2013 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 22–25, 2013, SOUTHAPMTON, UK

COMPLEX SUPPORT VECTOR MACHINES FOR QUATERNARY CLASSIFICATION

P. Bouboulis *E. Theodoridi* *S. Theodoridis**

Department of Informatics
and Telecommunications,
University of Athens,
Athens, Greece.
panbouboulis@gmail.com eva.theodoridou86@gmail.com stheodor@di.uoa.gr

ABSTRACT

We present a support vector machines (SVM) rationale suitable for quaternary classification problems that use complex data, exploiting the notions of widely linear estimation and pure complex kernels. The recently developed Wirtinger's calculus on complex RKHS is employed in order to compute the Lagrangian and derive the dual optimization problem. We show that this approach is equivalent with solving two real SVM tasks exploiting a specific real kernel, which is induced by the chosen complex kernel.

***Index Terms*—** Complex SVM, Quaternary Classification, complex kernels, RKHS

1. INTRODUCTION

Support vector machines (SVM) have become a popular toolbox for addressing non-linear classification tasks. The excellent performance of SVMs was firmly grounded in the context of statistical learning theory, which ensures their fine generalization properties. In the context of regression, this toolbox is usually known as Support Vector Regression (SVR).

In the SVM framework, the notion of the Reproducing Kernel Hilbert Space (RKHS) plays a significant role. The original data are transformed into a higher dimensional RKHS $\mathcal{H}$ (possibly of infinite dimension) and linear tools are applied to the transformed data in the so called feature space $\mathcal{H}$. This is equivalent to solving a non-linear problem in the original space. Furthermore, inner products in $\mathcal{H}$ can efficiently be computed via the specific kernel function κ associated to the RKHS $\mathcal{H}$, disregarding the actual structure of the function space.

Although the theory of RKHS has been developed by the mathematicians for general complex spaces, most kernel-based methods employ real kernels. This is largely due to the fact that many of them originated as variants of the original SVM formulation, which was targeted to treat real data. However, in modern applications complex data arise frequently in areas as diverse as biomedicine, communications, radar, etc. Hence, the design of SVMs suitable for treating problems of complex and/or multidimensional outputs has, recently, attracted attention in the machine learning community. Perhaps, till now, the most complete works, which attempt to generalize the SVM rationale in this context, are a) the Clifford SVMs [1] and b) the division algebraic SVR [2, 3].

It is important to emphasize that the aforementioned efforts to generalize the SVM rationale to complex and hypercomplex numbers are limited to the case of the output data. These methods consider a multidimensional output, which can be represented, for example, as a complex number or a quaternion, while the input data are real vectors. In the following, they employ real valued kernels to model the input-output relationship, breaking it down to its multidimensional components. In this short paper we propose a different approach. Our modeling takes place directly into complex RKHS, which are generated by pure complex kernels (meaning that they take complex inputs and return a complex value), instead of real ones. In that fashion, besides complex (or multidimensional) outputs, we can exploit complex inputs as well. To be inline with the current trend in complex signal processing, we employ the widely linear estimation rationale, which has been shown to perform better than the standard linear one [4, 5, 6, 7, 8, 9]. This means that we model the input-output relationship as a sum of two parts. The first is linear with respect to the input vector, while the second is linear with respect to its conjugate. The widely linear approach is a necessity, as the alternative would lead to a significantly restricted model. In order to compute the gradients, which are required by the Karush-Kuhn-Tucker conditions and the dual, we employ the generalized Wirtinger Calculus introduced in [10].

Following the proposed rationale, i.e., working in a RKHS

*This work was carried out under the 621 ARISTEIA program, co-financed by the Greek Secretariat for Research and Development and the EU.

978-1-4673-1026-0/12/$31.00

$\mathbb{H}$ induced by a pure complex kernel $\kappa_{\mathbb{C}}$, it can be shown that the problem is equivalent to two problems in a real RKHS $\mathcal{H}$, albeit with a specific real kernel $\kappa_{\mathbb{R}}$, which is induced by the complex $\kappa_{\mathbb{C}}$. Hence, well established algorithms that solve the SVM problem can be easily modified to this case also. It must be pointed out that these induced kernels are not trivial. For example, the exploitation of the complex gaussian kernel results to an induced kernel which is different from the standard real gaussian RBF. Moreover, it turns out that the proposed complex SVM rationale, is suitable for quaternary classification (i.e., four classes problem), in contrast to the binary classification carried out by the traditional real SVM approach. This is because, the widely linear concept turns out to be equivalent to two, instead of one, hyperplanes in the corresponding RKHS.

The paper is organized as follows. In Section 2 the main mathematical background regarding RKHS is outlined and the differences between real and complex RKHS's are highlighted. Section 3 describes the standard real SVM algorithm. The main contributions of the paper can be found in Section 4, where the theory and the generalized complex algorithms are developed. Experiments are presented in Section 5. Throughout the paper, we will denote the set of all integers, real and complex numbers by $\mathbb{N}$, $\mathbb{R}$ and $\mathbb{C}$ respectively. The imaginary unit is denoted as $\mathbb{i}$. Vector or matrix valued quantities appear in boldfaced symbols.

2. MATHEMATICAL BACKGROUND

A RKHS [11] is a Hilbert space $\mathcal{H}$ over a field $\mathbb{F}$ for which there exists a positive definite function $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{F}$ with the following two important properties: a) For every $x \in \mathcal{X}$, $\kappa(\cdot, x)$ belongs to $\mathcal{H}$ and b) κ has the so called *reproducing property*, i.e., $f(x) = \langle f, \kappa(\cdot, x)\rangle_{\mathcal{H}}$, for all $f \in \mathcal{H}$, in particular $\kappa(x, y) = \langle \kappa(\cdot, y), \kappa(\cdot, x)\rangle_{\mathcal{H}}$. The map $\Phi : \mathcal{X} \to \mathcal{H} : \Phi(x) = \kappa(\cdot, x)$ is called the *feature map* of $\mathcal{H}$. In the case of complex Hilbert spaces (i.e., $\mathbb{F} = \mathbb{C}$) the inner product is sesqui-linear (i.e., linear in one argument and antilinear in the other) and Hermitian. In the following, we will denote by $\mathbb{H}$ a complex RKHS and by $\mathcal{H}$ a real one. Moreover, in order to distinguish the two cases, we will use the notations $\kappa_{\mathbb{R}}$ and $\Phi_{\mathbb{R}}$ to refer to a real kernel and its corresponding feature map, instead of the notation $\kappa_{\mathbb{C}}$, $\Phi_{\mathbb{C}}$, which is reserved for pure complex kernels.

Although there are many kernels to choose from, in this paper the experiments are focused on the real *Gaussian radial basis function*, i.e., $\kappa_{\mathbb{R}^\nu,t}(\boldsymbol{x}, \boldsymbol{y}) := \exp\left(-t\sum_{k=1}^{\nu}(x_k - y_k)^2\right)$, defined for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^\nu$, and the *complex Gaussian kernel*: $\kappa_{\mathbb{C}^\nu,t}(\boldsymbol{z}, \boldsymbol{w}) := \exp\left(-t\sum_{k=1}^{\nu}(z_k - w_k^*)^2\right)$, where $\boldsymbol{z}, \boldsymbol{w} \in \mathbb{C}^\nu$, z_k denotes the k-th component of the complex vector $\boldsymbol{z}$ and $\exp(\cdot)$ is the extended exponential function in the complex domain. In both cases, t, is a free positive parameter that defines the shape of the kernel function.

Besides the complex RKHS produced by a complex kernel, such as the Gaussian one, one may construct a complex RKHS as a cartesian product of a real RKHS with itself, in a fashion similar to the identification of the field of complex numbers, $\mathbb{C}$, to $\mathbb{R}^2$. This technique is called *complexification* of a real RKHS and the respective Hilbert space is called *complexified* RKHS. This can be done , if we enrich $\mathcal{H}^2$, where $\mathcal{H}$ is a real RKHS associated with a real kernel $\kappa_{\mathbb{R}}$, with a complex structure using the complex inner product:

$$\langle f, g\rangle_{\mathbb{H}} = \langle f^r, g^r\rangle_{\mathcal{H}} + \langle f^i, g^i\rangle_{\mathcal{H}} + \mathbb{i}\left(\langle f^i, g^r\rangle_{\mathcal{H}} - \langle f^r, g^i\rangle_{\mathcal{H}}\right), \tag{1}$$

for $f = f^r + \mathbb{i}f^i$, $g = g^r + \mathbb{i}g^i$, where $\mathbb{H} = \{f = f^r + \mathbb{i}f^i;\ f^r, f^i \in \mathcal{H}\}$ is the respective complex RKH space. We call $\mathbb{H}$ the complexification of $\mathcal{H}$. We map the data samples from the complex input space to the complexified RKHS $\mathbb{H}$ using the following simple rule:

$$\bar{\Phi}_{\mathbb{C}}(\boldsymbol{z}) = \Phi_{\mathbb{R}}(\boldsymbol{x}, \boldsymbol{y}) + \mathbb{i}\Phi_{\mathbb{R}}(\boldsymbol{x}, \boldsymbol{y}), \tag{2}$$

where $\Phi_{\mathbb{R}}$ is the feature map of the real reproducing kernel $\kappa_{\mathbb{R}}$, i.e., $\Phi_{\mathbb{R}}(\boldsymbol{x}, \boldsymbol{y}) = \kappa_{\mathbb{R}}(\cdot, (\boldsymbol{x}, \boldsymbol{y}))$ and $\boldsymbol{z} = \boldsymbol{x} + \mathbb{i}\boldsymbol{y}$. It must be emphasized that a complex RKHS, $\mathbb{H}$, (whether it is constructed through the complexification procedure, or it is produced by a complex kernel) can, always, be represented as a cartesian product of a Hilbert space with itself, i.e., we can, always, identify $\mathbb{H}$ with a *doubled real space* $\mathcal{H}^2$. Furthermore, the complex inner product of $\mathbb{H}$ can always be related to the real inner product of $\mathcal{H}$ as in (1).

In order to compute the gradients of real valued cost functions, which are defined on complex domains, we adopt the rationale of Wirtinger's calculus [12]. This was brought into light recently [5, 6, 13], as a means to compute, in an efficient and elegant way, gradients of real valued cost functions that are defined on complex domains ($\mathbb{C}^\nu$), in the context of widely linear processing [7, 14]. It is based on simple rules and principles, which bear a great resemblance to the rules of the standard complex derivative, and it greatly simplifies the calculations of the respective derivatives. In [10], the notion of Wirtinger's calculus was extended to general complex Hilbert spaces, providing the tool to compute the gradients that are needed to develop kernel-based algorithms for treating complex data. In [15] the notion of Wirtinger calculus was extended to include subgradients in RKHS.

3. REAL SVM

In this section, we briefly describe the standard SVM rationale. Suppose we are given training data, which belong to two separate classes C_+, C_- and have the form $\{(\boldsymbol{x}_n, d_n);\ n = 1, \dots, N\} \subset \mathcal{X} \times \{\pm 1\}$, where if $d_n = +1$, then the n-th sample belongs to C_+, while if $d_n = -1$, then the n-th sample belongs to C_-. Consider the real RKHS $\mathcal{H}$ with respective kernel $\kappa_{\mathbb{R}}$. We transform the input data from $\mathcal{X}$ to $\mathcal{H}$, via the feature map $\Phi_{\mathbb{R}}$. The goal of the SVM task is to estimate the maximum margin hyperplane, that separates the points of the two classes as best as possible [16, 17, 18]. This is usually cast as

$$\begin{array}{ll} \underset{w\in\mathcal{H}, c\in\mathbb{R}}{\text{minimize}} & \frac{1}{2}\|w\|_{\mathcal{H}}^2 + \frac{C}{N}\sum_{n=1}^{N}\xi_n \\ \text{subject to} & \begin{cases} d_n\left(\langle\Phi_{\mathbb{R}}(\boldsymbol{x}_n), w\rangle_{\mathcal{H}} + c\right) & \geq & 1-\xi_n \\ \xi_n & \geq & 0 \end{cases} \\ & \quad \text{for } n=1,\dots,N, \end{array} \tag{3}$$

for some $C > 0$. This is a constant that determines the trade-off between the two conflicting goals of the SVM task: maximizing the margin (i.e., $2/\|w\|^2$) and minimizing the training error (i.e., $\sum_{n=1}^{N}\xi_n$).

Introducing the Lagrangian and exploiting the KKT conditions we find that the dual problem is casted as:

$$\begin{array}{ll} \underset{\boldsymbol{a}\in\mathbb{R}^N}{\text{maximize}} & \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n,m=1}^{N} a_n a_m d_n d_m \kappa_{\mathbb{R}}(\boldsymbol{x}_m, \boldsymbol{x}_n) \\ \text{subject to} & \sum_{n=1}^{N} a_n d_n = 0 \text{ and } a_n \in [0, C/N]. \end{array} \tag{4}$$

Furthermore, the solution can be shown to have an expansion:

$$w = \sum_{n=1}^{N} a_n d_n \kappa_{\mathbb{R}}(\cdot, \boldsymbol{x}_n),$$

while the threshold c can be computed by averaging

$$c = d_m - \sum_{n=1}^{N} a_n d_n \kappa_{\mathbb{R}}(\boldsymbol{x}_m, \boldsymbol{x}_n),$$

over all points with $0 < a_m < C$, for $m = 1, \dots, N$.

4. COMPLEX SVM

Recall that in any real Hilbert space $\mathcal{H}$, a hyperplane consists of all the elements $f \in \mathcal{H}$ that satisfy

$$\langle f, w\rangle_{\mathcal{H}} + b = 0, \tag{5}$$

for some $w \in \mathcal{H}$, $b \in \mathbb{R}$. Moreover, as figure 1 shows, any hyperplane of $\mathcal{H}$ divides the space into two parts, $\mathcal{H}_+ = \{f \in \mathcal{H};\ \langle f, w\rangle_{\mathcal{H}} + b > 0\}$ and $\mathcal{H}_- = \{f \in \mathcal{H};\ \langle f, w\rangle_{\mathcal{H}} + b < 0\}$. In the traditional SVM classification task, which has been outlined in section 3, the goal is to separate two distinct classes of data by a maximum margin hyperplane, so that one class falls into $\mathcal{H}_+$ and the other into $\mathcal{H}_-$ (excluding some outliers). In order to be able to generalize the SVM rationale to complex spaces, we need first to develop an appropriate definition for a complex hyperplane. The difficulty is that the set of complex numbers is not an ordered one, and thus one may not assume that a complex version of (5) divides the space into two parts, as $\mathcal{H}_+$ and $\mathcal{H}_-$ cannot be defined. To

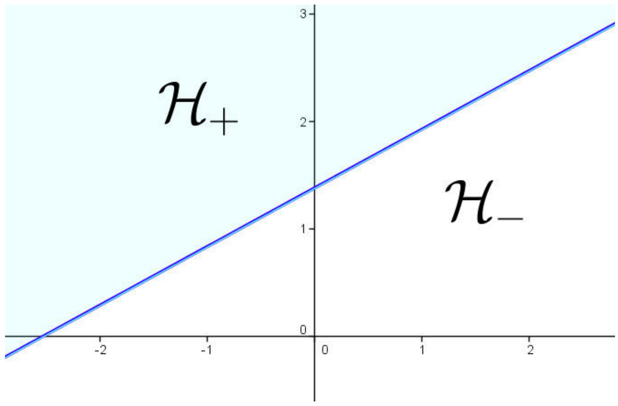

Fig. 1. A hyperplane separates the space $\mathcal{H}$ into two parts, $\mathcal{H}_+$ and $\mathcal{H}_-$.

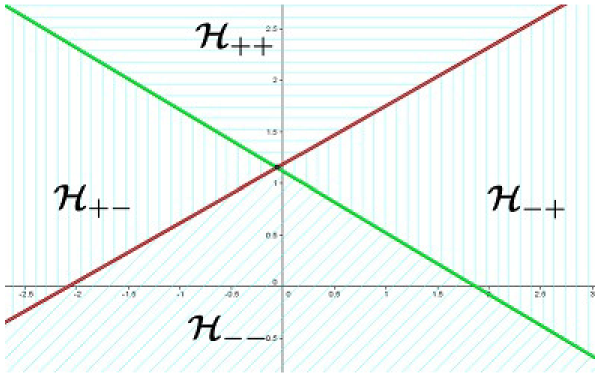

Fig. 2. A complex couple of hyperplanes separates the space of complex numbers (i.e., $\mathbb{H} = \mathbb{C}$) into four parts.

circumvent this obstacle, we will provide a novel definition of complex hyperplanes, which will divide the complex space into four parts. This will be our kick off point for deriving the complex SVM rationale, which classifies objects into four (instead of two) classes.

Let us begin by considering the following two relations,

$$\mathrm{Re}\left(\langle f, w\rangle_{\mathbb{H}} + c\right) = 0, \tag{6a}$$
$$\mathrm{Im}\left(\langle f, w\rangle_{\mathbb{H}} + c\right) = 0, \tag{6b}$$

for some $w \in \mathbb{H}$, $c \in \mathbb{C}$, where $f \in \mathbb{H}$. It is not difficult to see, that this couple of relations represents two *orthogonal* hyperplanes of the doubled real space, i.e., $\mathcal{H}^2$. To overcome this constraint and be able to define arbitrarily placed hyperplanes, we need to employ the widely linear estimation functions, i.e.,

$$\mathrm{Re}\left(\langle f, w\rangle_{\mathbb{H}} + \langle f^*, v\rangle_{\mathbb{H}} + c\right) = 0, \tag{7a}$$
$$\mathrm{Im}\left(\langle f, w\rangle_{\mathbb{H}} + \langle f^*, v\rangle_{\mathbb{H}} + c\right) = 0, \tag{7b}$$

for some $w, v \in \mathbb{H}$, $c \in \mathbb{C}$, where $f \in \mathbb{H}$. Depending on the values of w, v, these hyperplanes may be placed *arbitrarily* on $\mathcal{H}^2$. We define this complex couple of hyperplanes as the set of all $f \in \mathbb{H}$ that satisfy either one of the relations (7), for some $w, v \in \mathbb{H}$, $c \in \mathbb{C}$.

The aforementioned arguments demonstrate the significant difference between complex linear estimation and widely linear estimation functions, which has been pointed out by many other authors, in the context of regression tasks. In the current context of classification, we have just seen that confining to complex linear modeling is quite restrictive, as the corresponding couple of complex hyperplanes are always orthogonal. On the other hand, the widely linear case is more general and covers all cases. The complex couple of hyperplanes (as defined above) divides the space into four parts, i.e.,

$$\mathcal{H}_{++} = \left\{ f \in \mathcal{H};\ \begin{matrix} \mathrm{Re}\left(\langle f, w\rangle_{\mathbb{H}} + \langle f^*, v\rangle_{\mathbb{H}} + c\right) > 0, \\ \mathrm{Im}\left(\langle f, w\rangle_{\mathbb{H}} + \langle f^*, v\rangle_{\mathbb{H}} + c\right) > 0 \end{matrix} \right\},$$
$$\mathcal{H}_{+-} = \left\{ f \in \mathcal{H};\ \begin{matrix} \mathrm{Re}\left(\langle f, w\rangle_{\mathbb{H}} + \langle f^*, v\rangle_{\mathbb{H}} + c\right) > 0, \\ \mathrm{Im}\left(\langle f, w\rangle_{\mathbb{H}} + \langle f^*, v\rangle_{\mathbb{H}} + c\right) < 0 \end{matrix} \right\},$$
$$\mathcal{H}_{-+} = \left\{ f \in \mathcal{H};\ \begin{matrix} \mathrm{Re}\left(\langle f, w\rangle_{\mathbb{H}} + \langle f^*, v\rangle_{\mathbb{H}} + c\right) < 0, \\ \mathrm{Im}\left(\langle f, w\rangle_{\mathbb{H}} + \langle f^*, v\rangle_{\mathbb{H}} + c\right) > 0 \end{matrix} \right\},$$
$$\mathcal{H}_{--} = \left\{ f \in \mathcal{H};\ \begin{matrix} \mathrm{Re}\left(\langle f, w\rangle_{\mathbb{H}} + \langle f^*, v\rangle_{\mathbb{H}} + c\right) < 0, \\ \mathrm{Im}\left(\langle f, w\rangle_{\mathbb{H}} + \langle f^*, v\rangle_{\mathbb{H}} + c\right) < 0 \end{matrix} \right\}.$$

Figure 2 demonstrates a simple case of a complex couple of hyperplanes that divides $\mathbb{C}$ into four parts.

We are now ready to formulate the more general complex SVM classification task as follows. Suppose we are given training data, which belong to four separate classes $C_{++}, C_{+-}, C_{-+}, C_{--}$, i.e., $\{(\boldsymbol{z}_n, d_n);\ n = 1, \ldots, N\} \subset \mathcal{X} \times \{\pm 1 \pm \mathbb{i})\}$. If $d_n = +1 + \mathbb{i}$, then the n-th sample belongs to C_{++}, i.e., $\boldsymbol{z}_n \in C_{++}$, if $d_n = 1 - \mathbb{i}$, then $\boldsymbol{z}_n \in C_{+-}$, e.t.c. Consider the complex RKHS, $\mathbb{H}$, with respective kernel $\kappa_{\mathbb{C}}$. Following a similar rationale to the real case, we transform the input data from $\mathcal{X}$ to $\mathbb{H}$, via the feature map $\Phi_{\mathbb{C}}$. The goal of the SVM task is to estimate a complex couple of maximum margin hyperplanes, that separates the points of the four classes as best as possible. To this end, we formulate the primal complex SVM as

$$\begin{array}{ll} \min\limits_{w,v,c} & \frac{1}{2}\|w\|_{\mathbb{H}}^2 + \frac{1}{2}\|v\|_{\mathbb{H}}^2 + \frac{C}{N}\sum_{n=1}^{N}(\xi_n^r + \xi_n^i) \\ \text{s. to} & \begin{cases} d_n^r\,\mathrm{Re}\left(\langle \Phi_{\mathbb{C}}(\boldsymbol{z}_n), w\rangle_{\mathbb{H}} + \langle \Phi_{\mathbb{C}}^*(\boldsymbol{z}_n), v\rangle_{\mathbb{H}} + c\right) \geq 1 - \xi_n^r \\ d_n^i\,\mathrm{Im}\left(\langle \Phi_{\mathbb{C}}(\boldsymbol{z}_n), w\rangle_{\mathbb{H}} + \langle \Phi_{\mathbb{C}}^*(\boldsymbol{z}_n), w\rangle_{\mathbb{H}} + c\right) \geq 1 - \xi_n^i \\ \xi_n^r, \xi_n^i \geq 0 \\ \text{for } n = 1, \ldots, N. \end{cases} \end{array} \tag{8}$$

Consequently, the Lagrangian function becomes

$$\begin{aligned} L(w, v, \boldsymbol{a}, \hat{\boldsymbol{a}}, \boldsymbol{b}, \hat{\boldsymbol{b}}) &= \frac{1}{2}\|w\|_{\mathbb{H}}^2 + \frac{1}{2}\|v\|_{\mathbb{H}}^2 + \frac{C}{N}\sum_{n=1}^{N}(\xi_n^r + \xi_n^i) \\ &- \sum_{n=1}^{N} a_n\left(d_n^r\,\mathrm{Re}\left(\langle \Phi_{\mathbb{C}}(\boldsymbol{z}_n), w\rangle_{\mathbb{H}} + \langle \Phi_{\mathbb{C}}^*(\boldsymbol{z}_n), v\rangle_{\mathbb{H}} + c\right) - 1 + \xi_n^r\right) \\ &- \sum_{n=1}^{N} b_n\left(d_n^i\,\mathrm{Im}\left(\langle \Phi_{\mathbb{C}}(\boldsymbol{z}_n), w\rangle_{\mathbb{H}} + \langle \Phi_{\mathbb{C}}^*(\boldsymbol{z}_n), w\rangle_{\mathbb{H}} + c\right) - 1 + \xi_n^i\right) \\ &- \sum_{n=1}^{N}\eta_n\xi_n^r - \sum_{n=1}^{N}\theta_n\xi_n^i, \end{aligned}$$

where $a_n, b_n, \eta_n, \theta_n$ are the positive Lagrange multipliers of the respective inequalities, for $n = 1, \ldots, N$. Employing the notion of Wirtinger's calculus to derive the respective gradients and exploiting the saddle point conditions of the Lagrangian function, it turns out that the dual problem can be split into two separate maximization tasks:

$$\begin{array}{ll} \text{maximize}_{\boldsymbol{a}} & \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n,m=1}^{N} a_n a_m d_n^r d_m^r \kappa_{\mathbb{C}}^r(\boldsymbol{z}_m, \boldsymbol{z}_n) \\ \text{subject to} & \begin{cases} \sum_{n=1}^{N} a_n d_n^r = 0 \\ 0 \leq a_n \leq \frac{C}{N} \end{cases} \\ & \text{for } n = 1, \ldots, N \end{array} \tag{9a}$$

and

$$\begin{array}{ll} \text{maximize}_{\hat{\boldsymbol{a}}} & \sum_{n=1}^{N} b_n - \frac{1}{2}\sum_{n,m=1}^{N} b_n b_m d_n^i d_m^i \kappa_{\mathbb{C}}^r(\boldsymbol{z}_m, \boldsymbol{z}_n) \\ \text{subject to} & \begin{cases} \sum_{n=1}^{N} b_n d_n^i = 0 \\ 0 \leq b_n \leq \frac{C}{N} \end{cases} \\ & \text{for } n = 1, \ldots, N, \end{array} \tag{9b}$$

where

$$w = \sum_{n=1}^{N}(a_n d_n^r - \mathbb{i} b_n d_n^i)\Phi_{\mathbb{C}}(\boldsymbol{z}_n),$$
$$v = \sum_{n=1}^{N}(a_n d_n^r - \mathbb{i} b_n d_n^i)\Phi_{\mathbb{C}}^*(\boldsymbol{z}_n)$$
$$\text{and} \quad a_n + \eta_n = \frac{C}{N}, \quad b_n + \theta_n = \frac{C}{N},$$

for $n = 1, \ldots, N$. The function $\kappa_{\mathbb{C}}^r$, that appears in the dual problems, is the induced real kernel defined on $\mathbb{R}^{2\nu} \times \mathbb{R}^{2\nu}$ (or $\mathbb{C}^{\nu} \times \mathbb{C}^{\nu}$) as follows:

$$\kappa_{\mathbb{C}}^r(\boldsymbol{z}, \boldsymbol{z}') = \kappa_{\mathbb{C}}^r\left(\begin{pmatrix}\boldsymbol{x} \\ \boldsymbol{y}\end{pmatrix}, \begin{pmatrix}\boldsymbol{x}' \\ \boldsymbol{y}'\end{pmatrix}\right) = 2\,\mathrm{Re}(\kappa_{\mathbb{C}}(\boldsymbol{z}, \boldsymbol{z}')), \tag{10}$$

for $\boldsymbol{z} = \boldsymbol{x} + \mathbb{i}\boldsymbol{y}$, $\boldsymbol{z}' = \boldsymbol{x}' + \mathbb{i}\boldsymbol{y}'$.

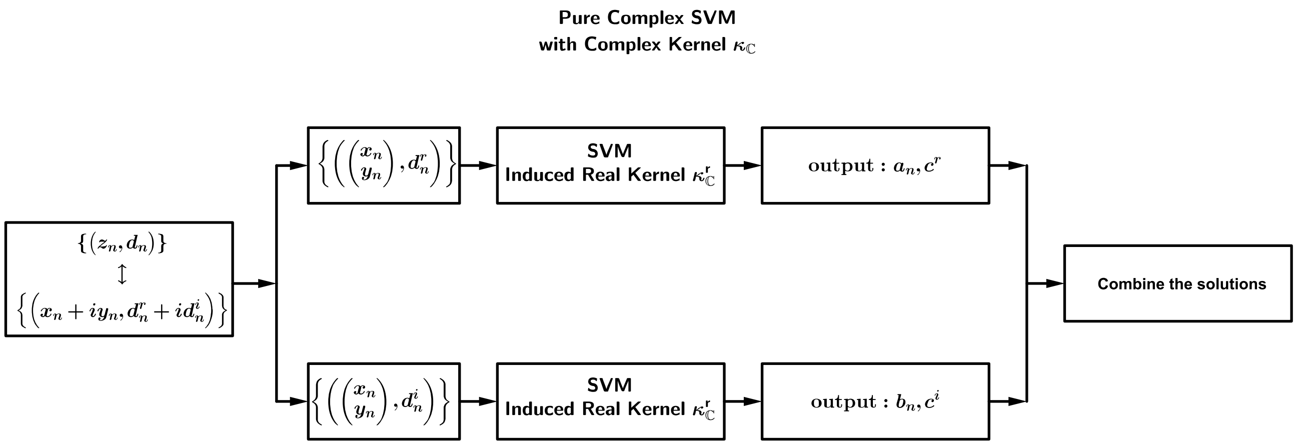

Fig. 3. Pure Complex Support Vector Machines.

We observe that these problems are equivalent with two distinct real SVM (dual) tasks employing the induced real kernel $\kappa_{\mathbb{C}}^r$. One may split the (output) data to their real and imaginary parts, as figure 3 demonstrates, solve two real SVM tasks employing any one of the standard algorithms and, finally, combine the solutions to take the complex labeling function:

$$g(\boldsymbol{z}) = \underset{\mathbb{i}}{\operatorname{sign}}\left(\sum_{n=1}^{N}(a_n d_n^r + \mathbb{i} b_n d_n^i)\kappa_{\mathbb{C}}^r(\boldsymbol{z}_n, \boldsymbol{z}) + c^r + \mathbb{i} c^i\right),$$

$$\text{where} \quad \underset{\mathbb{i}}{\operatorname{sign}}(z) = \operatorname{sign}(\operatorname{Re}(z)) + \mathbb{i}\operatorname{sign}(\operatorname{Im}(z)).$$

If we follow the complexification procedure, i.e., employ a real kernel $\kappa_{\mathbb{R}}$ and transform the input data from $\mathcal{X}$ to the complexified space $\mathbb{H}$, instead of the complex RKHS $\mathbb{H}$, we can similarly deduce that the dual of the complexified SVM task is equivalent to two real SVM tasks employing the kernel $2\kappa_{\mathbb{R}}$. Hence, in both cases, we end up with two real SVM tasks. However, while in the complexification procedure we directly employ a selected real kernel, in the pure complex case we exploit a real kernel that is induced by the selected complex kernel. Although both scenarios are developed for quaternary classification, they can be easily adapted to the binary case also. This can be done by considering that the labels of the data are real numbers (i.e., $d_n \in \mathbb{R}$) taking the values ± 1. In this case we solve one problem instead of two.

5. EXPERIMENTS

We perform two simple experiments to demonstrate the advantages of exploiting complex data, using the popular MNIST database of handwritten digits [19]. In both cases, the respective parameters of the SVM tasks were tuned to obtain the lowest error rate possible. The MNIST database contains 60000 handwritten digits (from 0 to 9) for training and 10000 handwritten digits for testing. Each digit is encoded as an image file with 28×28 pixels. The scenario, that it is typically used to quantify the performance of an SVM-like learning machine, is to employ a one-versus-all strategy to the training set (using the raw pixel values as input data) and then measure the success using the testing set [20, 21].

In the first experiment, we compare the aforementioned standard one-versus-all scenario with a classification task that exploits complex numbers. In the complex variant, we perform a Fourier transform to each training image and keep only the 100 most significant coefficients. As these coefficients are complex numbers, we employ a one-versus-all classification task using the binary complexified SVM rationale. In both scenarios we use the first 6000 digits of the MNIST training set to train the learning machines and test their performances using the 10000 digits of the testing set. In addition, we used the gaussian kernel with $t = 1/64$ and $t = 1/140^2$ respectively. The SVM parameter C has been set equal to 100. The error rate of the standard real-valued scenario is 3.79%, while the error rate of the complexified (one-versus-all) SVM is 3.46%. In both learning tasks we used the SMO algorithm to train the SVM. The total amount of time needed to perform the training of each learning machine is almost the same for both cases (the complexified task is slightly faster).

In section 4, we discussed how the 4-classes problem comes naturally to the complex SVM. Exploiting the notion of the complex couple of hyperplanes (see figure 2), we have shown that the generalization of the SVM rationale to complex spaces directly assumes quaternary classification. Using this approach, the 4 classes problem can be solved using only 2 distinct SVM tasks instead of the 4 tasks needed by the 1-versus-all or the 1-versus-1 strategies. The second experiment compares the quaternary complex SVM approach to the standard 1-versus-all scenario using the first four digits (0, 1, 2 and 3). In both cases we used the first 6000 such digits of the MNIST training set to train the learning machines. We tested their performance using the digits contained in the testing set. The error rate of the 1-versus-all SVM was 0.721%, while the error rate of the complex SVM was 0.866%. However, the 1-versus-all SVM task required about double the time for training, compared to the complex SVM. This is expected, as the latter solves half as many distinct SVM tasks as the first one. In both experiments we used the gaussian kernel with $t = 1/49$ and $t = 1/160^2$ respectively. The SVM parameter C has been set equal to 100 in this case also.

6. CONCLUSIONS

We presented two generalized SVM frameworks suitable for binary and quaternary classification of data with complex inputs. We showed that in both cases this problem is equivalent to solving two standard real SVM tasks, albeit with a specific induced real kernel. In the first proposed framework (using pure complex kernels as the complex gaussian one) this induced kernel is not trivial. Finally, we presented two simple experiments that demonstrated the advantages obtained by exploiting the complex structure of the input data.

7. REFERENCES

[1] E.J. Bayro-Corrochano and N. Arana-Daniel, "Clifford support vector machines for classification, regression, and recurrence," *Neural Networks, IEEE Transactions on*, vol. 21, no. 11, pp. 1731 –1746, nov. 2010.

[2] A. Shilton and D.T.H. Lai, "Quaternionic and complex-valued support vector regression for equalization and function approximation," in *Neural Networks, 2007. IJCNN 2007. International Joint Conference on*, aug. 2007, pp. 920 –925.

[3] A. Shilton, D.T.H. Lai, and M. Palaniswami, "A division algebraic framework for multidimensional support vector regression," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 40, no. 2, pp. 517 –528, april 2010.

[4] P. Bouboulis, S. Theodoridis, and M. Mavroforakis, "The augmented complex kernel LMS," *Signal Processing, IEEE Transactions on*, vol. 60, no. 9, pp. 4962–4967, 2012.

[5] T. Adali and H. Li, *Adaptive signal processing: next generation solutions*, chapter Complex-valued Adaptive Signal Processing, pp. 1–74, Wiley, NJ, 2010.

[6] M. Novey and T. Adali, "On extending the complex ICA algorithm to noncircular sources," *IEEE Transanctions on Signal Processing*, vol. 56, no. 5, pp. 2148–2154, 2008.

[7] D. Mandic and V.S.L Goh, *Complex Valued nonlinear Adaptive Filters*, Wiley, 2009.

[8] A. Kuh and D. P. Mandic, "Applications of complex augmented kernels to wind profile prediction," *Proceedings of ICASSP*, 2009.

[9] P. Bouboulis, S. Theodoridis, and Ch. Mavroforakis, "Complex Support Vector Regression," *CIP*, 2012.

[10] P. Bouboulis and S. Theodoridis, "Extension of Wirtinger's calculus to reproducing kernel Hilbert spaces and the complex kernel LMS," *IEEE Transactions on Signal Processing*, vol. 59, no. 3, pp. 964–978, 2011.

[11] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American Mathematical Society*, vol. 68, no. 3, pp. 337–404, May 1950.

[12] W. Wirtinger, "Zur formalen theorie der functionen von mehr complexen veranderlichen," *Mathematische Annalen*, vol. 97, pp. 357–375, 1927.

[13] H. Li, *Complex-valued adaptive signal processing using Wirtinger calculus and its application to Independent Component Analysis*, Ph.D. thesis, University of Maryland Baltimore County, 2008.

[14] B. Picinbono and P. Chevalier, "Widely linear estimation with complex data," *IEEE Transactions on Signal Processing*, vol. 43, no. 8, pp. 2030–2033, 1995.

[15] P. Bouboulis, K. Slavakis, and S. Theodoridis, "Adaptive learning in complex reproducing kernel Hilbert spaces employing Wirtinger's subgradients," *Neural Networks and Learning Systems, IEEE Transactions on*, vol. 23, no. 3, pp. 425 –438, march 2012.

[16] B. Scholkopf and A.J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*, MIT Press, 2002.

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

[18] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, Academic Press, 4^{th} edition, Nov. 2008.

[19] Yann LeCun and Corinna Cortes, "The MNIST database," .

[20] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[21] D. Decoste and B. Schölkopf, "Training invariant support vector machines," *Machine Learning, 46, 161190, 2002*, vol. 46, pp. 161–190, 2002.