Approximation errors of online sparsification criteria

Paul Honeine

Approximation Errors of Online Sparsification Criteria

Paul Honeine, *Member IEEE*

***Abstract*—Many machine learning frameworks, such as resource-allocating networks, kernel-based methods, Gaussian processes, and radial-basis-function networks, require a sparsification scheme in order to address the online learning paradigm. For this purpose, several online sparsification criteria have been proposed to restrict the model definition on a subset of samples. The most known criterion is the (linear) approximation criterion, which discards any sample that can be well represented by the already contributing samples, an operation with excessive computational complexity. Several computationally efficient sparsification criteria have been introduced in the literature with the distance and the coherence criteria. This paper provides a unified framework that connects these sparsification criteria in terms of approximating samples, by establishing theoretical bounds on the approximation errors. Furthermore, the error of approximating any pattern is investigated, by proposing upper bounds on the approximation error for each of the aforementioned sparsification criteria. Two classes of fundamental patterns are described in detail, the centroid (*i.e.*, empirical mean) and the principal axes in the kernel principal component analysis. Experimental results show the relevance of the theoretical results established in this paper.**

***Index Terms*—Sparse approximation, adaptive filtering, kernel-based methods, resource-allocating networks, Gram matrix, machine learning, pattern recognition, online learning, sparsification criteria.**

I. Introduction

DATA DELUGE in the era of "Big Data" brings new opportunities and challenges in the area of machine learning and signal processing [1], [2], [3]. This paradigm is often addressed in an online setting, by coupling a sparsification scheme with the learning machine under scrutiny. Indeed, many machine learning frameworks, such as resource-allocating networks [4], kernel-based methods for classification and regression [5], Gaussian processes [6], radial-basis-function networks [7] and kernel principal component analysis [8], share essentially the same underlying model, with as many parameters to be estimated as training samples, as defined by the "Representer Theorem" [9]. This model is inappropriate in online learning, where a new sample is available at each instant. To stay computationally tractable, one needs to restrict the incrementation in the model complexity, by selecting the subset of samples that contributes to a reduced-order model as an approximation of the full-order one. In order to overcome this bottleneck in online learning, sparsification schemes have been proposed for all the aforementioned machines.

P. Honeine is with the Institut Charles Delaunay (CNRS), Université de technologie de Troyes, 10000, Troyes, France. Phone: +33 (0) 3 25 71 56 25; Fax: +33 (0) 3 25 71 56 99; E-mail: paul.honeine@utt.fr

An online sparsification scheme operates as follows: at each instant, it determines if the new sample can be safely discarded from contributing to the order growth of the model; otherwise, the sample needs to take part in the order incrementation. The most known online sparsification criteria is the approximation criterion, also called approximate linear dependency. It has been widely investigated in the literature, for Gaussian processes [10], kernel recursive least squares algorithm [11], kernel least mean square algorithm [12], and kernel principal component analysis [8]. This criterion determines the relevance of discarding or accepting the current sample by comparing, to a predefined threshold, the residual error of approximating it with a representation (*i.e.*, linear combination) of samples — or nonlinearly mapped samples as in kernel methods — already contributing to the model. A crucial issue in the approximation criterion is its computational complexity, which scales cubically with the order of the model under scrutiny.

Several computationally efficient sparsification criteria have been introduced in the literature, with essentially the same computational complexity that scales linearly with the model order. These sparsification criteria rely on the topology of the samples in order to select the most relevant samples. The most widely investigated criteria are the distance and the coherence criteria, as well as several variants such as the Babel criterion. The distance criterion, introduced by Platt in [4] to control the complexity of resource-allocating networks in radial-basis-function networks, retains the most mutually distant samples; see also [13], [14] for recent advances on the distance criterion. The coherence criterion, introduced by Honeine, Richard, and Bermudez in [15], [16] with the recent advances in compressed sensing [17], [18], retains samples that are mutually least coherent. As an extension of the coherence criterion, the Babel criterion uses the cumulative coherence as a measure of diversity [19].

These sparsification criteria have been separately investigated in the literature. To the best of our knowledge, there is no work that studies all these sparsification criteria together. The conducted analyses have been often based on the computational complexity, as advocated in [16], [20] by criticizing the computational cost of the approximation criterion in favor of the other sparsification criteria. In [15], [16], [21], we have developed with colleagues several theoretical results that allows to compare the coherence to the approximation criterion. These results have not been extended to other sparsification criteria, and were demonstrated for the particular case of unit-norm data. This paper allows to cross-fertilize these results for several sparsification criteria and extends them to the general case, beyond unit-norm data.

TABLE I
A BIRDS EYE VIEW OF THIS PAPER. SOME OF THE RESULTS WERE PREVIOUSLY STUDIED FOR UNIT-NORM KERNELS, AS SHOWN WITH THE REFERENCES GIVEN IN THE TABLE (WHERE • DENOTES TRIVIALITY FOR THE APPROXIMATION CRITERION). IN THIS WORK, WE PROVIDE AN EXTENSIVE STUDY THAT COMPLETES THE ANALYSIS TO ALL SPARSIFICATION CRITERIA, OFTEN WITH SHARPER BOUNDS (SHOWN IN GRAY COLOR), AND WE ESTABLISH NEW THEORETICAL RESULTS. MOREOVER, WE GENERALIZE THESE RESULTS TO ANY TYPE OF ATOMS, BEYOND UNIT-NORM ATOMS.

	Distance	Approximation	Coherence	Section
Reference: most known work	[4]	[10]	[16]	
Reference: more recent work	[20]	[8]	[25]	
Approximation of any sample	✓	•	✓	**IV**
↳ Error on discarded samples	✓	•	[15]	IV-A
↳ Error on any atom	✓	•	[16]	IV-B
Approximation of any pattern	✓	✓	✓	**V**
↳ Error on the centroid	✓	✓	[21]	V-A
↳ Error on the principal axes	[11]	[15]	V-B	

This paper presents a unified framework in order to bridge the gap between online sparsification criteria, as follows. On one hand, we show that most known online sparsification criteria behave essentially in an identical mechanism as the approximation criterion. To this end, we provide upper bounds on the error of approximating, with the dictionary elements, any sample discarded by the sparsification criterion; secondly, we provide lower bounds on the error of approximating retained samples. On the other hand, we examine the relevance of approximating any full-order pattern with a sparse model obtained with any of the aforementioned sparsification criteria, including the approximation criterion. Within the proposed framework, we provide upper bounds on the error of approximating any pattern in the general case. Furthermore, we explore in detail two particular patterns, the centroid (*i.e.*, empirical mean, as studied for instance in [22], [23]) and the principal axes in the kernel principal component analysis (kernel-PCA, [24]).

The core contribution of this paper is to provide a unified presentation of kernel-related sparsification criteria, with the derivation of bounds for the approximation errors for samples and patterns, as described respectively in Sections IV and V; see Table I for an overview. The remainder of this paper is organized as follows. Next section introduces the kernel-based machines for online learning and presents the key issues studied in this work. Section III presents the aforementioned computationally efficient sparsification criteria. Section IV investigates bounds on the error of approximating samples, either discarded or accepted by any sparsification criterion. These results are extended in Section V to the problem of approximating any pattern. Experimental results are conducted in Section VI, illustrating the relevance of the obtained results. Section VII concludes this document with discussions and future works.

II. KERNEL-BASED MACHINES FOR ONLINE LEARNING

In this section, we introduce the kernel-based machines for online learning, by presenting the approximation criterion with the key issues studied in this paper.

A. Machine learning and online learning

Machine learning seeks a pattern $\psi(\cdot)$ connecting an input space $\mathbb{X} \subset \mathbb{R}^d$ to an output space $\mathbb{Y} \subset \mathbb{R}$, by using a set of training samples, denoted $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}$ with $(\boldsymbol{x}_k, y_k) \in \mathbb{X} \times \mathbb{Y}$. Considering a loss function $\mathcal{C}(\cdot,\cdot)$ defined on $\mathbb{Y} \times \mathbb{Y}$ that measures the error between the desired output and the estimated one with $\psi(\cdot)$, one seeks to minimize a regularized empirical risk of the form

$$\underset{\psi(\cdot)\in\mathbb{H}}{\operatorname{argmin}} \sum_{i=1}^{n} \mathcal{C}(\psi(\boldsymbol{x}_i), y_i) + \eta\, \mathcal{R}(\|\psi(\cdot)\|_{\mathbb{H}}^2), \tag{1}$$

where $\mathbb{H}$ is the hypothesis space of candidate solutions and η is a parameter that controls the tradeoff between the fitness error (first term) and the regularity of the solution (second term) with $\mathcal{R}(\cdot)$ being a monotonically increasing function. Examples of loss functions are the quadratic loss $|\psi(\boldsymbol{x}_i) - y_i|^2$, the hinge loss $(1 - \psi(\boldsymbol{x}_i) y_i)_+$ of the SVM [5], the logistic loss $\log(1 + \exp(-\psi(\boldsymbol{x}_i) y_i))$, as well as the unsupervised loss function $-|\psi(\boldsymbol{x}_i)|^2$ which is related to the PCA.

Let $\kappa : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ be a positive definite kernel, and $(\mathbb{H}, \langle\cdot,\cdot\rangle_{\mathbb{H}})$ the induced reproducing kernel Hilbert space (RKHS) with its inner product. The reproducing property states that any function $\psi(\cdot)$ of $\mathbb{H}$ can be evaluated at any sample $\boldsymbol{x}_i$ of $\mathbb{X}$ using $\psi(\boldsymbol{x}_i) = \langle \psi(\cdot), \kappa(\cdot, \boldsymbol{x}_i)\rangle_{\mathbb{H}}$. This property shows that any sample $\boldsymbol{x}_i$ of $\mathbb{X}$ is represented with $\kappa(\cdot, \boldsymbol{x}_i)$ in $\mathbb{H}$. Moreover, the reproducing property leads to the so-called kernel trick, that is for any pair of samples $(\boldsymbol{x}_i, \boldsymbol{x}_j)$, we have $\langle \kappa(\cdot, \boldsymbol{x}_i), \kappa(\cdot, \boldsymbol{x}_j)\rangle_{\mathbb{H}} = \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and $\|\kappa(\cdot, \boldsymbol{x}_i)\|_{\mathbb{H}} = \kappa(\boldsymbol{x}_i, \boldsymbol{x}_i)$. The most used kernels and their expressions are:

Kernel	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$
Linear	$\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$
Polynomial	$(\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle + c)^p$
Exponential	$\exp(\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle)$
Gaussian	$\exp\left(\frac{-1}{2\sigma^2}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2\right)$

Among these kernels, only the Gaussian kernel is unit-norm, that is $\|\kappa(\boldsymbol{x}, \cdot)\|_{\mathbb{H}} = 1$ for all $\boldsymbol{x}$. Other kernels can be unit-norm on some restricted $\mathbb{X}$, such as the linear kernel for unit-norm samples. In this paper, we do not restrict ourselves to any particular kernel or space $\mathbb{X}$. We denote

$$r^2 = \inf_{\boldsymbol{x}\in\mathbb{X}} \kappa(\boldsymbol{x}, \boldsymbol{x}) \qquad \text{and} \qquad R^2 = \sup_{\boldsymbol{x}\in\mathbb{X}} \kappa(\boldsymbol{x}, \boldsymbol{x}).$$

For unit-norm kernels, we get $R = r = 1$.

The Representer Theorem provides an essential result in kernel-based machines. It states that the solution of the optimization problem (1) takes the form

$$\psi(\cdot) = \sum_{i=1}^{n} \alpha_i\, \kappa(\boldsymbol{x}_i, \cdot). \tag{2}$$

This theorem, proven in [9], shows that the optimal solution has as many parameters α_i to be estimated as the number of

available samples $(\boldsymbol{x}_i, y_i)$. This result constitutes the principal bottleneck for online learning. Indeed, in an online setting, the solution should be adapted based on a new sample available at each instant, namely $(\boldsymbol{x}_t, y_t)$ at instant t. Thus, by including the new pair $(\boldsymbol{x}_t, y_t)$ in the training set, the corresponding parameter α_t is added to the set of parameters to be estimated, by following the Representer Theorem. As a consequence, the order of the model (2) is continuously increasing.

To overcome this bottleneck, one needs to control the growth of the model order at each instant, by keeping only a fraction of the kernel functions in the expansion (2). The reduced-order model takes the form

$$\psi(\cdot) = \sum_{j=1}^{m} \alpha_j \, \kappa(\dot{\boldsymbol{x}}_j, \cdot) \tag{3}$$

with $m \ll t$, predefined or dependent on t. In this expression, $\{\dot{\boldsymbol{x}}_1, \dot{\boldsymbol{x}}_2, \ldots, \dot{\boldsymbol{x}}_m\}$ is a subset of $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_t\}$, as often considered in the literature[1]. We denote by dictionary the set $\mathcal{D} = \{\kappa(\dot{\boldsymbol{x}}_1, \cdot), \kappa(\dot{\boldsymbol{x}}_2, \cdot), \ldots, \kappa(\dot{\boldsymbol{x}}_m, \cdot)\}$, and by atoms its elements. Throughout this paper, all quantities associated to the dictionary have an accent (by analogy to phonetics, where stress accents are associated to prominence). This is the case for instance of the m-by-m Gram matrix $\dot{\boldsymbol{K}}$ whose (i,j)-th entry is $\kappa(\dot{\boldsymbol{x}}_i, \dot{\boldsymbol{x}}_j)$. The eigenvalues of this matrix are denoted $\dot{\lambda}_1, \dot{\lambda}_2, \ldots, \dot{\lambda}_m$, given in non-increasing order.

The optimization problem is two-fold at each instant: selecting the proper dictionary $\mathcal{D} = \{\kappa(\dot{\boldsymbol{x}}_1, \cdot), \kappa(\dot{\boldsymbol{x}}_2, \cdot), \ldots, \kappa(\dot{\boldsymbol{x}}_m, \cdot)\}$ and estimating the corresponding parameters $\alpha_1, \alpha_2, \ldots, \alpha_m$. New challenges arise in an online learning setting. Determining the optimal dictionary at each instant is a combinatorial optimization problem, when optimality is measured by comparing the reduced-order solution (3) to its full-order form (2). The recursive update provides an elegant way to overcome this computationally intractable problem, by determining if the new kernel function $\kappa(\boldsymbol{x}_t, \cdot)$ needs to be included to the dictionary; otherwise, it can be discarded since it is efficiently approximated with atoms already belonging to the dictionary. This is the essence of the approximation criterion.

B. Approximation criterion

The (linear) approximation criterion was initially proposed in [29] for classification and regression, and in [30] for Gaussian processes. In online learning with kernels, as studied for system identification in [11] and more recently for kernel principal component analysis in [8], it operates as follows: the current sample is discarded (not included in the dictionary), if it can be sufficiently represented by a linear combination of atoms already belonging to the dictionary; otherwise, it is included in the dictionary. Formally, the kernel function $\kappa(\boldsymbol{x}_t, \cdot)$ is included in the dictionary if

$$\min_{\xi_1 \cdots \xi_m} \left\| \kappa(\boldsymbol{x}_t, \cdot) - \sum_{j=1}^{m} \xi_j \, \kappa(\dot{\boldsymbol{x}}_j, \cdot) \right\|_{\mathbb{H}}^2 \geq \delta^2, \tag{4}$$

where δ is a positive threshold parameter that controls the level of sparsity. The above norm is the residual error obtained by projecting $\kappa(\boldsymbol{x}_t, \cdot)$ onto the space spanned by the dictionary. The optimal value of each coefficient ξ_j is obtained by nullifying the derivative of the above cost function with respect to it, which leads to $\boldsymbol{\xi} = \dot{\boldsymbol{K}}^{-1} \dot{\boldsymbol{\kappa}}(\dot{\boldsymbol{x}}_t)$, where $\dot{\boldsymbol{\kappa}}(\boldsymbol{x}_t)$ is the column vector of entries $\kappa(\dot{\boldsymbol{x}}_j, \boldsymbol{x}_t)$, for $j = 1, 2, \ldots, m$. By inserting this expression into expression (4), we get the following condition of accepting the current kernel function:

$$\kappa(\boldsymbol{x}_t, \boldsymbol{x}_t) - \dot{\boldsymbol{\kappa}}(\boldsymbol{x}_t)^\top \dot{\boldsymbol{K}}^{-1} \dot{\boldsymbol{\kappa}}(\boldsymbol{x}_t) \geq \delta^2. \tag{5}$$

The resulting dictionary, called δ-approximate, satisfies the relation

$$\min_{i=1 \cdots m} \min_{\xi_1 \cdots \xi_m} \left\| \kappa(\dot{\boldsymbol{x}}_i, \cdot) - \sum_{\substack{j=1 \\ j \neq i}}^{m} \xi_j \, \kappa(\dot{\boldsymbol{x}}_j, \cdot) \right\|_{\mathbb{H}} \geq \delta.$$

One could also include a removal process, in the same spirit as the fixed-budget concept, by discarding the atom that can be well approximated with the other atoms, as investigated for instance in [31]. Nonetheless, the dictionary is still δ-approximate. The use of a removal process does not affect the results given in this paper.

C. Issues studied in this paper

In the following, we describe several issues that motivate (and structure) this work, illustrated here with respect to the approximation criterion.

Computational complexity

The approximation criterion requires the inversion of the Gram matrix associated to the dictionary, which is the most computationally expensive process. Its computational complexity scales cubically with the size of the dictionary, *i.e.*, $\mathcal{O}(m^3)$ operations, and can be reduced to $\mathcal{O}(m^2)$ by using a recursive rule when a single element is included in the dictionary. Moreover, the evaluation of the condition expressed in (5) requires two matrix multiplications at each instant. These computational costs counteract the benefits of several online learning techniques, such as gradient-based and least-mean-square algorithms (*e.g.*, LMS, NLMS, affine projection, ...).

To reduce the computational burden of the approximation criterion, several computationally efficient sparsification criteria have been proposed in the literature, sharing essentially the same computational complexity that scales linearly with the size of the dictionary, *i.e.*, $\mathcal{O}(m)$ operations at each instant. The most known criteria are the distance and the coherence criteria; see Section III for a description.

[1]One may also relax the constraint that dictionary elements must be a subset of the set of available samples $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_t\}$. This general case can also be considered within the framework proposed in this paper, as long as the investigated dictionary is quantified using any of the diversity/sparsity measures, such as the coherence, the approximation or the distance measures [26]. For instance in [27], [28], the dictionary elements are updated with a constrained stochastic gradient algorithm in order to minimize the prediction error subject to a bounded coherence measure.

Approximation error of any sample

The approximation criterion relies on establishing a dictionary such that the error of approximating each of its atoms, with a linear combination of the other atoms, cannot be smaller than the given threshold δ. Moreover, the decision of discarding any sample from the dictionary is defined by the same process, namely when its approximation error, with a linear combination of the other atoms, is smaller than the same threshold δ. While the approximation criterion possesses such duality between accepting and discarding samples at the very same value of thresholding the approximation error, this is not the case of the other sparsification criteria.

In Section IV, we bridge the gap between the approximation criterion and the other online sparsification criteria. For this purpose, on one hand, we establish in Section IV-A upper bounds on the error of approximating a discarded sample with atoms of a dictionary obtained by the distance or the coherence criterion. One the other hand, we provide in Section IV-B lower bounds on the error of approximating any atom with the other atoms of the dictionary under scrutiny.

From approximating samples to approximating patterns

All the aforementioned sparsification criteria operate in a pre-processing scheme, by selecting samples independently of the resulting sparse approximation of the full-order pattern. In other words, the selection of the relevant subset $\{\dot{\boldsymbol{x}}_1, \dot{\boldsymbol{x}}_2, \ldots, \dot{\boldsymbol{x}}_m\}$ from the set $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_t\}$ is only based on the topology of the samples; it is independent of the power of the dictionary to approximate accurately any pattern of the form (2) with the reduced-order model (3).

In Section V, we study the relevance of approximating any pattern with a dictionary obtained by any online sparsification criterion, including the approximation criterion. We establish upper bounds on the approximation error of any pattern, before examining in detail two particular classes of patterns, the centroid as studied in Section V-A and the most relevant principal axes in kernel-PCA investigated in Section V-B.

III. Online sparsification criteria

With a novel sample $\boldsymbol{x}_t$ available at instant t, a sparsification rule determines if $\kappa(\boldsymbol{x}_t, \cdot)$ should be included in the dictionary, by incrementing the model order m and setting $\dot{\boldsymbol{x}}_{m+1} = \boldsymbol{x}_t$. The sparsification criteria measure the relevance of such complexity-incrementation by comparing the current kernel function $\kappa(\boldsymbol{x}_t, \cdot)$ with the atoms of the dictionary. They are defined by either a dissimilarity measure, *i.e.*, constructing the dictionary with the most mutually distant atoms, or a similarity measure, *i.e.*, constructing the dictionary with the least coherent or correlated atoms. To this end, a threshold is used to control the level of sparsity of the dictionary. The most investigated criteria are outlined in the following.

A. Distance criterion

It is natural to propose a sparsification criterion that constructs a dictionary with large distances between its entries, thus discarding samples that are too close to any of the atoms already belonging to the dictionary. The current kernel function $\kappa(\boldsymbol{x}_t, \cdot)$ is included in the dictionary if

$$\min_{j=1\cdots m} \min_{\xi} \|\kappa(\boldsymbol{x}_t, \cdot) - \xi\, \kappa(\dot{\boldsymbol{x}}_j, \cdot)\|_{\mathbb{H}} \geq \delta, \tag{6}$$

for a predefined positive threshold δ; otherwise, it can be efficiently approximated, up to a multiplicative constant, with an atom of the dictionary. It is easy to see that the optimal value of the scaling factor ξ is $\kappa(\boldsymbol{x}_t, \dot{\boldsymbol{x}}_j)/\kappa(\dot{\boldsymbol{x}}_j, \dot{\boldsymbol{x}}_j)$, since the left-hand-side of (6) is the residual error on projecting $\kappa(\boldsymbol{x}_t, \cdot)$ onto $\kappa(\dot{\boldsymbol{x}}_j, \cdot)$ (in the same spirit as the approximation criterion). This allows to simplify the condition (6) to get

$$\min_{j=1\cdots m} \left(\kappa(\boldsymbol{x}_t, \boldsymbol{x}_t) - \frac{\kappa(\boldsymbol{x}_t, \dot{\boldsymbol{x}}_j)^2}{\kappa(\dot{\boldsymbol{x}}_j, \dot{\boldsymbol{x}}_j)} \right) \geq \delta^2. \tag{7}$$

The resulting dictionary, called δ-distant, satisfies for any pair $(\dot{\boldsymbol{x}}_i, \dot{\boldsymbol{x}}_j)$:

$$\kappa(\dot{\boldsymbol{x}}_i, \dot{\boldsymbol{x}}_i) - \frac{\kappa(\dot{\boldsymbol{x}}_i, \dot{\boldsymbol{x}}_j)^2}{\kappa(\dot{\boldsymbol{x}}_j, \dot{\boldsymbol{x}}_j)} \geq \delta^2. \tag{8}$$

For unit-norm atoms, this expression reduces to the condition $|\kappa(\dot{\boldsymbol{x}}_i, \dot{\boldsymbol{x}}_j)| \leq \sqrt{1-\delta^2}$. This sparsification criterion has been extensively used in the literature under different names, such as the novelty criterion proposed in [4] (where the scaling factor was dropped and a prediction error mechanism was included in a second stage; see also [32], [7]) and the quantized criterion described in [33].

B. Coherence criterion

The coherence measure has been extensively studied in the literature of compressed sensing in the particular case of the linear kernel with unit-norm samples [17], [18]. In the more general case with the kernel formalism, a dictionary is γ-coherent if

$$\max_{\substack{i,j=1\cdots m \\ i\neq j}} \frac{|\kappa(\dot{\boldsymbol{x}}_i, \dot{\boldsymbol{x}}_j)|}{\sqrt{\kappa(\dot{\boldsymbol{x}}_i, \dot{\boldsymbol{x}}_i)\, \kappa(\dot{\boldsymbol{x}}_j, \dot{\boldsymbol{x}}_j)}} \leq \gamma, \tag{9}$$

which corresponds to the largest value of the cosine angle between any pair of atoms, since the above objective function can be written as

$$\frac{|\langle \kappa(\dot{\boldsymbol{x}}_i, \cdot), \kappa(\dot{\boldsymbol{x}}_j, \cdot) \rangle_{\mathbb{H}}|}{\|\kappa(\dot{\boldsymbol{x}}_i, \cdot)\|_{\mathbb{H}} \|\kappa(\dot{\boldsymbol{x}}_j, \cdot)\|_{\mathbb{H}}}.$$

The coherence criterion, introduced in [15], [16] and studied more recently in [34], [35], constructs a dictionary with atoms that are mutually least coherent, by restricting this measure below some predefined value $\gamma \in [0\,;1]$, where the null value yields an orthogonal basis. This criterion includes the current kernel function $\kappa(\boldsymbol{x}_t, \cdot)$ in the dictionary if

$$\max_{j=1\cdots m} \frac{|\kappa(\boldsymbol{x}_t, \dot{\boldsymbol{x}}_j)|}{\sqrt{\kappa(\boldsymbol{x}_t, \boldsymbol{x}_t)\, \kappa(\dot{\boldsymbol{x}}_j, \dot{\boldsymbol{x}}_j)}} \leq \gamma. \tag{10}$$

It is worth noting that the denominator in each of the above expressions reduces to 1 when dealing with unit-norm atoms, thus expression (10) becomes

$$\max_{j=1\cdots m} |\kappa(\boldsymbol{x}_t, \dot{\boldsymbol{x}}_j)| \leq \gamma.$$

In this case, it turns out that this criterion is equivalent to the distance criterion with the threshold set to $\delta = \sqrt{1-\gamma^2}$.

IV. Approximation error of any sample

In this section, we study the elementary issue of approximating a sample by the span of a dictionary. To this end, this issue is considered in its two folds: the error of approximating a discarded sample, and the error of approximating any accepted sample, namely approximating any atom of the dictionary with all the other atoms. We provide upper bounds on the former and lower bounds on the latter, for each of the sparsification criteria studied in previous section. It is worth noting that only the approximation criterion relies on a duality of discarding and accepting samples at the very same value in thresholding the approximation error.

Let $\hat{\mathcal{P}}$ be the projection operator onto the subspace spanned by the atoms $\kappa(\hat{\boldsymbol{x}}_1,\cdot),\ldots,\kappa(\hat{\boldsymbol{x}}_m,\cdot)$ of a dictionary resulting from a sparsification criterion. Thus, for any sample $\boldsymbol{x}$, the projection of the kernel function $\kappa(\boldsymbol{x},\cdot)$ onto this subspace is given by $\hat{\mathcal{P}}\kappa(\boldsymbol{x},\cdot)$. The norm of $\hat{\mathcal{P}}\kappa(\boldsymbol{x},\cdot)$ is obtained by the maximum inner product $\langle\kappa(\boldsymbol{x},\cdot),\varphi(\cdot)\rangle_{\mathbb{H}}$ over all the unit-norm functions $\varphi(\cdot)$ of that subspace. By writing $\varphi(\cdot) = \sum_{j=1}^m \beta_j\kappa(\hat{\boldsymbol{x}}_j,\cdot)/\|\sum_{j=1}^m \beta_j\kappa(\hat{\boldsymbol{x}}_j,\cdot)\|_{\mathbb{H}}$, one gets

$$\begin{aligned}\|\hat{\mathcal{P}}\kappa(\boldsymbol{x},\cdot)\|_{\mathbb{H}} &= \max_{\boldsymbol{\beta}} \frac{\langle\sum_{j=1}^m \beta_j\kappa(\hat{\boldsymbol{x}}_j,\cdot),\kappa(\boldsymbol{x},\cdot)\rangle_{\mathbb{H}}}{\|\sum_{j=1}^m \beta_j\kappa(\hat{\boldsymbol{x}}_j,\cdot)\|_{\mathbb{H}}} \\ &= \max_{\boldsymbol{\beta}} \frac{\sum_{j=1}^m \beta_j\kappa(\boldsymbol{x},\hat{\boldsymbol{x}}_j)}{\|\sum_{j=1}^m \beta_j\kappa(\hat{\boldsymbol{x}}_j,\cdot)\|_{\mathbb{H}}}. \end{aligned} \tag{11}$$

Moreover, the Pythagorean Theorem allows to write $\|(\mathbf{I}-\hat{\mathcal{P}})\kappa(\boldsymbol{x},\cdot)\|_{\mathbb{H}}^2 = \kappa(\boldsymbol{x},\boldsymbol{x}) - \|\hat{\mathcal{P}}\kappa(\boldsymbol{x},\cdot)\|_{\mathbb{H}}^2$, where $\mathbf{I}$ is the identity operator. Thus, the quadratic approximation error is

$$\|(\mathbf{I}-\hat{\mathcal{P}})\kappa(\boldsymbol{x},\cdot)\|_{\mathbb{H}}^2 = \kappa(\boldsymbol{x},\boldsymbol{x}) - \max_{\boldsymbol{\beta}} \frac{\big(\sum_{j=1}^m \beta_j\kappa(\boldsymbol{x},\hat{\boldsymbol{x}}_j)\big)^2}{\|\sum_{j=1}^m \beta_j\kappa(\hat{\boldsymbol{x}}_j,\cdot)\|_{\mathbb{H}}^2}. \tag{12}$$

Next, we shall investigate this expression to establish bounds on approximating either discarded samples or included ones.

A. Approximation error of discarded samples

When the sample $\boldsymbol{x}_t$ is discarded, we propose to upper bound the quadratic approximation error (12) with

$$\|(\mathbf{I}-\hat{\mathcal{P}})\kappa(\boldsymbol{x}_t,\cdot)\|_{\mathbb{H}}^2 \leq \kappa(\boldsymbol{x}_t,\boldsymbol{x}_t) - \max_j \frac{\kappa(\boldsymbol{x}_t,\hat{\boldsymbol{x}}_j)^2}{\kappa(\hat{\boldsymbol{x}}_j,\hat{\boldsymbol{x}}_j)}, \tag{13}$$

where the inequality corresponds to the special choice of the coefficients, with $\beta_1 = \ldots = \beta_m = 0$ except for $\beta_j = \pm 1$. Next, we show that the above quotient is lower-bounded, by examining separately the distance and the coherence criteria.

Theorem 1 (Discarding error for the distance criterion): Let $\boldsymbol{x}_t$ be a sample not satisfying the distance condition (7) for some given threshold δ. The quadratic error of approximating $\kappa(\boldsymbol{x}_t,\cdot)$ with a linear combination of atoms from the resulting dictionary cannot exceed δ^2.

Proof: The proof is straightforward, since we have

$$\kappa(\boldsymbol{x}_t,\boldsymbol{x}_t) - \max_{j=1\ldots m} \frac{\kappa(\boldsymbol{x}_t,\hat{\boldsymbol{x}}_j)^2}{\kappa(\hat{\boldsymbol{x}}_j,\hat{\boldsymbol{x}}_j)} = \min_{j=1\ldots m} \kappa(\boldsymbol{x}_t,\boldsymbol{x}_t) - \frac{\kappa(\boldsymbol{x}_t,\hat{\boldsymbol{x}}_j)^2}{\kappa(\hat{\boldsymbol{x}}_j,\hat{\boldsymbol{x}}_j)},$$

which is upper-bounded by δ^2 for any $\boldsymbol{x}_t$ that does not satisfy the condition (6)-(7). As a consequence, the right-hand-side of inequality (13) is also upper-bounded by δ^2. ∎

Theorem 2 (Discarding error for the coherence criterion): Let $\boldsymbol{x}_t$ be a sample not satisfying the coherence (10) for some threshold γ. The quadratic error of approximating $\kappa(\boldsymbol{x}_t,\cdot)$ with a linear combination of the dictionary atoms cannot exceed $\kappa(\boldsymbol{x}_t,\boldsymbol{x}_t)(1-\gamma^2)$, and $R^2(1-\gamma^2)$ for all samples.

Proof: The unfulfilled coherence condition (10) can be written in the equivalent form

$$\max_{j=1\cdots m} \frac{|\kappa(\boldsymbol{x}_t,\hat{\boldsymbol{x}}_j)|}{\sqrt{\kappa(\hat{\boldsymbol{x}}_j,\hat{\boldsymbol{x}}_j)}} > \gamma\sqrt{\kappa(\boldsymbol{x}_t,\boldsymbol{x}_t)}.$$

By inserting this inequality into (13), we get $\|(\mathbf{I}-\hat{\mathcal{P}})\kappa(\boldsymbol{x}_t,\cdot)\|_{\mathbb{H}}^2 < \kappa(\boldsymbol{x}_t,\boldsymbol{x}_t) - \gamma^2\,\kappa(\boldsymbol{x}_t,\boldsymbol{x}_t)$. ∎

B. Approximation error of an atom from the dictionary

Next, we study the approximation error of an atom of a dictionary with a linear combination of its other atoms.

Consider projecting an atom $\kappa(\hat{\boldsymbol{x}}_i,\cdot)$ of the dictionary onto the span of the other $m-1$ atoms. By following the same derivations as in the beginning of Section IV, we have

$$\|(\mathbf{I}-\hat{\mathcal{P}})\kappa(\hat{\boldsymbol{x}}_i,\cdot)\|_{\mathbb{H}}^2 = \kappa(\hat{\boldsymbol{x}}_i,\hat{\boldsymbol{x}}_i) - \max_{\boldsymbol{\beta}_{\setminus\{i\}}} \frac{\big(\sum_{j=1,j\neq i}^m \beta_j\kappa(\hat{\boldsymbol{x}}_i,\hat{\boldsymbol{x}}_j)\big)^2}{\|\sum_{j=1,j\neq i}^m \beta_j\kappa(\hat{\boldsymbol{x}}_j,\cdot)\|_{\mathbb{H}}^2}.$$

On one hand, the numerator in the above expression is upper-bounded, since we have from the Cauchy-Schwarz inequality:

$$\left(\sum_{\substack{j=1\\j\neq i}}^m \beta_j\kappa(\hat{\boldsymbol{x}}_i,\hat{\boldsymbol{x}}_j)\right)^2 \leq \sum_{\substack{j=1\\j\neq i}}^m \beta_j^2 \sum_{\substack{j=1\\j\neq i}}^m |\kappa(\hat{\boldsymbol{x}}_i,\hat{\boldsymbol{x}}_j)|^2.$$

On the other hand, the denominator has a lower bound, since

$$\Big\|\sum_{\substack{j=1\\j\neq i}}^m \beta_j\kappa(\hat{\boldsymbol{x}}_j,\cdot)\Big\|_{\mathbb{H}}^2 = \boldsymbol{\beta}_{\setminus\{i\}}^\top \hat{\boldsymbol{K}}_{\setminus\{i\}}\boldsymbol{\beta}_{\setminus\{i\}} \geq \hat{\lambda}_{\setminus\{i\}m-1}\|\boldsymbol{\beta}_{\setminus\{i\}}\|^2,$$

where $\hat{\boldsymbol{K}}_{\setminus\{i\}}$ is the $(m-1)$-by-$(m-1)$ submatrix of the matrix $\hat{\boldsymbol{K}}$ obtained by removing its i-th row and its i-th column, *i.e.*, the entries associated to $\hat{\boldsymbol{x}}_i$, and $\hat{\lambda}_{\setminus\{i\}m-1}$ is its smallest eigenvalue. By combining these two inequalities, we get

$$\|(\mathbf{I}-\hat{\mathcal{P}})\kappa(\hat{\boldsymbol{x}}_i,\cdot)\|_{\mathbb{H}}^2 \geq \kappa(\hat{\boldsymbol{x}}_i,\hat{\boldsymbol{x}}_i) - \frac{1}{\hat{\lambda}_{\setminus\{i\}m-1}} \sum_{\substack{j=1\\j\neq i}}^m |\kappa(\hat{\boldsymbol{x}}_i,\hat{\boldsymbol{x}}_j)|^2. \tag{14}$$

For each sparsification criterion, we shall write this lower bound by using the corresponding summation expression and the appropriate lower bound on the eigenvalues, as obtained in [26, Section IV] and summarized in the appendix.

Theorem 3 (Acceptance error for the distance criterion): For a δ-distant dictionary, the quadratic error of approximating any atom $\kappa(\hat{\boldsymbol{x}}_i,\cdot)$ with a linear combination of the other atoms is lower-bounded by

$$\kappa(\hat{\boldsymbol{x}}_i,\hat{\boldsymbol{x}}_i) - \frac{\big(\kappa(\hat{\boldsymbol{x}}_i,\hat{\boldsymbol{x}}_i)-\delta^2\big)(m-1)R^2}{r^2-(m-2)R\sqrt{R^2-\delta^2}}.$$

For unit-norm atoms, we get a lower bound for all atoms.

Proof: The proof is split in two parts, by investigating expression (14). Firstly, the summation term is upper-bounded since, from (8), we have that any pair $(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_j)$ satisfies

$$\begin{aligned}\sum_{\substack{j=1\\j\neq i}}^{m} |\kappa(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_j)|^2 &\leq \sum_{\substack{j=1\\j\neq i}}^{m} \kappa(\grave{\boldsymbol{x}}_j, \grave{\boldsymbol{x}}_j)\left(\kappa(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_i) - \delta^2\right)\\ &= \left(\kappa(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_i) - \delta^2\right) \sum_{\substack{j=1\\j\neq i}}^{m} \kappa(\grave{\boldsymbol{x}}_j, \grave{\boldsymbol{x}}_j)\\ &= \left(\kappa(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_i) - \delta^2\right)(m-1) \max_{\substack{j=1\cdots m\\j\neq i}} \kappa(\grave{\boldsymbol{x}}_j, \grave{\boldsymbol{x}}_j)\\ &= \left(\kappa(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_i) - \delta^2\right)(m-1)R^2.\end{aligned}$$

Secondly, the eigenvalue in this expression is lower-bounded by $r^2-(m-2)R\sqrt{R^2-\delta^2}$ for a δ-distant dictionary of $m-1$ atoms, as shown in Lemma A.1 of the Appendix. ∎

Theorem 4 (Acceptance error for the coherence criterion): For a γ-coherent dictionary, the quadratic error of approximating any atom $\kappa(\grave{\boldsymbol{x}}_i, \cdot)$ with a linear combination of the other atoms is lower-bounded by

$$\kappa(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_i) - \frac{(m-1)\,\gamma^2 R^2 \kappa(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_i)}{r^2 - (m-2)\gamma R^2}.$$

For unit-norm atoms, this bounds is independent of $\grave{\boldsymbol{x}}_i$.

Proof: Following the same steps as in the previous proof,

$$\begin{aligned}\sum_{\substack{j=1\\j\neq i}}^{m} |\kappa(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_j)|^2 &\leq (m-1) \max_{\substack{j=1\cdots m\\j\neq i}} |\kappa(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_j)|^2\\ &\leq (m-1)\,\gamma^2 \max_{\substack{j=1\cdots m\\j\neq i}} \kappa(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_i)\,\kappa(\grave{\boldsymbol{x}}_j, \grave{\boldsymbol{x}}_j)\\ &\leq (m-1)\,\gamma^2 R^2 \kappa(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_i),\end{aligned}$$

where the second inequality follows from the coherence condition. On the other hand, we use the lower bound $r^2-(m-2)\gamma R^2$ on the eigenvalues associated to a γ-coherent dictionary of $m-1$ atoms, as given in Lemma A.2 of the Appendix. ∎

C. Analyzing the tightness of the obtained bounds

In this section, we study the relevance of these theorems in terms of tightness of the resulting bounds.

C.1 Revisiting the proofs given in Section IV-A

The theorems given in Section IV-A rely on (13), which is obtained from (12) as a special choice of the coefficient vector $\boldsymbol{\beta}$. It turns out that this choice provides relevant bounds that can be obtained using proofs independent of this choice, as demonstrated in the following alternative proof of Theorem 1:

Another proof of Theorem 1: The approximation error under scrutiny is upper-bounded as follows:

$$\begin{aligned}\min_{\xi_1\cdots\xi_m} \Big\|\kappa(\boldsymbol{x}_t, \cdot) - \sum_{i=1}^{m} \xi_i\,\kappa(\grave{\boldsymbol{x}}_i, \cdot)\Big\|_{\mathrm{H}} &\leq \min_{j=1\cdots m} \min_{\xi_j} \Big\|\kappa(\boldsymbol{x}_t, \cdot) - \xi_j\,\kappa(\grave{\boldsymbol{x}}_j, \cdot)\Big\|_{\mathrm{H}}\\ &< \delta,\end{aligned}$$

where the last inequality is due to the violation of (6). ∎

C.2 Revisiting Section IV-A with the Geršgorin Discs Theorem

Section IV-A studies the error of approximating discarded samples. Next, we show that other bounds (but not as sharp) can be obtained by using the Geršgorin Discs Theorem, by following the same steps as in Section IV-B. To this end, we write the quadratic approximation error as in Section II-B with $\kappa(\boldsymbol{x}_t, \boldsymbol{x}_t) - \grave{\boldsymbol{\kappa}}(\boldsymbol{x}_t)^\top \grave{\boldsymbol{K}}^{-1} \grave{\boldsymbol{\kappa}}(\boldsymbol{x}_t)$. If $\grave{\lambda}_1$ denotes the largest eigenvalue of $\grave{\boldsymbol{K}}$, then $1/\grave{\lambda}_1$ corresponds to the smallest eigenvalue of $\grave{\boldsymbol{K}}^{-1}$. Therefore, $\|\grave{\boldsymbol{\kappa}}(\boldsymbol{x}_t)\|^2/\grave{\lambda}_1 \leq \grave{\boldsymbol{\kappa}}(\boldsymbol{x}_t)^\top \grave{\boldsymbol{K}}^{-1} \grave{\boldsymbol{\kappa}}(\boldsymbol{x}_t)$, and

$$\kappa(\boldsymbol{x}_t, \boldsymbol{x}_t) - \grave{\boldsymbol{\kappa}}(\boldsymbol{x}_t)^\top \grave{\boldsymbol{K}}^{-1} \grave{\boldsymbol{\kappa}}(\boldsymbol{x}_t) \leq \kappa(\boldsymbol{x}_t, \boldsymbol{x}_t) - \frac{\|\grave{\boldsymbol{\kappa}}(\boldsymbol{x}_t)\|^2}{\grave{\lambda}_1}.$$

In the following, we consider the case of the coherence criterion with a unit-norm kernel, while extensions to the general case is straightforward. When dealing with the coherence criterion, we have $\|\grave{\boldsymbol{\kappa}}(\boldsymbol{x}_t)\|^2 \geq \gamma^2$ and $\grave{\lambda}_1 \leq 1+(m-1)\gamma$ as given in Lemma A.2. By combining these results, we get

$$\kappa(\boldsymbol{x}_t, \boldsymbol{x}_t) - \grave{\boldsymbol{\kappa}}(\boldsymbol{x}_t)^\top \grave{\boldsymbol{K}}^{-1} \grave{\boldsymbol{\kappa}}(\boldsymbol{x}_t) \leq 1 - \frac{\gamma^2}{1+(m-1)\gamma}.$$

It is easy to see that this upper bound is not as tight as the one obtained in Theorem 2 with $1-\gamma^2$. We also get similar results for the distance criterion, with the upper bound

$$1 - \frac{1-\delta^2}{1+(m-1)\sqrt{1-\delta^2}},$$

which is looser than the one given in Theorem 1.

These results demonstrate once again the relevance of the theorems given in Section IV-A.

C.3 Relevance of the bounds given in Section IV-B

Results given in Section IV-B rely on relation (14), and as a consequence on lower-bounding the eigenvalues of the Gram matrix. For this purpose, the well-known Geršgorin Discs Theorem [36, Chapter 6] is investigated (see the Appendix).

Unfortunately, the Geršgorin Discs Theorem may provide negative lower bounds, which yields meaningless results since the Gram matrix is positive definite. However, we can provide a natural condition to overcome this drawback, by imposing positive denominators in Theorems 3 and 4. Consider for instance Theorem 4, then condition $r^2-(m-2)\gamma R^2 > 0$ yields $m < 2-r^2/(\gamma R^2)$. It is worth noting that this sufficient condition is less restrictive than the sufficient condition to have a dictionary of linear independent atoms as given in [26, Theorem 7], which is $m < 1-r^2/(\gamma R^2)$. As a consequence, when one uses the latter to impose the linear independency condition, the bounds given in Theorems 3 and 4 are relevant.

C.4 On providing sharper bounds

The bounds obtained in Sections IV-A and IV-B are sharp as shown in Section VI with experimental results. The quality of the bounds given in Section IV-B depends on lower-bounding the eigenvalues. While we have used the well-known Geršgorin Discs Theorem for this purpose, one can easily substitute it with any novel bound in the literature, such as the positive lower bounds obtained for positive definite matrices in [37], [38], [39]. See also [40], [41], [42] for recent results. This active research activity is beyond the scope of this paper.

V. Approximation of any pattern

In this section, we study the relevance of approximating any pattern with its projection onto the subspace spanned by the atoms of a dictionary. An upper bound on the approximation error is derived in the following theorem for any sparsification criterion, while specific bounds in term of the threshold of each criterion are given in the following Theorem 6. Moreover, these results are explored in two particular kernel-based learning algorithms, with the centroid (see Section V-A) and the principal axes (see Section V-B) as patterns to be estimated.

Theorem 5: Consider the approximation of any $\psi(\cdot) = \sum_{i=1}^n \alpha_i \, \kappa(\boldsymbol{x}_i, \cdot)$ with its projection onto the subspace spanned by the m atoms of a given dictionary. The quadratic error of such approximation is upper-bounded by

$$(n-m) \; \|\boldsymbol{\alpha}\|^2 \, \epsilon^2,$$

where ϵ is an upper bound on the approximation of any $\kappa(\boldsymbol{x}_i, \cdot)$ with a linear combination of atoms from the dictionary.

Proof: The approximation error is upper-bounded, with

$$\begin{aligned} \|(\mathbf{I} - \hat{\mathcal{P}})\psi(\cdot)\|_{\mathbb{H}} &= \|\sum_{i=1}^n \alpha_i \, (\mathbf{I} - \hat{\mathcal{P}}) \, \kappa(\boldsymbol{x}_i, \cdot)\|_{\mathbb{H}} \\ &\leq \sum_{i=1}^n |\alpha_i| \, \|(\mathbf{I} - \hat{\mathcal{P}}) \, \kappa(\boldsymbol{x}_i, \cdot)\|_{\mathbb{H}}, \end{aligned} \tag{15}$$

where the inequality is due to the generalized triangular inequality. By applying the Cauchy-Schwarz inequality, we get the quadratic approximation error

$$\|(\mathbf{I} - \hat{\mathcal{P}})\psi(\cdot)\|_{\mathbb{H}}^2 \leq \sum_{i=1}^n \alpha_i^2 \;\; \sum_{i=1}^n \|(\mathbf{I} - \hat{\mathcal{P}}) \, \kappa(\boldsymbol{x}_i, \cdot)\|_{\mathbb{H}}^2. \tag{16}$$

The first summation is the quadratic ℓ_2-norm of the vector of coefficients, namely $\|\boldsymbol{\alpha}\|^2$. For the second summation, we separate it in two terms, entries belonging to the dictionary and those discarded thanks to the used sparsification criterion. While the former do not contribute to the error, only the latter take part in the summation, namely the $n-m$ discarded samples where m is the size of the dictionary. Let ϵ^2 be an upper bound on the quadratic error of discarding samples, as given in Section IV-A. Then, we get $\|(\mathbf{I} - \hat{\mathcal{P}})\psi(\cdot)\|_{\mathbb{H}}^2 \leq (n-m) \, \|\boldsymbol{\alpha}\|^2 \; \epsilon^2$, which concludes the proof. ∎

By revisiting the upper bounds given in Section IV-A, the proof of the following theorem is straightforward.

Theorem 6: The upper bound given in Theorem 5 can be specified for each sparsification criterion, as follows:

- $(n-m)\|\boldsymbol{\alpha}\|^2\delta^2$ for the δ-distant criterion.
- $(n-m)\|\boldsymbol{\alpha}\|^2\delta^2$ for the δ-approximate criterion.
- $(n-m)\|\boldsymbol{\alpha}\|^2 R^2(1-\gamma^2)$ for the γ-coherent criterion.

We shall explore next these results for two specific learning algorithms, in order to clarify the relevance of these bounds.

A. Approximation of the centroid

The centroid (*i.e.*, empirical mean) is a fundamental pattern of the set of sample, and its use is essential in many statistical methods such as in [23] for visualization and clustering and in [22], [43] for one-class classification. In the following, we study the relevance of approximating the centroid by its projection onto the subspace spanned by the dictionary atoms. Let $\psi(\cdot) = \frac{1}{n} \sum_{i=1}^n \kappa(\boldsymbol{x}_i, \cdot)$ be the centroid, namely $\alpha_i = 1/n$ for all i. From Theorem 5, we get

$$\|(\mathbf{I} - \hat{\mathcal{P}})\psi(\cdot)\|_{\mathbb{H}}^2 \leq \left(1 - \frac{m}{n}\right) \epsilon^2, \tag{17}$$

where $\max_i \|(\mathbf{I} - \hat{\mathcal{P}})\kappa(\boldsymbol{x}_i, \cdot)\|_{\mathbb{H}} \leq \epsilon$.

It turns out that we can give a sharper bound by relaxing the Cauchy-Schwarz inequality used in (16), since that the α_i are constant. As a result, we get by revisiting expression (15):

$$\begin{aligned} \|(\mathbf{I} - \hat{\mathcal{P}})\psi(\cdot)\|_{\mathbb{H}} &\leq \sum_{i=1}^n |\alpha_i| \|(\mathbf{I} - \hat{\mathcal{P}}) \, \kappa(\boldsymbol{x}_i, \cdot)\|_{\mathbb{H}} \\ &= \frac{1}{n} \sum_{i=1}^n \|(\mathbf{I} - \hat{\mathcal{P}}) \, \kappa(\boldsymbol{x}_i, \cdot)\|_{\mathbb{H}} \\ &\leq \frac{1}{n} \, (n-m) \, \epsilon, \end{aligned}$$

where we have followed the same decomposition as in the proof of Theorem 5, with only the $n-m$ discarded samples contribute to the summation term. Therefore, the quadratic approximation error is upper-bounded as follows:

$$\|(\mathbf{I} - \hat{\mathcal{P}})\psi(\cdot)\|_{\mathbb{H}}^2 \leq \left(1 - \frac{m}{n}\right)^2 \epsilon^2.$$

This bound is sharper than the one in (17) since $1 - \frac{m}{n} < 1$.

By revisiting Theorem 6 in the light of this result, the upper bound on the quadratic approximation error is given in terms of the threshold of each sparsification criterion, as follows:

- $\left(1 - \frac{m}{n}\right)^2 \delta^2$ for the δ-distant criterion.
- $\left(1 - \frac{m}{n}\right)^2 \delta^2$ for the δ-approximate criterion.
- $\left(1 - \frac{m}{n}\right)^2 R^2(1-\gamma^2)$ for the γ-coherent criterion.

These results generalize the work in [21], where only the case of the coherence criterion was studied for unit-norm atoms.

B. Approximation of the most relevant principal axes

Any sparsification criterion can be seen as a dimensionality reduction technique, because it identifies a subspace by selecting relevant samples from the available ones. Since it is an unsupervised approach, it is natural to connect it with PCA.

PCA seeks the principal axes that capture most of the data[2] variance. These axes correspond to the eigenvectors associated to the largest eigenvalues (called principal values) of the covariance matrix. In its kernel-PCA formulation, the k-th principal axis takes the form $\psi_k(\cdot) = \sum_{i=1}^n \alpha_{i,k} \, \kappa(\boldsymbol{x}_i, \cdot)$, where the $\alpha_{i,k}$ are the entries of the eigenvector associated to the k-th eigenvalue λ_k of the Gram matrix $\boldsymbol{K}$. Moreover, to get unit-norm principal axes, these coefficients are normalized such that $\sum_{i=1}^n \alpha_{i,k}^2 = 1/n\lambda_k$.

Theorem 7: Consider the principal axis associated to the eigenvalue λ_k of the corresponding Gram matrix. The quadratic error of approximating it by its projection onto the span of a dictionary of m atoms cannot exceed $\left(1 - \frac{m}{n}\right) \frac{\epsilon^2}{\lambda_k}$,

[2]Data are assumed centered; otherwise, use the connections given in [44].

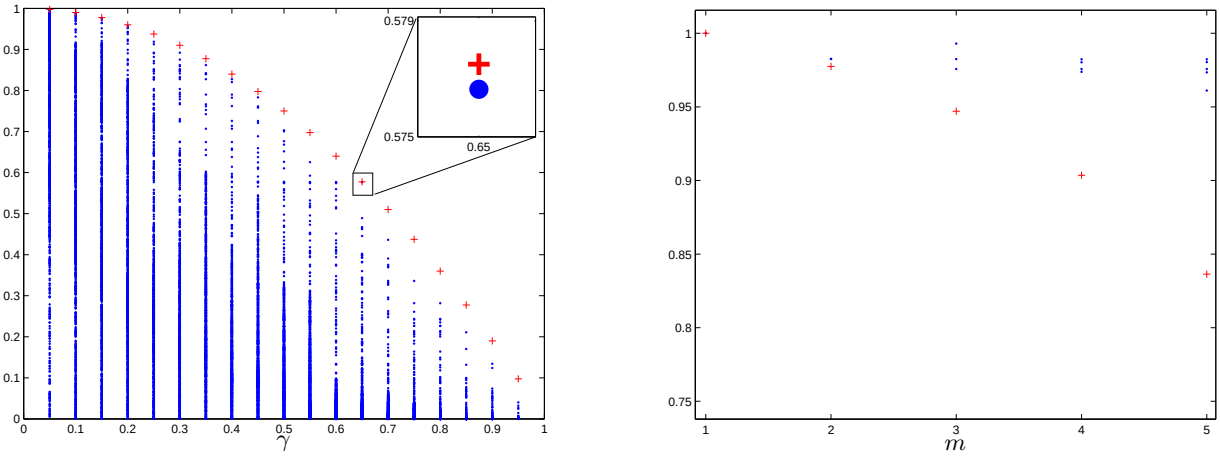

Fig. 2. Illustration of the bounds (+) on the error of approximating any sample (·). Left figure deals with the error of discarding samples as given in Theorem 2 for several values of the coherence threshold γ, while right figure illustrates the lower bounds on the error of approximating any atom with the other atoms of the dictionary, as given in Theorem 4, when $\gamma = 0.35$ and the dictionary size m is increasing.

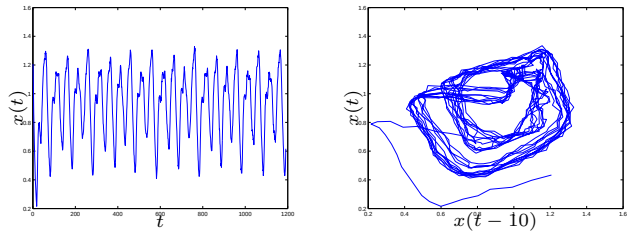

Fig. 1. The Mackey-Glass data, as a time series with $(t, x(t))$ (left figure) and as a time delay embedding $(x(t-10), x(t))$ (right figure).

where ϵ is an upper bound on the approximation of any $\kappa(\boldsymbol{x}_i, \cdot)$ with a linear combination of atoms from the dictionary.

The proof is due to Theorem 5, since $\|\boldsymbol{\alpha}\|^2 = 1/n\lambda_k$. Theorem 7 shows that, under the only condition that the used dictionary has an upper bound on the error of approximating each kernel function, the principal axes associated to the largest principal values have the smallest approximation errors. One can therefore say that the most relevant principal axes lie, with a small error, in the span of the dictionary.

Moreover, we establish expressions for each sparsification criterion, as given next in terms of the used threshold:

- $\left(1 - \frac{m}{n}\right) \frac{\delta^2}{\lambda_k}$ for the δ-distant criterion.
- $\left(1 - \frac{m}{n}\right) \frac{\delta^2}{\lambda_k}$ for the δ-approximate criterion.
- $\left(1 - \frac{m}{n}\right) R^2 \frac{1-\gamma^2}{\lambda_k}$ for the γ-coherent criterion.

These results generalize previous work on the approximation and the coherence criteria given for unit-norm atoms, and provide sharper bounds than the ones previously known in the literature. Indeed, the upper bound δ^2/λ_k was obtained for the approximation criterion in [11, Theorem 3.3] and in [8, Theorem 5], while the coherence criterion was studied in [15, Proposition 5] with the upper bound $(1-\gamma^2)/\lambda_k$.

VI. Experimental results

To illustrate the results obtained in this paper, we use the Mackey-Glass time-delay differential equation

$$\frac{dx}{dt} = \frac{0.2x(t-\tau)}{1+x^{10}(t-\tau)} - 0.1x(t),$$

with $\tau = 17$ the initial condition $x(0) = 1.2$ and $x(t) = 0$ for $t < 0$. The fourth-order Runge-Kutta method is used to get the corresponding time series at integer time points[3], for $t = 1, 2, \ldots, 1200$. It is well-known that this time series has a chaotic behavior, as shown in Fig. 1. The two-dimensional samples $\boldsymbol{x}_t = [x_{t-10} \;\; x_t]^\top$ are used as input data.

First, we illustrate the relevance of the results given in Section IV for the Gaussian kernel with the bandwidth set to $\sigma = 0.35$ and the coherence criterion. Fig. 2 (left) illustrates the error of discarded samples and the upper bound given in Theorem 2, namely $\kappa(\boldsymbol{x}_t, \boldsymbol{x}_t)(1-\gamma^2)$, for several values of the coherence threshold $\gamma = 0.05, 0.1, 0.15, \ldots, 0.95$. An enlargement at $\gamma = 0.65$ shows the tightness of this bound. By setting the threshold $\gamma = 0.35$, Fig. 2 (right) shows the lower bounds of Theorem 4 on the error of approximating any atom with the other atoms of the dictionary, namely $1 - \frac{(m-1)\gamma^2}{1-(m-2)\gamma}$, where m is the increasing dictionary size. It is worth noting that these results can also be obtained with the distance criterion with the threshold set to $\delta = \sqrt{1-0.35^2} \approx 0.937$. Fig. 3 shows the bounds given in Section V-B for the quadratic error of approximating the two most relevant principal axes, where these principal axes are estimated at each instant t on all samples available up to t. This figure also compares these results with the ones obtained from the approximation criterion, where the threshold is set to $\delta = 0.9$, yielding the same dictionary size and almost the same bounds as the coherence criterion.

[3]The data are available from the Fuzzy Logic Toolbox of Matlab with `mgdata.dat`

VII. Final remarks

In this paper, we studied the approximation errors of any sample when dealing with the distance or the coherence criterion, revealing that these criteria are roughly based on an approximation process, in the same sense as the approximation criterion. To this end, we first established an upper bound on the error of approximating a sample discarded from the dictionary. As a consequence, we explored that the atoms are "sufficient" to represent any sample. Then, we considered the dual condition, namely that each atom of the dictionary is "necessary", by providing a lower bound on the error of approximating any atom of the dictionary with the other atoms. Moreover, beyond the analysis of a single sample, we extended these results to the estimation of any pattern, by describing in detail two classes of patterns, the centroid and the principal axes in kernel-PCA, and including the approximation criterion in our analysis.

This work did not devise any particular sparsification criterion. It provided a unified framework to study online sparsification criteria. We argued that these criteria behave essentially in an identical mechanism, and share many interesting and desirable properties. Without loss of generality, we considered the framework of kernel-based learning algorithms. It is worth noting that these machines are intimately connected with the Gaussian processes [6], where the approximation criterion was initially proposed [10].

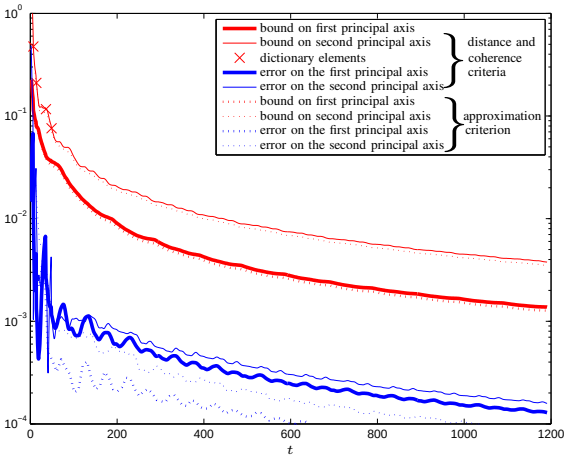

Fig. 3. Illustration of the bounds (the two upper curves with solid lines) on the quadratic errors (the two lower curves with solid lines) of approximating the first (bold lines) and second (thin lines) most relevant principal axes, as given in Theorem 7. Results obtained with the approximation criterion are given in dotted lines.

Appendix

This appendix provides bounds on the eigenvalues of a Gram matrix associated to a dictionary, for each of the sparsity measures investigated in this paper. For completeness, these bounds are put here in a nutshell; see [26, Section IV] for more details. A cornerstone of these results is the well-known Geršgorin Discs Theorem [36, Chapter 6]. Revisited here for a Gram matrix associated to a dictionary, it states that any of its eigenvalues lies in the union of the m discs, centered on each diagonal entry of $\grave{\boldsymbol{K}}$ with a radius given by the sum of the absolute values of the other $m-1$ entries from the same row. In other words, for each $\grave{\lambda}_i$, there exists at least one $j \in \{1, 2, \ldots, m\}$ such that

$$|\grave{\lambda}_i - \kappa(\grave{\boldsymbol{x}}_j, \grave{\boldsymbol{x}}_j)| \leq \sum_{\substack{j=1 \\ j \neq i}}^{m} |\kappa(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_j)|. \tag{18}$$

This theorem provides upper and lower bounds on the eigenvalues of a Gram matrix associated to a dictionary, as described in the following for each sparsity measure.

Lemma A.1: The eigenvalues of a Gram matrix associated to a δ-distant dictionary of m atoms are bounded as follows:

$$r^2 - (m-1)R\sqrt{R^2-\delta^2} \leq \grave{\lambda}_m \leq \cdots \leq \grave{\lambda}_1 \leq R^2 + (m-1)R\sqrt{R^2-\delta^2}.$$

Proof: From (8), a δ-distant dictionary satisfies

$$|\kappa(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_j)| \leq \sqrt{\kappa(\grave{\boldsymbol{x}}_j, \grave{\boldsymbol{x}}_j)\left(\kappa(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_i) - \delta^2\right)},$$

for any $i = 1, 2, \ldots, m$, which yields

$$\begin{aligned}\sum_j |\kappa(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_j)| &\leq \sum_j \sqrt{\kappa(\grave{\boldsymbol{x}}_j, \grave{\boldsymbol{x}}_j)\left(\kappa(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_i) - \delta^2\right)} \\ &= \sqrt{\kappa(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_i) - \delta^2} \sum_j \sqrt{\kappa(\grave{\boldsymbol{x}}_j, \grave{\boldsymbol{x}}_j)}.\end{aligned}$$

By substituting this relation in (18), we get that, for each eigenvalue $\grave{\lambda}_k$, there exists at least one i such that

$$|\grave{\lambda}_k - \kappa(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_i)| \leq \sqrt{\kappa(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_i) - \delta^2} \sum_{\substack{j=1 \\ j \neq i}}^{m} \sqrt{\kappa(\grave{\boldsymbol{x}}_j, \grave{\boldsymbol{x}}_j)}.$$

■

Lemma A.2: The eigenvalues of a Gram matrix associated to a γ-coherent dictionary of m atoms are bounded as follows:

$$r^2 - (m-1)\gamma R^2 \leq \grave{\lambda}_m \leq \cdots \leq \grave{\lambda}_1 \leq R^2 + (m-1)\gamma R^2.$$

Proof: A γ-coherent dictionary satisfies, for any $i, j = 1, 2, \ldots, m$,

$$\begin{aligned}\max_{\substack{j=1\cdots m \\ j \neq i}} |\kappa(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_j)| &\leq \gamma \max_{\substack{j=1\cdots m \\ j \neq i}} \sqrt{\kappa(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_i)\,\kappa(\grave{\boldsymbol{x}}_j, \grave{\boldsymbol{x}}_j)} \\ &= \gamma \sqrt{\kappa(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_i)} \max_{\substack{j=1\cdots m \\ j \neq i}} \sqrt{\kappa(\grave{\boldsymbol{x}}_j, \grave{\boldsymbol{x}}_j)} \\ &\leq \gamma R \sqrt{\kappa(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_i)}.\end{aligned}$$

By inserting this expression into (18), we get

$$\begin{aligned}\sum_{\substack{j=1 \\ j \neq i}}^{m} |\kappa(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_j)| &\leq (m-1) \max_{\substack{j=1\cdots m \\ j \neq i}} |\kappa(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_j)| \\ &\leq (m-1)\gamma R \sqrt{\kappa(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_i)}.\end{aligned}$$

Since $\kappa(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_i) \leq R^2$, this completes the proof. ■

REFERENCES

[1] "Special issue on signal processing for big data," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 8, pp. 507–507, June 2014.

[2] "Special issue on signal processing for big data," *Signal Processing, IEEE Transactions on*, vol. 62, pp. 1899–1899, April 2014.

[3] G. Giannakis, F. Bach, R. Cendrillon, M. Mahoney, and J. Neville, "Signal processing for big data [from the guest editors]," *Signal Processing Magazine, IEEE*, vol. 31, pp. 15–16, Sept 2014.

[4] J. Platt, "A resource-allocating network for function interpolation," *Neural Comput.*, vol. 3, pp. 213–225, June 1991.

[5] V. N. Vapnik, *Statistical Learning Theory*. New York, NY, USA: Wiley, September 1998.

[6] C. E. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[7] G. bin Huang, P. Saratch, S. Member, and N. Sundararajan, "A generalized growing and pruning rbf (ggap-rbf) neural network for function approximation," *IEEE Transactions on Neural Networks*, vol. 16, pp. 57–67, 2005.

[8] P. Honeine, "Online kernel principal component analysis: a reduced-order model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, pp. 1814–1826, September 2012.

[9] B. Schölkopf, R. Herbrich, and A. J. Smola, "A generalized representer theorem," in *Proc. 14th Annual Conference on Computational Learning Theory and 5th European Conference on Computational Learning Theory*, (London, UK), pp. 416–426, Springer-Verlag, 2001.

[10] L. Csató and M. Opper, "Sparse online gaussian processes," *Neural Computation*, vol. 14, pp. 641–668, 2002.

[11] Y. Engel, S. Mannor, and R. Meir, "The kernel recursive least squares algorithm," *IEEE Trans. Signal Processing*, vol. 52, no. 8, pp. 2275–2285, 2004.

[12] P. P. Pokharel, W. Liu, and J. C. Principe, "Kernel least mean square algorithm with constrained growth," *Signal Processing*, vol. 89, no. 3, pp. 257 – 265, 2009.

[13] G. S. Babu and S. Suresh, "Meta-cognitive rbf network and its projection based learning algorithm for classification problems," *Appl. Soft Comput.*, vol. 13, pp. 654–666, Jan. 2013.

[14] Y.-K. Yang, T.-Y. Sun, C.-L. Huo, Y.-H. Yu, C.-C. Liu, and C.-H. Tsai, "A novel self-constructing radial basis function neural-fuzzy system," *Applied Soft Computing*, vol. 13, no. 5, pp. 2390 – 2404, 2013.

[15] P. Honeine, C. Richard, and J. C. M. Bermudez, "On-line nonlinear sparse approximation of functions," in *Proc. IEEE International Symposium on Information Theory*, (Nice, France), pp. 956–960, June 2007.

[16] C. Richard, J. C. M. Bermudez, and P. Honeine, "Online prediction of time series data with kernels," *IEEE Transactions on Signal Processing*, vol. 57, pp. 1058–1067, March 2009.

[17] J. A. Tropp, "Greed is good: algorithmic results for sparse approximation," *IEEE Trans. Information Theory*, vol. 50, pp. 2231–2242, 2004.

[18] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer, 2010.

[19] H. Fan, Q. Song, and S. B. Shrestha, "Online learning with kernel regularized least mean square algorithms," *Knowledge-Based Systems*, vol. 59, no. 0, pp. 21 – 32, 2014.

[20] W. Liu, J. C. Principe, and S. Haykin, *Kernel Adaptive Filtering: A Comprehensive Introduction*. Wiley Publishing, 1st ed., 2010.

[21] Z. Noumir, P. Honeine, and C. Richard, "One-class machines based on the coherence criterion," in *Proc. IEEE workshop on Statistical Signal Processing*, (Ann Arbor, Michigan), pp. 600–603, 5–8 August 2012.

[22] Z. Noumir, P. Honeine, and C. Richard, "On simple one-class classification methods," in *Proc. IEEE International Symposium on Information Theory*, (MIT, Cambridge (MA), USA), pp. 2022–2026, 1–6 July 2012.

[23] R. Jenssen, "Mean vector component analysis for visualization and clustering of nonnegative data," *Neural Networks and Learning Systems, IEEE Transactions on*, vol. 24, pp. 1553–1564, Oct 2013.

[24] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, pp. 1299–1319, July 1998.

[25] C. Saidé, R. Lengellé, P. Honeine, and R. Achkar, "Online kernel adaptive algorithms with dictionary adaptation for mimo models," *IEEE Signal Processing Letters*, vol. 20, pp. 535–538, May 2013.

[26] P. Honeine, "Analyzing sparse dictionaries for online learning with kernels," *IEEE Transactions on Signal Processing*, (submitted) 2015.

[27] C. Saidé, R. Lengellé, P. Honeine, C. Richard, and R. Achkar, "Dictionary adaptation for online prediction of time series data with kernels," in *Proc. IEEE workshop on Statistical Signal Processing*, (Ann Arbor, Michigan, USA), pp. 604–607, 5–8 August 2012.

[28] C. Saidé, P. Honeine, R. Lengellé, C. Richard, and R. Achkar, "Adaptation en ligne d'un dictionnaire pour les méthodes à noyau," in *Actes du 24-ème Colloque GRETSI sur le Traitement du Signal et des Images*, (Brest, France), September 2013.

[29] G. Baudat and F. Anouar, "Kernel-based methods and function approximation," in *In International Joint Conference on Neural Networks (IJCNN)*, vol. 5, (Washington, DC, USA), pp. 1244–1249, July 2001.

[30] L. Csató and M. Opper, "Sparse representation for gaussian process models," in *Advances in Neural Information Processing Systems 13*, pp. 444–450, MIT Press, 2001.

[31] D. Nguyen-Tuong and J. Peters, "Incremental online sparsification for model learning in real-time robot control," *Neurocomputing*, vol. 74, no. 11, pp. 1859 – 1867, 2011.

[32] R. Rosipal, M. Koska, and I. Farkas, "Prediction of chaotic time-series with a resource-allocating RBF network," in *Neural Processing Letters*, pp. 185–197, 1997.

[33] B. Chen, S. Zhao, P. Zhu, and J. Principe, "Quantized kernel least mean square algorithm," *Neural Networks and Learning Systems, IEEE Transactions on*, vol. 23, pp. 22–32, Jan 2012.

[34] M. Yukawa and R. Ishii, "An efficient kernel adaptive filtering algorithm using hyperplane projection along affine subspace," in *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, pp. 2183–2187, Aug 2012.

[35] W. Gao, J. Chen, C. Richard, and J. Huang, "Online dictionary learning for kernel lms," *Signal Processing, IEEE Transactions on*, vol. 62, pp. 2765–2777, June 2014.

[36] R. A. Horn and C. R. Johnson, *Matrix analysis*. New York, NY, USA: Cambridge University Press, 2nd edition ed., December 2012.

[37] E. Ma and C. Zarowski, "On lower bounds for the smallest eigenvalue of a hermitian positive-definite matrix," *Information Theory, IEEE Transactions on*, vol. 41, pp. 539–540, Mar 1995.

[38] W. Sun, "Lower bounds of the minimal eigenvalue of a hermitian positive-definite matrix," *Information Theory, IEEE Transactions on*, vol. 46, pp. 2760–2762, Nov 2000.

[39] D. Park and B. G. Lee, "On determining upper bounds of maximal eigenvalue of hermitian positive-definite matrix," *Signal Processing Letters, IEEE*, vol. 10, pp. 267–269, Sept 2003.

[40] R. Turkmen and H. Civciv, "Some bounds for the singular values of matrices," *Applied Mathematical Sciences*, vol. 1, no. 49, pp. 2443–2449, 2007.

[41] K. Hlavackova-Schindler, "A new lower bound for the minimal singular value for real non-singular matrices by a matrix norm and determinant," *Applied Mathematical Sciences*, vol. 4, no. 64, pp. 3189–3193, 2010.

[42] L. Zou, "A lower bound for the smallest singular value," *Journal of Mathematical Inequalities*, vol. 6, no. 4, pp. 625–629, 2012.

[43] Z. Noumir, P. Honeine, and C. Richard, "Online one-class machines based on the coherence criterion," in *Proc. 20th European Conference on Signal Processing*, (Bucharest, Romania), pp. 664–668, 27–31 August 2012.

[44] P. Honeine, "An eigenanalysis of data centering in machine learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (submitted) 2015.

PLACE PHOTO HERE

Paul Honeine (M'07) was born in Beirut, Lebanon, on October 2, 1977. He received the Dipl.-Ing. degree in mechanical engineering in 2002 and the M.Sc. degree in industrial control in 2003, both from the Faculty of Engineering, the Lebanese University, Lebanon. In 2007, he received the Ph.D. degree in Systems Optimisation and Security from the University of Technology of Troyes, France, and was a Postdoctoral Research associate with the Systems Modeling and Dependability Laboratory, from 2007 to 2008. Since September 2008, he has been an assistant Professor at the University of Technology of Troyes, France. His research interests include nonstationary signal analysis and classification, nonlinear and statistical signal processing, sparse representations, machine learning. Of particular interest are applications to (wireless) sensor networks, biomedical signal processing, hyperspectral imagery and nonlinear adaptive system identification. He is the co-author (with C. Richard) of the 2009 Best Paper Award at the IEEE Workshop on Machine Learning for Signal Processing. Over the past 5 years, he has published more than 100 peer-reviewed papers.