A Generalized Representer Theorem for Hilbert Space - Valued Functions

Sanket Diwale SANKET.DIWALE@EPFL.CH
Colin N. Jones COLIN.JONES@EPFL.CH
Automatic Control Laboratory
École Polytechnique Fédérale de Lausanne
Lausanne, Switzerland

Abstract

The necessary and sufficient conditions for existence of a generalized representer theorem are presented for learning Hilbert space - valued functions. Representer theorems involving explicit basis functions and Reproducing Kernels are a common occurrence in various machine learning algorithms like generalized least squares, support vector machines, Gaussian process regression, and kernel-based deep neural networks to name a few. Due to the more general structure of the underlying variational problems, the theory is also relevant to other application areas like optimal control, signal processing and decision making. The following presents a generalized representer theorem using the theory of closed, densely defined linear operators and subspace valued maps as a means to address variational optimization problems in learning and control. The implications of the theorem are presented with examples of multi-input - multi-output problems from kernel-based deep neural networks, stochastic regression and sparsity learning problems.

Keywords: Linear Operators, Adjoints, Kernels, Representer Theorems

1. Introduction

The development of kernel-based methods for regression and machine learning has a long history with several algorithms basing themselves on the Reproducing Kernel Hilbert Space (RKHS) theory. Some of the early works in the field include Aronszajn (1950); Tikhonov (1963); Wahba (1990), which looked at problems of spline interpolation and smoothing in the RKHS setting. Several practical learning algorithms like linear regression, support vector machines, Bayesian regression were also developed in their kernel forms to allow more complex nonlinear representations of data (see Bishop (2006) for some examples). Kernel-based stochastic models are also popular in the form of Gaussian Process models (see Rasmussen (2006)). RKHS based neural networks are investigated in Cho and Saul (2009); Rebai et al. (2016); Damianou and Lawrence (2013).

Most such problems (see example 6) from learning and control in their general form can be written as a variational optimization problem of the following form,

$$f_{opt} = \text{argmin}_{f \in \mathcal{H}} \quad C(Lf) + \Omega(f) \tag{1}$$

where $\mathcal{H}$ and $\mathcal{Z}$ are some separable Hilbert spaces (possibly infinite dimensional, e.g. spaces of square integrable functions), $L : \mathcal{H} \to \mathcal{Z}$ is a closed, densely defined linear operators, and

$C : \mathcal{Z} \to \mathbb{R} \cup \{+\infty\}$ and $\Omega : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ are general nonlinear functionals encoding the cost functions, regularizers and constraints in the problem. The functionals C and Ω are written separately as different properties are assumed to hold for the two functionals (see Section 3). We also hide the fact (in the notation of $C : \mathcal{Z} \to \mathbb{R} \cup \{+\infty\}$) that the functional can be dependent on additional inputs like the data set used for learning which are fixed during the optimization and thus not explicitly shown in the notation.

Let $\mathcal{V}(\mathcal{H})$ denote the collection of all closed vector subspaces in $\mathcal{H}$ and $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ denote a map from a vector in $\mathcal{H}$ to a closed subspace of $\mathcal{H}$. Also let S have a subspace valued extension $S : \mathcal{V}(\mathcal{H}) \to \mathcal{V}(\mathcal{H})$ given by the union operation, i.e, for any $A \in \mathcal{V}(\mathcal{H})$, $S(A) = \cup_{a \in A} S(a)$, must belong to $\mathcal{V}(\mathcal{H})$. Let $L^* : \mathcal{Z} \to \mathcal{H}$, defined on a dense subset $\mathrm{dom}(L^*) \subseteq \mathcal{Z}$, denote the adjoint operator to the closed, densely defined operator $L : \mathcal{H} \to \mathcal{Z}$. Let $\mathrm{range}(L^*)$ denote the range of the operator L^* given by the set $\{L^*(z) : z \in \mathrm{dom}(L^*)\}$. The generalized representer theorem (Theorem 18) then states under certain assumptions on C, Ω and S, that an optimal solution for (1) can be found in a subspace of $\mathcal{H}$ (often finite dimensional) given by $S(\mathrm{range}(L^\star))$, i.e.,

$$f_{opt} \in S(\mathrm{range}(L^*)) \tag{2}$$

Representer theorems thus provide a means to reduce infinite dimensional optimization problems for learning in the Hilbert space $\mathcal{H}$ to an equivalent and often tractable finite dimensional optimization in $\mathcal{Z}$ of the form,

$$\begin{aligned} f_{opt} &= L^* z_{opt} \\ z_{opt} &= \mathrm{argmin}_{f \in S(\mathrm{range}(L^*))} \quad C \circ Lf + \Omega(f) \end{aligned} \tag{3}$$

If $\mathcal{Z}$ and $S(\mathrm{range}(L^*))$ are finite dimensional then (3) is a finite dimensional optimization.

A key underlying tool in the use of RKHS methods is the Riesz Representer Theorem (Conway, Theorem 3.3.1) and the existence and uniqueness of adjoint operators for bounded linear operators given by (Conway, Theorem 5.4.2). The above two theorems combined with restrictions on the forms of the objective and constraint functionals in learning problems have led to several variants of representer theorems. Early variants of representer theorems are presented in Wahba (1990) for variational problems in learning real valued functions with least squares regularization. Representer theorems for kernel versions of different learning algorithms like SVM, PCA, CCA, ICA can be found in Suykens et al. (2010). Works like Micchelli and Pontil (2005); Minh and Sindhwani (2011); Minh et al. (2016) present representer theorems for kernel based learning methods for vector valued functions in Hilbert spaces. While these works cover a large set of learning algorithms, the representer theorem needed to be proven individually for each problem. This has prompted investigation into unifying representer theorems into a single generalized theorem and characterizing the class of problems for which a representer theorem can be guaranteed to exist.

The first such results appear to have come from Schölkopf et al. (2001), where the problem is addressed for learning real valued functions with functionals of the form (4).

$$f_{opt} = \mathrm{argmin}_{f \in \mathcal{H}} \quad C(f(x_1), \ldots, f(x_m)) + \Omega(||f||_{\mathcal{H}}) \tag{4}$$

where $\mathcal{H}$ is a reproducing kernel Hilbert space of $\mathbb{R}$-valued functions with kernel K, $f(x_1), \ldots, f(x_m)$ are function evaluations for f at given points $x_1, \ldots, x_m$. The functional C is of the

form, $C : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$, and $\Omega : [0, \infty) \to \mathbb{R}$ is a strictly monotonically increasing function. The strictly increasing monotonic property of Ω was shown to be a sufficient condition for the existence of a representer such that,

$$f_{opt} \in \left\{ \sum_{i=1}^{m} c_i K(\cdot, x_i) : c_i \in \mathbb{R} \right\} \tag{5}$$

The regularizers (written as a function of the norm of f) showed how kernel versions of the least squares algorithms in linear regression, SVMs and others are covered by a single generalized theorem.

Dinuzzo and Schölkopf (2012) relaxed the restriction on the regularizer further and provided necessary and sufficient conditions for the existence of representer theorems. Dinuzzo and Schölkopf (2012) considers problems of the form

$$f_{opt} = \text{argmin}_{f \in \mathcal{H}} \quad C(\langle w_1, f \rangle_{\mathcal{H}}, \ldots, \langle w_m, f \rangle_{\mathcal{H}}) + \Omega(f) \tag{6}$$

where $\mathcal{H}$ is a separable Hilbert space, $w_1, \ldots, w_m \in \mathcal{H}$ are given vectors corresponding to bounded functionals on $\mathcal{H}$ and functionals $C : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ and $\Omega : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ are lower semi-continuous functionals. If for all orthogonal vectors $f, g \in \mathcal{H}$ ($\langle f, g \rangle_{\mathcal{H}} = 0$), $\Omega(f + g) \geq \max\{\Omega(f), \Omega(g)\}$, the functional Ω is called "orthomonotone". It was also shown that this orthomonotone property is necessary and sufficient for the existence of a representer in the form,

$$f_{opt} \in \left\{ \sum_{i=1}^{m} c_i w_i : c_i \in \mathbb{R} \right\} \tag{7}$$

Schölkopf et al. (2001); Dinuzzo and Schölkopf (2012) restricted the scope of their theorem to learning $\mathbb{R}$-valued functions. The generalized theorem was extended to learning multi-output functions in Argyriou and Dinuzzo (2014) with the help of subspace valued maps $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$. Argyriou and Dinuzzo (2014) considers problems of the form,

$$f_{opt} = \text{argmin}_{f \in \mathcal{H}} \quad C(\langle w_1, f \rangle_{\mathcal{H}}, \ldots, \langle w_m, f \rangle_{\mathcal{H}}) + \Omega(f) \tag{8}$$

where $\mathcal{H}, w_1, \ldots, w_m$, $C : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ and $\Omega : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ are as before from Dinuzzo and Schölkopf (2012). However, Ω satisfies the orthomonotone property with respect to a subspace valued map $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$, defined as, for any $f \in \mathcal{H}$ and $g \in S(f)^{\perp}$, $\Omega(f + g) \geq \Omega(f)$. The representer theorem then provides that,

$$f_{opt} \in \sum_{i=1}^{m} S(w_i) \tag{9}$$

(the summation over sets $S(w_i) + S(w_j)$ being considered as the pairwise addition $a + b$ of all possible pairs $(a, b) \in S(w_i) \times S(w_j)$).

The results from Schölkopf et al. (2001); Dinuzzo and Schölkopf (2012) can be viewed under this framework as Ω being orthomonotone with respect to a trivial map $S_{\mathbb{R}}(w_i) = \{\lambda w_i : \lambda \in \mathbb{R}\}$. The inclusion of orthomonotonicity with respect to non trivial subspace valued maps allows the consideration of a larger class of regularizers for Ω including regularizers like the ℓ_1-norm, Frobenius norm, trace norm and general spectral norms in matrix

learning problems Argyriou et al. (2009). For learning a matrix $f \in \mathcal{H} = \mathbb{R}^{m\times n}$, with Ω being a monotonically increasing penalty on $f^T f$, (Argyriou and Dinuzzo, 2014, Example 4.2) shows the representer is of the form $\sum_{i=1}^{m} S(w_i)$ with $S(w_i) = \{w_i c_i : c_i \in \mathbb{R}^{n\times n}\}$ for given $w_i \in \mathcal{H} = \mathbb{R}^{m\times n}$, thus showing the role of S in extending the result from Dinuzzo and Schölkopf (2012) to a multi-output scenario.

Argyriou and Dinuzzo (2014) however makes an assumption of "r-regularity" (see appendix for definition) on the allowed subspace valued map $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$, which requires for all $w \in \mathcal{H}$, the dimension of $S(w) \leq r$ for some finite $r \leq m$. We show in Section 4.3 that for ℓ_1-regularization on function spaces the functional Ω is orthomonotone with respect to a non r-regular subspace valued map S, i.e. no such finite r exists for all $w \in \mathcal{H}$. Theorem 18 eliminates the r-regularity assumption and enables the generalized representer theorem to be applied to such problems.

The prior counter parts of Theorem 18 Schölkopf et al. (2001); Argyriou et al. (2009); Dinuzzo and Schölkopf (2012); Argyriou and Dinuzzo (2014) also consider functionals $C : \mathbb{R}^m \to \mathbb{R}\cup\{+\infty\}$ defined on $\mathbb{R}^m$ instead of an arbitrary separable Hilbert space $\mathcal{Z}$. In Section 4.2, we show with an example of stochastic process regression the utility of considering loss functionals $C : \mathcal{Z} \to \mathbb{R} \cup \{\infty\}$ over an infinite dimensional Hilbert space $\mathcal{Z}$. The learning problems for stochastic processes require loss functionals to be defined over a Hilbert space of measurable functions (not isomorphic to $\mathbb{R}^m$) and were thus outside the scope of previous generalized representer theorems from Schölkopf et al. (2001); Argyriou et al. (2009); Dinuzzo and Schölkopf (2012); Argyriou and Dinuzzo (2014).

We thus present here an extension for the generalized representer theorem where the functional $C : \mathcal{Z} \to \mathbb{R} \cup \{+\infty\}$ is a lower semi-continuous non-linear functional over an arbitrary Hilbert space $\mathcal{Z}$, in terms of non r-regular subspace valued maps and adjoints of closed, densely defined linear operators.

The chapter is structured as follows. Section 2 presents some preliminary definitions and results of existing notions required to establish the generalized representer theorem. Section 2.1 presents some background material on linear operators and their adjoints. Section 2.2 presents the notion of a subspace valued map and Section 2.3 presents the notion of orthomonotone functionals with respect to a subspace valued map. The generalized representer theorem giving necessary and sufficient conditions for the existence of a representer is then presented in Section 3. Section 4 presents examples of some simple learning problems to highlight extensions made by the representer theorem. The appendix provides proofs for some lemmas and discussion with regards to the subspace valued maps considered in the chapter and their relation to properties of quasilinear, idempotent and r-regular subspace valued maps used in previous works.

2. Preliminaries

The notions of adjoints and closed operators are known to be crucial in determining solutions to linear inverse problem of the form $Lx = y$ (find x given y) Kulkarni and Nair (2000). It is thus natural for them to be important in the theory for a generalized representer theorem (which cover problems of the form $Lx = y$ as a special case). Section 2.1 presents some preliminary, well known results that will be useful in proving the generalized representer theorem.

2.1 Closed linear operators and adjoint operators

Let $\mathcal{H}$ and $\mathcal{Z}$ be two arbitrary separable Hilbert spaces. Let $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, $\langle \cdot, \cdot \rangle_{\mathcal{Z}}$ be the inner products defined on $\mathcal{H}$ and $\mathcal{Z}$ respectively. A closed linear operator from $\mathcal{H}$ to $\mathcal{Z}$ is defined as follows.

Definition 1 *(Closed linear operator)*
Let $\mathcal{H}, \mathcal{Z}$ be two separable Hilbert spaces. Let $\mathrm{dom}(L) \subseteq \mathcal{H}$ be the domain for a linear operator $L : \mathrm{dom}(L) \to \mathcal{Z}$. L is called a closed operator if the graph of the operator, $\mathrm{graph}(L) = \{(x, Lx) : x \in \mathrm{dom}(L)\}$ is a closed subset of $\mathcal{H} \times \mathcal{Z}$.

An operator L is called closable if there exists an extension to L that is closed.

A linear operator (not necessarily closed) is said to be densely defined on $\mathcal{H}$ if $\mathrm{dom}(L)$ is a dense subset of $\mathcal{H}$. Let $L : \mathrm{dom}(L) \to \mathcal{Z}$ be a linear operator, densely defined on $\mathcal{H}$. Then an adjoint operator can be defined as follows,

Definition 2 *(Adjoint for densely defined operators)*
Let $\mathrm{dom}(L)$ be a dense subset of $\mathcal{H}$ and $L : \mathrm{dom}(L) \to \mathcal{Z}$ be a densely defined operator (also denoted as $L : \mathcal{H} \to \mathcal{Z}$). Let

$$\mathrm{dom}(L^*) := \{z \in \mathcal{Z} : f(h) = \langle Lh, z \rangle_{\mathcal{Z}} \text{ is bounded linear functional on } dom(L)\}$$

. The adjoint $L^* : \mathrm{dom}(L^*) \to \mathcal{H}$ is defined as the operator mapping $z \in \mathrm{dom}(L^*)$ to a dual in $\mathcal{H}$ such that,

$$\forall f \in \mathrm{dom}(L), z \in \mathrm{dom}(L^*) \qquad \langle Lf, z \rangle_{\mathcal{Z}} = \langle f, L^* z \rangle_{\mathcal{H}} \tag{10}$$

By (Conway, 1994, Chapter 10, Proposition 1.6), if the operator $L : \mathrm{dom}(L) \to \mathcal{Z}$ is closable and densely defined then the adjoint L^* is a closed, densely defined operator, i.e., $\mathrm{dom}(L^*)$ is a dense subset of $\mathcal{Z}$. For a closed densely defined operator $L : \mathcal{H} \to \mathcal{Z}$, $L^*L : \mathrm{dom}(L^*L) \subseteq \mathcal{H} \to \mathcal{H}$ and $LL^* : \mathrm{dom}(LL^*) \subseteq \mathcal{Z} \to \mathcal{Z}$ are closed, densely defined, self-adjoint operators Sandovici (2018). Also, for a closed and bounded operator $L : \mathrm{dom}(L) \subseteq \mathcal{H} \to \mathcal{Z}$ the domain is the entire space, i.e. $\mathrm{dom}(L) = \mathcal{H}$ and the adjoint L^* is also closed and bounded.

Further, by Banach's closed range theorem (Yoshida, 2013, Chapter 7.5), the null space of a densely defined, closed linear operator $\mathcal{N}_L = \{f \in \mathrm{dom}(L) : Lf = 0\}$ is a closed subset in $\mathcal{H}$ and can be characterized in terms of the orthogonal complementary space $\mathcal{N}_L^{\perp}$ and the adjoint operator $L^* : \mathrm{dom}(L^*) \to \mathcal{H}$ as follows,

Lemma 3 *Let $\mathcal{N}_L$ be the null space of a closed, densely defined operator $L : \mathrm{dom}(L) \to \mathcal{Z}$ and $\mathcal{N}_L^{\perp}$ be its orthogonal complementary space, then,*

$$\mathcal{N}_L^{\perp} = \mathrm{range}(L^*) = \{L^* z : z \in \mathrm{dom}(L^*)\}$$

The above lemma is a direct result of the closed range theorem and we refer the reader to (Yoshida, 2013, Chapter 7.5) for the proof.

Corollary 4 *For some finite $m \in \mathbb{N}$, let $\{L_i : \mathrm{dom}(L_i) \to \mathcal{Z}_i : i = 1, \ldots, m\}$ be a set of closed, densely defined operators with $\mathcal{Z}_i$ being separable Hilbert spaces and $\mathrm{dom}(L_i) \subseteq \mathcal{H}$ for some separable Hilbert space $\mathcal{H}$. Let $\cap_{i=1}^m \mathrm{dom}(L_i)$ be a dense subset of $\mathcal{H}$. The joint*

null space is $\mathcal{N}_{L_1,\ldots,L_m} = \mathcal{N}_{L_1} \cap \cdots \cap \mathcal{N}_{L_m}$ *and* $\mathcal{N}^{\perp}_{L_1,\ldots,L_m} = \mathcal{N}^{\perp}_{L_1} + \cdots + \mathcal{N}^{\perp}_{L_m} = \{\sum_{i=1}^{m} L_i^* z_i : z_i \in \mathrm{dom}(L_i^*)\}$.

Proof *Consider the Hilbert space* $\mathcal{Z}$ *given as the direct sum of* $\mathcal{Z}_i$, $i = 1, \ldots, m$, *i.e.* $\mathcal{Z} = \mathcal{Z}_1 \oplus \cdots \oplus \mathcal{Z}_m$. *The inner product on* $\mathcal{Z}$ *is given by* $\langle (z_1, \ldots, z_m), (y_1, \ldots, y_m) \rangle_{\mathcal{Z}} = \sum_{i=1}^{m} \langle z_i, y_i \rangle_{\mathcal{Z}_i}$. *Consider then the linear operator* $L : \cap_{i=1}^{m} \mathrm{dom}(L_i) \to \mathcal{Z}$ *given as* $Lf = (L_1 f, \ldots, L_m f)$. *By assumption,* $\cap_{i=1}^{m} \mathrm{dom}(L_i)$ *is a dense subset of* $\mathcal{H}$ *and thus* L *is a densely defined operator. Further* $\mathrm{graph}(L) = \{(x, Lx) : x \in \mathrm{dom}(L)\}$ *is a closed subset since for every converging sequence* $x_n \in \mathrm{dom}(L)$, $Lx_n = (L_1 x_n, \ldots, L_m x_n)$ *converges to a point* $(L_1 x, \ldots, L_m x)$ *with* $(x, L_i x) \in \mathrm{graph}(L_i)$ *(since* L_i *is a closed operator). Thus* L *is a closed, densely defined operator with* $\mathrm{dom}(L) = \cap_{i=1}^{m} \mathrm{dom}(L_i)$.

Clearly $\mathcal{N}_L = \mathcal{N}_{L_1,\ldots,L_m} = \cap_{i=1}^{m} \mathcal{N}_{L_i}$. *The adjoint domain* $\mathrm{dom}(L^*) = \{(z_1, \ldots, z_m) \in \mathcal{Z} : f(h) = \sum_{i=1}^{m} \langle L_i f, z_i \rangle_{\mathcal{Z}_i}$ *is bounded*$\} = \mathrm{dom}(L_1^*) \times \cdots \times \mathrm{dom}(L_m^*)$. *The adjoint* $L^* : \mathrm{dom}(L^*) \to \mathcal{H}$, *is such that, for all* $f \in \mathrm{dom}(L)$ *and* $z = (z_1, \ldots, z_m) \in \mathrm{dom}(L^*)$, $\langle L^* z, f \rangle_{\mathcal{H}} = \langle z, Lf \rangle_{\mathcal{Z}} = \sum_{i=1}^{m} \langle z_i, L_i f \rangle_{\mathcal{Z}_i} = \langle \sum_{i=1}^{m} L_i^* z_i, f \rangle_{\mathcal{H}}$. *Thus* $L^* z = \sum_{i=1}^{m} L_i^* z_i$. *Then by Lemma 3,* $\mathcal{N}_L^{\perp} = \mathrm{range}(L^*) = \{\sum_{i=1}^{m} L_i^* z_i : z_i \in \mathrm{dom}(L_i)\}$. ■

When rewriting functionals of the form $C(L_1 f, \ldots, L_m f)$ as $C(Lf)$, Corollary 4 gives the required characterization of the orthogonal null space. Thus the adjoint operator plays a key role in characterizing the null space of an operator $\mathcal{N}_L$ and its orthogonal complementary space $\mathcal{N}_L^{\perp}$.

Closed range characterization for bounded linear operators in terms of the operator spectrum are given in (Kulkarni and Nair, 2000, Theorem 2.5) or equivalently by (Conway, Lemma 5.6.13). Characterization of closed, densely defined operators is given by (Kulkarni et al., 2008, Theorem 3.3).

By (Conway, Proposition 5.6.13), a bounded adjoint $T^* : \mathcal{Z} \to \mathcal{H}$ is a closed range if and only if

$$\inf\{||L^* z||_{\mathcal{H}} : ||z||_{\mathcal{Z}} = 1\} > 0 \tag{11}$$

or equivalently

$$\inf\{\langle z, LL^* z \rangle_{\mathcal{Z}} : ||z||_{\mathcal{Z}} = 1\} > 0 \tag{12}$$

By (Kulkarni et al., 2008, Theorem 3.3) a densely defined operator is closed if and only if, there exists a $\gamma > 0$ such that the spectrum $\sigma(L^* L) \subseteq \{0\} \cup [\gamma, \infty)$.

2.1.1 Adjoint for operators of common interest

Below we show a few examples of adjoint operator for densely defined, closed linear operators commonly seen in learning and control algorithms.

Example 1 *(Evaluation Operators)*
Let $\mathcal{Z}$ *be a separable Hilbert space and* $\mathcal{C}_b(\mathcal{X}, \mathcal{Z})$ *be the separable Banach space of* $\mathcal{Z}$*-valued continuous and bounded functions with a domain set* $\mathcal{X}$. *Let* $\mathcal{H}$ *be a reproducing kernel Hilbert space with kernel* $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}_{\mathcal{Z},\mathcal{Z}}$ *induced by a Gaussian measure on* $\mathcal{C}_b(\mathcal{X}, \mathcal{Z})$. *A parametric linear evaluation operator* $L_x : \mathcal{H} \to \mathcal{Z}$, *given by* $L_x(f) = f(x)$ *for some fixed parameter* $x \in \mathcal{X}$ *is then a bounded linear operator and* $\mathcal{H}$ *is a dense subset in* $\mathcal{C}_b(\mathcal{X}, \mathcal{Z})$ *(Bogachev, 2015, Theorem 3.9.5). The operator commonly occurs in machine learning and*

data fitting problems where x is the training input data and $f(x)$ gives a predicted value for the output in $\mathcal{Z}$. The adjoint $L_x^* : \mathcal{Z} \to \mathcal{H}$ can be found as follows.

Note that by definition of L_x and its adjoint L_x^*, $\forall g \in \mathcal{H}, z \in \mathcal{Z}$, $\langle L_x^* z, g\rangle_{\mathcal{H}} = \langle L_x g, z\rangle_{\mathcal{Z}}$, i.e., $\langle L_x^* z, g\rangle_{\mathcal{H}} = \langle g(x), z\rangle_{\mathcal{Z}}$. When $\mathcal{H}$ is a reproducing kernel Hilbert space with kernel K, L_x^* is well defined and coincides with the definition of the RKHS kernel (see (Micchelli and Pontil, 2005, Definition 2.1)). Thus RKHS spaces provide a case where the adjoint operator for evaluation operators is well defined and $L_x^* = K(\cdot, x)$, i.e. we have $\mathrm{dom}(L_x) = \mathcal{H}$ and $\mathrm{dom}(L_x^*) = \mathcal{Z}$.

The closed range property for L_x^* thus corresponds to the closed range property of the kernel. Using (12), this corresponds to checking $\inf\{\langle z, K(x,x)z\rangle_{\mathcal{Z}} : ||z||_{\mathcal{Z}} = 1\} > 0$. For a positive definite kernel K, this is automatically satisfied and the adjoint is a closed range, bounded linear operator.

Example 2 *(Linear Transformations of an explicit basis ϕ)*
Let $\mathcal{H}, \mathcal{Y}, \mathcal{Z}$ be arbitrary Hilbert spaces. Let $\mathcal{L}_{\mathcal{Y},\mathcal{Z}}$ be the space of bounded, closed range operators from $\mathcal{Y}$ to $\mathcal{Z}$. Let $\phi : \mathcal{X} \to \mathcal{Y}$ be some given $\mathcal{Y}$-valued function and $x \in \mathcal{X}$ be an evaluation point such that $||\phi(x)||_{\mathcal{Y}} < \infty$. Let $\ell : \mathcal{H} \to \mathcal{L}_{\mathcal{Y},\mathcal{Z}}$ be a bounded linear map from $\mathcal{H}$ to $\mathcal{L}_{\mathcal{Y},\mathcal{Z}}$ such that there exists a $\kappa_\ell \in [0, \infty)$ satisfying, for all $W \in \mathcal{H}$, $||\ell(W)||_{\mathcal{L}_{\mathcal{Y},\mathcal{Z}}} \leq \kappa_\ell ||W||_{\mathcal{H}}$. Then we can define a bounded, closed range linear operator $L_{x,\phi} : \mathcal{H} \to \mathcal{Z}$ given as $L_{x,\phi}(W) := \ell(W)\phi(x)$ for any $W \in \mathcal{H}$. The boundedness for the operator follows from the fact that $||L_{x,\phi}(W)||_{\mathcal{Z}} = ||\ell(W)\phi(x)||_{\mathcal{Z}} \leq ||\ell(W)||_{\mathcal{L}_{\mathcal{Y},\mathcal{Z}}} ||\phi(x)||_{\mathcal{Y}} \leq \kappa_\ell ||\phi(x)||_{\mathcal{Y}} ||W||_{\mathcal{H}}$. The adjoint operator satisfies $\langle L_{x,\phi}^* z, W\rangle_{\mathcal{H}} = \langle \ell(W)\phi(x), z\rangle_{\mathcal{Z}}$ and its form depends on further specification of ℓ.

The operator $L_{x,\phi}$ is closed range, if $\inf\{\langle L_{x,\phi} L_{x,\phi}^* z, z\rangle_{\mathcal{Z}} : ||z||_{\mathcal{Z}} = 1\} > 0$, i.e.,

$$\inf\{\langle \ell(L_{x,\phi}^* z)\phi(x), z\rangle_{\mathcal{Z}} : ||z||_{\mathcal{Z}} = 1\} > 0$$

We look at two examples below giving ℓ explicitly and making the adjoint and closed range characterization for the given cases.

Example 2(a) *Finite dimensional $\mathcal{Z}$ example*
Let $\mathcal{Y} = \mathbb{R}^n$, $\mathcal{Z} = \mathbb{R}^k$, $\mathcal{H} = \mathbb{R}^{n \times k}$ and $\phi : \mathcal{X} \to \mathcal{Y}$ is a given basis function and $x \in \mathcal{X}$ with $||\phi(x)||_{\mathcal{Y}} < \infty$. Let $\ell(W) := W^T$ be the bounded operator from $\mathcal{H}$ to $\mathcal{L}_{\mathcal{Y},\mathcal{Z}}$. Then for any $W \in \mathcal{H}$, $L_{x,\phi}(W) = W^T\phi(x)$ and $L_{x,\phi}$ is a bounded operator. Such an operator is common when W represent weights or coefficients to be learned and ϕ is a given vector of basis functions.

Let the inner product on $\mathcal{H}$ be the Frobenius inner product of matrices, i.e, $\langle w_1, w_2\rangle_{\mathcal{H}} = \mathrm{trace}(w_1^T w_2)$. Let inner product on $\mathcal{Z}$ be $\langle z_1, z_2\rangle_{\mathcal{Z}} = z_1^T z_2$. Then for the adjoint operator $\langle L_{x,\phi}^* z, W\rangle_{\mathcal{H}} = \langle W^T\phi(x), z\rangle_{\mathcal{Z}}$, $\forall z \in \mathcal{Z}$ implying $\mathrm{trace}(W^T L_{x,\phi}^* z) = \phi(x)^T W z$. Noting then that $\phi(x)^T W z = \mathrm{trace}(\phi(x)^T W z) = \mathrm{trace}(z^T W^T \phi(x)) = \mathrm{trace}(W^T \phi(x) z^T)$, we can define $L_{x,\phi}^* z := \phi(x) z^T$ with $\mathrm{dom}(L_{x,\phi}) = \mathcal{H}$ and $\mathrm{dom}(L_{x,\phi}^\star) = \mathcal{Z}$.

The operator $L_{x,\phi}$ is closed range if $\inf\{\phi(x)^T\phi(x) z^T z : ||z||_{\mathcal{Z}} = 1\} = \phi(x)^T\phi(x) > 0$.

Example 2(b) *Infinite dimensional $\mathcal{Z}$ example*
Let $\mathcal{X} = \mathbb{R}^n$, $\mathcal{U} = \mathbb{R}^m$ and $\mathcal{H} = \mathbb{R}^{m \times N}$. Let $\{\mathcal{Y}_i : i = 1, \ldots, N\}$ be a collection of RKHS spaces of functions $f : \mathcal{X} \to \mathcal{U}$ with kernels $K_1, \ldots, K_N$. Let $\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_N$ and

$\phi(x) = \begin{pmatrix} K_1(\cdot,x) \\ K_2(\cdot,x) \\ \vdots \\ K_N(\cdot,x) \end{pmatrix}$ *and let* $\mathcal{Z}$ *be some infinite dimensional Hilbert space of functions* $g : \mathcal{X} \to \mathcal{U}$ *with inner product* $\langle g_1, g_2 \rangle_{\mathcal{Z}} = \int_{\mathcal{X}} \langle g_1(x), g_2(x) \rangle_{\mathcal{U}} dx$. *Then we can define a continuous linear operator* $L_{x,\phi} : \mathcal{H} \to \mathcal{Z}$ *for any* $W \in \mathcal{H}$ *as* $L_{x,\phi}(W) := \sum_{i=1}^{N} K_i(\cdot,x) W_i$, *with* W_i *denoting the* i^{th} *column of* W. *Using the Frobenius inner product on* $\mathcal{H}$, $\langle L^*_{x,\phi} g, W \rangle_{\mathcal{H}} = \langle L_{x,\phi}(W), g \rangle_{\mathcal{Z}} = \sum_{i=1}^{N} \int_{\mathcal{X}} \langle K_i(y,x) W_i, g(y) \rangle_{\mathcal{U}} dy = \sum_{i=1}^{N} \int_{\mathcal{X}} W_i^T K_i(x,y) g(y) dy$. *Also note that* $\langle L^*_{x,\phi} g, W \rangle_{\mathcal{H}} = \text{trace}(W^T L^*_{x,\phi} g) = \sum_{i=1}^{N} W_i^T [L^*_{x,\phi} g]_i$. *Thus* $[L^*_{x,\phi} g]_i = \int_{\mathcal{X}} K_i(x,y) g(y) dy$ *gives the adjoint.*

The operator $L_{x,\phi}$ *is closed range, if* $\inf\{\langle L^*_{x,\phi}(L_{x,\phi} W), W \rangle_{\mathcal{H}} : ||W||_{\mathcal{H}} = 1\} = \inf\{\langle L^*_{x,\phi} W, L^*_{x,\phi} W \rangle_{\mathcal{Z}} : ||W||_{\mathcal{H}} = 1\} = \inf\{\sum_{i=1}^{N} \sum_{j=1}^{N} (\int_{\mathcal{X}} W_i^T K_i(x,y)^T K_j(x,y) W_j dy) : ||W||_{\mathcal{H}} = 1\} > 0$.

Such an operator can be used to pose an optimization for learning with weighted kernels.

Example 3 *(Derivative operator in Sobolov Hilbert spaces)*
Let $\Omega \subset \mathbb{R}^n$ *be a open subset of* $\mathbb{R}^n$ *with a smooth boundary* $\partial\Omega$. *Let* $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}^n$ *be a multi-index and* $\partial^\alpha f = \partial^{\alpha_1}_{x_1}, \ldots, \partial^{\alpha_n}_{x_n} f$. *Let* $L^2(\Omega, \mathbb{R}, \mu)$ *be the space of* $\mathbb{R}$*-valued functions, square integrable on* Ω *with respect to a non-negative measure* μ *and* $H^k(\Omega, \mathbb{R}, \mu)$ *be the Sobolov Hilbert space such that* $\partial^\alpha f \in L^2(\Omega, \mu)$ *for all multi-index* $\alpha \in \mathbb{N}^n$ *such that* $|\alpha| \leq k$. *The inner product on* $H^k(\Omega, \mathbb{R}, \mu)$ *is given by* $\langle f, g \rangle_{H^k(\mu)} = \sum_{i=1}^{k} \sum_{\alpha : |\alpha| \leq k} \langle \partial^\alpha f, \partial^\alpha g \rangle_{L^2(\Omega,\mu)}$. *It is also known that* $H^k(\Omega, \mathbb{R}, \mu) \subset L^2(\Omega, \mathbb{R}, \mu)$ *is a dense subset of* $L^2(\Omega, \mathbb{R}, \mu)$ *(Sudan et al., 2012, Prop. 3.10). Thus any differential operator* $D : H^k(\Omega, \mathbb{R}, \mu) \to L^2(\Omega, \mathbb{R}, \mu)$ *defined on* $H^k(\Omega, \mathbb{R}, \mu)$ *is densely defined on* $L^2(\Omega, \mathbb{R}, \mu)$ *with* $\text{dom}(D) = H^k(\Omega, \mathbb{R}, \mu)$. *Consider then a differential operator* $Df = \phi(\cdot)^T \nabla f + \Delta f$ *for a given smooth function* $\phi \in C^\infty(\mathbb{R}^n, \mathbb{R}^n)$, ∇ *and* Δ *are the gradient and Laplace operators respectively. Such an operator* D *is closable (Yoshida, 2013, Page 78). Thus we have* D *as a closable, densely defined operator on* $L^2(\Omega, \mathbb{R}, \mu)$, *implying the adjoint* D^* *is closed and densely defined on* $L^2(\Omega, \mathbb{R}, \mu)$. *For any* g *with differentiability upto order two and* $f \in \text{dom}(D)$ *we have, using integration by parts,*

$$\begin{aligned} \langle Df, g \rangle_{L^2(\Omega,\mathbb{R},\mu)} &= \int_\Omega (\phi(x)^T \nabla f(x) + \Delta f(x)) g(x) d\mu(x) && (13) \\ &= \int_\Omega f(x)(-\nabla \cdot (g\phi)(x) + \Delta g(x)) d\mu(x) && (14) \\ &\quad + \int_{\partial\Omega} (\phi f g + g \nabla f - f \nabla g) \cdot dS && (15) \end{aligned}$$

Then for the boundary conditions $g(x) = 0$ *and* $\nabla g(x) = 0$ *for all* $x \in \partial\Omega$, *and defining*

$$D^* g = -\nabla \cdot (g\phi)(x) + \Delta g(x)$$

we have the integral terms over the boundary going to zero and,

$$\langle Df, g \rangle_{L^2(\Omega,\mathbb{R},\mu)} = \langle f, D^* g \rangle_{L^2(\Omega,\mathbb{R},\mu)} \quad (16)$$

for all $f \in H^k(\Omega, \mathbb{R}, \mu)$ *and* $g \in \text{dom}(D^*) = \{g \in H^2(\mathbb{R}^n, \mathbb{R}, \mu) : \forall x \in \partial\Omega,\ g(x) = 0, \nabla g(x) = 0\}$.

Example 3 shows an example of an unbounded operator where the domain $\mathrm{dom}(L^*)$ is a strict subset of $\mathcal{Z}$ unlike in the case of bounded, closed operators in Examples 1 and 2. Derivative operators with boundary conditions are common in numerical methods for control, signal processing and partial differential equation applications.

2.2 Subspace Valued Maps

The notion of subspace valued maps expands the class of regularizers that a generalized representer theorem can explain and was introduced in Argyriou and Dinuzzo (2014). Let $\mathcal{H}$ be a separable Hilbert space, $2^{\mathcal{H}}$ be the power set on $\mathcal{H}$ and $\mathcal{V}(\mathcal{H})$ be a set of all closed vector subspaces of $\mathcal{H}$. Also for any subsets $A, B \subseteq \mathcal{H}$, let $A + B$ denote the set $\{a + b : a \in A, b \in B\}$. A map $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ is then called a subspace valued map. For evaluation on any set $A \subseteq \mathcal{H}$, we denote $S(A)$ to mean, $S(A) = \cup_{x \in A} S(x)$. The union operation, thus, extends the map $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$, in general, to a set valued map $S : \mathcal{V}(\mathcal{H}) \to 2^{\mathcal{H}}$ (as the union of vector spaces is not necessarily a vector space). Below we present a few definitions of terms we will use in the context of subspace valued maps and show conditions under which the union leads to closed vector spaces.

Definition 5 *(Subspace valued map)*
Let $\mathcal{H}$ be a separable Hilbert space and $\mathcal{V}(\mathcal{H})$ be a set of all closed vector subspaces of $\mathcal{H}$. A map $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ is called ***subspace valued****.*

Definition 6 *(Union extension)*
Let $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ be a subspace valued map. Then the extension of $S : \mathcal{V}(\mathcal{H}) \to 2^{\mathcal{H}}$ given by an union operation $S(A) = \cup_{x \in A} S(x)$ is called the union extension of S.

Definition 7 *(Inclusive map)*
A subspace valued map $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ is called ***inclusive****, if, for all $x \in \mathcal{H}$, $x \in S(x)$*

Definition 8 *(Super additive map)*
A map $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ is called ***super additive*** *if its union extension $S : \mathcal{V}(\mathcal{H}) \to 2^{\mathcal{H}}$ is super-additive, i.e. for all vector subspaces $A, B \in \mathcal{V}(\mathcal{H})$,*

$$S(A) + S(B) \subseteq S(A + B)$$

Note the the above name is a misnomer since we do not require $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ to be super-additive, but only its union extension to be super-additive. The misnomer is used for the purposes of brevity.

Definition 9 *(Closed map)*
A map $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ is called closed if its union extension $S : \mathcal{V}(\mathcal{H}) \to 2^{\mathcal{H}}$ maps closed subspaces from $\mathcal{V}(\mathcal{H})$ to closed subsets in $2^{\mathcal{H}}$.

Definition 10 *(Orthogonal subspace)*
For any $A \subseteq \mathcal{H}$, we define $S(A)^{\perp} := \{b \in \mathcal{H} : \forall a \in S(A), \langle a, b \rangle_{\mathcal{H}} = 0\}$

The following shows a few examples of inclusive and super-additive subspace valued maps that are used for application examples in Section 4,

Example 4 *Subspace valued maps*

1. $S_{\mathbb{R}}(a) := \{\lambda a : \lambda \in \mathbb{R}\}$ *is a closed, inclusive, super additive subspace valued map. Inclusivity of* $S_{\mathbb{R}}$ *is straightforward to see since* $a = 1 \cdot a \in S(a) = \{\lambda \cdot a : \lambda \in \mathbb{R}\}$. *Further for any* $A, B \in \mathcal{V}(\mathcal{H})$, $S(A)+S(B) = \{\lambda_1 a+\lambda_2 b : \lambda_1, \lambda_2 \in \mathbb{R}, a \in A, b \in B\} = \{\lambda a : \lambda \in \mathbb{R}, a \in A+B\} = S(A+B)$. *Also for any closed subspace* $A \in \mathcal{V}(\mathcal{H})$, *the union extension is such that* $S_{\mathbb{R}}(A) = \cup_{a\in A} S_{\mathbb{R}}(a) = A$ *and thus maps closed subspaces to closed subspaces.*

2. *Let* $K = \{L_i : \mathcal{H} \to \mathcal{H} : i = 1, \ldots, n\}$ *be a finite set of linearly independent, closed and bounded linear operators with the identity operator* $I \in \text{span}(K)$. *Then* $S_{\mathcal{L}}(a) := \{\sum_{i=1}^{n} \lambda_i L_i a : \lambda_i \in \mathbb{R}\}$ *is a closed, inclusive and super additive subspace valued map. The fact that* $S_{\mathcal{L}}$ *is closed, can be seen by noting that for any closed subspace* A, *we have* $\mathcal{S}_{\mathcal{L}}(A) = \sum_{i=1}^{n} L_i A$. *Since* L_i *are closed linear operators, the sets* $L_i A$ *are closed and the sum of finitely many closed sets remains closed.* $S_{\mathcal{L}}$ *being inclusive follows from the fact that the identity operator* $Ia = a$ *belongs to* $\text{span}(K)$, *and thus* $a \in S_{\mathcal{L}}(a)$, *implying* $S_{\mathcal{L}}$ *is inclusive. Also for any closed vector subspaces* $A, B \in \mathcal{V}(\mathcal{H})$, $S(A)+S(B) = \{\sum_{i=1}^{\infty} \lambda_i L_i a+\lambda_i' L_i b : \lambda_i, \lambda_i' \in \mathbb{R}, a \in A, b \in B\} = \{\sum_{i=1}^{\infty} L_i(\lambda_i a+\lambda_i' b) : \lambda_i, \lambda_i' \in \mathbb{R}, a \in A, b \in B\} = \{\sum_{i=1}^{\infty} L_i a : a \in A+B\} = S(A+B)$.

3. *A special case of the above example is the case when* $\mathcal{H} = \mathbb{R}^n$ *and* $E = \{e_1, \ldots, e_n\}$ *is the standard orthonormal basis for* $\mathbb{R}^n$. *Then* $S_{proj}(a) := \{\sum_{i=1}^{n} \lambda_i \langle a, e_i\rangle_{\mathcal{H}} e_i : e_i \in E, \lambda_i \in \mathbb{R}\}$ *is an inclusive, super additive subspace valued map. The* S_{proj} *corresponds to* $S_{\mathcal{L}}$ *from the previous example, with* $L_i : \mathcal{H} \to \mathcal{H}$, *being a set of projections onto the orthonormal basis, given as* $L_i a = \langle a, e_i\rangle_{\mathcal{H}} e_i$

4. *A countable counterpart of the example above can be presented for the space of square summable sequences,* $\mathcal{H} = \ell^2(\mathbb{N}, \mathbb{R})$, *taking values in* $\mathbb{R}$ *and indexed by natural numbers* $\mathbb{N}$. *Let* $\{\delta_i \in \ell^2(\mathbb{N}, \mathbb{R}) : i \in \mathbb{N}\}$ *with* $\delta_i(j) = 1$ *if* $i = j$ *and* 0 *otherwise, be the orthonormal basis for* $\ell^2(\mathbb{N}, \mathbb{R})$. *Let* $f(i)$ *denote the* i^{th} *member of a sequence and let* $\langle f, g\rangle_{\mathcal{H}} = \sum_{i=1}^{\infty} f(i)g(i)$. *Then* $S_{proj}(f) = \left\{\sum_{i=1}^{\infty} \lambda(i) \frac{\langle f, \delta_i\rangle_{\mathcal{H}} \delta_i}{||f||_{\mathcal{H}}} : \lambda \in \ell^2(\mathbb{N}, \mathbb{R})\right\}$ *for* $||f||_{\mathcal{H}} \neq 0$ *and* $S_{proj}(f) = \{0\}$ *if* $||f||_{\mathcal{H}} = 0$, *is an inclusive, closed and super additive subspace valued map. The* S_{proj} *defined can be seen to be inclusive as for any* $f \in \ell^2(\mathbb{N}, \mathbb{R})$, *there exists a representation for* f *in terms of the orthonormal basis* $f = \sum_{i=1}^{\infty} a(i)\delta_i$ *for some coefficients sequence* $a \in \ell^2$. $S_{proj}(f) = \{\sum_{i=1}^{\infty} \lambda(i)\delta_i : \lambda \in \ell^2(\mathbb{N}, \mathbb{R}), \lambda(i) = 0$ *if* $a(i) = 0\}$ *and thus* $f = \sum_{i=1}^{n} a(i)\delta_i$ *belongs to* $S_{proj}(f)$. *Similarly for any* $f = \sum_{i=1} a(i)\delta_i$ *and* $g = \sum_{i=1} b(i)\delta_i$ *with* $a, b \in \ell^2(\mathbb{N}, \mathbb{R})$, *we have* $S_{proj}(f) + S_{proj}(g) = \{\sum_{i=1}^{\infty} \lambda(i)\delta_i : \lambda \in \ell^2(\mathbb{N}, \mathbb{R}), \lambda(i) = 0$ *if* $a(i) = 0\} + \{\sum_{i=1}^{\infty} \lambda(i)\delta_i : \lambda \in \ell^2(\mathbb{N}, \mathbb{R}), \lambda(i) = 0$ *if* $b(i) = 0\} = \{\sum_{i=1}^{\infty} \lambda(i)\delta_i : \lambda \in \ell^2(\mathbb{N}, \mathbb{R}), \lambda(i) = 0$ *if* $b(i) = a(i) = 0\} = S_{proj}(f+g)$. *Thus* $S_{proj}(A) + S_{proj}(B) = S_{proj}(A+B)$ *for all* $A, B \in \mathcal{V}(\mathcal{H})$ *and thus it is trivially super-additive. Also* S_{proj} *is closed as it maps any* $A \in \mathcal{V}(\mathcal{H})$, *to* $S_{proj}(A) = \{\sum_{i=1}^{\infty} \lambda(i)\delta_i : \lambda \in \ell^2(\mathbb{N}, \mathbb{R})$ *and* $\lambda(i) = 0$ *if* $a(i) = 0$ *for all* $a \in A\}$ *which is a closed vector subspace of* $\ell^2(\mathbb{N}, \mathbb{R})$

Examples 4-3 and 4-4 are used to construct representers for regularizers given by ℓ_1 norm.

Noting that the union extension of a subspace valued map $S : \mathcal{V}(\mathcal{H}) \to 2^{\mathcal{H}}$, in general, is not subspace valued, the following Lemma shows that a subspace valued union extension

$S : \mathcal{V}(\mathcal{H}) \to \mathcal{V}(\mathcal{H})$ to a subspace valued map $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ exists, if and only if, the union extension $S : \mathcal{V}(\mathcal{H}) \to 2^{\mathcal{H}}$ is super-additive.

Lemma 11 *(Extending $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ to $S : \mathcal{V}(\mathcal{H}) \to \mathcal{V}(\mathcal{H})$)*
Let $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ be a subspace valued map and its union extension $S : \mathcal{V}(\mathcal{H}) \to 2^{\mathcal{H}}$ be given by $S(A) = \cup_{x \in A} S(x)$. Then the extension maps into $\mathcal{V}(\mathcal{H})$, if and only if, S is super-additive and closed.

Proof *We first prove that if S is super-additive and closed then for any $A \in \mathcal{V}(\mathcal{H})$, $S(A) \in \mathcal{V}(\mathcal{H})$ and thus the extension $S : \mathcal{V}(\mathcal{H}) \to 2^{\mathcal{H}}$ has range in $\mathcal{V}(\mathcal{H})$.*

To show $S(A) \in \mathcal{V}(\mathcal{H})$, we need to show that for any $a, b \in S(A)$, $\lambda a + \mu b \in S(A)$ for all $\lambda, \mu \in \mathbb{R}$ and that any converging sequence $\{a_n \in S(A)\}$ converges within $S(A)$.

First, we show that $S(A)$ is a vector space if S is super-additive.

Since $S(A) = \cup_{x \in A} S(x)$, for any $a \in S(A)$, there exists a $x_a \in A$ such that $a \in S(x_a)$. Further, $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ maps $x_a \in \mathcal{H}$ to a closed vector space $S(x_a) \in \mathcal{V}(\mathcal{H})$. Thus $a \in S(x_a)$ implies $\lambda a \in S(x_a)$ for all $\lambda \in \mathbb{R}$, also implying $\lambda a \in S(A)$. By the same arguments, for all $\mu \in \mathbb{R}$, $b \in S(A)$, implies $\mu b \in S(A)$. Thus the one dimensional closed vector spaces $K_a = \{\lambda a : \lambda \in \mathbb{R}\}$ and $K_b = \{\mu b : \mu \in \mathbb{R}\}$ are subspaces in $S(A)$ i.e., $K_a \subseteq S(A)$ and $K_b \subseteq S(A)$. Thus $K_a + K_b \subseteq S(A) + S(A)$. By super-additive property of S, $S(A) + S(A) \subseteq S(A + A) = S(A)$ (because for vector space A, $A + A = A$). Also, $\lambda a + \mu b \in K_a + K_b \subseteq S(A)$, implying for all $a, b \in S(A)$, $\lambda, \mu \in \mathbb{R}$, $\lambda a + \mu b \in S(A)$.

$S(A)$ is also closed, as S is taken to be a closed subspace valued map. Thus we have shown that S being super-additive and closed implies for all $A \in \mathcal{V}(\mathcal{H})$, $S(A) \in \mathcal{V}(\mathcal{H})$. Thus the union extension can be written as $S : \mathcal{V}(\mathcal{H}) \to \mathcal{V}(\mathcal{H})$.

Next we show the reverse statement that a union extension $S : \mathcal{V}(\mathcal{H}) \to \mathcal{V}(\mathcal{H})$ implies S is super-additive and closed.

For all $A, B \in \mathcal{V}(\mathcal{H})$, we have $A + B \in \mathcal{V}(\mathcal{H})$, as the sum of two closed vector spaces is a closed vector space. Also $A \subseteq A + B$ and $B \subseteq A + B$. Thus $S(A) = \cup_{x \in A} S(x) \subseteq \cup_{x \in A+B} S(x) = S(A + B)$. Similarly, $S(B) \subseteq S(A + B)$. Given S maps $\mathcal{V}(\mathcal{H})$ into $\mathcal{V}(\mathcal{H})$, we have for $A, B, A + B \in \mathcal{V}(\mathcal{H})$, $S(A), S(B), S(A + B) \in \mathcal{V}(\mathcal{H})$. Since $S(A) \subseteq S(A + B)$ and $S(B) \subseteq S(A + B)$, $S(A) + S(B) \subseteq S(A + B)$ implying S is super-additive. S being closed follows from the assumption that $S(A)$ was in $\mathcal{V}(\mathcal{H})$ which is a space of closed vector spaces. ■

The notions of quasilinear and idempotent maps from Argyriou and Dinuzzo (2014) are related to the notion of super additivity by noting that for any quasilinear, idempotent S, $S_{sup}(A) := \sum_{w \in A} S(w)$ can be defined as the corresponding super additive map. Also the representers from Argyriou and Dinuzzo (2014) are of the form $\sum_{i=1}^{m} S(w_i)$ and thus equivalently can be written as $S_{sup}(\text{span}(\{w_1, \ldots, w_m\}))$. Thus considering a super-additive subspace valued map does not lead to any loss of generality. Furthermore Argyriou and Dinuzzo (2014) assumed the maps to be idempotent, i.e., $S(S(x)) = S(x)$, which implicitly assumes that S has a subspace valued union extension and thus all idempotent subspace valued maps are implicitly required to be super-additive.

Another property that is of interest for us is the preservation of $\mathcal{N}_L^{\perp} = \text{range}(L^*)$ for a given operator $L : \mathcal{H} \to \mathcal{Z}$ under a subspace valued map, i.e., we want $\text{range}(L^*) \subseteq$

$S(\text{range}(L^*))$. S being inclusive is a sufficient condition for such a range preserving property. Formally we define this property as follows,

Definition 12 *(Range preserving map)*
*Let $L : \mathcal{H} \to \mathcal{Z}$ be a closable, densely defined operator as considered in Section 2.1 and let $\mathcal{N}_L^\perp = \text{range}\,(L^*)$ be the null space orthogonal of L. Then a subspace valued map $S : \mathcal{V}(\mathcal{H}) \to \mathcal{V}(\mathcal{H})$ is called **range preserving** with respect to L if*

$$\mathcal{N}_L^\perp \subseteq S(\mathcal{N}_L^\perp)$$

or equivalently, $S(\mathcal{N}_L^\perp)^\perp \subseteq \mathcal{N}_L$.

Given that $\mathcal{N}_L^\perp$ and $\mathcal{N}_L$ are closed, orthogonal complementary subspaces in $\mathcal{H}$, the subspace valued extension $S : \mathcal{V}(\mathcal{H}) \to \mathcal{V}(\mathcal{H})$ implies $S(\mathcal{N}_L^\perp)$ and $S(\mathcal{N}_L^\perp)^\perp$ are also closed, orthogonal complementary spaces in $\mathcal{H}$.

The range preserving property $S(\mathcal{N}_L^\perp)^\perp \subseteq \mathcal{N}_L$ implies that any $g \in S(\mathcal{N}_L^\perp)^\perp$, $g \in \mathcal{N}_L$, i.e., $Lg = 0$. This property will be useful later when proving the generalized theorem.

Lemma 13 *(Inclusive implies range preserving)*
If $S : \mathcal{V}(\mathcal{H}) \to \mathcal{V}(\mathcal{H})$ is inclusive then it is range preserving with respect to any closable, densely defined operator $L : \mathcal{H} \to \mathcal{Z}$.
Proof *If S is inclusive, then for all $A \in \mathcal{V}(\mathcal{H})$, $A \subseteq S(A)$. For a closable, densely defined operator, the orthogonal to the null space $\mathcal{N}_L^\perp = \text{range}\,(L^*)$ is a closed vector subspace in $\mathcal{V}(\mathcal{H})$ and thus inclusivity implies $\mathcal{N}_L^\perp = \text{range}\,(L^*) \subseteq S(\mathcal{N}_L^\perp)$.* ■

The range preserving property and orthogonal complementary nature of $S(\mathcal{N}_L^\perp)$ and $S(\mathcal{N}_L^\perp)^\perp$ will be key in characterizing the conditions for the existence of a representer theorem.

2.3 Orthomonotone Functionals

Dinuzzo and Schölkopf (2012); Argyriou and Dinuzzo (2014) introduced orthomonotone functionals as a way to expand the class of regularizers. The following reiterates the notions introduced there in the context of subspace valued maps of the form $S : \mathcal{V}(\mathcal{Z}) \to \mathcal{V}(\mathcal{Z})$ and separates out the notions of orthomonotonicity with respect to a single closed subspace (which gives a sufficient condition for the existence of a representer) and orthomonotonicity with respect to a subspace valued map, which gives as a necessary and sufficient condition when considering existence of representers for a family of minimization problems.

Definition 14 *(Orthomonotonicity with respect to a subspace)*
*Let $\mathcal{Z}$ be a Hilbert space and $\mathcal{K} \subseteq \mathcal{Z}$ be a closed subspace of $\mathcal{Z}$. Let $\mathcal{K}^\perp$ denote the orthogonal complementary space to $\mathcal{K}$. A functional $\Omega : \mathcal{Z} \to \mathbb{R} \cup \{+\infty\}$ is called **orthomonotone** with respect to the subspace $\mathcal{K}$ if*

$$\forall f \in \mathcal{K}, g \in \mathcal{K}^\perp, \qquad \Omega(f + g) \geq \Omega(f)$$

Definition 15 *(Orthomonotonicity with respect to a subspace valued map)*
*Let $\mathcal{Z}$ be a Hilbert space. A functional $\Omega : \mathcal{Z} \to \mathbb{R} \cup \{+\infty\}$ is called **orthomonotone** with respect to a subspace valued map $S : \mathcal{V}(\mathcal{Z}) \to \mathcal{V}(\mathcal{Z})$ if*

$$\forall A \in \mathcal{V}(\mathcal{Z}), f \in S(A), g \in S(A)^\perp, \qquad \Omega(f + g) \geq \max\{\Omega(f), \Omega(g)\}$$

Consider the subspace valued map $S_{\mathbb{R}}$ from Example 4. (Dinuzzo and Schölkopf, 2012, Theorem 1) showed that a functional Ω is orthomonotone with respect to $S_{\mathbb{R}}$ if and only if there exists a monotonically increasing functional $h : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ such that $\Omega(z) = h(||z||), \forall z \in \mathcal{Z}$. Note that while the above characterization with a monotonically increasing functional restricts its analysis to inner product induced norms, other kinds of orthomonotone functionals can be constructed as well, and orthomonotonicity with respect to a subspace valued map $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ was introduced in Argyriou and Dinuzzo (2014) as a means to expands the class of regularizers to non inner product terms. Example 5 belows shows a few examples of orthomonotone regularizers.

Example 5 *Orthomonotone functionals*

1. $\Omega(z) = ||z||^p_{\mathcal{Z}}$, *for any* $p > 0$ *is orthomonotone w.r.t.* $S_{\mathbb{R}}$
2. *Let* $\mathcal{Z} = \mathbb{R}^n$ *and* $||\cdot||_1$ *denote the* ℓ_1 *norm. Then,* $\Omega(z) = ||z||_1$ *is orthomonotone w.r.t.* S_{proj} *(*S_{proj} *as defined in Example 4-3).*

The proof for the first statement follows directly from (Dinuzzo and Schölkopf, 2012, Theorem 1) since $\Omega(z) = ||z||^p_{\mathcal{Z}}$, for any $p > 0$ is a monotonically increasing function of the inner product induced norm. The proof for the second statement follows from Theorem 16.

The second statement in the example above shows how sparse regularization problems involving the ℓ_1 norm are also covered by the notion of orthomonotone functionals.

The orthomonotonicity of ℓ_1 regularizers is formalized with the following theorem,

Theorem 16 *Orthomonotonicity of* ℓ_1 *regularizers*
Let $\mathcal{Z} = \mathbb{R}^n$, S_{proj} *be the subspace valued map defined in Example 4 and let* $h : [0, \infty] \to \mathbb{R} \cup \{+\infty\}$ *be a monotonic increasing function. Then* $\Omega(z) = h(||z||_1)$ *is orthomonotone with respect to* S_{proj}.
Proof *We first show* $\Omega(z) = ||z||_1$ *is orthomonotone w.r.t.* S_{proj}. *The result for monotonic increasing* h *follows from there.*

Let $E = \{e_1, \ldots, e_n\}$ *be the standard basis for* $\mathbb{R}^n$. *Note that for any* $z \in \mathbb{R}^n$, $S_{proj}(z) = \{\sum_{i=1}^n \lambda_i \langle z, e_i\rangle_{\mathbb{R}^n} e_i : e_i \in E, \lambda_i \in \mathbb{R}\}$ *and* $(S_{proj}(z))^\perp = \{\sum_j \lambda_j e_j : \langle z, e_j\rangle_{\mathbb{R}^n} = 0, e_j \in E, \lambda_j \in \mathbb{R}\}$. *Similarly for a set* $A \subset \mathbb{R}^n$, $S_{proj}(A) = \{\sum_{i=1}^n \lambda_i \langle z, e_i\rangle_{\mathbb{R}^n} e_i : e_i \in E, \lambda_i \in \mathbb{R}, z \in A\}$ *and* $(S_{proj}(A))^\perp = \{\sum_j \lambda_j e_j : e_j \in E, \lambda_j \in \mathbb{R}, \forall z \in A, \langle z, e_j\rangle_{\mathbb{R}^n} = 0\}$. *Now for any* $z \in S_{proj}(A)$ *and* $c \in S_{proj}(A)^\perp$, $||z + c||_1 = \sum_{\{i:\langle z,e_i\rangle_{\mathbb{R}^n} \neq 0\}} |z_i| + \sum_{\{i:\langle z,e_i\rangle_{\mathbb{R}^n} = 0\}} |c_i|$ *with* $z_i = \langle z, e_i\rangle_{\mathbb{R}^n}$ *and* $c_i = \langle c, e_i\rangle_{\mathbb{R}^n}$. *Also* $||z||_1 = \sum_{i=1}^n |z_i| = \sum_{\{i:\langle z,e_i\rangle_{\mathbb{R}^n} \neq 0\}} |z_i|$ *and* $||c||_1 = \sum_{i=1}^n |c_i| = \sum_{\{i:\langle z,e_i\rangle_{\mathbb{R}^n} = 0\}} |c_i|$. *Thus we see* $||z + c||_1 = ||z||_1 + ||c||_1 \geq \max\{||z||_1, ||c||_1\}$ $\implies$ $\Omega(z) = ||z||_1$ *is orthomonotone with respect to* S_{proj}.

For any monotonically increasing function h, *for any* $a, b \in [0, \infty)$, $a > b$ *implies* $h(a) > h(b)$. *Thus* $||z + c||_1 \geq \max\{||z||_1, ||c||_1\}$ *implies* $h(||z + c||_1) \geq \max\{h(||z||_1), h(||c||_1)\}$. *And thus* $\Omega(z) = h(||z||_1)$ *is orthomonotone with respect to* S_{proj} *for any monotonically increasing function* h. ■

The theorem can also be extended to a countable space of sequences as follows,

Theorem 17 *(Orthomonotonicity of ℓ_1 regularizers in countable spaces)*
Let $\mathcal{Z} = \ell^2(\mathbb{N})$ be the Hilbert space of $\mathbb{R}-$valued square summable sequences on $\mathbb{N}$. Let $||f||_1 = \begin{cases} \sum_{i=1}^{\infty} |f_i| & \text{if summation is bounded} \\ +\infty & \text{otherwise} \end{cases}$. Let S_{proj} be the subspace valued map considered in Example 4-4 and $h : [0, \infty] \to \mathbb{R} \cup \{\infty\}$ be a monotonic increasing function. Then $\Omega(f) = h(||f||_1)$ is orthomonotone with respect to S_{proj}.
Proof *For any $A \in \mathcal{V}(\mathcal{Z})$, $f \in S_{proj}(A)$, $g \in S_{proj}(A)^{\perp}$, we have $f = \sum_{i \in K_A} \lambda_i \delta_i$ and $g = \sum_{j \in \mathbb{N} \setminus K_A} \lambda_j \delta_j$, for δ_i being the orthonormal basis of $\ell^2(\mathbb{N})$ considered in Example 4-4 and K_A being some subset of indices in $\mathbb{N}$ for which A has a non-zero projection on δ_i, written as $K_A = \{i \in \mathbb{N} :$ there exists some $a \in A$ such that $\langle a, \delta_i \rangle_{\ell^2} \neq 0\}$. Thus we have $||f+g||_1 = ||f||_1 + ||g||_1$ (including the case when any of them takes the value of ∞) as both f and g have disjoint supports. Thus we have $||f+g||_1 \geq \max\{||f||_1, ||g||_1\}$ for all $A \in \mathcal{V}(\mathcal{Z})$ and $f \in S_{proj}(A)$, $g \in S_{proj}(A)^{\perp}$. Then for any monotonically increasing function h, we have $h(||f+g||_1) \geq \max\{h(||f||_1), h(||g||_1)\}$ and thus Ω is orthomonotone with respect to S_{proj}.* ■

For more properties of orthomonotone functional regarding compositions and sums we refer the reader to Argyriou and Dinuzzo (2014). With the notions of linear and adjoint operators combined with subspace valued maps and orthomonotone functionals, we are now ready to present the main result for the generalized representer theorem.

3. Generalized representer theorem

Let $\mathcal{H}$ and $\mathcal{Z}$ be separable Hilbert spaces. Let $L : \mathcal{H} \to \mathcal{Z}$ be closed, densely defined operators on $\mathcal{H}$. Let $C : \mathcal{Z} \to \mathbb{R} \cup \{+\infty\}$ and $\Omega : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be some lower semi-continuous functionals.

Functionals of the form $C' : \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_m \to \mathbb{R} \cup \{\infty\} := C'(L_1 f, \ldots, L_m f)$ are written without loss of generality in terms of a Hilbert space $\mathcal{Z}$ considered above, as follows. For any $m \in \mathbb{N}$ and $i \in \{1, \ldots, m\}$, let $L_i : \mathcal{H} \to \mathcal{Z}_i$ be closed, densely defined linear operators from $\mathcal{H}$ to separable Hilbert spaces $\mathcal{Z}_i$. Let $\mathcal{Z} = \mathcal{Z}_1 \times \mathcal{Z}_2 \times \cdots \times \mathcal{Z}_m$ and let $L : \mathcal{H} \to \mathcal{Z}$ be given by $Lf = (L_1 f, \ldots, L_m f)$, thus equivalently writing C' as a functional $C : \mathcal{Z} \to \mathbb{R} \cup \{+\infty\}$.

Now, consider the optimization problem,

$$f_{opt} = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \quad C(Lf) + \Omega(f) \tag{17}$$

The inclusion of $\{+\infty\}$ in the range of lower semi-continuous C and Ω allows one to consider constrained optimization problems. A few examples of learning problems written in this form are shown below,

Example 6 *(Learning and control problems)*

1. *Let $\mathcal{H}$ be an RKHS space of functions taking values in $\mathcal{Z}_i = \mathbb{R}^n$. Consider the evaluation operator from Example 1 such that $L_x : \mathcal{H} \to \mathcal{Z}_i$ is given by $L_x f := f(x)$. Let $\{(x_i, y_i) : i = 1, \ldots, m\}$ be a training data set. Let $L_1, \ldots, L_m$ be given by $L_{x_1}, \ldots, L_{x_m}$ and $L' : \mathcal{H} \to \mathcal{H}$ be the identity operator. Let $C(L_1 f, \ldots, L_m f) :=$*

$\sum_{i=1}^{m} ||y_i - \sigma(L_{x_i} f)||_{\mathcal{Z}}^2$ *for some activation function* $\sigma : \mathbb{R}^n \to \mathbb{R}^n$. *Let* $\Omega(L'f) := ||f||_{\mathcal{H}}^2$. *Then for* $J(f) = \sum_{i=1}^{m} ||y_i - \sigma(L_{x_i} f)||_{\mathcal{Z}_i}^2 + ||f||_{\mathcal{H}}^2$ *we get a regularized least squares problem in the RKHS space if* σ *is linear and an RKHS based neural network layer for some nonlinear* σ.

2. *Let* $\Omega(f) = ||f||_1^2$ *in the above example and we get a* ℓ_1 *regularized problem.*

3. *Let* $\mathcal{Z}_i = \mathbb{R}$, $y_i \in \{+1, -1\}$, $C(L_1 f, \ldots, L_m f) := \begin{cases} 0 & \forall i \in \{1, \ldots, m\}; \quad y_i L_i f > 0 \\ +\infty & \text{otherwise} \end{cases}$ *and* $\Omega(f) = ||f||^2$. *Then* $J(f) = C(L_1 f, \ldots, L_m f) + \Omega(f)$ *gives the hard margin support vector machine objective for binary classification.*

4. *Let* μ *be a positive measure on the measurable space* $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. *Let* f, u *be functions in* $L^2([0, \infty), \mu; \mathbb{R}^n)$ *and* $L^2([0, \infty), \mu; \mathbb{R}^m)$ *respectively. Consider the regularizer* $\Omega(f, u) = ||f||_{L^2}^2 + ||u||_{L^2}^2$ *and* $C(L(f, u)) = \begin{cases} 0 & \text{if } \substack{\partial_t f(t_i) - \phi(f(t_i), u(t_i)) = 0 \text{ for all } i = 1, \ldots, m \\ f(0) = x_0, u(0) = u_0} \\ +\infty & \text{otherwise} \end{cases}$ *for some known nonlinear function* $\phi : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$, *finite set of points* $\{t_i \in [0, \infty) : i = 1, \ldots, m\}$ *and* $(x_0, u_0) \in \mathbb{R}^n \times \mathbb{R}^m$. *Then* $J((f, u)) = C(L(f, u)) + \Omega(f, u)$ *gives the objective function for solving a collocation based approximation to a continuous time nonlinear optimal control problem, where* ϕ *is a known function for the dynamics of the system,* f *denotes the continuous time trajectory and* u *denotes the continuous time control signal. The measure* μ *is used as a weighting measure to determine the growth rate of the functions considered in the hypothesis space for the solutions. Note also that the derivative operator* ∂_t *is only a closed, densely defined operator and not a bounded one.*

Given a learning problem in the form of (17), let Ω be orthomonotone with respect to an inclusive, super-additive subspace valued map $S : \mathcal{V}(\mathcal{H}) \to \mathcal{V}(\mathcal{H})$. The generalized representer theorem states that a minimizer for (17) exists in the subspace given by $S(\mathcal{N}_L^{\perp})$ and the problem (17) is said to be linearly representable.

The notion of linear representability is significant as it often allows one to reformulate infinite dimensional optimization problems in $\mathcal{H}$ into equivalent finite dimensional optimization in $\mathcal{Z}$ given as

$$f_{opt} = \operatorname{argmin}_{f \in S(\operatorname{range}(L^*))} \quad C(Lf) + \Omega(f) \tag{18}$$

(18) gives a finite dimensional optimization if $\mathcal{Z}$ is finite dimensional and $S(\operatorname{range}(L^*))$ is a finite dimensional subspace.

Below we state and prove, first the sufficient condition for linear representability of a functional $J(f) = C(Lf) + \Omega(f)$ and then the complete statement of necessary and sufficient condition for linear representability over a given family of functionals.

3.1 Sufficient conditions for linear representability

Theorem 18 *Generalized Representer Theorem (Sufficient condition)*
Let $\mathcal{H}$ *and* $\mathcal{Z}$ *be separable Hilbert spaces and* $L : \mathcal{H} \to \mathcal{Z}$ *be a closed, densely defined linear operator with the null space orthogonal* $\mathcal{N}_L^{\perp} = \operatorname{range}(L^*)$. *Let* $S : \mathcal{V}(\mathcal{H}) \to \mathcal{V}(\mathcal{H})$ *be*

a closed and super additive subspace valued map, range preserving with respect to L. Let $\Omega : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ and $C : \mathcal{Z} \to \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous functionals, with Ω orthomonotone with respect to the subspace $S(\mathcal{N}_L^\perp)$. Then for the problem,

$$f_{opt} = \underset{f \in \mathcal{H}}{argmin} \quad C(Lf) + \Omega(f) \tag{19}$$

if the minimizers are attainable, atleast one minimizer is linearly representable with respect to S, such that $f_{opt} \in S(\mathcal{N}_L^\perp)$.

Proof *Since Ω is orthomonotone with respect to the closed subspace $S(\mathcal{N}_L^\perp)$, $\forall f \in S(\mathcal{N}_L^\perp), g \in S(\mathcal{N}_L^\perp)^\perp$, $\Omega(f+g) \geq \Omega(f)$. Also by Definition 12, S is range preserving with respect to L, implies $S(\mathcal{N}_L^\perp)^\perp \subseteq \mathcal{N}_L$. Thus for all $g \in S(\mathcal{N}_L^\perp)^\perp$, $Lg = 0$.*

For a closed, densely defined operator L, $\mathcal{N}_L^\perp = \text{range}(L^*)$ is a closed vector subspace and thus by definition is mapped to a closed subspace $S(\mathcal{N}_L^\perp)$ by the subspace valued map. Thus $S(\mathcal{N}_L^\perp)$ and $S(\mathcal{N}_L^\perp)^\perp$ form an orthogonal complementary pair for $\mathcal{H}$ and for any $F \in \mathcal{H}$ we can find a decomposition $F = f + g$, with $f \in S(\mathcal{N}_L^\perp)$, $g \in S(\mathcal{N}_L^\perp)^\perp$. Then

$$\begin{aligned} J(F) &= C(L(f+g)) + \Omega(f+g) && (20)\\ &= C(Lf) + \Omega(f+g) && (21)\\ &\geq C(Lf) + \Omega(f) && (22) \end{aligned}$$

Thus $\forall F \in \mathcal{H}$, $\exists f \in S(\mathcal{N}_L^\perp)$ such that $J(f) \leq J(F)$. Thus if J admits a minimizer in $\mathcal{H}$, a minimizer must exists in $S(\mathcal{N}_L^\perp)$, implying J is linearly representable w.r.t. S. ■

3.2 Necessary and sufficient conditions for linear representability

The generalized representer theorem we present here differs from its previous counterpart (Argyriou and Dinuzzo, 2014, Theorem 3.1) in three significant ways. Firstly, there is no assumption for a finite dimensional r-regularity property on the subspace valued map and secondly, the loss functional C can be defined on arbitrary infinite dimensional Hilbert spaces $\mathcal{Z}$. These two changes become significant since when dealing with stochastic regression problems the output space $\mathcal{Z}$ is an infinite dimensional Hilbert space of random variables (or measurable functions) and when dealing with ℓ_1 regularization problems in function spaces, the corresponding subspace valued map S_{proj} is not r-regular for any finite r. We will expand upon these differences in Section 4 with corresponding application examples. Lastly, we consider closed and densely defined operators in the loss function which allows for unbounded, derivative like operators in learning and control problems.

Now note that problems of the form (17) are typically considered over families of linear operators $L : \mathcal{H} \to \mathcal{Z}$ where L depends on training data for the learning problem and scaled regularizers $\{\gamma\Omega : \gamma \in (0, \infty)\}$, and if (17) is linearly representable for some choice of L and γ, it is natural to expect the problem to be linearly representable for all possible problems in this family. In fact if Ω is orthomonotone with respect to a closed, inclusive and super-additive subspace valued map S, this follows from Theorem 18 for all closed, densely defined linear operators (since an inclusive S is null space preserving for any operator L, by Lemma 13). The necessary condition in the representer theorem considers the reverse

proposition, that is, if (17) is linearly representable with respect to a closed, inclusive and super-additive subspace valued map S for all closed, densely defined operators L and all $\gamma \in (0, \infty)$, then under certain additional assumptions on C and Ω it can be concluded that Ω must be orthomonotone with respect to S.

Thus, consider the family of functionals, given a closed, inclusive and super-additive subspace valued map S,

$$\mathcal{J}_S = \{C \circ L + \gamma\Omega \mid \gamma \in (0, \infty), L : \mathcal{H} \to \mathcal{Z} \text{ is closed, densely defined}\} \tag{23}$$

and for fixed functionals $C : \mathcal{Z} \to \mathbb{R} \cup \{\infty\}$ and $\Omega : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$ such that C admits a unique non-zero minimizer $z^\star$ in $\mathcal{Z}\backslash\{0\}$ with compact sub-level sets in its neighborhood and Ω admits a minimizer at 0. Note that the assumption on Ω is not a new one. Any Ω orthomonotone with respect to a subspace valued map must admit a minimizer at 0 and thus was not explicitly stated in Theorem 18. For the reverse proposition of the representer theorem, however Ω is not assumed to be orthomonotone and thus for it to be orthomonotone by the reverse proposition, a minimizer at 0 must be assumed (the minimizer at 0 need not be a unique minimizer).

The necessary and sufficient conditions for the generalized representer theorem can then be stated as follows,

Theorem 19 *Generalized Representer Theorem (Necessary and Sufficient Conditions) Let $\mathcal{H}$ and $\mathcal{Z}$ be separable Hilbert spaces. Let $S : \mathcal{V}(\mathcal{H}) \to \mathcal{V}(\mathcal{H})$ be a closed, inclusive and super additive subspace valued map. Let $\Omega : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$ and $C : \mathcal{Z} \to \mathbb{R} \cup \{+\infty\}$ be lower semicontinuous functionals, such that Ω admits a minimizer at 0 and C admits a unique minimizer $z^\star$ in $\mathcal{Z}\backslash\{0\}$ with sequentially compact sub-level sets around $z^\star$. Let $\mathcal{J}_S = \{J_{L,\gamma} = C \circ L + \gamma\Omega \mid \gamma \in (0, \infty), L : \mathcal{H} \to \mathcal{Z}$ is a closed, densely defined linear operator$\}$ be the family of functionals corresponding to all closed, densely defined linear operators $L : \mathcal{H} \to \mathcal{Z}$ and constants $\gamma \in (0, \infty)$. For each functional in $J_{L,\gamma} \in \mathcal{J}_S$ consider the problem,*

$$f_{opt} = \underset{f \in \mathcal{H}}{argmin} \quad C(Lf) + \Omega(f) \tag{24}$$

Then, each problem in the family $\{\min_{f\in\mathcal{H}} J(f) : J \in \mathcal{J}_S\}$ is linearly representable with respect to S if and only if, Ω is orthomonotone with respect to S

Proof *The proof for sufficiency (i.e. orthomonotone $\Omega \implies$ existence of linear representer) follows from Theorem 18 and Lemma 13.*

To prove necessity of orthomonotone Ω, assume that all functionals $J_{L,\gamma} \in \mathcal{J}_S$ corresponding to a linear operator L and constant γ are linear representable w.r.t. to S, i.e., for all functionals $J_{L,\gamma} = C \circ L + \gamma\Omega \in \mathcal{J}_S$ a minimizer exists in $S(\mathcal{N}_L^\perp)$. Note that a minimizer $J_{L,\gamma}$ exists because both C and Ω admit minimizers in $\mathcal{Z}\backslash\{0\}$ and $\mathcal{H}$ respectively and $\text{range}(L)$ *is a closed subset in $\mathcal{Z}$.*

We first show that for all closed densely defined operators $L : \mathcal{H} \to \mathcal{Z}$ we must have $\Omega(f+g) \geq \max\{\Omega(f), \Omega(g)\}$ for all $f \in S(\mathcal{N}_L^\perp)$ and $g \in S(\mathcal{N}_L^\perp)^\perp$ for a family of functionals $\{J_{L,\gamma} \in \mathcal{J}_S : \gamma \in (0, \infty)\}$ to be linearly representable with respect to S. We show this in two parts, first we show $\Omega(f + g) \geq \Omega(f)$ and then $\Omega(f + g) \geq \Omega(g)$ for $f \in S(\mathcal{N}_L^\perp)$ and $g \in S(\mathcal{N}_L^\perp)^\perp$.

Finally we show that there exists a one to one correspondence between the space of all closed vector subspaces $A \in \mathcal{V}(\mathcal{H})$ and a set of closed and bounded linear operators (which is a subset of closed, densely defined linear operators) and thus for all $A \in \mathcal{V}(\mathcal{H})$, we must have a closed, bounded operator $L : \mathcal{H} \to \mathcal{Z}$ such that $A = \mathcal{N}_L^\perp$. Thus for all $A \in \mathcal{V}(\mathcal{H})$, $f \in S(A)$ and $g \in S(A)^\perp$ we have $\Omega(f+g) \geq \max\{\Omega(f), \Omega(g)\}$. Thus we show that if the family of functionals $\mathcal{J}_S = \{J_{L,\gamma} = C \circ L + \gamma\Omega \mid \gamma \in (0,\infty), L : \mathcal{H} \to \mathcal{Z}$ is a closed, densely defined linear operator$\}$ for a given tuple of functionals and subspace valued map (C, Ω, S) are all linearly representable with respect to S then Ω must be orthomonotone with respect to S.

We start by proving the result that $\Omega(f+g) \geq \max\{\Omega(f), \Omega(g)\}$ for all $f \in S(\mathcal{N}_L^\perp)$ and $g \in S(\mathcal{N}_L^\perp)^\perp$.

(i) Consider first the corner case for $f \in S(\mathcal{N}_L^\perp)$ such that $f = 0$, and $g \in S(\mathcal{N}_L^\perp)^\perp$. Then, we have $\Omega(f+g) = \Omega(g) \geq \Omega(g)$ (trivially true) and $\Omega(f+g) = \Omega(g) \geq \Omega(0) = \Omega(f)$ (since Ω admits a minimizer at 0 and $f = 0$). Thus for the case of $f \in S(\mathcal{N}_L^\perp)$, $f = 0$, we have shown that $\Omega(f+g) \geq \max\{\Omega(f), \Omega(g)\}$ for all $g \in S(\mathcal{N}_L^\perp)^\perp$.

(ii) Next, consider the corner case, where the operator $L : \mathcal{H} \to \mathcal{Z}$ is such that $S(\mathcal{N}_L^\perp) = \{0\}$, i.e. there exists no $f \neq 0$ in $S(\mathcal{N}_L^\perp)$. If $S(\mathcal{N}_L^\perp) = \{0\}$, then result (i) implies that for all $f \in S(\mathcal{N}_L^\perp)$ and $g \in S(\mathcal{N}_L^\perp)^\perp$, $\Omega(f+g) \geq \max\{\Omega(f), \Omega(g)\}$.

(iii) Now for the general case where $S(\mathcal{N}_L^\perp) \neq \{0\}$, there exist a $f \neq 0$ in $S(\mathcal{N}_L^\perp)$. Let $z^\star \neq 0 \in \mathcal{Z}$ denote the unique minimizer for functional C. From result (i) we already have the result that for $f = 0$, $\Omega(f+g) \geq \max\{\Omega(f), \Omega(g)\}$ for all $g \in S(\mathcal{N}_L^\perp)^\perp$. Thus, consider the case for $f \neq 0$. By Proposition 20 below, we have shown that for any closed, densely defined operator $L : \mathcal{H} \to \mathcal{Z}$ for which $S(\mathcal{N}_L^\perp) \neq \{0\}$, given a $f \neq 0 \in S(\mathcal{N}_L^\perp)$, we have a closed and bounded linear operator $L'_f : \mathcal{H} \to \mathcal{Z}$, such that $L'_f f = z^\star$ and for any $g \in S(\mathcal{N}_L^\perp)^\perp$, $L'_f g = 0$. Since L' is closed and bounded we have the functional $J_{L'_f,\gamma} = C \circ L'_f + \gamma\Omega$ in $\mathcal{J}_S$. Let $h^\star_{f,\gamma} \in S(\mathcal{N}_L^\perp)$ be a minimizer for $J_{L'_f,\gamma}$.

Next, note that since $z^\star$ is a minimizer for C, we have $C(z^\star) \leq C(L'_f h^\star_{f,\gamma})$ and thus

$$C(z^\star) + \gamma\Omega(h^\star_{f,\gamma}) \leq C(L'_f h^\star_{f,\gamma}) + \gamma\Omega(h^\star_{f,\gamma}) = J_{L'_f,\gamma}(h^\star_{f,\gamma}) \tag{25}$$

Also $h^\star_{f,\gamma}$ is the minimizer for $J_{L'_f,\gamma}$ and thus

$$J_{L'_f,\gamma}(h^\star_{f,\gamma}) = C(L'_f h^\star_{f,\gamma}) + \gamma\Omega(h^\star_{f,\gamma}) \leq C(L'_f(f+g)) + \gamma\Omega(f+g)$$

for all $g \in S(\mathcal{N}_L)^\perp$.

But from Proposition 20, we have $L'_f(g) = 0$ and $L'_f f = z^\star$, implying $C(L'_f(f+g)) = C(z^\star)$, giving

$$C(L'_f h^\star_{f,\gamma}) + \gamma\Omega(h^\star_{f,\gamma}) \leq C(z^\star) + \gamma\Omega(f+g) \tag{26}$$

Thus we have the inequality

$$C(z^\star) + \gamma\Omega(h^\star_{f,\gamma}) \leq J_{L'_f,\gamma}(h^\star_{f,\gamma}) \leq C(z^\star) + \gamma\Omega(f+g) \tag{27}$$

for all $\gamma \in (0,\infty)$ or equivalently,

$$\Omega(h^\star_{f,\gamma}) \leq \Omega(f+g) \tag{28}$$

for all $\gamma \in (0,\infty)$, $g \in S(\mathcal{N}_L^\perp)^\perp$ and all minimizers $h^\star_{f,\gamma}$.

Also, from (25), we have the inequality $C(L'_f h^\star_{f,\gamma}) - C(z^\star) \geq 0$ and from (26) we have $C(L'_f h^\star_{f,\gamma}) - C(z^\star) \leq \gamma(\Omega(f+g) - \Omega(h^\star_{f,\gamma}))$. Thus we have an inequality

$$0 \leq C(L'_f h^\star_{f,\gamma}) - C(z^\star) \leq \gamma(\Omega(f+g) - \Omega(h^\star_{f,\gamma})) \tag{29}$$

for all $\gamma \in (0,\infty)$, $g \in S(\mathcal{N}_L^\perp)^\perp$ and minimizers $h^\star_{f,\gamma}$. For the case where $\Omega(f+g) = \infty$, $\Omega(f+g) \geq \max\{\Omega(f), \Omega(g)\}$ is trivially satisfied. When $\Omega(f+g) < \infty$, so is $\Omega(h^\star_{f,\gamma})$ (by (29)). Thus for the case of $\Omega(f+g) < \infty$, we have $\Omega(f+g) - \Omega(h^\star_{f,\gamma}) < \infty$ and thus by (29),

$$\gamma \to 0 \implies C(L'_f h^\star_{f,\gamma}) \to C(z^\star)$$

Since the sub-level sets around $C(z^\star)$, $V_\epsilon = \{z \in \mathcal{Z} : C(z) \leq C(z^\star) + \epsilon\}$ are given to be sequentially compact, and $z^\star$ is the unique minimizer, this implies $L'_f h^\star_{f,\gamma} \to z^\star$ as $\gamma \to 0$. But $f, h^\star_{f,\gamma} \in S(\mathcal{N}_L^\perp)$ and thus by Proposition 20, we have strong convergence $h^\star_{f,\gamma} \to f$.

Thus from (28), under the limit $\gamma \to 0$, we have $\Omega(f) \leq \Omega(f+g)$ for all $g \in S(\mathcal{N}_L^\perp)^\perp$. Since the above argument holds for all $f \neq 0$, $f \in S(\mathcal{N}_L^\perp)$ and we have the result from (i) for $f = 0$, we have for all $f \in S(\mathcal{N}_L^\perp)$ and all $g \in S(\mathcal{N}_L^\perp)^\perp$, the result that

$$\Omega(f+g) \geq \Omega(f)$$

(iv) To show the remaining inequality $\Omega(f+g) \geq \Omega(g)$ for all $f \in S(\mathcal{N}_L^\perp)$ and $g \in S(\mathcal{N}_L^\perp)^\perp$, similar arguments to result (i), (ii) and (iii) are required and are presented in the following.

(iv-i) Firstly note that for $g = 0$, $g \in S(\mathcal{N}_L^\perp)^\perp$ and for any $f \in S(\mathcal{N}_L^\perp)$, we have $\Omega(f+g) = \Omega(f) \geq \Omega(f)$ (trivially true) and $\Omega(f+g) \geq \Omega(0) = \Omega(g)$ (because Ω admits a minimizer at 0 and $g = 0$). Thus for $g = 0$, we have $\Omega(f+g) \geq \max\{\Omega(f), \Omega(g)\}$ for all $f \in S(\mathcal{N}_L^\perp)$.

(iv-ii) Now consider the corner case, where $S(\mathcal{N}_L^\perp)^\perp = \{0\}$. In such a case, using result (iv-i), we have for all $f \in S(\mathcal{N}_L^\perp)$ and $g \in S(\mathcal{N}_L^\perp)^\perp$, $\Omega(f+g) \geq \Omega(g)$

(iv-iii) For the general case when $S(\mathcal{N}_L^\perp)^\perp \neq \{0\}$, there exists a $g \in S(\mathcal{N}_L^\perp)^\perp$ such that $g \neq 0$. For $g = 0$, we already have the required inequality from (iv-i). Thus we consider the case for $g \in S(\mathcal{N}_L^\perp)^\perp$ and $g \neq 0$. From Proposition 20, using the $A = S(\mathcal{N}_L^\perp)^\perp$, we have a closed and bounded operator $L'_g : \mathcal{H} \to \mathcal{Z}$ such that $L'_g g = z^\star$ and for all $f \in S(\mathcal{N}_L^\perp)$, $L'_g f = 0$. Thus we have the functional $J_{L'_g,\gamma} = C \circ L'_g + \gamma\Omega$ in $\mathcal{J}_S$. Let $h^\star_{g,\gamma}$ be a minimizer for $J_{L'_g,\gamma}$. Then following the same arguments as before from (iii), we have the analogous inequality

$$\Omega(h^\star_{g,\gamma}) \leq \Omega(f+g) \tag{30}$$

for all $f \in S(\mathcal{N}_L^\perp)$, $\gamma \in (0,\infty)$ and minimizers $h^\star_{g,\gamma}$.

As $\gamma \to 0$, we have as before, a sequence of minimizers $h^\star_{g,\gamma} \to g$ and thus in the limit, we have $\Omega(f+g) \geq \Omega(g)$ for all $f \in S(\mathcal{N}_L^\perp)$ and $g \in S(\mathcal{N}_L^\perp)^\perp$, $g \neq 0$. Combining with the result from (iv-i) for $g = 0$, we have for all $f \in S(\mathcal{N}_L^\perp)$ and $g \in S(\mathcal{N}_L^\perp)^\perp$,

$$\Omega(f+g) \geq \Omega(g)$$

(v) Thus from (iii) and (iv), we have shown that for all $f \in S(\mathcal{N}_L^\perp)$ and $g \in S(\mathcal{N}_L^\perp)^\perp$, we have,

$$\Omega(f+g) \geq \max\{\Omega(f), \Omega(g)\}$$

for all closed, densely defined operators $L : \mathcal{H} \to \mathcal{Z}$.

(vi) Finally, we show that there is a one to one correspondence between the set of closed vector spaces $A \in \mathcal{V}(\mathcal{H})$ and a set of closed, bounded linear operators $L : \mathcal{H} \to \mathcal{Z}$, such that for any $A \in \mathcal{V}(\mathcal{H})$ there exists a closed bounded operator L satisfying $A = \mathcal{N}_L^{\perp}$. Using this correspondence and the result from (v), we have the final result stating that for all closed vector subspaces $A \in \mathcal{V}(\mathcal{H})$, and for all $f \in S(A)$ and $g \in S(A)^{\perp}$,

$$\Omega(f+g) \geq \max\{\Omega(f), \Omega(g)\}$$

implying Ω is orthomonotone with respect to S.

To show the correspondence between A and L consider the following.

For any closed vector subspace $A \in \mathcal{V}(\mathcal{H})$, let $P_A : \mathcal{H} \to \mathcal{H}$ denote the orthogonal projection onto the closed vector subspace A. Since A is a closed vector subspace of $\mathcal{H}$, A and $A^{\perp}$ form an orthogonal complementary pair of subspaces such that $\text{range}(P_A) = A$ *and null space of P_A is $A^{\perp}$, and thus by the closed graph theorem, it follows that P_A is a closed operator. Since for any $F \in \mathcal{H}$, there exists an unique decomposition $F = f + g$ such that $f \in A$ and $g \in A^{\perp}$ and $P_A F = P_A(f+g) = f$, it also follows that $||P_A F||_{\mathcal{H}} = ||f||_{\mathcal{H}} \leq ||F||_{\mathcal{H}}$ and P_A is thus a closed, bounded linear operator. It is also easy to see that $\langle P_A F_1, F_2 \rangle_{\mathcal{H}} = \langle F_1, P_A F_2 \rangle_{\mathcal{H}}$ for all $F_1, F_2 \in \mathcal{H}$ and thus P_A is a self-adjoint, closed, bounded operator.*

Let $L : \mathcal{H} \to \mathcal{Z}$ be any closed, bounded linear operator with null space $\mathcal{N}_L = \{0\}$. Then the composition $L'_A = L \circ P_A$ is also closed and bounded, and $\mathcal{N}_{L'_A}^{\perp} =$ $\text{range}((L'_A)^*) =$ $\text{range}(P_A L^*) = A$. *Thus for every $A \in \mathcal{V}(\mathcal{H})$, we have a closed and bounded operator given by L'_A such that $\mathcal{N}_{L'_A}^{\perp} = A$. The result for orthomonotonicity of Ω then follows, as stated above.*

■

To prove the necessary part of Theorem 19, the following proposition is considered.

Proposition 20 *Let $\mathcal{H}$ and $\mathcal{Z}$ be separable Hilbert spaces. Let there exist a minimizer $z^\star \neq 0 \in \mathcal{Z}$ for the lower semicontinuous functional $C : \mathcal{Z} \to \mathbb{R} \cup \{\infty\}$. Let $A \in \mathcal{V}(\mathcal{H})$ be a closed vector subspace of $\mathcal{H}$. Let there exist a $f \in A$ such that $f \neq 0$. Let $\mathcal{H}$ be spanned by a orthonormal basis $\{f/||f||, \phi_1, \phi_2, \dots\}$, let $\mathbb{N}_A$ be a subset of $\mathbb{N}$ such that A is spanned by $\{f/||f||\} \cup \{\phi_k : k \in \mathbb{N}_A\}$ and $A^{\perp}$ is spanned by $\{\phi'_k : k' \in \mathbb{N} \backslash \mathbb{N}_A\}$. Then there exists a closed and bounded linear operator $L'_f : \mathcal{H} \to \mathcal{Z}$ given by*

$$L'_f h = z^\star \left\langle \sum_{k \in \mathbb{N}_A} \frac{\phi_k}{k^2} + \frac{f}{||f||^2}, h \right\rangle_{\mathcal{H}}$$

such that

1. *$L'_f g = 0$ for all $g \in A^{\perp}$*
2. *$L'_f f = z^\star$*
3. *$h \in S(A)$ and $L'_f h = z^\star$, implies $h = f$*

Proof *Firstly, note that the existence of a countable basis* $\{f/||f||, \phi_1, \phi_2, \cdots\}$ *is guaranteed by E. Schmidt's orthogonalization (Yoshida, 2013, Chapter III-5) for a separable Hilbert space. Since* A *and* $A^\perp$ *are orthogonal complementary subspaces and* $f \in A$*, they split the orthonormal basis into two disjoint countable subset as mentioned in the statement of the proposition given by the index set* $\mathbb{N}_A$*.*

To see that L'_f *is bounded, note that for any* $h \in \mathcal{H}$*,* $||L'_f h||_{\mathcal{Z}} = ||z^\star||_{\mathcal{Z}} |\langle \sum_{k\in\mathbb{N}_A} \phi_k/k^2 + f/||f||^2, h\rangle_{\mathcal{H}}|$*. Then note that* $|\langle \sum_{k\in\mathbb{N}_A} \phi_k/k^2 + f/||f||^2, h\rangle_{\mathcal{H}}| \leq \sum_{k\in\mathbb{N}_A} |\langle \phi_k/k^2, h\rangle_{\mathcal{H}}| + |\langle f/||f||^2, h\rangle_{\mathcal{H}}| \leq (\sum_{k\in\mathbb{N}_A} 1/k^2 + 1/||f||)||h||_{\mathcal{H}}$*. Now since* $\sum_{k\in\mathbb{N}_A} 1/k^2 \leq \sum_{k\in\mathbb{N}} 1/k^2 = \pi^2/6 < \infty$ *(since summation of a series* $1/k^2$ *over* $\mathbb{N}$ *is known to be bounded),* $0 < ||f||_{\mathcal{H}} < \infty$ *and* $||z^\star||_{\mathcal{Z}} < \infty$*, we have* $||L'_f h||_{\mathcal{Z}} \leq M||h||_{\mathcal{H}}$ *for some bounded constant* $M = ||z^\star||(\sum_{k\in\mathbb{N}_A} 1/k^2 + 1/||f||) < \infty$*.*

Also since the null space of L'_f *denoted* $ker(L'_f)$ *is* $A^\perp$*,* $\inf\{||L'_f h||_{\mathcal{Z}} : h \in ker(L'_f)^\perp, ||h||_{\mathcal{H}} = 1\} > 0$ *and thus* L'_f *is closed by (Conway, Proposition 6.5.5).*

Then for any $g \in S(A)^\perp$*, we have* $L'_f g = z^\star \langle \sum_{k\in\mathbb{N}_A} \phi_k/k^2 f/||f||^2, g\rangle_{\mathcal{H}} = 0$*, showing the first property stated for* L'_f*.*

The second statement $L'_f f = z^\star$*, follows by substituting* f *into the definition for* $L'_f f$*. Since* $\{f/||f||, \phi_1, \phi_2 \dots\}$ *are orthonormal basis,* f *is orthogonal to all* ϕ_k *and thus* $\langle \phi_k, f\rangle_{\mathcal{H}} = 0$*, which leaves the term* $z^\star \langle f/||f||^2, f\rangle_{\mathcal{H}} = z^\star$*.*

The last statement can be seen from the fact that $L'_f h = z^\star$ *implies* $L'_f h = L'_f f$ *or* $L'_f(h - f) = 0$*, i.e.,* $\langle \sum_{k\in\mathbb{N}_A} \phi_k + f, f - h\rangle_{\mathcal{H}} = 0$ *implying* $f - h \in S(A)^\perp$ *(since* ϕ_k *and* f *span* $S(A)$*). But both* f *and* h *are given to be in* $S(A)$ *and thus they must be in* $S(A) \cap S(A)^\perp = \{0\}$*. Thus* $h = f$*.*

■

3.3 Related work

We presented here a generalized version of representer theorems for problems of the form

$$f_{opt} = \text{argmin}_{f\in\mathcal{H}} \; C(Lf) + \Omega(f) \tag{31}$$

for a loss function $C : \mathcal{Z} \to \mathbb{R} \cup \{+\infty\}$ on a separable Hilbert space $\mathcal{Z}$ and closed, densely defined operator $L : \mathcal{H} \to \mathcal{Z}$ and Ω orthomonotone with respect to a subspace valued map S. The assumption of "r-regularity" on subspace valued maps from previous counterparts of the theorem was dropped to allow for more general regularization like the ℓ_1 norm on function spaces, $\mathcal{Z}$ was considered as separable Hilbert spaces to allow for loss functional on infinite dimensional Hilbert space, as occurring in examples from learning in Hilbert spaces of stochastic processes and the linear operators were considered to be closed and densely defined to allow for unbounded operators like the derivative operators that occur commonly in optimal control problems.

Special cases of the theorem addressing learning with bounded functionals like the least squares regularization for vector valued functions in Reproducing Kernel Hilbert Space (RKHS) framework can be found in (Micchelli and Pontil, 2005, Theorems 3.1, 4.1). Special cases of the theorem for ℓ_1 regularization can be found in Unser et al. (2016). A generalized version of the representer theorems for general loss functions but still restricted

to Hilbert spaces of real valued functions and bounded functionals can be found in Dinuzzo and Schölkopf (2012); Schölkopf et al. (2001). The far more general framework of subspace valued maps was introduced in (Argyriou and Dinuzzo, 2014, Theorem 3.1) and a variant of the presented theorem with an assumption of $r-$regularity, for bounded linear functionals and with the loss functional C on $\mathcal{Z} = \mathbb{R}^m$ can be found there.

4. Application examples

4.1 Deep neural networks

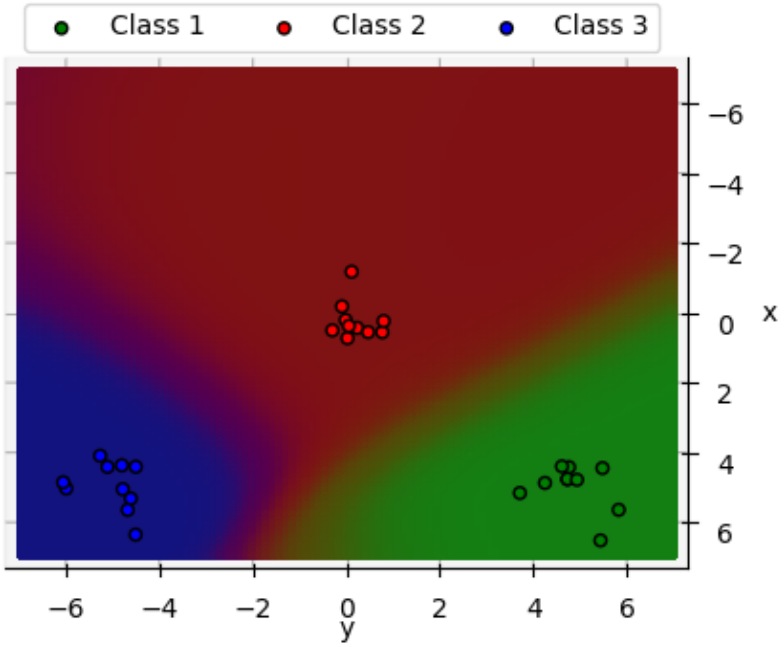

Figure 1: Multi class classification with a 3 layer, squared exponential kernel based neural network. Class probabilities shaded as red, blue, green values. Training data shown as point clusters.

4.1.1 Motivation

Consider, first, a single layer perceptron with an activation function σ, with input x and output y. Given m training samples $\{(x_i, y_i) : i \in \mathbb{N}_m\}$ consider the variational learning problem

$$\min_{f \in \mathcal{H}} \quad \sum_{i=1}^{m} ||y_i - \sigma(L_{x_i} f)||_{\mathcal{Z}}^2 + \lambda ||f||_{\mathcal{H}}^2 \tag{32}$$

Let $\mathcal{Z} = \mathbb{R}^n$, $\mathcal{H}$ be an RKHS space with kernel K and $L_{x_i} : \mathcal{H} \to \mathcal{Z}$ be a closed bounded linear evaluation operator $L_{x_i} f = f(x_i)$ on the RKHS space. This minimization problem fits exactly the form of (17) by taking $C(L_{x_1} f, \ldots, L_{x_m} f) = \sum_{i=1}^{m} ||y_i - \sigma(L_{x_i}(\cdot))||^2$ and Ω to be $||f||_{\mathcal{H}}^2$. Since Ω is orthomonotone with respect to $S_{\mathbb{R}}$, we know a minimizer of the

form $\sum_{i=1}^{m} L^*_{x_i} z_i$ must exist. Substituting this form into the minimization above we can get a finite dimensional minimization problem.

$$\min_{z_j \in \mathcal{Z}} \quad \sum_{i=1}^{m} ||y_i - \sigma(L_{x_i} \sum_{j=1}^{m} L^*_{x_j} z_j)||^2_{\mathcal{Z}} + \lambda || \sum_{j=1}^{m} L^*_{x_j} z_j ||^2_{\mathcal{H}} \tag{33}$$

On the RKHS $\mathcal{H}$, the adjoint L^*_x is known to be the kernel section $K(\cdot, x)$ (see Example 1) and $L_{x_i} L^*_{x_j} = K(x_i, x_j)$. Thus we have a nonlinear program to solve for a kernel based single layer perceptron with $z_i \in \mathcal{Z}$ being the new decision variables. Note that the program becomes nonlinear due to a nonlinear activation function σ and only thus differs from a generalized least squares setting.

So far we see nothing new as the problem is simply a least squares like problems in the RKHS space with finite dimensional outputs. Such problems can easily be covered by representer theorems from Argyriou and Dinuzzo (2014).

Now consider a N-layer concatenation of such perceptrons. Let the inputs for the first layer be denoted as $y^{(0)} = (y^{(0)}_1, \ldots, y^{(0)}_m) \in \mathbb{R}^{n_0 \times m}$ taking values $y^{(0)}_i = X_i$ from a training data set $\mathcal{D} = \{(X_i, Y_i) \in \mathbb{R}^{n_0} \times \mathbb{R}^{n_N} : i = 1, \ldots, m\}$. Let the function $f^{(1)}$ for the first layer be learned from an RKHS space $\mathcal{H}^{(1)}$ of $\mathbb{R}^{n_1}$-valued functions and let the output for the first layer be the unknown latent variables $y^{(1)} = (y^{(1)}_1, \ldots, y^{(1)}_m) \in \mathbb{R}^{n_1 \times m}$. Let $\mathcal{Z}^{(1)}$ denote the separable Hilbert space $\mathbb{R}^{n_1 \times m} \times \mathcal{H}^{(1)}$ for notational convenience. The learning of the function $f^{(1)}$ can thus be considered as the variational problem,

$$y^{(1)}_{opt}, f^{(1)}_{opt} = \text{argmin}_{(y^{(1)}, f^{(1)}) \in \mathcal{Z}^{(1)}} \, C_1(L^{(1)}(y^{(1)}, f^{(1)})) + \Omega_1((y^{(1)}, f^{(1)})) \tag{34}$$

with $L^{(1)}(y^{(1)}, f^{(1)}) = (y^{(1)}_1 - L_{y^{(0)}_1} f^{(1)}, \ldots, y^{(1)}_m - L_{y^{(0)}_m} f^{(1)})$ being the bounded linear operator $L^{(1)} : \mathcal{Z}^{(1)} \to \mathbb{R}^{n_1 \times m}$, $C_1 : \mathbb{R}^{n_1 \times m} \to \mathbb{R} \cup \{\infty\}$ being the loss functional such that $C_1(L^{(1)}(y^{(1)}, f^{(1)})) = \begin{cases} 0 & , \text{ if } y^{(1)} = L_{y^{(0)}} f^{(1)} \\ \infty & , \text{otherwise} \end{cases}$ and $\Omega_1((y^{(1)}, f^{(1)})) = ||(f^{(1)})||^2_{\mathcal{H}^{(1)}}$ being the regularizer. Again, nothing new so far, we have a Hilbert search space $\mathcal{Z}^{(1)}$ and a finite dimensional domain for the loss functional, $\mathbb{R}^{n_1 \times m}$. Also, note that this variational problem is ill posed since only the input data is fixed and the output data is left free and thus the minimizer for the above problem is at $y^{(1)}_{opt} = 0$ and $f^{(1)}_{opt} = 0$. We ignore the ill-posed nature of the optimization for now, as additional concatenated layers connecting to the final output data will force the minimizer to become non trivial.

Consider next the second layer for the network. Let $y^{(2)} \in \mathbb{R}^{n_2 \times m}$ be the latent variables, $\mathcal{H}^{(2)}$ be a an RKHS space of $\mathbb{R}^{n_2}$-valued functions and $f^{(2)} \in \mathcal{H}^{(2)}$ be the learned function for this layer. Let $\mathcal{Z}^{(2)}$ denote the Hilbert space $\mathbb{R}^{n_2 \times m} \times \mathcal{H}^{(2)}$. The learning problem for the second layer can then be posed as,

$$y^{(1)}_{opt}, y^{(2)}_{opt}, f^{(2)}_{opt} = \text{argmin}_{y^{(1)} \in \mathbb{R}^{n_1 \times m}, (y^{(2)}, f^{(2)}) \in \mathcal{Z}^{(2)}} \, C_2((y^{(1)}, y^{(2)}, f^{(2)})) + \Omega_2((y^{(2)}, f^{(2)})) \tag{35}$$

with $C_2((y^{(1)}, y^{(2)}, f^{(2)})) = \begin{cases} 0 & , \text{ if } y^{(2)} = L_{y^{(1)}} f^{(2)} \\ \infty & , \text{otherwise} \end{cases}$ and $\Omega_2((y^{(2)}, f^{(2)})) = ||f^{(2)}||^2_{\mathcal{H}^{(2)}}$.

This is where we see a significant difference from the standard least squares like problem for the first time. Here $y^{(1)}$ being an unknown latent variable, is considered as a decision

variable for the problem and thus $L_{y^{(1)}}$ is not a linear operator on the search space $\mathbb{R}^{n_1\times m}\times \mathcal{Z}^{(2)}$. Thus unlike the first layer we cannot write the loss functional for the second layer as $C_2(L(y^{(1)},y^{(2)},f^{(2)}))$ for some linear operator $L:\mathbb{R}^{n_1\times m}\times\mathcal{Z}^{(2)}\to\mathbb{R}^{n_2\times m}$. The operator $L_{y^{(1)}}$ makes the operator $L(y^{(1)},y^{(2)},f^{(2)})=y^{(2)}-L_{y^{(1)}}f^{(2)}$ a non-linear operator. Instead we consider a non-linear loss functional $C:\mathbb{R}^{n_1\times m}\times\mathbb{R}^{n_2\times m}\times\mathcal{H}^{(2)}\to\mathbb{R}\cup\{\infty\}$ as given in (35).

The problem for learning the first and second layer together can then be written as

$$\begin{aligned} y^{(1)}_{opt},f^{(1)}_{opt},y^{(2)}_{opt},f^{(2)}_{opt} = \operatorname{argmin}_{(y^{(1)},f^{(1)})\in\mathcal{Z}^{(1)},(y^{(2)},f^{(2)})\in\mathcal{Z}^{(2)}} C_1(L^{(1)}(y^{(1)},f^{(1)})) \\ + C_2(y^{(1)},y^{(2)},f^{(2)})+\Omega_1((y^{(1)},f^{(1)}))+\Omega_2((y^{(2)},f^{(2)})) \end{aligned} \tag{36}$$

Also note that we did not use any activation functions σ in the construction above. This was done to show clearly that the nonlinearity of the operation $L_{y^{(1)}}f^{(2)}$ present in C_2 has nothing to do with the activation function. Even with a simple interpolation or least squares like loss function we have to treat C_2 as a nonlinear functional on the Hilbert space $\mathbb{R}^{n_1\times m}\times\mathcal{Z}^{(2)}$. Having shown that C_2 is a nonlinear functional on $\mathbb{R}^{n_1\times m}\times\mathcal{Z}^{(2)}$ in any case, we can reintroduce the activation function and write $C^{(2)}:\mathbb{R}^{n_1\times m}\times\mathcal{Z}^{(2)}\to\mathbb{R}\cup\{\infty\}$ as the functional

$$C_2((y^{(1)},y^{(2)},f^{(2)}))=\begin{cases}0 & ,\text{ if } y^{(2)}=\sigma(L_{y^{(1)}}f^{(2)})\\ \infty & ,\text{otherwise}\end{cases} \tag{37}$$

For the functional C_1, reintroducing σ makes the operator $L^{(1)}:\mathbb{R}^{n_1\times m}\times\mathcal{H}^{(1)}\to\mathbb{R}^{n_1\times m}$ defined above, nonlinear. We can instead view the operator $L^{(1)}$ as the linear operator $L^{(1)}:\mathbb{R}^{n_1\times m}\times\mathcal{H}^{(1)}\to\mathbb{R}^{n_1\times m}\times\mathbb{R}^{n_1\times m}$ given as the mapping

$$L^{(1)}(y^{(1)},f^{(1)})=(y^{(1)},L_{y^{(0)}}f^{(1)})$$

and C_1 as a corresponding nonlinear functional on $\mathbb{R}^{n_1\times m}\times\mathbb{R}^{n_1\times m}$. Thus we can view C_1 as the functional $C_1:\mathbb{R}^{n_1\times m}\times\mathbb{R}^{n_1\times m}\to\mathbb{R}\cup\{\infty\}$, given as,

$$C_1(L^{(1)}(y^{(1)},f^{(1)}))=\begin{cases}0 & ,\text{ if } y^{(1)}=\sigma(L_{y^{(0)}}f^{(1)})\\ \infty & ,\text{otherwise}\end{cases} \tag{38}$$

A similar construction can be done for each layer upto the $(N-1)^{th}$ layer. Note also that, while we introduced $\mathcal{H}^{(l)}$ as a RKHS space and $L_{y^{(l-1)}}$ as the linear evaluation operator evaluating functions at the point $y^{(l-1)}$, the same construction remains valid for any separable Hilbert space $\mathcal{H}^{(l)}$ and any closed, densely defined linear operator $L_{y^{(l-1)}}$, where the subscript $y^{(l-1)}$ denotes that the operators action depends on the output of the previous layer. The following describes the construction of the full N-layer neural network.

4.1.2 Formal construction

Let $y^{(l)}\in\mathbb{R}^{n_l\times m}$ be the latent output variable for each layer $l=1,\dots,N-1$. Let $y^{(0)}=(X_1,\dots,X_m)$ and $y^{(N)}=(Y_1,\dots,Y_m)$ be the known input and output data respectively, used for training the network. Let $f^{(l)}$ denote the function learned for the l^{th} layer from a separable Hilbert space $\mathcal{H}^{(l)}$ of $\mathbb{R}^{n_l}-$valued functions. Let $\mathcal{O}_{\mathcal{H}^{(l)},\mathbb{R}^{n_l\times m}}$ be a set of closed,

densely defined operators from $\mathcal{H}^{(l)}$ to $\mathbb{R}^{n_l \times m}$. Let $\phi_l : \mathbb{R}^{n_{l-1} \times m} \to \mathcal{O}_{\mathcal{H}^{(l)}, \mathbb{R}^{n_l \times m}}$ be known functions mapping the output, $y^{(l-1)}$, of the $(l-1)^{th}$ layer to some closed, densely defined operator, $L_{y^{(l-1)}} \in \mathcal{O}_{\mathcal{H}^{(l)}, \mathbb{R}^{n_l \times m}}$, denoted as $L_{y^{(l-1)}} = \phi_l(y^{(l-1)})$. Let $L^*_{y^{(l-1)}} : \mathbb{R}^{n_l \times m} \to \mathcal{H}^{(l)}$ denote the adjoint to $L_{y^{(l-1)}}$ and ϕ_l^* denote the map $\phi_l^*(y^{(l-1)}) = L^*_{y^{(l-1)}}$. An example for $\mathcal{O}_{\mathcal{H}^{(l)}, \mathbb{R}^{n_l \times m}}$ is the set of all evaluation operators on an RKHS space and the function ϕ maps $y^{(l-1)}$ to the linear operator evaluating a function in the RKHS space at $y^{(l-1)}$. Another example for $\mathcal{O}_{\mathcal{H}^{(l)}, \mathbb{R}^{n_l \times m}}$ is the set of gradient operators ∇_x computing the gradient of a function in $\mathcal{H}^{(l)}$ at a point $x \in \mathbb{R}^{n_{l-1} \times m}$ with $\phi(y^{(l-1)}) = \nabla_{y^{(l-1)}}$.

For notational convenience, let $z^{(l)} = (y^{(l)}, f^{(l)})$ and $\mathcal{Z}^{(l)} = \mathbb{R}^{n_l \times m} \times \mathcal{H}^{(l)}$.
Let

$$C_l(y^{(l-1)}, z^{(l)}) = \begin{cases} 0 & y^{(l)} = \sigma_l(\phi_l(y^{(l-1)}) f^{(l)}) \\ \infty & \text{otherwise} \end{cases} \text{ for } l = 1, \ldots, N-1 \tag{39}$$

be the lower semi-continuous functional $C_l : \mathbb{R}^{n_l \times m} \times \mathcal{Z}^{(l)} \to \mathbb{R} \cup \{\infty\}$, with $\sigma_l : \mathbb{R} \to \mathbb{R}$ being a lower semi-continuous function, interpreted as acting on each component for a matrix in $\mathbb{R}^{n_l \times m}$. Let,

$$\Omega_l(z^{(l)}) = ||f^{(l)}||^2_{\mathcal{H}^{(l)}} \text{ for } l = 1, \ldots, N$$

be the regularizer $\Omega_l : \mathcal{Z}^{(l)} \to \mathbb{R} \cup \{\infty\}$.

For the final N^{th} layer, let $y^{(N)} = (Y_1, \ldots, Y_m) \in \mathbb{R}^{n_N \times m}$ be a known output vector. Let the loss functional $C_N : \mathbb{R}^{n_{N-1} \times m} \times \mathcal{H}^{(N)} \to \mathbb{R} \cup \{\infty\}$ be given as

$$C_N(y^{(N-1)}, f^{(N)}) = ||y^{(N)} - L_{y^{(N-1)}} f^{(N)}||^2_{\mathbb{R}^{n_N \times m}}$$

Given a training data set $\mathcal{D} = \{(X_i, Y_i) : i = 1, \ldots, m\}$ of input-output pairs, we can write the full N-layer neural network learning problem as

$$\begin{aligned} z^{(1)}_{opt}, \ldots, z^{(N-1)}_{opt}, f^{(N)}_{opt} = \operatorname{argmin}_{\substack{z^{(1)}, \ldots, z^{(N-1)}, f^{(N)} \\ \in \mathcal{Z}^{(1)} \times \cdots \times \mathcal{Z}^{(N-1)} \times \mathcal{H}^{(N)}}} & C_N(y^{(N-1)}, f^{(N)}) + \sum_{l=1}^{N-1} C_l(y^{(l-1)}, z^{(l)}) \\ & + \sum_{l=1}^{N} \Omega_l(z^{(l)}) \end{aligned} \tag{40}$$

4.1.3 Applying the generalizer representer theorem to the neural network

(40) written in the standard form for the representer theorem,

$$F_{opt} = \operatorname{argmin}_{F \in \mathcal{H}} \quad C(LF) + \Omega(F) \tag{41}$$

is a problem considered on the Hilbert space $\mathcal{H} = \mathcal{Z}^{(1)} \times \cdots \times \mathcal{Z}^{(N-1)} \times \mathcal{H}^{(N)}$. Let $F \in \mathcal{H}$, be the concatenated vector $F = (z^{(1)}, \ldots, z^{(N-1)}, f^{(N)})$. The operator $L : \mathcal{H} \to \mathbb{R}^{n_1 \times m} \times \mathbb{R}^{n_1 \times m}$ be a closed, densely defined operator, given by the oblique projection $L(F) = (y^{(1)}, L_{y_0} f^{(1)})$. Given the adjoint operator $L^*_{y_0} : \mathbb{R}^{n_1 \times m} \to \mathcal{H}^{(1)}$, we can write the adjoint $L^* : \mathbb{R}^{n_1 \times m} \times \mathbb{R}^{n_1 \times m} \to \mathcal{H}$ as $L^*(y, c) = ((y, L^*_{y_0} c), 0, 0, \ldots, 0)$. Thus L is an operator $L : \mathcal{H} \to (\mathbb{R}^{n_1 \times m} \times \mathcal{H}^{(1)})$ with the null space orthogonal

$$\mathcal{N}_L^{\perp} = \mathbb{R}^{n_l \times m} \times \text{range}(L^*_{y_0}) \times \{0\} \times \{0\} \cdots \times \{0\} \tag{42}$$

with the $\{0\}$ sets corresponding to $\mathcal{Z}^{(2)} \times \mathcal{Z}^{(3)} \times \ldots \mathcal{Z}^{(N-1)} \times \mathcal{H}^{(N)}$. The functional $C(LF) = C_1(y^{(0)}, LF)$ with C_1 as defined by (39) and

$$\Omega(F) = \Omega_1(z^{(1)}) + \sum_{l=2}^{N} (C_l(y^{(l-1)}, z^{(l)}) + \Omega_l(z^{(l)})) \tag{43}$$

Let $S_{\mathbb{R}}$ be the inclusive, closed, super-additive subspace valued map $S_{\mathbb{R}}(a) = \{\lambda a : a \in \mathbb{R}\}$, considered in Example 1.

Then, consider the subspace valued maps,

$$S_1(z^{(1)}) = S_{\mathbb{R}}(y^{(1)}) \times S_{\mathbb{R}}(\text{range}(L^*_{y^{(0)}})) \tag{44}$$

For $l = 1, \ldots, N-1$, let $\mathcal{Y}_l \subseteq \mathbb{R}^{n_l \times m}$ be a Borel measurable subset of $\mathbb{R}^{n_l \times m}$ given by the range of the function σ_l, i.e., $\mathcal{Y}_l = \{\sigma_l(y) : y \in \mathbb{R}^{n_l \times m}\} \subseteq \mathbb{R}^{n \times m}$. Let $\mathcal{B}(\mathcal{Y}_l)$ be the Borel $\sigma-$algebra on $\mathcal{Y}_l$ (inherited from the Borel $\sigma-$algebra on $\mathbb{R}^{n_l \times m}$).

For $l = 1, \ldots, N-1$, recall that $C_l(z^{(l)})$, forces $y^{(l)} = \sigma(L_{y^{(l-1)}} f^{(l)})$ for a non-infinite cost. Then, the range of values for $y^{(l)}$, $\mathcal{Y}_l$ restricts the possible input values for ϕ_{l+1} and shrinks the solution space in which a minimizer may lie. For measurable, bounded variation functions ϕ^*_l, we can exploit this fact by considering the following subspace valued map over the $\mathcal{Z}^{(l)}$, for $l = 2, \ldots, N-1$,

$$S_l(z^{(l)}) = S_{\mathbb{R}}(y^{(l)}) \times \text{closure}\left(\left\{\int_{\mathcal{Y}_l} \phi^*_l(y) dc_l(y) : c_l \in M_\sigma(\mathcal{Y}_{l-1}, \mathcal{B}(\mathcal{Y}_{l-1}); \mathbb{R}^{n_l \times m})\right\}\right) \tag{45}$$

where $M_\sigma(\mathcal{Y}_{l-1}, \mathcal{B}(\mathcal{Y}_{l-1}); \mathbb{R}^{n_l \times m})$ is the Banach space of signed, $\mathbb{R}^{n_l \times m}-$valued Borel measures with bounded total variation (see Zaanen (2012)) on the measurable space $(\mathcal{Y}_l, \mathcal{B}(\mathcal{Y}_l))$.

Lemma 21 shows that $S_l : \mathcal{Z}^{(l)} \to \mathcal{V}(\mathcal{Z}^{(l)})$ defined in (45) under a certain regularity assumptions for the map ϕ_l and ϕ^*_l over the domain $\mathcal{Y}_{l-1}$, is a closed and super-additive subspace valued map.

Lemma 21 *(S_l is closed and super-additive)*
Let $\mathcal{Y}_{l-1}$ be a Borel measurable subset of $\mathbb{R}^{n_{l-1} \times m}$. Let $||T||_{\mathcal{L}_{\mathbb{R}^{n_l \times m}}} = \inf\{c \geq 0 : ||Tv||_{\mathbb{R}^{n_l \times m}} \leq c||v||_{\mathbb{R}^{n_l \times m}}$ for all $v \in \mathbb{R}^{n_l \times m}\}$ be the standard operator norm for bounded operators mapping $\mathbb{R}^{n_l \times m}$ into itself. Let ϕ_l, ϕ^*_l be measurable functions such that, ϕ^*_l is a function of bounded variation and the self-adjoint operator given by $\phi(y)\phi^*(y) = L_y L^*_y$ is a closed and bounded linear operator, for all $y \in \mathcal{Y}_{l-1}$ and there exists a constant $M < \infty$ satisfying the bound $\sup\{||\phi_l(y)\phi^*_l(y)||_{\mathcal{L}_{\mathbb{R}^{n_l \times m}}} : y \in \mathcal{Y}_{l-1}\} = M$ for all $y \in \mathcal{Y}$. Let $M_\sigma(\mathcal{Y}_{l-1}, \mathcal{B}(\mathcal{Y}_{l-1}); \mathbb{R}^{n_l \times m})$ be the Banach space of signed, $\mathbb{R}^{n_l \times m}-$valued Borel measures with finite total variation. Then $S_l : \mathcal{Z}^{(l)} \to \mathcal{V}(\mathcal{Z}^{(l)})$ as defined by (45) is a closed and super-additive subspace valued map.
Proof *The map S_l is a product of $S_{\mathbb{R}}$ with the set*

$$K = \text{closure}\left(\left\{\int_{\mathcal{Y}_l} \phi^*_l(y) dc_l(y) : c_l \in M_\sigma(\mathcal{Y}_{l-1}, \mathcal{B}(\mathcal{Y}_{l-1}); \mathbb{R}^{n_l \times m})\right\}\right)$$

$S_{\mathbb{R}}$ is already known to be closed and super-additive (from Example 1). Thus it only remains to be shown that the set K is closed, super-additive and actually subspace valued i.e. $K \subseteq \mathcal{H}^{(l)}$ and $K \in \mathcal{V}(\mathcal{H}^{(l)})$.

If ϕ_l^* is assumed to be integrable with respect to every $c_l \in M_\sigma(\mathcal{Y}_{l-1}, \mathcal{B}(\mathcal{Y}_{l-1}); \mathbb{R}^{n_l \times m})$, it is easy to see that K is a vector space since for any $f_1, f_2 \in K$, there exist $c_l^1, c_l^2 \in M_\sigma(\mathcal{Y}_{l-1}, \mathcal{B}(\mathcal{Y}_{l-1}); \mathbb{R}^{n_l \times m})$ such that $f_i = \int_{\mathcal{Y}_l} \phi_l^*(y) dc_l^i(y)$ for $i = 1, 2$. Thus by linearity for the integral we have for any $\alpha, \beta \in \mathbb{R}$, $\alpha f_1 + \beta f_2 = \int_{\mathcal{Y}_{l-1}} \phi_l^*(y) d(\alpha c_l^1(y) + \beta c_l^2(y))$. Since $M_\sigma(\mathcal{Y}_{l-1}, \mathcal{B}(\mathcal{Y}_{l-1}); \mathbb{R}^{n_l \times m})$ is a closed vector space (actually a Banach space), we have a $c_l' = \alpha c_l^1 + \beta c_l^2 \in M_\sigma(\mathcal{Y}_{l-1}, \mathcal{B}(\mathcal{Y}_{l-1}); \mathbb{R}^{n_l \times m})$ and thus $\alpha f_1 + \beta f_2$ belongs to K. Next we show that under the conditions mentioned for ϕ_l, ϕ_l^*, ϕ_l^* is actually integrable with respect to every $c_l \in M_\sigma(\mathcal{Y}_{l-1}, \mathcal{B}(\mathcal{Y}_{l-1}); \mathbb{R}^{n_l \times m})$ and that $K \subseteq \mathcal{H}^{(l)}$ is actually a closed subspace, i.e. $K \in \mathcal{V}(\mathcal{H}^{(l)})$.

By (Dobrakov, 1988, Definition 1), the measurable function $\phi_l^* : \mathcal{Y}_{l-1} \to \mathcal{L}_{\mathbb{R}^{n_l \times m}, \mathcal{H}^{(l)}}$ is integrable with respect to a $\mathbb{R}^{n_l \times m}-$valued measure $c_l : \mathcal{B}(\mathcal{Y}_{l-1}) \to \mathbb{R}^{n_l \times m}$ if there exists a $h \in \mathcal{H}^{(l)}$ such that for every $\epsilon > 0$, and every countable partition $\{E_i\}$ of $\mathcal{Y}_{l-1}$ with maximum volume ϵ for any cell in the partition, and any selection of points $y_i \in E_i$, we have $||h - \sum_{E_i} \phi_l^*(y_i) c_l(E_i)||_{\mathcal{H}^{(l)}} < \epsilon$. For this to be true we need that $||\sum_{E_i} \phi_l^*(y_i) c_l(E_i)||_{\mathcal{H}^{(l)}} < \infty$ for any partition $\{E_i\}$ and selection $y_i \in E_i$, and then convergence of $\sum_{E_i} \phi_l^*(y_i) c_l(E_i)$ to a unique $h \in \mathcal{H}^{(l)}$.

First we show that $||\sum_{E_i} \phi_l^*(y_i) c_l(E_i)||_{\mathcal{H}^{(l)}} < \infty$. Note that $||\sum_{E_i} \phi_l^*(y_i) c_l(E_i)||_{\mathcal{H}^{(l)}} \leq \sum_{E_i} ||\phi_l^*(y_i) c_l(E_i)||_{\mathcal{H}^{(l)}} = \sum_{E_i} (\langle \phi_l^*(y_i) c_l(E_i), \phi_l^*(y_i) c_l(E_i) \rangle_{\mathcal{H}^{(l)}})^{1/2} = \sum_{E_i} (\langle c_l(E_i), \phi_l(y_i) \phi_l^*(y_i) c_l(E_i) \rangle_{\mathbb{R}^{n_l \times m}})^{1/2} \leq \sum_{E_i} (||c_l(E_i)||_{\mathbb{R}^{n_l \times m}} ||\phi_l(y_i) \phi_l^*(y_i) c_l(E_i)||_{\mathbb{R}^{n_l \times m}})^{1/2} \leq \sum_{E_i} (||c_l(E_i)||^2_{\mathbb{R}^{n_l \times m}} ||\phi_l(y_i) \phi_l^*(y_i)||_{\mathcal{L}_{\mathbb{R}^{n_l \times m}}})^{1/2} = \sum_{E_i} ||c_l(E_i)||_{\mathbb{R}^{n_l \times m}} ||\phi_l(y_i) \phi_l^*(y_i)||^{1/2}_{\mathcal{L}_{\mathbb{R}^{n_l \times m}}}$. But note from the assumptions on $\phi_l(y) \phi_l^*(y)$, that $||\phi_l(y_i) \phi_l(y_i)||_{\mathcal{L}_{\mathbb{R}^{n_l \times m}}} < M$. Thus $||\sum_{E_i} \phi_l^*(y_i) c_l(E_i)||_{\mathcal{H}^{(l)}} \leq M \sum_{E_i} ||c_l(E_i)||_{\mathbb{R}^{n_l \times m}} \leq M ||c_l||_{M_\sigma(\mathcal{Y}_{l-1}, \mathcal{B}(\mathcal{Y}_{l-1}), \mathbb{R}^{n_l \times m})}$ (the inequality follows from the definition of the norm on the space of vector measures Schwarz (1967), $||c_l||_{M_\sigma(\mathcal{Y}_{l-1}, \mathcal{B}(\mathcal{Y}_{l-1}), \mathbb{R}^{n_l \times m})} = \sup \sum_{E_i} ||c_l(E_i)||_{\mathbb{R}^{n_l \times m}}$, over all partitions $\{E_i\}$ of $\mathcal{Y}_{l-1}$). Thus for any $c_l \in M_\sigma(\mathcal{Y}_{l-1}, \mathcal{B}(\mathcal{Y}_{l-1}), \mathbb{R}^{n_l \times m})$ and any partition $\{E_i\}$ of $\mathcal{Y}_{l-1}$, $||\sum_{E_i} \phi_l^*(y_i) c_l(E_i)||_{\mathcal{H}^{(l)}} < \infty$.

Now for the uniqueness in convergence, note that for any refinement $\{E_i'\}$ of a partition $\{E_i\}$ such that $E_i = E_{2i}' \cup E_{2i+1}'$ for all $i = 0, \ldots, \infty$, and selection of points without loss of generality, as $y_i = y_{2i}'$, we have $||\sum_{E_i} \phi_l^*(y_i) c_l(E_i) - \sum_{E_i'} \phi_l^*(y_i') c_l(E_i')|| = ||\sum_{E_{2i}'} \phi_l^*(y_{2i}') c_l(E_{2i}') + \sum_{E_{2i+1}'} \phi_l^*(y_{2i}') c_l(E_{2i+1}') - \sum_{E_{2i}'} \phi_l^*(y_{2i}') c_l(E_{2i}') - \sum_{E_{2i+1}'} \phi_l^*(y_{2i+1}') c_l(E_{2i+1}')|| = ||\sum_{E_{2i+1}'} (\phi_l^*(y_{2i}') - \phi_l^*(y_{2i+1})) c_l(E_{2i+1}')|| \leq \sum_{E_{2i+1}'} ||(\phi_l^*(y_{2i}') - \phi_l^*(y_{2i+1}'))|| ||c_l(E_{2i+1}')||$. Since ϕ_l^* has bounded total variation for any partition E_i', we have $\sum_{E_{2i+1}'} ||(\phi_l^*(y_{2i}') - \phi_l^*(y_{2i+1}'))|| \leq \sup \sum_{E_i'} ||(\phi_l^*(y_i') - \phi_l^*(y_{i+1}'))|| < \infty$. Then as the partition refinements converge $E_i' \to E_i$, $c_l(E_{2i+1}')$ tends to 0 and thus we have $\sum_{E_{2i+1}'} ||(\phi_l^*(y_{2i}') - \phi_l^*(y_{2i+1}'))|| ||c_l(E_{2i+1}')||$ converging to 0.

Thus we have shown that the conditions of ϕ_l^* and $\phi_l(y) \phi_l^*(y)$ ensure that ϕ_l^* is integrable with respect to all c_l and thus the integral $h = \int_{\mathcal{Y}_l} \phi_l^*(y) dc_l(y)$ belongs to $\mathcal{H}^{(l)}$ for all c_l, implying K is a vector subspace of $\mathcal{H}^{(l)}$.

Finally taking the closure of K, makes the subspace a closed subspace of $\mathcal{H}^{(l)}$ (since $\mathcal{H}^{(l)}$ is closed).

Now, since $S_l(y^{(l)}, f) = S_{\mathbb{R}}(y^{(l)}) \times K$ for any $f \in \mathcal{H}^{(l)}$ (K does not depend on f), we have for any closed subspaces $A, B \in \mathcal{V}(\mathcal{Z}^{(l)})$, we have $S_l(A) = A_y \times K$ and $S_l(B) = B_y \times K$, where A_y, B_y are the vector subspaces in $\mathcal{V}(\mathbb{R}^{n_{l-1} \times m} \times \mathbb{R}^{n_l \times m})$ corresponding to the additive

subspace valued map $S_{\mathbb{R}}$. Thus we have $S_l(A)+S_l(B) = A_y \times K + B_y \times K = (A_y+B_y)\times K = S_l(A+B)$ (since we have $S_{\mathbb{R}}(A_y) = A_y, S_{\mathbb{R}}(A_y) = B_y$ and $S_{\mathbb{R}}(A_y+B_y) = A_y+B_y$), implying S_l is closed and super-additive (trivially, since its additive). ■

For the last layer, consider,

$$S_N(f^{(N)}) = \text{closure}\left(\left\{\int_{\mathcal{Y}_l} \phi_l^*(y) dc_l(y) : c_l \in M_\sigma(\mathcal{Y}_{N-1}, \mathcal{B}(\mathcal{Y}_{N-1}); \mathbb{R}^{n_N \times m})\right\}\right) \tag{46}$$

which is again a closed, super-additive subspace valued map by the arguments in Lemma 21.

Now consider the subspace valued map $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ for the complete network given by

$$S(F) = S_1(z^{(1)}) \times S_2(z^{(2)}) \times \cdots \times S_{N-1}(z^{(N-1)}) \times S_N(f^{(N)}) \tag{47}$$

Theorems 22 below, shows that the functional Ω in (43) is orthomonotone with respect to the subspace $S(\mathcal{N}_L^\perp)$ for $\mathcal{N}_L^\perp$ as defined in (42).

Theorem 22 *Let $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ be the closed, super-additive subspace valued map defined in (46). Let $\mathcal{N}_L^\perp$ be the closed subspace as defined in (42) and $\Omega : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be the functional from (43). Then Ω is orthomonotone with respect to the subspace $S(\mathcal{N}_L^\perp)$, i.e. for all $f \in S(\mathcal{N}_L^\perp)$ and $g \in S(\mathcal{N}_L^\perp)^\perp$, $\Omega(f+g) \geq \Omega(f)$.*

Proof *For the vector subspace $\mathcal{N}_L^\perp = \mathbb{R}^{n_l \times m} \times \text{range}(L^*_{y_0}) \times \{0\} \times \{0\} \cdots \times \{0\}$ in $\mathcal{V}(\mathcal{H})$, we have $S(\mathcal{N}_L^\perp) = \mathbb{R}^{n_1 \times m} \times \text{range}(L^*_{y^{(0)}}) \times (\Pi_{l=2}^{N-1} \mathbb{R}^{n_l \times m} \times K_l) \times K_N$, for the subspace $K_l = \text{closure}\left(\left\{\int_{\mathcal{Y}_l} \phi_l^*(y) dc_l(y) : c_l \in M_\sigma(\mathcal{Y}_{l-1}, \mathcal{B}(\mathcal{Y}_{l-1}); \mathbb{R}^{n_l \times m})\right\}\right)$ for each $l = 2, \ldots, N$. Also $S(\mathcal{N}_L^\perp)^\perp = \{0\} \times \text{range}(L^*_{y_0})^\perp \times (\Pi_{l=2}^{N-1}\{0\} \times K_l^\perp) \times K_N^\perp$. Also recall that $\Omega(F) = \Omega_1(z^{(1)}) + \sum_{l=2}^N (C_l(y^{(l-1)}, z^{(l)}) + \Omega_l(z^{(l)}))$*

Thus for $F = (z^{(1)}, \ldots, z^{(N-1)}, f^{(N)}) \in S(\mathcal{N}_L^\perp)$, we have $z^{(1)} = (y^{(1)}, f^{(1)}) \in \mathbb{R}^{n_1 \times m} \times \text{range}(L^*_{y_0})$, $z^{(l)} = (y^{(l)}, f^{(l)}) \in \mathbb{R}^{n_l \times m} \times K_l$ for $l = 2, \ldots, N-1$ and $z^{(N)} \in K_N$.

Similarly for $G = (x^{(1)}, \ldots, x^{(N-1)}, g^{(N)}) \in S(\mathcal{N}_L^\perp)^\perp$, we have $x^{(l)} = (y'^{(l)}, g^{(l)})$ for $y'^{(l)} = 0$ and $g^{(l)} \in K_l^\perp$, for each $l = 2, \ldots, N-1$, $g^{(N)} \in K_N^\perp$. And $x^{(1)} = (y'^{(1)}, g^{(1)})$ for $y'^{(1)} = 0$ and $g^{(1)} \in \text{range}(L^*_{y_0})^\perp$.

Now for $l = 1$, note that the only term in Ω depending on $z^{(1)}$ and $x^{(1)}$ is Ω_1 defined as $\Omega(f) = ||f||^2_{\mathcal{H}^{(1)}}$. The squared norm functional is orthomonotone for any pair of orthogonal subspaces (from Example 5). Thus for any orthogonal $z^{(1)}$ and $x^{(1)}$ as defined above we have $\Omega_1(z^{(1)} + x^{(1)}) = ||z^{(1)} + x^{(1)}||^2 = ||z^{(1)}||^2 + ||x^{(1)}||^2 \geq ||z^{(1)}||^2 = \Omega_1(z^{(1)})$ (the equality of square of sum, to sum of squares, follows from orthogonality of the two vectors).

Similarly for each $l = 2, \ldots, N-1$, for the orthogonal vectors $z^{(l)}$ and $x^{(l)}$, we have $\Omega_l(z^{(l)} + x^{(l)}) \geq \Omega_l(z^{(l)})$ and for $f^{(N)}, g^{(N)}$, $\Omega(f^{(N)} + g^{(N)}) \geq \Omega(f^{(N)})$.

The terms remaining to be shown orthomonotone are the functional C_l. Note that for all $l = 2, \ldots, N$, we have $y^{(l-1)} \in \mathbb{R}^{n_{l-1} \times m}$, $y^{(l)} \in \mathbb{R}^{n_l \times m}$ and $f^{(l)} \in K_l$, and we have $y'^{(l-1)} = 0$, $y'^{(l)} = 0$ and $g^{(l)} \in K_l^\perp$.

Then, $C_l(y^{(l-1)}+y'^{(l-1)}, y^{(l)}+y'^{(l)}, f^{(l)}+g^{(l)}) = C_l(y^{(l-1)}, y^{(l)}, f^{(l)}+g^{(l)})$ (since $y'^{(l-1)} = 0$ and $y'^{(l-1)} = 0$).

Now note that for any $l = 1, \ldots, N-1$, if $y^{(l)} \notin \mathcal{Y}_l$, then $C_l(y^{(l-1)}, y^{(l)}, f) = \infty$ for any $f \in \mathcal{H}^{(l)}$. Then we trivially have $\Omega(F+G) = \Omega(F) = \infty$ (thus satisfying the orthomonotone inequality trivially).

For all $y^{(l)} \in \mathcal{Y}_l$, for $f^{(l)} \in K_l$, $g^{(l)} \in K_l^\perp$, we have $C_l(y^{(l-1)}, y^l, f^{(l)} + g^{(l)}) = C_l(y^{(l)} - \sigma_l(\phi_l(y^{(l-1)})f^{(l)} + \phi_l(y^{(l-1)})g^{(l)}))$ for some $c_l \in M_\sigma(\mathcal{Y}_{l-1})$.

Now since for all $z \in \mathbb{R}^{n_l \times m}$, $c_l = z\delta_{y^{(l-1)}}$ (where $\delta_{y^{(l-1)}}$ is the dirac measure centered on $y^{(l-1)}$) belongs to $M_\sigma(\mathcal{Y}_{l-1}, \mathcal{B}(\mathcal{Y}_{l-1}); \mathbb{R}^{n_l \times m})$, $\mathcal{N}^\perp_{\phi_l(y^{(l-1)})} = \text{range}(\phi_l^*(y^{(l-1)})) \subseteq K_l$, implying $K_l^\perp \subseteq \mathcal{N}_{\phi_l(y^{(l-1)})}$ for all $y^{(l-1)} \in \mathcal{Y}_{l-1}$. Thus for any $g \in K_l^\perp$, $\phi_l(y^{(l-1)})g = 0$. Thus $C_l(y^{(l-1)}, y^l, f^{(l)} + g^{(l)}) = C_l(y^{(l-1)}, y^l, f^{(l)})$, for all $y^{(l)} \in \mathcal{Y}_l$, $y^{(l-1)} \in \mathcal{Y}_{l-1}$, $f^{(l)} \in K_l$, $g^{(l)} \in K_l^\perp$.

Thus for all $l = 2, \ldots, N$, we have shown $C_l(y^{(l-1)} + y'^{(l-1)}, y^{(l)} + y'^{(l)}, f^{(l)} + g^{(l)}) + \Omega_l(y^{(l)} + y'^{(l)}, f^{(l)} + g^{(l)}) \geq C_l(y^{(l-1)}, y^{(l)}, f^{(l)}) + \Omega_l(y^{(l)}, f^{(l)})$ and $\Omega_1(z^{(l)} + x^{(l)}) \geq \Omega_1(z^{(l)})$.

Thus $\Omega(F+G) \geq \Omega(F)$ for all $F \in S(\mathcal{N}_L^\perp)$, $G \in S(\mathcal{N}_L^\perp)^\perp$, i.e. Ω is orthomonotone with respect to $S(\mathcal{N}_L^\perp)$. ■

Corollary 23 *(S is range preserving with respect to L)*
For $L : \mathcal{H} \to \mathbb{R}^{n_1 \times m} \times \mathbb{R}^{n_1 \times m}$ defined as $LF = (y^{(1)}, L_{y_0} f^{(1)})$ for $F = (z^{(1)}, \ldots, z^{(N-1)}, f^{(N)}) \in \mathcal{H}$, $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ as defined in (46) and $\mathcal{N}_L^\perp$ as defined in (42), we have $\mathcal{N}_L^\perp \subseteq S(\mathcal{N}_L^\perp)$.
Proof

$$\mathcal{N}_L^\perp = \mathbb{R}^{n_l \times m} \times \text{range}(L^*_{y_0}) \times \{0\} \times \{0\} \cdots \times \{0\}$$

and

$$S(\mathcal{N}_L^\perp) = \mathbb{R}^{n_1 \times m} \times \text{range}(L^*_{y^{(0)}}) \times (\Pi_{l=2}^{N-1} \mathbb{R}^{n_l \times m} \times K_l) \times K_N$$

for the subspace $K_l = \text{closure}\left(\left\{\int_{\mathcal{Y}_l} \phi_l^*(y) dc_l(y) : c_l \in M_\sigma(\mathcal{Y}_{l-1}, \mathcal{B}(\mathcal{Y}_{l-1}); \mathbb{R}^{n_l \times m})\right\}\right)$ for each $l = 2, \ldots, N$. From the above expressions, it is visible that $\mathcal{N}_L^\perp \subseteq S(\mathcal{N}_L^\perp)$ ■

$\mathcal{N}_L^\perp = \mathbb{R}^{n_l \times m} \times \text{range}(L^*_{y_0}) \times \{0\} \times \{0\} \cdots \times \{0\}$ in $\mathcal{V}(\mathcal{H})$, we have

Corollary 24 *(Linear representer for the neural network exists in $S(\mathcal{N}_L^\perp)$)*
There exists an optimal set of representers $c_{1,opt} \in \mathbb{R}^{n_1 \times m}$, $y^{(l)}_{opt} \in \mathcal{Y}_l$ for $l = 1, \ldots, N-1$, and $c_{l,opt} \in M_\sigma(\mathcal{Y}_{l-1}, \mathcal{B}(\mathcal{Y}_{l-1}); \mathbb{R}^{n_l \times m})$ for $l = 2, \ldots, N$ such that a minimizer for (41) of the form

$$F_{opt} = \left(y^{(1)}_{opt}, \ldots, y^{(N-1)}_{opt}, L^*_{y_0} dc_{1,opt}, \int_{\mathcal{Y}_1} \phi_2^* dc_{2,opt}, \ldots, \int_{\mathcal{Y}_{N-1}} \phi_N^* dc_{N,opt}\right) \tag{48}$$

exists.
Proof *Theorem 22 showed that Ω is orthomonotone with respect to the subspace $S(\mathcal{N}_L^\perp)$ and Corollary 23 showed that S is range preserving with respect to L. Thus by the sufficient condition for existence of representers (Theorem 18) a linear representer exists in the subspace $S(\mathcal{N}_L^\perp)$, written as (48).* ■

Now since we know that, given the optimal solution $y^{(l)}_{opt}, y^{(l-1)}_{opt}$, for the $(l-1)^{th}$ and $(l)^{th}$ layer, we have,

$$f^{(l)}_{opt} = \text{argmin}_{h^{(l)} \in \mathcal{H}^{(l)}} C_l(y^{(l)}_{opt} - \sigma_l(L_{y^{(l-1)}_{opt}} f^{(l)})) + ||f^{(l)}||^2_{\mathcal{H}^{(l)}} \tag{49}$$

for all $l = 2, \ldots, N$, we have $f^{(l)}_{opt} = L^*_{y^{(l-1)}_{opt}} p_{l,opt}$ for some $p_{l,opt} \in \mathbb{R}^{n_l \times m}$.

This implies that an optimal solution of the form

$$F_{opt} = \left(y^{(1)}_{opt}, \ldots, y^{(N-1)}_{opt}, L^*_{y_0} c_{1,opt}, \int_{\mathcal{Y}_1} \phi^*_2 p_{2,opt} d\delta_{y^{(1)}_{opt}}, \ldots, \int_{\mathcal{Y}_{N-1}} \phi^*_N p_{N,opt} d\delta_{y^{(N-1)}_{opt}} \right) \tag{50}$$

exists for some $p_{l,opt} \in \mathbb{R}^{n_l \times m}$, i.e. we know that there exist dirac measures $\delta_{y^{(l)}_{opt}}$ corresponding to the optimal measures $c_{l,opt}$ from (48). Such a representer in terms of diracs is not directly useful, since the points at which the optimal diracs are centered $y^{(l)}_{opt}$ are unknown apriori. We can however use this knowledge to guide our search for measures converging towards diracs.

We can thus design a scheme to iteratively optimize over the space of measures $c^{(l)}$ and outputs $y^{(l)}$ such that the measures converge to dirac's centered at the predicted output. In particular if the maps ϕ_l and ϕ^*_l are differentiable (in addition to the regularity conditions of Theorem 21), we can design a scheme to optimize directly over the centers of dirac measures. We show in the next subsection a numerical example for such a scheme for a squared exponential kernel, satisfying the regularity and differentiability conditions.

4.1.4 Numerical example

Consider a N-layer network with each layer given by an RKHS space $\mathcal{H}^{(l)}$ with a matrix valued square exponential kernel,

$$K_l(x,y) = \begin{pmatrix} e^{-a^{(l)}_{11}||x-y||^2} & \ldots & e^{-a^{(l)}_{1n_l}||x-y||^2} \\ \vdots & \vdots & \vdots \\ e^{-a^{(l)}_{n_l 1}||x-y||^2} & \ldots & e^{-a^{(l)}_{n_l n_l}||x-y||^2} \end{pmatrix} \tag{51}$$

for some known constants $a^{(l)}_{11}, \ldots, a^{(l)}_{n_l n_l}$, mapping $x, y \in \mathbb{R}^{n_l}$ to a matrix in $\mathbb{R}^{n_l \times n_l}$.

Given m training samples, we denote the output of a layer as $y^{(l)} = (y^{(l)}_1, \ldots, y^{(l)}_m) \in \mathbb{R}^{n_l \times m}$. The kernel function is extended to inputs from $\mathbb{R}^{n_l \times m}$ by computing the matrix,

$$K_l(x,y) = \begin{pmatrix} K_l(x_1,y_1) & \ldots & K_l(x_1,y_m) \\ \vdots & \vdots & \vdots \\ K_l(x_m,y_1) & \ldots & K_l(x_m,y_m) \end{pmatrix} \tag{52}$$

where x_i, y_i denotes the i^{th} column of x and y respectively.

Let $E_y : \mathcal{H}^{(l)} \to \mathbb{R}^{n_l \times m}$ denote the evaluation operator such that $E_y f = (f(y_1), \ldots, f(y_m))$. The adjoint to the evaluation operator on the RKHS space is given by the kernel function and thus we have the adjoint $E^*_y = K_l(\cdot, y)$.

Let $\mathcal{O}_l = \{E_y : y \in \mathbb{R}^{n_l \times m}\}$ be the set of all the linear evaluation operators on $\mathcal{H}^{(l)}$. Similarly, let $\mathcal{O}_l^* = \{E_y^* : y \in \mathbb{R}^{n_l \times m}\}$ denote the set of all adjoints to the linear evaluation operators on $\mathcal{H}^{(l)}$.

Thus we have the function $\phi^* : \mathbb{R}^{n_l \times m} \to \mathcal{O}_l^*$ given by $\phi^*(y) = K_l(\cdot, y)$ and a function $\phi : \mathbb{R}^{n_l \times m} \to \mathcal{O}_l$ given by $\phi(y) = E_y$.

Let $\sigma_l : \mathbb{R} \to \mathbb{R}$ be the hyperbolic tangent function $\sigma(x) = \tanh(x)$, extended to inputs from $\mathbb{R}^{n_l \times m}$, as

$$\sigma(X) = \begin{pmatrix} \tanh(X_{11}) & \cdots & \tanh(X_{1m}) \\ \vdots & \vdots & \vdots \\ \tanh(X_{m1}) & \cdots & \tanh(X_{mm}) \end{pmatrix}$$

where X_{ij} denotes the $(i,j)^{th}$ component of the matrix X.
Thus the activation function restricts the output of the $l^{th}-$layer to the set

$$\mathcal{Y}_l = \{X \in \mathbb{R}^{n_l \times m} : X_{ij} \in (-1, 1) \text{ for all } i, j\}$$

for all $l = 1, \ldots, N$

Since the function $\phi_l^* : \mathcal{Y}_{l-1} \to \mathcal{O}_l^*$, given by $\phi^*(y) = K_l(\cdot, y)$ has a bounded domain $\mathcal{Y}_{l-1}$ and $K_l(\cdot, y)$ is a smooth bounded function in y, ϕ_l^* is a function of bounded variation on $\mathcal{Y}_{l-1}$ and thus satisfies the regularity condition for Lemma 21. Similarly $||\phi_l(y)\phi_l^*(y)||_{\mathcal{L}_{\mathbb{R}^{n_l \times m}}} = ||K_l(y, y)||_{\mathcal{L}_{\mathbb{R}^{n_l \times m}}} = ||K_l(0, 0)||_{\mathcal{L}_{\mathbb{R}^{n_l \times m}}} < \infty$ for all $y \in \mathcal{Y}_{l-1}$, we have the regularity condition for $\phi_l \phi_l^*$ satisfied as well.

Now since $\phi^*(y)$ is a smooth function in y and we know that a optimal solution to the linear representer of the form (50) exists on the smooth manifold such that $(y_{opt}^{(l-1)}, f_{opt}^{(l)}) \in \{(y, \phi_l^*(y)p_l) : y \in \mathcal{Y}_{l-1}, p_l \in \mathbb{R}^{n_l \times m}\}$, we can instead solve the smooth finite dimensional optimization problem

$$\begin{matrix} p_{1,opt}, \ldots, p_{N,opt}, \\ y_{1,opt}, \ldots, y_{N-1,opt} \end{matrix} = \text{argmin}_{p_l \in \mathbb{R}^{n_l \times m}, y_l \in \mathcal{Y}_l \subseteq \mathbb{R}^{n_l \times m}} \sum_{l=1}^{N} C_l(y^{(l)} - \sigma_l(L_{y^{(l-1)}} L_{y^{(l-1)}}^* p_l)) + \Omega_l(f^{(l)}) \tag{53}$$

Figure 1 shows the output of a three layer neural network trained in such a way for 3 way classification of a given set of points in $\mathbb{R}^2$. The output of the network is in $\mathbb{R}^3$, with the training data given such that the i^{th} component is set to 1 if a point is in the i^{th} class and the other components are set to 0. The trained network provides an output in $\mathbb{R}^3$ and the output is passed through a soft-max function (to rescale values in each component to [0,1]) and interpreted as class probability for points in $\mathbb{R}^2$ shaded with corresponding RGB color values (a small problem in $\mathbb{R}^2$ is chosen to allow for easy visualization of the results). Also note that the optimization scheme in (53) can only guarantee convergence to a local minimizer, but this is most often the case in neural networks due to the non-convex nature of the problem.

4.2 Multi-output stochastic regression with uncertain observations

Let $\mathcal{Z} = \mathcal{C}_b(\mathcal{X})$ be the Banach space of continuous and bounded $\mathbb{R}^n-$valued functions on some domain set $\mathcal{X}$. Let $\mathcal{B}(\mathcal{Z})$ be the Borel σ-algebra on $\mathcal{Z}$ and $\mu : \mathcal{B}(\mathcal{Z}) \to [0, 1]$ be a Gaussian measure on $\mathcal{Z}$. Let $\mathcal{Z}_\mu$ be the Banach space of all affine measurable functions

$X : \mathcal{Z} \to \mathcal{Z}$ with $\mathcal{B}(\mathcal{Z}_\mu)$ being the Borel σ-algebra on $\mathcal{Z}_\mu$. $\mathcal{Z}_\mu$ defines a space of Gaussian processes on the probability measure space $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), \mu)$ (see Section **??**). Let $\nu : \mathcal{B}(\mathcal{Z}_\mu) \to [0, 1]$ be a Gaussian measure on $\mathcal{Z}_\mu$ and let $\mathcal{H}_{\mu,\nu}$ be the RKHS space of Gaussian processes induced by the measure ν on $\mathcal{Z}_\mu$ as defined in Section **??**. Let $\mathcal{Y} = \mathbb{R}^n$ and $\mathcal{B}(\mathcal{Y})$ be the Borel σ-algebra on $\mathbb{R}^n$. Let $L_x : \mathcal{Z} \to \mathcal{Y}$ be the closed, bounded linear evaluation operator $L_x f = f(x)$. The linear operator L_x induces induces a push forward Gaussian measure $\mu \circ L_x^{-1}$ on $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$ (L_x^{-1} denoting the preimage operation, not the linear inverse). Let $\mathcal{Y}_{x,\mu}$ denote the Hilbert space of affine measurable functions $y : \mathcal{Y} \to \mathcal{Y}$ induced by the push forward measure $\mu \circ L_x^{-1}$ with the inner product $\langle y_1, y_2 \rangle_{\mathcal{Y}_{x,\mu}} = \int_{\mathcal{Y}} y_1(\omega) y_2(\omega) d(\mu \circ L_x^{-1})(\omega)$. $\mathcal{Y}_{x,\mu}$ is thus denotes a Hilbert space of $\mathbb{R}^n-$valued Gaussian random vectors. The extension of L_x to $\mathcal{H}_{\mu,\nu}$, for any affine function $X : \mathcal{Z} \to \mathcal{Z}$ in $\mathcal{Z}_\mu$, is given as $L_x X(f) = X(L_x f) = X(f(x))$, and defines a linear operator $L_x : \mathcal{H}_{\mu,\nu} \to \mathcal{Y}_{x,\mu}$, to space of Gaussian random vectors in $\mathcal{Y}_{x,\mu}$. The extension $L_x : \mathcal{H}_{\mu,\nu} \to \mathcal{Y}_{x,\mu}$ also preserves the closed and bounded property of $L_x : \mathcal{Z} \to \mathcal{Y}$ (by Lemma **??**).

Assuming that the $L_x : \mathcal{H}_{\mu,\nu} \to \mathcal{Y}_{x,\mu}$ induces equivalent Gaussian measures on $\mathcal{Y}$ for all $x \in \mathcal{X}$, we can write the map as $L_x : \mathcal{H}_{\mu,\nu} \to \mathcal{Y}_{c,\mu}$, mapping into a common probability measure space on $\mathcal{Y}$. The adjoint L_x^* can then be specified by a kernel function $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}^+_{\mathcal{Y}_{c,\mu}, \mathcal{Y}_{c,\mu}}$ for the RKHS space $\mathcal{H}_{\mu,\nu}$ such that for all $y \in \mathcal{Y}_{c,\mu}$ and $f \in \mathcal{H}_{\mu,\nu}$, $\langle L_x^* y, f \rangle_{\mathcal{H}_{\mu,\nu}} = \langle K(\cdot, x) y, f \rangle_{\mathcal{H}_{\mu,\nu}} = \langle y, L_x f \rangle_{\mathcal{Y}_{c,\mu}}$. Note that $\mathcal{L}^+_{\mathcal{Y}_{c,\mu}, \mathcal{Y}_{c,\mu}}$ denotes the space of closed, bounded symmetric positive definite linear operators from $\mathcal{Y}_{c,\mu}$ into itself. Since $\mathcal{Y}_{c,\mu}$ is a Banach space of Gaussian random vectors given by all affine transformations of $\mathcal{Y}$, we must have the kernel as a deterministic function taking values in $\mathcal{L}^+_{\mathcal{Y},\mathcal{Y}}$ (else the Gaussianity will be lost), i.e., $K(x_1, x_2)(\omega) = K'(x_1, x_2)$ for all $\omega \in \mathcal{Y}$ and $K' : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{n \times n}$ being a deterministic kernel of the kind usually used in non stochastic variants of kernel regression (see for example the squared exponential kernel used in Section 4.1). The form of the kernel is determined by the choice of the Gaussian measure ν and vice versa (in general the kernel function is chosen and the measure ν is as a result determined implicitly as there is a one to one correspondence between Gaussian measures on separable Banach spaces and the induced RKHS spaces).

Now with the spaces and adjoint defined we can consider a regression problem on the RKHS space of Gaussian processes $\mathcal{H}_{\mu,\nu}$. Let $\mathcal{H}_{\mu,\nu}$ be the RKHS space of Gaussian process with a kernel K. Let $\mathcal{D} = \{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}_{c,\mu} : i = 1, \ldots, m\}$ be a given training data set with observations $y_i \in \mathcal{Y}_{c,\mu}$ given as $\mathbb{R}^n-$valued Gaussian random vectors. Then consider the regression problem,

$$f_{opt} = \underset{f \in \mathcal{H}_{\mu,\nu}}{\operatorname{argmin}} \quad \sum_{i=1}^{m} ||y_i - L_{x_i} f||^2_{\mathcal{Y}_{c,\mu}} + \lambda ||f||^2_{\mathcal{H}_{\mu,\nu}} \tag{54}$$

Note that the observations $y_i \in \mathcal{Y}_{\mu,\nu}$ are now $\mathbb{R}^n-$valued Gaussian random vectors and not points in $\mathbb{R}^n$, making the loss functional $C_i : \mathcal{Y}_{c,\mu} \to \mathbb{R} \cup \{\infty\}$, given as $||y_i - L_{x_i} f||^2_{\mathcal{Y}_{c,\mu}}$, an example of a loss functional defined on a separable Hilbert space different from $\mathbb{R}^n$. Also note that even though the functional $C_i(y_i - L_{x_i}(f))$ can be written as in terms of the mean and covariance of a $\mathbb{R}^n-$valued Gaussian random vector, i.e. a functional of the form $C_i' : \mathbb{R}^n \times \mathbb{R}^{n \times n} \to \mathbb{R} \cup \{\infty\}$, as we will see below, we cannot write this reformulated objective as an equivalent functional $C_i' \circ L'_{x_i} : \mathcal{H}_{\mu,\nu} \to \mathbb{R} \cup \{\infty\}$ for a linear operator

$L'_{x_i} : \mathcal{H}_{\mu,\nu} \to \mathbb{R}^n \times \mathbb{R}^{n\times n}$, mapping the stochastic process f to its mean and covariance at x_i (since the mapping from f to its covariance will be a nonlinear operator). Thus (54) presents an example of a regression problem where the functional $C_i : \mathcal{Y}_{c,\mu} \to \mathbb{R} \cup \{\infty\}$ must be considered on the infinite dimensional Hilbert space of measurable affine maps given by $\mathcal{Y}_{c,\mu}$ in order to establish a representer in terms of the adjoint $L^*_{x_i}$.

Since, $\Omega(f) = ||f||^2_{\mathcal{H}_{\mu,\nu}}$ is known to be orthomonotone with respect to the the subspace valued map $S_{\mathbb{R}}$, we can write the representer for (54) as

$$S_{\mathbb{R}}\left(\sum_{i=1}^{m} \text{range}(L^*_{x_i})\right) = \left\{\sum_{i=1}^{m} K(\cdot, x_i)z_i : z_i \in \mathcal{Y}_{c,\mu}\right\} \tag{55}$$

Substituting a representer into (54), we can write the equivalent optimization problem,

$$\begin{aligned} f_{opt} &= \sum_{i=1}^{m} K(\cdot, x_i) z_i^{opt} \\ z_1^{opt}, \ldots, z_m^{opt} &= \text{argmin}_{z_i \in \mathcal{Y}_{c,\mu}} \sum_{i=1}^{m} ||y_i - \sum_{j=1}^{m} K(x_i, x_j) z_j||^2_{\mathcal{Y}_{c,\mu}} + \sum_{i=1}^{m}\sum_{j=1}^{m} \langle z_i, K(x_i, x_j) z_j \rangle_{\mathcal{Y}_{c,\mu}} \\ &= \text{argmin}_{z_i \in \mathcal{Y}_{c,\mu}} \sum_{i=1}^{m} \mathbb{E}_\mu \left[||y_i - \sum_{j=1}^{m} K(x_i, x_j) z_j||^2_{\mathbb{R}^n} + \sum_{j=1}^{m} z_i^T K(x_i, x_j) z_j \right] \end{aligned} \tag{56}$$

where $\mathbb{E}_\mu$ is the expectation with respect to the μ. We can expand the expectation from (56) as

$$\begin{aligned} &\mathbb{E}_\mu[||y_i||^2_{\mathbb{R}^n}] + \mathbb{E}_\mu \left[|| \sum_{j=1}^{m} K(x_i, x_j) z_j||^2_{\mathbb{R}^n} \right] \\ -2\mathbb{E}_\mu &\left[y_i^T \sum_{j=1}^{m} K(x_i, x_j) z_j \right] + \mathbb{E}_\mu \left[\sum_{j=1}^{m} z_i^T K(x_i, x_j) z_j \right] \end{aligned} \tag{57}$$

For the terms involving the decision variables $z_i \in \mathcal{Y}_{c,\mu}$, let $K^{xx} \in \mathbb{R}^{nm\times nm}$ denote the symmetric positive definite kernel matrix such that its block $K^{xx}_{i,j}$ is the kernel evaluation $K(x_i, x_j) \in \mathbb{R}^{n\times n}$ and let Z and y be the concatenation of all z_i and y_i respectively in to the vectors $Z = (z_1, \ldots, z_m)$ and $y = (y_1, \ldots, y_m)$, i.e., we write,

$$K^{xx} = \begin{pmatrix} K(x_1, x_1) & K(x_1, x_2) & \cdots & K(x_1, x_m) \\ \vdots & \vdots & \vdots & \vdots \\ K(x_m, x_1) & K(x_m, x_2) & \cdots & K(x_m, x_m) \end{pmatrix}, \qquad Z = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_m \end{pmatrix} \quad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} \tag{58}$$

$$\tag{59}$$

And let the mean and covariance be denoted as

$$\mathbb{E}_\mu[Z] = \mu_Z, \qquad \mathbb{E}_\mu[y] = \mu_y \tag{60}$$

$$\mathbb{E}_\mu[(y - \mu_y)(y - \mu_y)^T] = \Sigma_y \qquad \mathbb{E}_\mu[(Z - \mu_Z)(Z - \mu_Z)^T] = LL^T \tag{61}$$

for some lower triangular matrix $L \in \mathbb{R}^{nm \times nm}$ and the given covariance matrix Σ_y for the observations. We can then, rewrite the terms involving the decision variables z_i, as

$$-2\mu_y^T K^{xx} \mu_Z + \mathbb{E}_\mu[(Z(K^{xx})^{1/2})^T (K^{xx})^{1/2} Z] + \mathbb{E}_\mu[(K^{xx} Z)^T K^{xx} Z] \tag{62}$$

$$-2\mathbb{E}_\mu[(Y - \mu_y)^T K^{xx} (Z - \mu_Z)] \tag{63}$$

and using the properties of Gaussian random vectors under affine transformations, we have,

$$\mathbb{E}_\mu[(K^{xx} Z)^T K^{xx} Z] = \mu_Z^T K^{xx} K^{xx} \mu_Z + \text{trace}(K^{xx} L L^T K^{xx}) \tag{64}$$

$$\mathbb{E}_\mu[(Z(K^{xx})^{1/2})^T (K^{xx})^{1/2} Z] = \mu_Z^T K^{xx} \mu_Z + \text{trace}(K^{xx} L L^T) \tag{65}$$

$$\mathbb{E}_\mu[(Y - \mu_y)^T K^{xx} (Z - \mu_Z)] = \text{trace}(K^{xx} L (\Sigma_y^{1/2})^T) \tag{66}$$

(66) follows from the fact that y and Z are jointly Gaussian under a common measure $\mu : \mathcal{B}(\mathcal{Y}) \to [0, 1]$ and are thus related to each other through the affine transformation

$$\begin{pmatrix} y(\zeta) \\ Z(\zeta) \end{pmatrix} = \begin{pmatrix} \Sigma_y^{1/2} \zeta \\ L\zeta \end{pmatrix} + \begin{pmatrix} \mu_y \\ \mu_Z \end{pmatrix}$$

(without loss of generality, taking μ to be the Gaussian measure for the standard normal distribution $\mathcal{N}(0, I)$ on $\mathbb{R}^n$)

Thus for the new decision variables $\mu_Z \in \mathbb{R}^{nm}$ and $L \in (\mathbb{R}^{nm \times nm})_{lt}$ (lower triangular matrix denoted with subscript lt), we can write the equivalent finite dimensional problem to (56) as,

$$\begin{aligned} {\mu_z}^{opt}, L^{opt} = \text{argmin}_{\mu_z \in \mathbb{R}^{nm}, L \in (\mathbb{R}^{nm \times nm})_{lt}} & (\mu_y - K^{xx} \mu_Z)^T (\mu_y - K^{xx} \mu_Z) + \mu_Z^T (K^{xx} K^{xx}) \mu_Z \\ & + \text{trace}(L^T K^{xx} L + (K^{xx} L - (\Sigma_y^{1/2}))^T (K^{xx} L - (\Sigma_y^{1/2}))) \end{aligned} \tag{67}$$

From (67) it is easy to see the problem is an unconstrained, convex quadratic program in μ_Z and L and thus has an unique minimizer.

The final function form of f_{opt} from (56), is then given by the affine transformation of the random vector Z^{opt} having mean μ_Z^{opt} and covariance $L^{opt}(L^{opt})^T$.

$$f^{opt}(\cdot) = K(\cdot, X) Z^{opt} \tag{68}$$

where $K(\cdot, X)$ is the matrix $\begin{pmatrix} K(\cdot, x_1) & K(\cdot, x_2) & \cdots & K(\cdot, x_m) \end{pmatrix}$.
Thus we have

$$\mathbb{E}_\mu[f_{opt}(\cdot)] = K(\cdot, X) \mu_Z^{opt}$$

and covariance,

$$\text{Covar}_\mu[f_{opt}(\cdot)] = K(\cdot, X) L^{opt} (L^{opt})^T K(\cdot, X)^T$$

Note that the mean coincides, as expected with the Bayesian posterior mean, however the covariance is quite different. Instead of acquiring certainty at points of observations, the regression model tries to fit the Gaussian process to the specified covariances of the observations.

Figure 2 shows an example for such a regression with a squared exponential kernel mapping with the output $y_i \in \mathcal{Z}$ being a two dimensional Gaussian random vector and $x_i \in \mathbb{R}$.

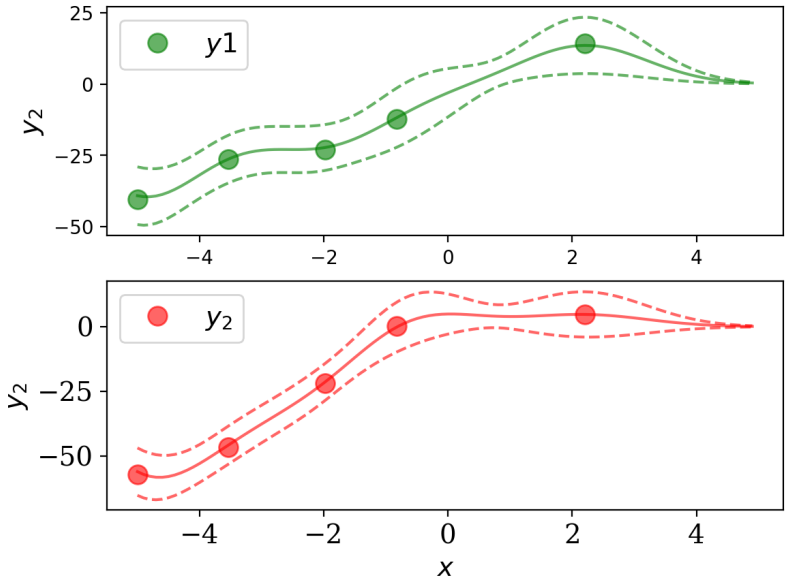

Figure 2: Least squares regression in an RKHS of Gaussian processes

Note that while we restricted our Banach space $\mathcal{Z}_\mu$ in the beginning to a space of Gaussian processes, there is no restriction from the point of view of the representer theorem, requiring Gaussianity. The above process can in principle be repeated for any given Banach space of stochastic processes (including non-Gaussian ones) and appropriate linear operator (as the evaluation operator may not be linear for non Gaussian cases). We limit ourselves to Gaussian processes in this case as it leads to simple analytically tractable computations. Also while we restricted ourselves to a simple regression problem, note that by virtue of the generalized representer theorem we can apply the above process to many other loss functionals and regularizers to create stochastic variants of any kernel based learning algorithms like the SVM, or the neural network example from Section 4.1, where the RKHS space of Gaussian processes alongside a moment matching constraint between the layers can be considered, to create a Gaussian process variant for the neural network example.

The example is left limited to this simple case, as it demonstrates the key issue being considered, which is the utility of extending the loss functional to $C : \mathcal{Z} \to \mathbb{R} \cup \{\infty\}$ for arbitrary separable Hilbert spaces $\mathcal{Z}$, like the Hilbert space of measurable functions $\mathcal{Y}_\mu$ considered above.

4.3 ℓ_1-Regularization

4.3.1 Motivating finite dimensional example

Consider first an example of the ℓ_1-regularization problem in a finite dimensional decision space. Let $\mathcal{X} = \mathbb{R}^l$, $\mathcal{Y} = \mathbb{R}^{n\times k}$, $\mathcal{Z} = \mathbb{R}^k$ and $\mathcal{H} = \mathbb{R}^n$. Let $\phi : \mathcal{X} \to \mathcal{Y}$ be a given collection of features and let $\{e_1, \ldots, e_n\}$ be the standard basis for $\mathbb{R}^n$. Consider the continuous linear operator $L_{x,\phi} : \mathcal{H} \to \mathcal{Z}$ from Example 2(a), where $L_{x,\phi}(w) = \phi(x)^T w$. Then consider the ℓ_1-regularization problem for feature selection given a set of observations $\mathcal{D} = \{(x_i, y_i) :$

$x_i \in \mathcal{X}, y_i \in \mathcal{Z}, i = 1, \ldots, m\}$ given by,

$$\min_{w \in \mathcal{H}} \quad \sum_{i=1}^{m} ||y_i - L_{x_i,\phi} w||_{\mathcal{Z}}^2 + \lambda ||w||_1^2 \tag{69}$$

where the $||w||_1 = (\sum_{i=1}^{n} |w_i|)$ is the standard ℓ_1-norm on $\mathbb{R}^n$. Let w_i denote the i^{th} component of a vector $w \in \mathbb{R}^n$. From Theorem 17 we know that the ℓ_1 norm is orthomonotone with respect to a subspace valued map $S_{proj} : \mathcal{H} \to \mathcal{V}(\mathcal{H})$, given by

$$S_{proj}(w) = \left\{ \sum_{i=1}^{n} \lambda_i \langle w, e_i \rangle_{\mathcal{H}} e_i : \lambda_i \in \mathbb{R} \right\} = \left\{ \sum_{\{i : w^T e_i \neq 0\}} \lambda_i e_i : \lambda_i \in \mathbb{R} \right\} \tag{70}$$

Example 3 showed that S_{proj} is an inclusive, quasilinear subspace valued map with the union extension $S_{proj} : \mathcal{V}(\mathcal{H}) \to \mathcal{V}(\mathcal{H})$. Then from the representer theorem (Theorem 18) we know that a minimizer for (69) must exist in $S_{proj}(\sum_{i=1}^{m} \text{range}(L^*_{x_i,\phi})) = S_{proj}(\{\sum_{i=1}^{m} L^*_{x_i,\phi} z_i : z_i \in \mathbb{R}^k\})$.

From Example 2(a), we also know that $L^*_{x_i,\phi} z_i = \phi(x_i) z_i$. Thus we have

$$S_{proj}\left(\sum_{i=1}^{m} \text{range}(L^*_{x_i,\phi}) \right) = \sum_{i=1}^{m} S_{proj}(\text{range}(L^*_{x_i,\phi})) = \sum_{i=1}^{m} S_{proj}(\{\phi(x_i) z_i : z_i \in \mathbb{R}^k\}) \tag{71}$$

$$= \sum_{i=1}^{m} \left\{ \sum_{\{j : \phi(x_i)^T e_j \neq 0\}} \lambda_j e_j : \lambda_j \in \mathbb{R} \right\} \tag{72}$$

$$= \left\{ \sum_{\{j :\ \phi(x_i)^T e_j \neq 0\ \forall i = 1, \ldots, m\}} \lambda_j e_j : \lambda_j \in \mathbb{R} \right\} \tag{73}$$

Substituting this form of the minimizer into (69), we can then find the optimal λ_js. The above problem is often used as a means for sparse feature selection in learning problems.

The subspace valued map S_{proj} defined above is a $n-$regular subspace valued map as it is quasilinear, idempotent, inclusive and $S_{proj}(w)$ for any $w \in \mathcal{H}$ has dimension at most n. However if we let $n \to \infty$, S_{proj} will lose the $r-$regularity property. This does not however mean that the representer for the case of $n \to \infty$ will be infinite dimensional. In fact since the dimension of $\sum_{i=1}^{m} \text{range}(L^*_{x_i,\phi})$ is at most m, the dimension for the representer is at most $\max\{n, m\}$, even when S_{proj} is not $r-$regular for any finite r, i.e., for any $n > m$, the representer dimension is limited to m.

We show below an example of ℓ_1 regularization in an infinite dimensional space ($n = \infty$) and show an example of applying the representer theorem to a problem with a non $r-$regular subspace valued map.

4.3.2 A non $r-$regular example

To show an application of a non $r-$regular subspace valued map, consider an analogue of the finite dimensional example presented above over an infinite dimensional Hilbert space. For this purpose, let $\mathcal{X} = \mathbb{N}$ be the set of natural numbers and $\mathcal{Z} = \mathbb{R}$. Let $\mathcal{H} = \ell^2(\mathbb{N}, \mathbb{R})$

be the space of square summable sequences taking values in $\mathbb{R}$. For any sequence $f \in \mathcal{H}$, let $f(i)$ denote the i^{th} member of the sequence f and let $||f||_2 = (\sum_{i\in\mathbb{N}} |f(i)|^2)^{1/2} < \infty$ be the ℓ_2 norm. Let $\langle f, g\rangle_{\mathcal{H}} = \sum_{i\in\mathbb{N}} f(i)g(i)$ be the inner product on $\mathcal{H}$. Let $\langle z_1, z_2\rangle_{\mathcal{Z}} = z_1 z_2$ be the scalar product on $\mathcal{Z} = \mathbb{R}$.

As an analogue to the orthonormal basis in $\mathbb{R}^n$, consider a set of orthonormal basis for $\mathcal{H}$ given by $\{\delta_i \in \mathcal{H} : i \in \mathbb{N}\}$ with δ_i defined as $\delta_i(j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$. The above space of ℓ_2 functions forms a separable Hilbert space as shown by (Rudin, 1964, Riesz-Fischer Theorem).

For all $f \in \mathcal{H}$, let $||f||_1 = \sum_{i\in\mathbb{N}} |f(i)|$ denote the ℓ_1 norm for the sequence. If a sequence $f \in \mathcal{H}$ is not absolutely summable, i.e. $\sum_{i\in\mathbb{N}} |f(i)|$ is not bounded, then we set $||f||_1 = \infty$.

Further note that the evaluation operator $L_x : \mathcal{H} \to \mathcal{Z}$ defined as $L_x f = f(x)$ for any $x \in \mathbb{N}$ is a bounded linear operator on $\ell_2(\mathbb{N}, \mathbb{R})$ with the adjoint L_x^* given by $\delta_x(\cdot)$, since for all $z \in \mathbb{R}$, $\langle z, L_x f\rangle_{\mathcal{Z}} = zf(x) = \langle z\delta_x, f\rangle_{\mathcal{H}} = \langle L_x^* z, f\rangle_{\mathcal{H}}$.
Then for the problem,

$$\min_{f\in\mathcal{H}} \quad \sum_{i=1}^{m} ||y_i - L_{x_i} f||_{\mathcal{Z}}^2 + \lambda ||f||_1^2 \tag{74}$$

we have $\Omega : \mathcal{H} \to \mathbb{R} \cup \{\infty\}$ given by $\Omega(f) = ||f||_1^2$. The functional Ω is orthomonotone with respect to the subspace valued map

$$S_{proj}(f) = \left\{ \sum_{i=1}^{\infty} \lambda(i) \frac{\langle f, \delta_i\rangle_{\mathcal{H}} \delta_i}{||f||_{\mathcal{H}}} : \lambda \in \ell^2(\mathbb{N}, \mathbb{R}) \right\}$$

Example 4 shows that $S_{proj} : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ defined above is an inclusive, quasilinear and superadditive subspace valued map with a union extension $S_{proj} : \mathcal{V}(\mathcal{H}) \to \mathcal{V}(\mathcal{H})$. Theorem 17 shows that $\Omega(f) = \begin{cases} ||f||_1^2 & \sum_{i=1}^{\infty} |f(i)| < \infty \\ +\infty & \text{otherwise} \end{cases}$ is orthomonotone with respect to the S_{proj} defined above.

Note also that $S_{proj}(f)$ in general can be infinite dimensional and thus is not r-regular for any finite r. However by Theorem 19 we know the minimizer for (74) must be of the form

$$S_{proj}\left(\left\{\sum_{i=1}^{m} L_{x_i}^* z_i : z_i \in \mathbb{R}\right\}\right) = S_{proj}\left(\left\{\sum_{i=1}^{m} \delta_{x_i}(\cdot) z_i : z_i \in \mathbb{R}\right\}\right) = \left\{\sum_{i=1}^{m} \delta_{x_i}(\cdot) z_i : z_i \in \mathbb{R}\right\}$$

The above representer can then be substituted for f in (74) and the optimization can be posed as a finite dimension optimization over $\{z_1, \ldots, z_m\}$. Thus (74) provides an example of problems where a non r-regular subspace valued map is required and thus was not be covered by previous counterparts of the generalized representer theorem. Also note that the non $r-$regularity of S_{proj} does not lead to an infinite dimensional representer as the dimension of the space is limited by the range of the adjoint.

5. Conclusion

We presented here an extension to existing work on generalized representer theorems by extending the result to apply to learning arbitrary Hilbert space-valued function spaces

with loss functionals composed with closed, densely defined operators on separable Hilbert spaces. Subspace valued maps with a super additive property were introduced and the property was shown to be necessary and sufficient for preserving a vector space structure for the union extension of a subspace valued map. The assumption of "r-regularity" was removed from the generalized theorem in order to allow more general subspace valued maps and its implications were shown for the ℓ_1 regularization problem in function spaces. The formalism of linear operators and adjoints was introduced into the generalized representer theorem and infinite dimensional representer spaces were treated as part of the result. The ℓ_1 norm was shown to be orthomonotone with respect to a projection based subspace valued map that shows the sparsity inducing nature of the ℓ_1 norm regularizers. An example from regression in a space of stochastic processes was shown to demonstrate the utility of the theorem when dealing with loss functionals on infinite dimensional Hilbert spaces and linear operators from one infinite dimensional Hilbert space to another. Finally, an example from kernel based neural networks was presented to show an approximation scheme based on the representer theorem to a kernel based neural network.

6. Appendix

6.1 Subspace Valued Maps

Definition 25 *(Quasilinear map)*
A subspace valued map $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ is called ***quasilinear*** *if*

$$\forall x, y \in \mathcal{H}, \lambda_1, \lambda_2 \in \mathbb{R}, \qquad S(\lambda_1 x + \lambda_2 y) \subseteq S(x) + S(y)$$

For any $A \in \mathcal{V}(\mathcal{H})$, let $S(A) = \cup_{x \in A} S(x)$. Then idempotence can be defined as,

Definition 26 *(Idempotent map)*
A map $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ is called ***idempotent*** *if*

$$\forall x \in \mathcal{H}, \qquad S(S(x)) = S(x)$$

Definition 27 *($r-$regular maps)*
*For some $r \in \mathbb{N}$, we call a map $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$, r-**regular** if*

1. *it is quasilinear and idempotent*
2. *for all $a \in U$, dimension of $S(a)$ is at most r*
3. *$\forall x \in \mathcal{H}$, $x \in S(x)$*

Lemma 28 *(Summation of subspace valued maps are super-additive)*
Let $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ be a subspace valued map. Let $S_{sup}(A) = \sum_{x \in A} S(x)$. Then for any $A, B \in \mathcal{V}(\mathcal{H})$, we have $S_{sup}(A) + S_{sup}(B) \subseteq S_{sup}(A + B)$.
Proof *The proof follows directly from the definition of S_{sup}, $S_{sup}(A)+S_{sup}(B) = \sum_{x \in A} S(x)+ \sum_{y \in B} S(y) = \sum_{x \in A \cup B} S(x)$. For vector spaces $A, B \in \mathcal{V}(\mathcal{H})$, we must have $A \cup B \subseteq A + B$. Thus $S_{sup}(A) + S_{sup}(B) = \sum_{x \in A \cup B} S(x) \subseteq \sum_{x \in A+B} S(x) = S_{sup}(A + B)$.* ■

The following example shows how addition of sets works in practice and shows a non $r-$regular example of a super-additive subspace valued map.

Example 7 *(Summation of subspace valued maps are super-additive)*
Let $\mathcal{H} = \ell^2(\mathbb{N})$ be the space of square summable sequences and let $||a||_\ell^2 = (\sum_{i=1}^\infty a_i^2)^{1/2}$. Consider the non r-regular subspace valued map $S^{proj} : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ given by

$$S^{proj}(a) = \begin{cases} \{\sum_{i=1}^\infty \lambda_i \frac{\langle a, \delta_i \rangle_\mathcal{H}}{||a||_{\ell^2}} \delta_i : \{\lambda_i\} \in \ell^2(\mathbb{N})\} & , ||a||_{\ell^2} \neq 0, \\ \{0\} & , otherwise \end{cases}$$

where $\delta_i(j) = 1$ for $j = i$ and 0 elsewhere. Let $S_{sup}^{proj}(A) = \sum_{a \in A} S^{proj}(a)$. Consider the subspaces $A = \{\sum_{i=1}^\infty \lambda_{2i}\delta_{2i} : \lambda_i \in \ell^2(\mathbb{N})\}$ and $B = \{\sum_{i=1}^\infty \lambda_{3i}\delta_{3i} : \lambda_i \in \ell^2(\mathbb{N})\}$. We have $S_{sup}^{proj}(A) = \{\sum_{i=1}^\infty \lambda_{2i}\delta_{2i} : \lambda_i \in \ell^2(\mathbb{N})\} = A$ and likewise $S_{sup}^{proj}(B) = B$ and $S_{sup}^{proj}(A+B) = A+B$. $S_{sup}^{proj}(A) + S_{sup}^{proj}(B) = A + B = \{\sum_{i=1}^\infty \lambda_{2i}\delta_{2i} + \lambda'_{3i}\delta_{3i} : \lambda_i, \lambda'_i \in \ell^2(\mathbb{N})\} = A + B = S_{sup}^{proj}(A+B)$ (equality trivially implies the inclusion required for super-additivity).

Note that the representers in Argyriou and Dinuzzo (2014) are given as $\sum_{i=1}^m S(w_i)$ for some r-regular subspace valued map $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$. The consideration of super-additive subspace valued maps does not lead to any loss of generality as we can consider the map $S_{sup}(A) = \sum_{x \in A} S(x)$ as given by the above lemma as our super-additive subspace valued map and then the representer is equivalently written as $S_{sup}(\text{span}\{w_1, \ldots, w_m\}) = \sum_{i=1}^m S(w_i)$. Note also that the super-additivity of S_{sup} does not contradict the sub-additive property of S required by quasi linearity, as we are considering S_{sup} as a new subspace valued map, entirely different from S, thus while S may be sub-additive, its summation S_{sup} is super-additive (in fact additive, in such a case, as shown below).

Lemma 29 *(Summation of quasilinear maps is additive)*
Let $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ be a quasilinear subspace valued map. Let $S_{sup}(A) = \sum_{x \in A} S(x)$ be the corresponding summation map defined as $S_{sup} : \mathcal{V}(\mathcal{H}) \to \mathcal{V}(\mathcal{H})$. Then S_{sup} is additive, i.e., for any $A, B \in \mathcal{V}(\mathcal{H})$, $S(A) + S(B) = S(A+B)$.
Proof $S_{sup}(A) + S_{sup}(B) = \sum_{x \in A} S(x) + \sum_{y \in B} S(y) = \sum_{x,y \in A \cup B} S(x) + S(y) \supseteq \sum_{x,y \in A \cup B} S(x+y) = S_{sup}(A+B)$. *Thus using a quasilinear S we get $S_{sup}(A) + S_{sup}(B) \supseteq S_{sup}(A+B)$. From Lemma 28, we already have $S_{sup}(A) + S_{sup}(B) \subseteq S_{sup}(A+B)$. Thus combining the two results, we have additivity, $S_{sup}(A) + S_{sup}(B) = S_{sup}(A+B)$.* ■

Example 8 *(A non-idempotent, non-r-regular, subspace valued map)*
Let $E_m = \{e_1, \ldots, e_m\}$ be the standard orthonormal basis for $\mathbb{R}^m$ and $E_{mn} = \{e_{11}, \ldots, e_{mn}\}$ be the standard orthonormal basis for $\mathbb{R}^{m \times n}$. Let $\mathcal{H}$ be a Hilbert space of $\mathbb{R}^m$-valued smooth, square integrable polynomial functions supported on $[-1,1]^n \subseteq \mathbb{R}^n$ with the Legendre polynomials, given as

$$\{p_{ij}e_i \in \mathcal{H} : p_{ij}(x) = c_j \partial_{x_i}^j [(x_i^2 - 1)^j], c_j = (j + 0.5)^{\frac{1}{2}} (2^j j!)^{-1}, j \in \mathbb{N}, e_i \in E_m, x_i = \langle x, e_i \rangle_{\mathbb{R}^n}\}$$

as the orthonormal basis for $\mathcal{H}$, where p_{ij} is a polynomial of order j. Let $\mathcal{Y}$ be the space of $\mathbb{R}^{m \times n}$-valued functions and $\nabla : \mathcal{H} \to \mathcal{Y}$ be the Jacobian operator, computing the Jacobian

for a $\mathbb{R}^m$-valued function. Let $\ell^2(\{1,\ldots,m\}\times\mathbb{N})$ be the space of dual indexed sequences $\{\lambda_i j : i \in \{1,\ldots,m\}, j \in \mathbb{N}\}$ that are square summable. Consider the subspace valued map $S_{proj} : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ given by,

$$S_{proj}(a) = \begin{cases} \left\{ \sum_{j=0}^{\infty} \sum_{i=1}^{m} \lambda_{ij} \frac{\langle a e_i, p_{ij} e_i \rangle_{\mathcal{H}} e_i}{||a||_{\mathcal{H}} ||p_{ij}||_{\mathcal{H}}} : \lambda_{ij} \in \ell^2(\{1,\ldots,m\}\times\mathbb{N}) \right\} & , ||a||_{\mathcal{H}} \neq 0 \\ \{0\} & , otherwise \end{cases}$$

Let for a matrix valued function $f \in \mathcal{Y}$, let f_i denote the i^{th} row of the matrix. Let $\nabla\cdot$ be the divergence operator and $\nabla^* : \mathcal{Y} \to \mathcal{H}$ be the adjoint operator to ∇, given as $\nabla^* f = -(\nabla \cdot f_1, \ldots, \nabla \cdot f_m)$. A subspace valued map $S' : \mathcal{Y} \to \mathcal{V}(\mathcal{Y})$ is then induced by the jacobian operator ∇ given as,

$$S'(f) = \nabla(S_{proj}(\nabla^* f))$$

Let $f_n \in \mathcal{Y}$ denote a polynomial of maximum order n. Then note that $S'(f_n)$ contains polynomials of order at most $n-2$. Thus clearly $f_n \notin S'(f_n)$. Thus S' is not inclusive.

Also $S'(S'(f_n))$ contains polynomials of order at most $n-4$ and thus $S'(S'(f_n)) \neq S'(f_n)$, implying S' is not idempotent. Also in general $f \in \mathcal{Y}$ can be an infinite order polynomial and thus the dimension of $S'(f)$ can be infinity.

Note that $S' : \mathcal{Y} \to \mathcal{V}(\mathcal{Y})$ is however a quasilinear, super-additive subspace valued map and can still be used to establish a representer theorem, provided it is range preserving with respect to the L being used with the loss functional. An example of such an operator would be the smooth kernel of an RKHS defined on $\mathcal{H}$. Since range(L^*) *then contains polynomials of order upto infinity, $\mathcal{N}_L^{\perp} \subseteq S'(\mathcal{N}_L^{\perp})$.*

References

Andreas Argyriou and Francesco Dinuzzo. A unifying view of representer theorems. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages II–748–II–756. JMLR.org, 2014. URL `http://dl.acm.org/citation.cfm?id=3044805.3044976`.

Andreas Argyriou, Charles A Micchelli, and Massimiliano Pontil. When is there a representer theorem? vector versus matrix regularizers. *Journal of Machine Learning Research*, 10(Nov):2507–2529, 2009.

N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc. 68 (1950), 337-404*, 1950. doi: https://doi.org/10.1090/S0002-9947-1950-0051437-7.

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 0387310738.

V.I. Bogachev. *Gaussian Measures*. Mathematical Surveys and Monographs. American Mathematical Society, 2015. ISBN 9781470418694. URL `https://books.google.ch/books?id=UtufBwAAQBAJ`.

Youngmin Cho and Lawrence K. Saul. Kernel methods for deep learning. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 342–350. Curran Associates, Inc., 2009. URL http://papers.nips.cc/paper/3628-kernel-methods-for-deep-learning.pdf.

J.B. Conway. *A Course in Abstract Analysis*. Graduate studies in mathematics. American Mathematical Soc. ISBN 9780821891599. URL https://books.google.be/books?id=GD7QxvMOFUcC.

J.B. Conway. *A Course in Functional Analysis*. Graduate Texts in Mathematics. Springer New York, 1994. ISBN 9780387972459.

Andreas C. Damianou and Neil D. Lawrence. Deep gaussian processes. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2013, Scottsdale, AZ, USA, April 29 - May 1, 2013*, pages 207–215, 2013. URL http://jmlr.org/proceedings/papers/v31/damianou13a.html.

F. Dinuzzo and B. Schölkopf. The representer theorem for hilbert spaces: a necessary and sufficient condition. In *Advances in Neural Information Processing Systems 25*, pages 189–196. Curran Associates Inc., 2012.

Ivan Dobrakov. On integration in banach spaces, vii. *Czechoslovak Mathematical Journal*, 38(3):434–449, 1988.

SH Kulkarni and MT Nair. A characterization of closed range operators. *Indian Journal of Pure and Applied Mathematics*, 31(4):353–362, 2000.

SH Kulkarni, MT Nair, and G Ramesh. Some properties of unbounded operators with closed range. *Proceedings Mathematical Sciences*, 118(4):613–625, 2008.

Charles A. Micchelli and Massimiliano A. Pontil. On learning vector-valued functions. *Neural Comput.*, 17(1):177–204, January 2005. ISSN 0899-7667. doi: 10.1162/0899766052530802. URL http://dx.doi.org/10.1162/0899766052530802.

Hà Quang Minh and Vikas Sindhwani. Vector-valued manifold regularization. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 57–64, USA, 2011. Omnipress. ISBN 978-1-4503-0619-5. URL http://dl.acm.org/citation.cfm?id=3104482.3104490.

Hà Quang Minh, Loris Bazzani, and Vittorio 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(25):1–72, 2016. URL http://jmlr.org/papers/v17/14-036.html.

Christopher K. I. Williams Rasmussen, Carl Edward. *Gaussian processes for machine learning*. MIT Press, 2006.

Ilyes Rebai, Yassine BenAyed, and Walid Mahdi. Deep multilayer multiple kernel learning. *Neural Computing and Applications*, 27(8):2305–2314, Nov 2016. ISSN 1433-3058. doi: 10.1007/s00521-015-2066-x. URL https://doi.org/10.1007/s00521-015-2066-x.

W. Rudin. *Principles of mathematical analysis*. International series in pure and applied mathematics. McGraw-Hill, 1964. URL https://books.google.ch/books?id=iifvAAAAMAAJ.

Adrian Sandovici. Von neumann's theorem for linear relations. *Linear and Multilinear Algebra*, 66(9):1750–1756, 2018. doi: 10.1080/03081087.2017.1369930. URL https://doi.org/10.1080/03081087.2017.1369930.

Bernhard Schölkopf, Ralf Herbrich, and Alex J. Smola. A generalized representer theorem. In *Proceedings of the 14th Annual Conference on Computational Learning Theory and and 5th European Conference on Computational Learning Theory*, COLT '01/EuroCOLT '01, pages 416–426, London, UK, UK, 2001. Springer-Verlag. ISBN 3-540-42343-5. URL http://dl.acm.org/citation.cfm?id=648300.755324.

Gideon Schwarz. Variations on vector measures. *Pacific Journal of Mathematics*, 23(2): 373–375, 1967.

Arjun Sudan, OW van Gaans, and PJC Spreij. Infinite order sobolev spaces and the schwartz space. 2012.

Johan A.K. Suykens, Carlos Alzate, and Kristiaan Pelckmans. Primal and dual model representations in kernel-based learning. *Statist. Surv.*, 4:148–183, 2010. doi: 10.1214/09-SS052. URL https://doi.org/10.1214/09-SS052.

A. N. Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Dokl.*, 4:1035–1038, 1963.

Michael Unser, Julien Fageot, and Harshit Gupta. Representer theorems for sparsity-promoting ℓ_1 regularization. *IEEE Transactions on Information Theory*, 62(9):5167–5180, 2016.

G. Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, 1990. doi: 10.1137/1.9781611970128. URL http://epubs.siam.org/doi/abs/10.1137/1.9781611970128.

K. Yoshida. *Functional Analysis*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2013. ISBN 9783662257623. URL https://books.google.ch/books?id=2Mb3CAAAQBAJ.

A.C. Zaanen. *Introduction to Operator Theory in Riesz Spaces*. Springer Berlin Heidelberg, 2012. ISBN 9783642606373. URL https://books.google.ch/books?id=cgvpCAAAQBAJ.