arXiv:1603.05876v1 [math.OC] 18 Mar 2016

Generalized support vector regression: duality and tensor-kernel representation.

Saverio Salzo[1] and Johan A.K. Suykens[2]

[1]LCSL, Istituto Italiano di Tecnologia and Massachusetts Institute of Technology
Bldg. 46-5155, 77 Massachusetts Avenue, Cambridge, MA 02139, USA
Email: saverio.salzo@iit.it

[2]KU Leuven, ESAT-STADIUS
Kasteelpark Arenberg 10, B-3001 Leuven (Heverlee), Belgium
Email: johan.suykens@esat.kuleuven.be

Abstract

In this paper we study the variational problem associated to support vector regression in Banach function spaces. Using the Fenchel-Rockafellar duality theory, we give explicit formulation of the dual problem as well as of the related optimality conditions. Moreover, we provide a new computational framework for solving the problem which relies on a tensor-kernel representation. This analysis overcomes the typical difficulties connected to learning in Banach spaces. We finally present a large class of tensor-kernels to which our theory fully applies: power series tensor kernels. This type of kernels describe Banach spaces of analytic functions and include generalizations of the exponential and polynomial kernels as well as, in the complex case, generalizations of the Szegö and Bergman kernels.

Keywords: support vector regression, regularized empirical risk, reproducing kernel Banach spaces, tensors, Fenchel-Rockafellar duality.

1 Introduction

Support vector regression is a kernel-based estimation technique which allows to estimate a function belonging to an infinite dimensional function space based on a finite number of pointwise observations [7, 21, 23, 24]. The (primal) problem is classically formulated as an empirical risk minimization on a reproducing kernel Hilbert space of functions, the regularization term being the square of the Hilbert norm. This infinite dimensional optimization problem is approached through its dual problem which turns out to be finite dimensional, quadratic (possibly constrained), and involving the kernel function only, evaluated at the available data points [7, 20, 24]. Therefore, the knowledge of the kernel suffices to completely describe and

solve the dual problem as well as to compute the solution of the primal (infinite dimensional) problem. This is what it is known as the *kernel trick* and makes support vector regression effective and so popular in applications.

Learning in Banach spaces of functions is an emerging area of research which in principle permits to consider learning problems with more general types of norms than Hilbert norms [5, 10, 27]. The main motivation for this generalization comes from the need of finding more effective sparse representations of data or for feature selection. To that purpose, several types of alternative regularization schemes have been proposed in the literature, and we mention, among others, ℓ^1 regularization (lasso), elastic net, and bridge regression [8, 11]. Moreover, the statistical consistency of such more general regularization schemes have been addressed in [5, 6, 8, 15]. However, moving to Banach spaces of functions and Banach norms pose serious difficulties from the computational point of view [22]. Indeed, even though, in this more general setting, it is still possible to introduce appropriate reproducing kernels [27], they fail to properly represent the solution of the dual and primal problem, so that the dual approach becomes cumbersome. For this reason, the above mentioned estimation techniques are often implemented by directly tackling the primal problem and therefore, as a matter of fact, reduces to a finite dimensional estimation methods (that is to parametric models).

In this work we address support vector regression in Banach function spaces and we provide a new computational framework for solving the associated optimization problem, overcoming the difficulties we discussed above. Our model is described in the primal by means of an appropriate feature map in Banach spaces of features and a general regularizer. We first study, in great generality, the interplay between the primal and the dual problem through the Fenchel-Rockafellar duality. We obtain an explicit formulation of the dual problem, as well as of the related optimality conditions, in terms of the feature map and the subdifferentials of the loss function and of the regularizer. As a byproduct we also provide a general representer theorem.

Next, we consider the setting of a linear model described through a countable dictionary of functions with the regularization term being the ℓ^r-norm of the related coefficients, with $r = m/(m-1)$ and m an even integer. This choice allows $r > 1$ to be close to 1 and hence to approximate ℓ^1 regularization, possibly keeping the stability properties of the ℓ^2 regularization based estimation. Then we introduce a new type of kernel function which turns to be a symmetric positive-definte tensor of order m, and we prove that it allows to formulate the dual problem without any reference to the underlying feature map as well as to evaluate the optimal solution function at any point in the input space. In this way, the dual problem becomes a finite dimensional convex homogeneous m-degree-polynomial minimization problem which can be solved by standard smooth optimization algorithms, e.g., the conjugate gradient method. In the end, we show that the kernel trick can be fully extended to *tensor-kernels* and makes the dual approach in the Banach setting still viable for computing the solution of the primal (infinite dimensional) problem. Finally, we illustrate the theoretical framework above by presenting an entire class of tensor-kernel functions, that is *power series tensor-kernels*, which are extensions of the analogue matrix-type power series kernels considered in [29]. We show that this class includes kernels of exponential and polynomial type as well as, in the complex case, generalizations of the Szegö and Bergman kernels.

The rest of the paper is organized as follows. Section 2 gives basic definitions and facts. Section 3 presents the dual framework for SVR in general Banach spaces of features. In Section 4 we introduce tensor kernels and explain their role in making Banach space problems more practical numerically. Section 5 treats tensor kernels of power series type, which give rise to a general class of function Banach spaces to which the theory applies. Finally Section 6 contains conclusions.

2 Basic definitions and facts

Let $\mathcal{F}$ be a real Banach space. We denote by $\mathcal{F}^*$ its dual space and by $\langle \cdot, \cdot \rangle$ the canonical pairing between $\mathcal{F}$ and $\mathcal{F}^*$, meaning that, for every $(w, w^*) \in \mathcal{F} \times \mathcal{F}^*$, $\langle w, w^* \rangle = w^*(w)$. We denote by $\|\cdot\|$ the norm of $\mathcal{F}$ as well as the norm of $\mathcal{F}^*$. Let $F \colon \mathcal{F} \to]-\infty, +\infty]$. The *domain of* F is $\operatorname{dom} F = \{w \in \mathcal{F} \mid F(w) < +\infty\}$ and F is *proper* if $\operatorname{dom} F \neq \varnothing$. Suppose that F is proper and convex. The *subdifferential* of F is the set-valued operator $\partial F \colon \mathcal{F} \to 2^{\mathcal{F}^*}$ such that,

$$(\forall\, w \in \mathcal{F}) \quad \partial F(w) = \big\{ w^* \in \mathcal{F}^* \;\big|\; (\forall v \in \mathcal{F})\; F(w) + \langle v - w, w^* \rangle \leq F(v) \big\},$$

and its domain is $\operatorname{dom} \partial F = \{w \in \mathcal{F} \mid \partial F(w) \neq \varnothing\}$. The *Fenchel conjugate* of F is the function $F^* \colon \mathcal{F}^* \to]-\infty, +\infty] : w^* \in \mathcal{F}^* \mapsto \sup_{w \in \mathcal{F}} \langle w, w^* \rangle - F(w)$. We denote by $\Gamma_0(\mathcal{F})$ the set of proper, convex, and lower semicontinuous functions on $\mathcal{F}$. If $C \subset \mathcal{F}$, we denote by ι_C the *indicator function* of C, that is $\iota_C \colon \mathcal{F} \to]-\infty, +\infty]$, such that, for every $w \in \mathcal{F}$, $\iota_C(w) = 0$ if $w \in C$, and $\iota_C(w) = +\infty$ if $w \notin C$. Let $F \in \Gamma_0(\mathcal{F})$. Then the following duality relation between the subdifferentials of F and its conjugate F^* holds [26, Theorem 2.4.4(iv)]

$$(\forall\, (w, w^*) \in \mathcal{F} \times \mathcal{F}^*) \qquad w^* \in \partial F(w) \;\Leftrightarrow\; w \in \partial F^*(w^*). \tag{2.1}$$

Let $p \in [1, +\infty[$. The *conjugate* exponent of p is $p^* \in]1, +\infty]$ such that $1/p + 1/p^* = 1$. If $(\mathcal{Z}, \mathfrak{A}, \mu)$ is a finite measure space, we denote by $\langle \cdot, \cdot \rangle_{p,p^*}$ the canonical pairing between the Lebesgue spaces $L^p(\mu)$ and $L^{p^*}(\mu)$, i.e., $\langle f, g \rangle_{p,p^*} = \int_{\mathcal{Z}} f g \, \mathrm{d}\mu$. If $\mathbb{K}$ is a countable set, we define the sequence space

$$\ell^p(\mathbb{K}) = \left\{ (w_k)_{k \in \mathbb{K}} \in \mathbb{R}^{\mathbb{K}} \;\middle|\; \sum_{k \in \mathbb{K}} |w_k|^p < +\infty \right\}$$

endowed with the norm $\|w\|_p = \big(\sum_{k \in \mathbb{K}} |w_k|^p \big)^{1/p}$. The pairing between $\ell^p(\mathbb{K})$ and $\ell^{p^*}(\mathbb{K})$ is $\langle w, w^* \rangle_{p,p^*} = \sum_{k \in \mathbb{K}} w_k w_k^*$.

The Banach space $\mathcal{F}$ is called *smooth* [4] if, for every $w \in \mathcal{F}$ there exists a unique $w^* \in \mathcal{F}^*$ such that $\|w^*\| = 1$ and $\langle w, w^* \rangle = 1$. The smoothness property is equivalent to the Gâteaux differentiability of the norm on $\mathcal{F} \setminus \{0\}$. We say that $\mathcal{F}$ is *strictly convex* if, for every w and every v in $\mathcal{F}$ such that $\|w\| = \|v\| = 1$ and $w \neq v$, one has $\|(w + v)/2\| < 1$. Let $\mathcal{F}$ be a reflexive, strictly convex and smooth real Banach space and let $p \in]1, +\infty[$. Then the *p-duality map* of $\mathcal{F}$ is the mapping [4]

$$J_p \colon \mathcal{F} \to \mathcal{F}^* \text{ such that } (\forall\, w \in \mathcal{F}) \quad \langle w, J_p(w) \rangle = \|w\|^p \quad \text{and} \quad \|J_p(w)\| = \|w\|^{p-1}. \tag{2.2}$$

This map is a bijection from $\mathcal{F}$ onto $\mathcal{F}^*$ and its inverse is the p^*-duality map of $\mathcal{F}^*$. Moreover, for every $w \in \mathcal{F}$ and every $\lambda \in \mathbb{R}_+$, $J_p(\lambda w) = \lambda^{p-1} J_p(w)$ and $J_p(-w) = -J_p(w)$. The mapping J_2 is called the *normalized duality map* of $\mathcal{F}$. The Banach space $\ell^p(\mathbb{K})$ is reflexive, strictly convex, and smooth, and, it is immediate to verify from (2.2) that, its p-duality map is

$$J_p \colon \ell^p(\mathbb{K}) \to \ell^{p^*}(\mathbb{K}) \colon w = (w_k)_{k\in\mathbb{K}} \mapsto (|w_k|^{p-1} \operatorname{sign}(w_k))_{k\in\mathbb{K}}. \tag{2.3}$$

Moreover, $J_p^{-1} \colon \ell^{p^*}(\mathbb{K}) \to \ell^p(\mathbb{K})$ is the p^*-duality map of $\ell^{p^*}(\mathbb{K})$, hence it has the same form as (2.3) with p replaced by p^*.

Fact 2.1 ([1, Example 13.7]). *Let $\mathcal{F}$ be a reflexive, strictly convex, smooth, and real Banach space, let $p \in]1, +\infty[$, and let $\varphi \in \Gamma_0(\mathbb{R})$ be even. Then $(\varphi \circ \|\cdot\|)^* = \varphi^* \circ \|\cdot\|$ and*

$$(\forall\, w \in \mathcal{F}) \qquad \partial(\varphi \circ \|\cdot\|)(w) = \begin{cases} \dfrac{\partial\varphi(\|w\|)}{\|w\|^{p-1}} J_p(x) & \text{if } w \neq 0 \\ \{w^* \in \mathcal{F}^* \mid \|w^*\| \in \partial\varphi(0)\} & \text{if } w = 0. \end{cases}$$

Fact 2.2 (Fenchel-Rockafellar duality [26, Corollary 2.8.5 and Theorem 2.8.3(vi)]). *Let $\mathcal{F}$ and $\mathcal{B}$ be two real Banach spaces. Let $f \in \Gamma_0(\mathcal{F})$, let $g \in \Gamma_0(\mathcal{B})$, and let $B \colon \mathcal{F} \to \mathcal{B}$ be a bounded linear operator. Suppose that $0 \in \operatorname{int}\big(B(\operatorname{dom} f) - \operatorname{dom} g\big)$. Then the dual problem*

$$\min_{y^* \in \mathcal{B}} f^*(-B^* y^*) + g^*(y^*) \tag{2.4}$$

admits solutions and strong duality holds, that is

$$\inf_{x\in\mathcal{F}} f(x) + g(Bx) = -\min_{y^*\in\mathcal{B}} f^*(-B^*y^*) + g^*(y^*).$$

Moreover, if in addition $f + g \circ B$ admits a minimizer, then, for every $(\bar{x}, \bar{y}^*) \in \mathcal{F} \times \mathcal{B}^*$, $\bar{x}$ is a minimizer for $f + g \circ B$ and $\bar{y}^*$ is a solution of (2.4) iff $-B^*\bar{y}^* \in \partial f(\bar{x})$ and $\bar{y}^* \in \partial g(B\bar{x})$.

3 General SVR in Banach spaces of features.

We start by describing the problem setting. We consider the following optimization problem

$$\min_{(w,b)\in\mathcal{F}\times\mathbb{R}} \gamma \int_{\mathcal{X}\times\mathcal{Y}} L\big(y - \langle w, \Phi(x)\rangle - b\big)\, \mathrm{d}P(x,y) + G(w), \tag{3.1}$$

where the following assumptions are made:

A1 $\mathcal{X}$ and $\mathcal{Y}$ are two nonempty sets such that $\mathcal{Y} \subset \mathbb{R}$. P is a probability distribution on $\mathcal{X} \times \mathcal{Y}$, defined on some underlying σ-algebra $\mathfrak{A}$ on $\mathcal{X} \times \mathcal{Y}$. $\mathcal{F}$ is a real separable reflexive Banach space and $\Phi \colon \mathcal{X} \to \mathcal{F}^*$ is a measurable function. The function $L \colon \mathbb{R} \to \mathbb{R}_+$ is positive and convex, $p \in [1, +\infty[$, $\gamma \in \mathbb{R}_{++}$, and $G \colon \mathcal{F} \to]-\infty, +\infty]$ is proper, lower semicontinuous, and convex.

A2 $(\exists\,(a,b) \in \mathbb{R}_+^2)(\forall\, t \in \mathbb{R}) \quad L(t) \leq a + b|t|^p$.

A3 $\displaystyle\int_{\mathcal{X}\times\mathcal{Y}} |y|^p \,\mathrm{d}P(x,y) < +\infty \quad \text{and} \quad \int_{\mathcal{X}\times\mathcal{Y}} \|\Phi(x)\|^p \,\mathrm{d}P(x,y) < +\infty.$

In this context $\mathcal{F}$ and Φ are respectively the *feature space* and the *feature map*, and L is the *loss* function [5, 27].[1] Problem (3.1) can be considered as a continuous version of support vector regression, for general loss L and regularizer G. Indeed, if P is chosen as a discrete distribution, say $P = (1/n)\sum_{i=1}^n \delta_{(x_i,y_i)}$, for some sample $(x_i,y_i)_{1\le i\le n} \in (\mathcal{X}\times\mathcal{Y})^n$, then one obtains

$$\min_{(w,b)\in\mathcal{F}\times\mathbb{R}} \frac{\gamma}{n}\sum_{i=1}^n L\big(y_i - \langle w, \Phi(x_i)\rangle - b\big) + G(w),$$

which is the way support vector regression is formulated in [12]. Assumption **A2** corresponds to an upper growth condition for the loss L, whereas assumption **A3** includes a moment condition for the distribution P and an integrable condition for the feature map Φ, with respect to P — they are both standard assumptions in support vector machines [21]. In the following we consider the Lebesgue space

$$L^p(P) = \left\{ u\colon \mathcal{X}\times\mathcal{Y}\to\mathbb{R} \;\middle|\; u \text{ is } \mathfrak{A}\text{-measurable and } \int_{\mathcal{X}\times\mathcal{Y}} |u(x,y)|^p dP(x,y) < +\infty \right\}.$$

Problem (3.1) is a convex optimization problem of a composite form on an infinite dimensional space. The following result first recasts the problem in a constrained form, as done in [7, 23], then presents its dual problem, with respect to the Fenchel-Rockafellar duality, and the related optimality conditions.

Theorem 3.1. *Let assumptions* **A1**, **A2**, *and* **A3** *hold. Then problem* (3.1) *is equivalent to*

$$\left[\begin{array}{l} \displaystyle\min_{(w,b,e)\in\mathcal{F}\times\mathbb{R}\times L^p(P)} \gamma\int_{\mathcal{X}\times\mathcal{Y}} L(e(x,y))\,\mathrm{d}P(x,y) + G(w), \\ \text{subject to}\quad y - \langle w,\Phi(x)\rangle - b = e(x,y) \quad \text{for } P\text{-a.a. } (x,y)\in\mathcal{X}\times\mathcal{Y} \end{array}\right. \tag{$\mathcal{P}$}$$

and its dual is

$$\left[\begin{array}{l} \displaystyle\min_{u\in L^{p^*}(P)} G^*\left(\int_{\mathcal{X}\times\mathcal{Y}} u(x,y)\Phi(x)\,\mathrm{d}P(x,y)\right) \\ \qquad\qquad + \gamma\displaystyle\int_{\mathcal{X}\times\mathcal{Y}} L^*\left(\frac{u(x,y)}{\gamma}\right)\mathrm{d}P(x,y) - \int_{\mathcal{X}\times\mathcal{Y}} y\,u(x,y)\,\mathrm{d}P(x,y) \\ \text{subject to}\quad \displaystyle\int_{\mathcal{X}\times\mathcal{Y}} u\,\mathrm{d}P = 0. \end{array}\right. \tag{$\mathcal{D}$}$$

[1] Usually one requires that L is also even. In that case it is easy to see that necessarily 0 is a minimizer of L and that L is increasing on $\mathbb{R}_+$. Indeed for every $t\in\mathbb{R}_+$, we have $-t\le 0\le t$, and hence $0 = (1-\alpha)(-t)+\alpha t$, for some $\alpha\in[0,1]$. Then, by convexity $L(0)\le(1-\alpha)L(-t)+\alpha L(t) = L(t)$, for $L(-t) = L(t)$. Moreover, for every $s,t\in\mathbb{R}$, with $0\le s\le t$, we have $s = (1-\alpha)0+\alpha t$, for some $\alpha\in[0,1]$, and hence $L(s)\le(1-\alpha)L(0)+\alpha L(t)$ which yields $L(s)-L(0)\le\alpha(L(t)-L(0))\le L(t)-L(0)$.

Moreover, the dual problem $(\mathcal{D})$ *admits solutions, strong duality holds, and for every* $(w, b, e) \in \mathcal{F} \times \mathbb{R} \times L^p(P)$ *and every* $u \in L^{p^*}(P)$, (w, b, e) *is a solution of* $(\mathcal{P})$ *and* u *is a solution of* $(\mathcal{D})$ *if and only if the following optimality conditions hold*

$$\begin{cases} w \in \partial G^* \left(\displaystyle\int_{\mathcal{X}\times\mathcal{Y}} u(x,y)\Phi(x)\,\mathrm{d}P(x,y) \right) \\ \displaystyle\int_{\mathcal{X}\times\mathcal{Y}} u\,\mathrm{d}P = 0 \\ \dfrac{u(x,y)}{\gamma} \in \partial L(e(x,y)) \quad \textit{for } P\textit{-a.a. } (x,y) \in \mathcal{X}\times\mathcal{Y} \\ y - \langle w, \Phi(x)\rangle - b = e(x,y) \quad \textit{for } P\textit{-a.a. } (x,y) \in \mathcal{X}\times\mathcal{Y}. \end{cases} \tag{3.2}$$

Proof. The Banach spaces $L^p(P)$ and $L^{p^*}(P)$ are put in duality by means of the pairing

$$\langle \cdot, \cdot \rangle_{p,p^*} \colon L^p(P) \times L^{p^*}(P) \to \mathbb{R} \colon (e, u) \mapsto \int_{\mathcal{X}\times\mathcal{Y}} e(x,y)u(x,y)\,\mathrm{d}P(x,y). \tag{3.3}$$

In virtue of **A3**, the following linear operator

$$A \colon \mathcal{F}\times\mathbb{R} \to L^p(P) \text{ s.t. } (\forall\,(w,b) \in \mathcal{F}\times\mathbb{R})\ A(w,b) \colon \mathcal{X}\times\mathcal{Y} \to \mathbb{R} \colon (x,y) \mapsto \langle w, \Phi(x)\rangle + b \tag{3.4}$$

is well-defined and the function

$$\mathrm{pr}_2 \colon \mathcal{X}\times\mathcal{Y} \to \mathbb{R} \colon (x,y) \mapsto y,$$

is in $L^p(P)$. Then problem (3.1) can be written in the following constrained form

$$\begin{cases} \displaystyle\min_{(w,b,e)\in\mathcal{F}\times\mathbb{R}\times L^p(P)} \gamma \int_{\mathcal{X}\times\mathcal{Y}} L(e(x,y))\,\mathrm{d}P(x,y) + G(w), \\ \text{subject to } \mathrm{pr}_2 - A(w,b) = e \end{cases} \tag{3.5}$$

— where, in the constraint, the equality is meant to be in $L^p(P)$ — and hence $(\mathcal{P})$ follows. Now, define the following integral functional

$$R_P \colon L^p(P) \to \mathbb{R} \colon e \mapsto \int_{\mathcal{X}\times\mathcal{Y}} L(e(x,y))\,\mathrm{d}P(x,y),$$

the linear operator

$$B \colon \mathcal{F}\times\mathbb{R}\times L^p(P) \to L^p(P) \colon (w,b,e) \mapsto A(w,b) + e,$$

and the functional

$$f \colon \mathcal{F}\times\mathbb{R}\times L^p(P) \to \left]-\infty, +\infty\right] \colon (w,b,e) \mapsto \gamma R_P(e) + G(w).$$

We note that the functional R_P is well defined, convex, and continuous. This follows from from the convexity and continuity of L and from the fact that, because of **A1**, for every

$(x,y) \in \mathcal{X} \times \mathcal{Y}$, $L(e(x,y)) \leq a + b|e(x,y)|^p$. Then, problem (3.5) can be equivalently written as

$$\min_{(w,b,e)\in\mathcal{F}\times\mathbb{R}\times L^p(P)} f(w,b,e) + \iota_{\{-\mathrm{pr}_2\}}(-B(w,b,e)), \quad \text{with } f(w,b,e) = \gamma R(e) + G(w). \tag{3.6}$$

This form of problem (3.1) is amenable by the Fenchel-Rockafellar duality theory. In view of Fact 2.2 we need only to check that $0 \in \operatorname{int}\big(-B(\operatorname{dom} f) + \mathrm{pr}_2\big)$. This is almost immediate. Indeed, since $\operatorname{dom} f = \operatorname{dom} G \times \mathbb{R} \times L^p(P)$, we have

$$B(\operatorname{dom} f) = \big\{A(w,b) + e \,\big|\, (w,b) \in \operatorname{dom} G \times \mathbb{R} \text{ and } e \in L^p(P)\big\} = L^p(P).$$

Now we compute the dual of (3.6). We have

$$(\forall\, u \in L^{p^*}(P)) \qquad (\iota_{\{-\mathrm{pr}_2\}})^*(u) = \langle -\mathrm{pr}_2, u\rangle_{p,p^*} \tag{3.7}$$

and, for every $(w^*, b^*, u) \in \mathcal{F}^* \times \mathbb{R} \times L^{p^*}(P)$,

$$\begin{aligned} f^*(w^*,b^*,u) &= \sup_{(w,b,e)\in\mathcal{F}\times\mathbb{R}\times L^p(P)} \langle (w,b,e),(w^*,b^*,u)\rangle - f(w,b,e) \\ &= \sup_{w\in\mathcal{F}} \sup_{b\in\mathbb{R}} \sup_{e\in L^p(P)} \langle w, w^*\rangle - G(w) + \langle u, e\rangle_{p,p^*} - \gamma R_P(e) + bb^* \\ &= \begin{cases} G^*(w^*) + \gamma R_P^*(u/\gamma) & \text{if } b^* = 0 \\ +\infty & \text{if } b^* \neq 0. \end{cases} \end{aligned} \tag{3.8}$$

Moreover, we need also to compute $A^*\colon L^{p^*}(P) \to \mathcal{F}^* \times \mathbb{R}$ and $B^*\colon L^{p^*}(P) \to \mathcal{F}^* \times \mathbb{R} \times L^{p^*}(P)$. To that purpose, we note that for every $(w,b,e) \in \mathcal{F} \times \mathbb{R} \times L^p(P)$ and every $u \in L^{p^*}(P)$,

$$\begin{aligned} \langle B(w,b,e), u\rangle_{p,p^*} &= \langle A(w,b) + e, u\rangle_{p,p^*} = \langle (w,b), A^*u\rangle + \langle e, u\rangle_{p,p^*} \\ &= \langle (w,b,e), (A^*u, u)\rangle \end{aligned}$$

and

$$\begin{aligned} \langle (w,b), A^*u\rangle &= \langle A(w,b), u\rangle_{p,p^*} \\ &= \int_{\mathcal{X}\times\mathcal{Y}} (\langle w, \Phi(x)\rangle + b) u(x,y)\,\mathrm{d}P(x,y) \\ &= \Big\langle w, \int_{\mathcal{X}\times\mathcal{Y}} u(x,y)\Phi(x)\,\mathrm{d}P(x,y)\Big\rangle + b\int_{\mathcal{X}\times\mathcal{Y}} u\,\mathrm{d}P \\ &= \Big\langle (w,b), \Big(\int_{\mathcal{X}\times\mathcal{Y}} u(x,y)\Phi(x)\,\mathrm{d}P(x,y), \int_{\mathcal{X}\times\mathcal{Y}} u\,\mathrm{d}P\Big)\Big\rangle, \end{aligned}$$

which yields

$$A^*u = \Big(\int_{\mathcal{X}\times\mathcal{Y}} u\Phi\,\mathrm{d}P, \int_{\mathcal{X}\times\mathcal{Y}} u\,\mathrm{d}P\Big) \tag{3.9}$$

and

$$B^*u = (A^*u, u) = \Big(\int_{\mathcal{X}\times\mathcal{Y}} u\Phi\,\mathrm{d}P, \int_{\mathcal{X}\times\mathcal{Y}} u\,\mathrm{d}P,\, u\Big), \tag{3.10}$$

where, for brevity, we put $\int_{\mathcal{X}\times\mathcal{Y}} u\Phi \,\mathrm{d}P = \int_{\mathcal{X}\times\mathcal{Y}} u(x,y)\Phi(x)\,\mathrm{d}P(x,y)$. Thus, taking into account (3.8),(3.9), and (3.10), we have that, for every $u \in L^{p^*}(P)$,

$$f^*(B^*u) = f^*(A^*u, u) = \begin{cases} G^*\left(\int_{\mathcal{X}\times\mathcal{Y}} u\Phi \,\mathrm{d}P\right) + \gamma R_P^*(u/\gamma) & \text{if } \int_{\mathcal{X}\times\mathcal{Y}} u \,\mathrm{d}P = 0 \\ +\infty & \text{otherwise.} \end{cases}$$

Moreover, it follows from [18, Theorem 21(a)] that the Fenchel conjugate of R_P is still an integral operator, more precisely

$$(\forall\, u \in L^p(P)) \qquad R_P^*(u/\gamma) = \int_{\mathcal{X}\times\mathcal{Y}} L^*(u(x,y)/\gamma)\,\mathrm{d}P(x,y).$$

Therefore, recalling (3.7), the final form $(\mathcal{D})$ is obtained. The corresponding optimality conditions for problem (3.6) and its dual $(\mathcal{D})$ are (see Fact 2.1)

$$B^*u \in \partial f(w,b,e) = \partial G(w) \times \{0\} \times \gamma\partial R(e) \quad \text{and} \quad B(w,b,e) = \mathrm{pr}_2. \tag{3.11}$$

Now, recalling (3.10), conditions (3.11) can be gathered together as follows

$$\begin{cases} \displaystyle\int_{\mathcal{X}\times\mathcal{Y}} u\Phi \,\mathrm{d}P \in \partial G(w) \\ \displaystyle\int_{\mathcal{X}\times\mathcal{Y}} u \,\mathrm{d}P = 0 \\ \dfrac{u}{\gamma} \in \partial R(e) \\ y - \langle w, \Phi(x)\rangle - b = e(x,y) \quad \text{for } P\text{-a.a. } (x,y) \in \mathcal{X}\times\mathcal{Y}. \end{cases} \tag{3.12}$$

Thus, subdifferentiating under the integral sign [18, Theorem 21(c)] and recalling (2.1), (3.2) follows. □

Remark 3.2.

(*i*) The form $(\mathcal{P})$ resembles the way the problem of support vector machines for regression is often formulated [23, eq. (3.51)] and the optimality conditions (3.2) are the continuous versions of the one stated in [23, eq. (3.52)] for RKHS, differentiable loss functions, and square norm regularizers.

(*ii*) If $b = 0$, condition $\int_{\mathcal{X}\times\mathcal{Y}} u\,\mathrm{d}P = 0$ in (3.2) should be omitted.

(*iii*) If G is strictly convex on every convex subset of $\mathrm{dom}\,\partial G$ and $\mathrm{int}(\mathrm{dom}\,G^*) = \mathrm{dom}\,\partial G^*$, then G^* is Gâteaux differentiable (hence ∂G^* is single valued) on $\mathrm{dom}\,\partial G^*$ [1, Proposition 18.9] and, if a solution w of the primal problem $(\mathcal{P})$ exists, then the first of (3.2) yields

$$w = \nabla G^*\left(\int_{\mathcal{X}\times\mathcal{Y}} u\Phi \,\mathrm{d}P\right), \tag{3.13}$$

where u is any solution of the dual problem $(\mathcal{D})$. This constitutes a general nonlinear representer theorem, since the solution of problem $(\mathcal{P})$ is expressed in terms of the values of the feature map Φ. In the special case that $\mathcal{F}$ is a Hilbert space and $G = (1/2)\|\cdot\|^2$, $\nabla G^* = \mathrm{Id}$ and the first and third condition in (3.2) reduce to the ones obtained in [9, Corollary 3]. When P is the discrete distribution $P = (1/n)\sum_{i=1}^n \delta_{(x_i,y_i)}$, for some sample $(x_i, y_i)_{1\le i\le n} \in (\mathcal{X}\times\mathcal{Y})^n$, then (3.13) becomes

$$w = \nabla G^*\bigg(\sum_{i=1}^n u_i \Phi(x_i)\bigg). \tag{3.14}$$

We note that in (3.13)-(3.14) the nonlinearity relies on the mapping ∇G^* only.

The optimality conditions (3.2) in Theorem 3.1 directly yield a continuous representer theorem in Banach space setting.

Corollary 3.3 (Continuous representer theorem)**.** *Let assumptions* **A1**, **A2**, *and* **A3** *hold. Suppose that $\mathcal{F}$ is strictly convex and smooth and let $r \in]1,+\infty[$. In problem $(\mathcal{P})$, suppose that $G = \varphi\circ\|\cdot\|$, for some convex and even function $\varphi\colon \mathbb{R}\to\mathbb{R}_+$ such that* $\operatorname{argmin}\varphi = \{0\}$*. Then the solution w of problem $(\mathcal{P})$ admits the following representation*

$$J_r(w) = \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\Phi(x)\,\mathrm{d}P(x,y), \tag{3.15}$$

for some function $c \in L^{p^*}(P)$, where $J_r\colon \mathcal{F}\to\mathcal{F}^*$ is the r-duality map of $\mathcal{F}$.

Proof. Let $t > 0$. We first note that, since 0 is the unique minimizer of φ and $t > 0$, then $0 \notin \partial\varphi(t)$; moreover, for every $\xi \in \partial\varphi(t)$, we have $\xi t \ge \varphi(t) - \varphi(0) > 0$, hence, $\xi > 0$. Now, if $w = 0$, then (3.15) holds trivially. Suppose that $w \ne 0$. Then, it follows from Fact 2.1 that when $w \ne 0$,

$$\partial G(w) = \frac{\partial\varphi(\|w\|)}{\|w\|^{r-1}} J_r(w).$$

Therefore, it follows from the first of (3.12) that

$$\int_{\mathcal{X}\times\mathcal{Y}} u\Phi\,\mathrm{d}P = \frac{\xi}{\|w\|^{r-1}} J_r(w), \qquad \xi \in \partial\varphi(\|w\|).$$

Hence, since $\xi > 0$,

$$J_r(w) = \frac{\|w\|^{r-1}}{\xi}\int_{\mathcal{X}\times\mathcal{Y}} u\Phi\,\mathrm{d}P$$

and the statement follows. □

Remark 3.4. If in Corollary 3.3, $r = 2$ and P is a discrete measure, say $P = (1/n)\sum_{i=1}^n \delta_{(x_i,y_i)}$, for some sample $(x_i, y_i)_{1\le i\le n} \in (\mathcal{X}\times\mathcal{Y})^n$, then (3.15) becomes

$$J_2(w) = \sum_{i=1}^n c_i\Phi(x_i), \quad (c_i)_{1\le i\le n} \in \mathbb{R}^n, \tag{3.16}$$

where J_2 is the normalized duality map. Formula (3.16) is the way the representer theorem is usually presented in reproducing kernel Banach spaces [10, 27, 28]. Here it is a simple consequence of the more general Theorem 3.1 and Corollary 3.3. Moreover, we stress that our derivation of (3.16) relies on convex analysis arguments only, while in the above cited literature it is proved as a consequence of a representer theorem for function interpolation, ultimately using different techniques and stronger hypotheses. We finally note that, if $\mathcal{F}$ is a Hilbert space and $r = 2$, then J_2 is the identity map of $\mathcal{F}$ and (3.16) becomes

$$w = \sum_{i=1}^{n} c_i \Phi(x_i).$$

This is the classical representer theorem in Hilbert spaces [19].

Example 3.5. We consider the case of the Vapnik's ε-*insensitive loss* [20, 24]. Let $\varepsilon > 0$ and define

$$L_\varepsilon \colon \mathbb{R} \to \mathbb{R}_+ \colon t \mapsto \max\{0, |t| - \varepsilon\}. \tag{3.17}$$

This loss clearly satisfies **A2** for every $p > 1$. We note that (3.17) is the distance function from the set $[-\varepsilon, \varepsilon]$, that is, using the notation in [13], we have $L_\varepsilon = d_{[-\varepsilon,\varepsilon]}$. Then, the Fenchel conjugate of L_ε is (see [13, Example 13.24(i)])

$$L_\varepsilon^* = \sigma_{[-\varepsilon,\varepsilon]} + \iota_{[-1,1]} = \varepsilon|\cdot| + \iota_{[-1,1]}.$$

Therefore, for the loss (3.17), the dual problem $(\mathcal{D})$ becomes

$$\left[\begin{array}{l} \displaystyle\min_{u \in L^{p^*}(P)} G^*\left(\int_{\mathcal{X}\times\mathcal{Y}} u(x,y)\Phi(x)\,\mathrm{d}P(x,y)\right) \\ \qquad\qquad + \varepsilon \displaystyle\int_{\mathcal{X}\times\mathcal{Y}} |u(x,y)|\,\mathrm{d}P(x,y) - \int_{\mathcal{X}\times\mathcal{Y}} y\,u(x,y)\,\mathrm{d}P(x,y) \\ \text{subject to } \displaystyle\int_{\mathcal{X}\times\mathcal{Y}} u\,\mathrm{d}P = 0 \quad \text{and} \quad |u(x,y)| \le \gamma \text{ for } P\text{-a.a. } (x,y) \in \mathcal{X}\times\mathcal{Y}. \end{array}\right.$$

This is a generalization of the dual problem that arises in classical support vector regression when the linear ε-insensitive loss is considered [7, Proposition 6.21] and [20] — here we have a general regularizer and a Banach feature space.

Remark 3.6. Let us consider the case that $\mathcal{F}$ is a Hilbert space. Then $\mathcal{F}$ is isomorphic to its dual and the pairing reduces to the inner product in $\mathcal{F}$. Moreover, suppose that $G = (1/2)\|\cdot\|^2$, that $L = (1/2)|\cdot|^2$, and that $b = 0$, so that in (3.2) the condition $\int_{\mathcal{X}\times\mathcal{Y}} u\,\mathrm{d}P = 0$ is not present. Then it follows from the first and the third in (3.2) that

$$w = \int_{\mathcal{X}\times\mathcal{Y}} u\Phi\,\mathrm{d}P, \quad \frac{u}{\gamma} = e$$

and hence

$$\langle w, \Phi(x)\rangle = \int_{\mathcal{X}\times\mathcal{Y}} u(x',y')\langle\Phi(x'),\Phi(x)\rangle\,\mathrm{d}P(x',y').$$

Thus, the last of (3.2) yields the following integral equation

$$(\forall\,(x,y)\in\mathcal{X}\times\mathcal{Y})\qquad \frac{u(x,y)}{\gamma}+\int_{\mathcal{X}\times\mathcal{Y}}u(x',y')\langle\Phi(x'),\Phi(x)\rangle\,\mathrm{d}P(x',y')=y.$$

4 Tensor-kernel representation

We present our framework. For clarity we consider separately the real and complex case. We describe the real case with full details, whereas in the complex case we provide results with sketched proofs only.

4.1 The real case

Let $\mathcal{F}=\ell^r(\mathbb{K})$, with $\mathbb{K}$ a countable set and $r=m/(m-1)$ for some even integer $m\geq 2$. Thus, we have $r^*=m$. Let $(\phi_k)_{k\in\mathbb{K}}$ be a family of measurable functions from $\mathcal{X}$ to $\mathbb{R}$ such that, for every $x\in\mathcal{X}$, $(\phi_k(x))_{k\in\mathbb{K}}\in\ell^{r^*}(\mathbb{K})$ and define the feature map as

$$\Phi\colon\mathcal{X}\to\ell^{r^*}(\mathbb{K})\colon x\mapsto(\phi_k(x))_{k\in\mathbb{K}}. \tag{4.1}$$

Thus, we consider the following linear model

$$\big(\forall\,(w,b)\in\ell^r(\mathbb{K})\times\mathbb{R}\big)\qquad f_{w,b}=\langle w,\Phi(\cdot)\rangle_{r,r^*}+b=\sum_{k\in\mathbb{K}}w_k\phi_k+b\ \text{(pointwise)}, \tag{4.2}$$

where $\langle\cdot,\cdot\rangle_{r,r^*}$ is the canonical pairing between $\ell^r(\mathbb{K})$ and $\ell^{r^*}(\mathbb{K})$. The space

$$\mathcal{B}=\left\{f\colon\mathcal{X}\to\mathbb{R}\;\middle|\;(\exists(w,b)\in\ell^r(\mathbb{K})\times\mathbb{R})(\forall\,x\in\mathcal{X})\left(f(x)=\sum_{k\in\mathbb{K}}w_k\phi_k(x)+b\right)\right\} \tag{4.3}$$

is a *reproducing kernel Banach space* with norm

$$(\forall\,f\in\mathcal{B})\quad\|f\|_{\mathcal{B}}=\inf\left\{\|w\|_r+|b|\;\middle|\;(w,b)\in\ell^r(\mathbb{K})\times\mathbb{R}\text{ and }f=\sum_{k\in\mathbb{K}}w_k\phi_k+b\ \text{(pointwise)}\right\},$$

meaning that, with respect to that norm, the point-evaluation operators are continuous [5, 27]. We also consider the following regularization function

$$G(w)=\varphi(\|w\|_r), \tag{4.4}$$

for some convex and even function $\varphi\colon\mathbb{R}\to\mathbb{R}_+$, such that $\operatorname{argmin}\varphi=\{0\}$, and we set $P=(1/n)\sum_{i=1}^n\delta_{(x_i,y_i)}$, for some given sample $(x_i,y_i)_{1\leq i\leq n}\in(\mathcal{X}\times\mathcal{Y})^n$.

In such setting the primal and dual problems of support vector regression considered in Theorem 3.1 turn into

$$\left[\begin{array}{l}\displaystyle\min_{(w,b,e)\in\ell^r(\mathbb{K})\times\mathbb{R}\times\mathbb{R}^n}\frac{\gamma}{n}\sum_{i=1}^n L(e_i)+\varphi(\|w\|_r),\\[2ex]\text{subject to}\quad y_i-\langle w,\Phi(x_i)\rangle_{r,r^*}-b=e_i,\quad\text{for every }i\in\{1,\dots,n\}\end{array}\right. \tag{$\mathcal{P}_n$}$$

and, since $G^* = \varphi^* \circ \|\cdot\|_{r^*}$ (Fact 2.1),

$$\begin{cases} \min\limits_{u\in\mathbb{R}^n} \varphi^*\left(\left\|\dfrac{1}{n}\sum\limits_{i=1}^n u_i\Phi(x_i)\right\|_{r^*}\right) + \dfrac{\gamma}{n}\sum\limits_{i=1}^n L^*\left(\dfrac{u_i}{\gamma}\right) - \dfrac{1}{n}\sum\limits_{i=1}^n y_i u_i \\ \text{subject to } \sum\limits_{i=1}^n u_i = 0. \end{cases} \tag{$\mathcal{D}_n$}$$

Moreover, assuming that $w \neq 0$, Fact 2.1 and (3.2) yield the following optimality conditions[2]

$$\begin{cases} w \in \dfrac{\partial\varphi^*\left(\frac{1}{n}\left\|\sum_{i=1}^n u_i\Phi(x_i)\right\|_{r^*}\right)}{\left\|\sum_{i=1}^n u_i\Phi(x_i)\right\|_{r^*}^{r^*-1}} J_{r^*}\left(\sum\limits_{i=1}^n u_i\Phi(x_i)\right) \\ \sum\limits_{i=1}^n u_i = 0 \\ u_i/\gamma \in \partial L(e_i) \quad \text{for every } i \in \{1,\dots,n\} \\ y_i - \langle w, \Phi(x_i)\rangle_{r,r^*} - b = e_i \quad \text{for every } i \in \{1,\dots,n\}. \end{cases} \tag{4.5}$$

The dual problem $(\mathcal{D}_n)$ is a convex optimization problem and it is finite dimensional, since it is defined on $\mathbb{R}^n$. Once $(\mathcal{D}_n)$ is solved, expressions in (4.5) in principle allow to recover the primal solution (w, b) and eventually to compute the estimated regression function $\langle w, \Phi(x)\rangle + b$ at a generic point x of the input space $\mathcal{X}$. However, if $\mathbb{K}$ is an infinite set, that procedure is not feasible in practice, since it relies on the explicit knowledge of the feature map Φ, which is an infinite dimensional object. In the following we show that, in the dual problem $(\mathcal{D}_n)$, we can actually get rid of the feature map Φ and use instead a new type of kernel function evaluated at the sample points $(x_i)_{1\le i\le n}$. This will ultimately provide a new and effective computational framework for treating support vector regression in Banach spaces of type (4.3).

Remark 4.1. Consider the reproducing kernel Banach space

$$\mathcal{B} = \left\{ f\colon \mathcal{X} \to \mathbb{R} \;\middle|\; (\exists w \in \ell^r(\mathbb{K}))(\forall\, x \in \mathcal{X})\left(f(x) = \sum_{k\in\mathbb{K}} w_k\phi_k(x)\right)\right\}$$

endowed with norm $\|f\|_{\mathcal{B}} = \inf\left\{\|w\|_r \;\middle|\; w \in \ell^r(\mathbb{K}) \text{ and } f = \sum_{k\in\mathbb{K}} w_k\phi_k \text{ (pointwise)}\right\}$. Let $f \in \mathcal{B}$ and let $(w_k)_{k\in\mathbb{K}} \in \ell^r(\mathbb{K})$ be such that $f = \sum_{k\in\mathbb{K}} w_k\phi_k$ pointwise. Then, for every finite subset $\mathbb{J} \subset \mathbb{K}$ we have $f - \sum_{k\in\mathbb{J}} w_k\phi_k = \sum_{k\in\mathbb{K}\setminus\mathbb{J}} w_k\phi_k$ pointwise; hence, by definition

$$\left\|f - \sum_{k\in\mathbb{J}} w_k\phi_k\right\|_{\mathcal{B}} \le \left\|(w_k)_{k\in\mathbb{K}\setminus\mathbb{J}}\right\|_r = \left(\sum_{k\in\mathbb{K}\setminus\mathbb{J}} |w_k|^r\right)^{1/r} \to 0 \quad \text{as } |\mathbb{J}| \to +\infty.$$

[2]Note that $G^* = \varphi^* \circ \|\cdot\|_{r^*}$ and $\{0\} = \operatorname{argmin}\varphi = \partial\varphi^*(0)$. Thus, since, by (3.2), $w \in \partial G^*(\sum_{i=1}^n u_i\Phi(x_i))$, if $w \neq 0$, then Fact 2.1 yields $\sum_{i=1}^n u_i\Phi(x_i) \neq 0$.

Thus, the family $(w_k\phi_k)_{k\in\mathbb{K}}$ is summable in $(\mathcal{B}, \|\cdot\|_{\mathcal{B}})$ and it holds $f = \sum_{k\in\mathbb{K}} w_k\phi_k$ in $(\mathcal{B}, \|\cdot\|_{\mathcal{B}})$. Therefore, if the family of functions $(\phi_k)_{k\in\mathbb{K}}$ is pointwise ℓ^r-independent, in the sense that

$$(\forall\, (w_k)_{k\in\mathbb{K}} \in \ell^r(\mathbb{K})) \quad \sum_{k\in\mathbb{K}} w_k\phi_k = 0 \text{ (pointwise)} \;\Rightarrow (w_k)_{k\in\mathbb{K}} \equiv 0, \tag{4.6}$$

then $(\phi_k)_{k\in\mathbb{K}}$ is an unconditional Schauder basis of $\mathcal{B}$. Indeed if $\sum_{k\in\mathbb{K}} w_k\phi_k = 0$ in $(\mathcal{B}, \|\cdot\|_{\mathcal{B}})$, since the evaluation operators on $\mathcal{B}$ are continuous, we have $\sum_{k\in\mathbb{K}} w_k\phi_k = 0$ pointwise, and hence, by (4.6), $(w_k)_{k\in\mathbb{K}} \equiv 0$. We finally note that when $(\phi_k)_{k\in\mathbb{K}}$ is a (unconditional) Schauder basis of $\mathcal{B}$, then $\mathcal{B}$ is isometrically isomorphic to $\ell^r(\mathbb{K})$.

We start by first providing a generalized Cauchy-Schwartz inequality for sequences which is a consequence of a standard generalization of Hölder's inequality [2, Corollary 2.11.5] and that we prove for completeness. We use the following compact notation for the component-wise product of two sequences:

$$(\forall\, a \in \ell^r(\mathbb{K}))(\forall\, b \in \ell^{r^*}(\mathbb{K})) \qquad \sum_{k\in\mathbb{K}} ab := \sum_{k\in\mathbb{K}} a[k]b[k].$$

Proposition 4.2 (Generalized Cauchy-Schwartz inequality)**.** *Let $\mathbb{K}$ be a nonempty set. Let $m \in \mathbb{N}$ and let $a_1, a_2, \dots, a_m \in l^m_+(\mathbb{K})$. Then $a_1a_2\cdots a_m \in \ell^1_+(\mathbb{K})$ and*

$$\sum_{k\in\mathbb{K}} a_1a_2\cdots a_m \le \left(\sum_{k\in\mathbb{K}} a_1^m\right)^{1/m}\left(\sum_{k\in\mathbb{K}} a_2^m\right)^{1/m}\cdots\left(\sum_{k\in\mathbb{K}} a_m^m\right)^{1/m}.$$

Proof. We prove it by induction. The statement is true for $m = 2$. Suppose that the statement holds for $m \ge 2$ and let $a_1, a_2, \dots, a_m, a_{m+1} \in \ell^{m+1}_+(\mathbb{K})$. Then $a_1^{(m+1)/m}, a_2^{(m+1)/m}, \dots, a_m^{(m+1)/m} \in \ell^m_+(\mathbb{K})$ and by induction hypothesis $(a_1a_2\cdots a_m)^{(m+1)/m} \in \ell^1_+(\mathbb{K})$ and

$$\sum_{k\in\mathbb{K}} (a_1a_2\cdots a_m)^{(m+1)/m} \le \left(\sum_{k\in\mathbb{K}} a_1^{m+1}\right)^{1/m}\left(\sum_{k\in\mathbb{K}} a_2^{m+1}\right)^{1/m}\cdots\left(\sum_{k\in\mathbb{K}} a_m^{m+1}\right)^{1/m}.$$

Now, since $a_1a_2\cdots a_m \in \ell^{(m+1)/m}_+(\mathbb{K})$, $a_{m+1} \in \ell^{m+1}_+(\mathbb{K})$, and $(m+1)/m$ and $m+1$ are conjugate exponents, it follows from Hölder inequality that $a_1a_2\cdots a_ma_{m+1} \in \ell^1_+(\mathbb{K})$ and

$$\begin{aligned}\sum_{k\in\mathbb{K}} a_1a_2\cdots a_ma_{m+1} &\le \left(\sum_{k\in\mathbb{K}} (a_1a_2\cdots a_m)^{(m+1)/m}\right)^{m/(m+1)}\left(\sum_{k\in\mathbb{K}} a_{m+1}^{m+1}\right)^{1/(m+1)}\\ &\le \left(\sum_{k\in\mathbb{K}} a_1^{m+1}\right)^{1/m+1}\left(\sum_{k\in\mathbb{K}} a_2^{m+1}\right)^{1/m+1}\cdots\left(\sum_{k\in\mathbb{K}} a_{m+1}^{m+1}\right)^{1/m+1}.\end{aligned}$$

□

Now we are ready to define a tensor-kernel associated to the feature map (4.1) and give its main properties.

Proposition 4.3. *In the setting (4.1) described above, the following function is well-defined*

$$K\colon \mathcal{X}^m = \underbrace{\mathcal{X}\times\cdots\times\mathcal{X}}_{m\ times} \to \mathbb{R}\colon (x'_1,\dots,x'_m)\mapsto \sum_{k\in\mathbb{K}} \phi_k(x'_1)\cdots\phi_k(x'_m), \tag{4.7}$$

and the following hold.

(*i*) *For every* $(x'_1,\dots,x'_m)\in\mathcal{X}^m$, *and for every permutation* σ *of the indexes* $\{1,\dots,m\}$,

$$K(x'_{\sigma(1)}\dots x'_{\sigma(m)}) = K(x'_1,\dots x'_m).$$

(*ii*) *For every* $(x_i)_{1\le i\le n}\in\mathcal{X}^n$

$$(\forall\, u\in\mathbb{R}^n)\qquad \sum_{i_1,\dots,i_m=1}^{n} K(x_{i_1},\dots,x_{i_m})u_{i_1}\dots u_{i_m}\ge 0\,.$$

(*iii*) *For every* $(x_i)_{1\le i\le n}\in\mathcal{X}^n$

$$u\in\mathbb{R}^n\mapsto \left\|\sum_{i=1}^{n} u_i\Phi(x_i)\right\|_{r^*}^{r^*} = \sum_{i_1,\dots,i_m=1}^{n} K(x_{i_1},\dots,x_{i_m})u_{i_1}\dots u_{i_m} \tag{4.8}$$

is a homogeneous polynomial form of degree m *on* $\mathbb{R}^n$.

(*iv*) *For every* $x\in\mathcal{X}$, $K(x,\dots,x)\ge 0$.

(*v*) *For every* $(x'_1,\dots,x'_m)\in\mathcal{X}^m$

$$|K(x'_1,\dots,x'_m)|\le K(x'_1,\dots,x'_1)^{1/m}\cdots K(x'_m,\dots,x'_m)^{1/m}.$$

Proof. Since $(\phi_k(x'_1))_{k\in\mathbb{K}},(\phi_k(x'_2))_{k\in\mathbb{K}},\dots(\phi_k(x'_m))_{k\in\mathbb{K}}\in l^m(\mathbb{K})$, it follows from Proposition 4.2 that $(\phi_k(x'_1)\phi_k(x'_2)\cdots\phi_k(x'_m))_{k\in\mathbb{K}}\in l^1(\mathbb{K})$ and

$$\sum_{k\in\mathbb{K}}|\phi_k(x'_1)\cdots\phi_k(x'_m)|\le\left(\sum_{k\in\mathbb{K}}|\phi_k(x'_1)|^m\right)^{1/m}\cdots\left(\sum_{k\in\mathbb{K}}|\phi_k(x'_m)|^m\right)^{1/m}. \tag{4.9}$$

This shows that definition (4.7) is well-posed and moreover, since m is even we can remove the absolute values in the right hand side of (4.9) and get (*v*). Properties (*i*) and (*iv*) are immediate from the definition of K. Finally, since $r^*=m$ is even, for every $u\in\mathbb{R}^n$, we have

$$\begin{aligned}\left\|\sum_{i=1}^{n} u_i\Phi(x_i)\right\|_{r^*}^{r^*} &= \sum_{k\in\mathbb{K}}\left(\sum_{i=1}^{n} u_i\phi_k(x_i)\right)^m\\ &= \sum_{k\in\mathbb{K}}\sum_{i_1,\dots,i_m=1}^{n}\phi_k(x_{i_1})\cdots\phi_k(x_{i_m})u_{i_1}\dots u_{i_m}\\ &= \sum_{i_1,\dots,i_m=1}^{n}\left(\sum_{k\in\mathbb{K}}\phi_k(x_{i_1})\cdots\phi_k(x_{i_m})\right)u_{i_1}\dots u_{i_m}\,.\end{aligned} \tag{4.10}$$

Therefore, recalling the definition of K, (*ii*) and (*iii*) follow. ∎

Remark 4.4. Let $(x_i)_{1\le i\le n} \in \mathcal{X}^n$. Then $(K(x_{i_1},\dots,x_{i_m}))_{i\in\{1,\dots n\}^m}$ defines a tensor of degree m on $\mathbb{R}^n$. Then, properties (*i*) and (*ii*) establish that the tensor is symmetric and positive definite: they are natural generalization of the defining properties of standard positive (matrix) kernels.

Because of Proposition 4.3(*v*), tensor kernels, as defined in (4.7), can be normalized as for the matrix kernels.

Proposition 4.5 (normalized tensor kernel)**.** *Let K be defined as in (4.7) and suppose that, for every $x \in \mathcal{X}$, $K(x,\dots,x) > 0$. Define*

$$\begin{aligned} \tilde{K}\colon \mathcal{X}^m &\to \mathbb{R}, \\ (x'_1,\dots,x'_m) &\mapsto \frac{K(x'_1,\dots,x'_m)}{K(x'_1,\dots,x'_1)^{1/m}\cdots K(x'_m,\dots,x'_m)^{1/m}}. \end{aligned} \tag{4.11}$$

Then $\tilde{K}$ is still of type (4.7), for some family of functions $(\tilde{\phi}_k)_{k\in\mathbb{K}}$, $\tilde{\phi}_k\colon \mathcal{X}\to\mathbb{R}$, and the following hold.

(*i*) *For every $x\in\mathcal{X}$, $\tilde{K}(x,\dots,x)=1$.*

(*ii*) *For every $(x'_1,\dots x'_m)\in\mathcal{X}^m$, $|\tilde{K}(x'_1,\dots x'_m)|\le 1$.*

Proof. Just note that, for every $x\in\mathcal{X}$, $\|\Phi(x)\|_m^m = K(x,\cdots,x) > 0$. Then define $\tilde{\phi}_k(x) = \phi_k(x)/\|\Phi(x)\|_m^m$. □

We present the first main result of the section, which is a direct consequence of Proposition 4.3.

Theorem 4.6. *In the setting (4.1)-(4.4) described above, the dual problem $(\mathcal{D}_n)$ reduces to the following finite dimensional problem*

$$\left[\begin{array}{l} \displaystyle\min_{u\in\mathbb{R}^n} \varphi^*\left(\frac{1}{n}\left(\sum_{i_1,\dots,i_m=1}^n K(x_{i_1},\dots,x_{i_m})u_{i_1}\dots u_{i_m}\right)^{1/r^*}\right) + \frac{\gamma}{n}\sum_{i=1}^n L^*\left(\frac{u_i}{\gamma}\right) - \frac{1}{n}\sum_{i=1}^n y_i u_i \\ \\ \text{subject to } \displaystyle\sum_{i=1}^n u_i = 0. \end{array}\right. \tag{4.12}$$

Remark 4.7.

(*i*) Problem (4.12) is a convex optimization problem with linear constraints.

(*ii*) If the tensor kernel K is explicitly computable by means of (4.7), the dual problem (4.12) is a very finite dimensional problem, in the sense that it does not involve the feature map anymore. This is exactly how the kernel trick works within the kernel matrix.

Remark 4.8. The homogeneous polynomial form (4.8) can be written as follows

$$\sum_{\substack{\alpha\in\mathbb{N}^n\\|\alpha|=m}} \binom{m}{\alpha} K(\underbrace{x_1,\dots,x_1}_{\alpha_1},\dots,\dots,\underbrace{x_n,\dots,x_n}_{\alpha_n})u^\alpha \tag{4.13}$$

where, for every multi-index $\alpha=(\alpha_1,\dots,\alpha_n)\in\mathbb{N}^n$ and for every vector $u\in\mathbb{R}^n$, we used the standard notation $u^\alpha=u_1^{\alpha_1}\cdots u_n^{\alpha_n}$, $|\alpha|=\sum_{i=1}^n\alpha_i$, and the multinomial coefficient

$$\binom{m}{\alpha}=\binom{m}{\alpha_1,\dots,\alpha_n}=\frac{m!}{\alpha_1!\dots\alpha_n!}. \tag{4.14}$$

Indeed it follows from (4.10) and the multinomial theorem [3, Theorem 4.12] that

$$\begin{aligned}\Big\|\sum_{i=1}^n u_i\Phi(x_i)\Big\|_{r^*}^{r^*} &= \sum_{k\in\mathbb{K}}\Big(\sum_{i=1}^n u_i\phi_k(x_i)\Big)^m\\ &= \sum_{k\in\mathbb{K}}\sum_{\substack{\alpha\in\mathbb{N}^n\\|\alpha|=m}}\binom{m}{\alpha}\phi_k(x_1)^{\alpha_1}\dots\phi_k(x_n)^{\alpha_n}u^\alpha\\ &= \sum_{\substack{\alpha\in\mathbb{N}^n\\|\alpha|=m}}\binom{m}{\alpha}\Big(\sum_{k\in\mathbb{K}}\phi_k(x_1)^{\alpha_1}\dots\phi_k(x_n)^{\alpha_n}\Big)u^\alpha.\end{aligned}$$

Thus (4.13) follows from (4.7).

Corollary 4.9. *In Theorem 4.6, let $\varphi=(1/r)|\cdot|^r$ (which gives $G=(1/r)\|\cdot\|_r^r$). Then the dual problem (4.12) becomes*

$$\begin{cases}\displaystyle\min_{u\in\mathbb{R}^n}\frac{1}{r^*n^{r^*}}\sum_{i_1,\dots,i_m=1}^n K(x_{i_1},\dots,x_{i_m})u_{i_1}\dots u_{i_m}+\frac{\gamma}{n}\sum_{i=1}^n L^*\Big(\frac{u_i}{\gamma}\Big)-\frac{1}{n}\sum_{i=1}^n y_iu_i\\ \text{subject to }\displaystyle\sum_{i=1}^n u_i=0.\end{cases} \tag{4.15}$$

Proof. Just note that $\varphi^*=(1/r^*)|\cdot|^{r^*}$ and apply Theorem 4.6. □

Remark 4.10. The first term in the objective function in (4.15) is a positive definite homogeneous polynomial of order m. So, if the function L^* is smooth, which occurs when L is strictly convex, then the dual problem (4.15) is a smooth convex optimization problem with a linear constraint and can be approached by standard optimization techniques such as Newton-type or gradient-type methods — in the case of square loss, the dual problem (4.15) is a polynomial convex optimization problems and possibly more appropriate optimization methods may be

employed. We finally specialize (4.15) to the case of ε-insensitive loss (see Example 3.5)

$$\left[\begin{array}{l} \min\limits_{u\in\mathbb{R}^n} \dfrac{1}{mn^m} \displaystyle\sum_{i_1,\dots,i_m=1}^{n} K(x_{i_1},\dots,x_{i_m})u_{i_1}\dots u_{i_m} + \dfrac{\varepsilon}{n}\sum_{i=1}^{n}|u_i| - \dfrac{1}{n}\sum_{i=1}^{n} y_i u_i \\[2ex] \text{subject to } \displaystyle\sum_{i=1}^{n} u_i = 0 \quad \text{and} \quad |u_i| \le \gamma \text{ for every } i \in \{1,\dots,n\}. \end{array}\right. \tag{4.16}$$

This problem clearly shows similarities with the dual formulation of standard support vector regression [20, 24].

Once a solution $u \in \mathbb{R}^n$ of the dual problem (4.12) is computed, then one can compute the solution of the primal problem $(\mathcal{P}_n)$ by means of the equations in (4.5). In particular, if φ^* and L^* are differentiable, then the solution of the primal problem $(\mathcal{P}_n)$ is given by

$$w = \frac{(\varphi^*)'(\frac{1}{n}K[u]^{1/r^*})}{K[u]^{1/r}} J_{r^*}\Big(\sum_{i=1}^{n} u_i \Phi(x_i)\Big), \quad K[u] := \sum_{i_1,\dots,i_m=1}^{n} K(x_{i_1},\dots,x_{i_m})u_{i_1}\dots u_{i_m} > 0 \tag{4.17}$$

and

$$b = y_1 - \langle w, \Phi(x_1)\rangle_{r,r^*} - (L^*)'\Big(\frac{u_1}{\gamma}\Big), \tag{4.18}$$

where

$$J_{r^*}\colon \ell^{r^*}(\mathbb{K}) \to \ell^{r}(\mathbb{K})\colon u \mapsto \big(|u_k|^{r^*-1}\operatorname{sign}(u_k)\big)_{k\in\mathbb{N}}.$$

Now note that $r^* = m$ and $m-1$ is odd, therefore

$$J_m\colon \ell^{m}(\mathbb{K}) \to \ell^{r}(\mathbb{K})\colon u \mapsto (u_k^{m-1})_{k\in\mathbb{N}}$$

and hence (4.17) yields

$$(\forall\, k \in \mathbb{N}) \qquad w_k = \xi(u)\Big(\sum_{i=1}^{n} u_i \phi_k(x_i)\Big)^{m-1}, \qquad \xi(u) = \frac{(\varphi^*)'(\frac{1}{n}K[u]^{1/r^*})}{K[u]^{1/r}}. \tag{4.19}$$

Remark 4.11. It follows from the last two of (4.5) that in (4.18) any index $i \in \{1,\dots,n\}$ can be actually chosen to determine b. We chose $i = 1$.

The next issue is to evaluate the regression function corresponding to (w, b) at a general input point, without the explicit knowledge of the feature map but relying on the tensor-kernel K only. In the analogue case of matrix-kernels, this is what is usually called *kernel trick*. The following proposition shows that the kernel trick is still viable in our more general situation and that a tensor-kernel representation holds.

Proposition 4.12. *Under the assumptions (4.1)-(4.4), let K be defined as in (4.7). Suppose that φ^* is differentiable on $\mathbb{R}_{++}$ and that L^* is differentiable on $\mathbb{R}$. Let $u \in \mathbb{R}^n$ be a solution of the dual problem (4.12) and set (w, b) as in (4.19)-(4.18). Then, for every $x \in \mathcal{X}$,*

$$\begin{aligned}\langle w, \Phi(x)\rangle_{r,r^*} &= \frac{(\varphi^*)'(\frac{1}{n}K[u]^{1/r^*})}{K[u]^{1/r}} \sum_{i_1,\dots,i_{m-1}=1}^{n} K(x_{i_1},\dots,x_{i_{m-1}},x)u_{i_1}\cdots u_{i_{m-1}} \\ b &= y_1 - (L^*)'\left(\frac{u_1}{\gamma}\right) \\ &\quad - \frac{(\varphi^*)'(\frac{1}{n}K[u]^{1/r^*})}{K[u]^{1/r}} \sum_{i_1,\dots,i_{m-1}=1}^{n} K(x_{i_1},\dots,x_{i_{m-1}},x_1)u_{i_1}\cdots u_{i_{m-1}}.\end{aligned} \tag{4.20}$$

Proof. Let $x \in \mathcal{X}$. Then, we derive from (4.19) that

$$\begin{aligned}\langle w, \Phi(x)\rangle_{r,r^*} &= \sum_{k\in\mathbb{K}} w_k\phi_k(x) \\ &= \xi(u)\sum_{k\in\mathbb{K}}\left(\sum_{i=1}^{n} u_i\phi_k(x_i)\right)^{m-1}\phi_k(x) \\ &= \xi(u)\sum_{k\in\mathbb{K}}\sum_{i_1,\dots,i_{m-1}=1}^{n}\phi_k(x_{i_1})\cdots\phi_k(x_{i_{m-1}})\phi_k(x)u_{i_1}\cdots u_{i_{m-1}} \\ &= \xi(u)\sum_{i_1,\dots,i_{m-1}=1}^{n} K(x_{i_1},\dots,x_{i_{m-1}},x)u_{i_1}\cdots u_{i_{m-1}},\end{aligned}$$

where we used the definition (4.7) of K. □

Remark 4.13. In the case treated in Corollary 4.9, (4.20) yields the following representation formula

$$\begin{aligned}\langle w, \Phi(x)\rangle_{r,r^*} + b = \frac{1}{n^{m-1}} \sum_{i_1,\dots,i_{m-1}=1}^{n} \big(K(x_{i_1},\dots,x_{i_{m-1}},x) - K(x_{i_1},\dots,x_{i_{m-1}},x_1)\big)u_{i_1}\cdots u_{i_{m-1}} \\ + y_1 - (L^*)'\left(\frac{u_1}{\gamma}\right).\end{aligned}$$

Moreover, if in model (4.2) we assume no offset ($b = 0$), then we can avoid the requirement of the differentiability of L^* and the representation formula becomes

$$\langle w, \Phi(x)\rangle_{r,r^*} = \frac{1}{n^{m-1}} \sum_{i_1,\dots,i_{m-1}=1}^{n} K(x_{i_1},\dots,x_{i_{m-1}},x)u_{i_1}\cdots u_{i_{m-1}}.$$

Concluding we have shown that, the estimated regression function can be evaluated at every point of the input space by means of a finite summation formula, provided that the tensor-kernel K is explicitly available: we will show in Section 5 several significant examples in which this occurs.

4.2 The complex case

In this section we give the complex version of the theory developed in Section 4.1. Therefore, we let $\mathcal{F} = \ell^r(\mathbb{K};\mathbb{C})$, with $\mathbb{K}$ a countable set and $r = m/(m-1)$ for some even integer $m \geq 2$. Let $(\phi_k)_{k\in\mathbb{K}}$ be a family of measurable functions from $\mathcal{X}$ to $\mathbb{C}$ such that, for every $x \in \mathcal{X}$, $(\phi_k(x))_{k\in\mathbb{K}} \in \ell^{r^*}(\mathbb{K};\mathbb{C})$. The feature map is now defined as

$$\Phi\colon \mathcal{X} \to \ell^{r^*}(\mathbb{K};\mathbb{C})\colon x \mapsto (\overline{\phi_k(x)})_{k\in\mathbb{K}}, \tag{4.21}$$

which generates the model

$$(\forall\, w \in \ell^r(\mathbb{K};\mathbb{C}))(\forall\, b \in \mathbb{C}) \qquad x \mapsto \langle w, \Phi(x)\rangle_{r,r^*} + b = \sum_{k\in\mathbb{K}} w_k\phi_k(x) + b, \tag{4.22}$$

where $\langle w, w^*\rangle_{r,r^*} = \sum_{k\in\mathbb{N}} w_k\overline{w_k^*}$ is the canonical sesquilinear form between $\ell^r(\mathbb{K};\mathbb{C})$ and $\ell^{r^*}(\mathbb{K};\mathbb{C})$. This case can be treated as a vector-valued real case by identifying complex functions with $\mathbb{R}^2$-valued functions and the space $\ell^r(\mathbb{K};\mathbb{C})$ with $\ell^r(\mathbb{K};\mathbb{R}^2)$. Moreover, it is not difficult to generalize the dual framework presented in Section 3 to the case of vector-valued (and specifically to $\mathbb{R}^2$-valued) functions. Then, the (complex) feature map (4.21) defines an underlying real vector-valued feature map on $\ell^r(\mathbb{K};\mathbb{R}^2)$ [5], that is

$$\Phi_{\mathbb{R}}\colon \mathcal{X} \to \mathcal{L}(\mathbb{R}^2, \ell^{r^*}(\mathbb{K};\mathbb{R}^2)) \approxeq \ell^{r^*}(\mathbb{K};\mathbb{R}^{2\times 2})\colon x \mapsto (\phi_{\mathbb{R},k}(x))_{k\in\mathbb{K}}, \tag{4.23}$$

where $\mathcal{L}(\mathbb{R}^2, \ell^{r^*}(\mathbb{K};\mathbb{R}^2))$ is the spaces of linear continuous operators from $\mathbb{R}^2$ to $\ell^{r^*}(\mathbb{K};\mathbb{R}^2)$ (which is isomorphic to $\ell^{r^*}(\mathbb{K};\mathbb{R}^{2\times 2})$) and

$$(\forall\, x \in \mathcal{X})(\forall\, k \in \mathbb{K}) \quad \phi_{\mathbb{R},k}(x) = \begin{bmatrix} \mathfrak{Re}\phi_k(x) & \mathfrak{Im}\phi_k(x) \\ -\mathfrak{Im}\phi_k(x) & \mathfrak{Re}\phi_k(x) \end{bmatrix} \in \mathbb{R}^{2\times 2}. \tag{4.24}$$

This way, denoting, for every $x \in \mathcal{X}$, by $\phi_{\mathbb{R},k}(x)^*$ the transpose of the matrix $\phi_{\mathbb{R},k}(x)$, we have

$$(\forall\, x \in \mathcal{X})(\forall\, k \in \mathbb{K})(\forall\, w_k \in \mathbb{R}^2 \approxeq \mathbb{C}) \qquad \phi_{\mathbb{R},k}(x)^* w_k = w_k\phi_k(x), \tag{4.25}$$

hence $\Phi_{\mathbb{R}}(x)^* w = \langle w, \Phi(x)\rangle_{r,r^*}$. Moreover

$$(\forall\, x \in \mathcal{X})(\forall\, u \in \mathbb{R}^2 \approxeq \mathbb{C}) \qquad \Phi_{\mathbb{R}}(x)u = (\phi_{\mathbb{R},k}(x)u)_{k\in\mathbb{K}} = (u\overline{\phi_k(x)})_{k\in\mathbb{K}} = u\Phi(x). \tag{4.26}$$

Then, problems $(\mathcal{P}_n)$ and $(\mathcal{D}_n)$ become

$$\left[\begin{array}{l} \displaystyle\min_{(w,b,e)\in\ell^r(\mathbb{K};\mathbb{C})\times\mathbb{C}\times\mathbb{C}^n} \frac{\gamma}{n}\sum_{i=1}^n L(e_i) + \varphi(\|w\|_r), \\ \text{subject to} \quad y_i - \langle w, \Phi(x_i)\rangle_{r,r^*} - b = e_i, \quad \text{for every } i \in \{1,\dots,n\} \end{array}\right. \tag{$\mathcal{P}_n(\mathbb{C})$}$$

and

$$\left[\begin{array}{l} \displaystyle\min_{u\in\mathbb{C}^n} \varphi^*\left(\left\|\frac{1}{n}\sum_{i=1}^n u_i\Phi(x_i)\right\|_{r^*}\right) + \frac{\gamma}{n}\sum_{i=1}^n L^*\left(\frac{u_i}{\gamma}\right) - \frac{1}{n}\sum_{i=1}^n \mathfrak{Re}(u_i\overline{y_i}) \\ \text{subject to} \quad \displaystyle\sum_{i=1}^n u_i = 0, \end{array}\right. \tag{$\mathcal{D}_n(\mathbb{C})$}$$

where, $L^*\colon \mathbb{C} \to \mathbb{R}\colon z^* \mapsto \sup_{z\in\mathbb{C}} \mathfrak{Re}(z\overline{z^*}) - L(z)$. Moreover, assuming that $w \neq 0$, the optimality conditions (4.5) still hold, where now $J_{r^*}\colon \ell^{r^*}(\mathbb{K};\mathbb{C}) \to \ell^r(\mathbb{K};\mathbb{C})\colon w^* \mapsto (|w_k^*|^{r-1} w_k^*/|w_k^*|)_{k\in\mathbb{K}}$, and

$$(\forall\, e \in \mathbb{C}) \quad \partial L(e) = \big\{ z^* \in \mathbb{C} \;\big|\; (\forall\, z \in \mathbb{C})\ L(z) \geq L(e) + \mathfrak{Re}\big(\overline{z^*}(z-e)\big) \big\}.$$

In the following we give the result corresponding to Proposition 4.3.

Proposition 4.14. *In the setting described above, suppose that m is even and set $q = m/2$. Then, the following function is well-defined*

$$K\colon \mathcal{X}^q \times \mathcal{X}^q \to \mathbb{C}\colon (x'_1,\dots,x'_q;x''_1,\dots,x''_q) \mapsto \sum_{k\in\mathbb{K}} \phi_k(x'_1)\cdots\phi_k(x'_q)\overline{\phi_k(x''_1)}\cdots\overline{\phi_k(x''_q)}, \tag{4.27}$$

and the following hold.

(i) For every $(x'_1,\dots,x'_q;x''_1,\dots,x''_q) \in \mathcal{X}^q \times \mathcal{X}^q$, and for every permutation σ' and σ'' of the indexes $\{1,\dots,q\}$,

$$K(x'_{\sigma'(1)}\dots x'_{\sigma'(q)};x''_{\sigma''(1)}\dots x''_{\sigma''(q)}) = K(x'_1,\dots x'_q;x''_1\dots x''_q).$$

(ii) For every $(x';x'') \in \mathcal{X}^q \times \mathcal{X}^q$ $K(x';x'') = \overline{K(x'';x')}$;

(iii) For every $(x_i)_{1\leq i\leq n} \in \mathcal{X}^n$

$$(\forall\, u \in \mathbb{C}^n) \quad \sum_{\substack{i_1,\dots,i_q=1\\ j_1,\dots,j_q=1}}^{n} K(x_{j_1},\dots,x_{j_q};x_{i_1},\dots,x_{i_q}) u_{i_1}\dots u_{i_q}\overline{u_{j_1}}\dots\overline{u_{j_q}} \geq 0\,.$$

(iv) For every $(x_i)_{\ \leq i\leq n} \in \mathcal{X}^n$

$$u \in \mathbb{C}^n \mapsto \Big\| \sum_{i=1}^{n} u_i \Phi(x_i) \Big\|_{r^*}^{r^*} = \sum_{\substack{i_1,\dots,i_q=1\\ j_1,\dots,j_q=1}}^{n} K(x_{j_1},\dots,x_{j_q};x_{i_1},\dots,x_{i_q}) u_{i_1}\dots u_{i_q}\overline{u_{j_1}}\dots\overline{u_{j_q}}$$

is a positive homogeneous polynomial form of degree m on $\mathbb{C}^n$.

(v) For every $(x'_1,\dots,x'_q) \in \mathcal{X}^q$, $K(x'_1,\dots,x'_q;x'_1,\dots,x'_q) \geq 0$;

(vi) For every $(x'_1,\dots,x'_q;x''_1,\dots,x''_q) \in \mathcal{X}^q \times \mathcal{X}^q$,

$$|K(x'_1,\dots,x'_q;x''_1,\dots,x''_q)| \leq K(x'_1,\dots,x'_1;x'_1,\dots,x'_1)^{1/m}\cdots K(x''_q,\dots,x''_q;x''_q,\dots,x''_q)^{1/m}.$$

Remark 4.15. Item *(iii)* states that $\big(K(x_{i_1},\dots,x_{i_m})\big)_{i\in\{1,\dots,n\}^m}$ is a positive-definite tensor of degree m.

As in the real case, the dual problem $(\mathcal{D}_n)$ reduces to

$$\begin{cases} \min\limits_{u\in\mathbb{C}^n} \varphi^*\left(\dfrac{1}{n}\left(\sum\limits_{\substack{i_1,\dots,i_q=1\\ j_1,\dots,j_q=1}}^{n} K(x_{j_1},\dots,x_{j_q},x_{i_1},\dots,x_{i_q})u_{i_1}\dots u_{i_q}\overline{u_{j_1}}\dots\overline{u_{j_q}}\right)^{1/r^*}\right) \\ \qquad\qquad + \dfrac{\gamma}{n}\sum\limits_{i=1}^{n} L^*\left(\dfrac{u_i}{\gamma}\right) - \dfrac{1}{n}\mathfrak{Re}\sum\limits_{i=1}^{n}\overline{y_i}u_i \\ \text{subject to } \sum\limits_{i=1}^{n} u_i = 0 \end{cases}$$

and the homogeneous polynomial form in Proposition 4.14(*iv*) can be written as follows

$$\sum_{\substack{\alpha\in\mathbb{N}^n,\beta\in\mathbb{N}^n\\ |\alpha|=q,|\beta|=q}} \binom{q}{\alpha}\binom{q}{\beta} K(\underbrace{x_1,\dots x_1}_{\alpha_1},\dots\dots,\underbrace{x_n,\dots,x_n}_{\alpha_n},\underbrace{x_1,\dots x_1}_{\beta_1},\dots\dots,\underbrace{x_n,\dots,x_n}_{\beta_n})u^\alpha\,\overline{u}^\beta. \quad (4.28)$$

Finally, in the setting of Proposition 4.12, defining

$$K[u] = \sum_{\substack{i_1,\dots,i_q=1\\ j_1,\dots,j_q=1}}^{n} K(x_{j_1},\dots,x_{j_q},x_{i_1},\dots,x_{i_q})u_{i_1}\dots u_{i_q}\overline{u_{j_1}}\dots\overline{u_{j_q}}, \quad (4.29)$$

for every $x\in\mathcal{X}$, the following representation formulas hold

$$\begin{aligned} \langle w,\Phi(x)\rangle_{r,r^*} &= \frac{(\varphi^*)'(\frac{1}{n}K[u]^{1/r^*})}{K[u]^{1/r}} \sum_{\substack{i_1,\dots,i_q=1\\ j_1,\dots,j_{q-1}=1}}^{n} K(x_{j_1},\dots x_{j_{q-1}},x;x_{i_1},\dots x_{i_q})u_{i_1}\cdots u_{i_q}\overline{u_{j_1}}\cdots\overline{u_{j_{q-1}}} \\ b &= y_1 - \langle w,\Phi(x_1)\rangle_{r,r^*} - \nabla L^*\left(\frac{u_1}{\gamma}\right). \end{aligned} \quad (4.30)$$

where ∇L^* is the (real) gradient of L^*, considered as a function from $\mathbb{R}^2$ to $\mathbb{R}$.

Remark 4.16. In view of Proposition 4.14(*iv*), definitions (4.21) and (4.27) correspond to those given in [21, Lemma 4.2] and the concept of positive definiteness stated in (*iii*) is a natural generalization of the analogue notion given in [21, Definition 4.15].

5 Power series tensor-kernels

In this section we consider reproducing kernel Banach spaces of complex analytic functions which are generated through power series. We show that, for such spaces, the corresponding tensor kernel, defined according to (4.7), admits an explicit expression. We provide also

representation formulas. In this section we assume, for simplicity, that $\varphi = (1/r)|\cdot|^r$, therefore we address the support vector regression problem

$$\min_{(w,b)\in\ell^r(\mathbb{K};\mathbb{C})\times\mathbb{C}} \frac{\gamma}{n}\sum_{i=1}^{n} L\big(y_i - \langle w, \Phi(x_i)\rangle_{r,r^*} - b\big) + \|w\|_r^r,$$

for a specific choice of the feature map (4.21).

We first need to set special notation for multi-index powers of complex vectors. Let $d \in \mathbb{N}$ with $d \geq 1$. We will denote the component of a vector $x \in \mathbb{C}^d$, by x_t, with $t \in \{1, \dots, d\}$. For every $x \in \mathbb{C}^d$ and every $\nu \in \mathbb{N}^d$ we set

$$x^\nu = \prod_{t=1}^{d} x_t^{\nu_t}, \quad |x| = (|x_1|, \dots, |x_d|), \quad \text{and} \quad \nu! = \prod_{t=1}^{d} \nu_t!$$

so that $\forall\, \nu \in \mathbb{N}^d$ we have $|x^\nu| = \prod_{t=1}^{d} |x_t|^{\nu_t} = |x|^\nu$. Moreover, when the exponent of the vector $x \in \mathbb{C}^d$ is an index (not a multi-index), say $m \in \mathbb{N}$, we consider m as a constant multi-index, that is $(m, \dots, m)$, so that x^m means $\prod_{t=1}^{d} x_t^m$. Finally, we define the binary inner operation of pointwise multiplication in $\mathbb{C}^d$. For every $x, x' \in \mathbb{C}^d$, we set $x \star x' \in \mathbb{C}^d$, such that, for every $t \in \{1, \dots, d\}$, $(x \star x')_t = x_t x'_t$. Let $m \in \mathbb{N}$ and $x \in \mathbb{C}^d$. We set $x^{\star m} = x \star \cdots \star x$ (m-times), so that $x^{\star m} \in \mathbb{C}^d$ and, for every $t \in \{1, \dots, d\}$, $(x^{\star m})_t = x_t^m$.

Let $\rho = (\rho_\nu)_{\nu\in\mathbb{N}^d}$ be a multi-sequence in $\mathbb{R}_+$, let $r = m/(m-1)$ for some even integer $m \geq 2$. Let $\mathcal{D}_\rho$ be the domain of (absolute) convergence of the power series $\sum_{\nu\in\mathbb{N}^d} \rho_\nu z^\nu$, that is the interior of the set $\big\{z \in \mathbb{C}^d \;\big|\; \sum_{\nu\in\mathbb{N}^d} \rho_\nu |z^\nu| < +\infty\big\}$. The set $\mathcal{D}_\rho$ is a complete Reinhardt domain[3] and we assume that $\mathcal{D}_\rho \neq \{0\}$. Let $\kappa\colon \mathcal{D}_\rho \to \mathbb{C}$ be the sum of the series $\sum_{\nu\in\mathbb{N}^d} \rho_\nu z^\nu$, that is

$$(\forall\, z \in \mathcal{D}_\rho) \quad \kappa(z) = \sum_{\nu\in\mathbb{N}^d} \rho_\nu z^\nu.$$

Clearly κ is an analytic function on $\mathcal{D}_\rho$. Set

$$\mathcal{D}_\rho^{\star 1/m} = \big\{x \in \mathbb{C}^d \;\big|\; x^{\star m} = (x_1^m, \dots, x_d^m) \in \mathcal{D}_\rho\big\},$$

let $\mathcal{X} \subset \mathcal{D}_\rho^{\star 1/m}$, and define the dictionary

$$(\forall\, \nu \in \mathbb{N}^d) \qquad \phi_\nu\colon \mathcal{X} \to \mathbb{C}\colon x \mapsto \rho_\nu^{1/m} x^\nu. \tag{5.1}$$

Then, for every $x \in \mathcal{X}$, since $x^{\star m} \in \mathcal{D}_\rho$, we have

$$\sum_{\nu\in\mathbb{N}^d} |\phi_\nu(x)|^m = \sum_{\nu\in\mathbb{N}^d} \rho_\nu |x^{\star m}|^\nu < +\infty,$$

hence $(\phi_\nu(x))_{\nu\in\mathbb{N}^d} \in \ell^m(\mathbb{N}^d;\mathbb{C})$. Thus, we are in the framework described at the beginning of Section 4.2. We define

$$B^r_{\rho,b}(\mathcal{X}) = \left\{ f \in \mathbb{C}^{\mathcal{X}} \;\middle|\; (\exists\, (c_\nu)_{\nu\in\mathbb{N}^d} \in \ell^r(\mathbb{N}^d;\mathbb{C}))(\exists\, b \in \mathbb{C})(\forall\, x \in \mathcal{X})\Big(f(x) = \sum_{\nu\in\mathbb{N}^d} c_\nu \phi_\nu(x) + b\Big) \right\},$$

[3] It means that if $z \in \mathcal{D}_\rho$, then $\mathcal{D}_\rho$ contains the polydisk $\{t \in \mathbb{C}^d \mid (\forall\, j \in \{1, \dots, d\})\ |t_j| \leq |z_j|\}$.

which is a reproducing kernel Banach spaces with norm

$$\|f\|_{B^r_{\rho,b}(\mathcal{X})} = \inf\Big\{\|c\|_r + |b| \;\Big|\; (c_\nu)_{\nu\in\mathbb{N}^d} \in \ell^r(\mathbb{N}^d;\mathbb{C}) \text{ and } f = \sum_{\nu\in\mathbb{N}^d} c_\nu \rho_\nu^{1/m} x^\nu + b \text{ (pointwise)}\Big\}.$$

Suppose now that $b = 0$ and that, for every $\nu \in \mathbb{N}^d$, $\rho_\nu > 0$. Then, defining the weights $(\eta_\nu)_{\nu\in\mathbb{N}^d} = (\rho_\nu^{-r/m})_{\nu\in\mathbb{N}^d}$ and the corresponding weighted ℓ^r space

$$\ell^r_\eta(\mathbb{N}^d;\mathbb{C}) = \Big\{(a_\nu)_{\nu\in\mathbb{N}^d} \in \mathbb{C}^{\mathbb{N}^d} \;\Big|\; \sum_{\nu\in\mathbb{N}^d} \frac{1}{\rho_\nu^{r/m}}|a_\nu|^r < +\infty\Big\},$$

we can express the space $B^r_{\rho,0}(\mathcal{X})$ in the form of a weighted Hardy-like space [17, 25]

$$B^r_{\rho,0}(\mathcal{X}) = \Big\{f \in \mathbb{C}^{\mathcal{X}} \;\Big|\; (\exists\,(a_\nu)_{\nu\in\mathbb{N}^d} \in \ell^r_\eta(\mathbb{N}^d;\mathbb{C}))(\forall\, x \in \mathcal{X})\Big(f(x) = \sum_{\nu\in\mathbb{N}^d} a_\nu x^\nu\Big)\Big\}.$$

Moreover, for every $(x'_1,\ldots,x'_q,x''_1,\ldots,x''_q) \in \mathcal{X}^q \times \mathcal{X}^q$,

$$K(x'_1,\ldots,x'_q;x''_1,\ldots,x''_q) = \sum_{\nu\in\mathbb{N}^d} \rho_\nu x_1'^{\nu}\cdots x_q'^{\nu}\overline{x_1''^{\nu}}\cdots\overline{x_q''^{\nu}} = \kappa(x'_1\star\cdots\star x'_q\star\overline{x''_1}\star\cdots\star\overline{x''_q}). \quad (5.2)$$

Remark 5.1. Suppose that $\rho_\nu > 0$, for every $\nu \in \mathbb{N}^d$. Then $\sum_{\nu\in\mathbb{N}^d} c_\nu\rho_\nu^{1/m}x^\nu = 0$ (pointwise) implies $c_\nu\rho_\nu^{1/m} = 0$, for every $\nu \in \mathbb{N}^d$ and hence $c_\nu = 0$, for every $\nu \in \mathbb{N}^d$. Thus, in virtue of Remark 4.1 this yields that $(\phi_\nu)_{\nu\in\mathbb{N}^d}$ is an unconditional Schauder basis of $B^r_{\rho,0}(\mathcal{X})$ and that $B^r_{\rho,0}(\mathcal{X})$ is isometric to $\ell^r(\mathbb{N}^d;\mathbb{C})$.

Proposition 5.2. *Under the notation and assumption above, suppose that $\mathcal{X}$ is a compact subset of $D_\rho^{\star 1/m}$ and that, for every $\nu \in \mathbb{N}^d$, $\rho_\nu > 0$. Then $B^r_{\rho,b}(\mathcal{X})$ is dense in $\mathscr{C}(\mathcal{X};\mathbb{C})$, the space of continuous functions on $\mathcal{X}$ endowed with the uniform norm.*

Proof. It is enough to note that $B^r_{\rho,b}(\mathcal{X})$ contains the set

$$\mathcal{A} = \operatorname{span}\{\phi_\nu \mid \nu \in \mathbb{N}\} = \Big\{\sum_{\nu\in I} c_\nu x^\nu \;\Big|\; I \subset \mathbb{N}^d \text{ and } I \text{ finite } (c_\nu)_{\nu\in I} \in \mathbb{C}^I\Big\}$$

which is the algebra of polynomials on $\mathcal{X}$ in d variables with complex coefficients. Thus the statement is a consequence of the Stone-Weierstrass theorem. □

In the sequence we also assume that the offset b is zero. Because of (5.2), the representation given in (4.28) yields the following homogenous polynomial form

$$u \in \mathbb{C}^n \mapsto \Big\|\sum_{i=1}^n u_i\Phi(x_i)\Big\|_{r^*}^{r^*} = \sum_{\substack{\alpha\in\mathbb{N}^n,\beta\in\mathbb{N}^n\\ |\alpha|=q,|\beta|=q}} \binom{q}{\alpha}\binom{q}{\beta}\kappa(x_1^{\star\beta_1}\star\cdots\star x_n^{\star\beta_n}\star\overline{x}_1^{\star\alpha_1}\star\cdots\star\overline{x}_n^{\star\alpha_n})u^\alpha\overline{u}^\beta, \quad (5.3)$$

where $(x_i)_{1\le i\le n}\in\mathcal{X}^n$ is the training set and, according to the convention established at the beginning of the section, $x_i^{\star\alpha_i}=(x_{i,1}^{\alpha_i},\dots,x_{i,d}^{\alpha_i})$. Moreover, in this case, recalling (4.30) and (5.2), for every $x\in\mathcal{X}$, we have

$$\langle w,\Phi(x)\rangle_{r,r^*}=\frac{1}{n^{m-1}}\sum_{\substack{i_1,\dots,i_q=1\\ j_1,\dots,j_{q-1}=1}}^{n}\kappa(x_{j_1}\star\cdots\star x_{j_{q-1}}\star x\star\overline{x_{i_1}}\star\cdots\overline{\star x_{i_q}})u_{i_1}\cdots u_{i_q}\overline{u_{j_1}}\cdots\overline{u_{j_{q-1}}}. \tag{5.4}$$

We now treat two special cases of power series tensor-kernels. Let $(\gamma_k)_{k\in\mathbb{N}}\in\mathbb{R}_+^{\mathbb{N}}$ and suppose that the power series $\sum_{k\in\mathbb{N}}\gamma_k\zeta^k$ ($\zeta\in\mathbb{C}$) has radius of convergence $R_\gamma>0$ ($R_\gamma=1/\limsup_k\gamma_k^{1/k}>0$). We denote by $D(R_\gamma)=\{\zeta\in\mathbb{C}\mid|\zeta|<R_\gamma\}$ and by $\psi\colon D(R_\gamma)\to\mathbb{R}$ respectively the disk of convergence and the sum of the power series $\sum_{k\in\mathbb{N}}\gamma_k\zeta^k$.

Case 1. We set

$$(\forall\,\nu\in\mathbb{N}^d)\qquad \rho_\nu=\gamma_{|\nu|}\binom{|\nu|}{\nu}=\gamma_{|\nu|}\frac{|\nu|!}{\nu_1!\cdots\nu_d!}. \tag{5.5}$$

Then, the domain of absolute convergence of the series $\sum_{\nu\in\mathbb{N}^d}\rho_\nu z^\nu$ is the strip

$$\mathcal{D}_\rho=\left\{z\in\mathbb{C}^d\;\middle|\;\left|\sum_{t=1}^d z_t\right|<R_\gamma\right\}$$

and, it follows from the multinomial theorem [3, Theorem 4.12] that, for every $z\in\mathcal{D}_\rho$,

$$\kappa(z)=\sum_{\nu\in\mathbb{N}^d}\rho_\nu z^\nu=\sum_{k\in\mathbb{N}}\gamma_k\sum_{\substack{\nu\in\mathbb{N}^d\\|\nu|=k}}\frac{k!}{\nu_1!\cdots\nu_d!}z^\nu=\sum_{k\in\mathbb{N}}\gamma_k\left(\sum_{t=1}^d z_t\right)^k=\psi\left(\sum_{t=1}^d z_t\right). \tag{5.6}$$

Note also that $\mathcal{D}_\rho^{\star1/m}=\{z\in\mathbb{C}^d\mid\|z\|_m^m<R_\gamma\}$. Thus, it follows from (5.2) that

$$\begin{aligned}K(x_1',\dots,x_q';x_1'',\dots,x_q'')&=\kappa(x_1'\star\cdots x_q'\star\overline{x_1''}\star\cdots\overline{x_q''})\\&=\psi\left(\sum_{t=1}^d x_{1,t}'\cdots x_{q,t}'\overline{x_{1,t}''}\cdots\overline{x_{q,t}''}\right),\end{aligned} \tag{5.7}$$

for every $(x_1',\dots,x_q',x_1'',\dots,x_q'')\in\mathcal{X}^q\times\mathcal{X}^q$. For $q=1$, the right hand side of (5.7) reduces to

$$K(x',x'')=\psi(\langle x'\mid x''\rangle)=\sum_{k\in\mathbb{N}}\gamma_k\langle x'\mid x''\rangle^k,$$

where $\langle\cdot\mid\cdot\rangle$ is the Euclidean scalar product in $\mathbb{R}^d$. These kind of kernels have been also called *Taylor kernels* in [21]. Thus, in virtue of (5.7), (5.3) takes the form

$$\begin{aligned}u\in\mathbb{C}^n\mapsto\left\|\sum_{i=1}^n u_i\Phi(x_i)\right\|_{r^*}^{r^*}&=\sum_{\substack{\alpha\in\mathbb{N}^n,\beta\in\mathbb{N}^n\\|\alpha|=q,|\beta|=q}}\binom{q}{\alpha}\binom{q}{\beta}\psi\left(\sum_{t=1}^d\overline{x}_{1,t}^{\alpha_1}\cdots\overline{x}_{n,t}^{\alpha_n}x_{1,t}^{\beta_1}\cdots x_{n,t}^{\beta_n}\right)u^\alpha\overline{u}^\beta\\&=\sum_{\substack{\alpha\in\mathbb{N}^n,\beta\in\mathbb{N}^n\\|\alpha|=q,|\beta|=q}}\binom{q}{\alpha}\binom{q}{\beta}\psi\left(\sum_{t=1}^d(\overline{x}_{\cdot,t})^\alpha(x_{\cdot,t})^\beta\right)u^\alpha\overline{u}^\beta,\end{aligned}$$

where we put, for every $t \in \{1, \dots, d\}$, $x_{\cdot,t} = (x_{1,t}, \dots x_{n,t}) \in \mathbb{C}^n$.[4] The representation formula (5.4) turns to

$$\langle w, \Phi(x) \rangle_{r,r^*} = \frac{1}{n^{m-1}} \sum_{\substack{i_1,\dots,i_q=1 \\ j_1,\dots,j_{q-1}=1}}^{n} \psi\Big(\sum_{t=1}^{d} \overline{x_{i_1,t}} \cdots \overline{x_{i_q,t}} x_{j_1,t} \cdots x_{j_{q-1},t} x_t \Big) u_{i_1} \cdots u_{i_q} \overline{u_{j_1}} \cdots \overline{u_{j_{q-1}}}.$$

Case 2. We set

$$(\forall\, \nu \in \mathbb{N}^d) \qquad \rho_\nu = \prod_{t=1}^{d} \gamma_{\nu_t}. \tag{5.8}$$

Then the domain of absolute convergence of the series $\sum_{\nu \in \mathbb{N}^d} \rho_\nu z^\nu$ is

$$\mathcal{D}_\rho = \left\{ z \in \mathbb{C}^d \;\middle|\; (\forall\, t \in \{1, \dots, d\})\ |z_t| < R_\gamma \right\}$$

and

$$(\forall\, z \in \mathcal{D}_\rho) \quad \kappa(z) = \sum_{\nu \in \mathbb{N}^d} \rho_\nu z^\nu = \sum_{\nu \in \mathbb{N}^d} \prod_{t=1}^{d} \gamma_{\nu_j} z_t^{\nu_t} = \prod_{t=1}^{d} \sum_{k \in \mathbb{N}} \gamma_k z_t^k = \prod_{t=1}^{d} \psi(z_t).$$

In this case $\mathcal{D}_\rho^{\star 1/m} = \{z \in \mathbb{C}^d \mid (\forall\, t \in \{1, \dots, d\}) |z_t| < R_\gamma^{1/m}\}$ and (5.2) becomes,

$$\begin{aligned} K(x'_1, \dots, x'_q; x''_1, \dots, x''_q) &= \kappa(x'_1 \star \cdots x'_q \star \overline{x''_1} \star \cdots \overline{x''_q}) \\ &= \prod_{t=1}^{d} \psi\Big(x'_{1,t} \cdots x'_{q,t} \overline{x''_{1,t}} \cdots \overline{x''_{q,t}} \Big), \end{aligned} \tag{5.9}$$

for every $(x'_1, \dots, x'_q, x''_1, \dots, x''_q) \in \mathcal{X}^q \times \mathcal{X}^q$. Thus, as done before, relying on (5.9) we can obtain the corresponding expression for the homogeneous polynomial form (5.3)

$$u \in \mathbb{C}^n \mapsto \Big\| \sum_{i=1}^{n} u_i \Phi(x_i) \Big\|_{r^*}^{r^*} = \sum_{\substack{\alpha \in \mathbb{N}^n, \beta \in \mathbb{N}^n \\ |\alpha|=q, |\beta|=q}} \binom{q}{\alpha} \binom{q}{\beta} \prod_{t=1}^{d} \psi\big(\overline{x}_{1,t}^{\alpha_1} \cdots \overline{x}_{n,t}^{\alpha_n} x_{1,t}^{\beta_1} \cdots x_{n,t}^{\beta_n} \big) u^\alpha \overline{u}^\beta \tag{5.10}$$

and the representation formula (5.4),

$$\langle w, \Phi(x) \rangle_{r,r^*} = \frac{1}{n^{m-1}} \sum_{\substack{i_1,\dots,i_q=1 \\ j_1,\dots,j_{q-1}=1}}^{n} \prod_{t=1}^{d} \psi(x_{j_1,t} \cdots x_{j_{q-1},t} x_t \overline{x_{i_1,t}} \cdots \overline{x_{i_q,t}}) u_{i_1} \cdots u_{i_q} \overline{u_{j_1}} \cdots \overline{u_{j_{q-1}}}. \tag{5.11}$$

[4] If we consider the matrix of the data $\mathsf{X} = (x_{i,t})_{\substack{1 \le i \le n \\ 1 \le t \le d}} \in \mathbb{C}^{n \times d}$, having the training set $(x_i)_{1 \le i \le n}$ as rows, the vectors $x_{\cdot,t}$ are the columns of X.

Example 5.3. We list significant examples of power series tensor kernels and for each one we provide the corresponding representation formulas.

(*i*) In (5.8) set $(\gamma_k)_{k\in\mathbb{N}} \equiv 1$, hence $(\rho_\nu)_{\nu\in\mathbb{N}^d} \equiv 1$ too. Then $R_\gamma = 1$ and $\psi(\zeta) = 1/(1-\zeta)$. Therefore, relying on (5.9), we obtain the tensor-*Szegö* kernel

$$K(x'_1,\dots,x'_q;x''_1,\dots,x''_q) = \frac{1}{\prod_{t=1}^d (1 - x'_{1,t}\cdots x'_{q,t}\overline{x''_{1,t}}\cdots\overline{x''_{q,t}})}.$$

This kernel generates a reproducing kernel Banach space of multi-variable analytic functions [17, 25]

$$B^r_{\rho,0}(\mathcal{X}) = \left\{ f\in\mathbb{C}^{\mathcal{X}} \;\middle|\; (\exists\,(c_\nu)_{\nu\in\mathbb{N}^d}\in\ell^r(\mathbb{N}^d;\mathbb{C}))(\forall\, x\in\mathcal{X})\Big(f(x) = \sum_{\nu\in\mathbb{N}^d} c_\nu x^\nu\Big)\right\}$$

with norm $\|f\|_{B^r_{\rho,b}(\mathcal{X})} = \|c\|_r$, where $(c_\nu)_{\nu\in\mathbb{N}^d}\in\ell^r(\mathbb{N}^d;\mathbb{C})$ is such that $f = \sum_{\nu\in\mathbb{N}^d} c_\nu x^\nu$ (pointwise). This space reduces to the Hardy space when $r=2$. Moreover, (5.10) yields the following homogenous polynomial form

$$u\in\mathbb{C}^n \mapsto \Big\|\sum_{i=1}^n u_i\Phi(x_i)\Big\|^{r^*}_{r^*} = \sum_{\substack{\alpha\in\mathbb{N}^n,\beta\in\mathbb{N}^n\\ |\alpha|=q,|\beta|=q}} \binom{q}{\alpha}\binom{q}{\beta}\frac{u^\alpha\overline{u}^\beta}{\prod_{t=1}^d(1-(\overline{x}_{\cdot,t})^\alpha(x_{\cdot,t})^\beta)}.$$

Finally, in view of (5.11), we have the following tensor-kernel representation

$$\langle w,\Phi(x)\rangle_{r,r^*} = \frac{1}{n^{m-1}}\sum_{\substack{i_1,\dots,i_q=1\\ j_1,\dots,j_{q-1}=1}}^n \frac{u_{i_1}\cdots u_{i_q}\overline{u_{j_1}}\cdots\overline{u_{j_{q-1}}}}{\prod_{t=1}^d(1-x_{j_1,t}\cdots x_{j_{q-1},t}x_t,\overline{x}_{i_1,t}\cdots\overline{x}_{i_q,t})}.$$

(*ii*) Set $(\gamma_k)_{k\in\mathbb{N}} \equiv ((k+1)/\pi)_{k\in\mathbb{N}}$ in (5.8). Then $R_\gamma = 1$ and $\psi(\zeta) = 1/(\pi(1-\zeta)^2)$. We then obtain the following Taylor type tensor kernel

$$K(x'_1,\dots,x'_q;x''_1,\dots,x''_q) = \frac{1}{\pi^d\prod_{t=1}^d(1-x'_{1,t}\cdots x'_{q,t}\overline{x''_{1,t}}\cdots\overline{x''_{q,t}})^2}.$$

This kernel gives rise to a reproducing kernel Banach space of analytic functions which reduces to the Bergman space when $m=2$. Proceeding as in the previous point, the expression of the corresponding homogeneous polynomial form and the representation formula can be obtained.

(*iii*) Let $(\gamma_k)_{k\in\mathbb{N}} = \big(1/k!\big)_{k\in\mathbb{N}}$ in (5.8). Then $R_\gamma = +\infty$ and $\psi(\zeta) = e^\zeta$. Hence, by (5.9),

$$K(x'_1,\dots,x'_q;x''_1,\dots,x''_q) = \prod_{t=1}^d e^{x'_{1,t}\cdots x'_{q,t}\overline{x''_{1,t}}\cdots\overline{x''_{q,t}}},$$

which is the tensor-*exponential kernel* and the form (5.10) becomes

$$u \in \mathbb{C}^n \mapsto \left\| \sum_{i=1}^{n} u_i \Phi(x_i) \right\|_{r^*}^{r^*} = \sum_{\substack{\alpha \in \mathbb{N}^n, \beta \in \mathbb{N}^n \\ |\alpha|=q, |\beta|=q}} \binom{q}{\alpha} \binom{q}{\beta} e^{\sum_{j=1}^{d} (\overline{x}_{\cdot,j})^{\alpha} (x_{\cdot,j})^{\beta}} u^{\alpha} \overline{u}^{\beta}.$$

The corresponding tensor representation is

$$\langle w, \Phi(x) \rangle_{r,r^*} = \frac{1}{n^{m-1}} \sum_{\substack{i_1,\dots,i_q=1 \\ j_1,\dots,j_{q-1}=1}}^{n} \prod_{t=1}^{d} e^{x_{i_1,t} \cdots x_{i_{q-1},t} x_t, \overline{x}_{j_1,t} \cdots \overline{x}_{j_q,t}}.$$

(*iv*) Let $\alpha > 0$, set

$$(\forall\, k \in \mathbb{N}) \qquad \gamma_k = \binom{-\alpha}{k} (-1)^k = \prod_{i=1}^{k} \frac{\alpha + i - 1}{i} > 0,$$

and define $(\rho_\nu)_{\nu \in \mathbb{N}^d}$ according to (5.5). Then $R_\gamma = 1$ and $\psi(z) = (1-\zeta)^{-\alpha}$ and formula (5.7) yields the following tensorial version of the *binomial kernel* [21]

$$K(x'_1, \dots, x'_q; x''_1, \dots, x''_q) = \frac{1}{\left(1 - \sum_{t=1}^{d} x'_{1,t} \cdots x'_{q,t} \overline{x''_{1,t}} \cdots \overline{x''_{q,t}}\right)^{\alpha}}.$$

(*v*) Let $s \in \mathbb{N}$, set

$$(\forall\, k \in \mathbb{N}) \qquad \gamma_k = \begin{cases} \binom{s}{k} & \text{if } k \le s \\ 0 & \text{if } k > s, \end{cases}$$

and define $(\rho_\nu)_{\nu \in \mathbb{N}^d}$ according to (5.5). Then $R_\gamma = +\infty$ and $\psi(\zeta) = (1+\zeta)^s$. This way, by (5.7), we have

$$K(x'_1, \dots, x'_q; x''_1, \dots, x''_q) = \left(1 + \sum_{t=1}^{d} x'_{1,t} \cdots x'_{q,t} \overline{x''_{1,t}} \cdots \overline{x''_{q,t}}\right)^{s},$$

which is the *polynomial tensor-kernel* of order s. By (5.5) we have that $\rho_\nu > 0$ if $|\nu| \le s$ and $\rho_\nu = 0$ if $|\nu| > s$. Therefore, recalling (5.1), we have that

$$B^r_{\rho,0}(\mathcal{X}) = \left\{ f \in \mathbb{C}^{\mathcal{X}} \,\middle|\, (\exists\, (c_\nu)_{\nu \in \mathbb{N}^d} \in \ell^r(\mathbb{N}^d; \mathbb{C}))(\forall\, x \in \mathcal{X}) \Big(f(x) = \sum_{\nu \in \mathbb{N}^d} c_\nu \phi_\nu(x) \Big) \right\},$$

is the space of polynomials in d variables with coefficients in $\mathbb{C}$ of degree up to s.

6 Conclusion

In this work we first provided a complete duality theory for support vector regression in Banach function spaces with general regularizers. Then, we specialized the analysis to reproducing kernel Banach spaces that admit a representation in terms of a (countable) dictionary of functions with ℓ^r-summable coefficients and regularization terms of type $\varphi(\|\cdot\|_r)$, being $r = m/(m-1)$ and m an even integer. In this context we showed that the problem of support vector regression can be explicitly solved through the introduction of a new type of kernel of tensorial type (with degree m) which completely encodes the finite dimensional dual problem as well as the representation of the corresponding infinite dimensional primal solution (the regression function). This can provide a new and effective computational framework for solving support vector regression in Banach space setting. We finally study a whole class of reproducing kernel Banach spaces of analytic functions to which the theory applies and show significant examples which can become useful in applications.

Acknowledgments. The research leading to these results has received funding from the European Research Council (FP7/2007–2013) / ERC AdG A-DATADRIVE-B (290923) under the European Union's Seventh Framework Programme. This paper reflects only the authors' views, the Union is not liable for any use that may be made of the contained information; Research Council KUL: GOA/10/09 MaNet, CoE PFV/10/002 (OPTEC), BIL12/11T; PhD/Postdoc grants; Flemish Government: FWO: PhD/Postdoc grants, projects: G.0377.12 (Structured systems), G.088114N (Tensor based data similarity); IWT: PhD/Postdoc grants, projects: SBO POM (100031); iMinds Medical Information Technologies SBO 2014; Belgian Federal Science Policy Office: IUAP P7/19 (DYSCO, Dynamical systems, control and optimization, 2012–2017).

References

[1] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces.* Springer, New York 2011.

[2] V. I. Bogachev, *Measure Theory.* Springer, Berlin 2007.

[3] M. Bóna, *A Walk Through Combinatorics. 3rd Ed.* World Scientific, Singapore 2011.

[4] I. Cioranescu, *Geometry of Banach Spaces, Duality Mappings and Nonlinear Problems.* Kluwer, Dordrecht 1990.

[5] P. L. Combettes, S. Salzo, and S. Villa, Consistency of Regularized Learning Schemes in Banach Spaces. arXiv:1410.6847v3, 2015.

[6] P. L. Combettes, S. Salzo, and S. Villa, Consistent Learning by Composite Proximal Thresholding. arXiv:1504.04636v2, 2015.

[7] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines.* Cambridge University Press, Cambridge 2000.

[8] C. De Mol, E. De Vito, and L. Rosasco, Elastic-net regularization in learning theory, *J. Complexity,* vol. 25, pp. 201–230, 2009.

[9] E. De Vito, L. Rosasco, A. Caponnetto, M. Piana, and A. Verri, Some properties of regularized kernel methods, *J. Mach. Learn. Res.,* vol. 5, pp. 1363–1390, 2004.

[10] G. E. Fasshauer, F. J. Hickernell, and Q. Ye, Solving support vector machines in reproducing kernel Banach spaces with positive definite functions, *Appl. Comput. Harmon. Anal.,* vol. 38, pp. 115–139, 2015.

[11] W. Fu, Penalized regressions: the bridge versus the lasso, *J. Comput. Graph. Stat.,* vol. 7, pp. 397–416, 1998.

[12] F. Girosi, An Equivalence Between Sparse Approximation and Support Vector Machines, *Neural Comput.,* vol. 10(6), pp. 1455–1480, 1998.

[13] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms II.* Springer, Berlin 1996.

[14] T. Hofmann, B. Schölkopf, and A. J. Smola, Kernel methods in machine learning, *Ann. Statist.,* vol. 36, pp. 1171–1220, 2008.

[15] V. Koltchinskii, Sparsity in penalized empirical risk minimization, *Ann. Inst. Henri Poincaré Probab. Stat.,* vol. 45, pp. 7–57, 2009.

[16] B. S. Mendelson and J. Neeman, Regularization in Kernel Learning, *Ann. Statist.*, 38(1), pp. 526–565, 2010.

[17] V. I. Paulsen, *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces.* [On line]. Available: http://www.math.uh.edu/~vern/rkhs.pdf

[18] R. T. Rockafellar, *Conjugate Duality and Optimization.* SIAM, Philadelphia, PA 1974.

[19] B. Schölkopf, R. Herbrich, and A. J. Smola, A Generalized Representer Theorem. In *Computational Learning Theory: 14th Annual Conference on Computational Learning Theory, COLT 2001.* Springer Berlin Heidelberg, 2001.

[20] I. Steinwart and A. Christmann, Sparsity of SVMs that use the ε-insensitive loss. In *Advances in Neural Information Processing Systems 21.* Curran Associates, Inc., 2009.

[21] I. Steinwart and A. Christmann, *Support Vector Machines.* Springer, New York 2008.

[22] B. K. Sriperumbudur, K. Fukumizu, and G. R. G. Lanckriet, Learning in Hilbert vs. Banach spaces: a measure embedding viewpoint, in: *Advances in Neural Information Processing Systems 24.* Curran Associates, Inc., 2011.

[23] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle, *Least Squares Support Vector Machines.* World Scientific, Singapore 2002.

[24] V. N. Vapnik, *Statistical Learning Theory.* Wiley, New York 1998.

[25] R. M. Young, *An Introduction to Nonharmonic Fourier Series.* Academic Press, San Diego 2001.

[26] C. Zălinescu, *Convex Analysis in General Vector Spaces.* World Scientific, River Edge, NJ 2002.

[27] H. Zhang, Y. Xu, and J. Zhang, Reproducing kernel Banach spaces for machine learning, *J. Mach. Learn. Res.,* vol. 10, pp. 2741–2775, 2009.

[28] H. Zhang and J. Zhang, Regularized learning in Banach spaces as an optimization problem: representer theorems, *J. Global Optim.,* vol. 54, pp. 235–250, 2012.

[29] B. Zwicknagl, Power series kernels, *Constr. Approx.,* vol. 29, pp. 61–84, 2009.