On-line Nonlinear Sparse Approximation of Functions

Paul Honeine
Institut Charles Delaunay (FRE CNRS 2848)
Université de technologie de Troyes
BP 2060,10010 Troyes, France
paul.honeine@utt.fr

Cédric Richard
Institut Charles Delaunay (FRE CNRS 2848)
Université de technologie de Troyes
BP 2060,10010 Troyes, France
cedric.richard@utt.fr

José Carlos M. Bermudez
Department of Electrical Engineering
Federal University of Santa Catarina
88040-900, Florianópolis, SC - Brazil
j.bermudez@ieee.org

***Abstract*— This paper provides new insights into on-line nonlinear sparse approximation of functions based on the coherence criterion. We revisit previous work, and propose tighter bounds on the approximation error based on the coherence criterion. Moreover, we study the connections between the coherence criterion and both the approximate linear dependence criterion and the principal component analysis. Finally, we derive a kernel normalized LMS algorithm based on the coherence criterion, which has linear computational complexity on the model order. Initial experimental results are presented on the performance of the algorithm.**

I. Introduction

Over the last decades, sparse approximation of functions has become a commonly used tool for a wide variety of problems involving dynamic systems. Although most of the work done in this field applies linear methods, many situations require nonlinear processing of data. This can be done using the formalism of reproducing kernel Hilbert spaces (RKHS). Initially proposed in [1], [2], the latter has gained wide popularity in recent years with kernel-based methods such as support vector machines. A common characteristic of these techniques is that they deal with series expansions whose size equals the number of training data, making them unsuitable for on-line applications. To overcome this, sparsification techniques have been proposed [3], [4] to control the model order. In spite of an effective order control, these techniques suffer from high computational complexity.

In [5], we presented a framework for on-line nonlinear sparse approximation of functions based on RKHS. The sparsification technique has only linear complexity in the order of the model. It is based on the coherence parameter, a fundamental quantity for characterizing dictionaries of functions [6], [7]. This paper extends that framework by providing new properties of the coherence criterion and connections to other sparsification techniques. Moreover, we present the kernel normalized LMS (KNLMS) adaptive filtering algorithm, whose complexity is linear in the model order, as opposed to the kernel recursive least-squares (KRLS) algorithms proposed in [4], [5] which have a quadratic complexity.

This paper is organized as follows. In Section 2, we outline some basic principles of nonlinear filtering in RKHS. In Section 3, we present the coherence parameter. Its properties and connections to other sparsification criteria are investigated in Section 4. Finally, we propose the new coherence criterion based KNLMS algorithm for on-line nonlinear approximation of functions, and we evaluate its performance.

II. Foundations of nonlinear filtering in RKHS

Let $\mathcal{U}$ be a compact subspace of $\mathbb{R}^p$, $\kappa : \mathcal{U} \times \mathcal{U} \to \mathbb{R}$ a reproducing kernel, and $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ the induced RKHS with its inner product. The reproducing property states that any function $\psi(\cdot)$ of $\mathcal{H}$ can be evaluated at any point $\boldsymbol{u}_i$ of $\mathcal{U}$ using $\psi(\boldsymbol{u}_i) = \langle \psi(\cdot), \kappa(\cdot, \boldsymbol{u}_i)\rangle_{\mathcal{H}}$, where $\kappa(\cdot, \boldsymbol{u}_i)$ is a positive definite kernel that takes $\boldsymbol{u}_\ell$ into $\kappa(\boldsymbol{u}_\ell, \boldsymbol{u}_i)$. By setting $\mathcal{H}$ as the hypothesis space, we consider as a cost function the squared error between the model outputs $\psi(\boldsymbol{u}_i)$ and the desired responses d_i, that is,

$$\sum_{i=1}^{n} (d_i - \psi(\boldsymbol{u}_i))^2. \tag{1}$$

It is well known from the representer theorem [2] that the solution to such optimization problems can be expressed as a kernel expansion in terms of available training data, namely,

$$\psi(\cdot) = \sum_{j=1}^{n} \alpha_j \, \kappa(\cdot, \boldsymbol{u}_j).$$

The optimization problem is then reduced to the dual problem of determining $\boldsymbol{\alpha} = [\alpha_1 \ldots \alpha_n]^t$ such that

$$\min_{\boldsymbol{\alpha}} \|\boldsymbol{d} - \boldsymbol{K}\boldsymbol{\alpha}\|^2,$$

where $\boldsymbol{K}$ denotes the Gram matrix whose (i,j)-th entry is $\kappa(\boldsymbol{u}_i, \boldsymbol{u}_j)$, and $\boldsymbol{d} = [d_1 \ldots d_n]^t$. Solution to this problem is given by $\boldsymbol{\alpha} = \boldsymbol{K}^\dagger \boldsymbol{d}$, where $\boldsymbol{K}^\dagger$ is the pseudo-inverse of $\boldsymbol{K}$. Since the model order is equal to the number n of available data $\boldsymbol{u}_i$, this approach cannot be considered for on-line applications.

To overcome this barrier, one can control the order of the kernel expansion by considering, at each time instant n, the reduced model

$$\psi_n(\cdot) = \sum_{\omega_j \in \mathcal{J}_n} \alpha_{n,j} \, \kappa(\cdot, \boldsymbol{u}_{\omega_j}), \tag{2}$$

where $\mathcal{J}_n$ is a subset of m indices of $\{1, \ldots, n\}$. The m kernel functions $\kappa(\cdot, \boldsymbol{u}_{\omega_j})$ form the dictionary $\mathcal{D}_m$. Let $P_{\mathcal{D}_m}$ denote the projection operator onto the space they span. A commonly used technique to select the kernel functions in (2) is the approximate linear dependence (ALD) criterion [4]. At each time instant n, the kernel function $\kappa(\cdot, \boldsymbol{u}_n)$ is included in the dictionary $\mathcal{D}_m$ if it satisfies the condition

$$\min_{\boldsymbol{\gamma}} \|\kappa(\cdot, \boldsymbol{u}_n) - \sum_{\omega_j \in \mathcal{J}_{n-1}} \gamma_j \, \kappa(\cdot, \boldsymbol{u}_{\omega_j})\|^2_{\mathcal{H}} > \epsilon_0, \tag{3}$$

with κ a unit-norm kernel[1], and therefore cannot be represented, up to a small error, as a linear combination of previously selected elements. The threshold ϵ_0 determines the level of sparsity of the model. Solving this problem is, however, a computationally intensive task since it requires an m-by-m matrix inversion. We propose making use of another criterion for model order control, the coherence criterion, which has linear complexity with respect to m.

III. Coherence for dictionary analysis

A. Coherence parameter

The coherence parameter is a fundamental quantity used to characterize dictionaries for sparse approximation techniques, with dictionaries from union of orthonormal bases [8], [9], or more recently with arbitrary ones [6], [7]. Let $\kappa(\cdot,\boldsymbol{u}_{\omega_1}),\ldots,\kappa(\cdot,\boldsymbol{u}_{\omega_m})$ be a dictionary composed of m unit-norm kernel functions[1]. Coherence is defined as

$$\mu = \max_{i\neq j} |\langle\kappa(\cdot,\boldsymbol{u}_{\omega_i}),\kappa(\cdot,\boldsymbol{u}_{\omega_j})\rangle_{\mathcal{H}}| = \max_{i\neq j}|\kappa(\boldsymbol{u}_{\omega_i},\boldsymbol{u}_{\omega_j})|,$$

for all $i,j=1,\ldots,m$, and we say that the dictionary is μ-coherent. Note that the largest absolute off-diagonal entry of the Gram matrix is equal to zero for orthonormal dictionaries. In what follows, we demonstrate that the coherence is a powerful parameter to characterize dictionaries. As a warm up to proving this result and others that follow, we have the following result essentially due to [7].

Proposition 1: Consider a μ-coherent dictionary $\mathcal{D}_m$ of m kernel functions. The eigenvalues of its Gram matrix are greater than or equal to $1-(m-1)\,\mu$.

Proof: The Geršgorin disk theorem, applied to the Gram matrix, defines regions that contains its eigenvalues $\nu_1,\ldots,\nu_m$. Each eigenvalue ν_j verifies at least one of the m inequalities $|\nu_j-\kappa(\boldsymbol{u}_{\omega_k},\boldsymbol{u}_{\omega_k})| \leq \sum_{i\neq k}^m |\kappa(\boldsymbol{u}_{\omega_i},\boldsymbol{u}_{\omega_k})|$, for $k=1,\ldots,m$. From the definition of coherence and the normalization condition of κ, we obtain $|\nu_j-1|\leq(m-1)\,\mu$. This implies that $\nu_j\geq 1-(m-1)\,\mu$ for all $j=1,\ldots,m$. ■

The following proposition gives a sufficient condition for a set of kernel functions to be linearly independent [5].

Proposition 2: A sufficient condition for m kernel functions to be linearly independent is $(m-1)\,\mu<1$, where μ denotes their coherence.

Proof: Linear algebra tells us that a set of functions is linearly independent if, and only if, the eigenvalues of its Gram matrix are non-zero. From Proposition 1, a sufficient condition is given by $1-(m-1)\,\mu>0$. ■

New insights on the relationships between the kernel functions of a μ-coherent dictionary are developed next. In particular, we revisit Proposition 3 in [5] by deriving a tighter bound on the approximation error of a dictionary element by the others. The sufficient condition above is obtained by setting this lower bound to zero.

[1]This means that $\kappa(\boldsymbol{u}_k,\boldsymbol{u}_k)=1$ for every $\boldsymbol{u}_k\in\mathcal{U}$; otherwise, substitute $\kappa(\cdot,\boldsymbol{u}_k)/\sqrt{\kappa(\boldsymbol{u}_k,\boldsymbol{u}_k)}$ for $\kappa(\cdot,\boldsymbol{u}_k)$ in the expression.

B. Relation between elements of a μ-coherent dictionary

We shall now study the problem of approximating an element of a μ-coherent dictionary by its other elements. After deriving a new lower bound on the residual error, we compare this lower bound to the bound proposed in [5].

Proposition 3: Let $\mathcal{D}_m$ be a μ-coherent dictionary of m kernel functions with $(m-1)\,\mu<1$. The squared error incurred by approximating any element by its other elements is greater than $1-\sqrt{(m-1)\,\mu^2/(1-(m-2)\,\mu)}$.

Proof: Let $P_{\mathcal{D}_{m-1}}$ denote the projection operator onto the space spanned by the elements of $\mathcal{D}_{m-1}=\{\kappa(\cdot,\boldsymbol{u}_{\omega_j})\}_{j=1}^{m-1}$. The squared norm of $P_{\mathcal{D}_{m-1}}\kappa(\cdot,\boldsymbol{u}_{\omega_m})$ is the maximum, over all the unit functions $\psi(\cdot)$ that belong to the spanned space, of the inner product $\langle\kappa(\cdot,\boldsymbol{u}_{\omega_m}),\psi(\cdot)\rangle_{\mathcal{H}}$.

Writing $\psi(\cdot)=\frac{\sum_{i=1}^{m-1}\alpha_i\,\kappa(\cdot,\boldsymbol{u}_{\omega_i})}{\|\sum_{i=1}^{m-1}\alpha_i\,\kappa(\cdot,\boldsymbol{u}_{\omega_i})\|_{\mathcal{H}}}$, we then have:

$$\|P_{\mathcal{D}_{m-1}}\kappa(\cdot,\boldsymbol{u}_{\omega_m})\|_{\mathcal{H}}^2 = \max_{\boldsymbol{\alpha}} \frac{\sum_{i=1}^{m-1}\alpha_i\,\kappa(\boldsymbol{u}_{\omega_i},\boldsymbol{u}_{\omega_m})}{\|\sum_{i=1}^{m-1}\alpha_i\,\kappa(\cdot,\boldsymbol{u}_{\omega_i})\|_{\mathcal{H}}}. \quad (4)$$

The square of the numerator can be upper-bounded by

$$\begin{aligned}\Big(\sum_{i=1}^{m-1}\alpha_i\kappa(\boldsymbol{u}_{\omega_i},\boldsymbol{u}_{\omega_m})\Big)^2 &\leq \Big(\sum_{i=1}^{m-1}|\alpha_i\kappa(\boldsymbol{u}_{\omega_i},\boldsymbol{u}_{\omega_m})|\Big)^2\\ &\leq \sum_{i=1}^{m-1}\alpha_i^2\sum_{i=1}^{m-1}|\kappa(\boldsymbol{u}_{\omega_i},\boldsymbol{u}_{\omega_m})|^2\\ &\leq (m-1)\,\mu^2\sum_{i=1}^{m-1}\alpha_i^2,\end{aligned}$$

where the second inequality follows from Cauchy-Schwartz inequality, and the last one is due to the definition of coherence. A lower bound on the denominator is found by writing

$$\frac{\|\sum_{i=1}^{m-1}\alpha_i\kappa(\cdot,\boldsymbol{u}_{\omega_i})\|_{\mathcal{H}}^2}{\sum_{i=1}^{m-1}\alpha_i^2} = \frac{\boldsymbol{\alpha}^t\boldsymbol{K}\boldsymbol{\alpha}}{\|\boldsymbol{\alpha}\|^2}\geq\nu_{\min}\geq 1-(m-2)\,\mu.$$

The last inequality follows from proposition 1 applied to the smallest eigenvalue $\nu_{\min}$ of $\boldsymbol{K}$, which is the Gram matrix of the $(m-1)$ elements of $\mathcal{D}_{m-1}$. Finally, combining both inequalities yields the following lower bound on the squared norm of the residue

$$\begin{aligned}&\|(I-P_{\mathcal{D}_{m-1}})\,\kappa(\cdot,\boldsymbol{u}_{\omega_m})\|_{\mathcal{H}}^2\\ &\qquad= \|\kappa(\cdot,\boldsymbol{u}_{\omega_m})\|_{\mathcal{H}}^2-\|P_{\mathcal{D}_{m-1}}\kappa(\cdot,\boldsymbol{u}_{\omega_m})\|_{\mathcal{H}}^2\\ &\qquad\geq 1-\sqrt{\frac{(m-1)\,\mu^2}{1-(m-2)\,\mu}}.\end{aligned} \quad (5)$$

Note that this bound is valid, that is, $1-(m-2)\,\mu>0$, under condition $(m-1)\,\mu<1$. ■

The bound (5) is sharp in the sense that it spans the entire interval $]0,1]$, the upper limit being reached for $\mu=0$ and the lower one for $\mu=1/(m-1)$. Once again, we find the sufficient condition $(m-1)\,\mu<1$ of linear independency. It refers to the case where there is no element that can be represented by a linear combination of the others, without approximation error.

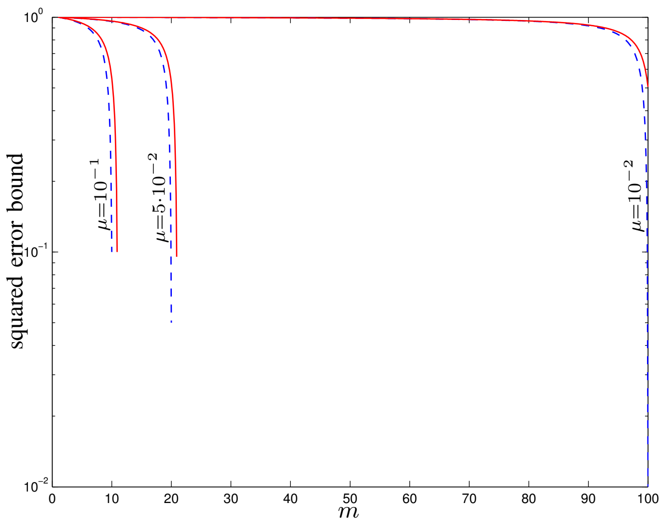

Fig. 1. Lower bounds on the squared error of approximation, in dashed-blue for the earlier work [5] and in solid-red for the one proposed in this paper.

It is sharper than bound $1-\sqrt{(m-1)\,\mu^2/(1-(m-1)\,\mu)}$, implicitly introduced in [6] and derived for filtering with kernel functions in [5]. This implies that the condition for validity, which is now $(m-1)\,\mu < 1$, is less restrictive than the original condition $(m-1)\,\mu < 1/2$ proposed in [5, Proposition 2]. Figure 1 plots these bounds as a function of the size m of the dictionary, for several values of the coherence parameter μ.

IV. Sparsification with the coherence criterion

On the basis of the above results, we shall now take advantage of coherence to produce a dictionary of linearly independent kernel functions. It suffices that the coherence of the latter does not exceed a given threshold $\mu_0 < 1/(m-1)$. Our sparsification rule consists of including, at each time instant n, the kernel function $\kappa(\cdot, \boldsymbol{u}_n)$ into $\mathcal{D}_m$ if

$$\max_{\omega_j \in \mathcal{J}_{n-1}} |\kappa(\boldsymbol{u}_n, \boldsymbol{u}_{\omega_j})| \leq \mu_0. \tag{6}$$

The level of sparsity of $\mathcal{D}_m$ is determined by μ_0. Note that small values induce quasi-orthonormal dictionaries. It has been shown in [5] that dictionaries determined by this rule are finite.

As shown below, the coherence criterion (6) has direct connection to other sparsification rules, in particular the ALD criterion (3) and principal component analysis (PCA). Relation with the entropy-maximization criterion [10] has been established in [5], where a lower bound on the entropy of a μ_0-coherent dictionary is derived.

A. The coherence criterion as an ALD criterion

We shall now establish connection between the coherence criterion (6) and the ALD criterion (3), that is, the relation between their respective thresholds μ_0 and ϵ_0. The lower bound in proposition 3 directly shows that the kernel functions of a μ_0-coherent dictionary $\mathcal{D}_m$ satisfy ALD rule with

$$\epsilon_0 = 1 - \sqrt{\frac{(m-1)\,\mu_0^2}{1-(m-2)\,\mu_0}}.$$

This result does not provide information about the rejection process of the coherence rule (6) and, in particular, the approximation error of discarded kernel functions. Note that the latter is upper-bounded by $\sqrt{\epsilon_0}$ in the case of the ALD rule. The following proposition states a similar result for the coherence criterion.

Proposition 4: Let $\mathcal{D}_m$ be a dictionary produced by rule (6), and $\kappa(\cdot, \boldsymbol{u}_n)$ a kernel function violating this rule. The squared approximation error of $\kappa(\cdot, \boldsymbol{u}_n)$ by the elements of $\mathcal{D}_m$ is lower than $1 - \mu_0$.

Proof: Consider the projection of $\kappa(\cdot, \boldsymbol{u}_n)$ onto the space spanned by the elements of $\mathcal{D}_m = \{\kappa(\cdot, \boldsymbol{u}_{\omega_j})\}_{j=1}^m$, and write its squared error from expression (4) as

$$\begin{aligned} \|(I - P_{\mathcal{D}_m})\kappa(\cdot, \boldsymbol{u}_n)\|_{\mathcal{H}}^2 &= 1 - \max_{\boldsymbol{\alpha}} \frac{\sum_{i=1}^m \alpha_i \kappa(\boldsymbol{u}_{\omega_i}, \boldsymbol{u}_n)}{\|\sum_{i=1}^m \alpha_i \kappa(\cdot, \boldsymbol{u}_{\omega_i})\|_{\mathcal{H}}} \\ &\leq 1 - \max_{\omega_k \in \mathcal{J}_{n-1}} \frac{|\kappa(\boldsymbol{u}_{\omega_k}, \boldsymbol{u}_n)|}{\|\kappa(\cdot, \boldsymbol{u}_{\omega_k})\|_{\mathcal{H}}}. \end{aligned}$$

The above inequality corresponds to the specific set of coefficients $\alpha_1, \ldots, \alpha_m = 0$, except $\alpha_k = \pm 1$ depending on the sign of $\kappa(\boldsymbol{u}_{\omega_k}, \boldsymbol{u}_n)$. Since $\kappa(\cdot, \boldsymbol{u}_n)$ violates condition (6), we have $\max_{\omega_j \in \mathcal{J}_{n-1}} |\kappa(\boldsymbol{u}_n, \boldsymbol{u}_{\omega_j})| > \mu_0$. Combining both inequalities yields the following expression

$$\|(I - P_{\mathcal{D}_m})\kappa(\cdot, \boldsymbol{u}_n)\|_{\mathcal{H}}^2 < 1 - \mu_0$$

because $\kappa(\cdot, \boldsymbol{u}_{\omega_k})$ is a unit-norm kernel function. ■

By combining this bound with the one derived in Proposition 3, we conclude the following about approximating a kernel function with a μ_0-coherent dictionary $\mathcal{D}_m$ with m elements. If the coherence rule (6) is verified, $\kappa(\cdot, \boldsymbol{u}_n)$ must be included in the dictionary. Its squared approximation error exceeds $1-\sqrt{m\,\mu_0^2/(1-(m-1)\,\mu_0)}$. If $\kappa(\cdot, \boldsymbol{u}_n)$ violates the coherence rule, it is discarded from the dictionary. Its squared approximation error is less than $1-\mu_0$. It is worth noting that the former bound is smaller than the latter, for all μ_0 and m. While these two bounds are reduced to a single one, ϵ_0, with ALD criterion, they are distinct with the coherence criterion as illustrated in Figure 2.

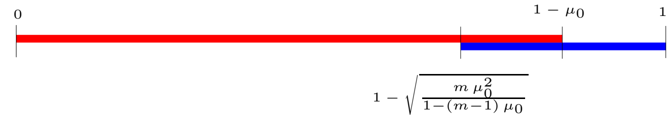

Fig. 2. Squared error bounds of approximating a kernel function from a μ_0-coherent dictionary of size m. The blue region corresponds to the kernel functions verifying (6), while the red one to its violation.

B. Connection to kernel-PCA

Our approach, whose main goal is to judiciously select a subspace spanned by m kernel functions from the original space of data, can be viewed as a dimensionality reduction technique. It seems natural now to consider its connection to kernel principal component analysis (kernel-PCA) [11], an elegant nonlinear extension of the mostly used dimensional reduction technique, the principal component analysis (PCA).

PCA consists of determining principal axes that capture the highest variance in the data, that is, useful information as opposed to noise. These principal axes are the eigenvectors Ψ_k associated with the largest eigenvalues λ_k of the covariance matrix $\boldsymbol{R}$ of data, which means that $\boldsymbol{R}\,\Psi_k = \lambda_k \Psi_k$. There exists a dual formulation of PCA involving only the inner products of the n training data, the kernel-PCA. The k-th principal axis is now given by $\Psi_k = \sum_{j=1}^n \beta_{j,k}\kappa(\cdot, \boldsymbol{u}_j)$, where the $\beta_{j,k}$'s are the components of the k-th eigenvector of the Gram matrix $\boldsymbol{K}$. To have unit-norm principal axes, expansion coefficients must be normalized such that $\sum_{j=1}^n \beta_{j,k}^2 = 1/n\lambda_k$ with $\lambda_k = \nu_k/n$.

Proposition 5: Let $\mathcal{D}_m$ be a dictionary produced by coherence rule (6) from n kernel functions. Let Ψ_k denote the k-th principal axis of these n kernel functions with eigenvalue λ_k. The squared approximation error of eigenvector Ψ_k by the m elements of $\mathcal{D}_m$ is lower than $(1-\mu_0)/\lambda_k$.

Proof: To prove this, we need to upper bound the norm of the residue $\|(I - P_{\mathcal{D}_m})\Psi_k\|_{\mathcal{H}}$. Let the $\kappa(\cdot, \boldsymbol{u}_{\omega_j})$'s be the elements of $\mathcal{D}_m$. From the expansion $\Psi_k = \sum_{j=1}^n \beta_{j,k}\kappa(\cdot, \boldsymbol{u}_j)$, we can write

$$\begin{aligned}\|(I - P_{\mathcal{D}_m})\Psi_k\|_{\mathcal{H}} &= \Big\|\sum_{j=1}^n \beta_{j,k}(I - P_{\mathcal{D}_m})\kappa(\cdot, \boldsymbol{u}_j)\Big\|_{\mathcal{H}} \\ &\leq \sum_{j=1}^n |\beta_{j,k}|\ \|(I - P_{\mathcal{D}_m})\kappa(\cdot, \boldsymbol{u}_j)\|_{\mathcal{H}} \\ &= \sum_{\substack{j=1\\ \omega_j \notin \mathcal{J}_n}}^n |\beta_{j,k}|\ \|(I - P_{\mathcal{D}_m})\kappa(\cdot, \boldsymbol{u}_j)\|_{\mathcal{H}}.\end{aligned}$$

The inequality comes from the generalized triangle inequality, and the last equality follows from $(I - P_{\mathcal{D}_m})\kappa(\cdot, \boldsymbol{u}_{\omega_j}) = 0$ for all $\kappa(\cdot, \boldsymbol{u}_{\omega_j}) \in \mathcal{D}_m$. Since the kernel functions in the right-hand-side of the above expression are discarded from the dictionary $\mathcal{D}_m$ by rule (6), we know from proposition 4 that $\|(I - P_{\mathcal{D}_m})\kappa(\cdot, \boldsymbol{u}_j)\|_{\mathcal{H}}^2 < 1 - \mu_0$. Thus,

$$\|(I - P_{\mathcal{D}_m})\Psi_k\|_{\mathcal{H}}^2 < (1-\mu_0)\Big(\sum_{\substack{j=1\\ \omega_j \notin \mathcal{J}_n}}^n |\beta_{j,k}|\Big)^2.$$

The above summation can be upper bounded by

$$\Big(\sum_{\substack{j=1\\ \omega_j \notin \mathcal{J}_n}}^n |\beta_{j,k}|\Big)^2 \leq \Big(\sum_{j=1}^n |\beta_{j,k}|\Big)^2 \leq n\sum_{j=1}^n \beta_{j,k}^2 = \frac{1}{\lambda_k},$$

where the second inequality follows from Cauchy-Schwartz inequality. The equality is due to normalization in kernel-PCA. Finally, by substituting this summation into the above expression, we get

$$\|(I - P_{\mathcal{D}_m})\Psi_k\|_{\mathcal{H}}^2 < \frac{1-\mu_0}{\lambda_k}.$$

■

From this upper bound, we conclude that the principal axes with the largest eigenvalues have small approximation errors. Then they can be viewed, up to a small error, as belonging to the space spanned by the elements of the dictionary. The coherence criterion can therefore be considered as a principal component technique, which estimates a subspace from observations without the computational burden of matrix inversion in both PCA and kernel-PCA algorithms. A similar result is derived in [4] for dictionaries derived from ALD criterion. While the latter has a computational complexity which is quadratic in the size of the dictionary, the coherence criterion provides a linear complexity.

V. The KNLMS algorithm with the coherence criterion

In [5], a KRLS algorithm is derived for solving (1) for the m-order model (2). The KRLS algorithm has a quadratic computational complexity with respect to m. Since the coherence criterion has a linear computational complexity, it is natural to propose a filtering algorithm that has a similar complexity. In this paper, we consider a simple stochastic-gradient method for solving the optimization problem, the kernel normalized least-mean-squares (KNLMS).

A. The KNLMS algorithm

Under the principle of minimal disturbance, we reformulate the optimization problem as follows: at each time instant n, we seek the coefficient vector $\boldsymbol{\alpha}_n$ that sets to zero the *a posteriori* error, namely

$$d_n - \boldsymbol{h}_n^t \boldsymbol{\alpha}_n = 0, \tag{7}$$

where $\boldsymbol{h}_n$ is an m-by-1 vector whose i-th entry is $\kappa(\boldsymbol{u}_n, \boldsymbol{u}_{\omega_i})$. Upon arrival of new data, two cases may occur, depending on the coherence rule (6).

Case 1. $\max_{j=1\dots m} |\kappa(\boldsymbol{u}_n, \boldsymbol{u}_{\omega_j})| > \mu_0$

In this case, the kernel function $\kappa(\cdot, \boldsymbol{u}_n)$ is not included in the dictionary. The model coefficients are updated according to the condition (7). Let us rewrite the *a priori* estimation error, defined by $e_n = d_n - \boldsymbol{h}_n^t \boldsymbol{\alpha}_{n-1}$, as follows:

$$e_n = \boldsymbol{h}_n^t(\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_{n-1}).$$

Minimizing e_n is an under-determined problem with 1 equation and m variables. Nevertheless, there exists a unique optimal solution in the least-squares sense that can be computed from the pseudo-inverse of $\boldsymbol{h}_n^t$. This leads to

$$\boldsymbol{\alpha}_n - \boldsymbol{\alpha}_{n-1} = \frac{1}{\|\boldsymbol{h}_n\|^2}\boldsymbol{h}_n e_n.$$

By introducing a step-size control parameter ρ, we obtain the recursion

$$\boldsymbol{\alpha}_n = \boldsymbol{\alpha}_{n-1} + \frac{\rho}{\|\boldsymbol{h}_n\|^2}\boldsymbol{h}_n\left(d_n - \boldsymbol{h}_n^t\boldsymbol{\alpha}_{n-1}\right). \tag{8}$$

The choice of an appropriate step-size for achieving optimal convergence rates is extensively investigated in the adaptive filtering literature [12].

Case 2. $\max_{j=1\ldots m}|\kappa(\boldsymbol{u}_n, \boldsymbol{u}_{\omega_j})| \leq \mu_0$
There may be considerable error in representing $\kappa(\cdot, \boldsymbol{u}_n)$ by the kernel functions of the dictionary. Therefore, $\kappa(\cdot, \boldsymbol{u}_n)$ must be included in the dictionary. For this, the model order is incremented, and both $\boldsymbol{\alpha}_n$ and $\boldsymbol{h}_n$ are updated to $(m+1)$-by-1 column vectors according

$$\begin{aligned}\boldsymbol{h}_n &= [\kappa(\boldsymbol{u}_n, \boldsymbol{u}_{\omega_1})\ldots\kappa(\boldsymbol{u}_n, \boldsymbol{u}_{\omega_{m+1}})]^t\\ \boldsymbol{\alpha}_n &= \begin{bmatrix}\boldsymbol{\alpha}_{n-1}\\ 0\end{bmatrix} + \frac{\rho}{\|\boldsymbol{h}_n\|^2}\boldsymbol{h}_n\left(d_n - \boldsymbol{h}_n^t\begin{bmatrix}\boldsymbol{\alpha}_{n-1}\\ 0\end{bmatrix}\right),\end{aligned}$$

where the recursion is obtained from expression (8) derived in Case 1.

B. Simulations

As an application, we consider the nonlinear dynamic system identification problem [3]

$$\begin{aligned}y_n &= 0.5\, y_{n-1}u_{n-1} + 0.2\, u_{n-1} + 0.05\, y_{n-1}^2 + 0.6\, u_{n-1}^2\\ d_n &= y_n + \epsilon_n\end{aligned}$$

where d_n is the observed output, corrupted by a measurement noise ϵ_n sampled from a zero-mean Gaussian distribution with a standard deviation of 0.1, which corresponds to a signal-to-noise ratio of about 30%. With an initial condition $y_1 = 0.1$, data were generated from a control sequence u_n sampled from a Gaussian distribution with a standard deviation of 0.1 and a mean of 0.2. We used these data to estimate a nonlinear model of the form $d_n = \psi(d_{n-1}, u_{n-1})$. We considered the Gaussian kernel $\kappa(\boldsymbol{u}_i, \boldsymbol{u}_j) = \exp(-\|\boldsymbol{u}_i - \boldsymbol{u}_j\|^2/\beta_0)$, with $\beta_0 = 0.02$, and a step-size ρ of 9×10^{-2}. Figure 3 illustrates the convergence behavior of both KNLMS and KRLS, for different values of threshold μ_0. Each curve represents the average over 100 runs, then smoothed by time averaging over 20 consecutive samples. The mean order of each model, over these runs, is given in the legend. Note that the sufficient condition for linearly independent kernel functions is verified. As expected, KRLS converges faster than KNLMS, but with a significantly larger computational complexity.

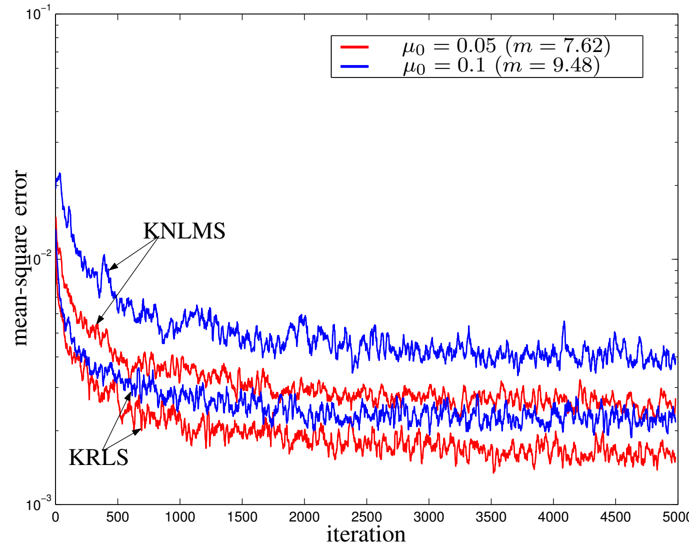

Fig. 3. Convergence behavior of KNLMS and KRLS algorithms for different values of μ_0. The mean order of each model is given by m in the legend.

VI. Conclusion

This paper considered nonlinear adaptive filtering in RKHS using coherence as a sparsification criterion. We studied the approximation problem from a dictionary constructed by this criterion, and provided new tighter bounds on its approximation error. We have also connected the coherence criterion to both the ALD criterion and the PCA technique. Finally, we presented initial experimental results on the kernel normalized LMS algorithm applied to nonlinear system identification.

References

[1] N. Aronszajn, "Theory of reproducing kernels," vol. 68, pp. 337–404, 1950.
[2] G. Kimeldorf and G. Wahba, "Some results on tchebycheffian spline functions," *Journal of Mathematical Analysis and Applications*, vol. 33, pp. 82–95, 1971.
[3] T. Dodd and R. Harrison, "Iterative solution to approximation in reproducing kernel hilbert spaces," in *Proceedings of the IFAC World Congress on Automatic Control*, Barcelona, Spain, 2002.
[4] Y. Engel, S. Mannor, and R. Meir, "The kernel recursive least squares algorithm," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2275–2285, 2004.
[5] C. Richard, J. C. M. Bermudez, and P. Honeine, "Nonlinear kernel-based adaptive filtering with order controlled by a coherence criterion," *submitted to IEEE Transactions on Signal Processing*, 2007.
[6] A. C. Gilbert, S. Muthukrishnan, and M. J. Strauss, "Approximation of functions over redundant dictionaries using coherence," in *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms (SODA'03)*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2003, pp. 243–252.
[7] J. A. Tropp, "Greed is good: algorithmic results for sparse approximation," *IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2231–2242, 2004.
[8] D. L. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Transactions on Information Theory*, vol. 47, no. 7, pp. 2845–2862, March 2001.
[9] R. Gribonval and M. Nielsen, "Sparse representations in unions of bases," *IEEE Transactions on Information Theory*, vol. 49, no. 12, pp. 3320–3325, 2003.
[10] L. Hoegaerts, "Eigenspace methods and subset selection in kernel based learning," PhD thesis, Faculty of Engineering, K.U.Leuven, Leuven, Belgium, Jun. 2005.
[11] B. Schölkopf, A. Smola, and K. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, no. 5, pp. 1299–1319, 1998.
[12] A. H. Sayed, *Fundamentals of adaptive filtering*. NY, USA: Wiley-IEEE Press, June 2003.