

EXTENSION OF WIRTINGER CALCULUS IN RKH SPACES AND THE COMPLEX KERNEL LMS

Pantelis Bouboulis, Sergios Theodoridis

*Department of Informatics and Telecommunications,
University of Athens,
Athens, Greece.*

{bouboulis, stheodor}@di.uoa.gr

ABSTRACT

Over the last decade, kernel methods for nonlinear processing have successfully been used in the machine learning community. However, so far, the emphasis has been on batch techniques. It is only recently, that online adaptive techniques have been considered in the context of signal processing tasks. To the best of our knowledge, no kernel-based strategy has been developed, so far, that is able to deal with complex valued signals. In this paper, we take advantage of a technique called *complexification* of real RKHSs to attack this problem. In order to derive gradients and subgradients of operators that need to be defined on the associated complex RKHSs, we employ the powerful tool of Wirtinger's Calculus, which has recently attracted much attention in the signal processing community. Wirtinger's calculus simplifies computations and offers an elegant tool for treating complex signals. To this end, in this paper, the notion of Wirtinger's calculus is extended, for the first time, to include complex RKHSs and use it to derive the Complex Kernel Least-Mean-Square (CKLMS) algorithm. Experiments verify that the CKLMS can be used to derive nonlinear stable algorithms, which offer significant performance improvements over the traditional complex LMS or Widely Linear complex LMS (WL-LMS) algorithms, when dealing with nonlinearities.

1. INTRODUCTION

Processing in Reproducing Kernel Hilbert Spaces (RKHSs) in the context of online adaptive processing is gaining in popularity within the Signal Processing community [1, 2, 3, 4, 5, 6]. The main advantage of mobilizing the tool of RKHSs is that the original nonlinear task is "transformed" into a linear one, where one can employ an easier "algebra". Moreover, different types of nonlinearities can be treated in a unifying way, that does not affect the derivation of the algorithms, except at the final implementation stage. The main concepts of this procedure can be summarized in the following two steps: 1) Map the finite dimensionality input data from the input space F (usually $F \subset \mathbb{R}^n$) into a higher dimensionality (possibly infinite) RKHS \mathcal{H} and 2) Perform a linear processing (e.g., adaptive filtering) on the mapped data in \mathcal{H} . The procedure is equivalent with a non-linear processing (non-linear filtering) in F .

An alternative way of describing this process is through the popular *kernel trick* [7, 8]: Given an algorithm, which is formulated in terms of dot products, one can construct an alternative algorithm by replacing each one of the dot products with a positive definite kernel κ . The specific choice of kernel implicitly

defines an RKHS with an appropriate inner product. Furthermore, the choice of kernel also defines the type of nonlinearity that underlies the model to be used. The main representatives of this class of algorithms are the celebrated *support vector machines* (SVMs), which have dominated the research in machine learning over the last decade. Besides SVMs and the more recent applications in adaptive filtering, there is a plethora of other scientific domains that have gained from adopting kernel methods (e.g., image processing and denoising [9, 10], principal component analysis [11], clustering [12], e.t.c.).

In this paper, we focus on the recently developed *Kernel Least Mean Squares Algorithm* (KLMS), which is the LMS algorithm in RKHSs [1, 13]. KLMS, as all the known kernel methods that use real-valued kernels, is able to deal with real valued data sequences only. To our knowledge, no kernel-based strategy has been developed, so far, that is able to effectively deal with complex valued signals. The main contributions of this paper are: a) the development of a wide framework that allows real-valued kernel algorithms to be extended to treat complex data effectively, taking advantage of a technique called *complexification* of real RKHSs, b) the extension of *Wirtinger's Calculus* in complex RKHSs as a means for the elegant and efficient computation of the gradients, that are involved in many adaptive filtering algorithms, and c) the development of the Complex Kernel LMS (CKLMS) algorithm, by exploiting the developed Wirtinger's calculus. Wirtinger's calculus [14] is enjoying increasing popularity, recently, mainly in the context of *Widely Linear* complex adaptive filters [15, 16, 17, 18, 19, 20, 21, 22], providing a tool for the derivation of gradients in the complex domain.

The paper is organized as follows. We start with a minimal introduction to RKHSs in Section 2, before we briefly review the KLMS algorithm in Section 3. In Section 4, we describe the complexification procedure of a real RKHS that provides the main framework for complex kernel methods, based on the popular real valued reproducing kernels (e.g., gaussian, polynomial, e.t.c.). The main notions of the extended Wirtinger's Calculus are summarized in Section 5 and the CKLMS is developed thereafter in Section 6. Finally, experimental results and conclusions are provided in Sections 7 and 8 respectively. We will denote the set of all integers, real and complex numbers by \mathbb{N} , \mathbb{R} and \mathbb{C} respectively. Vector or matrix valued quantities appear in boldfaced symbols.

2. REPRODUCING KERNEL HILBERT SPACES

We start with some basic definitions regarding RKHSs. Let X be a non empty set with $\mathbf{x}_1, \dots, \mathbf{x}_N \in X$. Consider a Hilbert space \mathcal{H} of real valued functions, f , defined on a set X , with a corresponding inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. We will call \mathcal{H} as a *Reproducing Kernel Hilbert Space* - RKHS, if there exists a function, known as kernel, $\kappa : X \times X \rightarrow \mathbb{R}$ with the following two properties:

1. For every $\mathbf{x} \in X$, $\kappa(\mathbf{x}, \cdot)$ belongs to \mathcal{H} .
2. κ has the so called *reproducing property*, i.e. $f(\mathbf{x}) = \langle f, \kappa(\mathbf{x}, \cdot) \rangle_{\mathcal{H}}$, for all $f \in \mathcal{H}$. In particular:

$$\kappa(\mathbf{x}, \mathbf{y}) = \langle \kappa(\mathbf{x}, \cdot), \kappa(\mathbf{y}, \cdot) \rangle_{\mathcal{H}}.$$

It can be shown that the kernel κ generates the entire space \mathcal{H} , i.e. $\mathcal{H} = \text{span}\{\kappa(\mathbf{x}, \cdot) | \mathbf{x} \in X\}$. There are several kernels that are used in practice (see [7]). Among the most widely used are the polynomial kernel: $\kappa(\mathbf{x}, \mathbf{y}) = (1 + \mathbf{x}^T \mathbf{y})^d$, $d \in \mathbb{N}$ and the gaussian kernel: $\kappa(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|^2 / \sigma^2)$, $\sigma > 0$.

Although there exist complex reproducing kernels that give rise to RKHSs of complex valued functions [23], in this paper we focus our attention on complexifying real valued ones, which have been extensively studied and contain several popular examples. Later on (in section 4), we will show how one can construct complex RKHSs from real ones, through a technique called complexification.

3. KERNEL LMS

In a typical LMS filter the goal is to learn a linear input output mapping $f : X \rightarrow \mathbb{R} : f(\mathbf{x}) = \mathbf{w}^T \mathbf{x}$, $X \subset \mathbb{R}^{\nu}$, based on a sequence of examples $(\mathbf{x}(1), d(1)), (\mathbf{x}(2), d(2)), \dots, (\mathbf{x}(N), d(N))$, so that to minimize the mean square error, $E[|d(n) - \mathbf{w}^T \mathbf{x}(n)|^2]$. To this end, the gradient descent rationale is employed and at each time instant, $n = 1, 2, \dots, N$, the gradient of $E[e(n)\mathbf{x}(n)]$ is estimated via its current measurement, i.e., $\hat{E}[e(n)\mathbf{x}(n)] = e(n)\mathbf{x}(n)$, where $e(n) = d(n) - \mathbf{w}(n-1)^T \mathbf{x}(n)$ is the error at instance $n = 2, \dots, N$. It takes a few lines of elementary algebra to deduce that the update of the unknown vector parameter is: $\mathbf{w}(n) = \mathbf{w}(n-1) + \mu e(n)\mathbf{x}(n)$, where μ is the step update. If we take the initial value of \mathbf{w} as $\mathbf{w}(0) = \mathbf{0}$, then the repeated application of the update equation yields:

$$\mathbf{w}(n) = \mu \sum_{k=1}^n e(k)\mathbf{x}(k) \quad (1)$$

Hence, for the filter output at instance n we have:

$$\hat{d}(n) = \mathbf{w}(n-1)^T \mathbf{x}(n) = \mu \sum_{k=1}^{n-1} e(k)\mathbf{x}(k)^T \mathbf{x}(n), \quad (2)$$

for $n = 1, 2, \dots, N$. Equation (2) is expressed in terms of inner products only, hence it allows for the application of the kernel trick. Thus, the filter output of the KLMS at instance n is

$$\hat{d}(n) = \langle \mathbf{x}(n), \mathbf{w}(n-1) \rangle = \mu \sum_{k=1}^{n-1} e(k)\kappa(\mathbf{x}(n), \mathbf{x}(k)), \quad (3)$$

$$\text{while } \mathbf{w}(n) = \mu \sum_{k=1}^n e(k)\kappa(\mathbf{x}(k), \cdot), \quad (4)$$

for $n = 1, 2, \dots, N$.

Another, more formal way of developing the KLMS is the following. First, we transform the input space X to a high dimensional feature space \mathcal{H} through the (implicit) mapping $\Phi : X \rightarrow \mathcal{H}$, $\Phi(\mathbf{x}) = \kappa(\mathbf{x}, \cdot)$. Thus, the training examples become

$$(\Phi(\mathbf{x}(1)), d(1)), \dots, (\Phi(\mathbf{x}(N)), d(N)).$$

We apply the LMS procedure on the transformed data, with the linear filter output $\hat{d}(n) = \langle \Phi(\mathbf{x}(n)), \mathbf{w} \rangle$. The model $\langle \Phi(\mathbf{x}), \mathbf{w} \rangle$ is more representative than the simple $\mathbf{w}^T \mathbf{x}$, since it includes the non-linear modeling through the presence of the kernel. The objective now becomes to minimize the cost function

$$E[|d(n) - \langle \Phi(\mathbf{x}(n)), \mathbf{w} \rangle|^2].$$

Using the notion of the Fréchet derivative, which has to be mobilized, since the dimensionality of the RKHS may be infinite, we are able to derive the gradient of the aforementioned cost function with respect to \mathbf{w} . It has to be emphasized, that now \mathbf{w} is not a vector, but a function, i.e., a point in the linear Hilbert space. It turns out that the update of the KLMS is given by $\mathbf{w}(n) = \mathbf{w}(n-1) + \mu e(n)\Phi(\mathbf{x}(n))$, where $e(n) = d(n) - \hat{d}(n)$. From this update, following the same procedure as in LMS and applying the reproducing property, we obtain equations (3) and (4), which are at the core of the KLMS algorithm. More details and the algorithmic implementation may be found in [13].

Note that, in a number of attempts to kernelize known algorithms, that are cast in inner products, the kernel trick is, usually, used in a "black box" rationale, without consideration of the problem in the RKH space, in which the (implicit) processing is carried out. Such an approach, often, does not allow for a deeper understanding of the problem, especially if a further theoretical analysis is required. Moreover, in our case, such a "blind" application of the kernel trick on a standard complex LMS form, can only lead to spaces defined by complex kernels. Complex RKH spaces, built around complexification of real kernels, do not result as a direct application of the standard kernel trick.

4. COMPLEXIFICATION OF A REAL RKHS

To generalize the kernel adaptive filtering algorithms on complex domains, we need a generalized framework regarding complex RKHSs. In this paper, we employ a simple technique called *complexification* of real RKHSs, which has the advantage of allowing modeling in complex RKHSs using popular well-established real kernels (e.g., gaussian, polynomial, e.t.c.).

Let $X \subseteq \mathbb{R}^{\nu}$. Define $X^2 = X \times X \subseteq \mathbb{R}^{2\nu}$ and $\mathbb{X} = \{\mathbf{x} + i\mathbf{y}, \mathbf{x}, \mathbf{y} \in X\}$ equipped with a complex product structure. Let \mathcal{H} be a real RKHS associated with a real kernel κ defined on $X^2 \times X^2$ and let $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ be its corresponding inner product. Then, every $f \in \mathcal{H}$ can be regarded as a function defined on either X^2 or \mathbb{X} , i.e., $f(\mathbf{z}) = f(\mathbf{x} + i\mathbf{y}) = f(\mathbf{x}, \mathbf{y})$.

Next, we define $\mathcal{H}^2 = \mathcal{H} \times \mathcal{H}$. It is easy to verify that \mathcal{H}^2 is also a Hilbert Space with inner product

$$\langle \mathbf{f}, \mathbf{g} \rangle_{\mathcal{H}^2} = \langle f_1, g_1 \rangle_{\mathcal{H}} + \langle f_2, g_2 \rangle_{\mathcal{H}}, \quad (5)$$

for $\mathbf{f} = (f_1, f_2)^T$, $\mathbf{g} = (g_1, g_2)^T$. Our objective is to enrich \mathcal{H}^2 with a complex structure. We address this problem using the complexification of the real RKHS \mathcal{H} . To this end, we define the

space $\mathbb{H} = \{\mathbf{f} = f_1 + if_2; f_1, f_2 \in \mathcal{H}\}$ equipped with the complex inner product:

$$\langle \mathbf{f}, \mathbf{g} \rangle_{\mathbb{H}} = \langle f_1, g_1 \rangle_{\mathcal{H}} + \langle f_2, g_2 \rangle_{\mathcal{H}} + i(\langle f_2, g_1 \rangle_{\mathcal{H}} - \langle f_1, g_2 \rangle_{\mathcal{H}}),$$

for $\mathbf{f} = f_1 + if_2$, $\mathbf{g} = g_1 + ig_2$. It is not difficult to verify that \mathbb{H} is a complex RKHS with kernel κ [23]. We call \mathbb{H} the complexification of \mathcal{H} .

To complete the presentation of the required framework for working on complex RKHSs, we need a technique to map the samples data from the complex input space to the complexified RKHS \mathbb{H} . This problem will be addressed in section 6.

5. WIRTINGER'S CALCULUS IN COMPLEX RKHS

Wirtinger's calculus [14] has become very popular in the signal processing community mainly in the context of complex adaptive filtering, as a means of computing, in an elegant way, gradients of real valued cost functions defined on complex domains (\mathbb{C}^ν). Such functions, obviously, are not holomorphic and therefore the complex derivative cannot be used. Instead, if we consider that the cost function is defined on a Euclidean domain with a double dimensionality ($\mathbb{R}^{2\nu}$), then the real derivatives may be employed. The price of this approach is that the computations become cumbersome and tedious. Wirtinger's calculus provides an alternative equivalent formulation, that is based on simple rules and principles and which bear a great resemblance to the rules of the standard complex derivative.

In the case of a simple non-holomorphic complex function T defined on $U \subseteq \mathbb{C}$, Wirtinger's calculus considers two forms of derivatives, the *\mathbb{R} -derivative* and the *conjugate \mathbb{R} -derivative*, which are defined as follows:

$$\frac{\partial T}{\partial z} = \frac{1}{2} \left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} \right) + \frac{i}{2} \left(\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} \right),$$

$$\frac{\partial T}{\partial z^*} = \frac{1}{2} \left(\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y} \right) + \frac{i}{2} \left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y} \right)$$

where $T(z) = T(x + iy) = T(x, y) = u(x, y) + iv(x, y)$. Note that any such non-holomorphic function can be written in the form $T(z, z^*)$, so that for fixed z^* , T is z -holomorphic and for fixed z , T is z^* -holomorphic [24] (assuming of course that $T(x, y)$ has partial derivatives of any order). This fact underlies the development of Wirtinger's calculus. Having this in mind, $\frac{\partial T}{\partial z}$, can be easily evaluated as the standard complex partial derivative taken with respect to z (thus treating z^* as a constant). Consequently, $\frac{\partial T}{\partial z^*}$ is evaluated as the standard complex partial derivative taken with respect to z^* (thus treating z as a constant). For example, if $T(z, z^*) = z(z^*)^2$, then

$$\frac{\partial T}{\partial z} = (z^*)^2, \quad \frac{\partial T}{\partial z^*} = 2zz^*.$$

Similar principles and rules hold for a function of many complex variables (i.e., $U \subseteq \mathbb{C}^\nu$) [24].

Wirtinger's calculus has been developed only for operators defined on finite dimensional spaces, \mathbb{C}^ν . Hence, this calculus cannot be used in RKH spaces, where the dimensionality of the function space can be infinite. To this end, Wirtinger's calculus needs to be generalized to a general Hilbert space, and this is one of the main contributions of the current paper. A rigorous presentation of this

extension is out of the scope of the paper (due to lack of space). Nevertheless, we will present the main ideas and results. At the heart of the generalization lies the notion of the Fréchet differentiability. Consider a Hilbert space H over the field F (typically \mathbb{R} or \mathbb{C}). The operator $T : H \rightarrow F$ is said to be *Fréchet differentiable* at f_0 , if there exists a $u \in H$, such that

$$\lim_{\|h\|_H \rightarrow 0} \frac{T(f_0 + h) - T(f_0) - \langle u, h \rangle_H}{\|h\|_H} = 0, \quad (6)$$

where $\langle \cdot, \cdot \rangle_H$ is the dot product of the Hilbert space H and $\|\cdot\|_H = \sqrt{\langle \cdot, \cdot \rangle_H}$ is the induced norm. The element u is usually called the gradient of T at f_0 .

Since our study involves mainly RKHS, we will present the necessary tools in that context. The generalization to a general Hilbert space has also been developed and follows a similar path. Consider the spaces \mathbb{H} and \mathcal{H}^2 defined in section 4. Let $\mathbf{T} : \mathbb{H} \rightarrow \mathbb{C}$, $\mathbf{T} = T_1 + iT_2$ be the operator we seek to differentiate. Assume that $\mathbf{T} = (T_1, T_2)^T$, $\mathbf{T}(\mathbf{f}) = \mathbf{T}(f_1 + if_2) = \mathbf{T}(f_1, f_2) = T_1(f_1, f_2) + iT_2(f_1, f_2)$, is differentiable as an operator defined on \mathcal{H}^2 and let $\nabla_1 T_1$, $\nabla_2 T_1$, $\nabla_1 T_2$ and $\nabla_2 T_2$ be the partial derivatives, with respect to the first (f_1) and the second (f_2) variable respectively. It turns out, proof is omitted due to lack of space, that if $\mathbf{T}(f_1, f_2)$ has derivatives of any order, then it can be written in the form $\mathbf{T}(\mathbf{f}, \mathbf{f}^*)$, where $\mathbf{f}^* = f_1 - if_2$, so that for fixed \mathbf{f}^* , \mathbf{T} is \mathbf{f} -holomorphic and for fixed \mathbf{f} , \mathbf{T} is \mathbf{f}^* -holomorphic. We may define the \mathbb{R} -derivative and the conjugate \mathbb{R} -derivative of \mathbf{T} as follows:

$$\nabla_{\mathbf{f}} \mathbf{T} = \frac{1}{2} (\nabla_1 T_1 + \nabla_2 T_2) + \frac{i}{2} (\nabla_1 T_2 - \nabla_2 T_1) \quad (7)$$

$$\nabla_{\mathbf{f}^*} \mathbf{T} = \frac{1}{2} (\nabla_1 T_1 - \nabla_2 T_2) + \frac{i}{2} (\nabla_1 T_2 + \nabla_2 T_1). \quad (8)$$

The following properties can be proved (among others):

1. if \mathbf{T} is \mathbf{f} -holomorphic (i.e., it has a Taylor series expansion with respect to \mathbf{f}), then $\nabla_{\mathbf{f}^*} \mathbf{T} = \mathbf{0}$.
2. if \mathbf{T} is \mathbf{f}^* -holomorphic (i.e., it has a Taylor series expansion with respect to \mathbf{f}^*), then $\nabla_{\mathbf{f}} \mathbf{T} = \mathbf{0}$.
3. $(\nabla_{\mathbf{f}} \mathbf{T})^* = \nabla_{\mathbf{f}^*} \mathbf{T}^*$.
4. $(\nabla_{\mathbf{f}^*} \mathbf{T})^* = \nabla_{\mathbf{f}} \mathbf{T}^*$.
5. If \mathbf{T} is real valued, then $(\nabla_{\mathbf{f}} \mathbf{T})^* = \nabla_{\mathbf{f}^*} \mathbf{T}$.
6. The first order Taylor expansion around $\mathbf{f} \in \mathbb{H}$ is given by
$$\begin{aligned} \mathbf{T}(\mathbf{f} + \mathbf{h}) = & \mathbf{T}(\mathbf{f}) + \langle \mathbf{h}, (\nabla_{\mathbf{f}} \mathbf{T}(\mathbf{f}))^* \rangle_{\mathbb{H}} \\ & + \langle \mathbf{h}^*, (\nabla_{\mathbf{f}^*} \mathbf{T}(\mathbf{f}))^* \rangle_{\mathbb{H}}. \end{aligned}$$
7. If $\mathbf{T}(\mathbf{f}) = \langle \mathbf{f}, \mathbf{w} \rangle_{\mathbb{H}}$, then $\nabla_{\mathbf{f}} \mathbf{T} = \mathbf{w}^*$, $\nabla_{\mathbf{f}^*} \mathbf{T} = \mathbf{0}$.
8. If $\mathbf{T}(\mathbf{f}) = \langle \mathbf{w}, \mathbf{f} \rangle_{\mathbb{H}}$, then $\nabla_{\mathbf{f}} \mathbf{T} = \mathbf{0}$, $\nabla_{\mathbf{f}^*} \mathbf{T} = \mathbf{w}$.
9. If $\mathbf{T}(\mathbf{f}) = \langle \mathbf{f}^*, \mathbf{w} \rangle_{\mathbb{H}}$, then $\nabla_{\mathbf{f}} \mathbf{T} = \mathbf{0}$, $\nabla_{\mathbf{f}^*} \mathbf{T} = \mathbf{w}^*$.
10. If $\mathbf{T}(\mathbf{f}) = \langle \mathbf{w}, \mathbf{f}^* \rangle_{\mathbb{H}}$, then $\nabla_{\mathbf{f}} \mathbf{T} = \mathbf{w}$, $\nabla_{\mathbf{f}^*} \mathbf{T} = \mathbf{0}$.
11. If $\mathbf{R}, \mathbf{S} : \mathbb{H} \rightarrow \mathbb{C}$ are \mathbf{f} -analytic and $\mathbf{T} = \mathbf{R} \cdot \mathbf{S}$ then:

$$\nabla_{\mathbf{f}} \mathbf{T} = \nabla_{\mathbf{f}} \mathbf{R} \cdot \mathbf{S} + \nabla_{\mathbf{f}} \mathbf{S} \cdot \mathbf{R}.$$

An important consequence of the above properties is that if \mathbf{T} is a real valued operator defined on \mathbb{H} , then its first order Taylor's expansion is given by:

$$\begin{aligned} \mathbf{T}(\mathbf{f} + \mathbf{h}) = & \mathbf{T}(\mathbf{f}) + \langle \mathbf{h}, (\nabla_{\mathbf{f}} \mathbf{T}(\mathbf{f}))^* \rangle_{\mathbb{H}} + \langle \mathbf{h}^*, (\nabla_{\mathbf{f}^*} \mathbf{T}(\mathbf{f}))^* \rangle_{\mathbb{H}} \\ = & \mathbf{T}(\mathbf{f}) + \langle \mathbf{h}, \nabla_{\mathbf{f}^*} \mathbf{T}(\mathbf{f}) \rangle_{\mathbb{H}} + \langle \langle \mathbf{h}, \nabla_{\mathbf{f}^*} \mathbf{T}(\mathbf{f}) \rangle_{\mathbb{H}}^* \rangle_{\mathbb{H}} \\ = & \mathbf{T}(\mathbf{f}) + 2 \cdot \Re [\langle \mathbf{h}, \nabla_{\mathbf{f}^*} \mathbf{T}(\mathbf{f}) \rangle_{\mathbb{H}}]. \end{aligned}$$

However, in view of the Cauchy Riemann inequality we have:

$$\begin{aligned} \Re[\langle \mathbf{h}, \nabla_{\mathbf{f}^*} \mathbf{T}(\mathbf{f}) \rangle_{\mathbb{H}}] &\leq |\langle \mathbf{h}, \nabla_{\mathbf{f}^*} \mathbf{T}(\mathbf{f}) \rangle_{\mathbb{H}}| \\ &\leq \|\mathbf{h}\|_{\mathbb{H}} \cdot \|\nabla_{\mathbf{f}^*} \mathbf{T}(\mathbf{f})\|_{\mathbb{H}}. \end{aligned}$$

The equality in the above relationship holds if $h \propto \nabla_{\mathbf{f}^*} \mathbf{T}$. Hence, the direction of increase of \mathbf{T} is $\nabla_{\mathbf{f}^*} \mathbf{T}(\mathbf{f})$. Therefore, any gradient descent based algorithm minimizing $\mathbf{T}(\mathbf{f})$ is based on the update scheme:

$$\mathbf{f}_n = \mathbf{f}_{n-1} - \mu \cdot \nabla_{\mathbf{f}^*} \mathbf{T}(\mathbf{f}_{n-1}). \quad (9)$$

6. COMPLEX KERNEL LMS

Consider the sequence of examples $(\mathbf{z}(1), d(1)), (\mathbf{z}(2), d(2)), \dots, (\mathbf{z}(N), d(N))$, where $d(n) \in \mathbb{C}$, $\mathbf{z}(n) \in V \subset \mathbb{C}^{\nu}$, $\mathbf{z}(n) = \mathbf{x}(n) + i\mathbf{y}(n)$, $\mathbf{x}(n), \mathbf{y}(n) \in \mathbb{R}^{\nu}$, for $n = 1, \dots, N$. We map the points $\mathbf{z}(n)$ to the RKHS \mathbb{H} using the mapping Φ :

$$\begin{aligned} \Phi(\mathbf{z}(n)) &= \Phi(\mathbf{z}(n)) + i\Phi(\mathbf{z}(n)) \\ &= \kappa\left(\left(\mathbf{x}(n), \mathbf{y}(n)\right)^T, \cdot\right) + i \cdot \kappa\left(\left(\mathbf{x}(n), \mathbf{y}(n)\right)^T, \cdot\right), \end{aligned}$$

for $n = 1, \dots, N$. The objective of the complex Kernel LMS is to minimize $E[\mathcal{L}_n(\mathbf{w})]$, where

$$\begin{aligned} \mathcal{L}_n(\mathbf{w}) &= |e(n)|^2 = |d(n) - \langle \Phi(\mathbf{z}(n)), \mathbf{w} \rangle_{\mathbb{H}}|^2 \\ &= (d(n) - \langle \Phi(\mathbf{z}(n)), \mathbf{w} \rangle_{\mathbb{H}}) (d(n) - \langle \Phi(\mathbf{z}(n)), \mathbf{w} \rangle_{\mathbb{H}})^* \\ &= (d(n) - \langle \mathbf{w}^*, \Phi(\mathbf{z}(n)) \rangle_{\mathbb{H}}) (d(n)^* - \langle \mathbf{w}, \Phi(\mathbf{z}(n)) \rangle_{\mathbb{H}}), \end{aligned}$$

at each instance n . We then apply the complex LMS to the transformed data, using the rules of Wirtinger's calculus to compute the gradient $\nabla_{\mathbf{w}^*} \mathcal{L}_n(\mathbf{w}) = -e(n)^* \cdot \Phi(\mathbf{z}(n))$. Therefore the CKLMS update rule becomes:

$$\mathbf{w}(n) = \mathbf{w}(n-1) + \mu e(n)^* \cdot \Phi(\mathbf{z}(n)), \quad (10)$$

where $\mathbf{w}(n)$ denotes the estimate at iteration n .

Assuming that $\mathbf{w}(0) = \mathbf{0}$, the repeated application of the weight-update equation gives:

$$\begin{aligned} \mathbf{w}(n) &= \mathbf{w}(n-1) + \mu e(n)^* \Phi(\mathbf{z}(n)) \\ &= \mathbf{w}(n-2) + \mu e(n-1)^* \Phi(\mathbf{z}(n-1)) \\ &\quad + \mu e(n)^* \Phi(\mathbf{z}(n)) \\ &= \sum_{k=1}^n e(k)^* \Phi(\mathbf{z}(k)). \end{aligned} \quad (11)$$

Thus, the filter output at iteration n becomes:

$$\begin{aligned} \hat{d}(n) &= \langle \Phi(\mathbf{z}(n)), \mathbf{w}(n-1) \rangle_{\mathbb{H}} \\ &= \mu \sum_{k=1}^{n-1} e(k) \langle \Phi(\mathbf{z}(n)), \Phi(\mathbf{z}(k)) \rangle_{\mathbb{H}} \\ &= 2\mu \sum_{k=1}^{n-1} e(k) \kappa(\mathbf{z}(n), \mathbf{z}(k)) \\ &= 2\mu \sum_{k=1}^{n-1} \Re[e(n)] \kappa(\mathbf{z}(n), \mathbf{z}(k)) \\ &\quad + 2\mu \cdot i \sum_{k=1}^{n-1} \Im[e(n)] \kappa(\mathbf{z}(n), \mathbf{z}(k)), \end{aligned} \quad (12)$$

where the evaluation of the kernel is done by replacing the complex vectors $\mathbf{z}(n)$, of \mathbb{C}^{ν} with the corresponding real vectors of $\mathbb{R}^{2\nu}$, i.e.,

$$\mathbf{z}(n) = \mathbf{x}(n) + i\mathbf{y}(n) = (\mathbf{x}(n)^T, \mathbf{y}(n)^T)^T.$$

It can readily be shown that, since the CKLMS is the complex LMS in RKHS, the important properties of the LMS (convergence in the mean, misadjustment, e.t.c.) carry over to CKLMS. Furthermore, we may also define a normalized version, which we call *Normalized Complex Kernel LMS* (NCKLMS). The weight-update of the NCKLMS is given by:

$$\mathbf{w}(n) = \mathbf{w}(n-1) + \frac{\mu}{2 \cdot \kappa(\mathbf{z}(n), \mathbf{z}(n))} e(n)^* \Phi(\mathbf{z}(n))$$

The NCKLMS algorithm is summarized in Algorithm 1.

Algorithm 1 Normalized Complex Kernel LMS

INPUT: $(\mathbf{z}(1), d(1)), \dots, (\mathbf{z}(N), d(N))$

OUTPUT: The expansion

$$\mathbf{w} = \sum_{k=1}^N a(k) \kappa(\mathbf{z}(k), \cdot) + i \cdot \sum_{k=1}^N b(k) \kappa(\mathbf{z}(k), \cdot).$$

Initialization: Set $\mathbf{a} = \{\}$, $\mathbf{b} = \{\}$, $\mathbf{Z} = \{\}$ (i.e., $\mathbf{w} = \mathbf{0}$). Select the step parameter μ and the kernel κ .

for $n=1:N$ **do**

 Compute the filter output:

$$\begin{aligned} \hat{d}(n) &= \sum_{k=1}^{n-1} (a(k) + b(k)) \cdot \kappa(\mathbf{z}(n), \mathbf{z}(k)) \\ &\quad + \sum_{k=1}^{n-1} (a(k) - b(k)) \cdot \kappa(\mathbf{z}(n), \mathbf{z}(k)). \end{aligned}$$

 Compute the error: $e(n) = d(n) - \hat{d}(n)$.

$$\gamma = 2\kappa(\mathbf{z}(n), \mathbf{z}(n)).$$

$$a(n) = \mu(\Re[e(n)] + \Im[e(n)])/\gamma.$$

$$b(n) = \mu(\Re[e(n)] - \Im[e(n)])/\gamma.$$

 Add the new center $\mathbf{z}(n)$ to the list of centers, i.e., add $\mathbf{z}(n)$ to the list \mathbf{Z} , add $a(n)$ to the list \mathbf{a} , add $b(n)$ to the list \mathbf{b} .

end for

6.1. Sparsification

The main drawback of kernel based adaptive filtering algorithms is that they require a growing network of training centers \mathbf{z}_n . They start from an empty set (usually called the *dictionary*) and gradually add new samples to that set, to form a summation similar to the one shown in equation (11). This results to an increasing memory and computational requirements, as time evolves. Several strategies have been proposed to cope with this problem and to produce sparse solutions. In this paper, we employ the well known *novelty criterion* [25, 13]. In novelty criterion online sparsification, whenever a new data pair $(\Phi(\mathbf{z}_n), d_n)$ is considered, a decision is immediately made of whether to add the new center $\Phi(\mathbf{z}_n)$ to the dictionary of centers \mathcal{C} . The decision is reached following two simple rules. First, the distance of the new center $\Phi(\mathbf{z}_n)$ from the current dictionary is evaluated: $dis = \min_{\mathbf{c}_k \in \mathcal{C}} \{\|\Phi(\mathbf{z}_n) - \mathbf{c}_k\|_{\mathbb{H}}\}$. If this distance is smaller than a given threshold δ_1 (i.e., the new center is close to the existing dictionary), then the center is not added

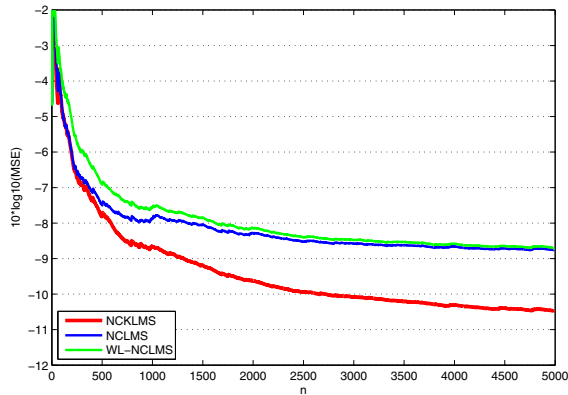


Fig. 1. Learning curves for KCLMS, ($\mu = 1/2$) CLMS ($\mu = 1/16$) and WL-CLMS ($\mu = 1/16$) (filter length $L = 5$, delay $D = 2$) in the nonlinear channel equalization, for the circular input case.

to \mathcal{C} . Otherwise, we compute the prediction error $e_n = d_n - \hat{d}_n$. If $|e_n|$ is smaller than a predefined threshold δ_2 , then the new center is discarded. Only if $|e_n| \geq \delta_2$ the new center $\Phi(z_n)$ is added to the dictionary.

Besides the previous scenario, other scenarios are also possible, that keep the number updated parameters, per recursion, fixed. For example, the sliding window LMS can be used. In [4, 5, 6], regularization, in the form of projections, has been used to cope efficiently with the problem. Results under such scenarios are available and will be presented elsewhere.

7. EXPERIMENTS

We tested the CKLMS on a nonlinear channel equalization problem (see figure 3). The nonlinear channel consists of a linear filter:

$$t(n) = (-0.9 + 0.8i) \cdot s(n) + (0.6 - 0.7i) \cdot s(n-1)$$

and a memoryless nonlinearity

$$q(n) = t(n) + (0.1 + 0.15i) \cdot t^2(n) + (0.06 + 0.05i) \cdot t^3(n).$$

At the receiver end of the channel, the signal is corrupted by white Gaussian noise and then observed as $r(n)$. The input signal that was fed to the channel had the form

$$s(n) = 0.70(\sqrt{1 - \rho^2}X(n) + i\rho Y(n)), \quad (13)$$

where $X(n)$ and $Y(n)$ are gaussian random variables. This input is circular for $\rho = \sqrt{2}/2$ and highly non-circular if ρ approaches 0 or 1 [17]. The aim of channel equalization is to construct an inverse filter which taking the output $r(n)$, reproduces the original input signal with as low an error rate as possible. To this end we apply the NCKLMS algorithm to the set of samples

$$\left((r(n+D), r(n+D-1), \dots, r(n+D-L))^T, s(n) \right),$$

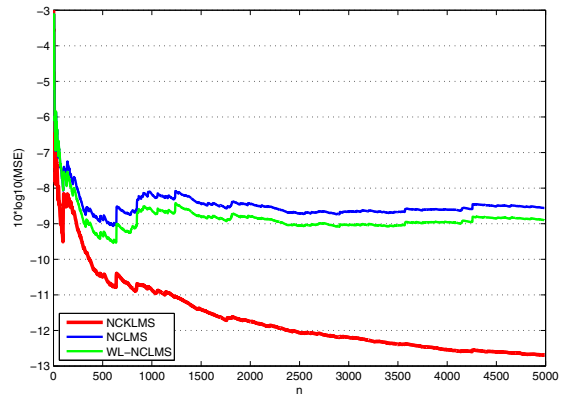


Fig. 2. Learning curves for KNCLMS ($\mu = 1/2$), NCLMS ($\mu = 1/16$) and WL-NCLMS ($\mu = 1/16$) (filter length $L = 5$, delay $D = 2$) in the nonlinear channel equalization, for the non-circular input case ($\rho = 0.1$).

where $L > 0$ is the filter length and D the equalization time delay.

Experiments were conducted on a set of 5000 samples of the input signal (13) considering both the circular and the non-circular case. The results are compared with the NCLMS and the WL-NCLMS algorithms. In all algorithms the step update parameter μ is tuned for best possible results. Time delay D was also set for optimality. Figures 1 and 2 show the learning curves of the NCKLMS using the Gaussian kernel $\kappa(x, y) = \exp(-\|x - y\|^2/\sigma^2)$ (with $\sigma = 5$), compared with the NCLMS and the WL-NCLMS algorithms. Novelty criterion was applied to the NCKLMS for sparsification with $\delta_1 = 0.15$ and $\delta_2 = 0.2$. In both examples, NCKLMS considerably outperforms both the NCLMS and the WL-NCLMS algorithms. However, this enhanced behavior comes at a price in computational complexity, since the NCKLMS requires the evaluation of the kernel function on a growing number of training examples.

8. CONCLUSIONS

A new framework for kernel adaptive filtering for complex signal processing was developed. The proposed methodology employs a technique called complexification of RKHSs to construct complex RKHSs from real ones, providing the advantage of working with some popular real kernels in the complex domain. It has to be pointed out, that our method is a general one and can be used on any type of complex kernels that have or can be developed. To the best of our knowledge, this is the first time that a methodology for complex adaptive processing in RKHSs is proposed. Wirtinger's calculus has been extended to cope with the problem of differentiation in the involved (infinite) dimensional Hilbert spaces. The derived rules and properties of the extended Wirtinger's calculus on complex RKHS turn out to be similar in structure to the special case of finite dimensional complex spaces. The proposed framework was applied on the complex LMS and the new complex Kernel LMS algorithm was developed. Experiments, which were performed on the equalization problem of a nonlinear channel for both circular and non-circular input data, showed a sig-

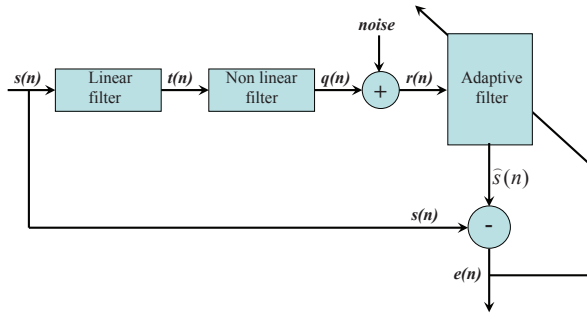


Fig. 3. The equalization problem.

nificant decrease in the steady state mean square error, compared with complex LMS and widely linear complex LMS.

9. REFERENCES

- [1] W. Liu, P. Pokharel, and J. C. Principe, "The kernel least-mean-square algorithm," *IEEE Trans. Sign. Proc.*, vol. 56, no. 2, pp. 543–554, 2008.
- [2] J. Kivinen, A. Smola, and R. C. Williamson, "Online learning with kernels," *IEEE Trans. Sign. Proc.*, vol. 52, no. 8, pp. 2165–2176, 2004.
- [3] Y. Engel, S. Mannor, and R. Meir, "The kernel recursive least-squares algorithm," *IEEE Trans. Sign. Proc.*, vol. 52, no. 8, 2004.
- [4] K. Slavakis, S. Theodoridis, and I. Yamada, "On line classification using kernels and projection based adaptive algorithm," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 2781–2797, 2008.
- [5] K. Slavakis, S. Theodoridis, and I. Yamada, "Adaptive constrained learning in reproducing kernel hilbert spaces: The robust beamforming case," *IEEE Trans. Signal Process.*, vol. 57, no. 12, pp. 4744–4764, 2009.
- [6] K. Slavakis and S. Theodoridis, "Sliding window generalized kernel affine projection algorithm using projection mappings," *Eurasip Journal on Advances in Signal Processing*, vol. art. no. 735351, 2008.
- [7] B. Schölkopf and A. J. Smola, *Learning with Kernels*, MIT Press, 2002.
- [8] S. Theodoridis and K. Koutroumbas, *Pattern Recognition, 4th edition*, Academic Press, 2009.
- [9] K. Kim, M. O. Franz, and B. Scholkopf, "Iterative kernel principal component analysis for image modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 9, pp. 1351–1366, 2005.
- [10] P. Bouboulis, K. Slavakis, and S. Theodoridis, "Adaptive kernel-based image denoising employing semi-parametric regularization," *IEEE Trans. Image Process.*, (to appear).
- [11] A. J. Smola B. Schölkopf and K. R. Muller, "Kernel principal component analysis," *Lecture notes in computer science*, vol. 1327, pp. 583–588, 1997.
- [12] M. Filippone, F. Camastra, F. Masulli, and S. Rovetta, "A survey of kernel and spectral methods for clustering. pattern recognition," *Pattern Recognition*, vol. 41, no. 1, pp. 176–190, 2008.
- [13] W. Liu, J. C. Principe, and S. Haykin, *Kernel Adaptive Filtering*, Wiley, 2010.
- [14] W. Wirtinger, "Zur formalen theorie der functionen von mehr complexen veränderlichen," *Math. Ann.*, vol. 97, pp. 357–375, 1927.
- [15] B. Picinbono and P. Chevalier, "Widely linear estimation with complex data," *IEEE Trans. Signal Process.*, vol. 43, no. 8, pp. 2030–2033, 1995.
- [16] D. Mandic and V. S. L. Goh, *Complex Valued Nonlinear Adaptive Filters*, Wiley, 2009.
- [17] T. Adali and H. Li, *Complex-valued adaptive signal processing*, Adaptive Signal Processing: Next Generation Solutions, T. Adali and S. Haykin, editors. Hoboken, NJ, Wiley, 2010.
- [18] T. Adali, H. Li, M. Novey, and J.F. Cardoso, "Complex ICA using nonlinear functions," *IEEE Trans. Signal Process.*, vol. 56, no. 9, pp. 4536–4544, 2008.
- [19] M. Novey and T. Adali, "On extending the complex fast ICA algorithm to noncircular sources," *IEEE Trans. Signal Process.*, vol. 56, no. 5, pp. 2148–2154, 2008.
- [20] D. Mattered, L. Paura, and F. Sterle, "Widely linear decision-feedback equalizer for time-dispersive linear MIMO channels," *IEEE Trans. Signal Process.*, vol. 53, no. 7, pp. 2525–2536, 2005.
- [21] A. S. Cacciapuoti, G. Gelli, L. Paura, and F. Verde, "Widely linear versus linear blind multiuser detection with subspace-based channel estimation: Finite sample-size effects," *IEEE Trans. Signal Process.*, vol. 57, no. 4, pp. 1426–1443, 2009.
- [22] J. Navarro-Moreno, "ARMA prediction of widely linear systems by using the innovations algorithm," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 3061–3068, 2008.
- [23] V. I. Paulsen, "An introduction to the theory of reproducing kernel hilbert spaces," <http://www.math.uh.edu/~vern/rkhs.pdf>.
- [24] K. Kreutz-Delgado, "The complex gradient operator and the $\mathbb{C}\mathbb{R}$ -calculus," <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.86.6515&rep=rep1&type=pdf>.
- [25] J. Platt, "A resource allocating network for function interpolation," *Neural Computation*, vol. 3, no. 2, pp. 213–225, 1991.