x_s' = 0$.

For eigenvectors of SKS associated with nonzero eigenvalues, the eigenvectors belong to V_S and K transforms x_s to μx_s in V_S with the component in V_A being unconstrained. For zero eigenvalues, K transforms x in V to V_A.

Nonzero eigenvalues of SKS as a principal submatrix problem

For the effective dimension we are only interested in nonzero eigenvalues. In what follows we transform this problem to another form where we can use a general theorem.

The eigenvectors of P form basis $B = \{x_{s1}, \dots, x_{sk}, x_{a(k+1)}, \dots, x_{an}\}$. Denote the standard basis as $E = \{e_1, \dots e_n\}$.

Express $x \in V_S$ in terms of B: $x = \alpha_1 x_{s1} + \cdots + \alpha_k x_{sk}$

For eigenvalues $\mu \neq 0$, we seek $\alpha_1, \dots, \alpha_k$ such that

$$Kx_s = \mu x_s + x'_a = \mu\alpha_1 x_{s1} + \cdots + \mu\alpha_k x_{sk} + \beta_{k+1} x_{a(k+1)} + \cdots + \beta_n x_{an}$$

We can write this in matrix form in terms of the basis B, where K_x denotes the transformation corresponding to K:

$$K_x \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_k \\ \overline{0} \end{pmatrix} = \begin{pmatrix} \mu\alpha_1 \\ \vdots \\ \mu\alpha_k \\ \overline{\beta} \end{pmatrix} \text{ coefficients w. r. t } \begin{bmatrix} x_{si} \\ x_{aj} \end{bmatrix}$$

Define $K_{(i,j,n,m)}$ as a submatrix (k_{xy}) where $x: i \dots n, y: j \dots m$. The above problem reduces to

$$\begin{pmatrix} k_{11} & \cdots & k_{1k} \\ \vdots & \ddots & \vdots \\ k_{k1} & \cdots & k_{kk} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_k \end{pmatrix} = \begin{pmatrix} \mu\alpha_1 \\ \vdots \\ \mu\alpha_k \end{pmatrix} \text{ and } \begin{pmatrix} k_{(k+1)1} & \cdots & k_{(k+1)k} \\ \vdots & \ddots & \vdots \\ k_{n1} & \cdots & k_{nk} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_k \end{pmatrix} = \begin{pmatrix} \beta_{k+1} \\ \vdots \\ \beta_n \end{pmatrix}$$

$\Leftrightarrow K_{(1,1,k,k)}\overline{\alpha} = \mu\overline{\alpha}$ and $K_{(1,k+1,k,n)}\overline{\alpha} = \overline{\beta}$

Since $\overline{\beta}$ is unconstrained, this is an eigenvalue and eigenvector problem of the principal submatrix:

$$K_{(1,1,k,k)}\overline{\alpha} = \mu\overline{\alpha}$$

Eigenvalues of the principal submatrix

Poincaré's separation theorem (Abadir & Magnus, 2005, s. 348):

Let A be $n \times n$ symmetric matrix with eigenvalues $\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_n(A)$ and let A_r be $r \times r$ principal submatrix of A with eigenvalues $\lambda_1(A_r) \geq \lambda_2(A_r) \geq \cdots \geq \lambda_r(A_r)$. Then $\lambda_k(A_r) \leq \lambda_k(A)$ for $k = 1 \ldots r$.

This theorem implies that for the eigenvalues of SKS, that is, the solutions to $K_{(1,1,k,k)}\,\overline{\alpha} = \mu\overline{\alpha}$ (principal submatrix of K_x): $\mu_i \leq \lambda_i'$

Define $T = (x_{s1}, \ldots, x_{sk}, x_{a(k+1)}, \ldots, x_{an})$, which is a basis transformation $B \to E$. Since the eigenvectors were orthogonal, we have $T^{-1} = T'$.

$K_x = TKT'$, which means that K_x is similar to K. Similar matrices share eigenvalues: $\lambda_i' = \lambda_i$ for all i.

Due to similarity, for eigenvalues λ_i of K: $\mu_i \leq \lambda_i$

The effective dimension

Each term $\frac{\lambda_i}{\lambda_i+\lambda}$ of the effective dimension is a monotonically increasing function in $\lambda_i \in [0, \infty)$ where $\lambda > 0$. Thefore $\frac{\lambda_x}{\lambda_x+\lambda} \geq \frac{\lambda_y}{\lambda_y+\lambda} \Leftrightarrow \lambda_x \geq \lambda_y$ $(\lambda_i \geq 0, \lambda > 0)$. For the eigenvalues of a PSD matrix, $\lambda_i \geq 0$.

Comparing the term expansions we have

$$D_\lambda(K) = D_\lambda(K_x) = \sum_{i=1}^{n} \frac{\lambda_i}{\lambda_i + \lambda} = \frac{\lambda_1}{\lambda_1 + \lambda} + \cdots + \frac{\lambda_r}{\lambda_r + \lambda} + \cdots + \frac{\lambda_n}{\lambda_n + \lambda}$$

$$D_\lambda(SKS) = D_\lambda\big(K_{(1,1,k,k)}\big) = \sum_{i=1}^{r} \frac{\mu_i}{\mu_i + \lambda} = \frac{\mu_1}{\mu_1 + \lambda} + \cdots + \frac{\mu_r}{\mu_r + \lambda}$$

where $\lambda_i \geq \mu_i \Longrightarrow \frac{\lambda_i}{\lambda_i+\lambda} \geq \frac{\mu_i}{\mu_i+\lambda},\ i = 1 \ldots r$ and $\frac{\lambda_i}{\lambda_i+\lambda} \geq 0,\ i = (r+1), \ldots n$.

Since every term is larger, we can conclude that

$$D_\lambda(SKS) \leq D_\lambda(K)$$

A.2 Effective dimension – a simple example

This section is based on the illustrative discussion in (Zhang, 2005).

Consider the a of observations x_i, associated signals $f(x_i)$ with i.i.d distributed zero mean, σ^2 variance noise $\{n_i\}$. For each observation, we observe a corrupted response

$$y_i = f(x_i) + n_i$$

The goal is to estimate $\hat{y}_i \approx f(x_i)$ such that the mean squared error to the true signal is small:

$$\|f - \hat{y}\|_m = \frac{1}{m}\sum_{i=1}^{m}(f(x_i) - \hat{y}_i)^2$$

Assume now that signal vector belongs to d -dimensional space, so that for $n \times d$ projection matrix P:

$$f = Pu$$

Assume P is orthogonal: $P^T P = I_d$. The best estimator projects the observation onto the signal subspace:

$$\hat{f} = PP^T y = PP^T(Pu + n) = Pu + PP^T n$$

For this estimator, the MSE is:

$$\left\|\hat{f} - f\right\|^2 = \frac{1}{n}\|PP^T n\|^2 = \frac{1}{n}(PP^T n)^T(PP^T n) = \frac{1}{n}n^T PP^T n = \frac{1}{n}\sum_i\sum_j n_i[PP^T]_{ij}n_j$$

We know that $E[n_i] = 0$, $E[n_i^2] = \sigma^2$ and $E[n_i n_j] = E[n_i]E[n_j] = 0$:

$$E_n\left[\left\|\hat{f} - f\right\|^2\right] = \frac{\sigma^2}{n}tr(PP^T) = \frac{d}{n}\sigma^2$$

In general, u may not belong to a fixed subspace. Extend the above analysis to any linear estimator $\hat{f} = Sy$:

$$\left\|\hat{f} - f\right\|^2 = \|S(f + n) - f\|^2 = \|Sf - f\|^2 + 2(Sf - f)^T n + n^T S^T S n$$
$$E_n\left[\left\|\hat{f} - f\right\|^2\right] = \|Sf - f\|^2 + \frac{\sigma^2}{n}tr(S^T S)$$

This is a bias-variance decomposition. A good estimator S optimizes the tradeoff; Sf is close to f and $tr(S^T S)$ is small.

For RLS kernel method, the estimator is

$$\hat{f} = (K + \lambda I)^{-1} K y = S y$$

Using the fact that (Zhang, 2005) $tr([K + \lambda I]^{-2} K^2) \leq tr([K + \lambda I]^{-1} K)$, expressing the trace as $\sum_i \frac{\lambda_i}{\lambda_i + \lambda}$, we have motivated the use of effective dimension as a measure of variance in this simple example.