ARTICLE IN PRESS

Journal of Complexity xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of Complexity

journal homepage: www.elsevier.com/locate/jco

Multi-task Learning in vector-valued reproducing kernel Banach spaces with the ℓ^1 norm☆

Rongrong Lin [a,1], Guohui Song [b,2], Haizhang Zhang [c,*,3]

[a] School of Data and Computer Science, Sun Yat-sen University, Guangzhou, PR China
[b] Department of Mathematics and Statistics, Old Dominion University, 2300 Engineering & Computational Sciences Building, Norfolk, VA 23529, United States of America
[c] School of Data and Computer Science, and Guangdong Province Key Laboratory of Computational Science, Sun Yat-sen University, Guangzhou, PR China

ARTICLE INFO

Article history:
Received 6 October 2019
Received in revised form 18 April 2020
Accepted 27 August 2020
Available online xxxx

Keywords:
Reproducing kernel Banach spaces
Admissible multi-task kernels
Lebesgue constants
Representer theorems

ABSTRACT

Targeting at sparse multi-task learning, we consider regularization models with an ℓ^1 penalty on the coefficients of kernel functions. In order to provide a kernel method for this model, we construct a class of vector-valued reproducing kernel Banach spaces with the ℓ^1 norm. The notion of multi-task admissible kernels is proposed so that the constructed spaces could have desirable properties including the crucial linear representer theorem. Such kernels are related to bounded Lebesgue constants of a kernel interpolation question. We study the Lebesgue constant of multi-task kernels and provide examples of admissible kernels. Furthermore, we present numerical experiments for both synthetic data and real-world benchmark data to demonstrate the advantages of the proposed construction and regularization models.

☆ Communicated by S. Pereverzyev.

* Corresponding author.
E-mail addresses: linrr@mail2.sysu.edu.cn (R. Lin), gsong@odu.edu (G. Song), zhhaizh2@mail.sysu.edu.cn (H. Zhang).

1 Supported in part by National Natural Science Foundation of China under grant 11901595.
2 Supported in part by NSF DMS-1521661 and DMS-1939203.
3 Supported in part by National Natural Science Foundation of China under grant 11971490, and by Natural Science Foundation of Guangdong Province under grant 2018A030313841.

https://doi.org/10.1016/j.jco.2020.101514
0885-064X/

Please cite this article as: R. Lin, G. Song and H. Zhang, Multi-task Learning in vector-valued reproducing kernel Banach spaces with the ℓ^1 norm, Journal of Complexity (2020) 101514, https://doi.org/10.1016/j.jco.2020.101514.

1. Introduction

Reproducing kernel Banach spaces (RKBSs) and their applications have attracted a lot of attention in the machine learning community [8,14,15,17,29,30,35,36,38,39]. In particular, RKBSs with the ℓ^1 norm [29,30] have proven to be useful in promoting sparsity in single task learning. The regularization models with an ℓ^1 penalty on the coefficients of kernel functions centered at the finite sample data were well understood in statistical learning theory [16,27–29]. It was proved in [30] that under certain conditions the ℓ^1-norm coefficient-based regularization over the finite sample-dependent space is equivalent to the one over the whole RKBSs with the ℓ^1 norm. On the other hand, vector-valued function spaces [1,6,23] provide a solid foundation for many models in multi-task learning. The purpose of this paper is to construct vector-valued RKBSs with the ℓ^1 norm and study regularization methods for multi-task learning in such spaces.

Reproducing kernel Hilbert spaces (RKHSs) are Hilbert spaces of functions on which point evaluation functionals are continuous [2]. In machine learning, RKHSs have been viewed as ideal spaces for kernel-based learning algorithms [9,27,31,34]. Thanks to the existence of an inner product, Hilbert spaces are well-understood in functional analysis. Most importantly, an RKHS has a reproducing kernel, which measures similarity between inputs and gives birth to the "kernel trick" in machine learning that significantly saves computations. Celebrated machine learning methods based on scalar-valued RKHSs include support vector machines and the regularization networks.

The RKBS is a recent and fast-growing research area. We mention two reasons that justify the need of RKBSs here. On one hand, Banach spaces possess richer geometrical structures and norms. It is standard knowledge in functional analysis that any two Hilbert spaces on a common number field of the same dimension are isometrically isomorphic to each other, and hence share the same norms and geometry. By contrast, for $1 \leq p \neq q \leq +\infty$, $L^p([0,1])$ and $L^q([0,1])$ are not isomorphic to each other. On the other hand, many important problems such as p-norm coefficient-based regularization [33], large-margin classification [12,36,37], lasso in statistics [32] and compressed sensing [4] had better be studied in Banach spaces.

There are various approaches to constructing scalar-valued RKBSs in the literature. For example, [37] employs the tool of semi-inner-products to build RKBSs and [35] constructs RKBSs based on certain feature mappings. In particular, a bilinear form has been used to develop RKBSs with the ℓ^1 norm in [30]. Moreover, a recent work [21] gives a unified definition of RKBSs that is more general than the aforementioned specific ones. It also proposed a unified framework of constructing scalar-valued RKBSs that covers all existing constructions [14,29,30,35,36,38,39] via a continuous bilinear form and a pair of feature maps.

We consider multi-task learning in this paper. Many real-world applications involve learning multiple tasks. A standard methodology in machine learning is to learn one task at a time. Large problems are hence broken into small and reasonably independent subproblems that are learned separately and then recombined. Multi-task learning where the unknown target function to be learned from finite sample data is vector-valued appears more often in practice [23]. Learning multiple related tasks simultaneously can be more beneficial. For instance, in certain circumstances, data for each task are not enough to avoid over-fitting and hence results in poor generalization ability. In this case, what is learned for each task can help other related tasks be learned better. There are numerical experiments in the literature [1,7] which demonstrate that multi-task learning can lead to better generalization performance than learning each task independently. Recent progress about multi-task learning in vector-valued RKHSs can be found in [5,6]. In such a framework, both the space of the candidate functions used for approximation and the output space are chosen as Hilbert spaces. The mathematical theory of learning on vector-valued RKBSs based on semi-inner-products has been proposed in [40]. The spaces considered there are reflexive and thus do not accommodate the ℓ^1 norm.

Motivated by sparse multi-task learning, we shall construct vector-valued RKBSs with the ℓ^1 norm in this paper. To ensure that the existence of a reproducing kernel, the construction starts directly with an admissible multi-task kernel satisfying three assumptions: non-singularity, boundedness, and independence. Then, we are able to obtain a vector-valued RKBS with the ℓ^1 norm and its associated reproducing kernel.

Please cite this article as: R. Lin, G. Song and H. Zhang, Multi-task Learning in vector-valued reproducing kernel Banach spaces with the ℓ^1 norm, Journal of Complexity (2020) 101514, https://doi.org/10.1016/j.jco.2020.101514.

Moreover, we will investigate the regularization model in such spaces. The classical linear representer theorem is a key to the mathematical analysis of kernel methods in machine learning [27]. It asserts that the minimizer is a linear combination of the kernel functions at the sampling points. The representer theorem in scalar-valued RKHSs was initially established by Kimeldorf and Wahba [20]. The result was generalized to other regularizers in [26]. Recent Refs. [14,21,35–37,39] developed representer theorems for various scalar-valued RKBSs. We shall present the representer theorem for machine learning schemes in vector-valued RKBSs with the ℓ^1 norm. We shall see that this is equivalent to requiring the Lebesgue constant of the admissible multi-task kernel to be exactly bounded by 1. To accommodate more kernels, we consider a relaxed representer theorem.

The outline of the paper is as follows. In Section 2, we present definitions of vector-valued RKBSs, the associated reproducing kernels, and admissible multi-task kernels. We next start constructing RKBSs of vector-valued functions with the ℓ^1 norm. Section 3 establishes representer theorems for minimal norm interpolation and regularization networks in the constructed spaces. Examples of admissible multi-task kernels are given in Section 4. To accommodate more kernel functions, a relaxed version of the linear representer theorem is discussed in Section 5. The relaxed representer theorem plays an important role in learning rate estimates. This is illustrated in the section. In Section 6, numerical experiments for both synthetic data and real-world benchmark data are presented to demonstrate the advantages of the proposed construction and regularization models in the vector-valued RKBSs with the Laplacian kernel and the exponential kernel.

2. Construction of vector-valued RKBSs with the ℓ^1 norm

We shall present the construction of vector-valued RKBSs with the ℓ^1 norm in this section. Specifically, we will first introduce the definition of general vector-valued RKBSs and then construct the specific vector-valued RKBSs with the ℓ^1 norm.

To give a formal definition of vector-valued RKBSs in our setting, we first review the definition of Banach spaces of vector-valued functions. A normed vector space V of functions from X to $Y \subseteq \mathbb{R}^d$ is called a *Banach space of vector-valued functions* if it is a Banach space whose elements are vector-valued functions on X and for each $f \in V$, $\|f\|_V = 0$ if and only if $f(x) = \mathbf{0}$ for all $x \in X$. Here, $\mathbf{0}$ denotes the zero vector of $\mathbb{R}^d$. For instance, $L^p([0, 1])$, $1 \leq p < +\infty$ is not a Banach space of functions while $C([0, 1])$ is. The definition of general vector-valued RKBSs is presented below.

Definition 2.1 (*Vector-valued RKBS*)**.** Let X be a prescribed nonempty set, and let Y be a Banach space. A vector-valued RKBS $\mathcal{B}$ of functions from X to Y is a Banach space of certain vector-valued functions $f : X \to Y$ such that every point evaluation functional δ_x, $x \in X$ on $\mathcal{B}$ is continuous. That is, for any $x \in X$, there exists a constant $C_x > 0$ such that

$$\|\delta_x(f)\|_Y = \|f(x)\|_Y \leq C_x \|f\|_{\mathcal{B}} \text{ for all } f \in \mathcal{B}.$$

Definition 2.1 is a natural "vectorized" generalization of the scalar-valued RKHS [2,27] and the scalar-valued RKBS in [21]. In [23], a Hilbert space $\mathcal{H}$ from X to a Hilbert space Y with inner product $\langle \cdot, \cdot \rangle_Y$ is called an RKHS of vector-valued functions if for any $y \in Y$ and $x \in X$, the linear functional which maps $f \in \mathcal{H}$ to $\langle y, f(x) \rangle_Y$ is continuous. With the tool of semi-inner product, Ref. [40] initially proposed the notion of vector-valued RKBS for multi-task learning in 2013. The prerequisite is that $\mathcal{B}$ and Y are uniform Banach spaces. Those requirements more or less seem unnatural. We are able to remove them by exploiting the definition of reproducing kernels via continuous bilinear forms.

We remark that there are no kernels directly mentioned in the above definition of the general vector-valued RKBSs. We will introduce a definition of the associated reproducing kernel through bilinear forms. Recall that a bilinear form between two normed vector spaces V_1 and V_2 is a function $(\cdot, \cdot)_{V_1 \times V_2}$ from $V_1 \times V_2$ to $\mathbb{R}$ that is linear about both arguments. It is said to be a *continuous bilinear form* if there exists a positive constant C such that

$$|(f, g)_{V_1 \times V_2}| \leq C \|f\|_{V_1} \|g\|_{V_2} \text{ for all } f \in V_1, g \in V_2.$$

For practical applications, it is natural and sufficient to restrict the output space to be the multi-dimensional Euclidean space. Thus, from now on, we assume the output space $Y = \mathbb{R}^d$. For any column vector $\boldsymbol{c} \in \mathbb{R}^d$, we denote by $\boldsymbol{c}^\top$ its transpose.

Please cite this article as: R. Lin, G. Song and H. Zhang, Multi-task Learning in vector-valued reproducing kernel Banach spaces with the ℓ^1 norm, Journal of Complexity (2020) 101514, https://doi.org/10.1016/j.jco.2020.101514.

Definition 2.2 (*Reproducing Kernel*)**.** Let X be a nonempty set, and let $\mathcal{B}$ be a vector-valued RKBS from X to $\mathbb{R}^d$. If there exist a Banach space $\mathcal{B}^\#$ of vector-valued functions from X to $\mathbb{R}^d$, a continuous bilinear form $(\cdot,\cdot)_{\mathcal{B}\times\mathcal{B}^\#}$, and a matrix-valued function $\mathbf{K}: X\times X\to\mathbb{R}^{d\times d}$ such that $\mathbf{K}(\cdot,x)\boldsymbol{c}\in\mathcal{B}^\#$ for all $x\in X$ and $\boldsymbol{c}\in\mathbb{R}^d$, and

$$(f,\mathbf{K}(\cdot,x)\boldsymbol{c})_{\mathcal{B}\times\mathcal{B}^\#}=f(x)^\top\boldsymbol{c}\quad\text{for all } x\in X,\boldsymbol{c}\in\mathbb{R}^d,f\in\mathcal{B},\tag{2.1}$$

then we call $\mathbf{K}$ a reproducing kernel for $\mathcal{B}$. If in addition, $\mathcal{B}^\#$ is also a vector-valued RKBS, $\mathbf{K}(x,\cdot)\boldsymbol{c}\in\mathcal{B}$ for all $x\in X$ and $\boldsymbol{c}\in\mathbb{R}^d$, and

$$(\mathbf{K}(x,\cdot)\boldsymbol{c},g)_{\mathcal{B}\times\mathcal{B}^\#}=\boldsymbol{c}^\top g(x)\text{ for all } x\in X,\boldsymbol{c}\in\mathbb{R}^d,g\in\mathcal{B}^\#,\tag{2.2}$$

then we call $\mathcal{B}^\#$ an adjoint vector-valued RKBS of $\mathcal{B}$, and call $\mathcal{B}$ and $\mathcal{B}^\#$ a pair of vector-valued RKBSs. In the latter case, $\tilde{K}(x,x'):=K(x',x)$ for $x,x'\in X$, is a reproducing kernel for $\mathcal{B}^\#$.

We call (2.1) and (2.2) the *reproducing properties* for the kernel $\mathbf{K}$ in vector-valued RKBSs $\mathcal{B}$ and $\mathcal{B}^\#$.

We shall next construct the specific vector-valued RKBSs with the ℓ^1 norm satisfying the above conditions of general vector-valued RKBSs. The construction is built on certain multi-task kernels. To this end, we first introduce admissible multi-task kernels and some related notations. For any vector $\boldsymbol{u}$ and $p\in[1,\infty]$, we use $\|\boldsymbol{u}\|_p$ to denote the ℓ^p norm of $\boldsymbol{u}$. For any matrix A and $p\in[1,\infty]$, we use $\|\mathsf{A}\|_p$ to denote the ℓ^p-induced matrix norm of A. We denote for any nonempty set Ω by $\ell^1_d(\Omega)$ the Banach space of vector-valued functions on Ω that is integrable with respect to the counting measure on Ω. Specifically,

$$\ell^1_d(\Omega):=\Big\{\boldsymbol{c}=(\boldsymbol{c}_t\in\mathbb{R}^d:t\in\Omega):\|\mathbf{c}\|_{\ell^1_d(\Omega)}=\sum_{t\in\operatorname{supp}\boldsymbol{c}}\|\boldsymbol{c}_t\|_1<+\infty\Big\}.\tag{2.3}$$

where $\operatorname{supp}\boldsymbol{c}:=\{t\in\Omega:\boldsymbol{c}_t\neq\mathbf{0}\}$ denotes the support of $\boldsymbol{c}\in\ell^1_d(\Omega)$. Note that Ω might be uncountable, but for every $\mathbf{c}\in\ell^1_d(\Omega)$, the support $\operatorname{supp}\boldsymbol{c}$ must be at most countable. Let us denote $\mathbb{N}_m:=\{1,2,\dots,m\}$ for any $m\in\mathbb{N}$.

Definition 2.3 (*Admissible Multi-task Kernel*)**.** Let X be a nonempty set and let $\mathbf{K}: X\times X\to\mathbb{R}^{d\times d}$ be a matrix-valued function such that $\mathbf{K}^\top=\mathbf{K}$. Such a kernel is an admissible multi-task kernel if the following assumptions are satisfied:

(A1) (Non-singularity) for all $m\in\mathbb{N}$ and all pairwise distinct sampling points $\mathbf{x}=\{x_j:j\in\mathbb{N}_m\}\subseteq X$, the matrix

$$\mathbf{K}[\mathbf{x}]:=\Big[\mathbf{K}(x_k,x_j):j,k\in\mathbb{N}_m\Big]\in\mathbb{R}^{md\times md}$$

is non-singular;

(A2) (Boundedness) there exists $\kappa>0$ such that $\|\mathbf{K}(x,x')\|_1\le\kappa$ for all $x,x'\in X$;

(A3) (Independence) for all pairwise distinct points $x_j\in X$, $j\in\mathbb{N}$ and $(\boldsymbol{c}_j\in\mathbb{R}^d:j\in\mathbb{N})\in\ell^1_d(\mathbb{N})$, if $\sum_{j\in\mathbb{N}}\mathbf{K}(x_j,x)\boldsymbol{c}_j=\mathbf{0}$ for all $x\in X$ then $\boldsymbol{c}_j=\mathbf{0}$ for all $j\in\mathbb{N}$.

We now present the construction of vector-valued RKBSs with the ℓ^1 norm based on admissible multi-task kernels. Suppose that $\mathbf{K}: X\times X\to\mathbb{R}^{d\times d}$ is an admissible multi-task kernel defined above. We then define

$$\mathcal{B}_{\mathbf{K}}:=\Big\{\sum_{x\in\operatorname{supp}\boldsymbol{c}}\mathbf{K}(x,\cdot)\boldsymbol{c}_x:\boldsymbol{c}=(\boldsymbol{c}_x\in\mathbb{R}^d:x\in X)\in\ell^1_d(X)\Big\}\tag{2.4}$$

with the norm

$$\Big\|\sum_{x\in\operatorname{supp}\boldsymbol{c}}\mathbf{K}(x,\cdot)\boldsymbol{c}_x\Big\|_{\mathcal{B}_{\mathbf{K}}}:=\|\boldsymbol{c}\|_{\ell^1_d(X)},\tag{2.5}$$

where $\ell^1_d(X)$ is given as in (2.3). Observe that $\mathcal{B}_{\mathbf{K}}$ defined by (2.5) is a Banach space since $\mathcal{B}_{\mathbf{K}}$ is isometric to the Banach space $\ell^1_d(X)$. By (**A3**), we should point out that the norm given by (2.5) is

Please cite this article as: R. Lin, G. Song and H. Zhang, Multi-task Learning in vector-valued reproducing kernel Banach spaces with the ℓ^1 norm, Journal of Complexity (2020) 101514, https://doi.org/10.1016/j.jco.2020.101514.

well-defined. In other words, for any $f \in \mathcal{B}_{\mathbf{K}}$, $\|f\|_{\mathcal{B}_{\mathbf{K}}} = 0$ if and only if $f = \mathbf{0}$ everywhere on X. It follows that $\mathcal{B}_{\mathbf{K}}$ is a Banach space of functions on X.

We next show that the Banach space of functions $\mathcal{B}_{\mathbf{K}}$ defined above is a vector-valued RKBS on X according to Definition 2.1.

Proposition 2.4. *If* $\mathbf{K}$ *is an admissible multi-task kernel, then the space* $\mathcal{B}_{\mathbf{K}}$ *as defined in Eq. (2.4) is a vector-valued RKBS on* X *in the sense that*

$$\|f(x)\|_1 \le \kappa \|f\|_{\mathcal{B}_{\mathbf{K}}}, \text{ for all } x \in X, f \in \mathcal{B}_{\mathbf{K}}.$$

Proof. Note that $\ell_d^1(X)$ is a Banach space. By definition (2.5) of the norm on $\mathcal{B}_{\mathbf{K}}$, $\mathcal{B}_{\mathbf{K}}$ is a vector-valued Banach space on X. For any $f \in \mathcal{B}_{\mathbf{K}}$, there exists $\boldsymbol{c} \in \ell_d^1(X)$ such that

$$f = \sum_{t \in \operatorname{supp} \boldsymbol{c}} \mathbf{K}(t, \cdot)\boldsymbol{c}_t.$$

For any $x \in X$, by Assumption (**A2**) and Eq. (2.5), we compute

$$\|\delta_x(f)\|_1 = \|f(x)\|_1 \le \sum_{t \in \operatorname{supp} \boldsymbol{c}} \left\| \mathbf{K}(t, x)\boldsymbol{c}_t \right\|_1 \le \sum_{t \in \operatorname{supp} \boldsymbol{c}} \left\| \mathbf{K}(t, x) \right\|_1 \|\boldsymbol{c}_t\|_1 \le \kappa \sum_{t \in \operatorname{supp} \boldsymbol{c}} \|\boldsymbol{c}_t\|_1 = \kappa \|f\|_{\mathcal{B}_{\mathbf{K}}}$$

for all $f \in \mathcal{B}_{\mathbf{K}}$. In other words, the point evaluation functional δ_x, $x \in X$ is continuous on $\mathcal{B}_{\mathbf{K}}$ in the sense that $\|\delta_x(f)\|_1 \le \kappa \|f\|_{\mathcal{B}_{\mathbf{K}}}$ for all $f \in \mathcal{B}_{\mathbf{K}}$. The proof is hence complete. □

We next show that $\mathbf{K}$ is a reproducing kernel of $\mathcal{B}_{\mathbf{K}}$ through checking the conditions in Definition 2.2. For this purpose, we introduce an adjoint vector-valued RKBS $\mathcal{B}_{\mathbf{K}}^{\#}$ below. Let

$$\mathcal{B}_0^{\#} := \left\{ \sum_{k=1}^{n} \mathbf{K}(\cdot, x_k)\boldsymbol{b}_k : x_k \in X, \boldsymbol{b}_k \in \mathbb{R}^d, k \in \mathbb{N}_n \text{ for all } n \in \mathbb{N} \right\}$$

endowed with the supremum norm

$$\left\| \sum_{k=1}^{n} \mathbf{K}(\cdot, x_k)\boldsymbol{b}_k \right\|_{\mathcal{B}_0^{\#}} := \sup_{t \in X} \left\| \sum_{k=1}^{n} \mathbf{K}(t, x_k)\boldsymbol{b}_k \right\|_{\infty}.$$

Generally, an abstract completion of $\mathcal{B}_0^{\#}$ might not consist of functions. We point out that the completion of an incomplete Hilbert space of functions being RKHS was given in ([2], Pages 347–349). Motivated by this, we present a Banach completion process that yields a space of vector-valued functions. Suppose $\{g_n : n \in \mathbb{N}\}$ is a Cauchy sequence in $\mathcal{B}_0^{\#}$. Observe that point evaluation functionals δ_x, $x \in X$ are continuous on $\mathcal{B}_0^{\#}$ in the sense that

$$\|g(x)\|_{\infty} \le \|g\|_{\mathcal{B}_0^{\#}} \text{ for all } g \in \mathcal{B}_0^{\#}.$$

Consequently, for any $x \in X$, the sequence $\{g_n(x) : n \in \mathbb{N}\}$ converges in $\mathbb{R}^d$. We define the limit by $g(x)$, which is a vector-valued function on X. By definition of $\mathcal{B}_0^{\#}$, two equivalent Cauchy sequences in $\mathcal{B}_0^{\#}$ give the same function. We let $\mathcal{B}_{\mathbf{K}}^{\#}$ be consisting of all such limit functions g with the norm

$$\|g\|_{\mathcal{B}_{\mathbf{K}}^{\#}} := \lim_{n \to +\infty} \|g_n\|_{\mathcal{B}_0^{\#}} = \lim_{n \to +\infty} \sup_{x \in X} \|g_n(x)\|_{\infty}.$$

It follows that $\mathcal{B}_{\mathbf{K}}^{\#}$ is a Banach space of vector-valued functions and for any $g \in \mathcal{B}_{\mathbf{K}}^{\#}$,

$$\|g\|_{\mathcal{B}_{\mathbf{K}}^{\#}} := \sup_{x \in X} \|g(x)\|_{\infty}.$$

Based on the above construction, we immediately have that $\mathcal{B}_{\mathbf{K}}^{\#}$ defined above is also a vector-valued RKBS.

Proposition 2.5. *If* $\mathbf{K}$ *is an admissible multi-task kernel, then the space* $\mathcal{B}_{\mathbf{K}}^{\#}$ *is a vector-valued RKBSs in the sense that*

$$\|g(x)\|_{\infty} \le \|g\|_{\mathcal{B}_{\mathbf{K}}^{\#}} \text{ for all } x \in X, g \in \mathcal{B}_{\mathbf{K}}^{\#}.$$

Please cite this article as: R. Lin, G. Song and H. Zhang, Multi-task Learning in vector-valued reproducing kernel Banach spaces with the ℓ^1 norm, Journal of Complexity (2020) 101514, https://doi.org/10.1016/j.jco.2020.101514.

We next characterize the reproducing properties in $\mathcal{B}_{\mathbf{K}}$ and $\mathcal{B}_{\mathbf{K}}^{\#}$ via a bilinear form. To this end, we define the linear space $\mathcal{B}_0$ by

$$\mathcal{B}_0 := \Big\{ \sum_{j=1}^{m} \mathbf{K}(x_j, \cdot)\boldsymbol{c}_j : x_j \in X, \boldsymbol{c}_j \in \mathbb{R}^d, m \in \mathbb{N} \Big\}. \tag{2.6}$$

By (2.4) and (2.6), $\mathcal{B}_0$ is dense in the Banach space $\mathcal{B}_{\mathbf{K}}$. It follows that $\mathcal{B}_{\mathbf{K}}$ is a Banach completion of $\mathcal{B}_0$ under the ℓ^1 norm. We define a bilinear form $(\cdot,\cdot)_{\mathbf{K}}$ on $\mathcal{B}_0 \times \mathcal{B}_0^{\#}$ by

$$\Big(\sum_{j=1}^{m} \mathbf{K}(x_j, \cdot)\boldsymbol{c}_j, \sum_{k=1}^{n} \mathbf{K}(\cdot, s_k)\boldsymbol{b}_k \Big)_{\mathbf{K}} = \sum_{k=1}^{n} \sum_{j=1}^{m} \boldsymbol{c}_j^{\top} \mathbf{K}(x_j, s_k)\boldsymbol{b}_k, \ \ s_j, t_k \in X, \ \ \boldsymbol{c}_j, \boldsymbol{b}_k \in \mathbb{R}^d.$$

According to the norms on $\mathcal{B}_0$ and $\mathcal{B}_0^{\#}$, we obtain

$$\begin{aligned}
\Big| \Big(\sum_{j=1}^{m} \mathbf{K}(x_j, \cdot)\boldsymbol{c}_j, \sum_{k=1}^{n} \mathbf{K}(\cdot, s_k)\boldsymbol{b}_k \Big)_{\mathbf{K}} \Big| &\le \sum_{j=1}^{m} \Big| \boldsymbol{c}_j^{\top} \Big(\sum_{k=1}^{n} \mathbf{K}(x_j, s_k)\boldsymbol{b}_k \Big) \Big| \\
&\le \sum_{j=1}^{m} \|\boldsymbol{c}_j\|_1 \Big\| \sum_{k=1}^{n} \mathbf{K}(x_j, s_k)\boldsymbol{b}_k \Big\|_{\infty} \\
&\le \Big(\sum_{j=1}^{m} \|\boldsymbol{c}_j\|_1 \Big) \Big(\sup_{t \in X} \Big\| \sum_{k=1}^{n} \mathbf{K}(t, s_k)\boldsymbol{b}_k \Big\|_{\infty} \Big) \\
&= \Big\| \sum_{j=1}^{m} \mathbf{K}(x_j, \cdot)\boldsymbol{c}_j \Big\|_{\mathcal{B}_0} \Big\| \sum_{k=1}^{n} \mathbf{K}(\cdot, s_k)\boldsymbol{b}_k \Big\|_{\mathcal{B}_0^{\#}}.
\end{aligned}$$

It implies that the bilinear form $(\cdot,\cdot)_{\mathbf{K}}$ is continuous on $\mathcal{B}_0 \times \mathcal{B}_0^{\#}$. By applying the Hahn–Banach extension theorem twice, the bilinear form can be extended to $\mathcal{B}_{\mathbf{K}} \times \mathcal{B}_{\mathbf{K}}^{\#}$ such that

$$|(f, g)_{\mathbf{K}}| \le \|f\|_{\mathcal{B}_{\mathbf{K}}} \|g\|_{\mathcal{B}_{\mathbf{K}}^{\#}} \text{ for all } f \in \mathcal{B}_{\mathbf{K}}, g \in \mathcal{B}_{\mathbf{K}}^{\#}.$$

Finally, we are ready to show that $\mathbf{K}$ is a reproducing kernel for $\mathcal{B}_{\mathbf{K}}$.

Theorem 2.6. *If $\mathbf{K} : X \times X \to \mathbb{R}^{d\times d}$ is an admissible multi-task kernel then $\mathcal{B}_{\mathbf{K}}$ and $\mathcal{B}_{\mathbf{K}}^{\#}$ are a pair of vector-valued RKBSs, and $\mathbf{K}$ is a reproducing kernel for $\mathcal{B}_{\mathbf{K}}$ and $\mathcal{B}_{\mathbf{K}}^{\#}$.*

Proof. It is sufficient to verify the reproducing properties for $\mathbf{K}$ in $\mathcal{B}_{\mathbf{K}}$ and $\mathcal{B}_{\mathbf{K}}^{\#}$. Recall that $(\cdot,\cdot)_{\mathbf{K}}$ is a continuous bilinear form on $\mathcal{B}_{\mathbf{K}} \times \mathcal{B}_{\mathbf{K}}^{\#}$. For any $f \in \mathcal{B}_{\mathbf{K}}$, there exist distinct points $x_j \in X$, $j \in \mathbb{N}$ and $\mathbf{c} \in \ell_d^1(\mathbb{N})$ such that $f = \sum_{j\in\mathbb{N}} \mathbf{K}(x_j, \cdot)\boldsymbol{c}_j$. Since $\mathbf{K} = \mathbf{K}^{\top}$, it follows from a direct computation

$$\begin{aligned}
(f, \mathbf{K}(\cdot, x)\boldsymbol{b})_{\mathbf{K}} &= \lim_{n\to+\infty} \Big(\sum_{j=1}^{n} \mathbf{K}(x_j, \cdot)\boldsymbol{c}_j, \mathbf{K}(\cdot, x)\boldsymbol{b} \Big)_{\mathbf{K}} = \lim_{n\to+\infty} \sum_{j=1}^{n} \boldsymbol{c}_j^{\top} \mathbf{K}(x_j, x)\boldsymbol{b} \\
&= \lim_{n\to+\infty} \Big(\sum_{j=1}^{n} \mathbf{K}(x_j, x)\boldsymbol{c}_j \Big)^{\top} \boldsymbol{b} = f(x)^{\top}\boldsymbol{b}
\end{aligned}$$

for any $x \in X$. The reproducing property for $\mathbf{K}$ in $\mathcal{B}_{\mathbf{K}}^{\#}$ follows in a similar way. □

3. Representer theorems

The linear representer theorems play a fundamental role in regularized learning schemes in machine learning. It helps us to turn the infinite-dimensional optimization problem to an equivalent optimization problem in a finite-dimensional subspace. We shall establish in this section representer theorems for the minimal norm interpolation problem and regularization networks in the vector-valued RKBSs $\mathcal{B}_{\mathbf{K}}$ with the ℓ^1 norm constructed in Section 2.

Please cite this article as: R. Lin, G. Song and H. Zhang, Multi-task Learning in vector-valued reproducing kernel Banach spaces with the ℓ^1 norm, Journal of Complexity (2020) 101514, https://doi.org/10.1016/j.jco.2020.101514.

3.1. Minimal norm interpolation

A minimal norm interpolation problem in a vector-valued RKBS $\mathcal{B}_{\mathbf{K}}$ with respect to a set of sampling data $\{(x_j, \boldsymbol{y}_j) : j \in \mathbb{N}_m\} \subseteq X \times \mathbb{R}^d$ is to solve

$$\min_{f \in \mathcal{I}_{\mathbf{x}}(\boldsymbol{y})} \|f\|_{\mathcal{B}_{\mathbf{K}}}, \text{ where } \mathcal{I}_{\mathbf{x}}(\boldsymbol{y}) := \Big\{ f \in \mathcal{B}_{\mathbf{K}} : f(\mathbf{x}) = \boldsymbol{y} \Big\}. \tag{3.1}$$

We always assume the minimizer of the above problem exists in this paper.

We say the vector-valued RKBS $\mathcal{B}_{\mathbf{K}}$ has the *linear representer theorem for minimal norm interpolation* if for any integer $m \in \mathbb{N}$ and any choice of sampling data $\{(x_j, \boldsymbol{y}_j) : j \in \mathbb{N}_m\} \subseteq X \times \mathbb{R}^d$, there always exists a minimizer of the problem (3.1) in the following finite-dimensional subspace:

$$\mathcal{S}^{\mathbf{x}} := \Big\{ \sum_{j=1}^{m} \mathbf{K}(x_j, \cdot)\boldsymbol{c}_j : \boldsymbol{c}_j \in \mathbb{R}^d,\ j \in \mathbb{N}_m \Big\}. \tag{3.2}$$

We point out that the finite-dimensional subspace $\mathcal{S}^{\mathbf{x}}$ has dimension md.

We shall next investigate the linear representer theorem for minimal norm interpolation. We remark that the interpolation space $\mathcal{I}_{\mathbf{x}}(\boldsymbol{y})$ is infinite-dimensional in general. We need to show the minimal norm interpolation in $\mathcal{B}_{\mathbf{K}}$ is equivalent to that in the finite-dimensional subspace in $\mathcal{S}^{\mathbf{x}}$. To this end, we first show the minimal norm interpolation in a finite-dimensional subspace containing $\mathcal{S}^{\mathbf{x}}$ could be reduced to that in $\mathcal{S}^{\mathbf{x}}$. We begin with the simplest case when a new point x_{m+1} is added to $\mathbf{x}$. In other words, we shall consider the finite-dimensional subspace $\mathcal{S}^{\tilde{\mathbf{x}}}$, where $\tilde{\mathbf{x}} := \mathbf{x} \cup \{x_{m+1}\}$ and $x_{m+1} \in X \setminus \mathbf{x}$. Observe that

$$\mathcal{I}_{\mathbf{x}}(\boldsymbol{y}) \cap \mathcal{S}^{\mathbf{x}} \subseteq \mathcal{I}_{\mathbf{x}}(\boldsymbol{y}) \cap \mathcal{S}^{\tilde{\mathbf{x}}} \subseteq \mathcal{I}_{\mathbf{x}}(\boldsymbol{y}).$$

For notational simplicity, we denote

$$\mathbf{K}_{\mathbf{x}}(t) := (\mathbf{K}(t, x_j) : j \in \mathbb{N}_m)^\top \in \mathbb{R}^{md \times d},\ \mathbf{K}^{\mathbf{x}}(t) := (\mathbf{K}(x_j, t) : j \in \mathbb{N}_m) \in \mathbb{R}^{d \times md},\ t \in X.$$

It is worthwhile to point out that $\mathbf{K}_{\mathbf{x}}$ is in general not the transpose of $\mathbf{K}^{\mathbf{x}}$. If, in addition, $\mathbf{K}$ is symmetric in the sense that $\mathbf{K}(x, x') = \mathbf{K}(x', x)$ for all $x, x' \in X$ then by $\mathbf{K} = \mathbf{K}^\top$ we have $\mathbf{K}_{\mathbf{x}}^\top = \mathbf{K}^{\mathbf{x}}$.

We will present a necessary and sufficient condition for the equivalence of the minimal norm interpolation in $\mathcal{S}^{\tilde{\mathbf{x}}}$ to that in $\mathcal{S}^{\mathbf{x}}$.

Lemma 3.1. *Suppose the multi-task kernel $\mathbf{K}$ is admissible. Let $\mathbf{x} = \{x_j : j \in \mathbb{N}_m\}$ be a set of pairwise distinct points, let x_{m+1} be an arbitrary point in $X \setminus \mathbf{x}$, and let $\tilde{\mathbf{x}} = \mathbf{x} \cup \{x_{m+1}\}$. Then*

$$\min_{f \in \mathcal{I}_{\mathbf{x}}(\boldsymbol{y}) \cap \mathcal{S}^{\tilde{\mathbf{x}}}} \|f\|_{\mathcal{B}_{\mathbf{K}}} = \min_{f \in \mathcal{I}_{\mathbf{x}}(\boldsymbol{y}) \cap \mathcal{S}^{\mathbf{x}}} \|f\|_{\mathcal{B}_{\mathbf{K}}}, \quad \text{for all } \boldsymbol{y} \in \mathbb{R}^{md} \tag{3.3}$$

if and only if $\|\mathbf{K}[\mathbf{x}]^{-1}\mathbf{K}_{\mathbf{x}}(x_{m+1})\|_1 \le 1$.

Proof. Notice that the set $\mathcal{I}_{\mathbf{x}}(\boldsymbol{y}) \cap \mathcal{S}^{\mathbf{x}}$ possesses only one interpolant $f(t) = \mathbf{K}^{\mathbf{x}}(t)\mathbf{K}[\mathbf{x}]^{-1}\boldsymbol{y}$, $t \in X$. Let $\tilde{f} \in \mathcal{I}_{\mathbf{x}}(\boldsymbol{y}) \cap \mathcal{S}^{\tilde{\mathbf{x}}}$ and $\boldsymbol{b} := \tilde{f}(x_{m+1})$. Note that $\tilde{f}$ is uniquely determined by $\boldsymbol{b}$ as it has already satisfied the interpolation condition $\tilde{f}(\mathbf{x}) = \boldsymbol{y}$. Specifically, by the admissible assumption (A1), we get $\tilde{f}(t) = \mathbf{K}^{\tilde{\mathbf{x}}}(t)\mathbf{K}[\tilde{\mathbf{x}}]^{-1}\tilde{\boldsymbol{y}}$, $t \in X$, where $\tilde{\boldsymbol{y}} := (\boldsymbol{y}^\top, \boldsymbol{b})^\top \in \mathbb{R}^{(m+1)d}$. It follows from a result (see for instance, [25], pages 201–202) concerning the inversion of 2×2 blockwise invertible matrix that

$$\begin{aligned}
\mathbf{K}[\tilde{\mathbf{x}}]^{-1}\tilde{\boldsymbol{y}} &= \begin{bmatrix} \mathbf{K}[\mathbf{x}] & \mathbf{K}_{\mathbf{x}}(x_{m+1}) \\ \mathbf{K}^{\mathbf{x}}(x_{m+1}) & \mathbf{K}(x_{m+1}, x_{m+1}) \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{b} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{K}[\mathbf{x}]^{-1} + \mathbf{K}[\mathbf{x}]^{-1}\mathbf{K}_{\mathbf{x}}(x_{m+1})\boldsymbol{p}^{-1}\mathbf{K}^{\mathbf{x}}(x_{m+1})\mathbf{K}[\mathbf{x}]^{-1} & -\mathbf{K}[\mathbf{x}]^{-1}\mathbf{K}_{\mathbf{x}}(x_{m+1})\boldsymbol{p}^{-1} \\ -\boldsymbol{p}^{-1}\mathbf{K}^{\mathbf{x}}(x_{m+1})\mathbf{K}[\mathbf{x}]^{-1} & \boldsymbol{p}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{b} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{K}[\mathbf{x}]^{-1}\boldsymbol{y} + \mathbf{K}[\mathbf{x}]^{-1}\mathbf{K}_{\mathbf{x}}(x_{m+1})\boldsymbol{p}^{-1}\boldsymbol{q} \\ -\boldsymbol{p}^{-1}\boldsymbol{q} \end{bmatrix}
\end{aligned}$$

Please cite this article as: R. Lin, G. Song and H. Zhang, Multi-task Learning in vector-valued reproducing kernel Banach spaces with the ℓ^1 norm, Journal of Complexity (2020) 101514, https://doi.org/10.1016/j.jco.2020.101514.

where the matrix

$$\boldsymbol{p} := \left[\mathbf{K}(x_{m+1}, x_{m+1}) - \mathbf{K}^{\mathbf{x}}(x_{n+1})\mathbf{K}[\mathbf{x}]^{-1}\mathbf{K}_{\mathbf{x}}(x_{m+1})\right] \in \mathbb{R}^{d\times d}$$

is non-singular and the column vector $\boldsymbol{q}$ is

$$\boldsymbol{q} := \mathbf{K}^{\mathbf{x}}(x_{m+1})\mathbf{K}[\mathbf{x}]^{-1}\boldsymbol{y} - \boldsymbol{b} \in \mathbb{R}^d.$$

We now show sufficiency. If $\|\mathbf{K}[\mathbf{x}]^{-1}\mathbf{K}_{\mathbf{x}}(x_{m+1})\|_1 \le 1$, then

$$\begin{aligned}\|\tilde{f}\|_{\mathcal{B}_{\mathbf{K}}} = \|\mathbf{K}[\tilde{\mathbf{x}}]^{-1}\tilde{\boldsymbol{y}}\|_1 &\ge \|\mathbf{K}[\mathbf{x}]^{-1}\boldsymbol{y} + \mathbf{K}[\mathbf{x}]^{-1}\mathbf{K}_{\mathbf{x}}(x_{n+1})\boldsymbol{p}^{-1}\boldsymbol{q}\|_1 + \|\boldsymbol{p}^{-1}\boldsymbol{q}\|_1\\ &\ge \|\mathbf{K}[\mathbf{x}]^{-1}\boldsymbol{y}\|_1 - \|\mathbf{K}[\mathbf{x}]^{-1}\mathbf{K}_{\mathbf{x}}(x_{n+1})\boldsymbol{p}^{-1}\boldsymbol{q}\|_1 + \|\boldsymbol{p}^{-1}\boldsymbol{q}\|_1\\ &\ge \|\mathbf{K}[\mathbf{x}]^{-1}\boldsymbol{y}\|_1 - \|\mathbf{K}[\mathbf{x}]^{-1}\mathbf{K}_{\mathbf{x}}(x_{n+1})\|_1\|\boldsymbol{p}^{-1}\boldsymbol{q}\|_1 + \|\boldsymbol{p}^{-1}\boldsymbol{q}\|_1\\ &\ge \|\mathbf{K}[\mathbf{x}]^{-1}\boldsymbol{y}\|_1\\ &= \|f\|_{\mathcal{B}_{\mathbf{K}}}\end{aligned}$$

which implies

$$\min_{f\in\mathcal{I}_{\mathbf{x}}(\boldsymbol{y})\cap\mathcal{S}^{\tilde{\mathbf{x}}}} \|f\|_{\mathcal{B}_{\mathbf{K}}} \ge \min_{f\in\mathcal{I}_{\mathbf{x}}(\boldsymbol{y})\cap\mathcal{S}^{\mathbf{x}}} \|f\|_{\mathcal{B}_{\mathbf{K}}}, \text{ for all } \boldsymbol{y} \subseteq \mathbb{R}^{md}.$$

Since $\mathcal{S}^{\mathbf{x}} \subseteq \mathcal{S}^{\tilde{\mathbf{x}}}$, the reverse direction of this inequality holds. Thus, (3.3) holds true.

Conversely, if (3.3) is true for all $\boldsymbol{y} \in \mathbb{R}^{md}$ then we must have

$$\|\mathbf{K}[\tilde{\mathbf{x}}]^{-1}\tilde{\boldsymbol{y}}\|_1 \ge \|\mathbf{K}[\mathbf{x}]^{-1}\boldsymbol{y}\|_1, \text{ for all } \boldsymbol{y} \in \mathbb{R}^{md} \text{ and all } \boldsymbol{b} \in \mathbb{R}^d.$$

Fix $j \in \mathbb{N}_d$. In particular, if we choose

$$\boldsymbol{y} := \mathbf{K}_{\mathbf{x}}(x_{m+1})\boldsymbol{e}_j, \text{ and } \boldsymbol{b} := \mathbf{K}^{\mathbf{x}}(x_{m+1})\mathbf{K}[\mathbf{x}]^{-1}\boldsymbol{y} + \boldsymbol{p}\boldsymbol{e}_j,$$

where $\boldsymbol{e}_j$ is a column vector in $\mathbb{R}^d$ whose jth component is 1 and other components are 0, then

$$\boldsymbol{q} = -\boldsymbol{p}\boldsymbol{e}_j, \ \boldsymbol{p}^{-1}\boldsymbol{q} = -\boldsymbol{e}_j \text{ and } \mathbf{K}[\mathbf{x}]^{-1}\boldsymbol{y} + \mathbf{K}[\mathbf{x}]^{-1}\mathbf{K}_{\mathbf{x}}(x_{m+1})\boldsymbol{p}^{-1}\boldsymbol{q} = \mathbf{0}_{md},$$

where $\mathbf{0}_{md}$ denotes the zero column vector in $\mathbb{R}^{md}$. Consequently

$$\|\mathbf{K}[\tilde{\mathbf{x}}]^{-1}\tilde{\boldsymbol{y}}\|_1 = \left\| \begin{matrix} \mathbf{0}_{md} \\ \boldsymbol{e}_j \end{matrix} \right\|_1 = 1 \text{ and } \|\mathbf{K}[\mathbf{x}]^{-1}\boldsymbol{y}\|_1 = \|\mathbf{K}[\mathbf{x}]^{-1}\mathbf{K}_{\mathbf{x}}(x_{m+1})\boldsymbol{e}_j\|_1.$$

Recalling the definition of the ℓ^1 norm of matrices, we get $\|\mathbf{K}[\mathbf{x}]^{-1}\mathbf{K}_{\mathbf{x}}(x_{m+1})\|_1 \le 1$. The proof is complete. □

We are now ready to present the following necessary and sufficient condition for the equivalence of the minimal norm interpolation in $\mathcal{B}_{\mathbf{K}}$ to that in $\mathcal{S}^{\mathbf{x}}$:

(**Lebesgue Constant Condition**) For all $m \in \mathbb{N}$ and all pairwise distinct sampling points $\mathbf{x} = \{x_j : j \in \mathbb{N}_m\} \subseteq X$,

$$\sup_{t\in X} \|\mathbf{K}[\mathbf{x}]^{-1}\mathbf{K}_{\mathbf{x}}(t)\|_1 \le 1, \tag{3.4}$$

where $\mathbf{K}_{\mathbf{x}}(t) = (\mathbf{K}(t, x_j) : j \in \mathbb{N}_m)^\top$, $t \in X$.

Theorem 3.2. *Suppose* $\mathbf{K}$ *is an admissible multi-task kernel. The space* $\mathcal{B}_{\mathbf{K}}$ *satisfies the linear representer theorem for minimal norm interpolation if and only if the Lebesgue constant condition* (3.4) *holds.*

Proof. We first prove the necessity. The space $\mathcal{B}_{\mathbf{K}}$ satisfies the linear representer theorem for minimal norm interpolation if and only if for any integer m and any choice of sampling data $\{(x_j, \boldsymbol{y}_j) : j \in \mathbb{N}_m\} \subseteq X \times \mathbb{R}^d$

$$\min_{g\in\mathcal{I}_{\mathbf{x}}(\boldsymbol{y})} \|g\|_{\mathcal{B}_{\mathbf{K}}} = \min_{f\in\mathcal{I}_{\mathbf{x}}(\boldsymbol{y})\cap\mathcal{S}^{\mathbf{x}}} \|f\|_{\mathcal{B}_{\mathbf{K}}}.$$

Choose a new point $x_{m+1} \in X \setminus \mathbf{x}$. Observe $\mathcal{I}_{\mathbf{x}}(\boldsymbol{y}) \cap \mathcal{S}^{\mathbf{x}} \subseteq \mathcal{I}_{\mathbf{x}}(\boldsymbol{y}) \cap \mathcal{S}^{\tilde{\mathbf{x}}} \subseteq \mathcal{I}_{\mathbf{x}}(\boldsymbol{y})$. By Lemma 3.1, we have $\|\mathbf{K}[\mathbf{x}]^{-1}\mathbf{K}_{\mathbf{x}}(x_{m+1})\|_1 \le 1$. Observe that $\|\mathbf{K}[\mathbf{x}]^{-1}\mathbf{K}_{\mathbf{x}}(t)\|_1 = 1$ for any $t \in \mathbf{x}$. Since we can pick an arbitrary integer m and an arbitrary point x_{m+1}, this leads to (3.4).

Please cite this article as: R. Lin, G. Song and H. Zhang, Multi-task Learning in vector-valued reproducing kernel Banach spaces with the ℓ^1 norm, Journal of Complexity (2020) 101514, https://doi.org/10.1016/j.jco.2020.101514.

We next show the sufficiency. Suppose the Lebesgue constant condition (3.4) holds. Fix $m \in \mathbb{N}$. Since $\mathcal{I}_{\mathbf{x}}(\boldsymbol{y}) \cap \mathcal{S}^{\mathbf{x}} \subseteq \mathcal{I}_{\mathbf{x}}(\boldsymbol{y})$,

$$\min_{g \in \mathcal{I}_{\mathbf{x}}(\boldsymbol{y})} \|g\|_{\mathcal{B}_{\mathbf{K}}} \le \min_{f \in \mathcal{I}_{\mathbf{x}}(\boldsymbol{y}) \cap \mathcal{S}^{\mathbf{x}}} \|f\|_{\mathcal{B}_{\mathbf{K}}}.$$

It remains to prove the reverse direction of this inequality. For this purpose, we shall first show $\|g\|_{\mathcal{B}_{\mathbf{K}}} \ge \min_{f \in \mathcal{I}_{\mathbf{x}}(\boldsymbol{y}) \cap \mathcal{S}^{\mathbf{x}}} \|f\|_{\mathcal{B}_{\mathbf{K}}}$ for all $g \in \mathcal{I}_{\mathbf{x}}(\boldsymbol{y}) \cap \mathcal{B}_0$, where the space $\mathcal{B}_0$ is defined by (2.6). By $\mathcal{I}_{\mathbf{x}}(\boldsymbol{y}) \cap \mathcal{S}^{\mathbf{x}} \subseteq \mathcal{I}_{\mathbf{x}}(\boldsymbol{y}) \cap \mathcal{B}_0$, the set $\mathcal{I}_{\mathbf{x}}(\boldsymbol{y}) \cap \mathcal{B}_0$ is non-empty. We write g as $g = \sum_{j=1}^{n} \mathbf{K}(x_j, \cdot)\boldsymbol{c}_j$ for some integer $n \ge m$ and distinct points $\{x_j : j \in \mathbb{N}_n\}$. This can always be done by adding some sampling points, setting the corresponding coefficients to be zero, and relabeling if necessary. Let $\boldsymbol{y}_j := g(x_j)$, $j \in \mathbb{N}_n$. Then set

$$\boldsymbol{u}_l := (\boldsymbol{y}_j : j \in \mathbb{N}_l)^T, \text{ and } \boldsymbol{v}_l := \{x_j : j \in \mathbb{N}_l\} \text{ for } l = m, m+1, \ldots, n.$$

Note that $\boldsymbol{y} = \boldsymbol{u}_m$ and $\mathbf{x} = \boldsymbol{v}_m$. By $g \in \mathcal{I}_{\boldsymbol{v}_n}(\boldsymbol{u}_n) \cap \mathcal{S}^{\boldsymbol{v}_n}$, it follows that $\|g\|_{\mathcal{B}_{\mathbf{K}}} \ge \min_{f \in \mathcal{I}_{\boldsymbol{v}_n}(\boldsymbol{u}_n) \cap \mathcal{S}^{\boldsymbol{v}_n}} \|f\|_{\mathcal{B}_{\mathbf{K}}}$. Since $\mathcal{I}_{\boldsymbol{v}_n}(\boldsymbol{u}_n) \subseteq \mathcal{I}_{\boldsymbol{v}_{n-1}}(\boldsymbol{u}_{n-1})$, we apply Lemma 3.1 to get

$$\min_{f \in \mathcal{I}_{\boldsymbol{v}_n}(\boldsymbol{u}_n) \cap \mathcal{S}^{\boldsymbol{v}_n}} \|f\|_{\mathcal{B}_{\mathbf{K}}} \ge \min_{f \in \mathcal{I}_{\boldsymbol{v}_{n-1}}(\boldsymbol{u}_{n-1}) \cap \mathcal{S}^{\boldsymbol{v}_n}} \|f\|_{\mathcal{B}_{\mathbf{K}}} = \min_{f \in \mathcal{I}_{\boldsymbol{v}_{n-1}}(\boldsymbol{u}_{n-1}) \cap \mathcal{S}^{\boldsymbol{v}_{n-1}}} \|f\|_{\mathcal{B}_{\mathbf{K}}}.$$

It follows that $\|g\|_{\mathcal{B}_{\mathbf{K}}} \ge \min_{f \in \mathcal{I}_{\boldsymbol{v}_{n-1}}(\boldsymbol{u}_{n-1}) \cap \mathcal{S}^{\boldsymbol{v}_{n-1}}} \|f\|_{\mathcal{B}_{\mathbf{K}}}$. Repeating this process, we get

$$\|g\|_{\mathcal{B}_{\mathbf{K}}} \ge \min_{f \in \mathcal{I}_{\boldsymbol{v}_m}(\boldsymbol{u}_m) \cap \mathcal{S}^{\boldsymbol{v}_m}} \|f\|_{\mathcal{B}_{\mathbf{K}}} = \min_{f \in \mathcal{I}_{\mathbf{x}}(\boldsymbol{y}) \cap \mathcal{S}^{\mathbf{x}}} \|f\|_{\mathcal{B}_{\mathbf{K}}} \text{ for all } g \in \mathcal{I}_{\mathbf{x}}(\boldsymbol{y}) \cap \mathcal{B}_0. \tag{3.5}$$

Now let $g \in \mathcal{I}_{\mathbf{x}}(\boldsymbol{y})$ be arbitrary but fixed. Recall that the completion of $\mathcal{B}_0$ with respect to the ℓ^1 norm is $\mathcal{B}_{\mathbf{K}}$. Then there exists a sequence of vector-valued functions $g_j \in \mathcal{B}_0$, $j \in \mathbb{N}$ that converges to g in $\mathcal{B}_{\mathbf{K}}$. We let f and f_j be the function in $\mathcal{S}^{\mathbf{x}}$ such that $f(\mathbf{x}) = \boldsymbol{y}$ and $f_j(\mathbf{x}) = g_j(\mathbf{x})$, $j \in \mathbb{N}$. They are explicitly given by

$$f = \mathbf{K}^{\mathbf{x}}(\cdot)\mathbf{K}[\mathbf{x}]^{-1} g(\mathbf{x}) \quad \text{and} \quad f_j = \mathbf{K}^{\mathbf{x}}(\cdot)\mathbf{K}[\mathbf{x}]^{-1} g_j(\mathbf{x}), \quad j \in \mathbb{N}.$$

Since g_j converges to g in $\mathcal{B}_{\mathbf{K}}$ and point evaluation functionals are continuous on $\mathcal{B}_{\mathbf{K}}$, $g_j(\mathbf{x}) \to g(\mathbf{x})$ as $j \to +\infty$. As a result,

$$\lim_{j \to +\infty} \|f - f_j\|_{\mathcal{B}_{\mathbf{K}}} = \lim_{j \to +\infty} \|\mathbf{K}[\mathbf{x}]^{-1}(g(\mathbf{x}) - g_j(\mathbf{x}))\|_1 \le \|\mathbf{K}[\mathbf{x}]^{-1}\|_1 \lim_{j \to +\infty} \|g(\mathbf{x}) - g_j(\mathbf{x})\|_1 = 0.$$

By (3.5), $\|g_j\|_{\mathcal{B}_{\mathbf{K}}} \ge \|f_j\|_{\mathcal{B}_{\mathbf{K}}}$ for all $j \in \mathbb{N}$. It follows that $\|g\|_{\mathcal{B}_{\mathbf{K}}} \ge \|f\|_{\mathcal{B}_{\mathbf{K}}}$ and thus,

$$\min_{g \in \mathcal{I}_{\mathbf{x}}(\boldsymbol{y})} \|g\|_{\mathcal{B}_{\mathbf{K}}} \ge \min_{f \in \mathcal{I}_{\mathbf{x}}(\boldsymbol{y}) \cap \mathcal{S}^{\mathbf{x}}} \|f\|_{\mathcal{B}_{\mathbf{K}}}.$$

The proof is complete. □

3.2. Regularization networks

We will present a representer theorem for regularization network in the following form:

$$\min_{f \in \mathcal{B}_{\mathbf{K}}} L(f(\mathbf{x}), \boldsymbol{y}) + \lambda \phi(\|f\|_{\mathcal{B}_{\mathbf{K}}}), \tag{3.6}$$

where $L : \mathbb{R}^{md} \times \mathbb{R}^{md} \to \mathbb{R}_+$ is a continuous loss function with the property $L(\boldsymbol{t}, \boldsymbol{t}) = 0$ for any $\boldsymbol{t} \in \mathbb{R}^{md}$, and $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ a non-decreasing continuous regularizer with $\lim_{t \to +\infty} \phi(t) = +\infty$. We will always assume a minimizer of the above model (3.6) exists.

Similarly, we say the vector-valued RKBS $\mathcal{B}_{\mathbf{K}}$ has *the linear representer theorem for regularization network* if for any integer m, any choice of sampling data $\{(x_j, \boldsymbol{y}_j) : j \in \mathbb{N}_m\}$ and any $\lambda > 0$, there exists a minimizer of (3.6) in the finite-dimensional subspace $\mathcal{S}^{\mathbf{x}}$ as defined in (3.2).

We shall establish a necessary and sufficient condition such that the linear representer theorem for the regularization network holds in the vector-valued RKBS $\mathcal{B}_{\mathbf{K}}$. This is achieved by showing the equivalence between the representer theorems for the minimal norm interpolation problem and the regularization network problem.

Please cite this article as: R. Lin, G. Song and H. Zhang, Multi-task Learning in vector-valued reproducing kernel Banach spaces with the ℓ^1 norm, Journal of Complexity (2020) 101514, https://doi.org/10.1016/j.jco.2020.101514.

Theorem 3.3. *Suppose* $\mathbf{K}$ *is an admissible multi-task kernel. The space* $\mathcal{B}_{\mathbf{K}}$ *satisfies the linear representer theorem for regularization network* (3.6) *if and only if it does so for minimal norm interpolation* (3.1).

Proof. Suppose we are given some sampling data $\{(x_j, \boldsymbol{y}_j) : j \in \mathbb{N}_m\} \subseteq X \times \mathbb{R}^d$ for some $m \in \mathbb{N}$. We first assume that the space $\mathcal{B}_{\mathbf{K}}$ satisfies the linear representer theorem for minimal norm interpolation. We want to prove

$$\min_{f\in\mathcal{B}_{\mathbf{K}}} L(f(\mathbf{x}), \boldsymbol{y}) + \lambda\phi(\|f\|_{\mathcal{B}_{\mathbf{K}}}) = \min_{f\in\mathcal{S}^{\mathbf{x}}} L(f(\mathbf{x}), \boldsymbol{y}) + \lambda\phi(\|f\|_{\mathcal{B}_{\mathbf{K}}}). \tag{3.7}$$

Note that the left hand side above is always bounded above by the right hand side as $\mathcal{S}^{\mathbf{x}} \subseteq \mathcal{B}_{\mathbf{K}}$. We only need to show the reverse inequality. For any $f \in \mathcal{B}_{\mathbf{K}}$, we consider the following minimal norm interpolation problem

$$\min_{g\in\mathcal{I}_{\mathbf{x}}(f(\mathbf{x}))} \|g\|_{\mathcal{B}_{\mathbf{K}}}.$$

Since $\mathcal{B}_{\mathbf{K}}$ satisfies the linear representer theorem for minimal norm interpolation, there exists a minimizer $f_0 \in \mathcal{S}^{\mathbf{x}}$ of the above problem. Note that $f \in \mathcal{I}_{\mathbf{x}}(f(\mathbf{x}))$. It follows that $f_0(\mathbf{x}) = f(\mathbf{x})$ and $\|f_0\|_{\mathcal{B}_{\mathbf{K}}} \le \|f\|_{\mathcal{B}_{\mathbf{K}}}$. As a result, $L(f_0(\mathbf{x}), \boldsymbol{y}) = L(f(\mathbf{x}), \boldsymbol{y})$ but $\phi(\|f_0\|_{\mathcal{B}_{\mathbf{K}}}) \le \phi(\|f\|_{\mathcal{B}_{\mathbf{K}}})$ as ϕ is nondecreasing. That is, for any $f \in \mathcal{B}_{\mathbf{K}}$, there exists a $f_0 \in \mathcal{S}^{\mathbf{x}}$ such that

$$L(f_0(\mathbf{x}), \boldsymbol{y}) + \lambda\phi(\|f_0\|_{\mathcal{B}_{\mathbf{K}}}) \le L(f(\mathbf{x}), \boldsymbol{y}) + \lambda\phi(\|f\|_{\mathcal{B}_{\mathbf{K}}}),$$

which implies the right hand side of (3.7) is bounded above by the left hand side of (3.7).

Moreover, we will prove the existence of the minimizer of right hand side of (3.7). Since $\lim_{t\to+\infty} \phi(t) = +\infty$, there exists a positive constant α such that

$$\min_{f\in\mathcal{S}^{\mathbf{x}}} L(f(\mathbf{x}), \boldsymbol{y}) + \lambda\phi(\|f\|_{\mathcal{B}_{\mathbf{K}}}) = \min_{f\in\mathcal{S}^{\mathbf{x}}, \|f\|_{\mathcal{B}_{\mathbf{K}}}\le\alpha} L(f(\mathbf{x}), \boldsymbol{y}) + \lambda\phi(\|f\|_{\mathcal{B}_{\mathbf{K}}}).$$

For instance, we can choose any $\alpha > 0$ satisfying $\lambda\phi(\alpha) \le L(\mathbf{0}, \boldsymbol{y}) + \lambda\phi(0)$, where $\mathbf{0}$ is the zero vector in $\mathbb{R}^{md}$. Note that the functional we are minimizing is continuous on $\mathcal{B}_{\mathbf{K}}$ by the assumption on V, ϕ and by the continuity of point evaluation functionals on $\mathcal{B}_{\mathbf{K}}$. By the elementary fact that a continuous function on a compact metric space attains its minimum in the space, the right hand side of (3.7) has a minimizer that belongs to $\{f \in \mathcal{S}^{\mathbf{x}} : \|f\|_{\mathcal{B}_{\mathbf{K}}} \le \alpha\}$.

We next show the contrary part. That is, assuming $\mathcal{B}_{\mathbf{K}}$ satisfies the linear representer theorem for regularization networks, we need to show it also does so for minimal norm interpolation. We will find a minimizer of the minimal norm interpolation problem (3.1) explicitly. To this end, consider the regularization network (3.6) with the following choices of L and ϕ:

$$L(f(\mathbf{x}), \boldsymbol{y}) = \|f(\mathbf{x}) - \boldsymbol{y}\|_2^2, \quad \text{and} \quad \phi(t) = t.$$

For any $\lambda > 0$, let $f_{0,\lambda}$ be a minimizer of (3.6) with the above choices of L and ϕ in $\mathcal{S}^{\mathbf{x}}$. We could then write $f_{0,\lambda} = \mathbf{K}^{\mathbf{x}}(\cdot)\boldsymbol{c}_\lambda$ for some $\boldsymbol{c}_\lambda \in \mathbb{R}^{md}$. It follows

$$\|\mathbf{K}[\mathbf{x}]\boldsymbol{c}_\lambda - \boldsymbol{y}\|_2^2 = \|f_{0,\lambda}(\mathbf{x}) - \boldsymbol{y}\|_2^2 \le L(f_{0,\lambda}, \boldsymbol{y}) + \lambda\phi(\|f_{0,\lambda}\|_{\mathcal{B}_{\mathbf{K}}}) \le L(0, \boldsymbol{y}) + \lambda\phi(\|0\|_{\mathcal{B}_{\mathbf{K}}}) = \|\boldsymbol{y}\|_2^2.$$

Since $\mathbf{K}[\mathbf{x}]$ is nonsingular, the above inequality implies that $\{\boldsymbol{c}_\lambda : \lambda > 0\}$ forms a bounded set in $\mathbb{R}^{md}$. By restricting to a subsequence if necessary, we may hence assume that $\boldsymbol{c}_\lambda$ converges to some $\boldsymbol{c}_0 \in \mathbb{R}^{md}$ as λ goes to infinity. We then define $f_{0,0} := \mathbf{K}^{\mathbf{x}}(\cdot)\boldsymbol{c}_0$. It is clear that $f_{0,0} \in \mathcal{S}^{\mathbf{x}}$. We will show that $f_{0,0}$ is a minimizer of the minimal norm interpolation problem (3.1).

Assume g is an arbitrary interpolant in $\mathcal{I}_{\mathbf{x}}(\boldsymbol{y})$. It is enough to show $f_{0,0} \in \mathcal{I}_{\mathbf{x}}(\boldsymbol{y})$ and $\|f_{0,0}\|_{\mathcal{B}_{\mathbf{K}}} \le \|g\|_{\mathcal{B}_{\mathbf{K}}}$. By the definition of $f_{0,\lambda}$, we have

$$\|f_{0,\lambda}(\mathbf{x}) - \boldsymbol{y}\|_2^2 + \lambda\|f_{0,\lambda}\|_{\mathcal{B}_{\mathbf{K}}} \le \|g(\mathbf{x}) - \boldsymbol{y}\|_2^2 + \lambda\|g\|_{\mathcal{B}_{\mathbf{K}}} = \lambda\|g\|_{\mathcal{B}_{\mathbf{K}}}. \tag{3.8}$$

We observe that

$$\lim_{\lambda\to0} \|f_{0,\lambda} - f_{0,0}\|_{\mathcal{B}_{\mathbf{K}}} = \lim_{\lambda\to0} \|\boldsymbol{c}_\lambda - \boldsymbol{c}_0\|_1 = 0. \tag{3.9}$$

Since point evaluation functionals are continuous on $\mathcal{B}_{\mathbf{K}}$, we have $f_{0,0}(x_j) = \lim_{\lambda\to0} f_{0,\lambda}(x_j)$ for all $j \in \mathbb{N}_m$. Letting $\lambda \to 0$ on both sides of the above inequality (3.8), we obtain $\|f_{0,0}(\mathbf{x}) - \boldsymbol{y}\|_2^2 = 0$. That is,

Please cite this article as: R. Lin, G. Song and H. Zhang, Multi-task Learning in vector-valued reproducing kernel Banach spaces with the ℓ^1 norm, Journal of Complexity (2020) 101514, https://doi.org/10.1016/j.jco.2020.101514.

$f_{0,0} \in \mathcal{I}_{\mathbf{x}}(\boldsymbol{y})$. Moreover, it also follows from (3.8) that $\|f_{0,\lambda}\|_{\mathcal{B}_{\mathbf{K}}} \le \|g\|_{\mathcal{B}_{\mathbf{K}}}$ for all $\lambda > 0$. This together with (3.9) implies $\|f_{0,0}\|_{\mathcal{B}_{\mathbf{K}}} \le \|g\|_{\mathcal{B}_{\mathbf{K}}}$, which finishes the proof. □

Corollary 3.4. *Suppose that* $\mathbf{K}$ *is an admissible multi-task kernel. The space* $\mathcal{B}_{\mathbf{K}}$ *satisfies the linear representer theorem for regularization network* (3.6) *if and only if* $\mathbf{K}$ *satisfies the Lebesgue constant condition* (3.4).

Proof. This is an immediate consequence of Theorems 3.2 and 3.3. □

4. Examples of admissible multi-task kernels

We shall present a few examples of admissible multi-task kernels in this section. In particular, we will investigate whether they satisfy the Lebesgue constant condition (3.4) such that the corresponding vector-valued RKBSs have the linear representer theorems. Positive together with some negative examples will both be given.

Consider a widely used class of multi-task kernels [1,5] in machine learning, which take the following form:

$$\mathbf{K}(x, x') = K(x, x')\mathbb{A}, \ \ x, x' \in X, \tag{4.1}$$

where $K : X \times X \to \mathbb{R}$ is a scalar-valued positive definite kernel and $\mathbb{A}$ denotes a $d \times d$ strictly positive definite symmetric matrix. We reserve the notation $\mathbf{K}$ in boldface type to denote the $d \times d$ matrix-valued function and K the scalar-valued function as usual, respectively.

Theorem 4.1. *The kernel* $\mathbf{K}$ *defined by* (4.1) *is an admissible multi-task kernel if and only if* K *is an admissible single-task kernel.*

Proof. Notice that $\mathbf{K}^\top = \mathbf{K}$ as $\mathbb{A}$ is a symmetric matrix. We shall verify the non-singularity, boundedness, and independence assumptions. By (4.1), the non-singularity condition holds by noting that

$$\mathbf{K}[\mathbf{x}] = [K(x_k, x_j)\mathbb{A} : j, k \in \mathbb{N}_m] = \operatorname{diag}(\mathbb{A}, \ldots, \mathbb{A})[K(x_k, x_j) : j, k \in \mathbb{N}_m].$$

Observe that $\|\mathbf{K}(x, x')\|_1 = |K(x, x')|\|\mathbb{A}\|_1$, $x, x' \in X$. As a result, K is bounded on $X \times X$ if and only if $\|\mathbf{K}(x, x')\|_1$ is. To prove the independence assumption, we let $x_j \in X$, $j \in \mathbb{N}$ be distinct points and $\boldsymbol{c} = (\boldsymbol{c}_j : j \in \mathbb{N}) \in \ell^1_d(\mathbb{N})$. One sees that

$$\sum_{j\in\mathbb{N}} \mathbf{K}(x_j, x)\boldsymbol{c}_j = \sum_{j\in\mathbb{N}} K(x_j, x)\mathbb{A}\boldsymbol{c}_j = \mathbb{A}\sum_{j\in\mathbb{N}} K(x_j, x)\boldsymbol{c}_j.$$

As $\mathbb{A}$ is nonsingular, $\sum_{j\in\mathbb{N}} \mathbf{K}(x_j, x)\boldsymbol{c}_j = \mathbf{0}$ for all $x \in X$ if and only if $\sum_{j\in\mathbb{N}} K(x_j, x)(\boldsymbol{c}_j)_k = 0$ for all $k \in \mathbb{N}_d$ and all $x \in X$, where $(\boldsymbol{c}_j)_k$ denotes the kth component of the column vector $\boldsymbol{c}_j$. The proof is complete. □

The above theorem provides a way of constructing admissible multi-task kernels via their single-task counterparts. So far there are two admissible single-task kernels found in the literature [30]. They are the Brownian bridge kernel $K(x, x') := \min\{x, x'\} - xx'$, $x, x' \in (0, 1)$ and the exponential kernel $K(x, x') := e^{-|x-x'|}$, $x, x' \in \mathbb{R}$. Here we are able to contribute another one. Specifically, we shall show that the *covariance of Brownian motion* ([24], Subsection 1.4) defined by

$$K(x, y) := \min\{x, y\}, \ \ x, y \in (0, 1), \tag{4.2}$$

is an admissible single-task kernel. The corresponding RKHS $\mathcal{H}_K$, also called the Cameron–Martin–Hilbert space, consists of continuous functions f on $[0, 1]$ such that their distributional derivatives $f' \in L^2([0, 1])$ and $f(0) = 0$. The inner product on $\mathcal{H}_K$ is defined by $\langle f, g\rangle_{\mathcal{H}_K} := \int_0^1 f'(x)g'(x)dx$, where $f, g \in \mathcal{H}_K$.

To verify the Lebesgue constant condition (3.4) for this kernel, we explore the connection between the Lebesgue constants of kernels K and $\mathbf{K}$.

Please cite this article as: R. Lin, G. Song and H. Zhang, Multi-task Learning in vector-valued reproducing kernel Banach spaces with the ℓ^1 norm, Journal of Complexity (2020) 101514, https://doi.org/10.1016/j.jco.2020.101514.

Lemma 4.2. *Let $\mathbf{x} = \{x_j \in X : j \in \mathbb{N}_m\}$ be a set of distinct points, and $\alpha > 0$. Then*

$$\sup_{t\in X} \|K[\mathbf{x}]^{-1}K_{\mathbf{x}}(t)\|_1 \le \alpha \text{ if and only if } \sup_{t\in X} \|\mathbf{K}[\mathbf{x}]^{-1}\mathbf{K}_{\mathbf{x}}(t)\|_1 \le \alpha.$$

Proof. For each $t \in X$, set $K[\mathbf{x}]^{-1}K_{\mathbf{x}}(t) := (b_1(t), b_2(t), \dots, b_d(t))^\top$. By (4.1), we compute

$$\begin{aligned}\mathbf{K}[\mathbf{x}]^{-1}\mathbf{K}_{\mathbf{x}}(\cdot) &= \big[K(x_j, x_k)\mathbb{A} : k, j \in \mathbb{N}_m\big]^{-1}(K(x_j, \cdot)\mathbb{A} : j \in \mathbb{N}_m)^\top \\ &= \big[K(x_j, x_k)\mathbb{I} : k, j \in \mathbb{N}_m\big]^{-1} \operatorname{diag}(\mathbb{A}, \dots, \mathbb{A})^{-1} \operatorname{diag}(\mathbb{A}, \dots, \mathbb{A})(K(x_j, \cdot)\mathbb{I} : j \in \mathbb{N}_m)^\top \\ &= \big[K(x_j, x_k)\mathbb{I} : k, j \in \mathbb{N}_m\big]^{-1}(K(x_j, \cdot)\mathbb{I} : j \in \mathbb{N}_m)^\top \end{aligned} \tag{4.3}$$

where $\operatorname{diag}(\mathbb{A}, \dots, \mathbb{A})$ is a block diagonal matrix with $\mathbb{A}$ as the diagonal entries. By (4.3), this leads to

$$\sup_{t\in X} \|\mathbf{K}[\mathbf{x}]^{-1}\mathbf{K}_{\mathbf{x}}(t)\|_1 = \max_{t\in X} \|(b_1(t)\mathbb{I}, b_2(t)\mathbb{I}, \dots, b_d(t)\mathbb{I})^\top\|_1.$$

The proof is completed by noting the definition of the norm $\|\cdot\|_1$ for a matrices. □

Now, we verify that the covariance of Brownian motion is an admissible kernel. The proof is analogous to the one for the Brownian bride kernel in [30].

Theorem 4.3. *The covariance of Brownian motion defined by (4.2) is an admissible single-task kernel and satisfies the Lebesgue constant condition (3.4).*

Proof. Obviously, $|K(x, y)|$ is bounded by 1 for all $x, y \in (0, 1)$. Let $m \in \mathbb{N}$. Without loss of generality, we choose $0 < x_1 < x_2 < \cdots < x_m < 1$ and let $\mathbf{x} := \{x_1, x_2, \dots, x_m\}$. An easy computation shows that the determinant of the kernel matrix

$$K[\mathbf{x}] := \big[\min\{x_j, x_k\} : j, k \in \mathbb{N}_m\big] = \begin{bmatrix} x_1 & x_1 & x_1 & \dots & x_1 & x_1 \\ x_1 & x_2 & x_2 & \dots & x_2 & x_2 \\ x_1 & x_2 & x_3 & \dots & x_3 & x_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ x_1 & x_2 & x_3 & \dots & x_{m-1} & x_{m-1} \\ x_1 & x_2 & x_3 & \dots & x_{m-1} & x_m \end{bmatrix}$$

is $x_1(x_2 - x_1)(x_3 - x_1)\cdots(x_m - x_1) \neq 0$, and thus non-singularity assumption (**A1**) holds. Notice that for any $t \in (0, 1)$,

$$K(x, t) := \min\{x, t\} = \begin{cases} x & 0 < x \le t, \\ t & t < x < 1, \end{cases} \quad x \in (0, 1).$$

It follows that K satisfies the independence assumption (**A3**).

There are three cases when we compute the Lebesgue constant:

Case 1: If $0 < t < x_1$ then $K_{\mathbf{x}}(t) = (t, t, \dots, t)^\top$ and $K[\mathbf{x}]^{-1}K_{\mathbf{x}}(t) = (\frac{t}{x_1}, 0, \dots, 0)^\top$.

Case 2: If $x_m < t < 1$ then $K_{\mathbf{x}}(t) = (x_1, x_2, \dots, x_m)^\top$ and $K[\mathbf{x}]^{-1}K_{\mathbf{x}}(t) = (0, 0, \dots, 0, 1)^\top$.

Case 3: If $x_j \le t < x_{j+1}$ for some $j \in \mathbb{N}_{m-1}$ then $K_{\mathbf{x}}(t) = (x_1, x_2, \dots, x_j, t, \dots, t)^\top$ and

$$K[\mathbf{x}]^{-1}K_{\mathbf{x}}(t) = \Big(0, 0, \dots, 0, \frac{x_{j+1} - t}{x_{j+1} - x_j}, \frac{t - x_j}{x_{j+1} - x_j}, 0, \dots, 0\Big)^\top.$$

In all three cases, it is straightforward to see that $\max_{t\in(0,1)} \|K[\mathbf{x}]^{-1}K_{\mathbf{x}}(t)\|_1 \le 1$. Namely, the Lebesgue constant condition (3.4) is satisfied. The proof is hence complete. □

At the end of this section, we give three negative examples. We shall show that the Laplacian kernel, the exponential kernel, and the Gaussian kernel are not admissible when the dimension is higher than 1. This forces us to look for a relaxed linear representer theorem in the next section.

Please cite this article as: R. Lin, G. Song and H. Zhang, Multi-task Learning in vector-valued reproducing kernel Banach spaces with the ℓ^1 norm, Journal of Complexity (2020) 101514, https://doi.org/10.1016/j.jco.2020.101514.

Theorem 4.4. *The Laplacian kernel $K(x,x') = e^{-\|x-x'\|_2}$, $x, x' \in \mathbb{R}^n$ does not satisfy the Lebesgue constant condition (3.4) for any $n \geq 2$.*

Proof. We choose three distinct points $x_1 = (0,0,0,\ldots,0)^\top$, $x_2 = (1/10,0,0,\ldots,0)^\top$, and $x_3 = (0,1/10,0,\ldots,0)^\top$ in $\mathbb{R}^d$. Let $\mathbf{x} := \{x_1,x_2,x_3\}$ and $t_0 = (1/10,1/10,0,\ldots,0)$. Then we estimate

$$\begin{aligned}\sup_{t\in X}\|K[\mathbf{x}]^{-1}K_{\mathbf{x}}(t)\|_1 &\geq \|K[\mathbf{x}]^{-1}K_{\mathbf{x}}(t_0)\|_1\\ &=0.068064+0.517323+0.517323=1.102711>1,\end{aligned}$$

which completes the proof. □

Theorem 4.5. *The multivariate exponential kernel $K(x,x') = e^{-\|x-x'\|_1}$, $x, x' \in \mathbb{R}^n$ does not satisfy the Lebesgue constant condition (3.4) when $n \geq 2$.*

Proof. We begin with the proof of the special case when $n = 2$. Choose three distinct points $x_1 = (0,0)^\top$, $x_2 = (1/2,0)^\top$, $x_3 = (0,1/2)^\top$ in $\mathbb{R}^2$. Let $\mathbf{x} := \{x_1,x_2,x_3\}$. Then we estimate the Lebesgue constant of bivariate exponential kernel

$$\begin{aligned}\operatorname{supp}_{t\in\mathbb{R}^2}\left\|K[\mathbf{x}]^{-1}K_{\mathbf{x}}(t)\right\|_1 &\geq \left\|K[\mathbf{x}]^{-1}K_{\mathbf{x}}((1/2,1/2)^\top)\right\|_1\\ &= \left\|\begin{bmatrix}1 & e^{-\frac12} & e^{-\frac12}\\ e^{-\frac12} & 1 & e^{-1}\\ e^{-\frac12} & e^{-1} & 1\end{bmatrix}^{-1}\begin{bmatrix}e^{-1}\\ e^{-\frac12}\\ e^{-\frac12}\end{bmatrix}\right\|_1\\ &= \left\|\frac{1}{1-e^{-1}}\begin{bmatrix}1+e^{-1} & -e^{-\frac12} & -e^{-\frac12}\\ -e^{-\frac12} & 1 & 0\\ -e^{-\frac12} & 0 & 1\end{bmatrix}\begin{bmatrix}e^{-1}\\ e^{-\frac12}\\ e^{-\frac12}\end{bmatrix}\right\|_1\\ &= \|(-e^{-1},e^{-\frac12},e^{-\frac12})^\top\|_1 = e^{-1}+2e^{-\frac12}>1.\end{aligned}$$

In general, for any $n \geq 3$, we choose $n+1$ points, $x_1 = \mathbf{0}$, $x_{l+1} = \boldsymbol{e}_l/2$ in $\mathbb{R}^n$ for all $l \in \mathbb{N}_n$. Here $\boldsymbol{e}_l$ is a column vector in $\mathbb{R}^n$ whose lth component is 1 and other components are 0. Let $\mathbf{x} := \{x_1,x_2,\ldots,x_{n+1}\}$. Then we compute

$$\begin{aligned}\operatorname{supp}_{t\in\mathbb{R}^n}\left\|K[\mathbf{x}]^{-1}K_{\mathbf{x}}(t)\right\|_1 &\geq \left\|K[\mathbf{x}]^{-1}K_{\mathbf{x}}\Big(\frac{\boldsymbol{e}_1+\boldsymbol{e}_2}{2}\Big)\right\|_1\\ &= \left\|(-e^{-1},e^{-\frac12},e^{-\frac12},0,\ldots,0)^\top\right\|_1 = e^{-1}+2e^{-\frac12}>1.\end{aligned}$$

In other words, the vector $K_{\mathbf{x}}(\frac{\boldsymbol{e}_1+\boldsymbol{e}_2}{2})$ can be exactly represented by the first three columns of $K[\mathbf{x}]$. The proof is hence complete. □

Theorem 4.6. *The Gaussian kernel $K(x,x') = e^{-\|x-x'\|_2^2}$, $x, x' \in \mathbb{R}^n$ does not satisfy the Lebesgue constant condition (3.4) for any $n \geq 1$.*

Proof. When $n = 1$, we choose two points $x_1 = 0$ and $x_2 = 1/2$ in $\mathbb{R}$. Let $\mathbf{x} := \{x_1,x_2\}$. Then we compute the Lebesgue constant of the Gaussian kernel on $\mathbb{R}$

$$\operatorname{supp}_{t\in\mathbb{R}}\left\|K[\mathbf{x}]^{-1}K_{\mathbf{x}}(t)\right\|_1 \geq \|K[\mathbf{x}]^{-1}K_{\mathbf{x}}(1)\|_1 = \|(-e^{-\frac12},e^{-\frac14}+e^{-\frac34})^\top\|_1 = e^{-\frac12}+e^{-\frac14}+e^{-\frac34}>1.$$

Generally, for any $n \geq 2$, we choose $n+1$ points, $x_1 = \mathbf{0}$, $x_{l+1} = \boldsymbol{e}_l/2$ in $\mathbb{R}^n$ for all $l \in \mathbb{N}_n$. Let $\mathbf{x} := \{x_1,x_2,\ldots,x_{n+1}\}$. Then we compute

$$\begin{aligned}\operatorname{supp}_{t\in\mathbb{R}^n}\left\|K[\mathbf{x}]^{-1}K_{\mathbf{x}}(t)\right\|_1 &\geq \left\|K[\mathbf{x}]^{-1}K_{\mathbf{x}}\Big(\frac{\boldsymbol{e}_1+\boldsymbol{e}_2}{2}\Big)\right\|_1\\ &= \left\|(-e^{-\frac12},e^{-\frac14},e^{-\frac14},0,\ldots,0)^\top\right\|_1 = e^{-\frac12}+2e^{-\frac14}>1.\end{aligned}$$

Please cite this article as: R. Lin, G. Song and H. Zhang, Multi-task Learning in vector-valued reproducing kernel Banach spaces with the ℓ^1 norm, Journal of Complexity (2020) 101514, https://doi.org/10.1016/j.jco.2020.101514.

In other words, the vector $K_{\mathbf{x}}(\frac{\boldsymbol{e}_1+\boldsymbol{e}_2}{2})$ can be exactly represented by the first three columns of $K[\mathbf{x}]$. The proof is hence complete. □

We remark that the Lebesgue constant for the kernel interpolation always satisfies

$$\operatorname{supp}_{t\in X}\left\|K[\mathbf{x}]^{-1}K_{\mathbf{x}}(t)\right\|_1 \geq \operatorname{supp}_{t\in \mathbf{x}}\left\|K[\mathbf{x}]^{-1}K_{\mathbf{x}}(t)\right\|_1 = 1.$$

Therefore, asking it to be exactly bounded below by 1 is a very strong condition and only a few kernels satisfy it. To address this problem, we are devoted to investigating a relaxed version of the representer theorem and the Lebesgue constant condition in the next section.

5. A relaxed representer theorem

As shown in Section 4, the Lebesgue constant condition (3.4) is a very strong condition and only a few commonly used kernels satisfy it. We will derive in this section relaxed representer theorems that need a weaker condition on the Lebesgue constant. We will also present a rich class of kernels that satisfy this weaker condition.

In particular, we will consider the *relaxed linear representer theorem* in the constructed vector-valued RKBS $\mathcal{B}_{\mathbf{K}}$ in the following form

$$\min_{f\in\mathcal{S}^{\mathbf{x}}} L(f(\mathbf{x}),\boldsymbol{y}) + \lambda\|f\|_{\mathcal{B}_{\mathbf{K}}} \leq \min_{f\in\mathcal{B}_{\mathbf{K}}} L(f(\mathbf{x}),\boldsymbol{y}) + \lambda\beta_m\|f\|_{\mathcal{B}_{\mathbf{K}}}, \tag{5.1}$$

where $\beta_m \geq 1$ is a constant depending on the number m of sampling points, the kernel $\mathbf{K}$ and the input space X. By the arguments in the proof of Theorem 3.3, we see that the left-hand side of (5.1) possesses a solution. The relaxed representer theorem tells that we can get a near-optimal minimizer of the regularization network over the finite sample-dependent function space in the sense of (5.1).

We point out that if we require $\beta_m = 1$ for all m, then it is exactly the same as the Lebesgue constant condition (3.4). Once allowing $\beta_m > 1$, there would be a large class of kernels included in our framework. On the other hand side, as long as β_m is bounded on m or does not increase too fast with respect to m, we will still get a reasonable learning rate estimate for the minimizer of the regularization network (3.6). This will be shown in Section 5.1. In Section 5.2, we present a weaker condition on the Lebesgue constant for the relaxed representer theorem (5.1).

5.1. Learning rate estimate under the relaxed representer theorem

We shall explain the important role of the relaxed representer theorem (5.1) by presenting an learning rate estimate for the regularization network (3.6). The scalar-valued case has been investigated in [29,30]. Here we shall work with the vector-valued RKBS $\mathcal{B}_{\mathbf{K}}$ with the ℓ^1 norm satisfying the relaxed representer theorem (5.1). We shall prove that after relaxing the requirement on the representer theorem, the generalization ability of the minimizer will not be much affected as long as the constant β_m is under control.

Before moving on, let us recall some existing work about the error analysis (or learning rate) for regularization networks in scalar-valued reproducing kernel spaces. The error analysis for regularization networks has been given in the scalar-valued RKHS [9], in the data dependent hypothesis space with ℓ^1 regularizer [28], and in the scalar-valued RKBS with the ℓ^1 norm satisfying the representer theorem (Theorem 3.5 of [29]) or satisfying the relaxed representer theorem (Page 113 of [30]).

To start with, some assumptions are needed. We let the loss function L in (5.1) be the least squared function and let X be a compact metric space. Suppose that $\mathbf{K} : X \times X \to \mathbb{R}^{d\times d}$ is a continuous positive definite kernel on X such that the relaxed representer theorem (5.1) is satisfied. For any matrix $A \in \mathbb{R}^{d\times d}$, we let $\|A\|_2 := \sup\{\|Ay\|_2 : y \in \mathbb{R}^d, \|y\|_2 \leq 1\}$ be the operator norm. We assume that there exists $C > 0$ such that $\|\mathbf{K}(x,t)\|_2 \leq C$ for all $x, t \in X$.

Please cite this article as: R. Lin, G. Song and H. Zhang, Multi-task Learning in vector-valued reproducing kernel Banach spaces with the ℓ^1 norm, Journal of Complexity (2020) 101514, https://doi.org/10.1016/j.jco.2020.101514.

Following a commonly used assumption in learning theory [9], we assume that the sample data $\boldsymbol{z} := \{(x_j, \boldsymbol{y}_j) : j \in \mathbb{N}_m\} \subseteq X \times \mathbb{R}^d$ is formed by independent and identically distributed instances of a random variable $(x, y) \in X \times \mathbb{R}^d$ subject to an unknown probability measure ρ on $X \times \mathbb{R}^d$. The minimizer $f_{\boldsymbol{z},\lambda}$ of the left-hand side of (5.1) takes the form

$$f_{\boldsymbol{z},\lambda} := \sum_{j=1}^{m} \mathbf{K}(x_j, \cdot)\boldsymbol{c}_j, \ \ \boldsymbol{c}_j \in \mathbb{R}^d. \tag{5.2}$$

We hope that $f_{\boldsymbol{z},\lambda}$ will well predict the outputs of new inputs from X. The performance of a general predictor $f : X \to \mathbb{R}^d$ is measured by

$$\mathcal{E}(f) := \int_{X\times\mathbb{R}^d} \|f(x) - y\|_2^2 d\rho(x, y),$$

where $\|\cdot\|_2$ denotes the standard Euclidean norm in $\mathbb{R}^d$. Suppose the probability measure ρ on $X \times \mathbb{R}^d$ can be factored as the product of the conditional probability measure $\rho(\cdot|x)$ at x and the marginal probability measure ρ_X on X. Then $\rho(x, y) = \rho(y|x)\rho_X(x)$ for all $x \in X$ and $y \in \mathbb{R}^d$. The vector-valued function $f_\rho : X \to \mathbb{R}^d$ minimizing $\mathcal{E}(f)$ is called the regression function defined by

$$f_\rho(x) = \int_{\mathbb{R}^d} y d\rho(y|x) = \Big(\int_{\mathbb{R}^d} y_j d\rho(y|x) : j \in \mathbb{N}_d\Big)^\top \in \mathbb{R}^d, \ \ x \in X, \tag{5.3}$$

where y_j the jth component of $y \in \mathbb{R}^d$. The definition and basic properties of vector integration can be found in [13]. Eq. (5.3) says that the function value $f_\rho(x)$ is the mean of the random variable $y \in \mathbb{R}^d$ with respect to the conditional probability measure $\rho(\cdot|x)$ at x. Equivalently, we have

$$\int_{\mathbb{R}^d} v^\top (f_\rho(x) - y) d\rho(y|x) = 0 \text{ for all } x \in X \text{ and } v \in \mathbb{R}^d.$$

A direct computation yields that

$$\mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f - f_\rho\|^2_{L^2(X,\rho_X;\mathbb{R}^d)}, \tag{5.4}$$

where $L^2(X, \rho_X; \mathbb{R}^d)$ denotes the Banach space of square-integrable vector-valued functions $f : X \to \mathbb{R}^d$ with respect to the measure ρ_X, that is,

$$\|f\|_{L^2(X,\rho_X;\mathbb{R}^d)} := \Big(\int_X \|f(x)\|_2^2 d\rho_X(x)\Big)^{1/2} < +\infty.$$

In general, the optimal predictor f_ρ is unknown. Hence, we shall approximate the regression function f_ρ with $f_{\boldsymbol{z},\lambda}$ defined by (5.2). By (5.4), we expect with a large confidence that the approximation error

$$\mathcal{E}(f_{\boldsymbol{z},\lambda}) - \mathcal{E}(f_\rho) = \|f_{\boldsymbol{z},\lambda} - f_\rho\|^2_{L^2(X,\rho_X;\mathbb{R}^d)}$$

would converge to zero rapidly as the number m of sampling points increases.

Similar to the scalar-valued case [9,29], the regression function f_ρ is assumed in the range $ran(L_{\mathbf{K}}^s)$ of $L_{\mathbf{K}}^s$ for some $s > 0$, where s represents the regularity of f_ρ. Here, $L_{\mathbf{K}}$ is the compact positive operator [11] on $L^2(X, \rho_X; \mathbb{R}^d)$ defined by

$$L_{\mathbf{K}}(f) := \int_X \mathbf{K}(\cdot, t) f(t) d\rho_X(t), \ \ f \in L^2(X, \rho_X; \mathbb{R}^d). \tag{5.5}$$

By the Mercer theorem (see Theorem 3.4 and Theorem A.1 in [11]) for matrix-valued kernels, there exists a sequence of orthonormal basis $\{\phi_j : j \in \mathbb{N}\}$ for $L^2(X, \rho_X; \mathbb{R}^d)$ consisting of eigenfunctions of $L_{\mathbf{K}}$ with the corresponding eigenvalues $\lambda_j \geq \lambda_{j+1}, j \in \mathbb{N}$. It follows that

$$\mathbf{K}(x, t) = \sum_{j\in\mathbb{N}} \lambda_j \phi_j(x) \cdot \phi_j(t)^\top \in \mathbb{R}^{d\times d}, \ \ x, t \in X$$

Please cite this article as: R. Lin, G. Song and H. Zhang, Multi-task Learning in vector-valued reproducing kernel Banach spaces with the ℓ^1 norm, Journal of Complexity (2020) 101514, https://doi.org/10.1016/j.jco.2020.101514.

where the series convergent absolutely and uniformly on $X \times X$. The assumption $f_\rho \in ran(L_{\mathbf{K}}^s)$ implies that there exists $h = \sum_{j\in\mathbb{N}} a_j\phi_j \in L^2(X, \rho_X; \mathbb{R}^d)$ such that

$$f_\rho = L_{\mathbf{K}}^s(h) = \sum_{j\in\mathbb{N}} \lambda_j^s a_j\phi_j. \tag{5.6}$$

We shall construct a new RKBS $\mathsf{B}_{\mathbf{K}}$ with the reproducing kernel $\mathbf{K}$ to accommodate $ranL_{\mathbf{K}}$. Moreover, we will have $\mathcal{B}_{\mathbf{K}} \subseteq \mathsf{B}_{\mathbf{K}}$. To do so, the last requirement we needed is that $\mathbf{K}$ satisfies the denseness condition

$$\overline{\text{span}}\{\mathbf{K}(x, \cdot)y : x \in X, y \in \mathbb{R}^d\} = C(X; \mathbb{R}^d), \tag{5.7}$$

where $C(X; \mathbb{R}^d)$ denotes the Banach space of bounded and continuous vector-valued functions on the compact metric space X with the maximum norm

$$\|f\|_{C(X;\mathbb{R}^d)} := \sup_{x\in X} \|f(x)\|_2, \ f \in C(X; \mathbb{R}^d).$$

The matrix-valued positive definite kernel $\mathbf{K}$ satisfying (5.7) is called a universal multi-task kernel [5]. Denote by $\mathcal{M}(X; \mathbb{R}^d)$ the Banach space of regular vector measures of bounded variation on X (see, for instance, [22]). By the Riesz representation theorem [22], every continuous linear functional T on $C(X; \mathbb{R}^d)$ is represented by a unique regular vector measure $\mu \in \mathcal{M}(X; \mathbb{R}^d)$ in the sense that

$$T(f) = \int_X \langle f(x), d\mu(x)\rangle \text{ for all } f \in C(X; \mathbb{R}^d),$$

and

$$\sup_{f\in C(X;\mathbb{R}^d), \|f\|_{C(X;\mathbb{R}^d)}\le 1} |T(f)| = \|\mu\|_{\mathcal{M}(X;\mathbb{R}^d)}$$

where $\|\mu\|_{\mathcal{M}(X;\mathbb{R}^d)}$ denotes the total variation of the vector measure μ.

For every $\mu \in \mathcal{M}(X; \mathbb{R}^d)$, we define the vector-valued function $\mathbf{K}_\mu : X \to \mathbb{R}^d$ by

$$y^\top \mathbf{K}_\mu(x) := \int_X \langle \mathbf{K}(x, t)y, d\mu(t)\rangle. \ x \in X, \ y \in \mathbb{R}^d. \tag{5.8}$$

Notice that for every $x \in X$ and $\mu \in \mathcal{M}(X; \mathbb{R}^d)$, the functional

$$y \in \mathbb{R}^d \mapsto \int_X \langle \mathbf{K}(x, t)y, d\mu(t)\rangle \in \mathbb{R}$$

is continuous on $\mathbb{R}^d$. By (5.7) and the Riesz representation theorem for the Hilbert space $\mathbb{R}^d$, there exists a unique element denoted by $\mathbf{K}_\mu(x)$ in $\mathbb{R}^d$ such that (5.8) is satisfied. The vector-valued function $\mathbf{K}_\mu$ can also be described by the weak-Bochner integral [13]

$$\mathbf{K}_\mu(x) = \int_X \mathbf{K}(x, t)d\mu(t), \ x \in X, \ \mu \in \mathcal{M}(X; \mathbb{R}^d). \tag{5.9}$$

We introduce the following linear space

$$\mathsf{B}_{\mathbf{K}} := \left\{\mathbf{K}_\mu : \mu \in \mathcal{M}(X; \mathbb{R}^d)\right\} \text{ with the norm } \|\mathbf{K}_\mu\|_{\mathsf{B}_{\mathbf{K}}} := \|\mu\|_{\mathcal{M}(X;\mathbb{R}^d)}, \tag{5.10}$$

where $\mathbf{K}_\mu$ is defined by (5.8) or (5.9). By the denseness condition (5.7), the norm $\|\cdot\|_{\mathsf{B}}$ is well-defined. Furthermore, $\mathsf{B}_{\mathbf{K}}$ is a Banach space as it is isometrically isomorphic to the Banach space $\mathcal{M}(X; \mathbb{R}^d)$. By (5.8), we compute for all $x \in X$ and $\mu \in \mathcal{M}(X; \mathbb{R}^d)$

$$\|\mathbf{K}_\mu(x)\|_2 = \sup_{\|y\|_2=1} y^\top \mathbf{K}_\mu(x) = \sup_{\|y\|_2=1} \int_X \langle \mathbf{K}(x, t)y, d\mu(t)\rangle \le \sup_{\|y\|_2=1} \|K(x, \cdot)y\|_{C(X;\mathbb{R}^d)} \|\mu\|_{\mathcal{M}(X;\mathbb{R}^d)}.$$

It follows that for all $x \in X$ and $\mu \in \mathcal{M}(X; \mathbb{R}^d)$

$$\|\mathbf{K}_\mu(x)\|_2 \le \sup_{t\in X} \|\mathbf{K}(x, t)\|_2 \|\mathbf{K}_\mu\|_{\mathsf{B}_{\mathbf{K}}}.$$

Please cite this article as: R. Lin, G. Song and H. Zhang, Multi-task Learning in vector-valued reproducing kernel Banach spaces with the ℓ^1 norm, Journal of Complexity (2020) 101514, https://doi.org/10.1016/j.jco.2020.101514.

By Definition 2.1, $\mathsf{B}_{\mathbf{K}}$ is an RKBS of vector-valued functions from X to $\mathbb{R}^d$ as every point evaluation functional on the space is continuous. Clearly, we obtain the inclusion relation

$$\mathcal{B}_{\mathbf{K}} \subseteq \mathsf{B}_{\mathbf{K}}$$

as $\ell^1_d(X)$ defined by (2.3) is a subset of $\mathcal{M}(X; \mathbb{R}^d)$.

For any function $f : X \to \mathbb{R}^d$, we denote the empirical risk with respect to the sample data $\boldsymbol{z} = \{(x_j, \boldsymbol{y}_j) : j \in \mathbb{N}_m\}$ by

$$\mathcal{E}_{\boldsymbol{z}}(f) := \frac{1}{m} \sum_{j=1}^{m} \|f(x_j) - \boldsymbol{y}_j\|_2^2.$$

Let g be an appropriate function in $\mathsf{B}_{\mathbf{K}}$ which will be specified later. Then the approximation error $\mathcal{E}(f_{\boldsymbol{z},\lambda}) - \mathcal{E}(f_\rho)$ can be decomposed into the sum of four quantities

$$\mathcal{E}(f_{\boldsymbol{z},\lambda}) - \mathcal{E}(f_\rho) = \mathcal{S}(\boldsymbol{z}, \lambda, g) + \mathcal{P}(\boldsymbol{z}, \lambda, g) + \mathcal{D}(\lambda, g) - \lambda \|f_{\boldsymbol{z},\lambda}\|_{\mathsf{B}_{\mathbf{K}}},$$

where the sampling error, the hypothesis error and the regularization error are respectively defined by

$$\begin{aligned}
&\mathcal{S}(\boldsymbol{z}, \lambda, g) := \mathcal{E}(f_{\boldsymbol{z},\lambda}) - \mathcal{E}_{\boldsymbol{z}}(f_{\boldsymbol{z},\lambda}) + \mathcal{E}_{\boldsymbol{z}}(g) - \mathcal{E}(g),\\
&\mathcal{P}(\boldsymbol{z}, \lambda, g) := (\mathcal{E}_{\boldsymbol{z}}(f_{\boldsymbol{z},\lambda}) + \lambda \|f_{\boldsymbol{z},\lambda}\|_{\mathsf{B}_{\mathbf{K}}}) - (\mathcal{E}_{\boldsymbol{z}}(g) + \lambda\beta_m \|g\|_{\mathsf{B}_{\mathbf{K}}}),\\
&\mathcal{D}(\lambda, g) := \mathcal{E}(g) - \mathcal{E}(f_\rho) + \lambda\beta_m \|g\|_{\mathsf{B}_{\mathbf{K}}}.
\end{aligned}$$

By the relaxed representer theorem (5.1) in $\mathsf{B}_{\mathbf{K}}$, we can choose $g \in \mathsf{B}_{\mathbf{K}}$ such that $\mathcal{P}(\boldsymbol{z}, \lambda, g) \le 0$. As a result, we have

$$\mathcal{E}(f_{\boldsymbol{z},\lambda}) - \mathcal{E}(f_\rho) \le \mathcal{S}(\boldsymbol{z}, \lambda, g) + \mathcal{D}(\lambda, g). \tag{5.11}$$

We restrict our attention to the case when $0 < s < 1$ as the other case $s \ge 1$ can be handled in a slightly different way. The estimation of the regularization error $\mathcal{D}(\lambda, g)$ in $\mathsf{B}_{\mathbf{K}}$ is similar to the one for Lemma 3.2 in [29].

Lemma 5.1. *If $0 < s < 1$ then*

$$\inf_{g \in \mathsf{B}_{\mathbf{K}}} \mathcal{D}(\lambda, g) \le \sqrt{d}\Big(\|h\|_{L^2(X,\rho_X;\mathbb{R}^d)} + \|h\|^2_{L^2(X,\rho_X;\mathbb{R}^d)}\Big)(\lambda\beta_m)^{\frac{2s}{1+s}}. \tag{5.12}$$

Proof. By (5.5) and (5.9), we have for each $\psi \in L^2(X, \rho; \mathbb{R}^d)$ that $L_{\mathbf{K}}(\psi) \in \mathsf{B}_{\mathbf{K}}$ and by the Cauchy–Schwartz inequality that

$$\begin{aligned}
\|L_{\mathbf{K}}\psi\|_{\mathsf{B}_{\mathbf{K}}} &= \int_X \|\psi(x)\|_1 d\rho_X(x) \le \Big(\int_X \|\psi(x)\|_1^2 d\rho_X(x)\Big)^{1/2}\\
&\le \sqrt{d}\Big(\int_X \|\psi(x)\|_2^2 d\rho_X(x)\Big)^{1/2} = \sqrt{d}\|\psi\|_{L^2(X,\rho_X;\mathbb{R}^d)}.
\end{aligned} \tag{5.13}$$

Note that $\|h\|^2_{L^2(X,\rho_X;\mathbb{R}^d)} = \sum_{j\in\mathbb{N}} a_j^2$, where $h = \sum_{j\in\mathbb{N}} a_j\phi_j$. If $\lambda_1 \le (\lambda\beta_m)^{\frac{1}{1+s}}$ then by (5.4) and (5.6)

$$\mathcal{D}(\lambda, 0) = \mathcal{E}(0) - \mathcal{E}(f_\rho) = \|f_\rho\|^2_{L^2(X,\rho_X;\mathbb{R}^d)} = \sum_{j\in\mathbb{N}} \lambda_j^{2s} a_j^2 \le (\lambda\beta_m)^{\frac{2s}{1+s}} \sum_{j\in\mathbb{N}} a_j^2 = (\lambda\beta_m)^{\frac{2s}{1+s}} \|h\|^2_{L^2(X,\rho_X;\mathbb{R}^d)},$$

which implies (5.12), where we have used the fact that $\{\phi_j : j \in \mathbb{N}\}$ is an orthonormal basis for $L^2(X, \rho_X; \mathbb{R}^d)$ in the last equality.

If $\lambda_1 > (\lambda\beta_m)^{\frac{1}{1+s}}$ then there exists some $N \in \mathbb{N}$ such that $\lambda_{N+1} < (\lambda\beta_m)^{\frac{1}{1+s}} \le \lambda_N \le \lambda_{N-1} \le \cdots \le \lambda_1$ as the eigenvalue λ_j decreases to zero as j tends to infinity. Notice that $s - 1 < 0$ for any

Please cite this article as: R. Lin, G. Song and H. Zhang, Multi-task Learning in vector-valued reproducing kernel Banach spaces with the ℓ^1 norm, Journal of Complexity (2020) 101514, https://doi.org/10.1016/j.jco.2020.101514.

$0 < s < 1$. Letting $\psi := \sum_{j\in\mathbb{N}_N} \lambda_j^{s-1} a_j \phi_j$, we have by (5.4), (5.6) and (5.13)

$$\begin{aligned}
\mathcal{D}(\lambda, L_{\mathbf{K}}\psi) &\le \|L_{\mathbf{K}}\psi - f_\rho\|^2_{L^2(X,\rho_X;\mathbb{R}^d)} + \lambda\beta_m\sqrt{d}\|\psi\|_{L^2(X,\rho_X;\mathbb{R}^d)} \\
&= \Big\| \sum_{j=N+1}^{\infty} \lambda_j a_j \phi_j \Big\|^2_{L^2(X,\rho_X;\mathbb{R}^d)} + \lambda\beta_m\sqrt{d} \Big\| \sum_{j\in\mathbb{N}_N} \lambda_j^{s-1} a_j \phi_j \Big\|_{L^2(X,\rho_X;\mathbb{R}^d)} \\
&= \sum_{j=N+1}^{\infty} \lambda_j^{2s} a_j^2 + \lambda\beta_m\sqrt{d}\Big(\sum_{j\in\mathbb{N}_m} \lambda_j^{2(s-1)} a_j^2\Big)^{1/2} \\
&\le (\lambda\beta_m)^{\frac{2s}{1+s}} \sum_{j=N+1}^{\infty} a_j^2 + \sqrt{d}\lambda\beta_m(\lambda\beta_m)^{\frac{s-1}{1+s}} \Big(\sum_{j\in\mathbb{N}_N} a_j^2\Big)^{1/2} \\
&\le (\lambda\beta_m)^{\frac{2s}{1+s}} \|h\|^2_{L^2(X,\rho_X;\mathbb{R}^d)} + \sqrt{d}(\lambda\beta_m)^{\frac{2s}{1+s}} \|h\|_{L^2(X,\rho_X;\mathbb{R}^d)}.
\end{aligned}$$

which implies (5.12). The proof is hence complete. □

The sampling error $\mathcal{S}(\boldsymbol{z}, \lambda, g)$ can be obtained by following the approach in [29]. Combining Lemma 5.1 and Lemmas 3.3 and 3.4 in [29], we are able to present an error estimate for (5.11).

Theorem 5.2. *For all $0 < \delta < 1$ and $0 < s < 1$, there exists a positive constant C_δ such that with confidence $1 - \delta$, we have*

$$\|f_{\boldsymbol{z},\lambda} - f_\rho\|^2_{L^2(X,\rho_X;\mathbb{R}^d)} \le C_\delta \Big(\sqrt{d}(\lambda\beta_m)^{\frac{2s}{1+s}} + \frac{\log\frac{2}{\delta}}{m}(\lambda\beta_m)^{\frac{2s-2}{1+s}} + \frac{\log\frac{2}{\delta}}{\sqrt{m}}(\lambda\beta_m)^{\frac{2s-1}{1+s}} + \frac{\log\frac{2}{\delta} + \log(1+m)}{(\lambda\beta_m)^2} m^{-\frac{1}{1+\theta}} \Big),$$

where $\theta > 0$ is a positive constant related to the assumptions on the kernel $\mathbf{K}$ and the input space.

We make some comments. If β_m^2 does not cancel the decay of the term $m^{-\frac{1}{1+\theta}}$, one can get a satisfactory learning rate when λ is appropriately chosen. We discuss two cases below:

(i) If β_m is uniformly bounded, then with a large confidence $1 - \delta$ we have

$$\|f_{\boldsymbol{z},\lambda} - f_\rho\|^2_{L^2(X,\rho_X;\mathbb{R}^d)} \le C_\delta \sqrt{d} m^{-\frac{s}{1+2s}\frac{1}{1+\theta}} \log\frac{2+2m}{\delta}.$$

(ii) If $\beta_m \le cm^\alpha$ for some positive constants c and $\alpha < \frac{1}{2+2\theta}$, then with a large confidence $1 - \delta$ we have

$$\|f_{\boldsymbol{z},\lambda} - f_\rho\|^2_{L^2(X,\rho_X;\mathbb{R}^d)} \le C_\delta \sqrt{d} m^{-\frac{s}{1+2s}(\frac{1}{1+\theta} - 2\alpha)} \log\frac{2+2m}{\delta}.$$

5.2. Characterization

We shall prove a weaker condition on the Lebesgue constant for the relaxed representer theorem (5.1). To this end, we first show a connection between the relaxed representer theorem for regularization networks and that for the minimal norm interpolation problem.

Lemma 5.3. *If there exists some $\beta_m \ge 1$ such that for all $\boldsymbol{y} \in \mathbb{R}^{md}$*

$$\min_{f\in\mathcal{I}_{\mathbf{x}}(\boldsymbol{y})} \|f\|_{\mathcal{B}_{\mathbf{K}}} \ge \frac{1}{\beta_m} \min_{\mathcal{I}_{\mathbf{x}}(\boldsymbol{y})\cap\mathcal{S}^{\mathbf{x}}} \|f\|_{\mathcal{B}_{\mathbf{K}}} \tag{5.14}$$

then the relaxed linear representer theorem (5.1) holds true for any continuous loss function L and any regularization parameter λ.

Proof. Suppose f_0 is a minimizer of $\min_{f\in\mathcal{B}_{\mathbf{K}}} L(f(\mathbf{x}), \boldsymbol{y}) + \lambda\beta_m\|f\|_{\mathcal{B}_{\mathbf{K}}}$. Let g be the unique function in $\mathcal{S}^{\mathbf{x}}$ that interpolates f_0 at $\mathbf{x}$, namely, $g(\mathbf{x}) = f_0(\mathbf{x})$. By (5.14), $\|g\|_{\mathcal{B}_{\mathbf{K}}} \le \beta_m\|f_0\|_{\mathcal{B}_{\mathbf{K}}}$. It implies

$$L(g(\mathbf{x}), \boldsymbol{y}) + \lambda\|g\|_{\mathcal{B}_{\mathbf{K}}} \le L(f_0(\mathbf{x}), \boldsymbol{y}) + \lambda\beta_m\|f_0\|_{\mathcal{B}_{\mathbf{K}}},$$

which finishes the proof. □

Please cite this article as: R. Lin, G. Song and H. Zhang, Multi-task Learning in vector-valued reproducing kernel Banach spaces with the ℓ^1 norm, Journal of Complexity (2020) 101514, https://doi.org/10.1016/j.jco.2020.101514.

The next result gives a characterization for condition (5.14). It is a weaker version of the Lebesgue constant condition (3.4).

Theorem 5.4. *Eq. (5.14) holds true for all $\boldsymbol{y} \in \mathbb{R}^{md}$ if and only if*

$$\sup_{t\in X} \|\mathbf{K}[\mathbf{x}]^{-1}\mathbf{K}_{\mathbf{x}}(t)\|_1 \le \beta_m. \tag{5.15}$$

Proof. Remember that the set $\mathcal{I}_{\mathbf{x}}(\boldsymbol{y}) \cap \mathcal{S}^{\mathbf{x}}$ consists of only one function $f_0 := \mathbf{K}^{\mathbf{x}}(\cdot)\mathbf{K}[\mathbf{x}]^{-1}\boldsymbol{y}$. Let g be an arbitrary function in $\mathcal{I}_{\mathbf{x}}(\boldsymbol{y}) \cap \mathcal{B}_0$, where $\mathcal{B}_0$ is defined in (2.6). By adding sampling points and assigning the corresponding coefficients to be zero if necessary, we may assume $g \in \mathcal{S}^{\mathbf{x}\cup \boldsymbol{t}} \cap \mathcal{I}_{\mathbf{x}}(\boldsymbol{y})$ for a set of new points $\boldsymbol{t} := \{t_k \in X : k \in \mathbb{N}_n\}$ disjoint with $\mathbf{x}$. Let $\boldsymbol{b} := g(\boldsymbol{t})$, and denote by $\mathbf{K}[\boldsymbol{t}, \mathbf{x}]$ and $\mathbf{K}[\mathbf{x}, \boldsymbol{t}]$ the $md \times nd$ and $nd \times md$ matrices given by

$$(\mathbf{K}[\boldsymbol{t}, \mathbf{x}])_{jk} := \mathbf{K}(t_k, x_j), \quad j \in \mathbb{N}_m, k \in \mathbb{N}_n, \quad \text{and} \quad (\mathbf{K}[\mathbf{x}, \boldsymbol{t}])_{jk} := \mathbf{K}(x_k, t_j): \quad j \in \mathbb{N}_n, k \in \mathbb{N}_m.$$

It then follows

$$\|g\|_{\mathcal{B}_{\mathbf{K}}} = \left\| \begin{bmatrix} \mathbf{K}[\mathbf{x}] & \mathbf{K}[\boldsymbol{t}, \mathbf{x}] \\ \mathbf{K}[\mathbf{x}, \boldsymbol{t}] & \mathbf{K}[\boldsymbol{t}] \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{b} \end{bmatrix} \right\|_1 = \left\| \begin{bmatrix} \mathbf{K}[\mathbf{x}]^{-1}\boldsymbol{y} - \mathbf{K}[\mathbf{x}]^{-1}\mathbf{K}[\boldsymbol{t}, \mathbf{x}]\tilde{\boldsymbol{b}} \\ \tilde{\boldsymbol{b}} \end{bmatrix} \right\|_1, \tag{5.16}$$

where

$$\tilde{\boldsymbol{b}} := \Big(\mathbf{K}[\boldsymbol{t}] - \mathbf{K}[\mathbf{x}, \boldsymbol{t}]\mathbf{K}[\mathbf{x}]^{-1}\mathbf{K}[\boldsymbol{t}, \mathbf{x}]\Big)^{-1}(\boldsymbol{b} - \mathbf{K}[\mathbf{x}, \boldsymbol{t}]\mathbf{K}[\mathbf{x}]^{-1}\boldsymbol{y}).$$

Note that as $\boldsymbol{b}$ is allowed to take any vector in $\mathbb{R}^{nd}$, so is $\tilde{\boldsymbol{b}}$.

If (5.14) holds true for all $\boldsymbol{y} \in \mathbb{R}^{md}$ then we choose $\boldsymbol{t}$ to be a singleton $\{t_1\}$, $\tilde{\boldsymbol{b}} = \boldsymbol{e}_j$, and $\boldsymbol{y} = \mathbf{K}[t_1, \mathbf{x}]\boldsymbol{e}_j = \mathbf{K}_{\mathbf{x}}(t_1)\boldsymbol{e}_j$ for some $j \in \mathbb{N}_d$. It follows

$$1 = \left\| \begin{bmatrix} \mathbf{0}_{md} \\ \boldsymbol{e}_j \end{bmatrix} \right\|_1 \ge \frac{1}{\beta_m}\|f_0\|_{\mathcal{B}_{\mathbf{K}}} = \frac{1}{\beta_m}\left\|\mathbf{K}[\mathbf{x}]^{-1}\boldsymbol{y}\right\|_1 = \frac{1}{\beta_m}\left\|\mathbf{K}[\mathbf{x}]^{-1}\mathbf{K}_{\mathbf{x}}(t_1)\boldsymbol{e}_j\right\|_1.$$

As $j \in \mathbb{N}_d$ is arbitrary, we get (5.15).

Conversely, suppose that (5.15) is satisfied. We need to show that for all $g \in \mathcal{I}_{\mathbf{x}}(\boldsymbol{y})$

$$\|g\|_{\mathcal{B}_{\mathbf{K}}} \ge \frac{1}{\beta_m}\|f_0\|_{\mathcal{B}_{\mathbf{K}}} = \frac{1}{\beta_m}\left\|\mathbf{K}[\mathbf{x}]^{-1}\boldsymbol{y}\right\|_1.$$

We shall discuss the case when $g \in \mathcal{I}_{\mathbf{x}}(\boldsymbol{y}) \cap \mathcal{B}_0$ only, as the general case will then follow from the same arguments as those in the last paragraph of the proof of Theorem 3.2. Let $g \in \mathcal{I}_{\mathbf{x}}(\boldsymbol{y}) \cap \mathcal{B}_0$ with the norm in Eq. (5.16). If $\|\mathbf{K}[\mathbf{x}]^{-1}\boldsymbol{y}\|_1 \le \beta_m\|\tilde{\boldsymbol{b}}\|_1$, it is direct to observe that

$$\|g\|_{\mathcal{B}_{\mathbf{K}}} \ge \|\tilde{\boldsymbol{b}}\|_1 \ge \frac{1}{\beta_m}\left\|\mathbf{K}[\mathbf{x}]^{-1}\boldsymbol{y}\right\|_1.$$

On the other hand, if $\|\mathbf{K}[\mathbf{x}]^{-1}\boldsymbol{y}\|_1 \ge \beta_m\|\tilde{\boldsymbol{b}}\|_1$, then by (5.15) we have

$$\begin{aligned}
\|g\|_{\mathcal{B}_{\mathbf{K}}} &\ge \|\mathbf{K}[\mathbf{x}]^{-1}\boldsymbol{y}\|_1 - \|\mathbf{K}[\mathbf{x}]^{-1}\mathbf{K}[\boldsymbol{t}, \mathbf{x}]\tilde{\boldsymbol{b}}\|_1 + \|\tilde{\boldsymbol{b}}\|_1 \\
&\ge \|\mathbf{K}[\mathbf{x}]^{-1}\boldsymbol{y}\|_1 - \left(\max_{k\in\mathbb{N}_n}\|\mathbf{K}[\mathbf{x}]^{-1}\mathbf{K}_{\mathbf{x}}(t_k)\|_1\right)\|\tilde{\boldsymbol{b}}\|_1 + \|\tilde{\boldsymbol{b}}\|_1 \\
&\ge \|\mathbf{K}[\mathbf{x}]^{-1}\boldsymbol{y}\|_1 - (\beta_m - 1)\|\tilde{\boldsymbol{b}}\|_1 \\
&> \|\mathbf{K}[\mathbf{x}]^{-1}\boldsymbol{y}\|_1 - (\beta_m - 1)\frac{1}{\beta_m}\|\mathbf{K}[\mathbf{x}]^{-1}\boldsymbol{y}\|_1 \\
&= \frac{1}{\beta_m}\|\mathbf{K}[\mathbf{x}]^{-1}\boldsymbol{y}\|_1.
\end{aligned}$$

The proof is hence complete. □

In the rest of this section, we discuss examples of admissible kernels that satisfy the weaker Lebesgue constant condition (5.15). The Lebesgue constants can measure the stability of kernel-based interpolation. Toward this research interest, it was proved in [18] that the Lebesgue constant

Please cite this article as: R. Lin, G. Song and H. Zhang, Multi-task Learning in vector-valued reproducing kernel Banach spaces with the ℓ^1 norm, Journal of Complexity (2020) 101514, https://doi.org/10.1016/j.jco.2020.101514.

associated with surface splines and Sobolev splines on a compact domain is uniformly bounded for quasi-uniform input points (see, Theorem 4.6 therein). A set S of points in a compact domain $\Omega \subseteq \mathbb{R}^n$ is said to be quasi-uniform with uniformity constant $\gamma > 1$ if

$$\frac{1}{\gamma} q_S \leq h_{S,\Omega} \leq \gamma q_S$$

where

$$h_{S,\Omega} := \sup_{x\in\Omega} \min_{x_j\in S} \|x - x_j\|_2 \text{ and } q_S := \frac{1}{2} \min_{x_i,x_j\in S, x_i\neq x_j} \|x_i - x_j\|_2$$

denote the fill distance and the separation distance, respectively. For a class of translation invariant kernels $K(x, x') = \phi(x - x')$, $x, x' \in \mathbb{R}^n$, the paper [10] showed that if the Fourier transform $\hat{\phi}$ of ϕ satisfies

$$0 < c_1(1 + \|\xi\|_2^2)^{-\tau} \leq \hat{\phi}(\xi) \leq c_2(1 + \|\xi\|_2^2)^{-\tau}$$

at infinity for some positive constants c_1, c_2, M and $\tau > \frac{n}{2}$, then the Lebesgue constant of K for quasi-uniform inputs is bounded by a multiple of $\sqrt{m}$. This includes, for example, Poisson radial functions, Matérn kernels and Wendland's compactly supported kernels [10,34]. In particular, the RKHS with the Laplacian kernel $K(x, x') = e^{-\|x-x'\|_2}$, $x, x' \in \mathbb{R}^n$ is norm equivalent to the Sobolev space of the smooth order $(n + 1)/2$ [34].

6. Numerical experiments

We shall perform four numerical experiments to show that the regularization network (3.6) in vector-valued RKBSs (VVRKBS) with the ℓ^1 norm is indeed able to yield sparsity compared to the one in vector-valued RKHSs (VVRKHS). Moreover, we can achieve better numerical performance for multi-task learning in the constructed spaces.

To begin with, let us compare the Lebesgue constant of the three aforementioned kernels in Section 4. They are the Laplacian kernel $e^{-\|x-x'\|_2}$, the exponential kernel $e^{-\|x-x'\|_1}$, and the Gaussian kernel $e^{-\|x-x'\|_2^2}$, $x, x' \in \mathbb{R}^n$. Fix $n = 2$. Notice that grid points

$$\left\{\left(-1 + \frac{2}{M-1} i, -1 + \frac{2}{M-1} j\right) : i, j = 0, 1, \ldots, M-1\right\}, \ M \in \mathbb{N}, \tag{6.1}$$

in $[-1, 1]^2$ are a rather extreme class of quasi-uniform points. Hence, to obtain a set of m quasi-uniform points in $[-1, 1]^2$, we randomly sample half of M^2 grid points defined by (6.1). In this case, we have $m = M^2/2$. We repeat the above sampling process 10 times. The average Lebesgue constant of the Gaussian kernel is not available for $m \geq 200$ as the corresponding kernel matrix is close to singular. We use N/A to indicate those cases. The average Lebesgue constant of three kernels was listed in Table 6.1 when $M = 10, 20, 30, 40, 50, 60$ or $m = 50, 200, 450, 800, 1250, 1800$. The numerical results in Table 6.1 show that Lebesgue constants of the Laplacian kernel and the exponential kernel grow moderately as the number of points increases. Hence, both of them are appropriate kernels admissible for the construction of vector-valued RKBSs with the ℓ^1 norm. By contrast, Lebesgue constants of the Gaussian kernel are quite large even for a small number of points. The same observations were mentioned in [10] that Lebesgue constants of the Gaussian kernel do not seem to be uniformly bounded. This results from the infinite smoothness of the Gaussian kernel.

Here and subsequently, we choose both the scaled Laplacian kernel and the scaled exponential kernel for the sparse multitask learning. Specifically, the multi-task kernel for numerical experiments takes the form

$$\mathbf{K}(x, x') = K_r(x, x')\mathbb{A}, \ x, x' \in \mathbb{R}^n, \tag{6.2}$$

where $\mathbb{A}$ denotes a $d \times d$ positive definite symmetric matrix, and K_r in (6.2) stands for the scaled Laplacian kernel $K_r(x, x') = e^{-\|x-x'\|_2/r}$, $x, x' \in \mathbb{R}^n$ or the scaled exponential kernel $e^{-\|x-x'\|_1/r}$,

Please cite this article as: R. Lin, G. Song and H. Zhang, Multi-task Learning in vector-valued reproducing kernel Banach spaces with the ℓ^1 norm, Journal of Complexity (2020) 101514, https://doi.org/10.1016/j.jco.2020.101514.

Table 6.1
Lebesgue constants of the Laplacian kernel $e^{-\|x-x'\|_2}$, the exponential kernel $e^{-\|x-x'\|_1}$, and the Gaussian kernel $e^{-\|x-x'\|_2^2}$ on a set of m quasi-uniform distributed points of $[-1, 1]^2$.

Kernels	m = 50	m = 200	m = 450	m = 800	m = 1250	m = 1800
Laplacian Kernel	1.401243	1.912003	2.482223	2.850599	3.262301	3.632585
Exponential Kernel	3.155898	4.770347	6.138831	6.291278	6.374352	7.512769
Gaussian Kernel	490.870640	N/A	N/A	N/A	N/A	N/A

$x, x' \in \mathbb{R}^n$, where $r > 0$. Observe that the function value of $K_r(x, x')$, $x, x' \in \mathbb{R}^n$ decays rapidly for a large $\|x - x'\|_2$ or $\|x - x'\|_1$. Hence, the choice of r should depend upon the data.

Let $\mathcal{B}_{\mathbf{K}}$ be the associated vector-valued RKBS with the ℓ^1 norm and $\mathcal{H}_{\mathbf{K}}$ the vector-valued RKHS with reproducing kernel $\mathbf{K}$ taking the form (6.2). For the sake of simplicity, the square loss function will be used. We compare the following regularization network models

$$\min_{f\in\mathcal{B}_{\mathbf{K}}} \|f(\mathbf{x}) - \boldsymbol{y}\|_2^2 + \lambda\|f\|_{\mathcal{B}_{\mathbf{K}}}$$

and

$$\min_{f\in\mathcal{H}_{\mathbf{K}}} \|f(\mathbf{x}) - \boldsymbol{y}\|_2^2 + \lambda\|f\|_{\mathcal{H}_{\mathbf{K}}}^2.$$

By the relaxed linear representer theorem for $\mathcal{B}_{\mathbf{K}}$ and the linear representer theorem for $\mathcal{H}_{\mathbf{K}}$, the minimizers of the previous models are

$$\mathbf{K}^{\mathbf{x}}(\cdot)\mathbf{b} = \mathbb{A}\sum_{j=1}^m K_r(x, x_j)\mathbf{b}_j \text{ with } \mathbf{b} := \arg\min_{\boldsymbol{c}\in\mathbb{R}^{md}} \left\{\|\mathbf{K}[\mathbf{x}]\boldsymbol{c} - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{c}\|_1\right\}$$

and

$$\mathbf{K}^{\mathbf{x}}(\cdot)\mathbf{h} = \mathbb{A}\sum_{j=1}^m K_r(x, x_j)\mathbf{h}_j \text{ with } \mathbf{h} := \arg\min_{\boldsymbol{c}\in\mathbb{R}^{md}} \left\{\|\mathbf{K}[\mathbf{x}]\boldsymbol{c} - \boldsymbol{y}\|_2^2 + \lambda\boldsymbol{c}^\top\mathbf{K}[\mathbf{x}]\boldsymbol{c}\right\},$$

respectively. The ℓ^1-regularized least square regression problem about $\mathbf{b}$ does not have a closed form solution. We employ the alternating direction method of multipliers (ADMM) [3] to solve it. We refer to the link (https://web.stanford.edu/~boyd/papers/admm/lasso/lasso.html) for the corresponding code of ADMM. The number of maximum iterations for ADMM is 5000. The coefficient vector $\mathbf{h}$ has the closed form $\mathbf{h} = (\mathbf{K}[\mathbf{x}] + \lambda I_{md})^{-1}\boldsymbol{y}$. The regularization parameter λ for each model will be optimally chosen from $\{10^j : j = -4, -3, \ldots, 1\}$ so that the mean square error (MSE) between predicted values and the observed values $\boldsymbol{y}$ will be minimized. We run all the experiments on a computer with a single NVIDIA Quadro P2000.

6.1. Experiment 1

The first numerical experiment is for synthetic data. In this experiment, we set $r = 1$ in (6.2). The training data is generated by a function $f : \mathbb{R}^2 \to \mathbb{R}^3$ defined as

$$f(x) := \mathbb{A}\sum_{j=1}^5 K_r(x, x_j)\boldsymbol{c}_j,\ x \in \mathbb{R}^2,$$

where $x_1 = (-1/3, -1/3)$, $x_2 = (-1/2, 1/2)$, $x_3 = (0, 0)$, $x_4 = (1/3, 2/3)$, $x_5 = (2/3, 2/3)$,

$$[\boldsymbol{c}_1, \boldsymbol{c}_2, \boldsymbol{c}_3, \boldsymbol{c}_4, \boldsymbol{c}_5] := \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 1 & 2 & 1 \\ 2 & 1 & 1 & 1 & 2 \end{bmatrix} \text{ and } \mathbb{A} := \begin{bmatrix} 1 & e^{-1} & e^{-2} \\ e^{-1} & 1 & e^{-1} \\ e^{-2} & e^{-1} & 1 \end{bmatrix}.$$

Let $\mathbf{x}$ be the set of 400 grid points in $[-1, 1]^2$ defined by (6.1) when $M = 20$, and use the output vector $\boldsymbol{y}$ at $\mathbf{x}$ which are then disturbed by some noise. We then compare the performance measured

Please cite this article as: R. Lin, G. Song and H. Zhang, Multi-task Learning in vector-valued reproducing kernel Banach spaces with the ℓ^1 norm, Journal of Complexity (2020) 101514, https://doi.org/10.1016/j.jco.2020.101514.

Table 6.2
Comparison of the least square regularization for synthetic data in the vector-valued RKBS with the ℓ^1 norm and in the vector-valued RKHS.

Kernels	Models	Gaussian noise		Uniform noise	
		MSE	Sparsity (Max)	MSE	Sparsity (Max)
Laplacian	VVRKBS	0.0020	68.5 (105)	0.0013	96.4 (102)
Kernel	VVRKHS	0.0039	1200 (1200)	0.0024	1200 (1200)
Exponential	VVRKBS	0.0020	59.8 (97)	0.0012	83.9 (90)
Kernel	VVRKHS	0.0040	1200 (1200)	0.0028	1200 (1200)

Table 6.3
Comparison of the classification for digits $6, 8, 9$ in the VVRKBS with the ℓ^1 norm and in the VVRKHS for the Laplacian kernel with $r = 10$ and the exponential kernel with $r = 70$.

Kernels	Models	Training accuracy	Sparsity	Testing accuracy
Laplacian	VVRKBS	100%	1735	98.05%
Kernel	VVRKHS	100%	5550	98.44%
Exponential	VVRKBS	100%	1455	98.44%
Kernel	VVRKHS	100%	5550	98.44%

by the MSE and the sparsity for two regularization models. The sparsity is measured by the number of nonzero components in the coefficient vectors $\mathbf{b}$ and $\mathbf{h}$. We test both models with two types of noise: Gaussian noise with variance 0.01, and uniform noise in $[-0.1, 0.1]$. For each type of noise, we run 10 times of numerical experiments and compute the average MSE, the average sparsity, and the maximum sparsity in the 10 experiments. We conclude that the regularization network in the vector-valued RKBS with the ℓ^1 norm outperforms the classical one for synthetic data. At the same time, the sparsity of data representation can be substantially promoted in our constructed spaces with the Laplacian kernel and the exponential kernel. The results are listed in Table 6.2.

6.2. Experiment 2

The second experiment is for the MNIST database (http://yann.lecun.com/exdb/mnist/) of handwritten digits from the machine learning repository. It possesses a training set of 7291 examples and a testing set of 2007 examples. Each digit is a vector in $[0, 1]^{255}$. Limited by the computation resource, we choose two classes $\{6, 8, 9\}$ and $\{2, 3, 7\}$ of handwritten digits which are relatively difficult to distinguish. For a set $\boldsymbol{z} \subseteq \mathbf{x}$ of 100 randomly chosen examples, $\{\|x - x'\|_2 : x, x' \in \boldsymbol{z}\}$ has mean 7.37 and standard deviation 1.50 and $\{\|x - x'\|_1 : x, x' \in \boldsymbol{z}\}$ has mean 74.60 and standard deviation 21.34. Therefore, we choose the $r = 10$ for the scaled Laplacian kernel and $r = 70$ for the scaled exponential kernel in the second experiment.

We first consider three digits 6,8, and 9. In this case, we have a set $\mathbf{x}$ of 1850 examples for training, and a set of 513 examples for testing. For the multi-task learning, three labels 6, 8, and 9 are transferred to the vectors $(1, 0, 0)^\top$, $(0, 1, 0)^\top$, and $(0, 0, 1)^\top$, respectively. We compute the prediction accuracy for training data and the sparsity of coefficients for both models. Then we apply learned coefficients of both models to testing data. The accuracy is measured by labels that are correctly predicted by models. The results are listed in Table 6.3. To be more specific, we pick out the digits from the testing data that are misclassified by models. We number the testing data with numbers from 1 to 513. For instance, the numbers of 8 misclassified digits, predicted labels, and true labels for each model with the exponential kernel are listed in Table 6.4. Both regularization models classify the numbers 56, 58, 73, 113, 212, 430, and 480 incorrectly. But the number 255 is misclassified only in VVRKBS and the number 64 is misclassified only in VVRKHS. The original images of 9 misclassified digits for both models with the scaled exponential kernel are displayed in Fig. 6.1. The numerical performances for both models are comparable.

Next, we study another class of three digits 2, 3, and 7. In this case, we have a set $\mathbf{x}$ of 2034 examples for training, and a set of 511 examples for testing. The results are listed in Table 6.5. Again,

Please cite this article as: R. Lin, G. Song and H. Zhang, Multi-task Learning in vector-valued reproducing kernel Banach spaces with the ℓ^1 norm, Journal of Complexity (2020) 101514, https://doi.org/10.1016/j.jco.2020.101514.

Fig. 6.1. Misclassified digits 6, 8, 9 for both regularization network models with the exponential kernel.

Table 6.4
Misclassified digits 6, 8, 9 in the VVRKBS with the ℓ^1 norm and in the VVRKHS for the exponential kernel with $r = 70$.

VVRKBS			VVRKHS		
Numbers	True labels	Predicted labels	Numbers	True labels	Predicted labels
56	6	8	56	6	8
58	8	6	58	8	6
73	9	8	**64**	8	6
113	8	9	73	9	8
212	8	9	113	8	9
255	8	9	212	8	9
430	8	9	430	8	9
480	9	8	480	9	8

Table 6.5
Comparison of the classification for digits 2, 3, 7 in the VVRKBS with the ℓ^1 norm and in the VVRKHS for the Laplacian kernel with $r = 10$ and the exponential kernel with $r = 70$.

Kernels	Models	Training accuracy	Sparsity	Testing accuracy
Laplacian	VVRKBS	100%	1778	97.46%
Kernel	VVRKHS	100%	6102	97.46%
Exponential	VVRKBS	100%	1607	97.26%
Kernel	VVRKHS	100%	6102	97.06%

we only present the misclassified examples for two models with the scaled exponential kernel. We number the testing data with numbers from 1 to 511. Both regularization models misclassify the numbers 38, 68, 119, 122, 203, 226, 232, 251, 277, 287, 392, 413, 419, 506. In addition, the VVRKHS classifies one more number 365 incorrectly. The original images of 15 misclassified digits for both models with the scaled exponential kernel are displayed in Fig. 6.2. Numerical results demonstrate that both models are comparable.

6.3. Experiment 3

The third experiment is for the Iris database (http://archive.ics.uci.edu/ml/datasets/Iris). This is the best known database to be found in the pattern recognition literature. The data set contains three classes of 50 instances each, where each class refers to a type of Iris plant. For each class, we choose 40 instances as the training set and the remaining 10 instances as the testing set. As a result, we have a training set of 120 instances and a testing set of 30 instances. Each Iris plant is measured by four attributes, namely, sepal length, sepal width, petal length, and petal width.

For the multi-task learning, the three classes, Iris setosa, Iris versicolor, and Iris virginica are transferred to the vectors $(1, 0, 0)^\top$, $(0, 1, 0)^\top$, and $(0, 0, 1)^\top$, respectively. For the Iris training set $\mathbf{x}$, $\{\|x - x'\|_2 : x, x' \in \mathbf{x}\}$ has mean 2.56 and standard deviation 1.66 and $\{\|x - x'\|_1 : x, x' \in \mathbf{x}\}$ has mean 4.29 and standard deviation 2.77. Therefore, we simply choose $r = 5$ for both the scaled Laplacian kernel and the scaled exponential kernel. The numerical results are listed in Table 6.6.

Please cite this article as: R. Lin, G. Song and H. Zhang, Multi-task Learning in vector-valued reproducing kernel Banach spaces with the ℓ^1 norm, Journal of Complexity (2020) 101514, https://doi.org/10.1016/j.jco.2020.101514.

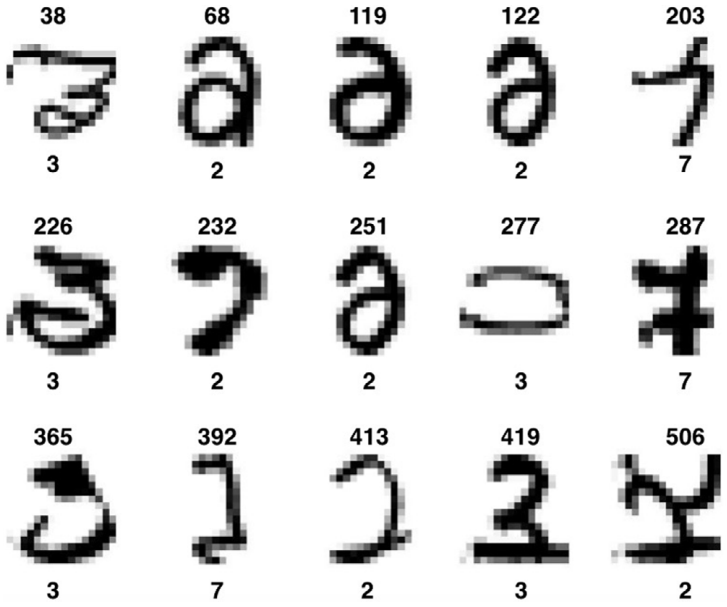

Fig. 6.2. Misclassified digits 2, 3, 7 for both regularization network models with the exponential kernel.

Table 6.6
Comparison of the classification for Iris database in the VVRKBS with the ℓ^1 norm and in the VVRKHS for the Laplacian kernel and the exponential kernel with $r = 5$.

Kernels	Models	Training accuracy	Sparsity	Testing accuracy
Laplacian	VVRKBS	100%	142	100%
Kernel	VVRKHS	100%	360	100%
Exponential	VVRKBS	100%	149	100%
Kernel	VVRKHS	100%	360	100%

6.4. Experiment 4

The last experiment is to classify an email as spam or non-spam. The database possesses 4601 instances (http://archive.ics.uci.edu/ml/datasets/Spambase). Among them, 1813 emails are spam. Each instance is a row vector in $\mathbb{R}^{58}$. The last component of the vector indicates whether the email is considered spam or not. For the multi-task learning, we coded spam as the vector $(1, 0)^\top$ and non-spam as $(0, 1)^\top$. A test set of size 1536 was randomly chosen, leaving the rest 3065 observations in the training set. The positive definite matrix $\mathbb{A}$ in (6.2) is given by

$$\mathbb{A} := \begin{bmatrix} 1 & 0.1 \\ 0.1 & 1 \end{bmatrix}.$$

Note that the scalar r in (6.2) is again determined by the data set. Let us set $r = 500$ for the scaled exponential kernel.

The numerical results are listed in Tables 6.7 and 6.8. We should mention the number of training set is the same as the one in numerical experiments for the spam dataset in [19]. The numerical results in [19] show that the test error rates of the linear logistic regression, the additive logistic regression, and the tree-based method for the spam dataset are 7.6%, 5.5% and 9.3%, respectively. By comparison, the testing accuracy of regularization networks for the spam dataset in the VVRKBS with the ℓ^1 norm and in the VVRKHS are 6.05% and 5.66%, respectively. The error analysis is presented in Table 6.8. Clearly, the accuracy of our kernel-based method is comparable to the aforementioned three methods. Furthermore, the numerical result in Table 6.7 implies that the spam predictor possesses a sparse representation in the VVRKBS with the ℓ^1 norm.

Please cite this article as: R. Lin, G. Song and H. Zhang, Multi-task Learning in vector-valued reproducing kernel Banach spaces with the ℓ^1 norm, Journal of Complexity (2020) 101514, https://doi.org/10.1016/j.jco.2020.101514.

Table 6.7
Comparison of the classification for spam database in the VVRKBS with the ℓ^1 norm and in the VVRKHS for the exponential kernel with $r = 500$.

Kernels	Models	Training accuracy	Sparsity	Testing accuracy
Exponential	VVRKBS	96.08%	1659	93.95%
Kernel	VVRKHS	99.09%	6130	94.34%

Table 6.8
Test data confusion matrix for the spam database in the VVRKBS with the ℓ^1 norm and in the VVRKHS. The overall test error rates of these two models are 6.05% and 5.66%, respectively.

	VVRKBS		VVRKHS	
True class	Predicted class		Predicted class	
	non-spam	spam	non-spam	spam
Non-spam	59.57%	1.56%	59.64%	1.50%
Spam	4.49%	34.38%	4.16%	34.70%

To sum up, numerical experiments for both synthetic data and real-world benchmark data have shown us the advantages of multi-task learning in the vector-valued RKBSs with the ℓ^1 norm.

References

[1] M.A. Alvarez, L. Rosasco, N.D. Lawrence, et al., Kernels for vector-valued functions: A review, Found. Trends Mach. Learn. 4 (3) (2012) 195–266.
[2] N. Aronszajn, Theory of reproducing kernels, Trans. Amer. Math. Soc. 68 (1950) 337–404.
[3] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, Found. Trends Mach. Learn. 3 (1) (2011) 1–122.
[4] E.J. Candès, J. Romberg, T. Tao, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information, IEEE Trans. Inform. Theory 52 (2) (2006) 489–509.
[5] A. Caponnetto, C.A. Micchelli, M. Pontil, Y. Ying, Universal multi-task kernels, J. Mach. Learn. Res. 9 (2008) 1615–1646.
[6] C. Carmeli, E. de Vito, A. Toigo, V. Umanità, Vector valued reproducing kernel Hilbert spaces and universality, Anal. Appl. (Singap.) 8 (1) (2010) 19–61.
[7] R. Caruana, Multitask learning, Mach. Learn. 28 (1) (1997) 41–75.
[8] J.G. Christensen, Sampling in reproducing kernel Banach spaces on Lie groups, J. Approx. Theory 164 (1) (2012) 179–203.
[9] F. Cucker, D.-X. Zhou, Learning Theory: an Approximation Theory Viewpoint, in: Cambridge Monographs on Applied and Computational Mathematics, vol. 24, Cambridge University Press, Cambridge, 2007, p. xii+224, With a foreword by Stephen Smale.
[10] S. De Marchi, R. Schaback, Stability of kernel-based interpolation, Adv. Comput. Math. 32 (2) (2010) 155–161.
[11] E. De Vito, V. Umanità, S. Villa, An extension of Mercer theorem to matrix-valued measurable kernels, Appl. Comput. Harmon. Anal. 34 (3) (2013) 339–351.
[12] R. Der, D. Lee, Large-margin classification in Banach spaces, in: M. Meila, X. Shen (Eds.), Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, in: Proceedings of Machine Learning Research, vol. 2, PMLR, San Juan, Puerto Rico, 2007, pp. 91–98.
[13] N. Dinculeanu, Vector Integration and Stochastic Integration in Banach Spaces, in: Pure and Applied Mathematics (New York), Wiley-Interscience, New York, 2000, p. xvi+424.
[14] G.E. Fasshauer, F.J. Hickernell, Q. Ye, Solving support vector machines in reproducing kernel Banach spaces with positive definite functions, Appl. Comput. Harmon. Anal. 38 (1) (2015) 115–139.
[15] K. Fukumizu, G.R. Lanckriet, B.K. Sriperumbudur, Learning in Hilbert vs. Banach spaces: A measure embedding viewpoint, in: J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira, K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems, Vol. 24, Curran Associates, Inc., 2011, pp. 1773–1781.
[16] Z.-C. Guo, L. Shi, Learning with coefficient-based regularization and ℓ^1-penalty, Adv. Comput. Math. 39 (3–4) (2013) 493–510.
[17] D. Han, M.Z. Nashed, Q. Sun, Sampling expansions in reproducing kernel Hilbert and Banach spaces, Numer. Funct. Anal. Optim. 30 (9–10) (2009) 971–987.
[18] T. Hangelbroek, F.J. Narcowich, J.D. Ward, Kernel approximation on manifolds I: bounding the Lebesgue constant, SIAM J. Math. Anal. 42 (4) (2010) 1732–1760.
[19] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, second ed., in: Springer Series in Statistics, Springer, New York, 2009, p. xxii+745.
[20] G. Kimeldorf, G. Wahba, Some results on Tchebycheffian spline functions, J. Math. Anal. Appl. 33 (1971) 82–95.

Please cite this article as: R. Lin, G. Song and H. Zhang, Multi-task Learning in vector-valued reproducing kernel Banach spaces with the ℓ^1 norm, Journal of Complexity (2020) 101514, https://doi.org/10.1016/j.jco.2020.101514.

[21] R. Lin, H. Zhang, J. Zhang, On reproducing kernel Banach spaces: Generic definitions and unified framework of constructions, arXiv:1901.01002.
[22] L. Meziani, On the dual space $C_0^*(S, X)$, Acta Math. Univ. Comenian. (N.S.) 78 (1) (2009) 153–160.
[23] C.A. Micchelli, M. Pontil, On learning vector-valued functions, Neural Comput. 17 (1) (2005) 177–204.
[24] P. Mörters, Y. Peres, Brownian Motion, in: Cambridge Series in Statistical and Probabilistic Mathematics, vol. 30, Cambridge University Press, Cambridge, 2010, p. xii+403, With an appendix by Oded Schramm and Wendelin Werner.
[25] C.E. Rasmussen, C.K.I. Williams, Gaussian Processes for Machine Learning, in: Adaptive Computation and Machine Learning, MIT Press, Cambridge, MA, 2006, p. xviii+248.
[26] B. Schölkopf, R. Herbrich, A.J. Smola, A generalized representer theorem, in: Computational Learning Theory (Amsterdam, 2001), in: Lecture Notes in Comput. Sci., vol. 2111, Springer, Berlin, 2001, pp. 416–426.
[27] B. Schölkopf, A.J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning), The MIT Press, Cambridge, 2001.
[28] L. Shi, Y.-L. Feng, D.-X. Zhou, Concentration estimates for learning with ℓ^1-regularizer and data dependent hypothesis spaces, Appl. Comput. Harmon. Anal. 31 (2) (2011) 286–302.
[29] G. Song, H. Zhang, Reproducing kernel Banach spaces with the ℓ^1 norm II: error analysis for regularized least square regression, Neural Comput. 23 (10) (2011) 2713–2729.
[30] G. Song, H. Zhang, F.J. Hickernell, Reproducing kernel Banach spaces with the ℓ^1 norm, Appl. Comput. Harmon. Anal. 34 (1) (2013) 96–116.
[31] I. Steinwart, A. Christmann, Support Vector Machines, in: Information Science and Statistics, Springer, New York, 2008, p. xvi+601.
[32] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1) (1996) 267–288.
[33] H. Tong, D.-R. Chen, F. Yang, Least square regression with ℓ^p-coefficient regularization, Neural Comput. 22 (2010) 3221–3235.
[34] H. Wendland, Scattered Data Approximation, Cambridge Monographs on Applied and Computational Mathematics, vol. 17, Cambridge University Press, Cambridge, 2005, p. x+336.
[35] Y. Xu, Q. Ye, Generalized Mercer kernels and reproducing kernel Banach spaces, Mem. Am. Math. Soc. 258 (1243) (2019) 1–122.
[36] Q. Ye, Support vector machines in reproducing kernel Hilbert spaces versus Banach spaces, in: Approximation Theory XIV: San Antonio 2013, in: Springer Proc. Math. Stat., vol. 83, Springer, Cham, 2014, pp. 377–395.
[37] H. Zhang, Y. Xu, J. Zhang, Reproducing kernel Banach spaces for machine learning, J. Mach. Learn. Res. 10 (2009) 2741–2775.
[38] H. Zhang, J. Zhang, Frames, Riesz bases, and sampling expansions in Banach spaces via semi-inner products, Appl. Comput. Harmon. Anal. 31 (1) (2011) 1–25.
[39] H. Zhang, J. Zhang, Regularized learning in Banach spaces as an optimization problem: representer theorems, J. Global Optim. 54 (2) (2012) 235–250.
[40] H. Zhang, J. Zhang, Vector-valued reproducing kernel Banach spaces with applications to multi-task learning, J. Complexity 29 (2) (2013) 195–215.

Please cite this article as: R. Lin, G. Song and H. Zhang, Multi-task Learning in vector-valued reproducing kernel Banach spaces with the ℓ^1 norm, Journal of Complexity (2020) 101514, https://doi.org/10.1016/j.jco.2020.101514.