Journal of Machine Learning Research 25 (2024) 1-47 Submitted 4/23; Revised 12/24; Published 12/24

Localisation of Regularised and Multiview Support Vector Machine Learning

Aurelian Gheondea A.GHEONDEA@IMAR.RO
Institute of Mathematics of the Romanian Academy
21 Calea Griviţei
010702 Bucharest, Romania
and
Department of Mathematics
Bilkent University
06800 Bilkent, Ankara, Turkey AURELIAN@FEN.BILKENT.EDU.TR

Cankat Tilki CANKAT@VT.EDU
Department of Mathematics and
Division of Computational Modeling and Data Analytics
Virginia Polytechnic Institute and State University
Blacksburg Virginia, 24061 U.S.A.

Editor: Chris Oates

Abstract

We prove some representer theorems for a localised version of a semisupervised, manifold regularised and multiview support vector machine learning problem introduced by H.Q. Minh, L. Bazzani, and V. Murino, *Journal of Machine Learning Research*, **17**(2016) 1–72, that involves operator valued positive semidefinite kernels and their reproducing kernel Hilbert spaces. The results concern general cases when convex or nonconvex loss functions and finite or infinite dimensional underlying Hilbert spaces are considered. We show that the general framework allows infinite dimensional Hilbert spaces and nonconvex loss functions for some special cases, in particular in case the loss functions are Gâteaux differentiable. Detailed calculations are provided for the exponential least squares loss functions that lead to systems of partially nonlinear equations for which some Newton's approximation methods based on the interior point method can be used. Some numerical experiments are performed on a toy model that illustrate the tractability of the methods that we propose.

Keywords: operator valued reproducing kernel Hilbert spaces, manifold co-regularised and multiview learning, support vector machine learning, loss functions, representer theorem

1. Introduction

Representer theorems are of a special interest in machine learning due to the fact that they reduce the problem of finding a minimiser for the learning map to the vector space spanned by the kernel functions, or operators, at the labeled and unlabeled input data. For classical

versions of representer theorem, we recommend the monographs of Schölkopf and Smola (2002) and Steinwart and Christmann (2008). There is a large literature on generalised representer theorems but in this article we refer to the unifying framework in vector valued reproducing kernel Hilbert spaces for semisupervised, manifold regularised and multiview machine learning, as investigated by Minh et al. (2016) and the vast literature cited there.

The article Minh et al. (2016) has remarkable contributions to the domain of representer theorems in support vector machine learning, firstly by unifying many variants of these theorems referring to semisupervised, regularised, manifold regularised, multiview machine learning and then by considering underlying Hilbert spaces that are infinite dimensional. Recently, infinite dimensional Hilbert spaces in learning with kernels have been of interest, e.g. see Lambert (2021). However, although the general representer theorem, Theorem 2 in Minh et al. (2016), is stated for infinite dimensional spaces, this turns out to be problematic, as we will see in Remark 10. Also, there is an interest for applications to learning problems in which loss functions may not be convex, cf. Zhao et al. (2010), or even indefinite, cf. Kwon and Zou (2023).

In this article we are concerned with questions triggered by the investigations in Minh et al. (2016) and Zhao et al. (2010), such as: to which extent can one allow the underlying input spaces be infinite dimensional and to which extent nonconvex loss functions can be used in the learning process. In this respect, we propose a localisation of the minimisation of the semisupervised regularised, manifold regularised, and multiview machine learning problem studied in Minh et al. (2016), in the sense that the output spaces and the loss functions may be different for each labeled and unlabeled input point. For this approach we use a generalised version of vector valued reproducing kernel Hilbert spaces with bundles of spaces and operators. We think that the localised framework offers more flexibility to the learning problem and it is quite natural, especially when semisupervised multiview learning is considered, that the output spaces and the loss functions depend locally on the input points.

There are a few reasons that motivate the localised versions that we consider in this article. Firstly, for some of the labeled input points, in the multivariable case, some of the components of the labels (properties) may be missing or some additional components of the labels may be necessary in order to allow reliable information. This means that the underlying vector spaces of labels may have different dimensions and hence, making the vector spaces of the labels depend on the input points solves this obstruction. Secondly, when the input set X shows a certain homogeneity, the localised version may not be needed but, when this set is more heterogeneous, the localised version brings the necessary flexibility. To be more precise, let us imagine the following scenario. The input set X is a finite union of sets X_i, $i = 1, \ldots, N$, where each X_i shows homogeneity. By homogeneity we mean that, for each i, the properties (labels) associated to $x \in X_i$ are of the same type, in particular they live in the same Hilbert space $\mathcal{Y}_i$. However, for different $i \neq j$, where $i, j \in \{1, \ldots, N\}$, the properties (labels) of the points $x \in X_i$ when compared to the properties of the points $x \in X_j$ may be different, meaning that the Hilbert spaces of labels $\mathcal{Y}_i$ and $\mathcal{Y}_j$ should be different. For example, if the set X is the collection of all the cells of a body, we can see X as the union of all its organs X_i, for $i = 1, \ldots, N$, and it is clear that due to special functions that different organs have in the body, the properties (labels) of cells in different organs are generally of different type. Moreover, the learning function, which is calculated

in terms of loss functions, may require different loss functions for different points that have different types of labels. For example, in the scenario of the body viewed as the union of its organs, comparison of properties of different cells should be performed differently for cells belonging to different organs. This justifies the dependence of the loss functions of the input points as well. As a further research project, the localised version of the machine learning problem might be used for investigating the process of spreading a malignant tumor (metastasis) in different organs of a body, provided that real data will be available and appropriate mathematical models will be obtained.

Following Minh et al. (2016), the direction of multiview learning we consider in this work is coregularisation, see e.g. Brefeld et al. (2006), Sindhwani and Rosenberg (2008), Rosenberg et al. (2009), and Sun (2011). In this approach, different hypothesis spaces are used to construct target functions based on different views of the input data, such as different features or modalities, and a data dependent regularisation term is used to enforce consistency of output values from different views of the same input example. The resulting target functions, each corresponding to one view, are then naturally combined together in a certain fashion to give the final solution.

The direction of semisupervised learning we follow here is manifold regularisation, cf. Belkin et al. (2006), Brouard et al. (2011), and Minh and Sindwhani (2011), which attempts to learn the geometry of the input space by exploiting the given unlabeled data. The latter two papers are recent generalisations of the original scalar version of manifold regularisation of Belkin et al. (2006) to the vector valued setting. In Brouard et al. (2011), a vector valued version of the graph Laplacian L is used while in Minh and Sindwhani (2011) L is a general symmetric, positive operator, including the graph Laplacian. The vector valued setting allows one to capture possible dependencies between output variables by the use of, for example, an output graph Laplacian. For a comprehensive discussion on semisupervised learning and a thorough comparison with supervised and unsupervised learning, see the collection Chapelle et al. (2006).

Because reproducing kernel Hilbert spaces make an essential ingredient in these representer type theorems, some historical considerations are in order. The classical article of Aronszajn (1950) provides an abstract formalisation of scalar reproducing kernel Hilbert spaces of many previous investigations and applications related to spaces of analytic functions, partial differential equations, and harmonic analysis, as of Mercer (1909), Bergman (1922), Moore (Part I, 1935, Part II, 1939), Bochner (1941), and Godement (1948). From the point of view of probability theory, Kolmogorov (1941) investigated linearisations, or feature spaces, associated to scalar positive semidefinite kernels. An equivalent formulation of reproducing kernel Hilbert spaces by Hilbert spaces continuously embeded in a quasicomplete locally convex space was investigated by Schwartz (1964) while a more group theoretical approach was performed in Krein (1949/1963) and Krein (1950/1963). Then, scalar valued reproducing kernel Hilbert spaces found many applications in machine learning, see Steinwart and Christmann (2008) and Schölkopf and Smola (2002) for a comprehensive list of literature in this direction.

Motivated by problems in operator theory and operator algebras, operator valued positive semidefinite kernels and either their linearisations (Kolmogorov decompositions) or their reproducing kernel Hilbert spaces have been considered by Sz.-Nagy (1955), Pedrick (1957), Parthasarathy and Schmidt (1972), Evans and Lewis (1977), and Constantinescu

(1996), to cite a few. Investigations of operator valued Hermitian kernels that yield Krein spaces have been performed by Constantinescu and Gheondea (1997), Constantinescu and Gheondea (2001). More general operator valued positive semidefinite kernels that yield reproducing kernel VH-spaces have been considered by Gheondea (2012), Ay and Gheondea (2015), Ay and Gheondea (2017), and Ay and Gheondea (2019). To make things more precise, recall that a vector space $\mathcal{S}$ endowed with a map $\mathcal{S} \ni s \mapsto s^* \in \mathcal{S}$ which is conjugate linear and involutive, in the sense that $(s^*)^* = s$ for all $s \in \mathcal{S}$, is called a *$*$-vector space*. If, in addition, a convex cone $\mathcal{S}^+$ on $\mathcal{S}$ is specified such that for any $s \in \mathcal{S}^+$ we have $s = s^*$, we call $\mathcal{S}$ an ordered $*$-space and, when a complete locally convex topology on $\mathcal{S}$ is specified and which is related in a certain fashion with the convex cone $\mathcal{S}^+$, one calls $\mathcal{S}$ an *admissible space*. The concept of admissible space is a generalisation of the concept of C^*-algebra which is a mainstream domain in functional analysis, partly due to its strong connections with quantum theories, e.g. see Davidson (1996) and Blackadar (2006). Admissible spaces $\mathcal{S}$ are then used to consider gramians $[\cdot,\cdot]\colon \mathcal{X} \times \mathcal{X} \to \mathcal{S}$, for some vector space $\mathcal{X}$, which are vector valued inner products and that induce a certain topology on $\mathcal{X}$. If this topology is complete then we call $\mathcal{X}$ a VH-space (Vector Hilbert space). The concept of VH-space is a generalisation of the concept of Hilbert modules over C^*-algebras, e.g. see Lance (1995), but it appeared independently and related to problems in probability theory, see Loynes (1965a), Loynes (1965b), and Loynes (1967). In case we have a positive semidefinite kernel $K\colon X \times X \to \mathcal{S}$, for some admissible space $\mathcal{S}$, one can define a reproducing kernel VH-space as a generalisation of the classical reproducing kernel Hilbert space. For details and many examples see, for example, Ay and Gheondea (2017).

During the last twenty years, operator valued positive semidefinite kernels and their reproducing kernel Hilbert spaces and feature spaces (linearisations, or Kolmogorov decompositions) became of interest in the theory of machine learning, see Micchelli and Pontil (2005), Carmelli et al. (2006), Caponnetto et al. (2008), Minh and Sindwhani (2011), Kadri et al. (2016), but the investigations have been somehow started from scratch, apparently unaware of the previous works. More recently, in a sequence of articles, Hashimoto et al. (2021), Hashimoto et al. (2022), Hashimoto et al. (2023a), Hashimoto et al. (2023b), and Hashimoto et al. (2024), it is shown that, using the C^*-algebra-valued gramians (vector valued inner products), one can learn function- and operator-valued maps, one can design positive definite kernels for structured data using the noncommutative product, one can use the norm of the *-algebra to alleviate the dependency of generalisation error bound on the output dimension using the generalisation of kernel mean embedding by means of C^*-algebras, one can analyse positive operator valued measures and spectral measures, one can continuously combine multiple models and use the tools for functions which can be applied to ensemble, multitask, and meta-learning (the noncommutative product structures in C^*-algebras induce interactions among models), and one can construct group equivariant neural networks using the products in group C^*-algebras.

We briefly describe the contents of this article. In order to introduce the localised version of the semisupervised, regularised, manifold regularised, and multiview learning problem, in Subsection 2.1 we firstly consider operator valued positive semidefinite kernels for which the entries are localised by a bundle of Hilbert spaces. For these kernels, we show how their reproducing kernel Hilbert spaces are constructed, their relation to the linearisations (Kolmogorov decompositions, feature spaces) and basic properties. Although

we have been inspired by the approach in Constantinescu and Gheondea (1997), we provide in the appendices proofs for all the statements we make since in this form they cannot be found elsewhere. In Pedrick (1957), reproducing kernel Hilbert spaces are considered for kernels over bundles of locally convex spaces and this might be a very interesting research problem in case applications to machine learning theory might be found.

Then, in Subsection 2.2 we present the localised version of the semisupervised, regularised, manifold regularised, and multiview machine learning problem inspired by Minh et al. (2016). Under rather general assumptions, we prove in Subsection 2.1 the representer theorem for general loss functions but assuming the input spaces at all input points be finite dimensional. The finite dimensionality assumption can be relaxed to the condition that the span of the input kernel operators at the input points is closed, as Proposition 6 shows. In this section, most of the underlying Hilbert spaces may be complex and only in some special cases we have to impose the condition that they are real.

Further on, in Subsection 3.2 we consider real underlying Hilbert spaces and loss functions that are Gâteaux differentiable and then, in Theorem 17 we show that, under this additional assumption, a general representer theorem can be obtained for infinite dimensional input spaces. In Subsection 3.3 we work out the details for the loss functions defined by the least squares which leads to a linear problem in terms of the unkown coefficients, as in Minh et al. (2016), while in Subsection 3.4 we work out the details for the loss functions defined by the exponential least squares which lead to a mixed problem, linear and nonlinear. Finally, in Subsection 3.5 we tackle algorithms to obtain approximations of the solutions for the latter machine learning problem by damped Newton approximation methods, as presented in Monteiro and Pang (1999), Byrd et al. (1999), Byrd et al. (2000), Waltz et al. (2006). Since, in the framework of systems of nonlinear equations, finding solutions is generally a hot topic in current research, much work remains to be done. Some numerical experiments are performed on a toy model for the latter case and the algorithm is tested. We show the robustness of the method on this toy model.

2. A General Semisupervised Regularised and Multiview Machine Learning Problem

2.1 Operator Valued Kernels.

The aim of this subsection is to introduce kernels that take values in bundles of bounded linear operators on different Hilbert spaces and which will make the theoretical foundations for the localised machine learning problem. Actually, we prove that passing from nonlocalised to localised machine learning problem has the advantage of being very flexible and allowing a very large nonhomogeneity of the input data while not bringing new obstructions. In particular, this shows that, with minimal changes, similar results and algorithms can be obtained for localised machine learning problems.

Let X be a nonempty set and $\boldsymbol{H} = \{\mathcal{H}_x\}_{x\in X}$ a bundle of Hilbert spaces over the field $\mathbb{F}$, that is either $\mathbb{R}$ or $\mathbb{C}$, with base X, that is, $\mathcal{H}_x$ is a Hilbert space over $\mathbb{F}$, for any $x \in X$. In order to avoid confusion, let us note that the Hilbert spaces $\mathcal{H}_x$ are actually tagged Hilbert spaces, meaning that for $x \neq y$ the spaces $\mathcal{H}_x$ and $\mathcal{H}_y$ are disjoint. An *$\boldsymbol{H}$-operator valued kernel* K is a mapping defined on $X \times X$ such that $K(y, x) \in \mathcal{B}(\mathcal{H}_x, \mathcal{H}_y)$ for all $x, y \in X$. Here and throughout this article, if $\mathcal{H}$ and $\mathcal{K}$ are two Hilbert spaces over the same filed $\mathbb{F}$,

we denote by $\mathcal{B}(\mathcal{H},\mathcal{K})$ the Banach space of all linear and bounded operators $T\colon \mathcal{H}\to\mathcal{K}$, endowed with the operator norm. We denote by $\mathfrak{K}(X;\boldsymbol{H})$ the collection of all $\boldsymbol{H}$-operator valued kernels on X and it is clear that $\mathfrak{K}(X;\boldsymbol{H})$ is a vector space over $\mathbb{F}$.

Given $K\in\mathfrak{K}(X;\boldsymbol{H})$, the adjoint kernel K^* is defined by $K^*(x,y)=K(y,x)^*$ for all $x,y\in X$. Clearly $K^*\in\mathfrak{K}(X;\boldsymbol{H})$. The kernel K is called *Hermitian* or *symmetric* if $K=K^*$. If $\mathbb{F}=\mathbb{C}$ then any kernel K is a linear combination of two Hermitian kernels, more precisely, letting

$$\Re(K):=(K+K^*)/2,\quad \Im(K):=(K-K^*)/2\mathrm{i}, \tag{2.1}$$

we have

$$K=\Re(K)+\mathrm{i}\Im(K). \tag{2.2}$$

It is easy to see that $K\mapsto K^*$ is an involution, that is, it is conjugate linear and involutive. In this way, $\mathfrak{K}(X;\boldsymbol{H})$ is a $*$-vector space, that is, $\mathfrak{K}(X;\boldsymbol{H})$ is a vector space endowed with an involution, e.g. see Ay and Gheondea (2015).

Let $\mathcal{F}(X;\boldsymbol{H})$ be the vector space over $\mathbb{F}$ of all cross-sections $f\colon X\to\bigcup_{x\in X}\mathcal{H}_x$, that is, $f(x)\in\mathcal{H}_x$ for all $x\in X$. Addition and multiplication with scalars in $\mathcal{F}(X;\boldsymbol{H})$ are defined pointwise. Equivalently, $\mathcal{F}(X;\boldsymbol{H})$ can be naturally identified with the vector space $\prod_{x\in X}\mathcal{H}_x$ and then any cross-section $f\in\prod_{x\in X}\mathcal{H}_x$ can be written $f=\{f_x\}_{x\in X}$. For each $x\in X$ and $h\in\mathcal{H}_x$ we consider the cross-section $\widehat{h}\in\mathcal{F}(X;\boldsymbol{H})$, defined by

$$\widehat{h}(y)=\begin{cases}h, & \text{if } y=x,\\ 0_{\mathcal{H}_y}, & \text{otherwise.}\end{cases} \tag{2.3}$$

In particular, since $\boldsymbol{H}$ consists of tagged Hilbert spaces this means that if either $h\in\mathcal{H}_x$ or $l\in\mathcal{H}_y$ are not null and $x\neq y$ then $\widehat{h}\neq\widehat{l}$. However, if $h=0_{\mathcal{H}_x}$ and $l=0_{\mathcal{H}_y}$ with $x\neq y$, then $\widehat{h}=\widehat{l}=0_{\mathcal{F}(X;\boldsymbol{H})}$. Also, for any $f\in\mathcal{F}(X;\boldsymbol{H})$ we have

$$f=\sum_{x\in X}\widehat{f_x}, \tag{2.4}$$

where $f_x:=f(x)$ for all $x\in X$. Clearly, for each $y\in X$ the sum $\sum_{x\in X}\widehat{f_x}(y)$ has at most one nonnull term and hence the sum in (2.4) converges pointwise.

Let $\mathcal{F}_0(X;\boldsymbol{H})$ be the vector subspace consisting of all $f\in\mathcal{F}(X;\boldsymbol{H})$ with finite support. Clearly, any cross-section of type $\widehat{h}$, for some $h\in\mathcal{H}_x$, belongs to $\mathcal{F}_0(X;\boldsymbol{H})$ and, for any $f\in\mathcal{F}_0(X;\boldsymbol{H})$ there exists uniquely distinct elements $x_1,\dots,x_n\in X$ and $h_i\in\mathcal{H}_{x_i}$, $i=1,\dots,n$, such that

$$f=\sum_{i=1}^n\widehat{h}_i.$$

An inner product $\langle\cdot,\cdot\rangle_0\colon\mathcal{F}_0(X;\boldsymbol{H})\times\mathcal{F}_0(X;\boldsymbol{H})\to\mathbb{F}$ can be defined by

$$\langle f,g\rangle_0=\sum_{x\in X}\langle f(x),g(x)\rangle_{\mathcal{H}_x},\quad f,g\in\mathcal{F}_0(X;\boldsymbol{H}). \tag{2.5}$$

In addition, let us observe that the sum in (2.5) makes sense in the more general case when $f, g \in \mathcal{F}(X; \boldsymbol{H})$ and at least one of f or g has finite support, the other can be arbitrary.

Associated to the kernel $K \in \mathfrak{K}(X; \boldsymbol{H})$ there is a sesquilinear form $\langle \cdot, \cdot \rangle_K \colon \mathcal{F}_0(X; \boldsymbol{H}) \times \mathcal{F}_0(X; \boldsymbol{H}) \to \mathbb{F}$ defined by

$$\langle f, g \rangle_K = \sum_{x,y \in X} \langle K(y,x) f(x), g(y) \rangle_{\mathcal{H}_y}, \quad f, g \in \mathcal{F}_0(X; \boldsymbol{H}), \tag{2.6}$$

that is, $\langle \cdot, \cdot \rangle_K$ is linear in the first variable and conjugate linear in the second variable. Also, the sesquilinear form $\langle \cdot, \cdot \rangle_K$ is Hermitian, that is, $\langle f, g \rangle_K = \overline{\langle g, f \rangle_K}$ for all $f, g \in \mathfrak{K}(X; \boldsymbol{H})$, if and only if the kernel K is Hermitian.

A *convolution operator* $C_K \colon \mathcal{F}_0(X; \boldsymbol{H}) \to \mathcal{F}(X; \boldsymbol{H})$ can be defined by

$$(C_K f)(y) = \sum_{x \in X} K(y,x) f(x), \quad f \in \mathcal{F}_0(X; \boldsymbol{H}),\ y \in X. \tag{2.7}$$

Clearly C_K is a linear operator and, with notation as in (2.5) and (2.6) we have

$$\langle C_K f, g \rangle_0 = \langle f, g \rangle_K, \quad f, g \in \mathcal{F}_0(X; \boldsymbol{H}). \tag{2.8}$$

By definition, the kernel K is *positive semidefinite* if the sesquilinear form $\langle \cdot, \cdot \rangle_K$ is nonnegative, that is, if $\langle f, f \rangle_K \geq 0$ for all $f \in \mathcal{F}_0(X; \boldsymbol{H})$, equivalently, if for all $n \in \mathbb{N}$, all $x_1, \ldots, x_n \in X$, and all $h_1 \in \mathcal{H}_{x_1}, \ldots, h_n \in \mathcal{H}_{x_n}$, we have

$$\sum_{i,j=1}^{n} \langle K(x_j, x_i) h_i, h_j \rangle_{\mathcal{H}_{x_j}} \geq 0. \tag{2.9}$$

An equivalent way of expressing (2.9) is to say that the operator block matrix $[K(x_j, x_i)]_{i,j=1}^n$, when viewed as a bounded linear operator acting in the orthogonal direct Hilbert sum $\mathcal{H}_{x_1} \oplus \cdots \oplus \mathcal{H}_{x_n}$, is a positive semidefinite operator. On the other hand, the kernel K is positive semidefinite if and only if the convolution operator, as defined in (2.7), is positive semidefinite when viewed as an operator on the inner product space $(\mathcal{F}_0(X; \boldsymbol{H}); \langle \cdot, \cdot \rangle_0)$, more precisely,

$$\langle C_K f, f \rangle_0 \geq 0, \quad f \in \mathcal{F}_0(X; \boldsymbol{H}). \tag{2.10}$$

It is easy to see that, if $\mathbb{F} = \mathbb{C}$ then, if the kernel K is positive semidefinite then it is Hermitian. If $\mathbb{F} = \mathbb{R}$ then this is not true and hence, for this case we confine to those positive semidefinite kernels that are Hermitian, more precisely, in this case, in addition to the property (2.9), by a positive semidefinite kernel we implicitly understand that it is Hermitian as well. The collection of all positive semidefinite $\boldsymbol{H}$-operator valued kernels on X is denoted by $\mathfrak{K}^+(X; \boldsymbol{H})$ and it is easy to see that $\mathfrak{K}^+(X; \boldsymbol{H})$ is a strict convex cone of the $*$-vector space $\mathfrak{K}(X; \boldsymbol{H})$.

Given an arbitrary bundle of Hilbert spaces $\boldsymbol{H} = \{\mathcal{H}_x\}_{x \in X}$ and an $\boldsymbol{H}$-operator valued Kernel k, a *Hilbert space linearisation*, or a *Kolmogorov decomposition*, or a *feature pair* of K is, by definition, a pair $(\mathcal{K}; V)$ subject to the following conditions.

(kd1) $\mathcal{K}$ is a Hilbert space over $\mathbb{F}$.

(kd2) $V = \{V(x)\}_{x\in X}$ is an operator bundle such that $V(x) \in \mathcal{B}(\mathcal{H}_x, \mathcal{K})$ for all $x \in X$.

(kd3) $K(x, y) = V(x)^*V(y)$ for all $x, y \in X$.

The linearisation $(\mathcal{K}; V)$ is called *minimal* if

(kd4) $\mathcal{K}$ is the closed span of $\{V(x)\mathcal{H}_x \mid x \in X\}$.

The following theorem is a general version of some classical results, e.g. Kolmogorov (1941), Parthasarathy and Schmidt (1972), Evans and Lewis (1977). This is also a special case of Constantinescu and Gheondea (1997). We include its proof in the Appendix A.

Theorem 1 *Given an arbitrary bundle of Hilbert spaces $\boldsymbol{H} = \{\mathcal{H}_x\}_{x\in X}$ and an $\boldsymbol{H}$-operator valued kernel K, the following assertions are equivalent.*

(a) *K is positive semidefinite.*

(b) *K has a Hilbert space linearisation.*

In addition, if K is positive semidefinite then a minimal Hilbert space linearisation $(\mathcal{K}; V)$ exists and it is unique, modulo unitary equivalence, that is, for any other minimal Hilbert space linearisation $(\mathcal{K}'; V')$ of K there exists a unitary operator $U\colon \mathcal{K}' \to \mathcal{K}$ such that $V(x) = UV'(x)$, for all $x \in X$.

Due to the uniqueness part in the previous theorem, for any $K \in \mathfrak{K}^+(X; \boldsymbol{H})$, we denote by $(\mathcal{K}_K; V_K)$ the minimal Hilbert space linearisation of K, as constructed during the proof of the implication (a)⇒(b).

Let X be a nonempty set and $\boldsymbol{H} = \{\mathcal{H}_x\}_{x\in X}$ a bundle of Hilbert spaces over $\mathbb{F}$. Given an $\boldsymbol{H}$-operator valued kernel K, a *reproducing kernel Hilbert space* associated to K is, by definition, a Hilbert space $\mathcal{R} \subseteq \mathcal{F}(X; \boldsymbol{H})$ subject to the following conditions.

(rk1) $\mathcal{R}$ is a subspace of $\mathcal{F}(X; \boldsymbol{H})$, with all induced algebraic operations.

(rk2) For all $x \in X$ and $h \in \mathcal{H}_x$, the cross-section $K_x h := K(\cdot, x)h$ belongs to $\mathcal{R}$.

(rk3) For all $f \in \mathcal{R}$ we have $\langle f(x), h\rangle_{\mathcal{H}_x} = \langle f, K_x h\rangle_{\mathcal{R}}$, for all $x \in X$ and all $h \in \mathcal{H}_x$.

Consequently, the following minimality condition holds as well:

(rk4) The span of $\{K_x h \mid x \in X,\ h \in \mathcal{H}_x\}$ is dense in $\mathcal{R}$.

Also, it is worth mentioning that by (rk2), for each $x \in X$, we actually have a bounded linear operator $K_x\colon \mathcal{H}_x \to \mathcal{R}$ defined by $K_x h := K(\cdot, x)h$, for all $h \in \mathcal{H}_x$. This operator is bounded, as proven in (4.1). The following result is a generalisation of Moore-Aronszajn Theorem, Moore (Part I, 1935, Part II, 1939), Aronszajn (1950), Micchelli and Pontil (2005), Carmelli et al. (2006), Caponnetto et al. (2008), Minh and Sindwhani (2011), Kadri et al. (2016). Also, it is a special case of Constantinescu and Gheondea (1997) and it is proven in the Appendix B. In Appendix C we present a more direct construction of the reproducing kernel Hilbert space induced by an operator valued positive semidefinite kernel.

Theorem 2 *Given an arbitrary bundle of Hilbert spaces $\boldsymbol{H} = \{\mathcal{H}_x\}_{x\in X}$ and an $\boldsymbol{H}$-operator valued kernel K, the following assertions are equivalent.*

(a) *K is positive semidefinite.*

(b) *There exists a reproducing kernel Hilbert space $\mathcal{R}$ having K its reproducing kernel.*

In addition, the reproducing kernel Hilbert space $\mathcal{R}$ is uniquely determined by its reproducing kernel K.

Remark 3 There is a natural bijective transformation between the unitary equivalency class of minimal linearisations $(\mathcal{K}; V)$ of K and the reproducing kernel Hilbert space $\mathcal{R}(K)$. The transformation from a minimal linearisations $(\mathcal{K}; V)$ to the reproducing kernel Hilbert space $\mathcal{R}(K)$ is described during the proof of the implication (a)⇒(b) of Theorem 2, see Appendix B. In the following we describe the inverse of this transformation.

Let $(\mathcal{R}; \langle\cdot,\cdot\rangle_{\mathcal{R}})$ be a reproducing kernel Hilbert space with reproducing kernel K. We define the operator bundle $V = \{V(x)\}_{x\in X}$ by

$$V(x)h = K_x h, \quad x \in X,\ h \in \mathcal{H}_x, \tag{2.11}$$

and remark that $V(x)\colon \mathcal{H}_x \to \mathcal{R}$ for all $x \in X$. By means of the reproducing property (rk3) of the kernel K, we have

$$\langle V(x)h, V(x)h\rangle_{\mathcal{R}} = \langle K_x h, K_x h\rangle_{\mathcal{R}} = \langle K(x,x)h, h\rangle_{\mathcal{H}_x} \leq \|K(x,x)\|\|h\|^2_{\mathcal{H}_x}, \quad x \in X,\ h \in \mathcal{H}_x,$$

hence $V(x) \in \mathcal{B}(\mathcal{H}_x, \mathcal{R})$. Also, using once more the reproducing property (rk3) of K, it follows that, for all $x, y \in X$, $h \in \mathcal{H}_x$, and $g \in \mathcal{H}_y$, we have

$$\langle V(y)^* V(x)h, g\rangle_{\mathcal{H}_y} = \langle V(x)h, V(y)g\rangle_{\mathcal{R}} = \langle K_x h, K_y g\rangle_{\mathcal{R}} = \langle K(y,x)h, g\rangle_{\mathcal{H}_y}.$$

Therefore, $K(y,x) = V(y)^* V(x)$ for all $x, y \in X$ and hence, $(\mathcal{R}; V)$ is a linearistion of K. In addition, using the minimality property (rk3), it is easy to see that the linearisation $(\mathcal{R}; V)$ is minimal as well.

One of the most important property of a reproducing kernel Hilbert space consists in the fact that, as a function space, its topology makes continuous all evaluation operators, see the proof in the Appendix D.

Theorem 4 *With notation as before, let $\mathcal{H}$ be a Hilbert space in the vector space $\mathcal{F}(X; \boldsymbol{H})$. The following assertions are equivalent.*

(a) *$\mathcal{H}$ is a reproducing kernel space of $\boldsymbol{H}$-valued maps on X.*

(b) *For any $x \in X$ the linear operator $\mathcal{H} \ni f \mapsto f(x) \in \mathcal{H}_x$ is bounded.*

In connection to the previous theorem it is worth mentioning that, for a reproducing kernel Hilbert space $\mathcal{H}_K \subseteq \mathcal{F}(X; \boldsymbol{H})$ and for arbitrary $x \in X$, the evaluation operator $\mathcal{H}_K \ni f \mapsto f(x) \in \mathcal{H}_x$ coincides with $K_x^*\colon \mathcal{H}_K \to \mathcal{H}_x$, where $K_x\colon \mathcal{H}_x \to \mathcal{H}_K$ is the bounded operator, see the axiom (rk2), defined by $K_x h := K(\cdot, x)h$, for all $h \in \mathcal{H}_x$.

2.2 Localisation of Semisupervised, Regularised, and Multiview Learning.

Let X be a nonempty set and $\mathbf{W} = \{\mathcal{W}_x\}_{x\in X}$ be a bundle of Hilbert spaces on X. In this section, it is not important whether the Hilbert spaces are complex or real, hence all Hilbert spaces are considered to be over the field $\mathbb{F}$, that is either $\mathbb{C}$ or $\mathbb{R}$. There is a difference between the complex and the real case consisting in the fact that in the latter case, for positive semidefiniteness we assume also the symmetry, or Hermitian, property, while in the complex case, the symmetry property is a consequence of the positive semidefiniteness. If K is a positive semidefinite $\mathbf{W}$-operator valued kernel, we let $\mathcal{H}_K$ be its reproducing kernel Hilbert space, as in the previous subsection. Also, let $\boldsymbol{Y} = \{\mathcal{Y}_x\}_{x\in X}$ be a bundle of Hilbert spaces.

For $l, u \in \mathbb{N}$, consider input distinct points $x_1, \ldots, x_{l+u} \in X$. Here $x_1, \ldots, x_l$ are the labeled input points while $x_{l+1}, \ldots, x_{l+u}$ are the unlabeled input points. More precisely, there are given $y_1, \ldots, y_l$ output points, such that $y_j \in \mathcal{Y}_{x_j}$ for all $j = 1, \ldots, l$. Then, for the general data let

$$\boldsymbol{x} := (x_j)_{j=1}^{l+u}, \quad \boldsymbol{y} := (y_j)_{j=1}^{l}, \quad \boldsymbol{z} := \big((x_j)_{j=l+1}^{l+u}, (y_j)_{j=1}^{l}\big).$$

The input points $x_1, \ldots, x_{l+u}$ are randomly selected with respect to an unknown probability and then, depending on the concrete problem, the labels $y_1, \ldots, y_l$ are produced in a certain way.

Let $\boldsymbol{W}^{l+u}$ denote the Hilbert space

$$\boldsymbol{W}^{l+u} = \bigoplus_{j=1}^{l+u} \mathcal{W}_{x_j}. \tag{2.12}$$

For $f \in \mathcal{H}_K$ let

$$\boldsymbol{f} := (f(x_1), \ldots, f(x_{l+u})) \in \boldsymbol{W}^{l+u}. \tag{2.13}$$

Also, there is given a (Hermitian, if $\mathbb{F} = \mathbb{R}$) positive semidefinite operator $M \in \mathcal{B}(\boldsymbol{W}^{l+u})$ represented as an operator block $(l+u)\times(l+u)$-matrix $M = [M_{j,k}]$, with $M_{j,k} \in \mathcal{B}(\mathcal{W}_{x_k}, \mathcal{W}_{x_j})$ for all $j, k = 1, \ldots, l+u$. Let $\boldsymbol{V} = \{V_x\}_{x\in X}$ be a bundle of maps, loss functions, where $V_x \colon \mathcal{Y}_x \times \mathcal{Y}_x \to \mathbb{R}$ is a function, for all $x \in X$. Also, $\boldsymbol{C} = \{C_x\}_{x\in X}$ is a bundle of bounded linear operators, where $C_x \colon \mathcal{W}_x \to \mathcal{Y}_x$ for all $x \in X$. The general minimisation problem is

$$f_{\boldsymbol{z},\gamma} = \operatorname{argmin}_{f\in\mathcal{H}_K} \frac{1}{l}\sum_{j=1}^{l} V_{x_j}(y_j, C_{x_j} f(x_j)) + \gamma_A \|f\|^2_{\mathcal{H}_K} + \gamma_I \langle \boldsymbol{f}, M\boldsymbol{f}\rangle_{\boldsymbol{W}^{l+u}}, \tag{2.14}$$

where $\gamma = (\gamma_A, \gamma_I)$ and $\gamma_A > 0$ and $\gamma_I \geq 0$ are the regularisation parameters.

The optimisation problem (2.14) is a localised version of the general vector valued reproducing kernel Hilbert space for semisupervised, regularised, manifold regularised and multiview learning as in Minh et al. (2016). It is also useful to introduce the map to be minimised

$$\mathcal{I}(f) := \frac{1}{l}\sum_{j=1}^{l} V_{x_j}(y_j, C_{x_j} f(x_j)) + \gamma_A \|f\|^2_{\mathcal{H}_K} + \gamma_I \langle \boldsymbol{f}, M\boldsymbol{f}\rangle_{\boldsymbol{W}^{l+u}}$$

and, since $f(x) = K_x^* f$ for all $f \in \mathcal{H}_K$ and all $x \in X$, it equals

$$= \frac{1}{l} \sum_{j=1}^{l} V_{x_j}(y_j, C_{x_j} K_{x_j}^* f) + \gamma_A \|f\|_{\mathcal{H}_K}^2 + \gamma_I \langle \boldsymbol{f}, M\boldsymbol{f} \rangle_{\boldsymbol{W}^{l+u}}. \tag{2.15}$$

In the following we explain the terms in the minimising map (2.15) and their significance from the point of view of machine learning. Firstly, why labeled and unlabeled data? Traditionally, in machine learning there are three fundamental approaches: supervised learning, unsupervised learning, and reinforcement learning, but the last one is out of our concern. We firstly recall the meaning and limitations of the first two approaches. *Supervised learning* means that the training of the machine learning model is using exclusively labeled dataset. The input points and the labels are selected according to a probability that is usually unknown or the input points are selected according to an unknown probability and then the labels are produced in a certain fashion. Basically, this means that a label is a description showing a model, what it is expected to predict. But supervised learning has some limitations since this process is: *slow*, because it requires human experts to either manually label training examples one by one or carefully supervise the procedure, and *costly*, because, in order to obtain reliable results, a model should be trained on the large volumes of labeled data to provide accurate predictions. *Unsupervised learning* is that approach when a model tries to find hidden patterns, differences, and similarities in unlabeled data by itself, without human supervision. Most of the time, in this approach, data points are grouped into clusters based on similarities. But, while unsupervised learning is a cheaper way to perform training tasks, it has other limitations: it has a *limited area of applications*, mostly for clustering purposes, and provides *less accurate results*.

Semisupervised learning combines supervised learning and unsupervised learning techniques to solve some important issues: we train an initial model on a few labeled samples and then iteratively apply it to a greater number of unlabeled data. Unlike unsupervised learning, semisupervised learning works for a larger variety of problems: classification, regression, clustering, and association. Unlike supervised learning, the method uses small amounts of labeled data but large amounts of unlabeled data, with the advantage that it reduces the costs on human work and the data preparation time, while the accuracy of results is not altered. Of course, some other issues show up: the unlabeled points should show certain consistency and for this some regularisation techniques are needed. A comprehensive discussion on this subject can be found in the collection Chapelle et al. (2006).

Secondly, the reproducing kernel Hilbert space $\mathcal{H}_K$ is associated to a vector valued positive semidefinite kernel for several reasons, but mainly because this is related to the multiview learning, cf. Evgeniou et al. (2005), Micchelli and Pontil (2005), Sindhwani and Rosenberg (2008), Rosenberg et al. (2009), Minh and Sindwhani (2011), Luo et al. (2013), Kadri et al. (2016), Minh et al. (2016), Hashimoto et al. (2021), Hashimoto et al. (2022), Hashimoto et al. (2023a), Hashimoto et al. (2023b). In this article we consider localised versions of these operator valued reproducing kernel Hilbert spaces that offers flexibility for a larger class of learning problems, as explained in the Introduction, and does not bring additional obstructions, as proven in Subsection 2.1.

Further on, the first term in (2.15) controls the distance, estimated by local loss (or cost) functions at the labeled input points with respect to the labels. More precisely, for

each label point x_j, $j = 1, \dots, l$ the label $y_j \in \mathcal{Y}_{x_j}$ is compared, through the cost function V_{x_j}, with $f(x_j) \in \mathcal{W}_{x_j}$ by a combination operator $C_{x_j} \colon \mathcal{W}_{x_j} \to \mathcal{Y}_{x_j}$, because the Hilbert spaces $\mathcal{Y}_{x_j}$ may be different from $\mathcal{W}_{x_j}$.

Example 1 Following Minh et al. (2016), for an input point $x \in X$ consider the label Hilbert space $\mathcal{Y}$ and let $\mathcal{W} = \mathcal{Y}^m$, the orthogonal direct sum of m copies of $\mathcal{Y}$. With this notation, the kernel K has values in $\mathcal{B}(\mathcal{W})$. A multiview $f(x)$ is then an m-tuple $(f^1(x), \dots, f^m(x))^T$, with each $f^i(x) \in \mathcal{Y}$ and let the combination operator $C = [C_1, \dots, C_m] \colon \mathcal{W} = \mathcal{Y}^m \to \mathcal{Y}$, that is, $Cf(x) = C_1 f^1(x) + \cdots + C_m f^m(x) \in \mathcal{Y}$.

In this article, the loss functions are also localised and one strong reason for this is that, depending on different purposes that this semisupervised learning is used for, there is a very large pool of choices for loss functions.

Example 2 We list in the following a few loss functions of interest in machine learning, see Zhao et al. (2010) and Kwon and Zou (2023) for a more comprehensive list and applications.

(1) *Least Squares.* The least squares loss function is

$$V(y, z) = (y - z)^2, \quad y, z \in \mathbb{R}.$$

It is convex, nonnegative, and differentiable.

(2) *Sigmoid.* The sigmoid loss function is

$$V(y, z) = \frac{1}{1 + \exp(z - y)}, \quad y, z \in \mathbb{R}.$$

It is nonnegative, differentiable, and nonconvex.

(3) *Hinge.* The hinge loss function is

$$V(y, z) = \max\{0, 1 - yz\}, \quad y, z \in \mathbb{R}.$$

It is nonnegative, continuous, convex, but not differentiable.

(4) *Exponential Least Squares.* The exponential least squares function is

$$V(y, z) = 1 - \exp(-(y - z)^2), \quad y, z \in \mathbb{R}.$$

It is nonnegative, upper bounded by 1, differentiable, and nonconvex.

(5) *Leaky Hockey Stick.* The leaky hockey stick loss function is

$$V(y, z) = \begin{cases} -\log(zy), & yz > 1, \\ 1 - yz, & yz \leq 1. \end{cases}$$

It is upper and lower unbounded, convex, and differentiable.

The second term in (2.15) is the usual regularisation penalty term, following the Tikhonov regularisation method, cf. Tikhonov (1963). This is used in order to avoid large target functions f and overfitting, that is, optimising functions that match very accurately the labeled data but perform badly for other data. Because of this, the regularisation parameter γ_A

is always positive. In the literature, sometimes the second term is replaced by $\varphi(\|f\|_{\mathcal{H}_K})$, where $\varphi\colon \mathbb{R}_+ \to \mathbb{R}_+$ is an increasing function so, in our case $\varphi(t) = \gamma_A t^2$.

The third term in (2.15) combines vector valued manifold regularisation, cf. Minh and Sindwhani (2011), with multiview regularisation, cf. Rosenberg et al. (2009) and Sun (2011). The parameter γ_I may be taken 1, without loss of generality, since it can be absorbed in M. Following Minh et al. (2016), the operator multiview regularisation term $\gamma_I \langle \boldsymbol{f}, M\boldsymbol{f}\rangle_{\boldsymbol{W}^{l+u}}$ is decomposed as

$$\gamma_I \langle \boldsymbol{f}, M\boldsymbol{f}\rangle_{\boldsymbol{W}^{l+u}} = \gamma_B \langle \boldsymbol{f}, M_B\boldsymbol{f}\rangle_{\boldsymbol{W}^{l+u}} + \gamma_W \langle \boldsymbol{f}, M_W\boldsymbol{f}\rangle_{\boldsymbol{W}^{l+u}}, \tag{2.16}$$

where $M_B, M_W \in \mathcal{B}(\boldsymbol{W}^{l+u})$ are selfadjoint positive operators and $\gamma_B, \gamma_W \geq 0$. As before, the regularising parameters γ_B and γ_W may be taken 1, without loss of generality, because they can be absorbed in M_B and M_W, respectively. The first term in (2.16) is the *localised between-view regularisation* while the latter term in (2.16) is the *localised within-view regularisation*. In the next example we show by some concrete situations the constructions of the operators M_B and M_W and their significance.

Example 3 This example follows closely the example of between-view regularisation as in Minh et al. (2016). With notation as in Example 1, let $M_m = mI_m - \mathbf{1}_m \mathbf{1}_m^T$, where $\mathbf{1}_m = (1, 1, \ldots, 1)^T$. More precisely, M_m is the $m \times m$ matrix with all entries equal to $m-1$ throughout its diagonal and -1 elsewhere. Then, for each $\mathbf{a} = (a_1, \ldots, a_m)^T \in \mathbb{R}^m$, we have

$$\mathbf{a}^T M_m \mathbf{a} = \sum_{j,k=1,\ j<k}^{m} (a_j - a_k)^2.$$

Then, for each $\mathbf{y} = (y_1, \ldots, y_m)^T \in \mathcal{Y}^m = \mathcal{W}$ we have

$$\mathbf{y}^T (M_m \otimes I_{\mathcal{Y}})\mathbf{y} = \sum_{j,k=1,\ j<k} \|y_j - y_k\|_{\mathcal{Y}}^2.$$

So, letting $M_B = I_{u+l} \otimes (M_m \otimes I_{\mathcal{Y}})$, for each $\mathbf{f} = (f(x_1), \ldots, f(x_{u+l}) \in \mathcal{Y}^{m(u+l)} = \mathcal{W}^{u+l}$, with $f(x_i) \in \mathcal{Y}^m = \mathcal{W}$, we have

$$\langle \mathbf{f}, M_B \mathbf{f}\rangle_{\mathcal{W}^{u+l}} = \sum_{i=1}^{u+l} \langle f(x_i), (M_m \otimes I_{\mathcal{Y}}) f(x_i)\rangle_{\mathcal{W}} = \sum_{i=1}^{u+l} \sum_{j,k=1,\ j<k} \|f^j(x_i) - f^k(x_i)\|_{\mathcal{Y}}^2.$$

This term is a control on the consistency between different components f^i's which represent the outputs on different views.

Example 4 This example follows essentially Minh et al. (2016) for a within-view manifold regularisation via multiview graph Laplacians in support vector machine learning, cf. Sun (2011). For manifold regularisation, a data adjaceny graph is defined in such a way that the entries measure the similarity or closeness of pairs of inputs. Given an undirected graph $G = (\mathcal{V}, E)$, where the vertices are $\mathcal{V} = \{1, \ldots, n\}$ and edges are simply pairs (j, k), assume that for each edge $(j, k) \in E$ there is a weight $w_{j,k}$, and to each edge $(j, k) \notin E$ we let $w_{j,k} = 0$, in such a way that the weight matrix $W = [w_{j,k}]$ is Hermitian and nonnegative (positive semidefinite).

For example, when each vertex j is associated to a vector $h_j \in \mathbb{R}^d$, we can use the Gaussian weights

$$w_{j,k} = \exp(-\|h_j - h_k\|^2/2\sigma^2). \tag{2.17}$$

In order to simplify the complexity of calculations, cf. Sun (2011), for most of the edges (j,k) we take $w_{j,k} = 0$ and only for neighbouring (j,k), that is, $\|h_j - h_k\|_2 \leq \epsilon$, for some ϵ, we define the weights by (2.17).

Further on, letting $v_{j,j} = \sum_{k=1}^n w_{j,k}$ and $v_{j,k} = 0$ if $j \neq k$, we make the diagonal matrix $V = [v_{j,k}]$. We work under the assumptions that $v_{j,j} > 0$ for all $j = 1, \ldots, n$. Then the *graph Laplacian* matrix is $L := V - W$, which is positive semidefinite. Sometimes it is useful to work with the *normalised graph Laplacian* $\widetilde{L} := V^{-1/2} L V^{-1/2}$.

But, because the learning is from multiviews, this should be performed for each view and then aggregated in a consistent fashion. From now on we use the same notations and settings as in Example 1 and Example 3. Assume that, to each view i, $1 \leq i \leq m$, we consider the undirected graph $G^i = (\mathcal{V}^i, E^i)$ where $\mathcal{V}^i = \{1, \ldots, u+l\}$, let $W^i = [w^i_{j,k}]$ be the corresponding weight matrix that is Hermitian and nonnegative, and let $L^i = [l^i_{j,k}]$ be the corresponding graph Laplacian. Then, for each vector $\mathbf{a} \in \mathbb{R}^{u+l}$ we have

$$\mathbf{a}^T L^i \mathbf{a} = \sum_{j,k=1,\ j<k}^{l+u} w^i_{j,k}(a_j - a_k)^2.$$

Now we aggregate the graph Laplacians into the multiview graph Laplacian as a block matrix $L = [L_{j,k}]$, where for each $j, k = 1, \ldots, u+l$ we define

$$L_{j,k} = \operatorname{diag}(l^1_{j,k}, \ldots, l^m_{j,k}).$$

This implies that for each vector $\mathbf{a} = (a_1, \ldots, a_{u+l})$, with $a_j = (a^1_j, \ldots, a^m_j) \in \mathbb{R}^m$ for each $j = 1, \ldots, u+l$, we have

$$\mathbf{a}^T L \mathbf{a} = \sum_{i=1}^m \sum_{j,k=1,\ j<k}^{l+u} w^i_{j,k}(a^i_j - a^i_k)^2.$$

Finally, letting $M_W := L \otimes I_{\mathcal{Y}}$, we have

$$\langle \mathbf{f}, M_W \mathbf{f} \rangle_{\mathcal{W}^{u+l}} = \sum_{i=1}^m \sum_{j,k=1,\ j<k}^{l+u} w^i_{j,k} \|f^i(x_j) - f^i(x_k)\|^2_{\mathcal{Y}},$$

for all $\mathbf{f} = \{f^i(x_j) \mid i = 1, \ldots, m,\ j = 1, \ldots, u+l\} \in \mathcal{W}^{u+l} = \mathcal{Y}^{m(u+l)}$. Each term in the leftmost sum is a manifold regularisation for the view i and hence the double sum is the aggregated manifold regularisation for all views. In this fashion, consistency is enforced for each view.

2.3 A Representer Theorem.

We continue to use the notation as in the previous subsection. Generally speaking, a representer theorem has the goal to prove that the optimal solution to the problem (2.14)

should belong to the space

$$\mathcal{H}_{K,\boldsymbol{x}} = \Big\{ \sum_{i=1}^{l+u} K_{x_i} w_i \mid w_i \in \mathcal{W}_{x_i} \Big\}. \tag{2.18}$$

Let $\overline{\mathcal{H}}_{K,\boldsymbol{x}}$ denote its closure in $\mathcal{H}_K$ and let $P_{\overline{\mathcal{H}}_{K,\boldsymbol{x}}}$ denote the orthogonal projection of $\mathcal{H}_K$ onto $\overline{\mathcal{H}}_{K,\boldsymbol{x}}$. Let the sampling operator $S_{\mathbf{x}} : \mathcal{H}_K \to \boldsymbol{W}^{l+u}$ be defined by

$$S_{\mathbf{x}} f = (K_{x_i}^* f)_{i=1}^{l+u} = (f(x_i))_{i=1}^{l+u} = \mathbf{f}, \quad f \in \mathcal{H}_K, \tag{2.19}$$

where $\boldsymbol{x} = (x_1, \ldots, x_{l+u})$ and we have taken into account that $f(x) = K_x^* f$ for all $f \in \mathcal{H}_K$ and all $x \in X$. Let also $E_{\boldsymbol{C},\mathbf{x}} : \mathcal{H}_K \to \boldsymbol{Y}^l$, where

$$\boldsymbol{Y}^l := \bigoplus_{j=1}^{l} \mathcal{Y}_{x_j}, \tag{2.20}$$

be defined by

$$E_{\boldsymbol{C},\boldsymbol{x}} f = \left(C_{x_1} K_{x_1}^* f, \ldots, C_{x_l} K_{x_l}^* f \right) = \left(C_{x_1} f(x_1), \ldots, C_{x_l} f(x_l) \right), \quad f \in \mathcal{H}_K. \tag{2.21}$$

The main technical fact used in this section is a lemma whose proof is inspired by the proof of Theorem 2 in Minh et al. (2016).

Lemma 5 *With notation and assumptions an in Subsection 2.2 and as before, for any $f \in \mathcal{H}_K$ the following inequality holds.*

$$\mathcal{I}(P_{\overline{\mathcal{H}}_{K,\boldsymbol{x}}} f) \leq \mathcal{I}(f).$$

Proof We have the decomposition

$$\mathcal{H}_K = \overline{\mathcal{H}}_{K,\boldsymbol{x}} \oplus \mathcal{H}_{K,\boldsymbol{x}}^{\perp}. \tag{2.22}$$

Let $f \in \mathcal{H}_{K,\boldsymbol{x}}^{\perp}$ be fixed. Then, for any $\mathbf{b} \in \boldsymbol{Y}^l$, since $C_{x_i}^* b_i \in \mathcal{W}_{x_i}$, for all $i = 1, \ldots, l+u$, and hence

$$\sum_{i=1}^{l} K_{x_i} C_{x_i}^* b_i \in \mathcal{H}_{K,\boldsymbol{x}},$$

we have

$$\langle E_{\boldsymbol{C},\mathbf{x}} f, \mathbf{b} \rangle_{\boldsymbol{Y}^l} = \big\langle f, E_{C,\mathbf{x}}^* \mathbf{b} \big\rangle_{\mathcal{H}_K} = \sum_{i=1}^{l} \big\langle f, K_{x_i} C_{x_i}^* b_i \big\rangle_{\mathcal{H}_K} = \langle f, \sum_{i=1}^{l} K_{x_i} C_{x_i}^* b_i \rangle_{\mathcal{H}_K} = 0.$$

Consequently,

$$E_{\boldsymbol{C},\mathbf{x}} f = (C_{x_1} K_{x_1}^* f, \ldots, C_{x_l} K_{x_l}^* f) = 0. \tag{2.23}$$

Similarly, by the reproducing property, letting $\boldsymbol{w} = (w_1, \ldots, w_{l+u})$ be an arbitrary vector in $\boldsymbol{W}^{l+u}$, we have

$$\langle S_{\mathbf{x}} f, \mathbf{w} \rangle_{\boldsymbol{W}^{l+u}} = \sum_{i=1}^{l+u} \langle f(x_i), \mathbf{w} \rangle_{\boldsymbol{W}^{l+u}} = \sum_{i=1}^{l+u} \langle f, K_{x_i} w_i \rangle_{\mathcal{H}_K} = \left\langle f, \sum_{i=1}^{l+u} K_{x_i} w_i \right\rangle_{\mathcal{H}_K} = 0,$$

hence

$$\mathbf{f} = S_{\mathbf{x}} f = (f(x_1), \dots, f(x_{l+u})) = 0. \tag{2.24}$$

For arbitrary $f \in \mathcal{H}_K$, in view of the decomposition (2.22), we have the unique decomposition $f = f_0 + f_1$ with $f_0 \in \overline{\mathcal{H}}_{K,\mathbf{x}}$ and $f_1 \in \mathcal{H}_{K,\mathbf{x}}^{\perp}$, that is, $f_0 = P_{\overline{\mathcal{H}}_{K,\boldsymbol{x}}} f$. Then,

$$\|f_0 + f_1\|_{\mathcal{H}_K}^2 = \|f_0\|_{\mathcal{H}_K}^2 + \|f_1\|_{\mathcal{H}_K}^2 ,$$

and consequently,

$$\begin{aligned}\mathcal{I}(f) = \mathcal{I}(f_0 + f_1) = \frac{1}{l}\sum_{i=1}^{l} V_{x_i}(y_i, C_{x_i} K_{x_i}^* f_0 + C_{x_i} K_{x_i}^* f_1) + \gamma_A \|f_0\|_{\mathcal{H}_\mathcal{K}}^2 + \gamma_A \|f_1\|_{\mathcal{H}_\mathcal{K}}^2 \\ + \gamma_I \langle S_{\mathbf{x}} f_0, M S_{\mathbf{x}} f_0 \rangle_{\boldsymbol{W}^{l+u}} + \gamma_I \langle S_{\mathbf{x}} f_0, M S_{\mathbf{x}} f_1 \rangle_{\boldsymbol{W}^{l+u}} \\ + \gamma_I \langle S_{\mathbf{x}} f_1, M S_{\mathbf{x}} f_0 \rangle_{\boldsymbol{W}^{l+u}} + \gamma_I \langle S_{\mathbf{x}} f_1, M S_{\mathbf{x}} f_1 \rangle_{\boldsymbol{W}^{l+u}} .\end{aligned} \tag{2.25}$$

By (2.23) we then see that

$$V_{x_i}(y_i, C_{x_i} K_{x_i}^* f_0 + C_{x_i} K_{x_i}^* f_1) = V_{x_i}(y_i, C_{x_i} K_{x_i}^* f_0),$$

and by (2.24) we see that

$$\langle S_{\mathbf{x}} f_0, M S_{\mathbf{x}} f_1 \rangle_{\boldsymbol{W}^{l+u}} = \langle S_{\mathbf{x}} f_0, 0 \rangle_{\boldsymbol{W}^{l+u}} = 0.$$

So,

$$\langle S_{\mathbf{x}} f_0, M S_{\mathbf{x}} f_1 \rangle_{\boldsymbol{W}^{l+u}} = \langle S_{\mathbf{x}} f_1, M S_{\mathbf{x}} f_0 \rangle_{\boldsymbol{W}^{l+u}} = \langle S_{\mathbf{x}} f_1, M S_{\mathbf{x}} f_1 \rangle_{\boldsymbol{W}^{l+u}} = 0 \tag{2.26}$$

and hence, by (2.25), we have

$$\begin{aligned}\mathcal{I}(f) = \mathcal{I}(f_0 + f_1) = \frac{1}{l}\sum_{i=1}^{l} V_{x_i}(y_i, C_{x_i} K_{x_i}^* f_0 + C_{x_i} K_{x_i}^* f_1) + \gamma_A \|f_0\|_{\mathcal{H}_\mathcal{K}}^2 + \gamma_A \|f_1\|_{\mathcal{H}_\mathcal{K}}^2 \\ + \gamma_I \langle S_{\mathbf{x}} f_0, M S_{\mathbf{x}} f_0 \rangle_{\boldsymbol{W}^{l+u}} + \gamma_I \langle S_{\mathbf{x}} f_0, M S_{\mathbf{x}} f_1 \rangle_{\boldsymbol{W}^{l+u}} + \\ + \gamma_I \langle S_{\mathbf{x}} f_1, M S_{\mathbf{x}} f_0 \rangle_{\boldsymbol{W}^{l+u}} + \gamma_I \langle S_{\mathbf{x}} f_1, M S_{\mathbf{x}} f_1 \rangle_{\boldsymbol{W}^{l+u}}\end{aligned}$$

and then, by (2.23) and (2.26) we get that

$$\begin{aligned}\mathcal{I}(f) &= \frac{1}{l}\sum_{i=1}^{l} V_{x_i}(y_i, C_{x_i} K_{x_i}^* f_0) + \gamma_A \|f_0\|_{\mathcal{H}_\mathcal{K}}^2 + \gamma_A \|f_1\|_{\mathcal{H}_\mathcal{K}}^2 + \gamma_I \langle S_{\mathbf{x}} f_0, M S_{\mathbf{x}} f_0 \rangle_{\boldsymbol{W}^{l+u}} \\ &\geq \frac{1}{l}\sum_{i=1}^{l} V_{x_i}(y_i, C_{x_i} K_{x_i}^* f_0) + \gamma_A \|f_0\|_{\mathcal{H}_\mathcal{K}}^2 + \gamma_I \langle S_{\mathbf{x}} f_0, M S_{\mathbf{x}} f_0 \rangle_{\boldsymbol{W}^{l+u}} = \mathcal{I}(f_0),\end{aligned} \tag{2.27}$$

and the proof is finished. ■

In order to get a conclusion in the spirit of the representer theorem, extra assumptions are needed.

Proposition 6 *Assume that the subspace $\mathcal{H}_{K,\boldsymbol{x}}$, see (2.18), is closed. This happens, for example, if all Hilbert spaces $\mathcal{W}_{x_1},\ldots,\mathcal{W}_{x_{l+u}}$ have finite dimensions. If the minimisation problem (2.14) has a solution $f_{\boldsymbol{z},\gamma}$ then there exist $a_1,\ldots,a_{l+u}$, with $a_j\in\mathcal{W}_{x_j}$ for all $j=1,\ldots,l+u$, such that*

$$f_{\boldsymbol{z},\gamma}=\sum_{j=1}^{l+u}K_{x_j}a_j.$$

Proof Since $\mathcal{H}_{K,\boldsymbol{x}}$ is closed, we have the decomposition

$$\mathcal{H}_K=\mathcal{H}_{K,\boldsymbol{x}}\oplus\mathcal{H}_{K,\boldsymbol{x}}^{\perp}. \tag{2.28}$$

If f is a solution to the minimisation problem (2.14), in view of Lemma 5, it follows that $f\in\mathcal{H}_{K,\boldsymbol{z}}$, and the conclusion follows. ■

The main theorem of this section is a representer theorem under certain general and natural assumptions.

Theorem 7 *Assume that the loss functions $V_{x_j}(y_j,\cdot)$ are bounded from below and continuous, for all $j=1,\ldots,l$, and that all Hilbert spaces $\mathcal{W}_{x_1},\ldots,\mathcal{W}_{x_{l+u}}$ have finite dimensions. Then the minimisation problem (2.14) has a solution $f_{\boldsymbol{z},\gamma}$ and, for any such a solution, there exist $a_1,\ldots,a_{l+u}$, with $a_j\in\mathcal{W}_{x_j}$ for all $j=1,\ldots,l+u$, such that*

$$f_{\boldsymbol{z},\gamma}=\sum_{j=1}^{l+u}K_{x_j}a_j.$$

Proof We first observe that, since all loss functions $V_{x_j}(y_j,\cdot)$, $j=1,\ldots,l+u$, are lower bounded and M is positive semidefinite, from (2.15) it follows that $\mathcal{I}(f)$ is lower bounded and hence its infimum exists as a real number. Since all the Hilbert spaces $\mathcal{W}_{x_1},\ldots,\mathcal{W}_{x_{l+u}}$ have finite dimensions it follows that the subspace $\mathcal{H}_{K,\boldsymbol{x}}$ has finite dimension and hence it is closed. Then, from Lemma 5 it follows that

$$-\infty<\inf_{f\in\mathcal{H}_K}\mathcal{I}(f)=\inf_{f\in\mathcal{H}_{K,\boldsymbol{x}}}\mathcal{I}(f).$$

So, it remains only to show that the latter infimum is attained.

Indeed, since $\gamma_A>0$ and the loss functions $V_{x_j}(y_j,\cdot)$ are bounded from below for all $j=1,\ldots,l$, it follows that

$$\lim_{\|f\|_{\mathcal{H}_K}\to\infty}\mathcal{I}(f)=+\infty. \tag{2.29}$$

Since all loss functions $V_{x_j}(y_j,\cdot)$ are continuous, for $j=1,\ldots,l$, and the evaluation functionals on $\mathcal{H}_K$ are continuous as well, see Theorem 4, it follows that $\mathcal{I}$ is continuous on $\mathcal{H}_K$. Let $f_0\in\mathcal{H}_{K,\boldsymbol{x}}$ be arbitrary but fixed. From (2.29), for $\epsilon>0$ there exists $\delta>0$ such that

$$\mathcal{I}(f)\geq\mathcal{I}(f_0)+\epsilon \text{ for all } f\in\mathcal{H}_{K,\boldsymbol{x}} \text{ with } \|f-f_0\|_{\mathcal{H}_K}>\delta. \tag{2.30}$$

We consider now the continuous function $\mathcal{I}$ restricted to the closed ball in $\mathcal{H}_{K,\boldsymbol{x}}$

$$\overline{B}_{\delta}^{\mathcal{H}_{K,\boldsymbol{x}}}(f_0)=\{f\in\mathcal{H}_{K,\boldsymbol{x}}\mid\|f-f_0\|_{\mathcal{H}_K}\leq\delta\},$$

which is compact, since the vector space $\mathcal{H}_{K,\boldsymbol{x}}$ is finite dimensional. This implies that the infimum of $\mathcal{I}$ on $\overline{B}_\delta^{\mathcal{H}_{K,\boldsymbol{x}}}(f_0)$ is attained. In view of (2.30) it follows that

$$\inf_{f\in\mathcal{H}_{K,\boldsymbol{x}}} \mathcal{I}(f) = \inf\{\mathcal{I}(f) \mid f \in \overline{B}_\delta^{\mathcal{H}_{K,\boldsymbol{x}}}(f_0)\},$$

and the proof is finished. ■

In view of Proposition 6, if the loss functions $V_{x_i}(y_i,\cdot)$ are convex for all $i = 1, 2, \ldots, l$, then the assumption on finite dimensionality of the spaces $\mathcal{W}_{x_i}$ for $i = 1, \ldots, l+u$, can be slightly weaker. We first record a result that is known but for which we include a proof, for the reader's convenience. To this end, we recall some basic definitions. If $\mathcal{V}$ is a vector space, a subset $A \subseteq \mathcal{V}$ is *convex* if for any $x, y \in A$ and $\lambda \in (0,1)$ we have $(1-\lambda)x+\lambda y \in A$. A function $f : \mathrm{Dom}(f) \to \mathbb{R}$ is *convex* if $\mathrm{Dom}(f)$ is a convex set in $\mathcal{V}$ and

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$$

for each $x, y \in \mathrm{Dom}(f)$ and $\lambda \in (0,1)$. In addition, f is *strictly convex* if

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda)f(y)$$

for each $x, y \in \mathrm{Dom}(f)$ such that $x \neq y$ and $\lambda \in (0,1)$.

Lemma 8 (a) *Given a vector space $\mathcal{V}$ that is endowed with a seminorm $\|\cdot\|$, the square of the seminorm $\|\cdot\|^2 : \mathcal{V} \to \mathbb{R}$ is a convex function on $\mathcal{V}$.*

(b) *If, in addition, $\|\cdot\|$ is a norm associated to an inner product on the real vector space $\mathcal{V}$, then $\|\cdot\|^2$ becomes strictly convex.*

Proof (a) Let $f_0, f_1 \in \mathcal{V}, \alpha \in (0,1)$. Then by triangle inequality we have

$$\|\alpha f_0 + (1-\alpha)f_1\| \leq \|\alpha f_0\| + \|(1-\alpha)f_1\| = \alpha\|f_0\| + (1-\alpha)\|f_1\|,$$

hence, by squaring both sides we get

$$\|\alpha f_0 + (1-\alpha)f_1\|^2 \leq \alpha^2\|f_0\|^2 + (1-\alpha)^2\|f_1\|^2 + 2\alpha(1-\alpha)\|f_0\|\|f_1\|.$$

Further on, if we add and subtract $-\alpha\|f_0\|^2 - (1-\alpha)\|f_1\|^2$ from the right hand side, we get

$$\begin{aligned}
\|\alpha f_0 + (1-\alpha)f_1\|^2 &\leq \alpha^2\|f_0\|^2 + (1-\alpha)^2\|f_1\|^2 + 2\alpha(1-\alpha)\|f_0\|\|f_1\| \\
&\quad - \alpha\|f_0\|^2 - (1-\alpha)\|f_1\|^2 + \alpha\|f_0\|^2 + (1-\alpha)\|f_1\|^2 \\
&= (\alpha^2-\alpha)\|f_0\|^2 + ((1-\alpha)^2 - (1-\alpha))\|f_1\|^2 \\
&\quad + 2\alpha(1-\alpha)\|f_0\|\|f_1\| + \alpha\|f_0\|^2 + (1-\alpha)\|f_1\|^2 \\
&= -\alpha(1-\alpha)(\|f_0\| - \|f_1\|)^2 + \alpha\|f_0\|^2 + (1-\alpha)\|f_1\|^2 \\
&\leq \alpha\|f_0\|^2 + (1-\alpha)\|f_1\|^2. \qquad (2.31)
\end{aligned}$$

This shows that $\|\cdot\|^2$ is convex.

(b) We assume now that $\|\cdot\|$ is a norm associated to an inner product $\langle\cdot,\cdot\rangle$ on a real vector space $\mathcal{V}$, that $f_0 \neq f_1$, $\alpha \in (0,1)$, and that

$$\|\alpha f_0 + (1-\alpha) f_1\|^2 = \alpha \|f_0\|^2 + (1-\alpha)\|f_1\|^2, \tag{2.32}$$

hence, by the last step in (2.31), it follows that $\|f_0\| = \|f_1\|$. Then, by (2.32) and since $\|f_0\| = \|f_1\|$, we get

$$\begin{aligned}\|f_0\|^2 = \|\alpha f_0 + (1-\alpha) f_1\|^2 &= \langle \alpha f_0 + (1-\alpha) f_1, \alpha f_0 + (1-\alpha) f_1\rangle \\ &= \alpha^2 \|f_0\|^2 + 2\alpha(1-\alpha)\langle f_0, f_1\rangle + (1-\alpha)^2 \|f_1\|^2.\end{aligned}$$

Taking into account that $\|f_0\| = \|f_1\|$ and that $\alpha(1-\alpha) \neq 0$, from here it follows that

$$\langle f_0, f_1\rangle = \|f_0\|^2 = \|f_1\|^2, \tag{2.33}$$

hence we have equality in the Schwarz inequality and, consequently, $f_0 = t f_1$ for some $t \in \mathbb{R}$. Since $\|f_0\| = \|f_1\|$ it follows that $t = \pm 1$. But $t = 1$ is not possible since $f_0 \neq f_1$, while $f_0 = -f_1$ is not possible because, by (2.33), this would imply $f_0 = 0 = f_1$. ■

Theorem 9 *Assume that all the underlying vector spaces are real, that the subspace $\mathcal{H}_{K,\boldsymbol{x}}$ is closed, and that the loss functions $V_{x_i}(y_i,\cdot)$ are convex for all $i = 1,2,\ldots,l$. Then, the minimisation problem* (2.14) *has a unique solution $f_{\boldsymbol{z},\gamma}$ and there exist $a_1,\ldots,a_{l+u}$, with $a_j \in \mathcal{W}_{x_j}$ for all $j = 1,\ldots,l+u$, such that*

$$f_{\boldsymbol{z},\gamma} = \sum_{j=1}^{l+u} K_{x_j} a_j.$$

Proof Consider the function $\boldsymbol{V}^l \colon \boldsymbol{Y}^l \times \boldsymbol{Y}^l \to \mathbb{R}$ defined by

$$\boldsymbol{V}^l(\boldsymbol{y},\boldsymbol{y}') := \sum_{j=1}^{l} V_{x_j}(y_j, y_j'), \quad \boldsymbol{y} = (y_1,\ldots,y_l),\ \boldsymbol{y}' = (y_1',\ldots,y_l'), \tag{2.34}$$

and observe that, for each fixed $\boldsymbol{y} \in \boldsymbol{Y}$, the function $\boldsymbol{V}^l(\boldsymbol{y},\cdot)$ is convex on $\boldsymbol{Y}$, since all maps V_{x_j} are convex in the second argument, $j = 1,\ldots,l$. Consequently, in the definition of $\mathcal{I}$ at (2.15), the first term is a convex function. Since the second term is a norm, it is a strictly convex function, while the third term is a seminorm, hence a convex function as well, by Lemma 8. Thus, $\mathcal{I}$ is a strictly convex function and hence the minimisation problem (2.14) has a unique solution. Then the conclusion follows from Proposition 6. ■

Remark 10 Theorem 9 contains Theorem 2 in Minh et al. (2016) in the case when the subspace $\mathcal{H}_{K,\boldsymbol{x}}$, see (2.18), is closed. This happens, for example, if the Hilbert space $\mathcal{W}$ in that theorem is finite dimensional. In Minh et al. (2016) the authors claim that the result

is true even in the case when $\mathcal{W}$ is an infinite dimensional space, which is not substantiated by the proof they provide. More precisely, the gap in that proof is that the subspace $\mathcal{H}_{K,\boldsymbol{x}}$ might not be finite dimensional and hence it might not be closed, which implies that, we have the decomposition (2.22) and not the decomposition (2.28). Consequently, the only conclusion that can be drawn is that the minimiser $f_{\boldsymbol{z},\gamma}$ belongs to the closure of $\mathcal{H}_{K,\boldsymbol{x}}$, and hence can only be approximated in the norm of $\mathcal{H}_K$ by sums of type $\sum_{j=1}^{l+u} K_{x_j} a_j$, but it may never equal such a sum.

3. Differentiable Loss Functions

3.1 Preliminary Results on Differentiable Optimisation.

Throughout this section, we assume that all vector spaces are real. The definitions and proofs of facts recalled in this subsection are from Peypouquet (2014). If $\mathcal{X}$ is a normed space, the *directional derivative* of a function $f\colon \mathrm{Dom}(f)(\subseteq \mathcal{X}) \to \mathbb{R}$ at an interior point $x \in \mathrm{Dom}(f)$ in the direction $h \in \mathcal{X}$ is given by

$$f'(x:h) = \lim_{t\to 0} \frac{f(x+th) - f(x)}{t},$$

provided that the limit exists. A function $f\colon \mathrm{Dom}(f)(\subseteq \mathcal{X}) \to \mathbb{R}$ is *Gâteaux differentiable* at an interior point $x \in \mathrm{Dom}(f)$ if f has directional derivatives for all directions at x and $\varphi_x(h) := f'(x:h)$ is linear and continuous in h. In this case, we denote the *Gâteaux derivative* $\nabla_x f \in \mathcal{B}(\mathcal{X}, \mathbb{R}) = X^*$ by the *gradient* notation

$$(\nabla_x f)h := \varphi_x(h), \quad h \in \mathcal{X}.$$

In general, if $\mathcal{X}$ and $\mathcal{Y}$ are Banach real spaces and $U \subseteq \mathcal{X}$ is open and $F : \mathcal{X} \to \mathcal{Y}$, then *$F$ has directional derivative for all directions at point $x \in U$* if

$$\lim_{\tau\to 0} \frac{F(x+\tau h) - F(x)}{\tau}$$

exists for any $h \in \mathcal{X}$. In this case we define the map $\nabla_x F\colon \mathcal{X} \to \mathcal{Y}$ as

$$(\nabla_x F)h := \lim_{\tau\to 0} \frac{F(x+\tau h) - F(x)}{\tau}, \quad h \in \mathcal{X}.$$

In the following we recall some basic facts.

Theorem 11 (Chain Rule for Directional Derivative) *Assume that $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ are Banach spaces, $F\colon \mathcal{X} \to \mathcal{Y}$, $G\colon \mathcal{Y} \to \mathcal{Z}$ and there exists $U \subseteq \mathcal{X}$ and $V \subseteq \mathcal{Y}$ open such that $F(U) \subseteq \mathcal{V}$, G has directional derivatives for all directions at $y \in \mathcal{V}$, and F has directional derivatives for all directions at $x \in U$. If $\nabla_x F$ and $\nabla_y G$ are continuous in $x \in U$ and $y \in V$, respectively, then, for any $h \in \mathcal{X}$, we have*

$$\nabla_x(G \circ F)(h) = \nabla_{F(x)} G(\nabla_x F(h)).$$

Let $\mathcal{X}$ be a Banach real space. For any fixed $x^* \in \mathcal{X}^* := \mathcal{B}(\mathcal{X}; \mathbb{R})$ and any $y \in \mathcal{X}$ we denote

$$\langle x^*, y\rangle := x^*(y) \in \mathbb{R}. \tag{3.1}$$

Remark 12 If $\mathcal{H}$ is a real Hilbert space, by Riesz-Fréchet Representation Theorem we have $\langle y, x^* \rangle = \langle y, f_{x^*} \rangle_{\mathcal{H}}$ for a unique $f_{x^*} \in \mathcal{H}$. For simplicity we denote f_{x^*} as x^*.

Let $f\colon \mathrm{Dom}(f)(\subseteq \mathcal{X}) \to \mathbb{R}$ be convex. A point $x^* \in \mathcal{X}^*$ is a *subgradient* of f at x if

$$f(y) \leq f(x) + \langle x^*, y - x \rangle$$

holds for all y in a neighbourhood of x. The set of all subgradients of f at x is the *subdifferential* of f at x and is denoted by $\partial f(x)$. If $\partial f(x) \neq \emptyset$, we say f is *subdifferentiable* at x. The *domain of subdifferential* is denoted as $\mathrm{Dom}(\partial f) = \{x \in X \mid \partial f(x) \neq \emptyset\}$. By definition $\mathrm{Dom}(\partial f) \subseteq \mathrm{Dom}(f)$.

Theorem 13 (Fermat's Rule) *Let $f\colon \mathcal{X} \to \mathbb{R}$ be convex. Then $\hat{x}$ is a global minimiser of f if and only if $0 \in \partial f(\hat{x})$.*

Let A be an open subset of $\mathcal{X}$ and $T : A \to \mathcal{Y}$. Then given a point $x \in A$, T is *Fréchet differentiable* at x if there exists a bounded linear operator $L_x\colon \mathcal{X} \to \mathcal{Y}$ such that

$$\lim_{h \to 0} \frac{\|T(x+h) - T(x) - L_x h\|_Y}{\|h\|_X} = 0.$$

In this case, due to the uniqueness of L_x we define the linear bounded operator $\mathrm{D}_x T : \mathcal{X} \to \mathcal{Y}$ as $\mathrm{D}_x T := L_x$ and call it the *Fréchet derivative* of T at x.

Theorem 14 *Let $(\mathcal{H}, \langle ., . \rangle_{\mathcal{H}})$ be a Hilbert real space. Then, given $\psi \in \mathcal{X}$ and $F, G\colon \mathcal{X} \to \mathcal{H}$ maps that are Fréchet differentiable at ψ, we have*

$$\mathrm{D}_\psi \langle F\cdot, G\cdot \rangle_{\mathcal{H}} (h) = \langle (\mathrm{D}_\psi F)(h), G(\psi) \rangle_{\mathcal{H}} + \langle F(\psi), (\mathrm{D}_\psi G)(h) \rangle_{\mathcal{H}}$$

for any $h \in \mathcal{H}$.

Remark 15 In Theorem 14 if we further assume that F, G are bounded linear operators, then we get

$$\mathrm{D}_\psi \langle F\cdot, G\cdot \rangle_{\mathcal{H}} (h) = \langle Fh, G\psi \rangle_{\mathcal{H}} + \langle F\psi, Gh \rangle_{\mathcal{H}}$$

for any $h \in \mathcal{H}$.

Theorem 16 *Let $\mathcal{X}, \mathcal{Y}$ be Banach spaces. Consider a nonempty open set $U \subset \mathcal{X}$ and a map $F\colon U \to \mathcal{Y}$. If F is Fréchet differentiable at $x \in U$ then it is Gâteaux differentiable at x and $\nabla_x F = \mathrm{D}_x F$.*

3.2 The Representer Theorem for Locally Differentiable Loss Functions.

The notation in this subsection is the same as in Section 2, only that all vector spaces are over the real field $\mathbb{R}$. In this subsection, we show that, if we add the assumption that for any $i = 1, \ldots, l$, $V_{x_i}(y_i, \cdot) : \mathcal{Y}_x \to \mathbb{R}$ is Gâteaux differentiable, then we can allow $\mathcal{W}_{x_i}$, for all $i = 1, \ldots, l$, to be infinite dimensional and get the same conclusion as in Proposition 6. The proof of this theorem is inspired to a certain extent by the proof of Theorem 3 in Minh et al. (2016), that was proven for the special case of the least squares loss function.

As in the proof of Lemma 5, let the sampling operator $S_{\mathbf{x}} : \mathcal{H}_K \to \boldsymbol{W}^{l+u}$ be defined as in (2.19) where $\boldsymbol{x} = (x_1, \ldots, x_{l+u})$. Let also $E_{\boldsymbol{C},\mathbf{x}} : \mathcal{H}_K \to \boldsymbol{Y}^l$ be defined as in (2.21) and the Hilbert space $\boldsymbol{Y}^l$ be defined as in (2.20). Define the function $\boldsymbol{V}^l \colon \boldsymbol{Y}^l \times \boldsymbol{Y}^l \to \mathbb{R}$ as in (2.34) and then denote $\boldsymbol{V}^l_{\boldsymbol{y}} \colon \boldsymbol{Y}^l \to \mathbb{R}$ as

$$\boldsymbol{V}^l_{\boldsymbol{y}}(\boldsymbol{y}') := \boldsymbol{V}^l(\boldsymbol{y}, \boldsymbol{y}'), \quad \boldsymbol{y}' = (y'_1, \ldots, y'_l) \in \boldsymbol{Y}^l, \tag{3.2}$$

where $\boldsymbol{y} = (y_1, y_2, \ldots, y_l) \in \boldsymbol{Y}^l$.

In the next theorem, which is the main result of this section, please note that all Hilbert spaces $\mathcal{W}_{x_1}, \ldots, \mathcal{W}_{x_{l+u}}$ are allowed to be infinite dimensional, without any restriction except that they are real.

Theorem 17 *Assume that for any $i = 1, \ldots, l$, the loss function $V_{x_i}(y_i, \cdot)$ is Gâteaux differentiable. If the minimisation problem (2.14) has a solution $f_{\boldsymbol{z},\gamma} \in \mathcal{H}_K$ then there exist $a_1, \ldots, a_{l+u}$, with $a_j \in \mathcal{W}_{x_j}$ for all $j = 1, \ldots, l+u$, such that*

$$f_{\boldsymbol{z},\gamma} = \sum_{j=1}^{l+u} K_{x_j} a_j, \tag{3.3}$$

where the vectors $a_i \in \mathcal{W}_{x_i}$, $i = 1, \ldots, l+u$, satisfy the following system

$$2\gamma_A l a_i + 2\gamma_I l \sum_{j,k=1}^{l+u} M_{i,j} K(x_i, x_j) a_j = -C^*_{x_i}(\nabla_{E_{\boldsymbol{C},\boldsymbol{x}} f_{\boldsymbol{z},\gamma}} \boldsymbol{V}^l_{\boldsymbol{y}})_i, \qquad \text{if } i = 1, \ldots, l, \tag{3.4}$$

$$\gamma_A a_i + \gamma_I \sum_{j,k=1}^{l+u} M_{i,j} K(x_i, x_j) a_j = 0, \qquad \text{if } i = l+1, \ldots, l+u. \tag{3.5}$$

Here, for arbitrary $y' \in \boldsymbol{Y}^l$, we abuse some notation and denote $\nabla_{y'} \boldsymbol{V}^l_{\boldsymbol{y}}$ both as the original functional (Gâteaux derivative) and the vector that represents this functional following the notation as in (3.1).

Proof We can rewrite the map $\mathcal{I}$ to be minimised, see (2.15), as

$$\mathcal{I}(f) = \frac{1}{l} \boldsymbol{V}^l_{\boldsymbol{y}}(E_{\boldsymbol{C},\boldsymbol{x}} f) + \gamma_A \|f\|^2_{\mathcal{H}_K} + \gamma_I \langle \boldsymbol{f}, M \boldsymbol{f} \rangle_{\boldsymbol{W}^{l+u}}, \quad f \in \mathcal{H}_K. \tag{3.6}$$

Assuming that the minimisation problem (2.14) has a solution $f_{\boldsymbol{z},\gamma}$, which is an interior point in the domain of $\mathcal{I}$ and that $\boldsymbol{V}^l_{\boldsymbol{y}}$ is Gâteaux differentiable, by Theorem 11, the Gâteaux derivative evaluated at $f \in \mathcal{H}_K$ is

$$\nabla_f (\boldsymbol{V}^l_{\boldsymbol{y}} \circ E_{\boldsymbol{C},\boldsymbol{x}})(h) = \nabla_{E_{\boldsymbol{C},\boldsymbol{x}} f} \boldsymbol{V}^l_{\boldsymbol{y}} (\nabla_f E_{\boldsymbol{C},\boldsymbol{x}}(h)), \quad h \in \mathcal{H}_K.$$

Then, by using the identification as in (3.1) and Theorem 16, for all $h \in \mathcal{H}_K$ we get

$$\begin{aligned} \nabla_{E_{\boldsymbol{C},\boldsymbol{x}} f} \boldsymbol{V}^l_{\boldsymbol{y}} (\nabla_f E_{\boldsymbol{C},\boldsymbol{x}}(h)) &= \left\langle \nabla_{E_{\boldsymbol{C},\boldsymbol{x}} f} \boldsymbol{V}^l_{\boldsymbol{y}}, \mathrm{D}_f E_{C,\mathbf{x}} h \right\rangle = \left\langle \nabla_{E_{\boldsymbol{C},\boldsymbol{x}} f} \boldsymbol{V}^l_{\boldsymbol{y}}, E_{C,\mathbf{x}} h \right\rangle_{\boldsymbol{Y}^l} \\ &= \left\langle E^*_{C,\mathbf{x}} \nabla_{E_{\boldsymbol{C},\boldsymbol{x}} f} \boldsymbol{V}^l_{\boldsymbol{y}}, h \right\rangle_{\mathcal{H}_K}. \end{aligned} \tag{3.7}$$

Since $\|f\|^2_{\mathcal{H}_K} = \langle f, f\rangle_{\mathcal{H}_K}$, by Theorem 14, the map $\mathcal{H}_k \mapsto \|f\|^2_{\mathcal{H}_K}$ is Fréchet differentiable and hence, by Theorem 16, it is Gâteaux differentiable. Thus, by Theorem 14 and since $\mathcal{H}_K$ is real we have

$$\gamma_A \nabla_f \|\cdot\|^2_{\mathcal{H}_K}(h) = 2\gamma_A \langle f, h\rangle_{\mathcal{H}_K}, \quad h \in \mathcal{H}_K. \tag{3.8}$$

Again by using Theorem 14 we get

$$\begin{aligned}\gamma_I \nabla_f \langle S_{\boldsymbol{x}}\cdot, MS_{\boldsymbol{x}}\cdot\rangle_{\boldsymbol{W}^{l+u}}(h) &= \gamma_I \langle S_{\boldsymbol{x}} f, MS_{\boldsymbol{x}} h\rangle_{\boldsymbol{W}^{l+u}} + \gamma_I \langle S_{\boldsymbol{x}} h, MS_{\boldsymbol{x}} f\rangle_{\boldsymbol{W}^{l+u}} \\ &= 2\gamma_I \langle S^*_{\boldsymbol{x}} MS_{\boldsymbol{x}} f, h\rangle_{\mathcal{H}_K}, \quad h \in \mathcal{H}_K. \end{aligned} \tag{3.9}$$

Summing up (3.7), (3.8) and (3.9) we get

$$\begin{aligned}\nabla_f \mathcal{I}(h) &= l^{-1}\left\langle E^*_{C,\mathbf{x}} \nabla_{E_{\boldsymbol{C},\boldsymbol{x}} f} \boldsymbol{V}^l_{\boldsymbol{y}}, h\right\rangle_{\mathcal{H}_K} + 2\gamma_A \langle f, h\rangle_{\mathcal{H}_K} + 2\gamma_I \langle S^*_{\boldsymbol{x}} MS_{\boldsymbol{x}} f, h\rangle_{\mathcal{H}_K} \\ &= \left\langle l^{-1} E^*_{C,\mathbf{x}} \nabla_{E_{\boldsymbol{C},\boldsymbol{x}} f} \boldsymbol{V}^l_{\boldsymbol{y}} + 2\gamma_A f + 2\gamma_I S^*_{\boldsymbol{x}} MS_{\boldsymbol{x}} f, h\right\rangle_{\mathcal{H}_K}, \quad h \in \mathcal{H}_K. \end{aligned} \tag{3.10}$$

Keeping Fermat's Theorem 13 in mind, since that minimiser is an interior point of the domain, we should have the minimiser $f_{\boldsymbol{z},\gamma} \in \mathcal{H}_K$ such that $\nabla_{f_{\boldsymbol{z},\gamma}} \mathcal{I}(\cdot) = 0$. By (3.10), this means

$$\nabla_{f_{\boldsymbol{z},\gamma}} \mathcal{I}(h) = \left\langle l^{-1} E^*_{C,\mathbf{x}} \nabla_{E_{\boldsymbol{C},\boldsymbol{x}} f_{\boldsymbol{z},\gamma}} \boldsymbol{V}^l_{\boldsymbol{y}} + 2\gamma_A f_{\boldsymbol{z},\gamma} + 2\gamma_I S^*_{\boldsymbol{x}} MS_{\boldsymbol{x}} f_{\boldsymbol{z},\gamma}, h\right\rangle_{\mathcal{H}_K} = 0 \tag{3.11}$$

for any $h \in \mathcal{H}_K$. Thus, we get

$$\frac{1}{l} E^*_{C,\mathbf{x}} \nabla_{E_{\boldsymbol{C},\boldsymbol{x}} f_{\boldsymbol{z},\gamma}} \boldsymbol{V}^l_{\boldsymbol{y}} + 2\gamma_A f_{\boldsymbol{z},\gamma} + 2\gamma_I S^*_{\boldsymbol{x}} MS_{\boldsymbol{x}} f_{\boldsymbol{z},\gamma} = 0$$

and hence

$$f_{\boldsymbol{z},\gamma} = \frac{-1}{2\gamma_A l} E^*_{C,\mathbf{x}} \nabla_{E_{\boldsymbol{C},\boldsymbol{x}} f_{\boldsymbol{z},\gamma}} \boldsymbol{V}^l_{\boldsymbol{y}} - \frac{\gamma_I}{\gamma_A} S^*_{\boldsymbol{x}} MS_{\boldsymbol{x}} f_{\boldsymbol{z},\gamma}.$$

Then by using the facts that $E^*_{\boldsymbol{C},\boldsymbol{x}} = [K_{x_1} C^*_{x_1} \quad K_{x_2} C^*_{x_2} \quad \dots \quad K_{x_l} C^*_{x_l}]$, see (2.21), and that, by definition, $S_{\boldsymbol{x}} = [K^*_{x_1} \quad K^*_{x_2} \quad \dots \quad K^*_{x_{l+u}}]^T$, see (2.19), we get

$$f_{\boldsymbol{z},\gamma} = \frac{-1}{2\gamma_A l} \sum_{i=1}^{l} K_{x_i} C^*_{x_i} (\nabla_{E_{\boldsymbol{C},\boldsymbol{x}} f_{\boldsymbol{z},\gamma}} \boldsymbol{V}^l_{\boldsymbol{y}})_i - \frac{\gamma_I}{\gamma_A} \sum_{i=1}^{l+u} K_{x_i} (MS_{\boldsymbol{x}} f_{\boldsymbol{z},\gamma})_i.$$

Thus, we can represent $f_{\boldsymbol{z},\gamma}$ as in (3.3), where

$$a_i = \begin{cases} \dfrac{-1}{2\gamma_A l} C^*_{x_i} (\nabla_{E_{\boldsymbol{C},\boldsymbol{x}} f_{\boldsymbol{z},\gamma}} \boldsymbol{V}^l_{\boldsymbol{y}})_i - \dfrac{\gamma_I}{\gamma_A} (MS_{\boldsymbol{x}} f_{\boldsymbol{z},\gamma})_i, & \text{if } i = 1, \dots, l \\ -\dfrac{\gamma_I}{\gamma_A} (MS_{\boldsymbol{x}} f_{\boldsymbol{z},\gamma})_i, & \text{otherwise.} \end{cases} \tag{3.12}$$

Since

$$(MS_{\boldsymbol{x}} f_{\boldsymbol{z},\gamma})_i = \sum_{k=1}^{l+u} M_{i,k} \sum_{j=1}^{l+u} K(x_k, x_j) a_j, \quad i = 1, \dots, l+u,$$

it follows that

$$a_i = \begin{cases} \dfrac{-1}{2\gamma_A l} C^*_{x_i}(\nabla_{E_{\boldsymbol{C},\boldsymbol{x}} f_{\boldsymbol{z},\gamma}} \boldsymbol{V}^l_{\boldsymbol{y}})_i - \dfrac{\gamma_I}{\gamma_A} \displaystyle\sum_{j,k=1}^{l+u} M_{i,k} K(x_k, x_j) a_j, & \text{if } i = 1, \dots, l, \\ \\ -\dfrac{\gamma_I}{\gamma_A} \displaystyle\sum_{j,k=1}^{l+u} M_{i,k} K(x_k, x_j) a_j, & \text{otherwise,} \end{cases}$$

which is equivalent to the system of equations (3.4) and (3.5). ■

Remark 18 The system of equations (3.4) and (3.5) can be reformulated by means of operators; to be compared with Theorem 4 in Minh et al. (2016) that was obtained for the special case of the least squares function. Let $K[\mathbf{x}]$ denote the $(l+u) \times (l+u)$ operator valued matrix whose (i,j) entry is $K(x_i, x_j)$, and let $\mathbf{v} = (v_1, \dots, v_{l+u})$ be the vector with entries

$$v_i = \begin{cases} -C^*_{x_i}(\nabla_{E_{\boldsymbol{C},\boldsymbol{x}} f_{\boldsymbol{z},\gamma}} \boldsymbol{V}^l_{\boldsymbol{y}})_i & \text{if } i = 1, \dots, l \\ 0, & \text{otherwise,} \end{cases}$$

where $C^*_x : \mathcal{Y}_x \to \mathcal{W}_x$ for any $x \in X$. Then, with notation and assumptions as in Theorem 17 and letting $\mathbf{a} = (a_1, \dots, a_{l+u})$, a simple algebraic calculation, that we leave to the reader, shows that the system of equations (3.4) and (3.5) coincides with the operator equation

$$(2l\gamma_I M K[\mathbf{x}] + 2l\gamma_A I)\, \mathbf{a} = \mathbf{v}, \tag{3.13}$$

where both $\mathbf{a}$ and $\mathbf{v}$ are considered as column vectors. However, at this level of generality, from here we cannot get the unknown vector $\mathbf{a}$ because it appears in the vector $\mathbf{v}$ as well. From this point of view, the equation (3.13) is more an implicit form.

Corollary 19 *With notation and assumptions as in Theorem 17, assume that the loss functions V_{x_i} are convex for all $i = 1, 2, \dots, l$. Then the minimisation problem* (2.14) *has a unique solution $f_{\boldsymbol{z},\gamma}$ and there exist $a_1, \dots, a_{l+u}$, with $a_j \in \mathcal{W}_{x_j}$ for all $j = 1, \dots, l+u$, such that*

$$f_{\boldsymbol{z},\gamma} = \sum_{j=1}^{l+u} K_{x_j} a_j,$$

where the vectors $a_1, \dots, a_{l+u}$ satisfy the system of equations (3.4) *and* (3.5)*.*

Proof The argument is the same as in Theorem 9 and then use Theorem 17. ■

Theorem 17 and its Corollary 19 have a theoretical significance and a less practical importance because, in the system (3.4) and (3.5), the unknowns $a_1, \dots, a_{l+u}$ appear also on the right hand side and hence it is more an implicit form of expressing them and not an explicit one. Because of that, for specified loss functions V, one should work further on these expressions in order to obtain explicit or, at least computable, solutions. In the next two subsections, we work out the details and show how Theorem 17 and its Corollary 19 can be improved for special loss functions, the least squares functions, similar to the results in Minh et al. (2016), and the exponential least squares functions, and compare the formulae.

3.3 The Least Squares Loss Function.

If all the loss functions are the least squares, see Example 2.(1), the minimisation function (2.15) becomes

$$\begin{aligned}\mathcal{I}(f) &:= \frac{1}{l}\sum_{j=1}^{l}\|y_j - C_{x_j}f(x_j)\|^2_{\mathcal{Y}_{x_j}} + \gamma_A\|f\|^2_{\mathcal{H}_K} + \gamma_I\langle \boldsymbol{f}, M\boldsymbol{f}\rangle_{\boldsymbol{W}^{l+u}} \\ &= \frac{1}{l}\sum_{j=1}^{l}\|y_j - C_{x_j}K^*_{x_j}f\|^2_{\mathcal{Y}_{x_j}} + \gamma_A\|f\|^2_{\mathcal{H}_K} + \gamma_I\langle \boldsymbol{f}, M\boldsymbol{f}\rangle_{\boldsymbol{W}^{l+u}}, \end{aligned} \tag{3.14}$$

Since, as functions of $f \in \mathcal{H}_K$, all the terms are convex and the middle term is strictly convex, see Lemma 8, the minimisation function $\mathcal{I}(\cdot)$ is strictly convex and hence the minimisation problem has unique solution. Also, $\mathcal{I}(\cdot)$ is Fréchet differentiable, hence Corollary 19 is applicable. According to Fermat's Theorem, this unique solution f should vanish the gradient. But, for each $h \in \mathcal{H}_K$, we have

$$\nabla_f\mathcal{I}(f)h = \frac{2}{l}\langle E^*_{\boldsymbol{C},\boldsymbol{x}}E_{\boldsymbol{C},\boldsymbol{x}}f, h\rangle_{\mathcal{H}_K} - \frac{2}{l}\langle E^*_{\boldsymbol{C},\boldsymbol{x}}\boldsymbol{y}, h\rangle_{\mathcal{H}_K} + 2\gamma_A\langle f, h\rangle_{\mathcal{H}_K} + 2\gamma_I\langle S_{\boldsymbol{x}}MS_{\boldsymbol{x}}\boldsymbol{f}, h\rangle_{\mathcal{H}_K},$$

hence,

$$E^*_{\boldsymbol{C},\boldsymbol{x}}E_{\boldsymbol{C},\boldsymbol{x}}f + l\gamma_A f + l\gamma_I S^*_{\boldsymbol{x}}MS_{\boldsymbol{x}}\boldsymbol{f} - E^*_{\boldsymbol{C},\boldsymbol{x}}\boldsymbol{y} = 0. \tag{3.15}$$

Since $\gamma_A > 0$ this is equivalent with

$$f = \frac{1}{l\gamma_A}E^*_{\boldsymbol{C},\boldsymbol{x}}\boldsymbol{y} - \frac{1}{l\gamma_A}E^*_{\boldsymbol{C},\boldsymbol{x}}E_{\boldsymbol{C},\boldsymbol{x}}f - \frac{\gamma_I}{\gamma_A}S^*_{\boldsymbol{x}}MS_{\boldsymbol{x}}\boldsymbol{f},$$

explicitly,

$$f = \sum_{i=1}^{l}K_{x_i}\Big(\frac{1}{l\gamma_A}C^*_{x_i}y_i\Big) + \sum_{i=1}^{l}K_{x_i}\Big(-\frac{1}{l\gamma_A}C^*_{x_i}C_{x_i}f(x_i)\Big) + \sum_{i=1}^{l+u}K_{x_i}\Big(-\frac{\gamma_I}{\gamma_A}\sum_{k=1}^{l+u}M_{i,k}f(x_k)\Big). \tag{3.16}$$

In this special case, the representation (3.16) improves Theorem 17 by obtaining the representation of the optimal solution as $f = \sum_{i=1}^{l+u}K_{x_i}a_i$, where,

$$a_i = \frac{1}{l\gamma_A}C^*_{x_i}y_i - \frac{1}{l\gamma_A}C^*_{x_i}C_{x_i}f(x_i) - \frac{\gamma_I}{\gamma_A}\sum_{k=1}^{l+u}M_{i,k}f(x_k), \quad \text{for all } i = 1,\ldots,l, \tag{3.17}$$

and,

$$a_i = -\frac{\gamma_I}{\gamma_A}\sum_{k=1}^{l+u}M_{i,k}f(x_k), \quad \text{for all } i = l+1,\ldots,l+u. \tag{3.18}$$

Then, since for all $k = 1,\ldots,l+u$, we have

$$f(x_k) = \sum_{j=1}^{l+u}K_{x_j}(x_k)a_j = \sum_{j=1}^{l+u}K(x_k,x_j)a_j,$$

and, consequently, from (3.17) and (3.18), we get

$$a_i + \sum_{j=1}^{l+u}\Big[\frac{1}{l\gamma_A}C^*_{x_i}C_{x_i}K(x_i,x_j) + \frac{\gamma_I}{\gamma_A}\sum_{k=1}^{l+u}M_{i,k}K(x_k,x_j)\Big]a_j = \frac{1}{l\gamma_A}C^*_{x_i}y_i,\ i=1,\dots,l, \quad (3.19)$$

$$a_i + \sum_{j=1}^{l+u}\Big[\frac{\gamma_I}{\gamma_A}\sum_{k=1}^{l+u}M_{i,k}K(x_k,x_j)\Big]a_j = 0,\ i=l+1,\dots,l+u. \quad (3.20)$$

Equations (3.19) and (3.20) make a system of linear equations which can be treated very efficiently by computational techniques, similarly as in Minh et al. (2016) and the literature cited there. This substantiates our claim that the localised versions offer more flexibility in modelling learning problems without bringing additional obstructions.

3.4 The Exponential Least Squares Loss Function.

If all the loss functions are the exponential least squares, see Example 2.(4), the minimisation function (2.15) becomes

$$\begin{aligned}\mathcal{I}(f) &:= 1 - \frac{1}{l}\sum_{j=1}^{l}\exp(-\|y_j - C_{x_j}f(x_j)\|^2_{\mathcal{Y}_{x_j}}) + \gamma_A\|f\|^2_{\mathcal{H}_K} + \gamma_I\langle \boldsymbol{f}, M\boldsymbol{f}\rangle_{\boldsymbol{W}^{l+u}}\\ &= 1 - \frac{1}{l}\sum_{j=1}^{l}\exp(-\|y_j - C_{x_j}K^*_{x_j}f\|^2_{\mathcal{Y}_{x_j}}) + \gamma_A\|f\|^2_{\mathcal{H}_K} + \gamma_I\langle \boldsymbol{f}, M\boldsymbol{f}\rangle_{\boldsymbol{W}^{l+u}}. \quad (3.21)\end{aligned}$$

In this case, since the second term is not, in general, convex, we cannot conclude that the function $\mathcal{I}$ is convex. However, $\mathcal{I}(f)\geq 0$ for all $f\in\mathcal{H}_K$ and $\lim_{\|f\|\to\infty}\mathcal{I}(f) = +\infty$, hence the minimisation problem has at least one solution f, but this solution might not be unique. Anyhow, because any solution f is an interior point in $\mathcal{H}_K$, Fermat's Rule is applicable, hence $\nabla_f\mathcal{I} = 0$. Taking advantage of the calculations performed in the previous subsection, see (3.15), by calculating the gradient of $\mathcal{I}$ and then, by Fermat's Rule, the optimal function f should satisfy the following equation

$$\exp(-\|y_j - C_{x_j}f(x_j)\|^2_{\mathcal{Y}_{x_j}})\big(E^*_{\boldsymbol{C},\boldsymbol{x}}E_{\boldsymbol{C},\boldsymbol{x}}f - E^*_{\boldsymbol{C},\boldsymbol{x}}\boldsymbol{y}\big) + l\gamma_A f + l\gamma_I S^*_{\boldsymbol{x}}MS_{\boldsymbol{x}}\boldsymbol{f} = 0. \quad (3.22)$$

Since $\gamma_A > 0$ this is equivalent with

$$f = \frac{\exp(-\|y_j - C_{x_j}f(x_j)\|^2_{\mathcal{Y}_{x_j}})}{l\gamma_A}\big(E^*_{\boldsymbol{C},\boldsymbol{x}}\boldsymbol{y} - E^*_{\boldsymbol{C},\boldsymbol{x}}E_{\boldsymbol{C},\boldsymbol{x}}f\big) - \frac{\gamma_I}{\gamma_A}S^*_{\boldsymbol{x}}MS_{\boldsymbol{x}}\boldsymbol{f},$$

explicitly,

$$\begin{aligned}f = \sum_{i=1}^{l}K_{x_i}\Big(&\frac{\exp(-\|y_j - C_{x_j}f(x_j)\|^2_{\mathcal{Y}_{x_j}})}{l\gamma_A}\big(C^*_{x_i}y_i - C^*_{x_i}C_{x_i}f(x_i)\big)\Big)\\ &+ \sum_{i=1}^{l+u}K_{x_i}\Big(-\frac{\gamma_I}{\gamma_A}\sum_{k=1}^{l+u}M_{i,k}f(x_k)\Big). \quad (3.23)\end{aligned}$$

In this special case, (3.23) improves Theorem 17 by obtaining the representation of the optimal solution as $f = \sum_{i=1}^{l+u} K_{x_i} a_i$, where,

$$a_i = \frac{\exp(-\|y_j - C_{x_j} f(x_j)\|^2_{\mathcal{Y}_{x_j}})}{l\gamma_A}\big(C^*_{x_i} y_i - C^*_{x_i} C_{x_i} f(x_i)\big), \quad \text{for all } i = 1, \ldots, l, \tag{3.24}$$

and,

$$a_i = -\frac{\gamma_I}{\gamma_A} \sum_{k=1}^{l+u} M_{i,k} f(x_k), \quad \text{for all } i = l+1, \ldots, l+u. \tag{3.25}$$

Then, since for all $j = 1, \ldots, l+u$, we have

$$f(x_j) = \sum_{k=1}^{l+u} K_{x_k}(x_j) a_j = \sum_{k=1}^{l+u} K(x_j, x_k) a_k,$$

and, consequently, from (3.24) and (3.25), we get

$$a_i + \sum_{j=1}^{l} \Big[\frac{\exp(-\|y_j - C_{x_j} \sum_{k=1}^{l+u} K(x_j, x_k) a_k\|^2_{\mathcal{Y}_{x_j}})}{l\gamma_A} C^*_{x_i} C_{x_i} K(x_i, x_j) + \frac{\gamma_I}{\gamma_A} \sum_{k=1}^{l+u} M_{i,k} K(x_k, x_j) \Big] a_j$$
$$= \frac{1}{l\gamma_A} C^*_{x_i} y_i, \ i = 1, \ldots, l, \tag{3.26}$$

$$a_i + \frac{\gamma_I}{\gamma_A} \sum_{j=1}^{l+u} \sum_{k=1}^{l+u} M_{i,k} K(x_k, x_j) a_j = 0, \ i = l+1, \ldots, l+u. \tag{3.27}$$

Equations (3.26) and (3.27) make a system of equations with respect to the unkowns a_i for $i = 1, \ldots, l+u$, which consists of nonlinear equations for the unknowns a_i corresponding to the labeled input points and of linear equations for the unknowns a_i corresponding to the unlabeled input points.

3.5 Available Numerical Methods for the Exponential Least Square Loss Function.

In this subsection we tackle the question of deriving algorithms to solve the system of equations defined by (3.26) and (3.27). We use the same notations and assumptions as in the previous subsection. In addition, we assume that all Hilbert spaces $\mathcal{W}_{x_i}$ have finite dimensions. To be more precise, $\mathcal{W}_{x_i}$ is identified with $\mathbb{R}^{d_i}$, for $i = 1, \ldots, l+u$. We first define the vector

$$\boldsymbol{a} = (a_1, a_2, \ldots, a_{l+u}) \in \mathbb{R}^N, \tag{3.28}$$

where $a_j \in \mathcal{W}_{x_j} = \mathbb{R}^{d_j}$, for each $j = 1, \ldots, l+u$, and

$$N = \sum_{i=1}^{l+u} \dim(\mathcal{W}_{x_i}) = \sum_{i=1}^{l+u} d_i.$$

Then we consider the function $H\colon \mathbb{R}^N \to \mathbb{R}^N$ with

$$H(\boldsymbol{a}) = (H_1(\boldsymbol{a}), H_2(\boldsymbol{a}), \dots, H_{l+u}(\boldsymbol{a})), \quad \boldsymbol{a} \in \mathbb{R}^N,$$

defined by

$$H_i(\boldsymbol{a}) = a_i + \frac{1}{l\gamma_A} \sum_{j=1}^{l} \exp\bigl(-\|y_j - C_{x_j} \sum_{k=1}^{l+u} K(x_j, x_k) a_k\|^2_{\mathcal{Y}_{x_j}}\bigr) C^*_{x_i} C_{x_i} K(x_i, x_j) a_j$$
$$+ \frac{\gamma_I}{\gamma_A} \sum_{j=1}^{l+u} \sum_{k=1}^{l+u} M_{i,k} K(x_k, x_j) a_j - \frac{1}{l\gamma_A} C^*_{x_i} y_i, \quad i = 1, \dots, l, \tag{3.29}$$

$$H_i(\boldsymbol{a}) = a_i + \frac{\gamma_I}{\gamma_A} \sum_{j=1}^{l+u} \sum_{k=1}^{l+u} M_{i,k} K(x_k, x_j) a_j, \quad i = l+1, \dots, l+u. \tag{3.30}$$

In view of the system of equations defined by (3.26) and (3.27) with the unknown vector solution $\boldsymbol{a} \in \mathbb{R}^N$ as in (3.28), we search for solutions of the equation $H(\boldsymbol{a}) = 0$. From the numerical analysis point of view, this problem can be approached by nonlinear optimisation techniques, more precisely, we search for an algorithm that yields a sequence $(\boldsymbol{a}_n)_{n\geq 0}$ of vectors in $\mathbb{R}^N$ with the property that for each $\epsilon > 0$ there exists an integer $n \geq 0$ such that $\|H(\boldsymbol{a}_n)\| < \epsilon$. One of the classical approaches for this kind of nonlinear problems is in the class of Newton's damped approximation methods, see Hazely et al. (2022) for an overview and advances in second-order approximation methods and machine learning.

There are different methods and algorithms for dealing with constrained nonlinear systems of differentiable functions. For example, the *potential reduction Newton's method* in Monteiro and Pang (1999) can be used in order to provide an algorithm to approximate solutions $\boldsymbol{a} \in \mathbb{R}^N$ of the equation $H(\boldsymbol{a}) = 0$ under the assumption that the Jacobi matrix $\nabla_{\boldsymbol{u}} H$ is nonsingular for all $\boldsymbol{u} \in \mathbb{R}^N$. Other methods based on the interior point methods are available, for example see Byrd et al. (1999), Byrd et al. (2000), and Waltz et al. (2006). The latter algorithm is implemented in the `fmincon` function in MATLAB, that we have used in our example.

Because the optimisation problem is nonlinear, multiple solutions may show up and this makes the choice of the initialisation vector $\boldsymbol{a}_0$ very important: for different choices of the initialisation vector different pools of solutions of the equations $H(\boldsymbol{a}) = 0$ might be found. Let us observe that, on the one hand, the components H_i for $i = l+1, \dots, l+u$, corresponding to unlabeled points x_i, are linear and homogeneous and hence that 0 is a solution. On the other hand, the components H_i for $i = 1, \dots, l$, corresponding to labeled points x_i, are nonlinear and nonhomogeneous and hence that 0 is not a solution. From here, we can see that, on the one hand, as the learning problem is semisupervised, l is significantly less than u and hence taking the initialisation vector $\boldsymbol{a}_0 = 0$ might be a good choice for the beginning. On the other hand, in order to find better minimisers, or even the global minimiser, some multigrid methods or stochastic methods for the choice of the initialisation vector $\boldsymbol{a}_0$ might be involved, e.g. see Gower et al. (2024). In our example, we have used the Latin Hypercube Sampling (LHS), see McKay et al. (1979), to solve this issue. Similarly to multigrid methods, the LHS divides the domain into small cubes. Without loss of generality, assume that the domain is $[0,1]^d$ a cube with dimension d. Then for a fixed

n, split the cube into n^d subcubes by splitting each interval $[0,1]$ into intervals with $1/n$ length, that is,

$$[0,1] = \bigcup_{i=0}^{n-1} \left[\frac{i}{n}, \frac{i+1}{n}\right].$$

Then we pick points such that any rectangle $R_{i,j}$ defined below contains only one point

$$R_{i,j} = [0,1]^{j-1} \times \left[\frac{i}{n}, \frac{i+1}{n}\right] \times [0,1]^{d-j}, \quad i,j = 1,\ldots,n.$$

This sampling strategy enjoys asymptotic bounds in expectation, see Stein (1987), with few points, hence it is a good candidate for our purposes.

Because there is no guarantee for uniqueness of solution, the algorithms for approximation of the solutions of the equation $H(\boldsymbol{a}) = 0$ should be complemented by further steps in which the obtained solutions should be tested whether they provide minimisers, of the learning function (3.45), or not and to which extent from the class of all minimisers one can get a global minimiser. But the most challenging problem refers to the assumption that the Jacobi matrix $\nabla_{\boldsymbol{u}} H$ is nonsingular for all $\boldsymbol{u} \in \mathbb{R}^N$. In order to tackle this question we explicitly calculate this Jacobi matrix. To this end, we first observe that for each $i,j = 1,\ldots,l+u$, letting $a_j = (a_j^{(1)},\ldots,a_j^{(d_j)})$, we have a partial Jacobi matrix

$$\frac{\partial H_i}{\partial a_j} = \Big(\frac{\partial H_i}{\partial a_j^{(1)}},\ldots,\frac{\partial H_i}{\partial a_j^{(d_j)}}\Big),$$

that is, a function matrix of dimension $d_i \times d_j = \dim(\mathcal{W}_{x_i}) \times \dim(\mathcal{W}_{x_j})$. Let $i = 1,\ldots,l$ and let I_{d_i} denote the identity operator on $\mathcal{W}_{x_i}$ identified with $\mathbb{R}^{d_i}$. Then, from (3.29), on the one hand we get

$$\begin{aligned}
\frac{\partial H_i}{\partial a_i} &= I_{d_i} + \frac{1}{l\gamma_A} \exp\Big(-\|y_i - C_{x_j} \sum_{k=1}^{l+u} K(x_j,x_k) a_k\|^2_{\mathcal{Y}_{x_j}}\Big) C^*_{x_i} C_{x_i} K(x_i,x_i) \\
&\quad + \frac{1}{l\gamma_A} \sum_{j=1}^{l+u} 2\langle y_j - C_{x_j} \sum_{k=1}^{l+u} K(x_j,x_k) a_k, C_{x_j} K(x_j,x_i) a_i\rangle_{\mathcal{Y}_{x_i}} \times \\
&\qquad \times \exp\Big(-\|y_j - C_{x_j} \sum_{k=1}^{l+u} K(x_j,x_k) a_k\|^2_{\mathcal{Y}_{x_j}}\Big) C^*_{x_i} C_{x_i} K(x_i,x_j) \\
&\quad + \frac{\gamma_I}{\gamma_A} \sum_{k=1}^{l+u} M_{i,k} K(x_k,x_i),
\end{aligned} \tag{3.31}$$

and, on the other hand, for each $j = 1, \ldots, l+u$, $j \neq i$, we get

$$\begin{aligned}\frac{\partial H_i}{\partial a_j} = {} & \frac{1}{l\gamma_A} \exp\bigl(-\|y_j - C_{x_j} \sum_{k=1}^{l+u} K(x_j, x_k) a_k\|^2_{\mathcal{Y}_{x_j}}\bigr) C^*_{x_i} C_{x_i} K(x_i, x_j) \\ & + \frac{1}{l\gamma_A} \sum_{m=1}^{l+u} 2\langle y_m - C_{x_m} \sum_{k=1}^{l+u} K(x_m, x_k) a_k, C_{x_m} K(x_m, x_j) a_j \rangle_{\mathcal{Y}_{x_m}} \times \\ & \qquad \times \exp\bigl(-\|y_m - C_{x_m} \sum_{k=1}^{l+u} K(x_m, x_k) a_k\|^2_{\mathcal{Y}_{x_m}}\bigr) C^*_{x_i} C_{x_i} K(x_i, x_m) \\ & + \frac{\gamma_I}{\gamma_A} \sum_{k=1}^{l+u} M_{i,k} K(x_k, x_j). \end{aligned} \tag{3.32}$$

Let now $i = l+1, \ldots, l+u$. Then, from (3.30), on the one hand we get

$$\frac{\partial H_i}{\partial a_i} = I_{d_i} + \frac{\gamma_I}{\gamma_A} \sum_{k=1}^{l+u} M_{i,k} K(x_k, x_i), \tag{3.33}$$

and, on the other hand, for $j = 1, \ldots, l+u$, $j \neq i$, we get

$$\frac{\partial H_i}{\partial a_j} = \frac{\gamma_I}{\gamma_A} \sum_{k=1}^{l+u} M_{i,k} K(x_k, x_j). \tag{3.34}$$

The partial Jacobi matrices obtained in (3.31) through (3.34) make a complete description of the Jacobi matrix of the Fréchet derivative $\nabla_{\boldsymbol{a}} H$. From these formulae, a few observations follow. A first observation is that we can write

$$\nabla_{\boldsymbol{a}} H = I_{\mathbb{R}^N} + \frac{1}{\gamma_A} R, \tag{3.35}$$

where R is an $N \times N$ matrix that can be calculated explicitly from (3.31) through (3.34). Then, one can use different extra assumptions on the kernels and data points in order to assure that $\nabla_{\boldsymbol{a}} H$ is a nonsingular matrix. For example, one can take into account the fact that the labeled points $x_1, \ldots, x_l$ are selected in a supervised manner while the unlabeled points $x_{l+1}, \ldots, x_{l+u}$ can be changed in a convenient manner that assures the Jacobi matrix $\nabla_{\boldsymbol{a}} H$ be nonsingular. Another observation is that in a semisupervised learning problem the number l of labeled points is significantly less than the number u of unlabeled points and hence the degree of nonlinearity of the system of equations defined by (3.26) and (3.27) is rather low, which can be used to search for reliable approximations by systems of linear equations.

Some of the available algorithms for constrained optimisation problems, such as that described before, depend heavily on the assumption that the Jacobi matrix $\nabla_{\boldsymbol{a}} H$ is nonsingular for all $\boldsymbol{a} \in \mathbb{R}^N$, a situation that may not be easy to get. In general, we can find a nonempty open set Ω in $\mathbb{R}^N$ on which the Jacobi matrix $\nabla_{\boldsymbol{a}} H$ is nonsingular, for example we can use (3.35) to show that Ω contains the set of those $\boldsymbol{a} \in \mathbb{R}^N$ with the property that $\|R\| < \gamma_A$. In view of (3.31) through (3.34), in the nonlinear terms of these partial Jacobi

matrices the exponentials with negative exponents tame the growth of $\|R\|$ when the vector $\boldsymbol{a}$ is far from the solution and, consequently, by manipulating the regularisation coefficient γ_A we can get very large sets Ω. Then, one can use other techniques to prevent the approximation sequence to get too close to the boundary of Ω as in the constrained version of the algorithm proposed in Monteiro and Pang (1999). More recent investigations refer to either adapting the algorithm in case the Jacobi matrix $\nabla_{\boldsymbol{u}} H$ is singular and pseudo-inverses replace the inverses, see e.g. the analyis of Kaczmarz type algorithms for ill-posed linear problems in Popa (2018), or using stochastic Bregman-Kaczmarz methods, see Gower et al. (2024). These allow us to modify correspondingly the algorithm that we presented here to the general case. We leave the details for a further research project on real data sets. As a practical approach, the simplest is to use perturbation theory, more precisely, for those values of $\boldsymbol{a}$ at which the Fréchet derivative is not invertible, a small perturbation of $\boldsymbol{a}$ changes the point to one for which the Fréchet derivative is invertible and then the stability of the problem to small perturbations guarantees the convergence of the iteration process.

3.6 A Toy Model.

In this subsection we provide a toy model for the localised version of the regularised machine learning problem in case the loss function is the exponential least square function as in Subsection 3.4 and test an algorithm following the discussion of the numerical methods as in Subsection 3.5. To this end, let $X = X_1 \cup X_2$, where

$$X_1 := \{(\alpha_1, \alpha_2) \mid 0.25 \leq \alpha_1 \leq 1,\ 0.25 \leq \alpha_2 \leq 1\}, \tag{3.36}$$

$$X_2 := \{(\alpha_1, \alpha_2) \mid -1 \leq \alpha_1 \leq -0.25,\ 0.25 \leq \alpha_2 \leq 1\}. \tag{3.37}$$

In the following we use the notation as in Subsection 2.2. We consider $x_1 \in X_1$ and $x_2 \in X_2$ randomly selected and let the labels $y_1 \in \mathcal{Y}_{x_1} = \mathbb{R}$ and $y_2 \in \mathcal{Y}_{x_2} = \mathbb{R}^2$ be randomly selected. Also, let $x_3, x_4 \in X_1 \setminus \{x_1\}$, $x_3 \neq x_4$, and $x_5, x_6 \in X_2 \setminus \{x_2\}$, $x_5 \neq x_6$, randomly selected as well, be unlabeled points. In particular, $l = 2$ and $u = 4$. Let $\mathcal{Y}_{x_3} = \mathcal{Y}_{x_4} = \mathbb{R}$ and $\mathcal{Y}_{x_5} = \mathcal{Y}_{x_6} = \mathbb{R}^2$. Then we take $\mathcal{W}_{x_j} = \mathcal{Y}_{x_j}$ for all $j = 1, \ldots, 6$, in particular the machine learning problem is single viewed. With respect to the notation (2.12) we have

$$\begin{aligned} \boldsymbol{W}^{l+u} &= \mathcal{Y}_{x_1} \oplus \mathcal{Y}_{x_2} \oplus \mathcal{Y}_{x_3} \oplus \mathcal{Y}_{x_4} \oplus \mathcal{Y}_{x_5} \oplus \mathcal{Y}_{x_6} \\ &= \mathbb{R} \oplus \mathbb{R}^2 \oplus \mathbb{R} \oplus \mathbb{R} \oplus \mathbb{R}^2 \oplus \mathbb{R}^2 = \mathbb{R}^9, \end{aligned}$$

hence $\dim(\boldsymbol{W}^{l+u}) = 9$.

We let the regularisation coefficients γ_A and γ_I unspecified and we will test different choices later. Since the problem is single-view we let $C_{x_i} = I_{\mathcal{Y}_{x_i}}$ for all $i = 1, \ldots, 6$ and $M_B = 0$, see Example 3. For the within-view operator, see Example 4, we proceed as follows. Let

$$w_{j,k} = \exp\big(-\frac{\|x_j - x_k\|^2}{2\sigma^2}\big), \quad j, k = 1, \ldots, 6, \tag{3.38}$$

and then, consider the 6×6 matrices $W = [w_{j,k}]_{j,k=1}^6$ and $V = \operatorname{diag}(v_{1,1}, \ldots, v_{6,6})$, where

$$v_{j,j} = \sum_{k=1}^{6} w_{j,k}, \quad j = 1, \ldots, 6, \tag{3.39}$$

and the Laplace matrix $L = V - W = [l_{j,k}]_{j,k=1}^6$. Taking into account that the labels of points in X_1 have dimension 1 and the labels of points in X_2 have dimension 2, the matrix $M_W = M$ looks like this, see Example 4.

$$M = \begin{bmatrix} l_{1,1} & l_{1,2} & 0 & l_{1,3} & l_{1,4} & l_{1,5} & 0 & l_{1,6} & 0 \\ l_{2,1} & l_{2,2} & 0 & l_{2,3} & l_{2,4} & l_{2,5} & 0 & l_{2,6} & 0 \\ 0 & 0 & l_{2,2} & 0 & 0 & 0 & 0 & 0 & 0 \\ l_{3,1} & l_{3,2} & 0 & l_{3,3} & l_{3,4} & l_{3,5} & 0 & l_{3,6} & 0 \\ l_{4,1} & l_{4,2} & 0 & l_{4,3} & l_{4,4} & l_{4,5} & 0 & l_{4,6} & 0 \\ l_{5,1} & l_{5,2} & 0 & l_{5,3} & l_{5,4} & l_{5,5} & 0 & l_{5,6} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & l_{5,5} & 0 & 0 \\ l_{6,1} & l_{6,2} & 0 & l_{6,3} & l_{6,4} & l_{6,5} & 0 & l_{6,6} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & l_{6,6} \end{bmatrix}. \tag{3.40}$$

With notation as in Section 2.2 we have $M = [M_{j,k}]_{j,k=1}^6$ with

$$M_{1,1} = l_{1,1},\ M_{1,2} = \begin{bmatrix} l_{1,2} & 0 \end{bmatrix},\ M_{1,3} = l_{1,3},\ M_{1,4} = l_{1,4},\ M_{1,5} = \begin{bmatrix} l_{1,5} & 0 \end{bmatrix},\ M_{1,6} = \begin{bmatrix} l_{1,6} & 0 \end{bmatrix},$$

$$M_{2,1} = \begin{bmatrix} l_{2,1} \\ 0 \end{bmatrix},\ M_{2,2} = \begin{bmatrix} l_{2,2} & 0 \\ 0 & l_{2,2} \end{bmatrix},\ M_{2,3} = \begin{bmatrix} l_{2,3} \\ 0 \end{bmatrix},\ M_{2,4} = \begin{bmatrix} l_{2,4} \\ 0 \end{bmatrix},\ M_{2,5} = \begin{bmatrix} l_{2,5} & 0 \\ 0 & 0 \end{bmatrix},\ M_{2,6} = \begin{bmatrix} l_{2,6} & 0 \\ 0 & 0 \end{bmatrix},$$

$$M_{3,1} = l_{3,1},\ M_{3,2} = \begin{bmatrix} l_{3,2} & 0 \end{bmatrix},\ M_{3,3} = l_{3,3},\ M_{3,4} = l_{3,4},\ M_{3,5} = \begin{bmatrix} l_{3,5} & 0 \end{bmatrix},\ M_{3,6} = \begin{bmatrix} l_{3,6} & 0 \end{bmatrix},$$

$$M_{4,1} = l_{4,1},\ M_{4,2} = \begin{bmatrix} l_{4,2} & 0 \end{bmatrix},\ M_{4,3} = l_{4,3},\ M_{4,4} = l_{4,4},\ M_{4,5} = \begin{bmatrix} l_{4,5} & 0 \end{bmatrix},\ M_{4,6} = \begin{bmatrix} l_{4,6} & 0 \end{bmatrix},$$

$$M_{5,1} = \begin{bmatrix} l_{5,1} \\ 0 \end{bmatrix},\ M_{5,2} = \begin{bmatrix} l_{5,2} & 0 \\ 0 & 0 \end{bmatrix},\ M_{5,3} = \begin{bmatrix} l_{5,3} \\ 0 \end{bmatrix},\ M_{5,4} = \begin{bmatrix} l_{5,4} \\ 0 \end{bmatrix},\ M_{5,5} = \begin{bmatrix} l_{5,5} & 0 \\ 0 & l_{5,5} \end{bmatrix},\ M_{5,6} = \begin{bmatrix} l_{5,6} & 0 \\ 0 & 0 \end{bmatrix},$$

$$M_{6,1} = \begin{bmatrix} l_{6,1} \\ 0 \end{bmatrix},\ M_{6,2} = \begin{bmatrix} l_{6,2} & 0 \\ 0 & 0 \end{bmatrix},\ M_{6,3} = \begin{bmatrix} l_{6,3} \\ 0 \end{bmatrix},\ M_{6,4} = \begin{bmatrix} l_{6,4} \\ 0 \end{bmatrix},\ M_{6,5} = \begin{bmatrix} l_{6,5} & 0 \\ 0 & 0 \end{bmatrix},\ M_{6,6} = \begin{bmatrix} l_{6,6} & 0 \\ 0 & l_{6,6} \end{bmatrix}.$$

We consider the kernel $K\colon X \times X \to \bigcup_{z,\zeta\in X} \mathcal{B}(\mathcal{W}_\zeta, \mathcal{W}_z)$ defined as follows.

$$K(z,\zeta) := \exp\Big(-\frac{\|z-\zeta\|^2}{\sigma^2}\Big), \quad z,\zeta \in X_1, \tag{3.41}$$

$$K(z,\zeta) := \begin{bmatrix} \exp\big(-\frac{\|z-\zeta\|^2}{\sigma^2}\big) & 0 \\ 0 & \exp(-\alpha\|z-\zeta\|) \end{bmatrix}, \quad z,\zeta \in X_2, \tag{3.42}$$

$$K(z,\zeta) := \begin{bmatrix} \exp\big(-\frac{\|z-\zeta\|^2}{\sigma^2}\big) & 0 \end{bmatrix}, \quad z \in X_1,\ \zeta \in X_2, \tag{3.43}$$

$$K(z,\zeta) := \begin{bmatrix} \exp\big(-\frac{\|z-\zeta\|^2}{\sigma^2}\big) \\ 0 \end{bmatrix}, \quad z \in X_2,\ \zeta \in X_1. \tag{3.44}$$

Here the coefficients σ and α remain unspecified for the moment.

With these data we have the system of equations (3.26) and (3.27), where a_1, a_3, and a_4 are scalars and a_2, a_5, and a_6 are 2-vectors. So, speaking in terms of scalars, we have a system of nine equations with nine unknowns. Three of these equations are nonlinear and the rest of six equations are linear.

In order to find a bounded set on which we can guarantee the existence of the global solution of the minimisation problem (2.14), we follow the idea of the proof of Theorem 7. With notation as in that theorem, we search for the minimiser

$$f = \sum_{j=1}^{6} K_{x_j} a_j \tag{3.45}$$

and we let $f_0 = 0$ and hence the corresponding vector $\boldsymbol{a} = 0$. Then, by (3.21) we have

$$\mathcal{I}(f_0) = 1 - \frac{1}{2}\big(\exp(-y_1^2) + \exp(-\|y_2\|^2)\big). \tag{3.46}$$

We search for $\delta > 0$ such that $\mathcal{I}(f) \geq \mathcal{I}(f_0)$, with f as in (3.45), for any vector $\boldsymbol{a} = (a_1, a_2, a_3, a_4, a_5, a_6)$ of dimension 9 with the property that $\boldsymbol{a} \notin [-\delta, \delta]^9$. By the proof of Theorem 7, the minimisation problem $\operatorname{argmin} \|H(\boldsymbol{a})\|$ has the solution in the cube $[-\delta, \delta]^9$. Taking into account that

$$\mathcal{I}(f) \geq \gamma_A \|f\|_{\mathcal{H}_K} \geq \gamma_A \langle K_{\boldsymbol{x}} \boldsymbol{a}, \boldsymbol{a}\rangle \geq \gamma_A \lambda_{K_{\boldsymbol{x}}} \|\boldsymbol{a}\|_2^2 \geq \gamma_A \lambda_{K_{\boldsymbol{x}}} \|\boldsymbol{a}\|_\infty^2, \tag{3.47}$$

where,

$$K_{\boldsymbol{x}} = [K(x_i, x_j)]_{i,j=1}^6, \tag{3.48}$$

is a 9×9 matrix and $\lambda_{K_{\boldsymbol{x}}} > 0$ is the least eigenvalue of the positive matrix $K_{\boldsymbol{x}}$. From (3.46) and (3.47), letting

$$\delta = \frac{1}{\sqrt{\gamma_A \lambda_{K_{\boldsymbol{x}}}}} \sqrt{1 - \frac{1}{2}\big(\exp(-y_1^2) + \exp(-\|y_2\|^2)\big)}, \tag{3.49}$$

it follows that in order to find the global minimiser for the problem (2.14), it is sufficient to search for the solutions $\boldsymbol{a}$ of the minimisation problem $\operatorname{argmin} \|H(\boldsymbol{a})\|$ in the cube $[-\delta, \delta]^9$.

There are two conditions to be verified, in order for δ to be consistent. Firstly, we work under the assumption that $\lambda_{K_{\boldsymbol{x}}} > 0$ which can be numerically checked, hence $\delta < \infty$. Secondly, note that if $\delta = 0$ this means that $y_1 = 0$ and $y_2 = 0$, hence $f = 0$ is the solution for the global minimiser and hence, in this case, the problem is trivial. So, we work under the hypothesis that $\delta > 0$.

Since the optimisation problem is sensitive to the choice of initial conditions, we explored various possibilities. Specifically, we found that generating a Latin Hypercube Sampling (LHS), see McKay et al. (1979) and the previous section, within the cube $[-\delta, \delta]^9$ provides a reliable estimate of the global solution while maintaining a reasonable runtime. Algorithm 1 below summarises the numerical implementation.

Remark 20 In this algorithm, our admissibility criterion at line 7 is 3-fold. We check whether $\mathbf{a}_0$ results in both a smaller loss $\mathcal{I}(f_{\mathbf{a}_0})$, and a smaller gradient norm $\|\mathbf{H}(\mathbf{a}_0)\|$ compared to $\mathbf{a}$. Additionally, we verify that the first-order optimality condition is sufficiently small to ensure that $\mathbf{a}_0$ corresponds to a local extremum. Among all local optima we select the global one by a careful investigation and using the Latin Hypercube Sampling implemented in the algorithm and the code run on MATLAB. An essential part of the algorithm is the use of the function `fmincon` of MATLAB that uses the method of interior points for constrained optimisation, see Byrd et al. (1999), Byrd et al. (2000), and Waltz et al. (2006).

Algorithm 1 Numerical Implementation

Require: $\mathbf{x}_i \in X_1 \cup X_2 \subset \mathbb{R}^2$, $\sigma, \alpha > 0$, $\gamma_I, \gamma_A > 0$.

$D \leftarrow \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_6\}$
Construct M in (3.40) and K in (3.48) using data D
Compute δ in (3.49)
Compute the LHS L on the cube $[-\delta, \delta]^9$
for each element $\mathbf{a}_0 \in L$ **do**
    Solve the minimisation problem (3.29) and (3.30) on the cube $[-\delta, \delta]^9$ with initial condition $\mathbf{a}_0$ using the function `fmincon`.
    if $\mathbf{a}_0$ is admissible in the sense of Remark 20 **then**
        $\mathbf{a} \leftarrow \mathbf{a}_0$
    end if
end for
return **a**

Example 5 In this example we used $\gamma_A = 0.25$, $\gamma_I = 10$, $\sigma = 0.1$ and $\alpha = 10$, randomly generated the data x and y and returned the result $\boldsymbol{a}$.

$$x = \begin{bmatrix} 0.5377 & 0.6342 & 0.3273 & 0.3472 & 0.6724 & 0.8174 \\ 0.3978 & -0.4584 & 0.3923 & 0.4305 & -0.7962 & -0.3601 \end{bmatrix} \quad y = (1.2108, 1.6636, 4.3843)$$

$$\boldsymbol{a} = (0.8433, 1.7226, 1.5475, 0.4395, 0.3944, 0.1926, -0.0055, 1.4116, -0.1589) \tag{3.50}$$

The implicit optimality tolerance is $\epsilon = 10^{-6}$. Each time the code produced the mesh of the solution f_{best} given by (3.45) for the corresponding coefficients given by the coordinates of the solution $\boldsymbol{a}$. In figures 1 and 2 there are the meshes of the optimal solution as in (3.50), the first one corresponds to the set X_1 where the function f_{best} is scalar valued while the second one corresponds to the set X_2 where the function f_{best} has 2 dimensional vector values and there are two meshes, one for each component. The circles are the labeled points.

In figures 1 and 2 we observe that the choice of the parameters avoids overfitting. Actually, by varying the coefficients γ_A, γ_I, and the others, one can obtain different degrees of overfitting. The code runs efficiently on a Mac laptop, for this example it takes only less than a minute, but for other combinations of coefficients it may be five or more minutes.

We also plot in Figure 3 the values of the learning function $\mathcal{I}(f_{best})$ for each choice of the initial point. In Figure 4 we plot the values of $H(\boldsymbol{a})$, in order to empirically show that the optimal solution $\boldsymbol{a}$ is a good approximation of the solution of the equation $H(\boldsymbol{a}) = 0$, for each choice of the initial point,

Finally, in Figure 5 we empirically check that the solution f_{best}, corresponding to the optimal $\boldsymbol{a}$ as in (3.45), is a good approximation of the minimiser of the learning function $\mathcal{I}$, given at (3.21), for example by meshing it when we keep the last seven entries fixed as the last seven entries of the optimal $\boldsymbol{a}$ and let each of the first two entries of $\boldsymbol{a}$ vary in the interval $[-\delta, \delta]$. The point represents the value of the learning function at f_{best}.

Although the picture in Figure 5 shows a convex surface this is misleading since it represents only a 2-dimensional section of the general 9-dimensional surface generated by the learning function $\mathcal{I}(f)$, when f is parametrised in terms of the 9-dimensional vector $\boldsymbol{a}$ as

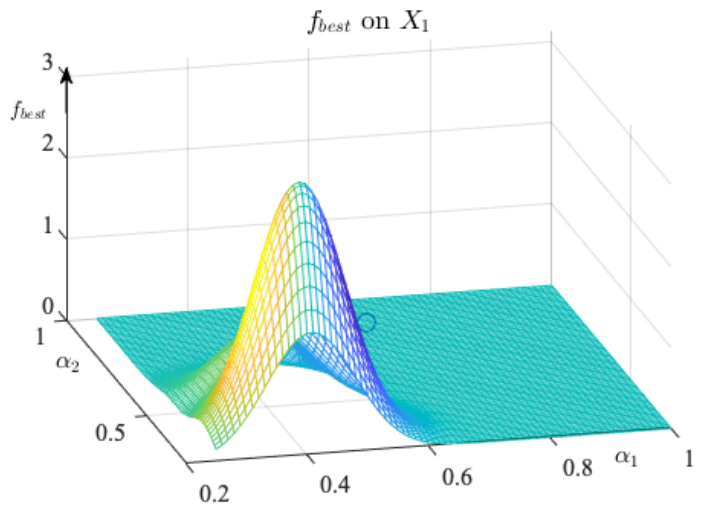

Figure 1: The plot of f_{best} over X_1 has only one component. The variables α_1 and α_2 correspond to the notation as in (3.36).

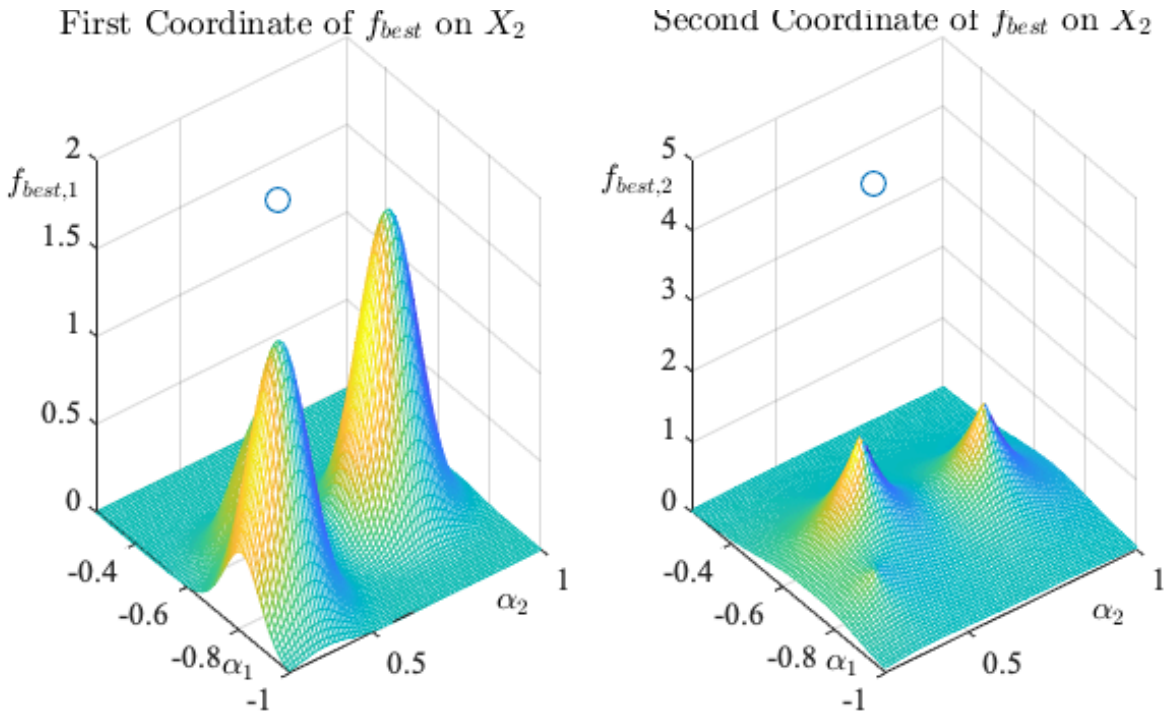

Figure 2: The plot of f_{best} over X_2 has two components. The variables α_1 and α_2 correspond to the notation as in (3.37).

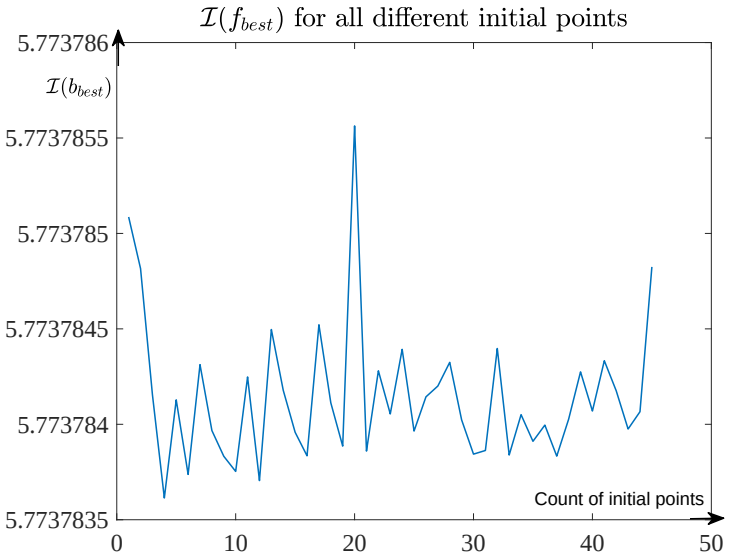

Figure 3: The values of the learning function $\mathcal{I}(f_{best})$ for each iteration corresponding to the 45 initial points are very close one to each other.

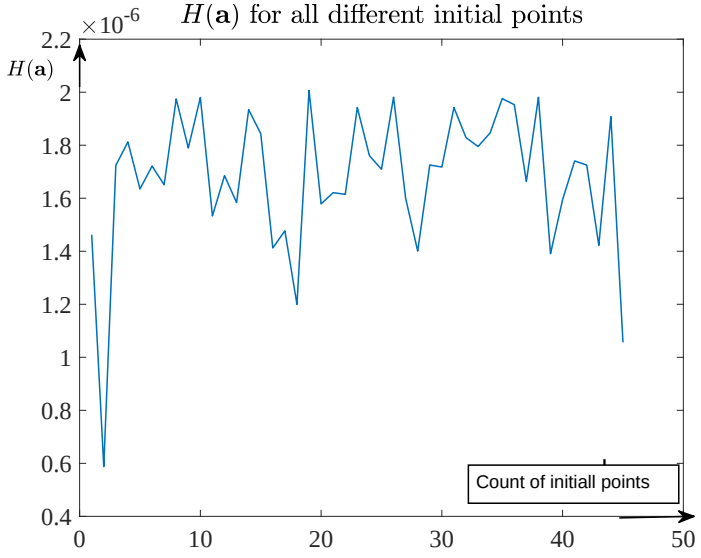

Figure 4: The values of $H(\boldsymbol{a})$ for each iteration corresponding to the 45 initial points are very close of 0.

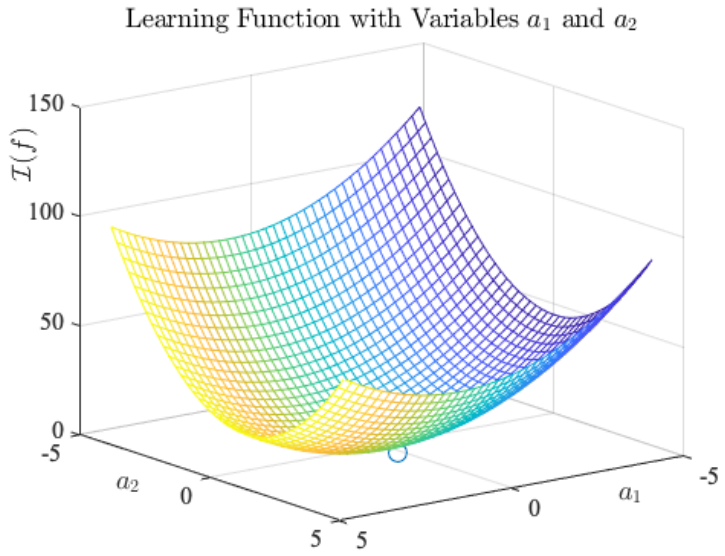

Figure 5: The mesh of the learning function $\mathcal{I}(f_{best})$ when the parameters a_1 and a_2 are varied and the other seven are fixed.

in (3.45). This behaviour of the learning map, as well as other traits, may vary for different choices of the coefficients γ_A, γ_I, σ, and α, as we performed test for different combinations.

In this example, one can observe that in the region that is trusted to contain the global minimum we actually have uniqueness of the solutions, both for the equation $H(\boldsymbol{a}) = 0$ and the minimiser of the learning map $\mathcal{I}$. But, due to the nonlinearity of the system $H(\boldsymbol{a}) = 0$ and the nonconvexity of the learning map $\mathcal{I}$, this does not mean that there may be no other solutions when the search is performed in a larger region, that is, local minima. This shows the importance of Theorem 7 that provides a bounded region where the solutions that are of interest live.

Acknowledgments

The authors thank Naci Saldi and Sorin Costiner for very helpful discussions on available algorithms to solve nonlinear systems of equations and optimisation problems. The authors thank the anonymous referees for the careful reading of the three versions of this manuscript and the recommendations and observations that helped us improve considerably the form and the contents of this manuscript. Last but no least we thank the editor for taking care so professionally of all revisions and the recommendations that eventually lead to the present final form.

This research is not supported by any financial activity. Authors declare that there exist no competing interests.

Appendix A. Proof of Theorem 1

In the following we use a formalisation of the quotient completion to a Hilbert space of an $\mathbb{F}$-vector space $\mathcal{V}$ with respect to a given nonnegative sesquilinear form $\mathcal{V} \times \mathcal{V} \ni (u, v) \mapsto q(u, v) \in \mathbb{F}$, as follows. A pair $(\mathcal{H}; \Pi)$ is called a *Hilbert space induced* by $(\mathcal{V}; q)$ if:

(ihs1) $\mathcal{H}$ is a Hilbert space.

(ihs2) $\Pi\colon \mathcal{V} \to \mathcal{H}$ is a linear operator with dense range.

(ihs3) $q(u, v) = \langle \Pi u, \Pi v\rangle_{\mathcal{H}}$, for all $u, v \in \mathcal{V}$.

Such an induced Hilbert space always exists and is unique, up to a unitary operator. More precisely, we will use the following construction. Consider the vector subspace of $\mathcal{V}$ defined by

$$\mathcal{N}_q := \{u \in \mathcal{V} \mid q(u, u) = 0\} = \{u, \in \mathcal{V} \mid q(u, v) = 0 \text{ for all } v \in \mathcal{V}\}, \tag{1.1}$$

where the equality holds due to the Schwarz Inequality for q, and then consider the quotient vector space $\mathcal{V}/\mathcal{N}_q$. Letting

$$\widetilde{q}(u + \mathcal{N}_q, v + \mathcal{N}_q) := q(u, v), \quad u, v \in \mathcal{V}, \tag{1.2}$$

we have a pre-Hilbert space $(\mathcal{V}/\mathcal{N}_q; \widetilde{q})$ that can be completed to a Hilbert space $(\mathcal{H}_q; \langle\cdot, \cdot\rangle_{\mathcal{H}})$. Letting $\Pi_q\colon \mathcal{V} \to \mathcal{H}_q$ be defined by

$$\Pi_q u := u + \mathcal{N}_q \in \mathcal{V}/\mathcal{N}_q \subseteq \mathcal{H}_q, \quad u \in \mathcal{V}, \tag{1.3}$$

it is easy to see that $(\mathcal{H}_q, \Pi_q)$ is a Hilbert space induced by $(\mathcal{V}; q)$.

(a)⇒(b). Assuming that the $\boldsymbol{H}$-operator valued kernel K is positive semidefinite, we consider the vector space $\mathcal{F}_0(X; \boldsymbol{H})$ of vector cross-sections with finite support and the Hermitian sesquilinear form $\langle\cdot, \cdot\rangle_K$ defined as in (2.6). We consider

$$\begin{aligned} \mathcal{N}_K &= \{f \in \mathcal{F}_0(X; \boldsymbol{H}) \mid \langle f, f\rangle_K = 0\} \\ &= \{f \in \mathcal{F}_0(X; \boldsymbol{H}) \mid \langle f, g\rangle_K = 0 \text{ for all } g \in F_0(X; \boldsymbol{H})\}, \end{aligned} \tag{1.4}$$

then consider the induced Hilbert space $(\mathcal{H}_K; \Pi_K)$ associated to $(\mathcal{F}_0(X; \boldsymbol{H}); \langle\cdot, \cdot\rangle_K)$, and let $\mathcal{K} := \mathcal{H}_K$. For each $x \in X$ let $V(x)\colon \mathcal{H}_x \to \mathcal{K}$ be the operator defined by

$$V(x)h = \Pi_K(\widehat{h}) = h + \mathcal{N} \in \mathcal{K}, \tag{1.5}$$

with notation as in (2.3). Since

$$\langle V(x)h, V(x)h\rangle_K = \langle K(x, x)h, h\rangle_{\mathcal{H}_x} \leq \|K(x, x)\|\|h\|^2_{\mathcal{H}_x}, \quad h \in \mathcal{H}_x,\ x \in X, \tag{1.6}$$

it follows that $V(x)$ is bounded for all $x \in X$. Note that, in this way, $\mathcal{K}$ is the closed span of $\{V(x)\mathcal{H}_x \mid x \in X\}$. On the other hand,

$$\langle K(y, x)h, g\rangle_{\mathcal{H}_y} = \langle h+\mathcal{N}, g+\mathcal{N}\rangle_K = \langle V(x)h, V(y)g\rangle_K, \quad h \in \mathcal{H}_x,\ g \in \mathcal{H}_y,\ x, y \in X, \tag{1.7}$$

hence, $K(y, x) = V(y)^*V(x)$ for all $x, y \in X$. We thus proved that $(\mathcal{K}; V)$ is a minimal Hilbert space linearisation of the $\boldsymbol{H}$-kernel K.

In the following we prove that $(\mathcal{K}; V)$ is unique, modulo unitary equivalence. To see this, let $(\mathcal{K}'; V')$ be another minimal linearisation of K. Then, for arbitrary $f \in \mathcal{F}_0(X; \boldsymbol{H})$ we have

$$\begin{aligned}\langle \sum_{x\in X} V(x)f_x, \sum_{y\in X} V(y)f_y\rangle_0 &= \sum_{x,y\in X} \langle V(y)^*V(x)f_x, f_y\rangle_0 \\ &= \sum_{x,y\in X} \langle K(y,x)f_x, f_y\rangle_0 \\ &= \langle f, f\rangle_K \\ &= \langle \sum_{x\in X} V'(x)f_x, \sum_{y\in X} V'(y)f_y\rangle_0,\end{aligned}$$

hence, defining $U(\sum_{x\in X} V'(x)f_x) = \sum_{x\in X} V(x)f_x$, for arbitrary $f \in \mathcal{F}_0(X; \boldsymbol{H})$, it follows that U is isometric and, taking into account of the minimality conditions, it follows that U can be uniquely extended to a unitary operator $U\colon \mathcal{K}' \to \mathcal{K}$, such that $UV'(x) = V(x)$ for all $x \in X$.

(b)⇒(a). Assuming that $(\mathcal{K}; V)$ is a Hilbert space linearisation of K, we have

$$\begin{aligned}\sum_{x,y\in X} \langle K(y,x)f_x, f_y\rangle_{\mathcal{H}_y} &= \sum_{x,y\in X} \langle V(y)^*V(x)f_x, f_y\rangle_{\mathcal{H}_y} \\ &= \| \sum_{x\in X} V(x)f_x\|^2_{\mathcal{K}}, \quad f \in \mathcal{F}_0(X; \boldsymbol{H}),\end{aligned}$$

hence K is positive semidefinite.

Appendix B. Proof of Theorem 2

(a)⇒(b). If K is positive definite then, by Theorem 1, there exists a minimal linearisation $(\mathcal{K}; V)$ of K. Define $\mathcal{R} = \{V(\cdot)^*f \mid f \in \mathcal{K}\}$, that is, $\mathcal{R}$ consists of all functions $X \ni x \mapsto V(x)^*f \in \mathcal{H}_x$, with $f \in \mathcal{K}$, in particular, $V(\cdot)^*f$ can be viewed as an $\boldsymbol{H}$-vector bundle, that is, $V(\cdot)^*f \in \mathcal{F}(X; \boldsymbol{H})$ for all $f \in \mathcal{K}$. Thus, we can view $\mathcal{R}$ as a linear subspace of $\mathcal{F}(X; \boldsymbol{H})$, with all its algebraic operations.

We now show that the mapping

$$\mathcal{K} \ni f \mapsto Uf = V(\cdot)^*f \in \mathcal{R} \tag{2.1}$$

is bijective. By definition, this mapping is surjective, hence it remains to prove that it is injective. To see this, let $f, g \in \mathcal{K}$ be such that $V(\cdot)^*f = V(\cdot)^*g$. Then for arbitrary $x \in X$ and $h \in \mathcal{H}_x$ we have $\langle V(x)^*f, h\rangle_{\mathcal{H}_x} = \langle V(x)^*g, h\rangle_{\mathcal{H}_x}$, equivalently, $\langle f - g, V(x(h\rangle_{\mathcal{K}} = 0$. Taking into account the minimality of the linearisation, it follows that $f = g$. Thus, U is bijective.

It is obvious that the bijective mapping U as in (2.1) is linear. On $\mathcal{R}$ we introduce an inner product $\langle \cdot, \cdot\rangle_{\mathcal{R}}$ defined by

$$\langle Uf, Uf\rangle_{\mathcal{R}} = \langle V(\cdot)^*f, V(\cdot)^*g\rangle_{\mathcal{K}}, \quad f, g \in \mathcal{K}, \tag{2.2}$$

in other words, U is now an isometric isomorphism between the Hilbert space $\mathcal{K}$ and the inner product space $\mathcal{R}$, hence $(\mathcal{R}; \langle\cdot,\cdot\rangle_{\mathcal{R}})$ is a Hilbert space as well.

We now show that $(\mathcal{R}; \langle\cdot,\cdot\rangle_{\mathcal{R}})$ is a reproducing kernel Hilbert space with reproducing kernel K. Indeed, since for all $x, y \in X$ and all $h \in \mathcal{H}_x$ we have $K_x(y)h = K(y,x)h = V(y)^*V(x)h$, it follows that $K_x \in \mathcal{R}$ for all $x \in X$. On the other hand, for arbitrary $f \in \mathcal{R}$, $x \in X$, and $h \in \mathcal{H}_x$, we have

$$\begin{aligned}\langle f, K_x h\rangle_{\mathcal{R}} &= \langle V(\cdot)^* g, K_x h\rangle_{\mathcal{R}} = \langle V(\cdot)g, V(\cdot)^*V(x)h\rangle_{\mathcal{R}}\\ &= \langle g, V(x)h\rangle_{\mathcal{K}} = \langle V(x)^* g, h\rangle_{\mathcal{H}_x},\end{aligned}$$

where $g \in \mathcal{H}$ is the unique vector such that $V(x)^*g = f$. Thus, we proved that K is the reproducing kernel of $\mathcal{R}$.

(b)⇒(a). Let $(\mathcal{R}; \langle\cdot,\cdot\rangle_{\mathcal{R}})$ be a reproducing kernel Hilbert space with reproducing kernel K. Using the reproducing property (rk3), for arbitrary $n \in \mathbb{N}$, $x_1, \ldots, x_n \in X$, and $h_1 \in \mathcal{H}_{x_1}, \ldots, h_n \in \mathcal{H}_{x_n}$, we have

$$\sum_{i,j=1}^{n} \langle K(x_j, x_i)h_i, h_j\rangle_{\mathcal{H}_{x_j}} = \sum_{i,j=1}^{n} \langle K_{x_i}h_i, K_{x_j}h_j\rangle_{\mathcal{R}} = \|\sum_{i=1}^{n} K_{x_i}h_i\|_{\mathcal{R}}^2 \geq 0,$$

hence K is positive semidefinite.

Due to the uniqueness property of the reproducing kernel Hilbert space associated to a positive semidefinite $\boldsymbol{H}$-operator valued kernel K, it is natural to denote this reproducing kernel Hilbert space by $\mathcal{R}(K)$.

Appendix C. A Direct Construction of $\mathcal{R}(K)$.

Given an arbitrary bundle of Hilbert spaces $\boldsymbol{H} = \{\mathcal{H}_x\}_{x\in X}$ and an $\boldsymbol{H}$-operator valued kernel K, we described the reproducing kernel Hilbert space $\mathcal{R}(K)$ through a minimal linearisation of K, as in the proof of the implication (a)⇒(b) of Theorem 2, while a minimal Kolmogorov decomposition of K was obtained as in the proof of the implication (a)⇒(b) of Theorem 1. One of the unpleasant trait of the mentioned construction of the Kolmogorov decomposition, a GNS type construction in fact, is that, at a certain step, it makes a factorisation and hence, the obtained Hilbert space consists of equivalence classes of vector cross-sections. On the other hand, the reproducing kernel Hilbert space $\mathcal{H}(K)$ consists solely of vector cross-sections and, as noted before, it is a Kolmogorov decomposition as well, hence it would be desirable to have a direct construction of it, independent of the Kolmogorov decomposition. Such a direct, but longer, construction, that yields simultaneously the reproducing kernel Hilbert space $\mathcal{R}(K)$ and a minimal Kolmogorov decomposition of K, is more illuminating from certain points of view, and we describe it in the following.

Let $\mathcal{R}_0$ be the range of the convolution operator K defined at (2.7), more precisely, with the definition of the convolution operator C_K as in (2.7),

$$\begin{aligned}\mathcal{R}_0 &= \{f \in \mathcal{F}(X; \boldsymbol{H}) \,|\, f = C_K g \text{ for some } g \in \mathcal{F}_0(X; \boldsymbol{H})\} \qquad (3.1)\\ &= \{f \in \mathcal{F}(X; \boldsymbol{H}) \,|\, f_y = \sum_{x\in X} K(y,x)g_x \text{ for some } g \in \mathcal{F}_0(X; \boldsymbol{H}),\ \text{all } y \in X\}.\end{aligned}$$

A pairing $\langle \cdot, \cdot \rangle_{\mathcal{R}_0}$ can be defined on $\mathcal{R}_0$ by

$$\langle e, f \rangle_{\mathcal{R}_0} = \langle g, h \rangle_K = \langle C_K g, h \rangle_0 = \sum_{y \in X} \langle e(y), h(y) \rangle_{\mathcal{H}_y} = \sum_{x,y \in X} \langle K(y,x) g(x), h(y) \rangle_{\mathcal{H}_y}, \quad (3.2)$$

where $f = C_K h$ and $e = C_K g$ for some $g, h \in \mathcal{F}_0(X; \boldsymbol{H})$. We observe that, with the previous notation,

$$\begin{aligned} \langle e, f \rangle_{\mathcal{R}_0} &= \sum_{y \in X} \langle e(y), h(y) \rangle_{\mathcal{H}_y} = \sum_{x,y \in X} \langle K(y,x) g(x), h(y) \rangle_{H_y} \\ &= \sum_{x,y \in X} \langle g(x), K(x,y) h(y) \rangle_{\mathcal{H}_x} = \sum_{x \in X} \langle g(x), f(x) \rangle_{\mathcal{H}_x}, \end{aligned} \quad (3.3)$$

which shows that the definition in (3.2) is correct, that is, it does not depend on g and h such that $e = C_K g$ and $f = C_K h$. In the following we prove that the pairing $\langle \cdot, \cdot \rangle_{\mathcal{R}_0}$ is an inner product. It is easy to verify the linearity in the first argument, conjugate symmetry, and nonnegativity. Hence, the Schwarz inequality holds as well. In order to verify its positive definiteness, let $f \in \mathcal{R}_0$ be such that $\langle f, f \rangle_{\mathcal{R}_0} = 0$. By the Schwarz inequality, it follows that $\langle f, f' \rangle_{\mathcal{R}_0} = 0$ for all $f' \in \mathcal{R}_0$. For arbitrary $x \in X$ and $h \in \mathcal{H}_x$ consider the cross-section $\widehat{h} \in \mathcal{F}_0(X; \boldsymbol{H})$ defined as in (2.3). Letting $f' = C_K \widehat{h} \in \mathcal{R}_0$, we thus have

$$0 = \langle f, f' \rangle_{\mathcal{R}_0} = \langle f, C_K \widehat{h} \rangle_0 = \sum_{x \in X} \langle f_y, (\widehat{h})_y \rangle_{\mathcal{H}_y} = \langle f_x, h \rangle_{\mathcal{H}_x},$$

hence, since $x \in X$ and $h \in \mathcal{H}_x$ are arbitrary, it follows that $f = 0$. Thus, $(\mathcal{R}_0; \langle \cdot, \cdot \rangle_{\mathcal{R}_0})$ is an inner product space contained in $\mathcal{F}(X; \boldsymbol{H})$.

For any $x \in X$ and $h \in \mathcal{H}_x$, we consider the vector cross-section $\widehat{h} \in \mathcal{F}_0(X; \boldsymbol{H})$ defined at (2.3) and note that

$$(C_K \widehat{h})(y) = \sum_{z \in X} K(y,z) (\widehat{h})(z) = K(y,x) h = K_x(y) h, \quad y \in X, \quad (3.4)$$

that is, $C_K \widehat{h} = K_x h$, which shows that $K_x h \in \mathcal{R}_0$. On the other hand, for any $f \in \mathcal{R}_0$, hence $f = C_K g$ for some $g \in \mathcal{F}_0(X; \boldsymbol{H})$, we have

$$f(y) = \sum_{x \in X} K(y,x) g(x) = \sum_{x \in X} K_x(y) g(x), \quad y \in X, \quad (3.5)$$

hence $\mathcal{R}_0 = \mathrm{Lin}\{K_x y \mid x \in X,\ h \in \mathcal{H}_x\}$. In addition,

$$\langle f, K_x h \rangle_{\mathcal{R}_0} = \langle f, C_K \widehat{h} \rangle_{\mathcal{R}_0} = \sum_{y \in X} \langle f(y), (\widehat{h})(y) \rangle_{\mathcal{H}_y} = \langle f(x), h \rangle_{\mathcal{H}_x}.$$

Thus, the inner product space $(\mathcal{R}_0; \langle \cdot, \cdot \rangle_{\mathcal{R}_0})$ has all properties (rk1)–(rk3), as well as a modified version of the minimality property (rk4), except the fact that it is a Hilbert space.

By the standard procedure, let $(\mathcal{R}; \langle \cdot, \cdot \rangle_{\mathcal{R}})$ be an abstract completion of the inner product space $(\mathcal{R}_0; \langle \cdot, \cdot \rangle_{\mathcal{R}})$ to a Hilbert space. In order to finish this construction, all we have to prove is that we can always choose $\mathcal{R} \subseteq \mathcal{F}(X; \boldsymbol{H})$, in other words, this Hilbert space abstract

completion can always be realised inside $\mathcal{F}(X;\boldsymbol{H})$. Once this done, after a moment of thought and taking into account that $(\mathcal{R}_0;\langle\cdot,\cdot\rangle_{\mathcal{R}})$ essentially has all properties (rk1)–(rk4), we can see that $(\mathcal{R};\langle\cdot,\cdot\rangle_{\mathcal{R}})$ is the reproducing kernel Hilbert space with reproducing kernel K.

Now, in order to prove that the Hilbert space abstract completion of $(\mathcal{R}_0;\langle\cdot,\cdot\rangle_{\mathcal{R}_0})$ can be realised within $\mathcal{F}(X;\boldsymbol{H})$, we can take at least two paths. One way is to use the existence part of the reproducing kernel Hilbert space associated to K, a consequence of Theorem 2. A second, more direct way, is to show that any Cauchy sequence, with respect to $\|\cdot\|_{\mathcal{R}_0}$, with elements in $\mathcal{R}_0$, converges pointwise on X to a vector cross-section in $\mathcal{F}(X;\boldsymbol{H})$ and that this vector cross-section can be taken as the strong limit of the sequence as well.

Appendix D. Proof of Theorem 4

(a)⇒(b). Let $x\in X$ be fixed, but arbitrary. It was already observed in Subsection 2.1 that, if $\mathcal{H}_K$ is the reproducing kernel Hilbert space in $\mathcal{F}(X;\boldsymbol{H})$ with kernel K, then by the reproducing property, we have

$$\langle f(x),h\rangle_{\mathcal{H}_x}=\langle f,K_xh\rangle_{\mathcal{H}_K},\quad f\in\mathcal{H}_K,\ h\in\mathcal{H}_x,$$

where $K_x\colon\mathcal{H}_x\to\mathcal{H}_K$ is the linear operator defined by $K_xh:=K(\cdot,x)h$, see the axiom (rk2). Since, by axiom (rk2), $K_xh\in\mathcal{H}_K$ for all $h\in\mathcal{H}_x$, the operator K_x is correctly defined. It is a bounded operator because

$$\|K_xh\|^2_{\mathcal{H}_K}=\langle K_xh,K_xh\rangle_{\mathcal{H}_K}=\langle(K_xh)(x),h\rangle_{\mathcal{H}_x}=\langle K(x,x)h\rangle_{\mathcal{H}_x}\leq\|K(x,x)\|\|h\|_{\mathcal{H}_x},\quad(4.1)$$

where we have used the reproducing property (rk3).

Finally, again by the reproducing property (rk3), for any $f\in\mathcal{H}_K$ and any $h\in\mathcal{H}_x$ we have

$$\langle f(x),h\rangle_{\mathcal{H}_x}=\langle f,K_xh\rangle_{\mathcal{H}_K}=\langle K_x^*f,h\rangle_{\mathcal{H}_x},$$

hence the evaluation operator $\mathcal{H}_K\ni f\mapsto f(x)\in\mathcal{H}_x$ coincides with K_x^* and hence it is bounded.

(b)⇒(a). For arbitrary $x\in X$, let $\mathrm{Ev}_x\colon\mathcal{H}\to\mathcal{H}_x$ be the evaluation operator $\mathrm{Ev}_xf:=f(x)$, for all $f\in\mathcal{H}$. By assumption, Ev_x is a bounded operator for all $x\in X$. We consider the $\boldsymbol{H}$-valued kernel

$$K(y,x)=\mathrm{Ev}_y\mathrm{Ev}_x^*,\quad x,y\in X.$$

From Theorem 1 it follows that K is a positive semidefinite $\boldsymbol{H}$-valued kernel and hence, by Theorem 2, there exists and it is unique, the reproducing kernel Hilbert space $\mathcal{H}_K$ with kernel K. In the following we show that $\mathcal{H}$ is the reproducing kernel Hilbert space with kernel K.

The axiom (rk1) holds, by assumption. For the axiom (rk2), let us observe that, for all $x\in X$ and $h\in\mathcal{H}_x$, we have

$$(K_xh)(y)=K(y,x)h=\mathrm{Ev}_y\mathrm{Ev}_x^*h=(\mathrm{Ev}_x^*h)(y),\quad y\in X,$$

hence $K_xh=\mathrm{Ev}_x^*h\in\mathcal{H}_K$. This proves that the axiom (rk2) holds and, in addition, that

$$K_x^*=\mathrm{Ev}_x,\quad x\in X.$$

Finally, for the axiom (rk3), let $f \in \mathcal{H}$, $x \in X$, and $h \in \mathcal{H}_x$ be arbitrary. Then,

$$\langle f(x), h\rangle_{H_x} = \langle \mathrm{Ev}_x f, h\rangle_{\mathcal{H}_x} = \langle f, \mathrm{Ev}_x^* h\rangle_{\mathcal{H}_K} = \langle f, K_x h\rangle_{\mathcal{H}_K}.$$

This shows that the axiom (rk3) holds as well.

Finally, by the uniqueness of the reproducing kernel Hilbert space associated to K, it follows that $\mathcal{H} = \mathcal{H}_K$.

References

N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

S. Ay and A. Gheondea. Representations of $*$-semigroups associated to invariant kernels with values adjointable operators. *Linear Algebra and Applications*, 486:361–388, 2015.

S. Ay and A. Gheondea. Representations of $*$-semigroups associated to invariant kernels with values continuously adjointable operators. *Integral Equations and Operator Theory*, 87(2):263–307, 2017.

S. Ay and A. Gheondea. Invariant weakly positive semidefinite kernels with values in topologically ordered $*$-spaces. *Studia Mathematica*, 248:255–294, 2019.

M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

S. Bergman. Über die Entwicklung der Harmonischen Funktionen der Ebene und des Raumes nach Orthogonalfunktionen. *Mathematische Annalen*, 86:238–271, 1922.

B. Blackadar. *Operator Algebras. Theory of C^*-Algebras and von Neumann Algebras.* Springer, 2006.

S. Bochner. Hilbert distances and positive definite functions. *Annals of Mathematics*, 42: 647–656, 1941.

U. Brefeld, T. Tärtner, T. Scheffer, and S. Wrobel. Efficient co-regularised least square regression. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2006.

C. Brouard, F. D'Alche-Buc, and M. Szafranski. Semi-supervised learning penalized output kernel regression for link prediction. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2011.

R.H. Byrd, Mary E. Hribar, and Jorge Nocedal. An interior point algorithm for large-scale nonlinear programming. *SIAM Journal on Optimization*, 9(4):877–900, 1999.

R.H. Byrd, J.C. Gilbert, and J. Nocedal. A trust region method based on interior point techniques for nonlinear programming. *Mathematical Programming*, 89(1):149–185, 2000.

A. Caponnetto, M. Pontil, C. Micchelli, and Y. Ying. Universal multi-task kernels. *Journal of Machine Learning Research*, 9:1615–1646, 2008.

C. Carmelli, E. De Vito, and A. Toigo. Vector-valued reproducing kernel hilbert spaces of integrable functions and mercer theorem. *Analysis and Applications*, 4:377–408, 2006.

O. Chapelle, B. Schölkopf, and A. Zien (Eds.). *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

T. Constantinescu. *Schur Parameters, Factorization and Dilation Problems*. Birkhäuser Verlag, Basel, 1996.

T. Constantinescu and A. Gheondea. Representations of Hermitian kernels by means of Krein spaces. *Publ. RIMS. Kyoto Univ.*, 33:917–951, 1997.

T. Constantinescu and A. Gheondea. Representations of Hermitian kernels by means of Krein spaces. II. invariant kernels. *Communications in Mathematical Physics*, 216:409–430, 2001.

K.R. Davidson. *C^*-Algebras by Example*. American Mathematical Society, 1996.

D.J. Evans and J.T. Lewis. *Dilations of Irreversible Evolutions in Algebraic Quantum Theory*. Communications of the Dublin Institute for Advanced Studies, Series A (Theoretical Physics), Vol. 24, Dublin, 1977.

T. Evgeniou, M. Pontil, and C.A. Micchelli. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6:615–637, 2005.

A. Gheondea. Dilations of some VH-spaces operator valued invariant kernels. *Integral Equations and Operator Theory*, 74:451–479, 2012.

R. Godement. Les fonctions de type positif et la théorie des groupes. *Transactions of the American Mathematical Society*, 63:1–84, 1948.

R. Gower, D.A. Lorenz, and M. Winkler. A Bregman-Kaczmarz method for nonlinear systems of equations. *Computational Optimization and Applications*, 87:1059–1098, 2024.

Y. Hashimoto, I. Ishikawa, M. Ikeda, F. Komura, T. Katsura, and Y. Kawahara T. Reproducing kernel hilbert C^*-module and kernel mean embeddings. *Journal of Machine Learning Research*, 22(267):1–56, 2021.

Y. Hashimoto, Z. Wang, and T.C. Matsui. $*$-Algebra net: A new approach generalizing neural network parameters to C^*-algebra. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

Y. Hashimoto, M. Ikeda, and H. Kadri. Learning in rkhm: a C^*-algebraic twist for kernel machines. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023a.

Y. Hashimoto, M. Ikeda, and H. Kadri. Deep learning with kernels through RKHM and the Perron-Frobenius operator. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*, 2023b.

Y. Hashimoto, M. Ikeda, and H. Kadri. C^*-algebraic machine learning: Moving in a new direction. In *arXiv:2402.02637 [cs.LG]*, 2024.

S. Hazely, D. Kamzolov, D. Pasechnyuk, A. Gasnikov, P. Richtarik, and M. Takac. A damped Newton method achieves global $\mathcal{O}(1/k^2)$ and local quadratic convergence rate. In *36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, 2022.

H. Kadri, E. Duflos, Ph. Preux, S. Canu, N. Rakotomamonjy, and N. Audifren. Operator-valued kernels for learning from functional response data. *Journal of Machine Learning Research*, 17:1–54, 2016.

N. Kolmogorov. Stationary sequences in Hilbert space [Russian]. *Bull. Math. Univ. Moscow*, 2:1–40, 1941.

M.G. Krein. Hermitian-positive kernels on homogenous spaces. I. *Ukrain Math. Zurnal/ Amer. Math. Soc. Translations (2)*, 1:4/34:64–98/109–164, 1949/1963.

M.G. Krein. Hermitian-positive kernels on homogenous spaces. II. *Ukrain Math. Zurnal/ Amer. Math. Soc. Translation (2)*, pages 10–59/109–164, 1950/1963.

O.R. Kwon and H. Zou. Leaky hockey stick loss: the first negatively divergent margin-based loss function for classification. *Journal of Machine Learning Research*, 24:1–40, 2023.

A. Lambert. *Learning Function-Valued Functions in Reproducing Kernel Hilbert Spaces with Integral Loss: Application to Infinite Task Learning*. PhD Thesis, Institut Polytechnique de Paris, 2021.

E.C. Lance. *Hilbert C^*-Modules: A Toolkit for Operator Algebraists*. Cambridge University Press, 1995.

R.M. Loynes. On generalized positive-definite functions. *Proc. London Math. Soc. III. Ser.*, 15:373–384, 1965a.

R.M. Loynes. Linear operators in VH-spaces. *Trans. Amer. Math. Soc.*, 116:167–180, 1965b.

R.M. Loynes. Some problems arising from spectral analysis. In *Symposium on Probability Methods in Analysis (Loutraki, 1966)*, pages 197–207. Springer Verlag, Berlin, 1967.

Y. Luo, D. Tao, C. Xu, C. Xu, H. Liu, and Y. Wen. Multiview vector-valued manifold regularization for multilabel image classification. *IEEE Transactions on Neural Networks and Learning Systems*, 24(5):709–722, 2013.

M. D. McKay, R. J. Beckman, and W. J. Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21(2):239–245, 1979.

J. Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philos. Trans. Roy. Soc. London Ser. A*, 209:415–446, 1909.

C. A. Micchelli and M. Pontil. On learning vector-valued functions. *Neural Computation*, 17:177–204, 2005.

H.Q. Minh and V. Sindwhani. Vector-valued manifold regularization. In *Proceedings of the 28th International Conference on Machine Learning*, 2011.

H.Q. Minh, L. Bazzani, and V. Murino. A unifying framework in vector-valued reproducing kernel hilbert spaces for manifold regularization and co-regularized multi-view learning. *Journal of Machine Learning Research*, 17:1–72, 2016.

R.D.C. Monteiro and J.-S. Pang. A potential reduction Newton method for constrained equations. *SIAM J. Optim.*, 9:729–754, 1999.

E.H. Moore. General Analysis. *Memoirs of the American Philosophical Society*, Part I, 1935, Part II, 1939.

K.R. Parthasarathy and K. Schmidt. *Positive-definite Kernels, Continuous Tensor Products and Central Limit Theorems of Probability Theory.* Lecture Notes in Mathematics vol. 272 Springer Verlag, Berlin, 1972.

G. Pedrick. *Theory of Reproducing Kernels for Hilbert Spaces of Vector Valued Functions.* PhD Thesis, Unversity of Kansas, 1957.

J. Peypouquet. *Convex Optimization in Normed Spaces.* Springer, Berlin, 2014.

C. Popa. Convergence rates for Kaczmarz type algorithms. *Numerical Algorithms*, 79:1: 1–17, 2018.

D. Rosenberg, V. Sindhwani, P. Bartlett, and P. Niyogi. A kernel for semi-supervised learning with multi-view point cloud regularization. *IEEE Sig. Proc. Mag.*, 26(5):145–150, 2009.

B. Schölkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* The MIT Press, Cambridge, 2002.

L. Schwartz. Sous espace hilbertiens d'espaces vectoriel topologiques et noyaux associés (noyaux reproduisants). *Journal d'Analys Mathematiques*, 13:115–256, 1964.

V. Sindhwani and D.S. Rosenberg. An RKHS for multi-view learning and manifold co-regularization. In *Proceedings of the 25th International Conference on Machine Learning,* Helsinki, Finland, 2008.

M. Stein. Large sample properties of simulations using latin hypercube sampling. *Technometrics*, 29(2):143–151, 1987.

I. Steinwart and A. Christmann. *Support Vector Machines.* Springer Verlag, Berlin, 2008.

S. Sun. Multi-view Laplacian support vector machines. In *Proceedings of the International Conference on Advanced Data Mining and Applications (ADMA)*, pages 209–211, 2011.

B. Sz.-Nagy. *Prolongement des transformations de l'espace de Hilbert qui sortent de cet espace, in* Appendice au livre "Leçons d'analyse fonctionnelle" par F. Riesz et B. Sz.-Nagy, pp. 439-573. Akademiai Kiado, Budapest, 1955.

A.N. Tikhonov. Solution of ill-posed formulated problems and the regularization method [Russian]. *Doklady Akademii Nauk SSSR*, 151:501–504, 1963.

R.A. Waltz, J. L. Morales, J. Nocedal, and D. Orban. An interior algorithm for nonlinear optimization that combines line search and trust region steps. *Mathematical Programming*, 107(3):391–408, 2006.

L. Zhao, M. Mammadov, and J. Yearwood. From convex to nonconvex: a loss function analysis for binary classification. In *IEEE International Conference on Data Mining Workshops*, pages 1281–1288, 2010.