Journal of Machine Learning Research (XXXX) Submitted XX; Published XXX

Universal Multi-Task Kernels

Andrea Caponnetto CAPONNET@CITYU.EDU.HK
Department of Mathematics
City University of Hong Kong
83 Tat Chee Avenue, Kowloon Tong, Hong Kong

Charles A. Micchelli CAM@MATH.ALBANY.EDU
Department of Mathematics and Statistics
State University of New York
The University at Albany
Albany, New York 12222, USA

Massimiliano Pontil M.PONTIL@CS.UCL.AC.UK
Department of Computer Science
University College London
Gower Street, London, WC1E 6BT, UK

Yiming Ying ENXYY@BRIS.AC.UK
Department of Engineering Mathematics
University of Bristol
Queen's Building, Bristol, BS8 1TR, UK

Editor:

Abstract

In this paper we are concerned with reproducing kernel Hilbert spaces $\mathcal{H}_K$ of functions from an input space into a Hilbert space $\mathcal{Y}$, an environment appropriate for multi-task learning. The reproducing kernel K associated to $\mathcal{H}_K$ has its values as operators on $\mathcal{Y}$. Our primary goal here is to derive conditions which ensure that the kernel K is universal. This means that on every compact subset of the input space, every continuous function with values in $\mathcal{Y}$ can be uniformly approximated by sections of the kernel. We provide various characterizations of universal kernels and highlight them with several concrete examples of some practical importance. Our analysis uses basic principles of functional analysis and especially the useful notion of vector measures which we describe in sufficient detail to clarify our results.

Keywords. Multi-task learning, multi-task kernels, universal approximation, vector-valued reproducing kernel Hilbert spaces.

©XXXX Andrea Caponnetto, Charles A. Micchelli, Massimiliano Pontil, and Yiming Ying.

1. Introduction

The problem of studying representations and methods for learning vector-valued functions has received increasing attention in Machine Learning in the recent years. This problem is motivated by several applications in which it is required to estimate a vector-valued function from a set of input/output data. For example, one is frequently confronted with situations in which multiple supervised learning tasks must be learned simultaneously. This problem can be framed as that of learning a vector-valued function $f = (f_1, f_2, \ldots, f_n)$, where each of its components is a real-valued function and corresponds to a particular task. Often, these tasks are dependent on each other in that they share some common underlying structure. By making use of this structure, each task is easier to learn. Empirical studies indicate that one can benefit significantly by learning the tasks simultaneously as opposed to learning them one by one in isolation (see e.g Evgeniou et al., 2005, and references therein).

In this paper, we build upon the recent work of Micchelli et al. (2006) by addressing the issue of universality of multi-task kernels. Multi-task kernels were recently discussed in Machine Learning context by Micchelli and Pontil (2005), however there is an extensive literature on multi-task kernels as there are important both in theory and practice (see Amodei, 1997; Burbea and Masani, 1984; Caponnetto and De Vito, 2006; Carmeli et al., 2006; Devinatz, 1960; Lowitzsh, 2005; Reisert and Burkhardt, 2007; Vazquez and Walter, 2003, and references therein for more information)

A multi-task kernel K is the reproducing kernel of a Hilbert space of functions from an input space $\mathcal{X}$ which takes values in a Hilbert space $\mathcal{Y}$. For example, in the discussion above, $\mathcal{Y} = \mathbb{R}^n$. Generally, the kernel K is defined on $\mathcal{X} \times \mathcal{X}$ and takes values as an *operator* from $\mathcal{Y}$ to itself [1]. When $\mathcal{Y}$ is n-dimensional, the kernel K takes values in the set of $n \times n$ matrix. The theory of reproducing kernel Hilbert spaces (RKHS) as described in (Aronszajn, 1950) for scalar-valued functions has extensions to any vector-valued $\mathcal{Y}$. Specifically, the RKHS is formed by taking the closure of the linear span of *kernel sections* $\{K(\cdot, x)y,\ x \in \mathcal{X},\ y \in \mathcal{Y}\}$, relative to the RKHS norm. We emphasize here that this fact is fundamentally tied to a norm induced by K and is generally non-constructive. Here, we are concerned with conditions on the kernel K which ensure that all continuous functions from $\mathcal{X}$ to $\mathcal{Y}$ can be uniformly approximated on any compact subset of $\mathcal{X}$ by the linear span of kernel sections.

As far as we are aware, the first paper which addresses this question in Machine Learning literature is (Steinwart, 2001). Steinwart uses the expression *universal kernel* and we follow that terminology here. The problem of identifying universal kernels was also discussed by Poggio et al. (2002). One of us was introduced to this problem in a lecture given at City University of Hong Kong by Ding-Xuan Zhou several years ago (Zhou, 2003). Subsequently, some aspects of this problem were treated in (Micchelli et al., 2003; Micchelli and Pontil, 2004) and then in detail in (Micchelli et al., 2006).

The question of identifying universal kernels has a practical basis. We wish to learn a *continuous* target function $f : \mathcal{X} \to \mathcal{Y}$ from a finite number of samples. The learning algorithm used for this purpose should be consistent. That is, as the samples size increases,

1. Sometimes, such a kernel is called operator-valued or matrix-valued kernel if $\mathcal{Y}$ is infinite of finite dimensional, respectively. However, for simplicity sake we adopt the terminology multi-task kernel throughout the paper.

the discrepancy between the target function and the function learned from the data should tend to zero. Kernel-based algorithms (Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004) generally use the representer theorem and learn a function in the linear span of kernel sections. Therefore, here we interpret consistency to mean that, for *any compact* subset $\mathcal{Z}$ of the input space $\mathcal{X}$ and every continuous target function $f : \mathcal{X} \to \mathcal{Y}$, the discrepancy between the target function and the learned function goes to zero uniformly on $\mathcal{Z}$ as the sample size goes to infinity. It is important to keep in mind that our input space is *not* assumed to be compact. However, we do assume that it is a Hausdorff topological space so that there is an abundance of compact subsets, for example any finite subset of the input space is compact.

Consistency in the sense we described above is important in order to study the statistical performance of learning algorithms based on RKHS. For example, Chen et al. (2004) and Steinwart et al. (2006) studied statistical analysis of soft margin SVM algorithms, Caponnetto and De Vito (2006) gave a detailed analysis of the regularized least-squares algorithm over vector-valued RKHS and proved universal consistency of this algorithm assuming that the kernel is universal and fulfills the additional condition that the operators $K(x, x)$ have finite trace. The results in these papers imply universal consistency of learning algorithms. One more interesting application of universal kernels is described in (Gretton et al., 2006).

This paper is organized as follows. In Section 2, we review the basic definition and properties of multi-task kernels, define the notion of universal kernel and describe some examples. In Section 3, we introduce the notion of feature map associated to a multi-task kernel and show its relevance to the question of universality. The main result in this section is Theorem 4, which establishes that the closure of the RKHS in the space of continuous functions is the same as the closure of the space generated by the feature map. The importance of this result is that universality of a kernel can be established directly by considering its features. In Section 4 we provide an alternate proof of Theorem 4 which uses the notion of vector measures and discuss ancillary results useful for several concrete examples of some practical importance highlighted in Section 5.

2. RKHS of Vector-Valued Functions

In this section, we review the theory of reproducing kernels for Hilbert spaces of vector-valued functions as in (Micchelli and Pontil, 2005) and introduce the notion of universal kernels.

We begin by introducing some notation. We let $\mathcal{Y}$ be a Hilbert space with inner product $(\cdot, \cdot)_{\mathcal{Y}}$ (we drop the subscript $\mathcal{Y}$ when confusion does not arise). The vector-valued functions will take values on $\mathcal{Y}$. We denote by $\mathcal{L}(\mathcal{Y})$ the space of all bounded linear operators from $\mathcal{Y}$ into itself, with the operator norm $\|A\| := \sup_{\|y\|=1} \|Ay\|$, $A \in \mathcal{L}(\mathcal{Y})$ and by $\mathcal{L}_+(\mathcal{Y})$ the set of all bounded, positive semi-definite linear operators, that is, $A \in \mathcal{L}_+(\mathcal{Y})$ provided that, for any $y \in \mathcal{Y}$, $(y, Ay) \geq 0$. We also denote, for any $A \in \mathcal{L}(\mathcal{Y})$, by A^* its adjoint. Finally, for every $m \in \mathbb{N}$, we define $\mathbb{N}_m = \{1, \ldots, m\}$. Table 1 summarizes the notation used in paper.

Definition 1 *We say that a function $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$ is a multi-task kernel on $\mathcal{X}$ if $K(x, t)^* = K(t, x)$ for any $x, t \in \mathcal{X}$, and it is* positive semi-definite*, that is, for any $m \in \mathbb{N}$,*

name	notation	information
input space	$\mathcal{X}$	a Hausdorff topological space
	$\mathcal{Z}$	compact subset of $\mathcal{X}$
	x, t, z	elements of $\mathcal{Z}$
	$\mathscr{B}(\mathcal{Z})$	Borel σ-algebra of $\mathcal{Z}$
	ν	signed scalar measure
	μ	vector measure, see Def. 8
	p, q	indices running from 1 to n
	i, j	indices running from 1 to m
output space	$\mathcal{Y}$	Hilbert space, with inner product $\langle \cdot, \cdot \rangle_{\mathcal{Y}}$
	$\mathcal{B}_1$	unit ball centered at the origin, in $\mathcal{Y}$
feature space	$\mathcal{W}$	Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{W}}$
	$\mathcal{L}(\mathcal{Y}, \mathcal{W})$	all bounded linear operators from $\mathcal{Y}$ into $\mathcal{W}$
	$\mathcal{L}(\mathcal{Y})$	all bounded linear operators from $\mathcal{Y}$ into itself
	A, B	elements of $\mathcal{L}(\mathcal{Y})$
	$\mathcal{L}_+(\mathcal{Y}) \subseteq \mathcal{L}(\mathcal{Y})$	subset of positive linear operators
multi-task kernel	K	a function from $\mathcal{X} \times \mathcal{X}$ to $\mathcal{L}(\mathcal{Y})$, see Def. 1
	$\mathcal{H}_K$	reproducing kernel Hilbert space of K
feature representation	Φ	mapping from $\mathcal{X}$ to $\mathcal{L}(\mathcal{Y}, \mathcal{W})$
	$\mathcal{C}(\mathcal{Z}, \mathcal{Y})$	space of continuous $\mathcal{Y}$-valued functions on $\mathcal{Z}$
	ι	isometric mapping from $\mathcal{C}(\mathcal{Z}, \mathcal{Y})$ to $\mathcal{C}(\mathcal{Z} \times \mathcal{B}_1)$
	$\mathcal{C}_K(\mathcal{Z}, \mathcal{Y})$	subset of $\mathcal{C}(\mathcal{Z}, \mathcal{Y})$ generated by K, see Eq. (2)
	$\mathcal{C}_\Phi(\mathcal{Z}, \mathcal{Y})$	subset of $\mathcal{C}(\mathcal{Z}, \mathcal{Y})$ generated by Φ, see Eq. (12)

Table 1: Notation.

$\{x_j : j \in \mathbb{N}_m\} \subseteq \mathcal{X}$ *and* $\{y_j : j \in \mathbb{N}_m\} \subseteq \mathcal{Y}$ *there holds*

$$\sum_{i,j \in \mathbb{N}_m} (y_i, K(x_i, x_j) y_j) \geq 0. \tag{1}$$

For any $t \in \mathcal{X}$ and $y \in \mathcal{Y}$, we introduce the mapping $K_t y : \mathcal{X} \to \mathcal{Y}$ defined, for every $x \in \mathcal{X}$ by $(K_t y)(x) := K(x, t)y$. In the spirit of Moore-Aronszjain's theorem, there is a one-to-one correspondence between the kernel K with property (1) and an RKHS of functions $f : \mathcal{X} \to \mathcal{Y}$ (Aronszajn, 1950), see also (Micchelli and Pontil, 2005; Carmeli et al., 2006).

Throughout this paper, we assume that the kernel K is *continuous* relative to the operator norm on $\mathcal{L}(\mathcal{Y})$. We now return to the formulation of the definition of universal kernel. For this purpose, we recall that $\mathcal{C}(\mathcal{Z}, \mathcal{Y})$ is the Banach space of continuous $\mathcal{Y}$-valued continuous function on a compact subset $\mathcal{Z}$ of $\mathcal{X}$ with the *maximum norm*, defined by $\|f\|_{\infty,\mathcal{Z}} := \sup_{x \in \mathcal{Z}} \|f(x)\|_{\mathcal{Y}}$. We also define, for every multi-task kernel K, the subspace of $\mathcal{C}(\mathcal{Z}, \mathcal{Y})$

$$\mathcal{C}_K(\mathcal{Z}, \mathcal{Y}) := \overline{\text{span}}\{K_x y : x \in \mathcal{Z}, y \in \mathcal{Y}\}, \tag{2}$$

where the closure is relative to the norm in the space $\mathcal{C}(\mathcal{Z}, \mathcal{Y})$.

Definition 2 *We say that a multi-task kernel K is a* universal kernel *if, for any compact subset $\mathcal{Z}$ of $\mathcal{X}$, $\mathcal{C}_K(\mathcal{Z}, \mathcal{Y}) = \mathcal{C}(\mathcal{Z}, \mathcal{Y})$.*

In the special case that $\mathcal{Y} = \mathbb{R}^n$, the kernel function K takes values as $n \times n$ matrices. The corresponding matrix elements can be identified by the formula

$$(K(x,t))_{pq} = \langle K_x e_p, K_t e_q \rangle_K, \quad \forall x, t \in \mathcal{X},$$

where e_p, e_q are the standard coordinate basis in $\mathbb{R}^n$, for $p, q \in \mathbb{N}_n$.

In order to describe some of the examples of multi-task kernels below, it is useful to first present the following generalization of Schur product of scalar kernels to matrix-valued kernels. For this purpose, for any $i \in \mathbb{N}_m$ we let $y_i = (y_{1i}, y_{2i}, \ldots, y_{ni}) \in \mathbb{R}^n$, so that equation (1) is equivalent to

$$\sum_{i,j \in \mathbb{N}_m} \sum_{p,q \in \mathbb{N}_n} y_{pi} (K(x_i, x_j))_{pq} y_{qj} \geq 0. \tag{3}$$

From the above observation, we conclude that K is a kernel if and only if $\big((K(x_i, x_j)_{p,q}\big)$ as the matrix with row index $(p, i) \in \mathbb{N}_n \times \mathbb{N}_m$ and column index $(q, j) \in \mathbb{N}_n \times \mathbb{N}_m$ is positive semi-definite.

Proposition 3 *Let G and K be $n \times n$ multi-task kernels. Then, the element-wise product kernel $K \circ G : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^n \times \mathbb{R}^n$ defined, for any $x, t \in \mathcal{X}$ and $p, q \in \mathbb{N}_n$, by $\big(K \circ G(x,t)\big)_{pq} := \big(K(x,t)\big)_{pq} \big(G(x,t)\big)_{pq}$ is an $n \times n$ multi-task kernel.*

Proof We have to check the positive semi-definiteness of $K \circ G$. To see this, for any $m \in \mathbb{N}$, $\{y_i \in \mathbb{R}^n : i \in \mathbb{N}_m\}$ and $\{x_i \in \mathcal{X} : i \in \mathbb{N}_m\}$ we observe that

$$\sum_{i,j \in \mathbb{N}_m} (y_i, K \circ G(x_i, x_j) y_j) = \sum_{p,i} \sum_{q,j} y_{pi} y_{qj} \big(K(x_i, x_j)\big)_{pq} \big(G(x_i, x_j)\big)_{pq}. \tag{4}$$

By inequality (3), it follows that the matrix $\big(\big(K(x_i, x_j)\big)_{pq}\big)$ is positive semi-definite as the matrix with (p, i) and (q, j) as row and column indices respectively, and so is $\big(\big(G(x_i, x_j)\big)_{pq}\big)$. Applying the Schur Lemma (Aronszajn, 1950) to these matrices implies that equation (4) is nonnegative, and hence proves the assertion. ■

We now present some examples of multi-task kernels. They will be used in Section 5 to illustrate the general results in Sections 3 and 4.

The first example is adapted from (Micchelli and Pontil, 2005).

Example 1 *If, for every $j \in \mathbb{N}_m$ the function $G_j : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a scalar kernel and $B_j \in \mathcal{L}_+(\mathcal{Y})$, then the function*

$$K(x,t) = \sum_{j \in \mathbb{N}_m} G_j(x,t) B_j, \quad \forall x, t \in \mathcal{X} \tag{5}$$

is a multi-task kernel.

The operators B_j model smoothness across the components of the vector-valued function. For example, in the context of multi-task learning (see e.g. Evgeniou et al., 2005, and references therein), we set $\mathcal{Y} = \mathbb{R}^n$, hence B_j are $n \times n$ matrices. These matrices model the relationships across the tasks. Evgeniou et al. (2005) considered kernels of the form (5) with $m = 2$, B_1 a multiple of the identity matrix and B_2 a low rank matrix. A specific case for $\mathcal{X} = \mathbb{R}^d$ is

$$(K(x,t))_{p,q} = \lambda x \cdot t + (1-\lambda)\delta_{pq}(x \cdot t)^2, \quad p,q \in \mathbb{N}_n,$$

where $x \cdot t$ is the standard inner product in $\mathbb{R}^d$ and $\lambda \in [0,1]$. This kernel has an interesting interpretation. Using only the first term on the right hand side of the above equation ($\lambda = 1$) corresponds to learning all tasks as the same task, that is, all components of the vector-valued function $f = (f_1, \ldots, f_n)$ are the same function, which will be a linear function since the kernel G_1 is linear. Whereas, using only the second term ($\lambda = 0$) corresponds to learning independent tasks, that is, the components of the function f will be generally different functions. These functions will be quadratic since G_2 is a quadratic polynomial kernel. Thus, the above kernel combines two heterogeneous kernels to form a more flexible one. By choosing the parameter λ appropriately, the learning model can be tailored to the data at hand.

We note that if K is a diagonal matrix-valued kernel, then each component of a vector-valued function in the associated RKHS of K can be represented, independently of the other components, as a function in the RKHS of a scalar kernel. However, in general, a multi-task kernel will not be diagonal and, more importantly, *will not* be reduced to a diagonal one by linearly transforming the output space. For example, the kernel in equation (5) cannot be reduced to a diagonal kernel, unless all the matrices $B_j, j \in \mathbb{N}_m$ can all be simultaneously transformed into a diagonal matrix. Therefore, in general, the component functions share some underlying structure which is reflected by the choice of the kernel and cannot be treated as independent objects. This fact is further illustrated by the next example.

Example 2 *If $\mathcal{X}_0$ is a compact Hausdorff space, for $p \in \mathbb{N}_n$, T_p is a map from $\mathcal{X}$ from $\mathcal{X}_0$ (not necessary linear) and $G : \mathcal{X}_0 \times \mathcal{X}_0 \to \mathbb{R}$ is a scalar kernel, then*

$$K(x,t) := \Big(G(T_p x, T_q t)\Big)_{p,q=1}^n, \quad \forall x, t \in \mathcal{X} \tag{6}$$

is a matrix-valued kernel on $\mathcal{X}$.

A specific instance of the above example is described by Vazquez and Walter (2003) in the context of system identification. It corresponds to the choices that $\mathcal{X}_0 = \mathcal{X} = \mathbb{R}$ and $T_p(x) = x + \tau_p$, where $\tau_p \in \mathbb{R}$. In this case, the kernel K models "delays" between the components of the vector-valued function. Indeed, it is easy to verify that, for this choice, for all $f \in \mathcal{H}_K$ and $p \in \mathbb{N}_n$,

$$f_p(x) := (f(x), e_p) = h(x - \tau_p), \quad \forall x \in \mathcal{X}$$

where h is a scalar-valued function in the RKHS of kernel G.

Other choices of the map T_p are possible and provide interesting extensions of scalar kernels. For instance, the choice $K(x,t) := (e^{\sigma_{pq}\langle x,t\rangle} : p,q \in \mathbb{N}_n)$, where $\sigma = (\sigma_{pq})$ is a positive

semi-definite matrix suggested by Example 2. Specifically, the eigenvalue decomposition of the matrix σ is given by $\sigma = \sum_{i=1}^n \lambda_i u_i u_i^T$ and, for any $x \in \mathcal{X}$ and $i \in \mathbb{N}_n$ the map $T_p^{(i)}$ is given by $T_p^{(i)} x := \sqrt{\lambda_i} u_{ip} x$. Therefore, we obtain that $K(x,t) = (\prod_{i=1}^n e^{\langle T_p^{(i)} x, T_q^{(i)} t\rangle} : p,q \in \mathbb{N}_n)$ and, so, by Proposition 3, we conclude that K is a matrix-valued kernel.

It is interesting to note, in passing, that, although one would expect the function

$$K(x,t) := \left(e^{-\sigma_{pq}\|x-t\|^2}\right)_{p,q=1}^n, \quad \forall x,t \in \mathcal{X} \tag{7}$$

to be a kernel over $\mathcal{X} = \mathbb{R}^d$, we will show later in Section 5 that this is not true, unless all entries of the matrix σ are the same.

Our next example called Hessian of Gaussian is motivated by the problem of learning gradients (Solak et al., 2002; Mukherjee and Zhou, 2006). In many applications, one wants to learn an unknown real-valued function $f(x)$, $x = (x^1, \ldots, x^d) \in \mathbb{R}^d$ and its gradient function $\nabla f = (\partial_1 f, \ldots, \partial_d f)$ where, for any $j \in \mathbb{N}_d$, $\partial_p f$ denotes the p-th partial derivative of f. Here the outputs y_{ip} denotes the observation of derivative of p-th derivative at sample x_i. Therefore, this problem is an appealing example of multi-task learning: learn the target function and its gradient function jointly.

To see why this problem related with Hessian of Gaussian, we adopt Gaussian process (Rasmussen and Williams, 2006) viewpoint of kernel methods. In this perspective, kernels are interpreted as covariance of Gaussian prior probability distributions over functions. More specifically, the target (unknown) function f is usually assumed as the realizations of random variables indexed by its input vectors in a zero-mean Gaussian process. The Gaussian process can be fully specified by giving the covariance matrix for any finite set of zero-mean random variables $\{f(x_i) : i \in \mathbb{N}_m\}$. The covariance between the functions corresponding to the inputs x_i and x_j can be defined by Mercer kernel functions, for example, the Gaussian kernel $G(x) = \exp(-\frac{\|x\|^2}{\sigma})$ with $\sigma > 0$, that is,

$$\operatorname{cov}(f(x_i), f(x_j)) = G(x_i - x_j).$$

Consequently, the covariance between $\partial_p f$ and $\partial_q f$ is given by

$$\operatorname{cov}(\partial_p f(x_i), \partial_q f(x_j)) = \partial_p \partial_q \operatorname{cov}(f(x_i), f(x_j)) = -\partial_p \partial_q G(x_i - x_j).$$

This suggests to us to use the Hessian of Gaussian to model the correlation of gradient function ∇f as we present in the following example.

Example 3 *We let $\mathcal{Y} = \mathcal{X} = \mathbb{R}^n$, and, for any $x = (x_p : p \in \mathbb{N}_n) \in \mathcal{X}$, $G(x) = \exp(-\frac{\|x\|^2}{\sigma})$ with $\sigma > 0$. Then, the Hessian matrix of G given by*

$$K(x,t) := (-(\partial_p \partial_q G)(x-t) : p,q \in \mathbb{N}_n) \quad \forall x,t \in \mathcal{X} \tag{8}$$

is a matrix-valued kernel.

To illustrate our final example we let $L^2(\mathbb{R})$ be the Hilbert space of square integrable functions on $\mathbb{R}$ with the norm $\|h\|_{L^2}^2 := \int_{\mathbb{R}} h^2(x)dx$. Moreover, we denote by $W^1(\mathbb{R})$ the Sobolev space of order one, which is defined as the space of real-valued functions h on $\mathbb{R}$ whose norm

$$\|h\|_{W^1} := \left(\|h\|_{L^2}^2 + \|h'\|_{L^2}^2\right)^{\frac{1}{2}}$$

is finite

Example 4 *Let $\mathcal{Y} = L^2(\mathbb{R})$, $\mathcal{X} = \mathbb{R}$ and consider the linear space of functions from $\mathbb{R}$ to $\mathcal{Y}$ which have finite norm*

$$\|f\|^2 = \int_{\mathbb{R}} \left(\|f(x,\cdot)\|_{W^1}^2 + \left\| \frac{\partial f(x,\cdot)}{\partial x} \right\|_{W^1}^2 \right) dx.$$

Then this is an RKHS with multi-task kernel given, for every $x, t \in \mathcal{X}$, by

$$(K(x,t)y)(r) = e^{-\pi|x-t|} \int_{\mathbb{R}} e^{-\pi|r-s|} y(s) ds, \quad \forall y \in \mathcal{Y},\ r \in \mathbb{R}.$$

This example may be appropriate to learn the heat distribution in a medium if we think of x as time. Another potential application extends the discussion following Example 1. Specifically, we consider the case that the input x represents both time and a task (e.g. the profile identifying a customer) and the output is the regression function associated to that task (e.g. the preference function of a customer, see (Evgeniou et al., 2005) for more information). So, this example may be amenable for learning the dynamics of the tasks.

Further examples for the case that $\mathcal{Y} = L^2(\mathbb{R}^d)$ will be provided in Section 5. We also postpone to that section the proof of the claims in Examples 1-4 as well as the discussion about the universality of the kernels therein.

We end this section with some remarks. It is well known that universality of kernels is a main hypothesis in the proof of the consistency of kernel-based learning algorithms. Universal consistency of learning algorithms and their error analysis also rely on the capacity of the RKHS. In particular, following the exact procedure for the scalar case in (Cucker and Smale, 2001), one sufficient condition for universal consistency of vector-valued (multi-task) learning algorithms is the compactness of the unit ball of vector-valued RKHS relative to the space of continuous vector-valued functions. Another alternate sufficient condition was proved in (Caponnetto and De Vito, 2006) for the regularized least-squares algorithm over vector-valued RKHS. There, it was assumed that, in addition to the universality of the kernel, the trace of the operators $K(x,x)$ is finite, for every $x \in \mathcal{X}$. Clearly, both conditions are fulfilled by the multi-task kernels presented above if the output space $\mathcal{Y}$ is finite dimensional, but they become non trivial in the infinite dimensional case. However, it is not clear to the authors whether either of these two conditions is necessary for universal consistency. We hope to come back to this problem in the future.

3. Universal Kernels by Features

In this section, we prove that a multi-task kernel is universal if and only if *its feature representation is universal*. To explain what we have in mind, we require some additional notation. We let $\mathcal{W}$ be a Hilbert space and $\mathcal{L}(\mathcal{Y}, \mathcal{W})$ be the set of all bounded linear operators from $\mathcal{Y}$ to $\mathcal{W}$. A *feature representation* associated with a multi-task kernel K is a continuous function

$$\Phi : \mathcal{X} \to \mathcal{L}(\mathcal{Y}, \mathcal{W})$$

such that, for every $x, t \in \mathcal{X}$

$$K(x,t) = \Phi^*(x)\Phi(t), \tag{9}$$

where, we recall, for each $x \in \mathcal{X}$, $\Phi^*(x)$ is the adjoint of $\Phi(x)$ and, therefore, it is in $\mathcal{L}(\mathcal{W}, \mathcal{Y})$. Hence, from now on we call $\mathcal{W}$ the *feature space*. In the case that $\mathcal{Y} = \mathbb{R}$, the condition that $\Phi(x) \in \mathcal{L}(\mathcal{Y}, \mathcal{W})$ can be merely viewed as saying that $\Phi(x)$ in an element of $\mathcal{W}$. Therefore, at least in this case we can rewrite equation (9) as

$$K(x,t) = (\Phi(x), \Phi(t))_{\mathcal{W}}. \tag{10}$$

Another example of practical importance corresponds to the choice $\mathcal{W} = \mathbb{R}^k$ and $\mathcal{Y} = \mathbb{R}^n$, both finite dimensional Euclidean spaces. Here we can identify $\Phi(x)$ relative to standard basis of $\mathcal{W}$ and $\mathcal{Y}$ with the $k \times n$ matrix $\Phi(x) = (\Phi_{rp}(x) : r \in \mathbb{N}_k, p \in \mathbb{N}_n)$, where Φ_{rp} are scalar-valued continuous functions on $\mathcal{X}$. Therefore, according to (9) the matrix representation of the multi-task kernel K is, for each $x, t \in \mathcal{X}$,

$$(K(x,t))_{pq} = \sum_{r \in \mathbb{N}_k} \Phi_{rp}(x)\Phi_{rq}(t), \quad p, q \in \mathbb{N}_n. \tag{11}$$

Returning to the general case, we emphasize that we *assume* that the kernel K has the representation in equation (9), although if it corresponds to a compact integral operator, such a representation will follow from the spectral theorem and Mercer Theorem (Micchelli et al., 2006, see e.g.).

Associated with a feature representation as described above is the following closed linear subspace of $\mathcal{C}(\mathcal{Z}, \mathcal{Y})$

$$\mathcal{C}_\Phi(\mathcal{Z}, \mathcal{Y}) := \overline{\left\{\Phi^*(\cdot)w : w \in \mathcal{W}\right\}}, \tag{12}$$

where the closure is taken relative to the norm of $\mathcal{C}(\mathcal{Z}, \mathcal{Y})$. The continuity of the functions $\Phi^*(\cdot)w$ follows from the assumed continuity of $K(\cdot, \cdot)$ by

$$\begin{aligned} \|\Phi^*(x)w - \Phi^*(t)w\|^2 &\leq \|\Phi^*(x) - \Phi^*(t)\|^2 \|w\|_{\mathcal{W}}^2 \\ &= \|(\Phi^*(x) - \Phi^*(t))(\Phi(x) - \Phi(t))\| \|w\|_{\mathcal{W}}^2 \\ &= \|K(x,x) + K(t,t) - K(x,t) - K(t,x)\| \|w\|_{\mathcal{W}}^2. \end{aligned}$$

Our definition of the phrase "the feature representation is universal" means that $\mathcal{C}_\Phi(\mathcal{Z}, \mathcal{Y}) = \mathcal{C}(\mathcal{Z}, \mathcal{Y})$ for every compact subset $\mathcal{Z}$ of the input space $\mathcal{X}$. The theorem below demonstrates, as we mentioned above, that the kernel K is universal if and only if its feature representation is universal.

Theorem 4 *If K is a continuous multi-task kernel with feature representation Φ, then for every compact subset $\mathcal{Z}$ of $\mathcal{X}$, we have that $\mathcal{C}_K(\mathcal{Z}, \mathcal{Y}) = \mathcal{C}_\Phi(\mathcal{Z}, \mathcal{Y})$.*

Proof The theorem follows straightforwardly from Lemmas 5, 6 and Proposition 7, which we present below.

■

As we know, the feature representation of a given kernel is not unique, therefore we conclude by Theorem 4 that if *some* feature representation of a multi-task kernel is universal then *every* feature representation is universal.

We shall give two different proofs of this general theorem. The first one will use a technique highlighted in (Micchelli and Pontil, 2005) and will be given in this section. The

second proof will be given in the next section and uses the notion of *vector measure*. Both approaches adopt the point of view of Micchelli et al. (2006), in which Theorem 4 is proved in the special case that $\mathcal{Y} = \mathbb{R}$.

We now begin to explain in detail our first proof. We denote the unit ball in $\mathcal{Y}$ by $\mathcal{B}_1 := \{y : y \in \mathcal{Y}, \|y\| \leq 1\}$ and let $\mathcal{Z}$ be a prescribed compact subset of $\mathcal{X}$. Recall that $\mathcal{B}_1$ is not compact in the norm topology on $\mathcal{Y}$ *unless* $\mathcal{Y}$ is finite dimensional, but it is compact in the *weak topology* on $\mathcal{Y}$ since $\mathcal{Y}$ is a Hilbert space (see e.g. Yosida, 1980). Remember that a basis for the open neighborhood of the origin in the weak topology is a set of the form $\{y : y \in \mathcal{Y}, |(y, y_i)| \leq 1, i \in \mathbb{N}_m\}$, where $y_1, \ldots, y_m$ are arbitrary vectors in $\mathcal{Y}$. We put on $\mathcal{B}_1$ the weak topology and conclude, by Tychonoff's theorem (see e.g. Folland, 1999, p.136), that the set $\mathcal{Z} \times \mathcal{B}_1$ is also compact in the product topology.

The above observation allows us to associate $\mathcal{Y}$-valued functions defined on $\mathcal{Z}$ to scalar-valued functions defined on $\mathcal{Z} \times \mathcal{B}_1$. Specifically, we introduce the map $\iota : \mathcal{C}(\mathcal{Z}, \mathcal{Y}) \to \mathcal{C}(\mathcal{Z} \times \mathcal{B}_1)$ which maps any function $f \in \mathcal{C}(\mathcal{Z}, \mathcal{Y})$ to the function $\iota(f) \in \mathcal{C}(\mathcal{Z} \times \mathcal{B}_1)$ defined by the action

$$\iota(f) : (x, y) \mapsto (f(x), y)_{\mathcal{Y}}, \qquad \forall (x, y) \in (\mathcal{Z} \times \mathcal{B}_1). \tag{13}$$

Consequently, it follows that the map ι is *isometric*, since

$$\sup_{x \in \mathcal{Z}} \|f(x)\|_{\mathcal{Y}} = \sup_{x \in \mathcal{Z}} \sup_{\|y\| \leq 1} |(f(x), y)_{\mathcal{Y}}| = \sup_{x \in \mathcal{Z}} \sup_{y \in \mathcal{B}_1} |\iota(f)(x, y)|, \tag{14}$$

where the first equality follows by Cauchy-Schwartz inequality. Moreover, we will denote by $\iota(\mathcal{C}(\mathcal{Z}, \mathcal{Y}))$ the image of $\mathcal{C}(\mathcal{Z}, \mathcal{Y})$ under the mapping ι. In particular, this space is a closed linear subspace of $\mathcal{C}(\mathcal{Z} \times \mathcal{B}_1)$ and, hence, a Banach space.

Similarly, to any multi-task kernel K on $\mathcal{Z}$ we associate a scalar kernel J on $\mathcal{Z} \times \mathcal{B}_1$ defined, for every $(x, y), (t, u) \in \mathcal{X} \times \mathcal{B}_1$, as

$$J((x, y), (t, u)) := (K(x, t)u, y). \tag{15}$$

Moreover, we denote by $\mathcal{C}_J(\mathcal{Z} \times \mathcal{B}_1)$ the closure in $\mathcal{C}(\mathcal{Z} \times \mathcal{B}_1)$ of the set of the sections of the kernel, $\{J((x, y), (\cdot, \cdot)) : (x, y) \in \mathcal{Z} \times \mathcal{B}_1\}$. It is important to realize that whenever K is a valid multi-task kernel, then J is a valid scalar kernel.

The lemma below relates the set $\mathcal{C}_K(\mathcal{Z}, \mathcal{Y})$ to the corresponding set $\mathcal{C}_J(\mathcal{Z} \times \mathcal{B}_1)$ for the kernel J on $\mathcal{Z} \times \mathcal{B}_1$.

Lemma 5 *If $\mathcal{Z}$ is a compact subset of $\mathcal{X}$ and K is a continuous multi-task kernel then $\iota(\mathcal{C}_K(\mathcal{Z}, \mathcal{Y})) = \mathcal{C}_J(\mathcal{Z} \times \mathcal{B}_1)$.*

Proof The assertion follows by equation (15) and the continuity of the map ι. ■

In order to prove Theorem 4, we also need to provide a similar lemma for the set $\mathcal{C}_\Phi(\mathcal{Z}, \mathcal{Y})$. Before we state the lemma, we note that knowing the features of the multi-task kernel K leads us to the features for the scalar-kernel J associated to K. Specifically, for every $(x, y), (t, u) \in \mathcal{X} \times \mathcal{B}_1$, we have that

$$J((x, y), (t, u)) = (\Psi(x, y), \Psi(t, u))_{\mathcal{W}}, \tag{16}$$

where the continuous function $\Psi : \mathcal{X} \times \mathcal{B}_1 \to \mathcal{W}$ is defined as

$$\Psi(x, y) = \Phi(x)y, \quad x \in \mathcal{X}, y \in \mathcal{B}_1.$$

Thus, equation (16) parallels equation (10) except that $\mathcal{X}$ is replaced by $\mathcal{X} \times \mathcal{B}_1$. We also denote by $\mathcal{C}_\Psi(\mathcal{Z} \times \mathcal{B}_1) = \overline{\{(\Psi(\cdot), w)_\mathcal{W} : w \in \mathcal{W}\}}$, the closed linear subspace of $\mathcal{C}(\mathcal{Z} \times \mathcal{B}_1)$.

Lemma 6 *If $\mathcal{Z}$ is a compact subset of $\mathcal{X}$ and K is a continuous multi-task kernel with feature representation Φ then $\iota(\mathcal{C}_\Phi(\mathcal{Z}, \mathcal{Y})) = \mathcal{C}_\Psi(\mathcal{Z} \times \mathcal{B}_1)$.*

Proof The proof is immediate. Indeed, for each $x \in \mathcal{X}$, $w \in \mathcal{W}$, $y \in \mathcal{Y}$, we have that $(\Phi^*(x)w, y)_\mathcal{Y} = (w, \Phi(x)y)_\mathcal{W} = (\Psi(x, y), w)_\mathcal{W}$. ■

We display the relationship among the spaces which appear in the previous results in the diagram below.

In order to complete the proof of Theorem 4, we review some facts about signed measures and bounded linear functionals on continuous functions. Let Ω be any prescribed compact Hausdorff space and $\mathcal{C}(\Omega)$ be the space of all real-valued continuous functions with norm $\|\cdot\|_{\infty,\Omega}$. We also use the notation $\mathscr{B}(\Omega)$ to denote the Borel σ-algebra on Ω. Now, we recall the description of the dual space of $\mathcal{C}(\Omega)$. By the Riesz representation theorem, any linear functional L in the dual space of $\mathcal{C}(\Omega)$ is uniquely identified as a regular signed Borel measure ν on Ω (see e.g. Folland (1999)), that is,

$$L(g) = \int_\Omega g(x) d\nu(x), \qquad \forall g \in \mathcal{C}(\Omega).$$

The variation of ν is given, for any $E \in \mathscr{B}(\Omega)$, by

$$|\nu|(E) := \sup\Big\{ \sum_{j \in \mathbb{N}} |\nu(A_j)| : \{A_j : j \in \mathbb{N}\} \text{ pairwise disjoint and } \cup_{j \in \mathbb{N}} A_j = E \Big\}.$$

Moreover, we have that $\|L\| = \|\nu\|$, where $\|\nu\| = |\nu|(\Omega)$ and $\|L\|$ is the operator norm of L defined by $\|L\| = \sup_{\|g\|_{\infty,\Omega}=1} |L(g)|$. Recall that a Borel measure ν is *regular* if, for any $E \in \mathscr{B}(\mathcal{X})$,

$$\nu(E) = \inf\big\{\nu(U) : E \subseteq U, U \text{ open}\big\} = \sup\big\{\nu(\bar{U}) : \bar{U} \subseteq E, \bar{U} \text{ compact}\big\}.$$

In particular, every finite Borel measure on Ω is regular, see Folland (1999, p.217). We denote by $\mathcal{M}(\Omega)$ the space of all regular signed measures on Ω with total variation norm.

Proposition 7 *For any compact set $\mathcal{Z} \subseteq \mathcal{X}$, and any continuous multi-task kernel K with feature representation Φ, we have that $\mathcal{C}_\Psi(\mathcal{Z} \times \mathcal{B}_1) = \mathcal{C}_J(\mathcal{Z} \times \mathcal{B}_1)$.*

Proof For any compact set $\mathcal{Z} \subseteq \mathcal{X}$, recall that $\mathcal{Z} \times \mathcal{B}_1$ is compact if $\mathcal{B}_1$ is endowed with the weak topology of $\mathcal{Y}$. Hence, the result follows by applying Theorem 4 in (Micchelli et al., 2006) to the scalar kernel J on the set $\mathcal{Z} \times \mathcal{B}_1$ with the feature representation given by equation (16). However, for the convenience of the reader we review the steps of the argument used to prove that theorem. The basic idea is the observation that two closed

$$\begin{array}{ccc} \mathcal{C}_\Psi(\mathcal{Z}\times\mathcal{B}_1) & = & \mathcal{C}_J(\mathcal{Z}\times\mathcal{B}_1) \\ {\scriptstyle\iota}\uparrow & & \uparrow{\scriptstyle\iota} \\ \mathcal{C}_K(\mathcal{Z},\mathcal{Y}) & = & \mathcal{C}_\Phi(\mathcal{Z},\mathcal{Y}) \end{array}$$

Table 2: The top equality is Proposition 7, the bottom equality is Theorem 4 and the left and right arrows are Lemma 5 and 6, respectively.

subspaces of a Banach space are equal if and only if whenever a continuous linear functional vanishes on either one of the subspaces, it must also vanish on the other one. This is a consequence of the Hahn-Banach Theorem (see e.g. Lax, 2002). In the case at hand, we know by the Riesz Representation Theorem that all continuous linear functionals L on $\mathcal{C}(\mathcal{Z}\times\mathcal{B}_1)$ are given by a regular signed Borel measure ν, that is for every $F\in\mathcal{C}(\mathcal{Z}\times\mathcal{B}_1)$, we have that

$$L(F)=\int_{\mathcal{Z}\times\mathcal{B}_1} F(x,y)d\nu(x,y).$$

Now, suppose that L vanishes on $\mathcal{C}_J(\mathcal{Z}\times\mathcal{B}_1)$, then we conclude, by (16), that

$$0=\int_{\mathcal{Z}\times\mathcal{B}_1}\int_{\mathcal{Z}\times\mathcal{B}_1}(\Psi(x,y),\Psi(t,u))_{\mathcal{W}}d\nu(x,y)d\nu(t,u).$$

However, we also have that

$$\int_{\mathcal{Z}\times\mathcal{B}_1}\int_{\mathcal{Z}\times\mathcal{B}_1}(\Psi(x,y),\Psi(t,u))_{\mathcal{W}}d\nu(x,y)d\nu(t,u)=\left\|\int_{\mathcal{Z}\times\mathcal{B}_1}\Psi(x,y)d\nu(x,y)\right\|^2_{\mathcal{W}} \tag{17}$$

and, so, we conclude that

$$\int_{\mathcal{Z}\times\mathcal{B}_1}\Psi(x,y)d\nu(x,y)=0. \tag{18}$$

The proof of equation (17) and the interpretation of the $\mathcal{W}$-valued integral appearing in equation (18) is explained in detail in Micchelli et al. (2006). So, we conclude that L vanishes on $\mathcal{C}_\Psi(\mathcal{Z}\times\mathcal{B}_1)$.

Conversely, if L vanishes on $\mathcal{C}_\Psi(\mathcal{Z}\times\mathcal{B}_1)$ then, for any $x\in\mathcal{Z}, y\in\mathcal{B}_1$, we have that

$$\int_{\mathcal{Z}\times\mathcal{B}_1}J((x,y),(t,u))d\nu(t,u)=\left(\Psi(x,y),\int_{\mathcal{Z}\times\mathcal{B}_1}\Psi(t,u)\nu(t,u)\right)=0$$

that is, L vanishes on $\mathcal{C}_J(\mathcal{Z}\times\mathcal{B}_1)$. ■

4. Further Perspectives for Universality

In this section, we provide an alternate proof of Theorem 4 using the notion of vector measure and also highlight the notion of the annihilator of a set, a useful tool for our examples of multi-task kernels in Section 5.

At first glance, the reduction of the question of when a multi-task kernel is universal to the scalar case, as explained in Section 3, seems compelling. However, there are several reasons to explore alternate approaches to this problem. Firstly, from a practical point of view, if we view multi-task learning as a scalar problem we may loose the ability to understand cross task interactions. Secondly, only one tool to resolve a problem may limit the possibility of success. Finally, as we demonstrated in Section 3 universality of multi-task kernels concerns density in the subspace $\mathcal{C}_J(\mathcal{Z} \times \mathcal{B}_1)$, not the full space $\mathcal{C}(\mathcal{Z} \times \mathcal{B}_1)$, an issue heretofore not considered. Therefore, we cannot directly employ the methods of the scalar case as presented in (Micchelli et al., 2003) to multi-task case.

As we shall see in this section, the concept of vector measure allows us to directly confront the set $\mathcal{C}(\mathcal{Z}, \mathcal{Y})$ rather than following a circuitous path to $\mathcal{C}_J(\mathcal{Z} \times \mathcal{B}_1)$. However, the basic principle which we employ is the same, namely, two closed linear subspaces of a Banach space are equal if and only if whenever a bounded linear functional vanishes on one of them, it also vanishes on the other one. Thus, to implement this principle we are led to consider the dual space of $\mathcal{C}(\mathcal{Z}, \mathcal{Y})$. We remark, in passing, that this space also arose also in the context of the feature space perspective for learning the kernel, see (Micchelli and Pontil, 2005a). For a description of the dual space of $\mathcal{C}(\mathcal{Z}, \mathcal{Y})$, we need the notion of vector measures and in this regard rely upon the information about them in (Diestel and Uhl, Jr., 1977).

To introduce our first definition, recall that throughout this paper $\mathcal{X}$ denotes a Hausdorff space, $\mathcal{Z} \subseteq \mathcal{X}$ any compact subset of $\mathcal{X}$ and $\mathscr{B}(\mathcal{Z})$ the Borel σ-algebra of $\mathcal{Z}$.

Definition 8 *A map $\mu : \mathscr{B}(\mathcal{Z}) \to \mathcal{Y}$ is called a Borel vector measure if μ is countably additive, that is, $\mu(\cup_{j=1}^{\infty} E_j) = \sum_{j=1}^{\infty} \mu(E_j)$ in the norm of $\mathcal{Y}$, for all sequences $\{E_j : j \in \mathbb{N}\}$ of pairwise disjoint sets in $\mathscr{B}(\mathcal{Z})$*

It is important to note that the definition of vector measure given in (Diestel and Uhl, Jr., 1977) only requires it to be finitely additive. For our purpose here, we only use countably additive measures and thus do not require the more general setting used in (Diestel and Uhl, Jr., 1977).

For any vector measure μ, the variation of μ is defined, for any $E \in \mathscr{B}(\mathcal{Z})$, by the equation

$$|\mu|(E) := \sup\left\{\sum_{j\in\mathbb{N}} \|\mu(A_j)\| : \{A_j : j \in \mathbb{N}\} \text{ pairwise disjoint and } \cup_{j\in\mathbb{N}} A_j = E\right\}.$$

In our terminology we conclude from (Diestel and Uhl, Jr., 1977, p.3) that μ is a vector measure if and only if the corresponding variation $|\mu|$ is a scalar measure as explained in Section 3. Whenever $|\mu|(\mathcal{Z}) < \infty$, we call μ a vector measure of *bounded variation* on $\mathcal{Z}$. Moreover, we say that a Borel vector measure μ on $\mathcal{Z}$ is *regular* if its variation measure $|\mu|$ is regular as defined in Section 3. We denote by $\mathcal{M}(\mathcal{Z}, \mathcal{Y})$ the Banach space of all vector measures with bounded variation and norm $\|\mu\| := |\mu|(\mathcal{Z})$.

For any scalar measure $\nu \in \mathcal{M}(\mathcal{Z} \times \mathcal{B}_1)$, we define a $\mathcal{Y}$-valued function on $\mathscr{B}(\mathcal{Z})$, by the equation

$$\mu(E) := \int_{E\times\mathcal{B}_1} y d\nu(x, y), \qquad \forall E \in \mathscr{B}(\mathcal{Z}). \tag{19}$$

Let us confirm that μ is indeed a vector measure. For this purpose, choose any sequence of pairwise disjoint subsets $\{E_j : j \in \mathbb{N}\} \subseteq \mathscr{B}(\mathcal{Z})$, and observe that

$$\sum_{j\in\mathbb{N}} \|\mu(E_j)\|_{\mathcal{Y}} \leq \sum_{j\in\mathbb{N}} \Big| \int_{E_j}\int_{\mathcal{B}_1} d\nu(x,y)\Big| \leq |\nu|(\mathcal{Z}\times\mathcal{B}_1),$$

which implies that $|\mu|(\mathcal{Z})$ is finite and, hence, μ is a regular vector measure. This observation suggests that we define, for any $f \in \mathcal{C}(\mathcal{Z},\mathcal{Y})$, the integral of f relative to μ as

$$\int_{\mathcal{Z}} (f(x), d\mu(x)) := \int_{\mathcal{Z}}\int_{\mathcal{B}_1} (f(x), y) d\nu(x,y). \tag{20}$$

Alternatively, by the standard techniques of measure theory, for any vector measure $\mu \in \mathcal{M}(\mathcal{Z},\mathcal{Y})$ the integral $\displaystyle\int_{\mathcal{Z}} (f(x), d\mu(x))$ is well-defined. One of our goals below is to show that given any vector measure μ, there corresponds a scalar measure ν such that equation (20) still holds. Before doing so, let us point out a useful property about the integral appearing in the left hand side of equation (20). Specifically, for any $y \in \mathcal{Y}$, we associate to any $\mu \in \mathcal{M}(\mathcal{Z},\mathcal{Y})$, a scalar measure μ_y defined, for any $E \in \mathscr{B}(\mathcal{Z})$, by the equation $\mu_y(E) := (y, \mu(E))$. Therefore, we conclude, for any $f \in \mathcal{C}(\mathcal{Z})$, that

$$\int_{\mathcal{Z}} (yf(x), d\mu(x)) = \int_{\mathcal{Z}} f(x) d\mu_y(x).$$

To prepare for our description of the dual space of $\mathcal{C}(\mathcal{Z},\mathcal{Y})$, we introduce, for each $f \in \mathcal{C}(\mathcal{Z},\mathcal{Y})$, a linear functional L_μ defined by,

$$L_\mu f := \int_{\mathcal{Z}} (f(x), d\mu(x)). \tag{21}$$

Then, we have the following useful lemmas, see the appendix for their proofs.

Lemma 9 *If $\mu \in \mathcal{M}(\mathcal{Z},\mathcal{Y})$ then $L_\mu \in \mathcal{C}^*(\mathcal{Z},\mathcal{Y})$ and $\|L_\mu\| = \|\mu\|$.*

Lemma 10 *(Dinculeanu-Singer) For any compact set $\mathcal{Z} \subseteq \mathcal{X}$, the map $\mu \mapsto L_\mu$ is an isomorphism from $\mathcal{M}(\mathcal{Z},\mathcal{Y})$ to $C^*(\mathcal{Z},\mathcal{Y})$.*

Lemma 10 is a vector-valued form of the Riesz representation theorem called *Dinculeanu-Singer theorem,* (see e.g. Diestel and Uhl, Jr., 1977, p.182). For completeness, we provide a self-contained proof in the appendix.

It is interesting to remark that, for any $\mu \in \mathcal{M}(\mathcal{Z},\mathcal{Y})$ we have established in the proof of Lemma 10 that there exists a regular scalar measure ν on $\mathcal{Z}\times\mathcal{B}_1$ such that

$$L_\mu f = \int_{\mathcal{Z}\times\mathcal{B}_1} (f(x), y) d\nu(x,y).$$

Since we established the isometry between $\mathcal{C}^*(\mathcal{Z},\mathcal{Y})$ and $\mathcal{M}(\mathcal{Z},\mathcal{Y})$, it follows that, for every regular vector measure there corresponds a scalar measure on $\mathcal{Z}\times\mathcal{B}_1$ for which equation (19) holds true.

In order to provide our alternate proof of Theorem 4, we need to attend to one further issue. Specifically, we need to define the integral $\int_{\mathcal{Z}} K(t,x)(d\mu(x))$ as an element in $\mathcal{Y}$. For this purpose, for any $\mu \in \mathcal{M}(\mathcal{Z},\mathcal{Y})$ and $t \in \mathcal{Z}$ we define a linear functional L_t on $\mathcal{Y}$ at $y \in \mathcal{Y}$ as

$$L_t y := \int_{\mathcal{Z}} (K(x,t)y, d\mu(x)). \tag{22}$$

Since its norm has the property $\|L_t\| \leq \big(\sup_{x\in\mathcal{Z}} \|K(x,t)\|\big)\|\mu\|$, by the Riesz representation lemma, we conclude that there exists a unique element $\bar{y}$ in $\mathcal{Y}$ such that

$$\int_{\mathcal{Z}} (K(x,t)y, d\mu(x)) = (\bar{y}, y).$$

It is this vector $\bar{y}$ which we denote by the integral $\int_{\mathcal{Z}} K(t,x)(d\mu(x))$.

Similarly, we define the integral $\int_{\mathcal{Z}} \Phi(x)(d\mu(x))$ as an element in $\mathcal{W}$. To do this, we note that $\|\Phi(x)\| = \|\Phi^*(x)\|$ and $\|\Phi^*(x)y\|^2 = \langle K(x,x)y, y\rangle$. Hence, we conclude that there exists a constant κ such that, for all $x \in \mathcal{X}$, $\|\Phi(x)\| \leq \|K(x,x)\|^{\frac{1}{2}} \leq \kappa$. Consequently, the linear functional L on $\mathcal{W}$ defined, for any $w \in \mathcal{W}$, by

$$L(w) := \int_{\mathcal{Z}} \Big(\Phi^*(x)w, d\mu(x)\Big)$$

satisfies the inequality $\|L\| \leq \kappa\|\mu\|$. Hence, we conclude that there exists a unique element $\bar{w} \in \mathcal{W}$ such that $L(w) = (\bar{w}, w)$ for any $w \in \mathcal{W}$. Now, we denote $\bar{w}$ by $\int_{\mathcal{Z}} \Phi(x)(d\mu(x))$ which means that

$$\Big(\int_{\mathcal{Z}} \Phi(x)(d\mu(x)), w\Big)_{\mathcal{W}} = \int_{\mathcal{Z}} \Big(\Phi^*(x)w, d\mu(x)\Big). \tag{23}$$

We have now assembled all the necessary properties of vector measures to provide an alternate proof of Theorem 4.

Alternate Proof of Theorem 4. We see from the feature representation (9) that

$$\int_{\mathcal{Z}} K(t,x)(d\mu(x)) = \int_{\mathcal{Z}} \Phi^*(t)\Phi(x)(d\mu(x)) = \Phi^*(t)\Big(\int_{\mathcal{Z}} \Phi(x)(d\mu(x))\Big), \quad \forall t \in \mathcal{Z}.$$

From this equation, we easily see that if $\int_{\mathcal{Z}} \Phi(x)(d\mu(x)) = 0$ then, for every $t \in \mathcal{Z}$, $\int_{\mathcal{Z}} K(t,x)(d\mu(x)) = 0$. On the other hand, applying (23) with the choice $w = \int_{\mathcal{Z}} \Phi(x)(d\mu(x))$ we get

$$\int_{\mathcal{Z}} \Big(\Phi^*(t) \int_{\mathcal{Z}} \Phi(x)(d\mu(x)), d\mu(t)\Big) = \Big\| \int_{\mathcal{Z}} \Phi(x)(d\mu(x))\Big\|_{\mathcal{W}}^2.$$

Therefore, if, for any $t \in \mathcal{Z}$, $\int_{\mathcal{Z}} K(t,x)(d\mu(x)) = 0$ then $\int_{\mathcal{Z}} \Phi(x)(d\mu(x)) = 0$, or equivalently, by equation (23),

$$\int_{\mathcal{Z}} (\Phi^*(x)w, d\mu(x)) = 0, \quad \forall w \in \mathcal{W}.$$

Consequently, a linear functional vanishes on $\mathcal{C}_K(\mathcal{Z},\mathcal{Y})$ if and only if it vanishes on $\mathcal{C}_\Phi(\mathcal{Z},\mathcal{Y})$ and thus, we obtained that $\mathcal{C}_K(\mathcal{Z},\mathcal{Y}) = \mathcal{C}_\Phi(\mathcal{Z},\mathcal{Y})$. ■

We end this section with a review of our approach to the question of universality of multi-task kernels. The principal tool we employ is a notion of functional analysis referred to as the *annihilator* set. Recall the notion of the annihilator of a set $\mathcal{V}$ which is defined by

$$\mathcal{V}^\perp := \big\{\mu \in \mathcal{M}(\mathcal{Z},\mathcal{Y}) \ : \ \int_{\mathcal{Z}} (v(x), d\mu(x)) = 0, \forall v \in \mathcal{V}\big\}.$$

Notice that the annihilator of the closed linear span of $\mathcal{V}$ is the same as that of $\mathcal{V}$. Consequently, by applying the basic principle stated at the beginning of this section , we conclude that the linear span of $\mathcal{V}$ is dense in $\mathcal{C}(\mathcal{Z},\mathcal{Y})$, that is, $\overline{\text{span}}(\mathcal{V}) = \mathcal{C}(\mathcal{Z},\mathcal{Y})$ if and only if the annihilator $\mathcal{V}^\perp = \{0\}$. Hence, applying this observation to the set of kernel sections $K(\mathcal{Z}) := \{K(\cdot,x)y : \ x \in \mathcal{Z}, \ y \in \mathcal{Y}\}$ or to the set of its corresponding feature sections $\Phi(\mathcal{Z}) := \{\Phi^*(\cdot)w : w \in \mathcal{W}\}$, we obtain from Lemma 10 and Theorem 4, the summary of our main result.

Theorem 11 *Let $\mathcal{Z}$ be a compact subset of $\mathcal{X}$, K a continuous multi-task kernel, and Φ its feature representation. Then, the following statements are equivalent.*

1. $\mathcal{C}_K(\mathcal{Z},\mathcal{Y}) = \mathcal{C}(\mathcal{Z},\mathcal{Y})$.
2. $\mathcal{C}_\Phi(\mathcal{Z},\mathcal{Y}) = \mathcal{C}(\mathcal{Z},\mathcal{Y})$.
3. $K(\mathcal{Z})^\perp = \Big\{\mu \in \mathcal{M}(\mathcal{Z},\mathcal{Y}) \ : \ \int_{\mathcal{Z}} K(t,x)(d\mu(x)) = 0, \quad \forall t \in \mathcal{Z}\Big\} = \big\{0\big\}$.
4. $\Phi(\mathcal{Z})^\perp = \Big\{\mu \in \mathcal{M}(\mathcal{Z},\mathcal{Y}) \ : \ \int_{\mathcal{Z}} \Phi(x)(d\mu(x)) = 0\Big\} = \big\{0\big\}$.

5. Universal Kernels

In this section, we prove the universality of some kernels, based on Theorem 11 developed above. Specifically, the examples highlighted in Section 2 will be discussed in detail.

It is well known that universality of the real-valued kernels is a main hypothesis in the proof of consistency of learning algorithms. Universal consistency of the regularized least-squares algorithm over vector-valued RKHS was proved in Caponnetto and De Vito (2006); there, it was assumed that, in addition to universality of the kernel, the trace of the operators $K(x,x)$ is finite. Clearly, this additional hypothesis is fulfilled when the output space $\mathcal{Y}$ is finite dimensional, however it becomes non trivial in the infinite dimensional case. In particular, this extra condition on the kernel holds, for the Example 1 highlighted in Section 2, when the operators B_j are trace class, and does not hold for Example 4. It is not clear to the authors whether the finite trace condition is necessary for consistency, however this question is outside the scope of the present paper.

5.1 Product of Scalar Kernels and Operators

Our first example is produced by coupling a scalar kernel with an operator in $\mathcal{L}_+(\mathcal{Y})$. Given a scalar kernel G on $\mathcal{X}$ and an operator $B \in \mathcal{L}_+(\mathcal{Y})$, we define the function $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$ by

$$K(x,t) = G(x,t)B, \quad \forall x,t \in \mathcal{X}. \tag{24}$$

For any $\{x_j \in \mathcal{X} : j \in \mathbb{N}_m\}$ and $\{y_j \in \mathcal{Y} : j \in \mathbb{N}_m\}$, we know that $(G(x_i,x_j))_{i,j\in\mathbb{N}_m}$ and $((By_i,y_j))_{i,j\in\mathbb{N}_m}$ are positive semi-definite. Applying Schur's lemma implies that the matrix $(G(x_i,x_j)(By_i,y_j))_{i,j\in\mathbb{N}_m}$ is positive semi-definite and hence, K is positive semi-definite. Moreover, $K^*(x,t) = K(x,t) = K(t,x)$ for any $x,t \in \mathcal{X}$. Therefore, we conclude by Definition 1 that K is a multi-task kernel.

Our goal below is to use the feature representation of the scalar kernel G to introduce the corresponding one for kernel K. To this end, we first let $\mathcal{W}$ be a Hilbert space and $\phi : \mathcal{X} \to \mathcal{W}$ a feature map of the scalar kernel G, so that

$$G(x,t) = (\phi(x),\phi(t))_{\mathcal{W}}, \qquad \forall x,t \in \mathcal{X}.$$

Then, we introduce the *tensor vector space* $\mathcal{W} \bigotimes \mathcal{Y}$. Algebraically, this vector space is spanned by elements of the form $w \otimes y$ with $w \in \mathcal{W}$, $y \in \mathcal{Y}$. For any $w_1 \otimes y_1, w_2 \otimes y_2 \in \mathcal{W} \bigotimes \mathcal{Y}$ and $c \in \mathbb{R}$, there holds the multi-linear relation

$$cw \otimes y = w \otimes cy = c(w \otimes y), \qquad (w_1 + w_2) \otimes y = w_1 \otimes y + w_2 \otimes y,$$

and

$$w \otimes (y_1 + y_2) = w \otimes y_1 + w \otimes y_2.$$

We can turn the tensor space into an inner product space by defining, for any $w_1 \otimes y_1, w_2 \otimes y_2 \in \mathcal{W} \bigotimes \mathcal{Y}$,

$$\langle w_1 \otimes y_1, w_2 \otimes y_2 \rangle = (w_1,w_2)_{\mathcal{W}}(y_1,y_2)_{\mathcal{Y}} \tag{25}$$

and extending by linearity. Finally, taking the completion under this inner product, the vector space $\mathcal{W} \bigotimes \mathcal{Y}$ becomes a Hilbert space. Furthermore, if $\mathcal{W}$ and $\mathcal{Y}$ have orthonormal bases $\{w_i : i \in \mathbb{N}\}$ and $\{y_i : i \in \mathbb{N}\}$ respectively, then $\mathcal{W} \bigotimes \mathcal{Y}$ is exactly the Hilbert space spanned by the orthonormal basis $\{w_i \otimes y_j : i,j \in \mathbb{N}\}$ under the inner product defined above. For instance, if $\mathcal{W} = \mathbb{R}^k$ and $\mathcal{Y} = \mathbb{R}^n$, then $\mathcal{W} \bigotimes \mathcal{Y} = \mathbb{R}^{kn}$.

The above tensor product suggests that we define the map $\Phi : \mathcal{X} \to \mathcal{L}(\mathcal{Y}, \mathcal{W} \otimes \mathcal{Y})$ of kernel K by

$$\Phi(x)y := \phi(x) \otimes \sqrt{B}y, \qquad \forall x \in \mathcal{X},\ y \in \mathcal{Y},$$

and it follows that $\Phi^* : \mathcal{X} \to \mathcal{L}(\mathcal{W} \otimes \mathcal{Y}, \mathcal{Y})$ is given by

$$\Phi^*(x)(w \otimes y) := (\phi(x),w)_{\mathcal{W}}\sqrt{B}y, \quad \forall x \in \mathcal{X}, w \in \mathcal{W}, \text{ and } y \in \mathcal{Y}. \tag{26}$$

From the above observation, it is easy to check, for any $x,t \in \mathcal{X}$ and $y,u \in \mathcal{Y}$, that $(K(x,t)y,u) = \langle \Phi(x)y, \Phi(t)u \rangle$. Therefore, we conclude that Φ is a feature map for the multi-task kernel K.

Finally, we say that an operator $B \in \mathcal{L}_+(\mathcal{Y})$ is *positive definite* if (By,y) is positive whenever y is nonzero. We are now ready to present the result on universality of kernel K.

Theorem 12 *Let $G : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a continuous scalar kernel, $B \in \mathcal{L}_+(\mathcal{Y})$ and K be defined by equation (24). Then, K is a multi-task universal kernel if and only if the scalar kernel G is universal and the operator B is positive definite.*

Proof By Theorem 11 and the feature representation (26), we only need to show that $\Phi(\mathcal{Z})^\perp = \{0\}$ if and only if G is universal and the operator B is positive definite.

We begin with the sufficiency. Suppose that there exists a nonzero vector measure μ such that, for any $w \otimes y \in \mathcal{W} \otimes \mathcal{Y}$, there holds

$$\int_{\mathcal{Z}} (\Phi^*(x)(w \otimes y), d\mu(x)) = \int_{\mathcal{Z}} (\phi(x), w)_{\mathcal{W}} (\sqrt{B}y, d\mu(x)) = 0. \tag{27}$$

Here, with a little abuse of notation we interpret, for a fixed $y \in \mathcal{Y}$, $(\sqrt{B}y, d\mu(x))$ as a scalar measure defined, for any $E \in \mathscr{B}(\mathcal{Z})$, by

$$\int_E (\sqrt{B}y, d\mu(x)) = (\sqrt{B}y, \mu(E)).$$

Since $\mu \in \mathcal{M}(\mathcal{Z}, \mathcal{Y})$, $(\sqrt{B}y, d\mu(x))$ is a regular signed scalar measure. Therefore, we see from (27) that $(\sqrt{B}y, d\mu(x)) \in \phi(\mathcal{Z})^\perp$ for any $y \in \mathcal{Y}$. Remember that G is universal if and only if $\phi(\mathcal{Z})^\perp = \{0\}$, and thus we conclude from (27) that $(\sqrt{B}y, d\mu(x)) = 0$ for any $y \in \mathcal{Y}$. It follows that $(\sqrt{B}y, \mu(E)) = 0$ for any $y \in \mathcal{Y}$ and $E \in \mathscr{B}(\mathcal{Z})$. Thus, for any fixed set E taking the choice $y = \sqrt{B}\mu(E)$ implies that $(B\mu(E), \mu(E)) = 0$. Since E is arbitrary, this means $\mu = 0$ and thus finishes the proof for the sufficiency.

To prove the necessity, suppose first that G is not universal and hence, we know that, for some compact subset $\mathcal{Z}$ of $\mathcal{X}$, there exists a nonzero scalar measure $\nu \in \mathcal{M}(\mathcal{Z})$ such that $\nu \in \phi(\mathcal{Z})^\perp$, that is, $\int_{\mathcal{Z}} (\phi(x), w) d\nu(x) = 0$ for any $w \in \mathcal{W}$. This suggests to us to choose, for a nonzero $y_0 \in \mathcal{Y}$, the nonzero vector measure $\mu = y_0\nu$ defined by $\mu(E) := y_0\nu(E)$ for any $E \in \mathscr{B}(\mathcal{Z})$. Hence, the integral in equation (27) equals

$$\int_{\mathcal{Z}} (\Phi^*(x)(w \otimes y), d\mu(x)) = (\sqrt{B}y, y_0) \int_{\mathcal{Z}} (\phi(x), w)_{\mathcal{W}} d\nu(x) = 0.$$

Therefore, we conclude that there exists a nonzero vector measure $\mu \in \Phi(\mathcal{Z})^\perp$, which implies that K is not universal.

If B is not positive definite, namely, there exists a nonzero element $y_1 \in \mathcal{Y}$ such that $(By_1, y_1) = 0$. However, we observe that $\|\sqrt{B}y_1\|^2 = (By_1, y_1)$ which implies that $\sqrt{B}y_1 = 0$. This suggests to us to choose a nonzero vector measure $\mu = y_1\nu$ with some nonzero scalar measure ν. Therefore, we conclude, for any $w \in \mathcal{W}$ and $y \in \mathcal{Y}$, that

$$\begin{aligned} \int_{\mathcal{Z}} (\Phi^*(x)(w \otimes y), d\mu(x)) &= (\sqrt{B}y, y_1) \int_{\mathcal{Z}} (\phi(x), w)_{\mathcal{W}} d\nu(x) \\ &= (y, \sqrt{B}y_1) \int_{\mathcal{Z}} (\phi(x), w)_{\mathcal{W}} d\nu(x) = 0, \end{aligned}$$

which implies that the nonzero vector measure $\mu \in \Phi(\mathcal{Z})^\perp$. This finishes the proof of the theorem. ■

In the special case $\mathcal{Y} = \mathbb{R}^n$, the operator B is an $n \times n$ positive semi-definite matrix. Then, Theorem 12 tells us that the matrix-valued kernel $K(x,t) := G(x,t)B$ is universal if and only if G is universal and the matrix B is of full rank.

We now proceed further and consider kernels produced by a finite combination of scalar kernels and operators. Specifically, we consider, for any $j \in \mathbb{N}_m$, that $G_j : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a scalar kernel and $B_j \in \mathcal{L}_+(\mathcal{Y})$. We are interested in the kernel defined, for any $x, t \in \mathcal{X}$, by

$$K(x,t) := \sum_{j \in \mathbb{N}_m} G_j(x,t) B_j. \tag{28}$$

Suppose also, for each scalar kernel G_j, that there exists a Hilbert feature space $\mathcal{W}_j$ and a feature map $\phi_j : \mathcal{X} \to \mathcal{W}_j$.

To explain the associated feature map of kernel K, we need to define its feature space. For this purpose, let H_j be a Hilbert space with inner products $(\cdot,\cdot)_j$ for $j \in \mathbb{N}_m$ and we introduce the *direct sum Hilbert space* $\oplus_{j \in \mathbb{N}_m} H_j$ as follows. The elements in this space are of the form $(h_1, \ldots, h_m)$ with $h_j \in H_j$, and its inner product is defined, for any $(h_1, \ldots, h_m), (h'_1, \ldots, h'_m) \in \oplus_{j \in \mathbb{N}_m} H_j$, by

$$\langle (h_1, \ldots, h_m), (h'_1, \ldots, h'_m) \rangle := \sum_{j \in \mathbb{N}_m} (h_j, h'_j)_j.$$

This observation suggests to us to define the feature space of kernel K by the direct sum Hilbert space $\mathcal{W} := \oplus_{j \in \mathbb{N}_m} (\mathcal{W}_j \otimes \mathcal{Y})$, and its the map $\Phi : \mathcal{X} \to \mathcal{L}(\mathcal{Y}, \mathcal{W})$, for any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, by

$$\Phi(x)y := (\phi_1(x) \otimes \sqrt{B_1}\ y, \ldots, \phi_m(x) \otimes \sqrt{B_m}\ y). \tag{29}$$

Hence, its adjoint operator $\Phi^* : \mathcal{X} \to \mathcal{L}(\oplus_{j \in \mathbb{N}_m} (\mathcal{W}_j \otimes \mathcal{Y}), \mathcal{Y})$ is given, for any $(w_1 \otimes y_1, \ldots, w_m \otimes y_m) \in \oplus_{j \in \mathbb{N}_m} (\mathcal{W}_j \otimes \mathcal{Y})$, by

$$\Phi^*(x)((w_1 \otimes y_1, \ldots, w_m \otimes y_m)) := \sum_{j \in \mathbb{N}_m} (\phi_j(x), w_j)_{\mathcal{W}_j} \sqrt{B_j}\ y_j.$$

Using the above observation, it is easy to see that, for any $x, t \in \mathcal{X}$, $K(x,t) = \Phi^*(x)\Phi(t)$. Thus K is a multi-task kernel and Φ is a feature map of K.

We are now in a position to state the result about the universality of the kernel K.

Theorem 13 *Suppose that $G_j : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a continuous scalar universal kernel, and $B_j \in \mathcal{L}_+(\mathcal{Y})$ for $j \in \mathbb{N}_m$. Then, $K(x,t) := \sum_{j \in \mathbb{N}_m} G_j(x,t) B_j$ is universal if and only if $\sum_{j \in \mathbb{N}_m} B_j$ is positive definite.*

Proof Following Theorem 11, we need to prove that $\Phi(\mathcal{Z})^\perp = \{0\}$ for any compact set $\mathcal{Z}$ if and only if $\sum_{j \in \mathbb{N}_m} B_j$ is positive definite. To see this, observe that $\mu \in \Phi(\mathcal{Z})^\perp$ implies, for any $(w_1 \otimes y_1, \ldots, w_m \otimes y_m) \in \oplus_{j \in \mathbb{N}_m} (\mathcal{W}_j \otimes \mathcal{Y})$, that

$$\int_{\mathcal{Z}} \sum_{j \in \mathbb{N}_m} (\phi_j(x), w_j)_{\mathcal{W}_j} (\sqrt{B_j}\ y_j, d\mu(x)) = 0.$$

Since $w_j \in \mathcal{W}_j$ is arbitrary, the above equation is equivalent to

$$\int_{\mathcal{Z}} (\phi_j(x), w_j)_{\mathcal{W}_j} (\sqrt{B_j}\ y_j, d\mu(x)) = 0, \quad \forall w_j \in \mathcal{W}_j, y_j \in \mathcal{Y} \text{ and } j \in \mathbb{N}_m, \tag{30}$$

which implies that $(\sqrt{B_j}\ y_j, d\mu(x)) \in \phi(\mathcal{Z})^\perp$ for any $j \in \mathbb{N}_m$. Recall that G_j is universal if and only if $\phi(\mathcal{Z})^\perp = \{0\}$. Therefore, equation (30) holds true if and only if

$$(\mu(E), \sqrt{B_j}\ y_j) = (\sqrt{B_j}\mu(E), y_j) = 0, \quad \forall E \in \mathscr{B}(\mathcal{Z}), \quad y_j \in \mathcal{Y}, j \in \mathbb{N}_m. \tag{31}$$

To move on to the next step, we will show that equation (31) is true if and only if

$$(\mu(E), B_j\mu(E)) = 0, \quad \forall E \in \mathscr{B}(\mathcal{Z}), j \in \mathbb{N}_m. \tag{32}$$

To see this, we observe, for any $j \in \mathbb{N}_m$, that $\|\sqrt{B_j}\mu(E)\|^2 = (\mu(E), B_j\mu(E))$. Hence, equation (32) implies equation (31). Conversely, applying equation (31) with the choice $y_j = \mu(E)$ directly yields equation (32).

Moreover, we know, for any $y \in \mathcal{Y}$ and $j \in \mathbb{N}_m$, that $(B_j y, y)$ is nonnegative. Therefore, equation (32) is equivalent to that

$$\big((\sum_{j\in\mathbb{N}_m} B_j)\mu(E), \mu(E)\big) = 0, \quad \forall E \in \mathscr{B}(\mathcal{Z}). \tag{33}$$

Therefore, we conclude that $\mu \in \Phi(\mathcal{Z})^\perp$ if and only if the above equation holds true.

Obviously, by equation (33), we see that if $\sum_{j\in\mathbb{N}_m} B_j$ is positive definite then $\mu = 0$. This means that kernel K is universal. Suppose that $\sum_{j\in\mathbb{N}_m} B_j$ is not positive definite, that is, there exists a nonzero $y_0 \in \mathcal{Y}$ such that $\|\big(\sum_{j\in\mathbb{N}_m} B_j\big)^{\frac{1}{2}} y_0\|^2 := \big((\sum_{j\in\mathbb{N}_m} B_j)y_0, y_0\big) = 0$. Hence, choosing a nonzero vector measure $\mu := y_0\nu$, with ν a nonzero scalar measure, implies that equation (33) holds true and, thus kernel K is not universal. This finishes the proof of the theorem. ■

Now we are in a position to analyze Examples 1 and 4 given in the Section 2. Since the function K considered in Example 1 is in the form of (29), we conclude that it is a multi-task kernel.

We now discuss a class of kernel which includes that presented in Example 4. To this end, we use the notation $\mathbb{Z}_+ = \{0\} \cup \mathbb{N}$ and, for any smooth function $f : \mathbb{R}^m \to \mathbb{R}$ and index $\alpha = (\alpha_1, \ldots, \alpha_m) \in \mathbb{Z}_+^m$, we denote the α-th partial derivative by $\partial^\alpha f(x) := \frac{\partial^{|\alpha|} f(x)}{\partial^{\alpha_1} x_1 \ldots \partial^{\alpha_m} x_m}$. Then, recall that the Sobolev space W^k with integer order k is the space of real valued functions with norm defined by

$$\|f\|_{W^k}^2 := \sum_{|\alpha|\le k} \int_{\mathbb{R}^m} |\partial^\alpha f(x)|^2 dx, \tag{34}$$

where $|\alpha| = \sum_{j\in\mathbb{N}_m} \alpha_j$, see Stein (1970). This space can be extended to any fractional index $s > 0$. To see this, we need the Fourier transform defined, for any $f \in L^1(\mathbb{R}^m)$, as

$$\hat{f}(\xi) := \int_{\mathbb{R}^m} e^{-2\pi i\langle x,\xi\rangle} f(x) dx, \qquad \forall \xi \in \mathbb{R}^m,$$

see Stein (1970). It has a natural extension to $L^2(\mathbb{R}^m)$ satisfying the Plancherel formula $\|f\|_{L^2(\mathbb{R}^m)} = \|\hat{f}\|_{L^2(\mathbb{R}^m)}$. In particular, we observe, for any $\alpha \in \mathbb{Z}^d$ and $\xi \in \mathbb{R}^m$, that $\widehat{\partial^\alpha f}(\xi) = (2\pi i\xi)^\alpha \hat{f}(\xi)$. Hence, by Plancherel formula, we see, for any $f \in W^k$ with $k \in \mathbb{N}$, that its norm $\|f\|_{W^k}$ is equivalent to

$$\Big(\int_{\mathbb{R}^m} (1 + 4\pi|\xi|^2)^k |\hat{f}(\xi)|^2 d\xi\Big)^{\frac{1}{2}}.$$

This observation suggests to us to introduce fractional Sobolev space W^s (see e.g. Stein (1970)) with any order $s > 0$ with norm defined, for any function f, by

$$\|f\|_{W^s}^2 := \int_{\mathbb{R}^m} (1 + 4\pi^2|\xi|^2)^s |\hat{f}(\xi)|^2 d\xi.$$

Finally, we need the Sobolev embedding lemma which states that, for any $s > \frac{m}{2}$, there exists an absolute constant c such that, for any $f \in W^s$ and any $x \in \mathbb{R}^m$, there holds

$$|f(x)| \le c\|f\|_{W^s},$$

see e.g. Folland (1999); Stein (1970).

The next result extends Example 4 to multivariate functions.

Proposition 14 *Let $\mathcal{Y} = L^2(\mathbb{R}^d)$, $\mathcal{X} = \mathbb{R}$ and $\mathcal{H}$ be the space of real-valued functions with norm*

$$\|f\|^2 := \int_{\mathbb{R}} \Big[\Big\|f(x,\cdot)\Big\|^2_{W^{\frac{d+1}{2}}} + \Big\|\frac{\partial f}{\partial x}(x,\cdot)\Big\|^2_{W^{\frac{d+1}{2}}}\Big] dx. \tag{35}$$

Then this is an RKHS with universal multi-task kernel given, for every $x, t \in \mathcal{X}$ by

$$(K(x,t)y)(r) = e^{-\pi|x-t|} \int_{\mathbb{R}^d} e^{-\pi\|r-s\|} y(s) ds, \quad \forall y \in \mathcal{Y},\ r \in \mathbb{R}^d. \tag{36}$$

Proof For any fixed $t \in \mathbb{R}^d$, it follows from the Sobolev embedding lemma that

$$|f(x,t)| \le c\|f(\cdot,t)\|_{W^1}.$$

Combining this with the definition of Sobolev space W^1 given by equation (34), we have that

$$\begin{aligned}\|f(x)\|_{\mathcal{Y}}^2 &\le c^2 \int_{\mathbb{R}^d} \|f(\cdot,t)\|_{W^1}^2 dt \\ &= c^2 \int_{\mathbb{R}^d} \Big(\int_{\mathbb{R}} \big|f(x,t)\big|^2 + \big|\frac{\partial f}{\partial x}(x,t)\big|^2 dt\Big) dx \le c^2\|f\|^2.\end{aligned}$$

Since, for any $y \in \mathcal{B}_1$ and $x \in \mathbb{R}$, $|(y, f(x))_{\mathcal{Y}}| \le \|y\|_{\mathcal{Y}}\|f(x)\|_{\mathcal{Y}} \le \|f(x)\|_{\mathcal{Y}}$, by the above equation there exists a constant c' such that, for any $y \in \mathcal{B}_1, x \in \mathbb{R}$ and $f \in \mathcal{H}$,

$$|(y, f(x))_{\mathcal{Y}}| \le c'\|f\|.$$

Hence, by the Riesz representation lemma, $\mathcal{H}$ is an RKHS (Micchelli and Pontil, 2005).

Next, we confirm equation (36) is the kernel associated to $\mathcal{H}$. To this end, it suffices to show that the reproducing property holds, that is, for any $f \in \mathcal{H}$, $y \in \mathcal{Y}$ and $x \in \mathcal{X}$

$$(y, f(x))_{\mathcal{Y}} = \langle K_x y, f \rangle, \tag{37}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product in $\mathcal{H}$.

By the Plancherel formula, we observe that the left-hand side of equation (37) equals

$$\int_{\mathbb{R}^d} \hat{y}(\tau) \Big[\overline{\int_{\mathbb{R}} e^{2\pi i \langle x, \xi \rangle} \hat{f}(\xi, \tau) d\xi} \Big] d\tau = \int_{\mathbb{R}^d} \int_{\mathbb{R}} \hat{y}(\tau) e^{-2\pi i \langle x, \xi \rangle} \overline{\hat{f}(\xi, \tau)} d\xi d\tau.$$

On the other hand, note that $K_x y(x') := (K(x', x)y \in \mathcal{Y}$, and consider its Fourier transform

$$\widehat{K}(\cdot, x)y)(\xi, \tau) = \int_{\mathbb{R}^d} \int_{\mathbb{R}} e^{-2\pi i \langle x', \xi \rangle} e^{-2\pi i \langle r, \tau \rangle} (K(x', x)y)(r) dr dx'.$$

Using equation (36) and the Plancherel formula, the integral on the right hand of the above equation is equal to

$$\frac{e^{-2\pi i \langle x, \xi \rangle}}{(1 + 4\pi^2 |\xi|^2)} \frac{\hat{y}(\tau)}{(1 + 4\pi^2 |\tau|^2)^{\frac{d+1}{2}}}. \tag{38}$$

However, the right-hand side of equation (37) is identical to

$$\int_{\mathbb{R}^d} \int_{\mathbb{R}} \widehat{K}(\cdot, x)y)(\cdot)(\xi, \tau) \overline{\hat{f}(\xi, \tau)} (1 + 4\pi^2 |\xi|^2)(1 + 4\pi^2 |\tau|^2)^{\frac{d+1}{2}} d\tau d\xi.$$

Putting (38) into the above equation, we immediately know that the reproducing property (37) holds true. This verifies that K is the reproducing kernel of the Hilbert space $\mathcal{H}$.

To prove the universality of this kernel, let $\mathcal{Z}$ be any prescribed compact subset of $\mathcal{X}$, we define the Laplace kernel, for any $x, t \in \mathbb{R}$, by $G(x, t) := e^{-|x-t|}$ and the operator $B : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ by

$$Bg(r) := \int_{\mathbb{R}^d} e^{-\|r-s\|} g(s) ds, \quad \forall\ r \in \mathbb{R}^d.$$

Then, $K(x, t) = e^{-|x-t|} B$ and moreover

$$\widehat{Bg}(\tau) = c_d \frac{\hat{g}(\tau)}{(1 + 4\pi^2 |\tau|^2)^{\frac{d+1}{2}}}. \tag{39}$$

By Theorem 12, it now suffices to prove that G is universal and B is positive definite. To this end, note that there exists c_d such that

$$G(x, t) = c_d \int_{\mathbb{R}} \frac{e^{-2\pi i \langle x-t, \xi \rangle}}{1 + 4\pi^2 |\xi|^2} d\xi.$$

Since the weight function $\frac{1}{1+4\pi^2|\xi|^2}$ is positive, G is universal according to Micchelli et al. (2003).

To show the positive definiteness of B, we obtain from equation (39) and the Plancherel formula that

$$(Bg, g) = c_d \int_{\mathbb{R}^d} \frac{|\hat{g}(\tau)|^2 d\tau}{(1 + 4\pi^2|\tau|^2)^{\frac{d+1}{2}}}.$$

From this observation, the assertion follows. ■

We now discuss continuous parameterized multi-task kernels. For this purpose, let Ω be a locally compact Hausdorff space and, for any $\omega \in \Omega$, $B(\omega)$ be an $n \times n$ positive semi-definite matrix. We are interested in the kernel of the following form

$$K(x, t) = \int_\Omega G(\omega)(x, t)B(\omega)dp(\omega), \quad \forall x, t \in \mathcal{X}, \tag{40}$$

where p is a measure on Ω. We investigate this kernel in the case that, for any $\omega \in \Omega$, $G(\omega)$ is a scalar kernel with a feature representation given, for any $x, t \in \mathcal{X}$, by the formula $G(x, t) = \langle \phi_\omega(x), \phi_\omega(t)\rangle_{\mathcal{W}}$. Now, we introduce the Hilbert space $\widetilde{\mathcal{W}} = L^2(\Omega, \mathcal{W} \otimes \mathcal{Y}, p)$ with norm defined, for any $f : \Omega \to \mathcal{W} \otimes \mathcal{Y}$, by

$$\|f\|^2_{\widetilde{\mathcal{W}}} := \int_\Omega \|f(\omega)\|^2_{\mathcal{W}\otimes\mathcal{Y}} dp(\omega).$$

Next, we define a map $\Phi : \mathcal{X} \to \mathcal{L}(\mathcal{Y}, L^2(\Omega, \mathcal{W} \otimes \mathcal{Y}, p))$, for any $x \in \mathcal{X}$ and $\omega \in \Omega$, by

$$\Phi(x)y(\omega) := \phi_\omega(x) \otimes (\sqrt{B(\omega)}y).$$

By an argument similar to that used just before Theorem 13, we conclude that K is a multi-task kernel and has the feature map Φ with feature space $\widetilde{\mathcal{W}}$.

We are ready to present a sufficient condition on the universality of K.

Theorem 15 *Let p be a measure on Ω and for every ω in the support of p, let $G(\omega)$ be a continuous universal kernel and $B(\omega)$ a positive definite operator. Then, the multi-task kernel K defined by equation (40) is universal.*

Proof Following Theorem 11, for a compact set $\mathcal{Z} \subseteq \mathcal{X}$ suppose that there exists a vector measure μ such that

$$\int_{\mathcal{Z}} \phi_\omega(x) \otimes \sqrt{B(\omega)}(d\mu(x)) = 0.$$

Therefore, there exists a $\omega_0 \in \text{support}(p)$ satisfying $\int_{\mathcal{Z}} \phi_{\omega_0}(x) \otimes \sqrt{B(\omega_0)}\big(d\mu(x)\big) = 0$. Equivalently, $\int_{\mathcal{Z}} \phi_{\omega_0}(x)\big(\sqrt{B(\omega_0)}d\mu(x), y\big) = 0$ for any $y \in \mathcal{Y}$. Since we assume $G(\omega_0)$ is universal, appealing to the feature characterization in the scalar case (Micchelli et al., 2006) implies that the scalar measure $(\sqrt{B(\omega_0)}d\mu(x), y) = 0$. Consequently, we obtain that $\mu \equiv 0$ since $y \in \mathcal{Y}$ is arbitrary. This completes the proof of this theorem. ■

Next we offer a concrete example of the above theorem.

Example 5 *Suppose the measure p over $[0,\infty)$ does not concentrate on zero and $B(\omega)$ be a positive definite $n \times n$ matrix for each $\omega \in (0,\infty)$. Then the kernel $K(x,t) = \int_0^\infty e^{-\omega\|x-t\|^2} B(\omega) dp(\omega)$ is a multi-task universal kernel.*

Further specializing this example, we choose the measure p to be the Lebesgue measure on $[0,\infty)$ and choose $B(\omega)$ in the following manner. Let A be $n \times n$ symmetric matrices. For every $\omega > 0$, we define the (i,j)-th entry of the matrix $B(\omega)$ as $e^{-\omega A_{ij}}$, $i,j \in \mathbb{N}_n$. Recall that a matrix A is *conditionally negative semi-definite* if, for any $c_i \in \mathbb{R}, i \in \mathbb{N}_n$ with $\sum_{i\in\mathbb{N}_n} c_i = 0$, then the quadratic form satisfies $\sum_{i,j\in\mathbb{N}_n} c_i A_{ij} c_j \le 0$. A well-known theorem of I. J. Schoenberg (see e.g. Micchelli (1986)) state that $B(\omega)$ is positive semi-definite for all $\omega > 0$ if and only if A is conditionally negative semi-definite. Moreover, if the elements of the conditionally negative semi-definite matrix A satisfy, for any $i,j \in \mathbb{N}_n$, the inequalities $A_{ij} > \frac{1}{2}(A_{ii} + A_{jj})$ and $A_{ii} > 0$, then $B(\omega)$ is positive definite (Micchelli, 1986). With this choice of A, the universal kernel in Example 5 becomes

$$\Big(K(x,t)\Big)_{ij} = \frac{1}{\|x-t\|^2 + A_{ij}}, \quad \forall i,j \in \mathbb{N}_n.$$

5.2 Transformation Kernels

In this subsection we explore matrix-valued kernels produced by transforming scalar kernels. To introduce this type of kernels, let $\mathcal{Y} = \mathbb{R}^n$, $\widetilde{\mathcal{X}}$ be a Hausdorff space and T_p be a map from $\mathcal{X}$ to $\widetilde{\mathcal{X}}$ (not necessary linear) for $p \in \mathbb{N}_n$. Then, given a continuous scalar kernel $G : \widetilde{\mathcal{X}} \times \widetilde{\mathcal{X}} \to \mathbb{R}$, we consider the matrix-valued kernel on $\mathcal{X}$ defined by

$$K(x,t) := \Big(G(T_p x, T_q t)\Big)_{p,q=1}^n, \quad \forall x,t \in \mathcal{X}. \tag{41}$$

Proposition 16 *Let G be a scalar kernel and K be defined by (41). Then, K is a matrix-valued kernel.*

Proof For any $m \in \mathbb{N}$, $\{y_i : y_i \in \mathbb{R}^n, i \in \mathbb{N}_m\}$ and $\{x_i : x_i \in \widetilde{\mathcal{X}}, i \in \mathbb{N}_m\}$ then

$$\sum_{i,j\in\mathbb{N}_m} (y_i, K(x_i,x_j) y_j) = \sum_{p,i}\sum_{q,j} y_{pi} y_{qj} G(T_p x_i, T_q x_j).$$

Since G is a scalar reproducing kernel on $\mathcal{Z}$, the last term of the above equality is nonnegative, and hence K is positive semi-definite matrix-valued kernel. This completes the proof of the assertion. ■

We turn our attention to the characterization of the universality of K defined by equation (41). To this end, we assume that the scalar kernel G has a feature map $\phi : \mathcal{X} \to \mathcal{W}$ and define the mapping $\Phi(x) : \mathbb{R}^n \to \mathcal{W}$, for any $y = (y_1,\dots,y_n) \in \mathbb{R}^n$, by $\Phi(x)y = \sum_{p\in\mathbb{N}_n} y_p \phi(T_p x)$. Its adjoint operator $\Phi(x)^* : \mathcal{W} \to \mathbb{R}^n$ is given, for any $w \in \mathcal{W}$, as $\Phi^*(x)w = (\langle\phi(T_1 x), w\rangle_{\mathcal{W}}, \dots, \langle\phi(T_n x), w\rangle_{\mathcal{W}})$. Then, for any $x,t \in \mathcal{X}$, the kernel $K(x,t) = \Phi^*(x)\Phi(t)$ and thus, we conclude that $\mathcal{W}$ is the feature space of K and Φ is its feature map.

We also need some further notation and definitions. For a map $T : \mathcal{X} \to \widetilde{\mathcal{X}}$, we denote its range space by $T\mathcal{X} := \{Tx : x \in \mathcal{X}\}$ and $T^{-1}(E) := \{x : Tx \in E\}$ for any $E \subset \widetilde{\mathcal{X}}$.

In addition, we say that T is continuous if $T^{-1}(U)$ is open whenever U is a open set in $\widetilde{\mathcal{X}}$. Finally, for any scalar Borel measure ν on $\mathcal{X}$ and a continuous map T from $\mathcal{X}$ to $\widetilde{\mathcal{X}}$, we introduce the *image measure* $\nu \circ T^{-1}$ on $\widetilde{\mathcal{X}}$ defined, for any $E \in \mathscr{B}(\widetilde{\mathcal{X}})$, by $(\nu \circ T^{-1})(E) := \nu(\{x \in \mathcal{X} : Tx \in E\})$.

We are ready to state the result about universality of the kernel K in equation (41).

Proposition 17 *Let G be a scalar universal kernel, $T_p : \mathcal{X} \to \widetilde{\mathcal{X}}$ be continuous for each $p \in \mathbb{N}_n$ and define the kernel K by equation (41). Then K is universal if and only the sets $T_q\mathcal{X}$, $q \in \mathbb{N}_n$, are pairwise disjoint and T_q is one-to-one for each $q \in \mathbb{N}_n$.*

Proof Following Theorem 11, for any compact set $\mathcal{Z} \subseteq \mathcal{X}$, it suffices to verify the equation $\Phi(\mathcal{Z})^\perp = \{0\}$. Before doing so, we recall that, by Lemma 10 and the remark which followed it, for any vector measure $\mu \in \mathcal{M}(\mathcal{Z}, \mathbb{R}^n)$, there exists a scalar regular measure $\nu \in \mathcal{M}(\mathcal{Z} \times \mathcal{B}_1)$ such that

$$d\mu(t) = \Big(\int_{\mathcal{B}_1} y_1 d\nu(t,y), \ldots, \int_{\mathcal{B}_1} y_n d\nu(t,y) \Big).$$

Hence, any vector measure μ can be represented as $\mu = (\mu_1, \ldots, \mu_n)$ where each μ_i is a scalar measure. Then, $\mu \in \Phi(\mathcal{Z})^\perp$ can be rewritten as

$$\sum_{q \in \mathbb{N}_n} \int_{\mathcal{Z}} \phi(T_q t) d\mu_q(t) = 0.$$

Equivalently, if $\widetilde{\mathcal{Z}} := \cup_{q \in \mathbb{N}_n} T_q \mathcal{Z}$ we conclude that

$$\int_{\widetilde{\mathcal{Z}}} \phi(z) d\big(\sum_{q \in \mathbb{N}_n} \mu_q \circ T_q^{-1} \big)(z) = 0. \tag{42}$$

Since T_q is continuous for any $q \in \mathbb{N}_n$, the range space $T_q\mathcal{Z}$ is compact and so is $\widetilde{\mathcal{Z}}$. Recall (Micchelli et al., 2006) that the scalar kernel G is universal on $\widetilde{\mathcal{Z}}$ if and only if its feature ϕ is universal on $\widetilde{\mathcal{Z}}$. Therefore, the above equation is reduced to the form

$$\sum_{q \in \mathbb{N}_n} \mu_q \circ T_q^{-1} = 0.$$

Consequently, we conclude that K is universal if and only if

$$\big\{ (\mu_1, \ldots, \mu_n) : \sum_{q \in \mathbb{N}_n} \mu_q \circ T_q^{-1} = 0 \big\} = \big\{ 0 \big\}. \tag{43}$$

With the above derivation, we can now prove the necessity. Suppose that $\{T_q\mathcal{X} : q \in \mathbb{N}_n\}$ is not pairwise disjoint. Without loss of generality, we assume that $T_1\mathcal{X} \cap T_2\mathcal{X} \neq \emptyset$. That means there exists $x_1, x_2 \in \mathcal{X}$ such that $T_1 x_1 = T_2 x_2 = z_0$. Let $\mu_q \equiv 0$ for $q \geq 3$, and denote by $\delta_{x=x'}$ the point distribution at $x' \in \mathcal{X}$. Then, choosing $\mu_1 = \delta_{x=x_1}$, and $\mu_2 = -\delta_{x=x_2}$ implies that equation (43) holds true. By Theorem 11 in Section 4, we know that K is not universal. This completes the first assertion.

Now suppose that there is a map, for example T_p, which is not one-to-one. This implies that there exists $x_1, x_2 \in \mathcal{X}$, $x_1 \neq x_2$, such that $T_p x_1 = T_p x_2$. Hence, if we let $\mu_q = 0$ for any $q \neq p$ and $\mu_p = \delta_{x=x_1} - \delta_{x=x_2}$ then $\sum_{q\in\mathbb{N}_n} \mu_q \circ T_q^{-1} = 0$,. But $\mu_p \neq 0$, hence, by Theorem 11, K is not universal. This completes the our assertion.

Finally, we prove the sufficiency. Since $\mu_q \circ T_q^{-1}$ only lives on $T_q\mathcal{X}$ and $\{T_q\mathcal{X} : q \in \mathbb{N}_n\}$ is pairwise disjoint, then $\sum_{q\in\mathbb{N}_n} \mu_q \circ T_q^{-1} = 0$ is equivalent to $\mu_q \circ T_q^{-1} = 0$ for each $q \in \mathbb{N}_n$. However, since T_q is one-to-one, $E = T_q^{-1}(T_q(E))$ for each Borel set $E \in \mathscr{B}(\mathcal{X})$. This means that $\mu_q(E) = \mu_q \circ T_q^{-1}(T_q(E)) = 0$ for any $E \in \mathscr{B}(\mathcal{X})$. This concludes the proof of the proposition. ■

We end this subsection with detailed proofs of our claims about the examples presented in Section 2. Indeed, we already proved the positive semi-definiteness of the kernel in Example 2 by Proposition 16. Below, we prove the claim that the function K given by equation (7) is not a kernel in general.

Proposition 18 *Let $\sigma_{pq} > 0$ and $\sigma_{pq} = \sigma_{qp}$ for any $p, q \in \mathbb{N}_n$. Then, the matrix-valued function defined by*

$$K(x,t) := \left(e^{-\sigma_{pq}\|x-y\|^2}\right)_{p,q=1}^{n}, \quad \forall x, t \in \mathcal{X}$$

is a multi-task kernel if and only if for some constant σ, $\sigma_{pq} = \sigma$ for any $p, q \in \mathbb{N}_n$.

Proof When $\left(\sigma_{pq}\right)_{p,q=1}^{n}$ is a constant matrix then K is positive semi-definite. Conversely, suppose K is a multi-task kernel which means, for any $m \in \mathbb{N}$ and $x_i \in \mathcal{X}$ with $i \in \mathbb{N}_m$, that the double-indexed $nm \times nm$ matrix

$$\left(G((i,p),(j,q)) = e^{-\sigma_{pq}\|x_i-x_j\|^2}\right)_{(i,p),(j,q)\in\mathbb{N}_m\times\mathbb{N}_n} \tag{44}$$

is positive semi-definite.

We choose any distinct positive integers p_0 and q_0. In equation (44), we specify any m, n with $m \geq n$ such that $p_0, q_0 \in \mathbb{N}_m$, $x_1, \ldots, x_n$ with $x_{p_0} \neq x_{q_0}$ and set $c = \|x_{p_0} - x_{q_0}\|^2$. Therefore, we conclude that the matrix

$$\begin{pmatrix} 1 & 1 & \exp\{-c\sigma_{p_0p_0}\} & \exp\{-c\sigma_{p_0q_0}\} \\ 1 & 1 & \exp\{-c\sigma_{q_0p_0}\} & \exp\{-c\sigma_{q_0q_0}\} \\ \exp\{-c\sigma_{p_0p_0}\} & \exp\{-c\sigma_{p_0q_0}\} & 1 & 1 \\ \exp\{-c\sigma_{q_0p_0}\} & \exp\{-c\sigma_{q_0q_0}\} & 1 & 1 \end{pmatrix} \tag{45}$$

is positive semi-definite. Consequently, the determinant of its 3×3 sub-matrix in the upper right hand corner, which equals $-\left[\exp\{-c\sigma_{p_0p_0}\} - \exp\{-c\sigma_{q_0p_0}\}\right]^2$, is nonnegative. Therefore, we conclude that $\sigma_{p_0p_0} = \sigma_{q_0p_0}$. ■

5.3 Hessian of Gaussian Kernels

In this subsection we consider the universal example of the Hessian of scalar Gaussian kernels (Example 3 in Section 2). To introduce this type of kernels, we let $\mathcal{X} = \mathcal{Y} = \mathbb{R}^n$ and G be a Gaussian kernel with deviation σ, that is, for any $x \in \mathbb{R}^n$, $G(x) = e^{-\|x\|^2/\sigma}$ with $\sigma > 0$. Then, the Hessian matrix of the kernel G is

$$K(x,t) := \Big(-(\partial_p \partial_q G)(x-t) \Big)_{p,q=1}^{n} \qquad \forall x, t \in \mathbb{R}^n. \tag{46}$$

and so alternatively, K has the form

$$K(x,t) = 4\pi \big(2\pi\sigma\big)^{n/2} \int_{\mathbb{R}^n} e^{2\pi i \langle x-t, \xi\rangle} \xi\xi^T e^{-\sigma\|\xi\|^2} d\xi. \tag{47}$$

Corollary 19 *Let $n \geq 1$ and $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^{n\times n}$ be defined by (46). Then, K is a matrix-valued kernel which is universal if and only if $n = 1$.*

Proof The fact that K is positive semi-definite directly follows from the observation, for any $m \in \mathbb{N}$, $\{y_i : y_i \in \mathbb{R}^n, i \in \mathbb{N}_m\}$ and $\{x_i : x_i \in \mathcal{X}, i \in \mathbb{N}_m\}$, that

$$\sum_{i,j,\in\mathbb{N}_m} (y_i, K(x_i,x_j)y_j) = 4\pi \big(2\pi\sigma\big)^{n/2} \int_{\mathbb{R}^d} \Big\| \sum_{i\in\mathbb{N}_m} \langle y_i, \xi\rangle e^{2\pi i\langle x_i,\xi\rangle} \Big\|^2 e^{-\sigma\|\xi\|^2} d\xi.$$

In order to prove the universality of K, we follow Theorem 11. For this purpose, we assume that $\mathcal{Z}$ is a compact subset of $\mathcal{X}$ and $\mu \in \mathcal{K}(\mathcal{Z})^\perp$, that is,

$$\int_{\mathcal{Z}} K(x,t)(d\mu(t)) = 0 \quad \forall x \in \mathcal{Z}. \tag{48}$$

By equation (47), this equation is equivalent to

$$\int_{\mathbb{R}^n} e^{2\pi i\langle x,\xi\rangle} \xi e^{-\sigma\|\xi\|^2} \int_{\mathcal{Z}} e^{-2\pi i\langle t,\xi\rangle} (\xi, d\mu(t)) d\xi = 0, \quad \forall x \in \mathcal{Z},$$

which implies, by integrating both sides of this equation with respect to $x \in \mathbb{R}^n$, that

$$\int_{\mathbb{R}^n} e^{-\sigma\|\xi\|^2} \Big\| \int_{\mathcal{Z}} e^{-2\pi i\langle t,\xi\rangle} (\xi, d\mu(t)) \Big\|^2 d\xi = 0.$$

Consequently, equation (48) is identical to the equation

$$\int_{\mathcal{Z}} e^{-2\pi i\langle t,\xi\rangle} (\xi, d\mu(t)) = 0, \qquad \forall \xi \in \mathbb{R}^n. \tag{49}$$

If $n = 1$, the above equation means that

$$\int_{\mathcal{Z}} e^{-2\pi i t\xi} d\mu(t) = 0, \qquad \forall \xi \in \mathbb{R}.$$

Taking the k-th derivative with respect to ξ of both sides of this equation and set $\xi = 0$, we have, for every $k \in \mathbb{N}$, that

$$\int_{\mathcal{Z}} t^k d\mu(t) = 0.$$

Since polynomials are dense in $\mathcal{C}(\mathcal{Z})$, we conclude from the above equation that $\mu = 0$. Hence, by Theorem 11, the kernel K is universal when $n = 1$.

If $n \geq 2$, we choose $\mu_q = 0$ for $q \geq 3$ and $d\mu_1(t) = dt_1(\delta_{t_2=1} - \delta_{t_2=-1}) \prod_{p=3}^n \delta_{t_p=0}$ and $d\mu_2(t) = (\delta_{t_1=-1} - \delta_{t_1=1})dt_2 \prod_{p=3}^n \delta_{t_p=0}$, and note that

$$\int_{[-1,1]^n} e^{-2\pi i\langle t,\xi\rangle} d\mu_1(t) = (-2\pi i \sin(2\pi\xi_2)) \frac{sin(2\pi\xi_1)}{\pi\xi_1},$$

and

$$\int_{[-1,1]^n} e^{-2\pi i\langle t,\xi\rangle} d\mu_2(t) = (2\pi i \sin(2\pi\xi_1)) \frac{sin(2\pi\xi_2)}{\pi\xi_2}.$$

Therefore, we conclude that

$$\int_{[-1,1]^n} e^{-2\pi i\langle t,\xi\rangle} (\xi, d\mu(t)) = \xi_1 \int_{[-1,1]^n} e^{-2\pi i\langle t,\xi\rangle} d\mu_1(t) + \xi_2 \int_{[-1,1]^n} e^{-2\pi i\langle t,\xi\rangle} d\mu_2(t) = 0.$$

Hence, the kernel K is not universal when $n \geq 2$. ■

5.4 Projection Kernels

In the final subsection we introduce a class of multi-task kernels associated with projection operators of scalar kernels.

We start with some notation and definitions. Let $\mathcal{X} \subseteq \mathbb{R}^d$, $\Omega \subseteq \mathbb{R}^m$ be a compact set and $\mathcal{Y} = L^2(\Omega)$. We also need a continuous scalar kernel $G : (\mathcal{X} \times \Omega) \times (\mathcal{X} \times \Omega) :\to \mathbb{R}$ with a feature representation given, for any $x, x' \in \mathcal{X}$ and $t, s \in \Omega$, by

$$G((x,t),(x',s)) = \langle \phi(x,t), \phi(x',s)\rangle_{\mathcal{W}}. \tag{50}$$

Then, the projection kernel $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y}, \mathcal{Y})$ is defined, for any $f \in L^2(\Omega)$, by

$$(K(x,x')f)(t) := \int_\Omega G((x,t),(x',s))f(s)ds, \quad \forall\, x, x' \in \mathcal{X}, \quad t \in \mathbb{R}. \tag{51}$$

We first show that K is a multi-task kernel. To see this, for any $m \in \mathbb{N}$, $\{x_i : x_i \in \mathcal{X}, i \in \mathbb{N}_m\}$, and $\{y_i : y_i \in L^2(\Omega), i \in \mathbb{N}_m\}$ there holds

$$\begin{aligned} \textstyle\sum_{i,j\in\mathbb{N}_m}(K(x_i,x_j)y_j, y_i) &= \textstyle\sum_{i,j\in\mathbb{N}_m} \int_\Omega \int_\Omega G((x_i,t),(x_j,s))y_j(s)y_i(t)dtds \\ &= \textstyle\sum_{i\in\mathbb{N}_m} \left\| \int_\Omega \phi(x_i,s)y_i(s)ds \right\|^2 \geq 0, \end{aligned}$$

which implies that K is a kernel.

To investigate its universality from the feature perspective, we define the mapping $\Phi : \mathcal{X} \to \mathcal{L}(\mathcal{Y}, \mathcal{W})$, for any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, by

$$\Phi(x)y := \int_\Omega \phi(x,s)y(s)ds,$$

and also its adjoint operator Φ^* is given, for any $w \in \mathcal{W}$, by $\Phi^*(x)w = \langle \phi(x,\cdot), w\rangle_{\mathcal{W}}$. Hence, for any $x, x' \in \mathcal{X}$, we conclude that $K(x,x') = \Phi^*(x)\Phi(x')$ which implies that K is a multi-task kernel and Φ is its associated feature map.

Our next goal is to prove the universality of K.

Theorem 20 *Let G and K be defined as in equations* (50) *and* (51)*. If G is a universal scalar kernel then K is a universal multi-task kernel.*

Proof By Theorem 11, it suffices to show that, for any compact $\mathcal{Z} \subseteq \mathcal{X}$, whenever there exists a vector measure μ such that

$$\int_{\mathcal{Z}} \Phi(x)(d\mu(x)) = 0,$$

then $\mu = 0$. Note that μ is an $L^2(\Omega)$-valued measure. Hence, μ can alternatively be interpreted as a measure $\mu(\cdot,\cdot)$ on $\mathcal{Z} \times \Omega$ defined, for any $E \in \mathscr{B}(\mathcal{Z})$ and $E' \in \mathscr{B}(\Omega)$, by $\mu(E,E') := \int_{E'} \mu(E)(s)ds$. From this observation, we know that

$$\int_{\mathcal{Z}} \Phi(x)(d\mu(x)) = \int_{\mathcal{Z}} \int_{\Omega} \phi(x,s) d\mu(x,s).$$

Since $\mathcal{Z}$ and Ω are both compact, then $\mathcal{Z}\times\Omega$ is also compact by Tychonoff theorem (Folland, 1999, p.136). By assumption, G is universal on $\mathcal{X} \times \Omega$ and ϕ is its feature map, and thus we conclude that the scalar measure $d\mu(x,s)$ is the zero measure. This means that, for any $E \in \mathscr{B}(\mathcal{Z})$ and $E' \in \mathscr{B}(\Omega)$,

$$\int_{E'} \mu(E)(s)ds = 0.$$

Since E, E' are arbitrary, we conclude that the vector measure $\mu = 0$ which completes the assertion. ■

6. Conclusion

We have presented a characterization of multi-task kernels from a Functional Analysis perspective. Our main result, Theorem 4 established the equivalence between two spaces associated with the kernel. The first space, $\mathcal{C}_K(\mathcal{Z},\mathcal{Y})$, is the closure of the linear span of kernel sections; the second space, $\mathcal{C}_\Phi(\mathcal{Z},\mathcal{Y})$, is the closure of the linear span of the features associated with the kernel. In both cases, the closure was relative to the Banach space of continuous vector-valued functions. This result is important in that it allows one to verify the universality of kernel directly by considering its features.

We have presented two alternate proofs of Theorem 4. The first proof builds upon the work of Micchelli et al. (2006) and the observation that a multi-task kernel can be reduced to a standard scalar-valued kernel on the cross product space $\mathcal{Z}\times\mathcal{Y}$. The second proof relies upon the theory of vector-measures. This proof is constructive and provides necessary and sufficient conditions on the universality of a multi-task kernels. They are summarized in Theorem 11, which is our main tool for verifying the universality of a multi-task kernel.

In both proofs, an important ingredient is a principle from Functional Analysis, which uses the notion of the *annihilator* set. This principle, which is a consequence of the Hahn-Banach Theorem, states that two closed linear subspaces of a Banach space — in our case $\mathcal{C}_K(\mathcal{Z}, \mathcal{Y})$ and $\mathcal{C}_\Phi(\mathcal{Z}, \mathcal{Y})$ — are equal if and only if whenever a bounded linear functional vanishes on one of them, it also vanishes on the other one.

A substantial part of the paper has been devoted to present several examples of multi-task kernels, some of which are valuable for applications. Although much remains to be done on developing applications of the theory of universal kernels, we hope that our theoretical findings, as they are illustrate through the examples, will motivate further work on multi-task learning in applied machine learning.

Acknowledgments

This work was supported by EPSRC Grants GR/T18707/01 and EP/D052807/1 by the IST Programme of the European Community, under the PASCAL Network of Excellence IST-2002-506778. The first author was supported by the NSF grant 0325113, the FIRB project RBIN04PARL, the EU Integrated Project Health-e-Child IST-2004-027749, and the City University of Hong Kong grant No.7200111(MA). The second author is supported by NSF grant DMS 0712827.

We are grateful to Alessandro Verri, Head of the Department of Computer Science at the University of Genova for providing us with the opportunity to complete part of this work in a scientifically stimulating and friendly environment. We also wish to thank the referees for their valuable comments.

References

L. Amodei. Reproducing kernels of vector–valued function spaces. Proc. of Chamonix, A. Le Meehaute et al. Eds., pp. 1–9, 1997.

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

S. K. Berberian. *Notes on Spectral Theory*. New York: Van Nostrand, 1966.

J. Burbea and P. Masani. *Banach and Hilbert Spaces of Vector-Valued Functions*. Pitman Research Notes in Mathematics Series, 90, 1984.

A. Caponnetto and E. De Vito. Optimal rates for regularized least-squares algorithm. *Foundations of Computational Mathematics*, DOI 10.1007/s10208-006-0196-8, August 2006.

C. Carmeli, E. De Vito, and A. Toigo. Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. To appear in *Analysis and Applications*, 2006.

F. Cucker and S. Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc.* 39: 1–49, 2001.

D. R. Chen, Q. Wu, Y. Ying, and D.X. Zhou. Support vector machine soft margin classifiers: error analysis. *Journal of Machine Learning Research* 5: 1143–1175, 2004.

A. Devinatz. On measurable positive definite operator functions. *J. Lonon Math. Soc.*, 35: 417–424, 1960.

N. Dinculeanu. *Vector Measures*. Pergamon, Berlin, 1967.

J. Diestel and J. J. Uhl, Jr. *Vector Measures*. AMS, Providence (Math Surveys 15), 1977.

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

G. B. Folland. *Real Analysis: Modern Techniques and Their Applications*. 2nd edition, New York, John Wiley & Sons, 1999.

A. Gretton, K.M. Borgwardt, M. Rasch, B. Schölkopf and A.J. Smola. A kernel method for the two-sample problem. NIPS 2006.

P. Lax. *Functional Analysis.*, John Wiley & Sons, 2002.

S. Lowitzsch. A density theorem for matrix-valued radial basis functions. *Numerical Algorithms*, 39: 253-256, 2005.

C. A. Micchelli, Interpolation of scattered data: distances matrices and conditionally positive definite functions. *Constructive Approximation*, 2: 11–22, 1986.

C. A. Micchelli and M. Pontil. A function representation for learning in Banach spaces. Proc. 17-th Annual Conf. on Learning Theory (COLT'04), 2004.

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

C.A. Micchelli and M. Pontil. Feature space perspectives for learning the kernel *Machine Learning*, 66: 297–319, 2007.

C. A. Micchelli, Y. Xu, and P. Ye. Cucker Smale learning theory in Besov spaces. *NATO Science Series sub Series III Computer and System Science*, 190: 47–68, 2003.

C. A. Micchelli, Y. Xu, and H. Zhang. Universal kernels. *J. Machine Learning Research*, 7: 2651-2667, 2006.

S. Mukherjee and D.X. Zhou. Learning coordinate covariances via gradients, *J. of Machine Learning Research* 7: 519-549, 2006.

T. Poggio, S. Mukherjee, R. Rifkin, A. Rakhlin, and A. Verri. b. In *Uncertainty in Geometric Computations,* J. Winkler and M. Niranjan (eds.), Kluwer, 131–141, 2002.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*, MIT Press, 2006.

M. Reisert and H. Burkhardt. Learning equivariant functions with matrix valued kernels. *J. Machine Learning Research*, 8: 385–408, 2007.

B. Schölkopf and A. J. Smola. *Learning with Kernels.* The MIT Press, Cambridge, MA, USA, 2002.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis.* Cambridge University Press, 2004.

E. Solak, R.Murray-Smith, W.E. Leithead, D.J. Leith and C.E. Rasmussen. Derivative observations in Gaussian Process models of dynamic Systems. NIPS 2002.

E. M. Stein. *Singular Integrals and Differential Properties of Functions*, Princeton University Press, Princeton, NJ, 1970.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *J. Machine Learning Research*, 2: 67–93, 2001.

I. Steinwart, D. Hush, and C. Scovel. Function classes that approximate the Bayes risk, *Proceeding of COLT*, 79–93, 2006.

K. Yosida. *Functional Analysis*. 6th edition, Springer-Verlag, 1980.

E. Vazquez and E. Walter. Multi-output support vector regression. 13th IFAC Symposium on System Identification, 2003.

D. X. Zhou. Density problem and approximation error in learning theory. Preprint, 2003.

Appendix

This appendix gives the proof of Lemmas 9 and 10 in Section 4.

Proof of Lemma 9 By the definition of the integral appearing in the right-hand side of equation it follows (21) (see e.g. Diestel and Uhl, Jr., 1977), for any $f \in \mathcal{C}(\mathcal{Z}, \mathcal{Y})$, that

$$|L_\mu f| \leq \|\mu\| \sup_{x \in \mathcal{Z}} \|f(x)\|_{\mathcal{Y}}. \tag{52}$$

Therefore, we obtain that $\|L_\mu\| \leq \|\mu\|$, and thus $L_\mu \in \mathcal{C}^*(\mathcal{Z}, \mathcal{Y})$.

To show that $\|L\| = \|\mu\|$, it remains to establish that $\|\mu\| \leq \|L_\mu\|$. To this end, for any $\varepsilon > 0$ and, by the definition of $\|\mu\|$, we conclude that there exist pairwise disjoint sets $\{A_j : j \in \mathbb{N}_n\}$ such that $\cup_{j \in \mathbb{N}_n} A_j \subseteq \mathcal{Z}$ and $\|\mu\| := |\mu|(\mathcal{Z}) \leq \varepsilon + \sum_{j \in \mathbb{N}_n} \|\mu(A_j)\|_{\mathcal{Y}}$. We introduce the function $g = \sum_{j \in \mathbb{N}_n} \frac{\mu(A_j)}{\|\mu(A_j)\|_{\mathcal{Y}}} \chi_{A_j}$ which satisfies, for any $x \in \mathcal{Z}$, the bound $\|g(x)\|_{\mathcal{Y}} \leq 1$. Since $|\mu|$ is a regular measure on $\mathcal{Z}$, applying Lusin's theorem (Folland, 1999, p.217) to the function χ_{A_j}, there exists a real-valued continuous function $f_j \in \mathcal{C}(\mathcal{Z})$ such that $|f_j(x)| \leq 1$ for any $x \in \mathcal{Z}$ and $f_j = \chi_{A_j}$, except on a set E_j with $|\mu|(E_j) \leq \frac{\varepsilon}{(n+1)2^{j+1}}$. We now define a function $h : \mathcal{Y} \to \mathcal{Y}$ by setting $h(y) = y$, if $\|y\|_{\mathcal{Y}} \leq 1$ and $h(y) = \frac{y}{\|y\|_{\mathcal{Y}}}$, if $\|y\|_{\mathcal{Y}} \geq 1$, and introduce another function in $\mathcal{C}(\mathcal{Z}, \mathcal{Y})$ given by $\bar{f} := \sum_{j \in \mathbb{N}_n} \frac{\mu(A_j)}{\|\mu(A_j)\|_{\mathcal{Y}}} f_j$. Therefore, the function $f = h \circ \bar{f}$ is in $C(\mathcal{Z}, \mathcal{Y})$ as well, because $\bar{f} \in \mathcal{C}(\mathcal{Z}, \mathcal{Y})$ and, for any $y, y' \in \mathcal{Y}$, $\|h(y) - h(y')\|_{\mathcal{Y}} \leq 2\|y - y'\|_{\mathcal{Y}}$. Moreover, we observe, for any $x \in \left(\cup_{j \in \mathbb{N}_n} E_j\right)^c$, that $f(x) = g(x)$ and, for any $x \in \mathcal{Z}$, that $\|f(x)\|_{\mathcal{Y}} \leq 1$.

We are now ready to estimate the total variation of μ. First, observe that

$$\int_{\mathcal{Z}} \|f(x) - g(x)\|_{\mathcal{Y}} \, d|\mu|(x) \leq \sum_{j \in \mathbb{N}_n} (n+1)|\mu|(E_j) \leq \varepsilon,$$

and consequently we obtain the inequality

$$\begin{aligned} \|\mu\| &\leq \textstyle\sum_{j \in \mathbb{N}_n} \|\mu(A_j)\|_{\mathcal{Y}} + \varepsilon = \int_{\mathcal{Z}} (g(x), d\mu(x)) + \varepsilon \\ &\leq \Big| \int_{\mathcal{Z}} (f(x) - g(x), d\mu(x)) \Big| + \Big| \int_{\mathcal{Z}} (f(x), d\mu(x)) \Big| + \varepsilon \\ &\leq \int_{\mathcal{Z}} \|f(x) - g(x)\|_{\mathcal{Y}} d|\mu|(x) + \|L_\mu\| + \varepsilon \leq 2\varepsilon + \|L_\mu\|. \end{aligned}$$

This finishes the proof of the lemma. ■

We proceed to the proof of Lemma 10.

Proof of Lemma 10 For each $\mu \in \mathcal{M}(\mathcal{X}, \mathcal{Y})$, there exists an $L_\mu \in \mathcal{C}^*(\mathcal{Z}, \mathcal{Y})$ given by equation (21). The isometry of the map $\mu \mapsto L_\mu$ follows from Lemma 9.

Therefore, it suffices to prove, for every $\bar{L} \in \mathcal{C}^*(\mathcal{Z}, \mathcal{Y})$, that there is a $\mu \in \mathcal{M}(\mathcal{Z}, \mathcal{Y})$ such that $L_\mu = \bar{L}$. To this end, note that $\bar{L} \circ \iota^{-1} \in \mathcal{C}_\iota^*(\mathcal{Z} \times \mathcal{B}_1)$ since ι is an isometric map from $\mathcal{C}(\mathcal{Z}, \mathcal{Y})$ onto $\mathcal{C}_\iota(\mathcal{Z} \times \mathcal{B}_1)$ defined by equation (13). Since $\mathcal{C}_\iota(\mathcal{Z} \times \mathcal{B}_1)$ is a closed subspace of $\mathcal{C}(\mathcal{Z} \times \mathcal{B}_1)$, applying the *Hahn-Banach extension theorem* (see e.g. Folland, 1999, p.222) yields that, for any $L \in \mathcal{C}_\iota^*(\mathcal{Z} \times \mathcal{B}_1)$, there exists an extension functional $\tilde{L} \in \mathcal{C}^*(\mathcal{Z} \times \mathcal{B}_1)$ such that $\tilde{L}(F) = L \circ \iota^{-1}(F)$ for any $F \in \mathcal{C}_\iota(\mathcal{Z} \times \mathcal{B}_1)$. Moreover, recalling that $\mathcal{Z} \times \mathcal{B}_1$ is compact if $\mathcal{B}_1$ is equipped with the weak topology, by the Riesz representation theorem, for any $\tilde{L}$, there exists a scalar measure ν on $\mathcal{Z} \times \mathcal{B}_1$ such that

$$\tilde{L}(F) = \int_{\mathcal{Z} \times \mathcal{B}_1} F(x, y) d\nu(x, y), \quad \forall F \in \mathcal{C}(\mathcal{Z} \times \mathcal{B}_1).$$

Equivalently, for any $f \in \mathcal{C}(\mathcal{Z}, \mathcal{Y})$ there holds

$$\bar{L}f = \bar{L} \circ \iota^{-1}(F) = \int_{\mathcal{Z} \times \mathcal{B}_1} F(x, y) d\nu(x, y) = \int_{\mathcal{Z}} (f(x), d\mu(x)) = L_\mu f,$$

where μ is defined i terms of ν as in Equation (19).

This finishes the identification between functionals in $\mathcal{C}(\mathcal{Z}, \mathcal{Y})$ and vector measures with bounded variation. ■