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 \( u_i \) of \( \mathcal{U} \) using \( \psi(u_i) = \langle \psi(\cdot), \kappa(\cdot, u_i) \rangle_{\mathcal{H}} \), where \( \kappa(\cdot, u_i) \) is a positive definite kernel that takes \( u_i \) into \( \kappa(u_i, u_i) \). By setting \( \mathcal{H} \) as the hypothesis space, we consider as a cost function the squared error between the model outputs \( \psi(u_i) \) and the desired responses \( d_i \), that is,

\[
\sum_{i=1}^{n} (d_i - \psi(u_i))^2.
\]

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,

\[
\psi(\cdot) = \sum_{j=1}^{n} \alpha_j \kappa(\cdot, u_j).
\]

The optimization problem is then reduced to the dual problem of determining \( \alpha = [\alpha_1 \ldots \alpha_n]^T \) such that

\[
\min_{\alpha} \|d - \mathcal{K} \alpha\|^2,
\]

where \( \mathcal{K} \) denotes the Gram matrix whose \( (i,j) \)-th entry is \( \kappa(u_i, u_j) \), and \( d = [d_1 \ldots d_n]^T \). Solution to this problem is given by \( \alpha = \mathcal{K}^+ d \), where \( \mathcal{K}^+ \) is the pseudo-inverse of \( \mathcal{K} \). Since the model order is equal to the number \( n \) of available data \( u_i \), 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

\[
\psi_n(\cdot) = \sum_{\omega_j \in \mathcal{J}_n} \alpha_{n,j} \kappa(\cdot, u_{\omega_j}),
\]

where \( \mathcal{J}_n \) is a subset of \( m \) indices of \{1, \ldots, n\}. The \( m \) kernel functions \( \kappa(\cdot, u_{\omega_j}) \) 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 \( \kappa(\cdot, u_n) \) is included in the dictionary \( \mathcal{D}_m \) if it satisfies the condition

\[
\min_{\gamma} \| \kappa(\cdot, u_n) - \sum_{\omega_j \in \mathcal{J}_{n-1}} \gamma_j \kappa(\cdot, u_{\omega_j}) \|^2_{\mathcal{H}} > \epsilon_0,
\]

where \( \mathcal{J}_n \) is the set of \( m \) indices of \{1, \ldots, n\}. The \( m \) kernel functions \( \kappa(\cdot, u_{\omega_j}) \) 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 \( \kappa(\cdot, u_n) \) is included in the dictionary \( \mathcal{D}_m \) if it satisfies the condition

\[
\min_{\gamma} \| \kappa(\cdot, u_n) - \sum_{\omega_j \in \mathcal{J}_{n-1}} \gamma_j \kappa(\cdot, u_{\omega_j}) \|^2_{\mathcal{H}} > \epsilon_0,
\]

where \( \mathcal{J}_n \) is the set of \( m \) indices of \{1, \ldots, n\}. The \( m \) kernel functions \( \kappa(\cdot, u_{\omega_j}) \) 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 \( \kappa(\cdot, u_n) \) is included in the dictionary \( \mathcal{D}_m \) if it satisfies the condition

\[
\min_{\gamma} \| \kappa(\cdot, u_n) - \sum_{\omega_j \in \mathcal{J}_{n-1}} \gamma_j \kappa(\cdot, u_{\omega_j}) \|^2_{\mathcal{H}} > \epsilon_0,
\]

where \( \mathcal{J}_n \) is the set of \( m \) indices of \{1, \ldots, n\}. The \( m \) kernel functions \( \kappa(\cdot, u_{\omega_j}) \) 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 \( \kappa(\cdot, u_n) \) is included in the dictionary \( \mathcal{D}_m \) if it satisfies the condition

\[
\min_{\gamma} \| \kappa(\cdot, u_n) - \sum_{\omega_j \in \mathcal{J}_{n-1}} \gamma_j \kappa(\cdot, u_{\omega_j}) \|^2_{\mathcal{H}} > \epsilon_0,
\]
with $\kappa$ 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 $\epsilon_0$ 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. COHESION 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 $\kappa(\cdot, u_{w_1}), \ldots, \kappa(\cdot, u_{w_m})$ be a dictionary composed of $m$ unit-norm kernel functions\(^1\). Coherence is defined as

$$\mu = \max_{i \neq j} |\kappa(\cdot, u_{w_i})| = \max_{i \neq j} |\kappa(u_{w_i}, u_{w_j})|,$$

for all $i, j = 1, \ldots, m$, and we say that the dictionary is $\mu$-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 $\mu$-coherent dictionary $D_m$ of $m$ kernel functions. The eigenvalues of its Gram matrix are greater than or equal to $(m - 1) \mu$.

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

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 $\mu$ 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 $1 - (m - 1) \mu > 0$. $\blacksquare$

New insights on the relationships between the kernel functions of a $\mu$-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.

**B. Relation between elements of a $\mu$-coherent dictionary**

We shall now study the problem of approximating an element of a $\mu$-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 $D_m$ be a $\mu$-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 or equal to $1 - \sqrt{(m - 1) \mu^2/(1 - (m - 2) \mu)}$.

**Proof:** Let $P_{D_m-1}$ denote the projection operator onto the space spanned by the elements of $D_{m-1} = \{\kappa(\cdot, u_{w_i})\}_{i=1}^{m-1}$. The squared norm of $P_{D_m-1} \kappa(\cdot, u_{w_m})$ is the maximum, over all the unit functions $\psi(\cdot)$ that belong to the spanned space, of the inner product $\langle \kappa(\cdot, u_{w_m}), \psi(\cdot) \rangle_{\mathcal{H}}$.

Writing $\psi(\cdot) = \sum_{i=1}^{m-1} \alpha_i \kappa(u_{w_i}, \cdot)$, we then have:

$$\|P_{D_m-1} \kappa(\cdot, u_{w_m})\|^2_{\mathcal{H}} = \max_{\alpha} \frac{\sum_{i=1}^{m-1} \alpha_i \kappa(u_{w_i}, u_{w_m})}{\| \sum_{i=1}^{m-1} \alpha_i \kappa(u_{w_i}, \cdot) \|^2_{\mathcal{H}}}.$$

(4)

The square of the numerator can be upper-bounded by

$$\left( \sum_{i=1}^{m-1} \alpha_i \kappa(u_{w_i}, u_{w_m}) \right)^2 \leq \left( \sum_{i=1}^{m-1} |\alpha_i| \kappa(u_{w_i}, u_{w_m}) \right)^2 \leq \sum_{i=1}^{m-1} \alpha_i^2 \sum_{i=1}^{m-1} \kappa(u_{w_i}, u_{w_m})^2 \leq (m - 1) \mu^2 \sum_{i=1}^{m-1} \alpha_i^2,$$

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

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

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

$$\| (I - P_{D_m-1}) \kappa(\cdot, u_{w_m}) \|^2_{\mathcal{H}} = \| \kappa(\cdot, u_{w_m}) \|^2_{\mathcal{H}} - \| P_{D_m-1} \kappa(\cdot, u_{w_m}) \|^2_{\mathcal{H}} \geq 1 - \sqrt{(m - 1) \mu^2/(1 - (m - 2) \mu)}.$$

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

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.

\(^1\)This means that $\kappa(u_{w_i}, u_k) = 1$ for every $u_k \in U$; otherwise, substitute $\kappa(\cdot, u_k)/\sqrt{\kappa(u_k, u_k)}$ for $\kappa(\cdot, u_k)$ in the expression.
bound in proposition 3 directly shows that the kernel functions small values induce quasi-orthonormal dictionaries. It has been criterion (3) and principal component analysis (PCA). Rela-
shon in [5] that dictionaries determined by this rule are finite. This implies that the coherence of functions in [5]. This implies that the condition for validity, in the case of the ALD rule. The following proposition states a similar result for the coherence criterion.

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

**Proof:** Consider the projection of \( \kappa(\cdot, \mathbf{u}_n) \) onto the space spanned by the elements of \( D_m = \{ \kappa(\cdot, \mathbf{u}_j) \}_{j=1}^m \), and write its squared error from expression (4) as

\[
\|(I - P_{D_m}) \kappa(\cdot, \mathbf{u}_n)\|_H^2 = 1 - \max {\sum_{i=1}^{m} \alpha_i \kappa(\mathbf{u}_i, \mathbf{u}_j) \over \sum_{i=1}^{m} \| \kappa(\mathbf{u}_i, \mathbf{u}_j) \|_H} \leq 1 - \max {\sum_{i=1}^{m} \alpha_i \kappa(\mathbf{u}_i, \mathbf{u}_n) \over \| \kappa(\mathbf{u}_n, \mathbf{u}_j) \|_H}.
\]

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 \( \kappa(\mathbf{u}_{\omega_k}, \mathbf{u}_n) \). Since \( \kappa(\cdot, \mathbf{u}_n) \) violates condition (6), we have \( \max_{\omega_j \in \mathcal{J}_{m-1}} |\kappa(\mathbf{u}_n, \mathbf{u}_{\omega_j})| > \mu_0 \). Combining both inequalities yields the following expression

\[
\|(I - P_{D_m}) \kappa(\cdot, \mathbf{u}_n)\|_H^2 < 1 - \mu_0
\]

because \( \kappa(\cdot, \mathbf{u}_{\omega_k}) \) 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 \( \mu_0 \)-coherent dictionary \( D_m \) with \( m \) elements. If the coherence rule (6) is verified, \( \kappa(\cdot, \mathbf{u}_n) \) must be included in the dictionary. Its squared approximation error exceeds \( 1 - \sqrt{m \mu_0^2/(1 - (m - 1) \mu_0)} \). If \( \kappa(\cdot, \mathbf{u}_n) \) 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 \( \mu_0 \) and \( m \).

While these two bounds are reduced to a single one, \( \epsilon_0 \), with ALD criterion, they are distinct with the coherence criterion as illustrated in Figure 2.

![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.](image1)

![Fig. 2. Squared error bounds of approximating a kernel function from a \( \mu_0 \)-coherent dictionary of size \( m \). The blue region corresponds to the kernel functions verifying (6), while the red one to its violation.](image2)
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 principal axis of these 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.

\[ \text{V. THE KNLMS ALGORITHM WITH THE COHERENCE CRITERION} \]

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 \( \alpha_n \) that sets to zero the a posteriori error, namely

\[ d_n - h_n^\top \alpha_n = 0, \]

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

Case 1. \( \max_{j=1 \ldots m} |\kappa(u_n, u_{\omega_j})| > \mu_0 \)

In this case, the kernel function \( \kappa(\cdot, u_n) \) is not included in the dictionary. The model coefficients are updated according to the condition \((7)\).

Let us rewrite the \textit{a priori} estimation error, defined by \( e_n = d_n - h_n^\top \alpha_{n-1} \), as follows:

\[ e_n = h_n^\top (\alpha_n - \alpha_{n-1}). \]

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 \( h_n^\top \). This leads to

\[ \alpha_n - \alpha_{n-1} = \frac{1}{||h_n||^2} h_n e_n. \]
By introducing a step-size control parameter $\rho$, we obtain the recursion
\[
\alpha_n = \alpha_{n-1} + \frac{\rho}{\|h_n\|^2} h_n (d_n - h_n^T \alpha_{n-1}). \tag{8}
\]

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

**Case 2.** \[\max_{j=1,\ldots,m} |\kappa(u_n, u_{w_j})| \leq \mu_0 \]

There may be considerable error in representing $\kappa(\cdot, u_n)$ by the kernel functions of the dictionary. Therefore, $\kappa(\cdot, u_n)$ must be included in the dictionary. For this, the model order is incremented, and both $\alpha_n$ and $h_n$ are updated to $(m+1)$-by-1 column vectors according
\[
\begin{align*}
    h_n &= [\kappa(u_n, u_{w_1}) \ldots \kappa(u_n, u_{w_{m+1}})]^T \\
    \alpha_n &= \begin{bmatrix} \alpha_{n-1} \\ 0 \end{bmatrix} + \frac{\rho}{\|h_n\|^2} h_n \begin{bmatrix} d_n - h_n^T \alpha_{n-1} \\ 0 \end{bmatrix},
\end{align*}
\]

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]
\[
y_n = 0.5 y_{n-1} u_{n-1} + 0.2 u_{n-1} + 0.05 y_{n-2}^2 + 0.6 u_{n-1}^2 \\
d_n = y_n + \epsilon_n
\]

where $d_n$ is the observed output, corrupted by a measurement noise $\epsilon_n$ 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 $\kappa(u, u_{ij}) = \exp(-\|u - u_{ij}\|^2/\beta_0)$, with $\beta_0 = 0.02$, and a step-size $\rho$ of $9 \times 10^{-2}$. Figure 3 illustrates the convergence behavior of both KNLSM and KRLS, for different values of threshold $\mu_0$. 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 KNLSM, but with a significantly larger computational complexity.

**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**