Analyzing sparse dictionaries for online learning with kernels

Paul Honeine

Analyzing sparse dictionaries for online learning with kernels

Paul Honeine, *Member IEEE*

***Abstract*—Many signal processing and machine learning methods share essentially the same linear-in-the-parameter model, with as many parameters as available samples as in kernel-based machines. Sparse approximation is essential in many disciplines, with new challenges emerging in online learning with kernels. To this end, several sparsity measures have been proposed in the literature to quantify sparse dictionaries and constructing relevant ones, the most prolific ones being the distance, the approximation, the coherence and the Babel measures. In this paper, we analyze sparse dictionaries based on these measures. By conducting an eigenvalue analysis, we show that these sparsity measures share many properties, including the linear independence condition and inducing a well-posed optimization problem. Furthermore, we prove that there exists a quasi-isometry between the parameter (i.e., dual) space and the dictionary's induced feature space.**

***Index Terms*—Sparse approximation, adaptive filtering, kernel-based methods, Gram matrix, machine learning, pattern recognition.**

I. Introduction

SPARSE approximation is essential in many disciplines due to the advent of data deluge in the era of "Big Data", as illustrated by the extensive literature of compressed sensing (see [1] and references therein). Sparsity promoting is crucial in signal processing and machine learning, such as Gaussian processes [2], kernel-based methods [3], Bayesian learning [4], as well as neural networks [5] with pruning [6] and the more recent dropout principle in deep learning [7].

Many learning machines share essentially the same model, in a linear or a nonlinear — kernel — form, including support vector machines [8], Gaussian processes [9] and radial-basis-function networks such as resource-allocating networks [5] and more recently neural networks for function approximation [10]; see also the seminal work of Poggio and Smale [11]. All these learning machines rely on the well-known "Representer Theorem" [12], which defines a linear-in-the-parameters model with as many parameters as training samples.

A sparse approximation of this model is often required for many interesting and desirable properties, such as enforcing the interpretation of the results and providing a computational tractable problem for large-scale datasets. Within the last 15 years in kernel-based machines, this issue has been largely investigated in an offline setting, with pursuit algorithms [13], [14] and more recently with sparse coding and dictionary learning algorithms [15], [16], [17]; see also [18], [19], [20] and references therein. Online learning brings new challenges to sparsity in signal processing and machine learning, when a new sample is available at each instant, which leads to an incrementation of the number of parameters. Therefore, one needs to control such complexity growth, by selecting samples that take part in the model formulation; in the literature, these contributing samples are called atoms and are collected in a set called dictionary[1].

The construction from available samples of a pertinent dictionary and the measure of its relevance have been investigated in the literature with several sparsification criteria, each being coupled with a sparsity measure that defines the diversity captured by the dictionary. The oldest sparsity criterion is the distance introduced in [5] for controlling the complexity of the structure of radial-basis-function networks in resource-allocating networks [27]; see also [28], [29] for recent advances. The criterion constructs a dictionary by lower-bounding the pairwise distance between its atoms. Another criterion, the approximation criterion, explores a deeper analysis of the atoms, by lower-bounding the error of approximating any atom by the other atoms, as investigated in [30] for Gaussian processes, in [31] for a kernel recursive least squares algorithm, and more recently in [32] for a kernel principal component analysis. A third criterion takes advantage of recent developments in the sparse approximation literature [33] and compressed sensing [23], by upper-bounding the coherence between any pair of atoms. Initially introduced for online learning with kernels [34], [35] and learning in sensor networks [36], [37], it has been extensively considered for one-class classification [38], [39], for online learning with multiple kernels [40], [41] and multiple dictionaries [42] and for multiple-output learning [43]. The Babel measure and

P. Honeine is with the Institut Charles Delaunay (CNRS), Université de technologie de Troyes, 10000, Troyes, France. Phone: +33(0)325715625; Fax: +33(0)325715699; E-mail: paul.honeine@utt.fr

[1] In dictionary learning literature, there exists two complementary concepts of sparsity of a dictionary, as illustrated next in the linear case.

The first one is the sparsity representation in the space of the learned dictionary, which is usually realized by an ℓ_p-norm approximation for $0 \leq p \leq 1$, namely by solving

$$\min_{\substack{\dot{\boldsymbol{x}}_j, \boldsymbol{\alpha}_j \\ j=1\cdots m}} \sum_i \|\boldsymbol{z}_i - [\dot{\boldsymbol{x}}_1 \;\; \dot{\boldsymbol{x}}_2 \;\; \cdots \;\; \dot{\boldsymbol{x}}_m]\,\boldsymbol{\alpha}_i\|^2 \qquad \text{subject to } \|\boldsymbol{\alpha}_i\|_p \leq c,$$

where $\dot{\boldsymbol{x}}_i$ is the i-th dictionary atom, all $\boldsymbol{z}_i$ are given signals, and c is a fixed threshold. This is a combinatorial and highly non-convex optimization problem. Currently used algorithms determine a local minimum in an offline setting [21], [22], by alternating between two steps, the sparse-coding (*i.e.*, estimation of the sparse vectors $\boldsymbol{\alpha}_i$) and the dictionary update (*i.e.*, estimation of the atoms $\dot{\boldsymbol{x}}_j$). See also [23] and references therein.

The second concept, which is the focus of this paper, consists in utilizing a specific dictionary structure which essentially serves as a pool of atoms from which a sub-dictionary could be efficiently selected, the sub-dictionary is also called "sparse dictionary". Algorithms have been proposed in either an offline setting, essentially with basis and matching pursuit algorithms [24], [13], [14], or an online setting as in resource-allocating networks [5] and online learning with kernels [31], [35], [32], as studied in this paper. It is worth noting that, while the atoms are selected from a fixed dictionary, one can also adapt them, as investigated in [25]. See also Footnote 3.

	Distance	Approximation	Coherence	Babel	Section
Reference: most known work	[5]	[30]	[35]	[33]	III
Reference: more recent work	[48]	[32]	[43]	[44]	III
Eigenvalues: lower bounds	✓	[32]	[34]	✓	IV-A
Eigenvalues: upper bounds	✓	[32]	✓	✓	IV-A
Linear independence	✓	✓	[34]	[35]	IV-B
Condition number	✓	✓	✓	✓	IV-C
Isometry property: distances	✓	✓	✓	✓	V-A
Isometry property: inner products	✓	✓	✓	✓	V-B

TABLE I
A BIRDS EYE VIEW OF THE THEORETICAL INSIGHTS STUDIED IN THIS PAPER. SOME OF THESE RESULTS WERE PREVIOUSLY DERIVED FOR UNIT-NORM ATOMS, AS SHOWN WITH THE REFERENCES GIVEN IN THE TABLE. IN THIS WORK, WE PROVIDE AN EXTENSIVE STUDY THAT COMPLETES THE ANALYSIS TO ALL SPARSITY MEASURES. WE DERIVE NEW THEORETICAL INSIGHTS ON CONNECTING THE DUAL SPACE WITH THE DICTIONARY'S INDUCED FEATURE SPACE. ALL THE RESULTS ARE GENERALIZED TO ANY TYPE OF KERNEL, BEYOND THE UNIT-NORM CASE.

its criterion provide a more comprehensive analysis of the dictionary structure, by limiting the cumulative coherence [44]. To the best of our knowledge, there is no work that studies all these sparsity measures and criteria.

Independently of the sparsification criterion and the resulting dictionary, many algorithms have been introduced to update the model. As it might be expected, the wide class of linear adaptive filters has been extensively investigated for online learning with kernels, by revisiting popular algorithms such as the least mean squares (LMS), the normalized LMS (NLMS), the affine projection (AP), and the recursive least squares (RLS) algorithms; see for instance [45] for a review of linear adaptive filters. There exists two frameworks to develop adaptive algorithms in online learning with kernels, owing to the underlying linear-in-the-parameters model: a functional (*i.e.*, feature) framework and a dual (*i.e.*, parameter) one. Within the functional framework, the optimization is operated in the feature space, by estimating and updating within the subspace spanned by the atoms of the dictionary. This framework has been widely investigated for online learning with kernels; see for instance [46], [40] as well as [47] for a theoretical analysis and [48] for a comprehensive study. The second framework is based on estimating the parameters of the model, thus solving an optimization problem in the so-called dual space. This framework has been extensively explored in the literature due to its simplicity, with a NLMS algorithm [34], an AP algorithm [35], and a RLS algorithm [31], [49]. For an overview of this framework, see [50] and references therein. To the best of our knowledge, only Yukawa pointed out the distinction between these two frameworks in [51, Section 6.6.4]. The relationship between the two frameworks has not been studied before, namely connecting the feature space to the dual space.

The aim of this paper is to study all the aforementioned sparsity measures and sparsification criteria (cf. Section III). To this end, we provide an analysis of the eigenvalues associated to a sparse dictionary, and provide upper and lower bounds in terms of the sparsity measures (cf. Section IV-A). We show that the lower bounds provide conditions on the linear independence of the atoms (cf. Section IV-B). Moreover, we show that the condition number of the Gram matrix associated to a sparse dictionary is upper-bounded, illustrating the impact of the sparsity measures on the conditioning of the optimization problem (cf. Section IV-C). A major result provided in this paper is the connection between the dictionary's induced feature space and the dual space, by showing that there exists a quasi-isometry between these spaces when dealing with sparse dictionaries. These results allow to bridge the gaps between the two aforementioned frameworks (cf. Sections V-A and V-B). The big picture is illustrated in TABLE I.

II. KERNEL-BASED LEARNING MACHINES

A learning problem aims to find the relation $\psi(\cdot)$ between a compact subspace of a Banach space $\mathbb{X}$ of $\mathbb{R}^d$ and a compact $\mathbb{Y}$ of $\mathbb{R}$ called output space, from a set of available samples, denoted $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}$ with $(\boldsymbol{x}_k, y_k) \in \mathbb{X} \times \mathbb{Y}$.

A. Batch learning with kernels

Considering a given 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 $\psi(\cdot)$, the optimization problem consists in minimizing a regularized empirical risk as follows

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

where $\mathbb{H}$ is the space of candidate functions and ϵ controls the tradeoff between the fitness error (first term) and the regularity of the solution (second term) where $\mathcal{R}(\cdot)$ is a monotonically increasing function. Examples of loss functions are the quadratic loss $|\psi(\boldsymbol{x}_i) - y_i|^2$ and the hinge loss $(1 - \psi(\boldsymbol{x}_i) y_i)_+$ of the support vector machines.

By using the formalism of the reproducing kernel Hilbert space (RKHS) as the space $\mathbb{H}$ of candidate functions, kernel-based machines incorporate prior knowledge by using a kernel. Let $\kappa : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ be a reproducing kernel, and $(\mathbb{H}, \langle \cdot, \cdot \rangle_{\mathbb{H}})$ the induced 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 $\psi(\boldsymbol{x}_i) = \langle \psi(\cdot), \kappa(\boldsymbol{x}_i, \cdot) \rangle_{\mathbb{H}}$. This property shows that any sample $\boldsymbol{x}_i$ of $\mathbb{X}$ is represented with $\kappa(\cdot, \boldsymbol{x}_i)$ in the space $\mathbb{H}$, also called feature space. 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 $\langle \kappa(\cdot, \boldsymbol{x}_i), \kappa(\cdot, \boldsymbol{x}_j) \rangle_{\mathbb{H}} = \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Commonly used kernels are the linear kernel with $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$, the polynomial kernel $(\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle + c)^p$ and the Gaussian kernel $\exp\left(\frac{-1}{2\sigma^2} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2\right)$.

The Representer Theorem is a cornerstone of kernel-based machines [12]. 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 shows that the functional optimization problem (1) is equivalent to the estimation of n unknowns, $\alpha_1, \alpha_2, \ldots, \alpha_n$ in (2). By injecting the above expression into (2), we get the (often called) dual problem. This duality is illustrated next for the kernel ridge regression problem.

B. Kernel ridge regression algorithms

In the kernel ridge regression, the quadratic loss and regularization are used in the optimization problem, namely

$$\underset{\psi(\cdot)\in\mathbb{H}}{\operatorname{argmin}} \tfrac{1}{2}\sum_{i=1}^{n} |\psi(\boldsymbol{x}_i) - y_i|^2 + \epsilon\, \tfrac{1}{2}\|\psi(\cdot)\|_{\mathbb{H}}^2. \tag{3}$$

By injecting the model (2) in the above expression, we get the following dual optimization problem:

$$\underset{\boldsymbol{\alpha}\in\mathbb{R}^n}{\operatorname{argmin}} \tfrac{1}{2}\|\boldsymbol{K}\boldsymbol{\alpha} - \boldsymbol{y}\|^2 + \epsilon\, \tfrac{1}{2}\boldsymbol{\alpha}^\top \boldsymbol{K}\boldsymbol{\alpha}, \tag{4}$$

where $\boldsymbol{K}$ is the Gram matrix whose (i,j)-th entry is $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $\boldsymbol{y}$ and $\boldsymbol{\alpha}$ are vectors whose i-th entries are y_i and α_i, respectively. In the above expression, we have used the relation

$$\|\psi(\cdot)\|_{\mathbb{H}}^2 = \Big\| \sum_{i=1}^{n} \alpha_i \kappa(\boldsymbol{x}_i, \cdot) \Big\|_{\mathbb{H}}^2 = \sum_{i,j=1}^{n} \alpha_i \alpha_j \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{\alpha}^\top \boldsymbol{K}\boldsymbol{\alpha}.$$

The solution of this optimization problem is given by the "normal equations" [52, Chapter 5], $(\boldsymbol{K}^\top\boldsymbol{K} + \epsilon\boldsymbol{K}^\top)\,\boldsymbol{\alpha} = \boldsymbol{K}^\top\boldsymbol{y}$, which yields[2]

$$\boldsymbol{\alpha} = \left(\boldsymbol{K}^\top\boldsymbol{K} + \epsilon\boldsymbol{K}^\top\right)^{-1}\boldsymbol{K}^\top\boldsymbol{y}. \tag{5}$$

Regularization: $\|\psi(\cdot)\|_{\mathbb{H}}$ versus $\|\boldsymbol{\alpha}\|$

The regularization in the dual optimization problem (4) is essentially a Tikhonov regularization of the form $\|\boldsymbol{\Gamma}\boldsymbol{\alpha}\|^2$ (where we have in our case $\boldsymbol{\Gamma}^\top\boldsymbol{\Gamma} = \epsilon\boldsymbol{K}$). In the literature, the Tikhonov matrix $\boldsymbol{\Gamma}$ is often chosen as the identity matrix, up to a multiplicative constant, giving preference to solutions with smaller norms. The kernel ridge regression becomes

$$\underset{\boldsymbol{\alpha}\in\mathbb{R}^n}{\operatorname{argmin}} \tfrac{1}{2}\|\boldsymbol{K}\boldsymbol{\alpha} - \boldsymbol{y}\|^2 + \epsilon\, \tfrac{1}{2}\|\boldsymbol{\alpha}\|^2. \tag{6}$$

With the "normal equations" $(\boldsymbol{K}^\top\boldsymbol{K} + \epsilon\,\mathbf{I})\,\boldsymbol{\alpha} = \boldsymbol{K}^\top\boldsymbol{y}$, we get

$$\boldsymbol{\alpha} = \left(\boldsymbol{K}^\top\boldsymbol{K} + \epsilon\,\mathbf{I}\right)^{-1}\boldsymbol{K}^\top\boldsymbol{y}.$$

Connections between the regularization in the functional space with $\|\psi(\cdot)\|_{\mathbb{H}}$ and the regularization in the dual space with $\|\boldsymbol{\alpha}\|$ are not straightforward. The only result is based on the fact that $\|\psi(\cdot)\|_{\mathbb{H}}^2 = \boldsymbol{\alpha}^\top\boldsymbol{K}\boldsymbol{\alpha}$, and therefore we have from the Rayleigh's quotient and the Courant-Fischer Minimax Theorem [52, Theorem 8.1.2]:

$$\lambda_{\min} \leq \frac{\|\psi(\cdot)\|_{\mathbb{H}}^2}{\|\boldsymbol{\alpha}\|^2} \leq \lambda_{\max}$$

where $\lambda_{\min}$ and $\lambda_{\max}$ are the smallest and largest eigenvalues of the Gram matrix $\boldsymbol{K}$. As a consequence, minimizing $\|\psi(\cdot)\|_{\mathbb{H}}^2$ yields the upper bound on the norm of the parameter vector with $\|\boldsymbol{\alpha}\|^2 \leq \lambda_{\min}^{-1}\|\psi(\cdot)\|_{\mathbb{H}}^2$, while minimizing $\|\boldsymbol{\alpha}\|^2$ yields the following upper bound on the norm in the functional space with $\|\psi(\cdot)\|_{\mathbb{H}}^2 \leq \lambda_{\max}\|\boldsymbol{\alpha}\|^2$.

It turns out that sparse dictionaries provide models with tighter bounds, as studied in detail in Section V.

C. Online learning with kernels

The Representer Theorem with its linear-in-the-parameters model (2) constitutes a bottleneck for online learning, which is required for real-time system identification, Big-Data processing and distributed optimization (*e.g.*, sensor networks). Indeed, in an online setting, the solution should be updated recursively based on a new information available at each instant, namely a novel $(\boldsymbol{x}_t, y_t)$ at instant t. Thus, by including the new pair $(\boldsymbol{x}_t, y_t)$ in the training set, the Representer Theorem dictates a new parameter α_t to be added to the set of unknowns. As a consequence, the order of the linear-in-the-parameters model is continuously increasing.

To overcome this drawback, 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 at instant t takes the form

$$\psi_t(\cdot) = \sum_{j=1}^{m} \alpha_{j,t}\, \kappa(\grave{\boldsymbol{x}}_j, \cdot), \tag{7}$$

for some order m, fixed or controlled, with $m \ll t$. Each $\grave{\boldsymbol{x}}_j$ is chosen from all available samples up to instant t, namely[3] $\{\grave{\boldsymbol{x}}_1, \grave{\boldsymbol{x}}_2, \ldots, \grave{\boldsymbol{x}}_m\} \subset \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_t\}$. We denote by dictionary the set $\mathcal{D} = \{\kappa(\grave{\boldsymbol{x}}_1, \cdot), \kappa(\grave{\boldsymbol{x}}_2, \cdot), \ldots, \kappa(\grave{\boldsymbol{x}}_m, \cdot)\}$, by atoms its elements, and by $\grave{\mathbb{H}}$ the space spanned by $\mathcal{D}$. In this paper, we do not restrictive ourselves to unit-norm[4] atoms. Let

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

The optimization problem is two-fold at each instant: selecting the proper dictionary $\mathcal{D} = \{\kappa(\grave{\boldsymbol{x}}_1, \cdot), \kappa(\grave{\boldsymbol{x}}_2, \cdot), \ldots, \kappa(\grave{\boldsymbol{x}}_m, \cdot)\}$ and estimating the corresponding parameters $\alpha_1, \alpha_2, \ldots, \alpha_m$. Before studying in detail the former in Section III, the latter is outlined next.

Notation

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-1 vector $\grave{\boldsymbol{\kappa}}(\cdot)$ whose j-th entry is $\kappa(\grave{\boldsymbol{x}}_j, \cdot)$ and the Gram matrix $\grave{\boldsymbol{K}}$ of size m-by-m whose (i,j)-th entry is $\kappa(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_j)$. The eigenvalues of this matrix are denoted $\grave{\lambda}_1, \grave{\lambda}_2, \ldots, \grave{\lambda}_m$, given in non-increasing order.

[2]The expression (5) is often simplified to $\boldsymbol{\alpha} = (\boldsymbol{K} + \epsilon\,\mathbf{I})^{-1}\boldsymbol{y}$. This equivalence is granted only when the matrix $\boldsymbol{K}$ is nonsingular, an assumption that is unfortunately not satisfied in general. This is due to the linear dependence of the training samples.

[3]We consider that each $\grave{\boldsymbol{x}}_j$ is a sample selected from available samples, that is $\grave{\boldsymbol{x}}_j$ is some $\boldsymbol{x}_{\omega_j}$ with $\omega_j \in \{1, 2, \ldots, t\}$. By using the notation $\grave{\boldsymbol{x}}_j$ in this paper, as opposed to $\boldsymbol{x}_{\omega_j}$, the elements $\grave{\boldsymbol{x}}_j$ in the expansion (7) need not be samples drawn from the distribution. This difference is investigated in [25], [26], by updating $\grave{\boldsymbol{x}}_j$ at each instant in order to minimize the prediction error.

[4]Throughout this paper, we outline the special case of unit-norm atoms since such setting is often considered in the literature. Unit-norm atoms arise when dealing either with the linear kernel when $\|\boldsymbol{x}\| = 1$ for any $\boldsymbol{x} \in \mathbb{X}$, or with unit-norm kernel, namely $\kappa(\boldsymbol{x}, \boldsymbol{x}) = 1$ for any $\boldsymbol{x} \in \mathbb{X}$.

D. Parameter estimation for online learning

Before studying in Section III the dictionary in terms of sparsity measures and sparsification criteria for constructing a relevant dictionary, we assume for now that the dictionary is known. From (7), the problem of determining the model can be solved in two ways: the functional framework where $\psi_t(\cdot)$ is updated from $\psi_{t-1}(\cdot)$, and the dual framework with the update of the parameter vector $\boldsymbol{\alpha}_t$ from $\boldsymbol{\alpha}_{t-1}$. These two frameworks are summarized next, starting with the latter since its vector-based formulation is straightforward.

We denote by $e_t = y_t - \psi_{t-1}(\boldsymbol{x}_t)$ the prediction error.

Dual framework

This framework explores the model (7) written, for any $\boldsymbol{x}$,

$$\psi_t(\boldsymbol{x}) = \boldsymbol{\alpha}_t^\top \grave{\boldsymbol{\kappa}}(\boldsymbol{x}), \tag{8}$$

where $\boldsymbol{\alpha}_t = [\alpha_{1,t} \quad \alpha_{2,t} \quad \cdots \quad \alpha_{m,t}]^\top$ is updated from the previous estimate, *i.e.*, $\boldsymbol{\alpha}_{t-1}$, in the dual space $\mathbb{R}^m$. It is easy to see in (8) the structure of a a finite-impulse-response filter, the filter input being $\grave{\boldsymbol{\kappa}}(\boldsymbol{x})$ and its coefficient vector $\boldsymbol{\alpha}_t$.

By considering the instantaneous risk $\frac{1}{2}|y_t - \boldsymbol{\alpha}^\top \grave{\boldsymbol{\kappa}}(\boldsymbol{x}_t)|^2 + \epsilon \frac{1}{2}\|\boldsymbol{\alpha}\|^2$, where the first term is the quadratic instantaneous error e_t^2, we get the stochastic gradient descent rule

$$\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1} + \eta_t \big(e_t \, \grave{\boldsymbol{\kappa}}(\boldsymbol{x}_t) - \epsilon \, \boldsymbol{\alpha}_{t-1}\big). \tag{9}$$

When dealing with the functional regularization $\|\psi(\cdot)\|_{\mathbb{H}}^2$ as in (4), this regularization is approximated with $\boldsymbol{\alpha}^\top \grave{\boldsymbol{K}} \boldsymbol{\alpha}$, which yields the modified version

$$\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1} + \eta_t \big(e_t \, \grave{\boldsymbol{\kappa}}(\boldsymbol{x}_t) - \epsilon \, \grave{\boldsymbol{K}} \boldsymbol{\alpha}_{t-1}\big). \tag{10}$$

The two rules (9) and (10) reduce to the LMS algorithm when $\epsilon = 0$. Another algorithm is the NLMS, which provides a scale insensitive version with

$$\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1} + \frac{\eta_t}{\|\grave{\boldsymbol{\kappa}}(\boldsymbol{x}_t)\|^2 + \epsilon} \, e_t \, \grave{\boldsymbol{\kappa}}(\boldsymbol{x}_t).$$

See [34] for more details. An extension to an AP algorithm is proposed in [35], while a RLS algorithm is presented in [31], [49]. A comprehensive study of adaptive filter algorithms in the dual framework is given in [50]. See also [40], [42], [44].

Functional framework

The functional framework considers the definition of the model (7) in the RKHS, with the form

$$\psi_t(\boldsymbol{x}) = \langle \psi_t(\cdot), \kappa(\boldsymbol{x}, \cdot) \rangle_{\mathbb{H}}, \tag{11}$$

for any $\boldsymbol{x} \in \mathbb{X}$. The estimation of $\psi_t(\cdot)$ from the previous estimate $\psi_{t-1}(\cdot)$ is operated in the RKHS $\mathbb{H}$, or more specifically in the span of the available dictionary, *i.e.*, $\psi_t(\cdot) \in \grave{\mathbb{H}} \subset \mathbb{H}$.

By considering the instantaneous risk $\frac{1}{2}|\psi(\boldsymbol{x}_t) - y_i|^2 + \epsilon \frac{1}{2}\|\psi(\cdot)\|_{\mathbb{H}}^2$, the stochastic gradient descent in $\mathbb{H}$ is

$$\psi_t(\cdot) = \psi_{t-1}(\cdot) + \eta_t \big(e_t \, \kappa(\boldsymbol{x}_t, \cdot) - \epsilon \, \psi_{t-1}(\cdot)\big).$$

By analogy with the dual framework, other algorithms can also be described such as an LMS, a NLMS, an AP, and a RLS algorithms. See [48] for more details.

Unfortunately, all these formulations assume the finiteness of the training set, as reported in [46] and [47]. This drawback is due to the fact that the model is fed with a new kernel function at each instant. In order to control this growth and restrict ourselves to the span of the dictionary, we replace[5] the current $\kappa(\boldsymbol{x}_t, \cdot)$ by its projection onto the subspace spanned by the dictionary, namely $\grave{\kappa}_{\boldsymbol{x}_t}(\cdot) = \grave{\boldsymbol{\kappa}}(\boldsymbol{x}_t)^\top \grave{\boldsymbol{K}}^{-1} \grave{\boldsymbol{\kappa}}(\cdot)$; see Appendix for details. This leads to the expression

$$\psi_t(\cdot) = (1 - \eta_t \, \epsilon) \, \psi_{t-1}(\cdot) + \eta_t \, e_t \, \grave{\kappa}_{\boldsymbol{x}}(\cdot).$$

To implement this formula, one needs to provide an update rule of the parameters, with an expression of the form

$$\boldsymbol{\alpha}_t = (1 - \eta_t \, \epsilon) \, \boldsymbol{\alpha}_{t-1} + \eta_t \, e_t \, \grave{\boldsymbol{\kappa}}(\boldsymbol{x}_t)^\top \grave{\boldsymbol{K}}^{-1}.$$

III. Online sparsification and sparsity measures

Independently of the investigated framework, online learning algorithms should be coupled with a sparsification scheme. At each instant, the dictionary is updated if necessary, or it is left unchanged. Indeed, the dictionary is augmented whenever the novel kernel function $\kappa(\boldsymbol{x}_t, \cdot)$ increases the diversity of the dictionary. There exists several sparsity measures to quantify this diversity, as described in the following.

Before detailing these sparsity measures, we outline the online sparsification scheme. Two cases may arise:

- **Case 1**: the dictionary is left unchanged.
  This case arises when the novel kernel function $\kappa(\boldsymbol{x}_t, \cdot)$ does not contribute significantly to the diversity of the dictionary, and therefore it could be discarded.
- **Case 2**: the kernel function is added to the dictionary.
  This case arises when the kernel function $\kappa(\boldsymbol{x}_t, \cdot)$ is significantly different from the atoms of the dictionary.

One may also use a removal process in the latter case in order to provide a fixed-budget learning [54], [55], by discarding the atom that has the least contribution to the diversity of the dictionary, as investigated for instance in [56].

A. The distance measure

A simple measure to characterize a sparse dictionary is the least distance between all pairs of its atoms. A dictionary is said to be δ-distant when

$$\min_{\substack{i,j=1\cdots m \\ i \neq j}} \min_{\xi} \|\kappa(\grave{\boldsymbol{x}}_i, \cdot) - \xi \, \kappa(\grave{\boldsymbol{x}}_j, \cdot)\|_{\mathbb{H}} \geq \delta, \tag{12}$$

where we have included a scaling factor ξ. This corresponds to the reconstruction error of projecting $\kappa(\grave{\boldsymbol{x}}_i, \cdot)$ onto $\kappa(\grave{\boldsymbol{x}}_j, \cdot)$, with $\xi = \kappa(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_j)/\kappa(\grave{\boldsymbol{x}}_j, \grave{\boldsymbol{x}}_j)$. By substituting this value in (12), we get for any pair $(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_j)$:

$$\kappa(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_i) - \frac{\kappa(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_j)^2}{\kappa(\grave{\boldsymbol{x}}_j, \grave{\boldsymbol{x}}_j)} \geq \delta^2. \tag{13}$$

[5] Besides the approximation with the projection which can be computationally expensive, one may replace the current kernel function with its most collinear atom. This leads to a quantization strategy [53].

A sparsification criterion based on this measure constructs a dictionary with a large distance measure, thus including the candidate kernel function $\kappa(\boldsymbol{x}_t, \cdot)$ in the dictionary if

$$\min_{j=1\cdots m} \left(\kappa(\boldsymbol{x}_t, \boldsymbol{x}_t) - \frac{\kappa(\boldsymbol{x}_t, \grave{\boldsymbol{x}}_j)^2}{\kappa(\grave{\boldsymbol{x}}_j, \grave{\boldsymbol{x}}_j)} \right) \geq \delta^2, \qquad (14)$$

for some threshold parameter δ. This sparsification criterion is related to the novelty criterion given in [5], which is the sparsification criterion without the scaling factor followed by a prediction error mechanism.

B. The approximation measure

The distance measure defined in (12)-(13) relies only on two atoms, that is the closest pair in the dictionary. A more comprehensive analysis of the dictionary composition is the capacity of approximating any atom by a linear combination of the other atoms. A dictionary is designated δ-approximate if the following is satisfied:

$$\min_{i=1\cdots m} \min_{\xi_1 \cdots \xi_m} \left\| \kappa(\grave{\boldsymbol{x}}_i, \cdot) - \sum_{\substack{j=1 \\ j\neq i}}^{m} \xi_j \, \kappa(\grave{\boldsymbol{x}}_j, \cdot) \right\|_{\mathbb{H}} \geq \delta. \qquad (15)$$

This corresponds to the reconstruction error of projecting any kernel function $\kappa(\grave{\boldsymbol{x}}_i, \cdot)$ onto the subspace spanned by the other kernel functions. Following the derivation given in Appendix

$$\boldsymbol{\xi} = \grave{\boldsymbol{K}}_{\backslash\{i\}}^{-1} \, \grave{\boldsymbol{\kappa}}_{\backslash\{i\}}(\grave{\boldsymbol{x}}_i), \qquad (16)$$

where $\grave{\boldsymbol{K}}_{\backslash\{i\}}$ and $\grave{\boldsymbol{\kappa}}_{\backslash\{i\}}(\grave{\boldsymbol{x}}_i)$ are obtained from $\grave{\boldsymbol{K}}$ and $\grave{\boldsymbol{\kappa}}(\grave{\boldsymbol{x}}_i)$, respectively, by removing the entries associated to $\grave{\boldsymbol{x}}_i$. As a consequence, expression (15) becomes

$$\min_{i=1\cdots m} \kappa(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_i) - \grave{\boldsymbol{\kappa}}_{\backslash\{i\}}(\grave{\boldsymbol{x}}_i)^\top \, \grave{\boldsymbol{K}}_{\backslash\{i\}}^{-1} \, \grave{\boldsymbol{\kappa}}_{\backslash\{i\}}(\grave{\boldsymbol{x}}_i) \geq \delta^2. \qquad (17)$$

The (linear) approximation criterion is based on constructing a dictionary with a high approximation measure, as investigated for Gaussian processes in [2], for a kernel-based filter in [31] and more recently for kernel principal component analysis in [32]. The kernel function $\kappa(\boldsymbol{x}_t, \cdot)$ is added to the dictionary if

$$\min_{\xi_1 \cdots \xi_m} \left\| \kappa(\boldsymbol{x}_t, \cdot) - \sum_{j=1}^{m} \xi_j \, \kappa(\grave{\boldsymbol{x}}_j, \cdot) \right\|_{\mathbb{H}}^2 \geq \delta^2, \qquad (18)$$

where δ is a positive threshold parameter that controls the level of sparseness. This leads to the following condition, written in matrix form $\kappa(\boldsymbol{x}_t, \boldsymbol{x}_t) - \grave{\boldsymbol{\kappa}}(\boldsymbol{x}_t)^\top \grave{\boldsymbol{K}}^{-1} \grave{\boldsymbol{\kappa}}(\boldsymbol{x}_t) \geq \delta^2$.

C. The coherence measure

The coherence is a fundamental measure to characterize a dictionary in the literature of sparse approximation. It corresponds to the largest correlation between atoms of a given dictionary, or mutually between atoms of two dictionaries. The coherence measure has been investigated for the analysis of the quality of representing a signal with a dictionary, initially with the work [57], [33], and more recently in the abundant publications on compressed sensing [23]. While most work consider the use of a linear measure, we explore in the following the coherence on kernel functions in order to derive the coherence criterion, as initially proposed in [34], [35].

A dictionary $\mathcal{D}$ is said γ-coherent if

$$\max_{\substack{i,j=1\cdots m \\ i\neq j}} \frac{|\kappa(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_j)|}{\sqrt{\kappa(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_i) \, \kappa(\grave{\boldsymbol{x}}_j, \grave{\boldsymbol{x}}_j)}} \leq \gamma. \qquad (19)$$

The coherence corresponds to the cosine of the angle between the kernel functions, since the above quotient can be written

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

For unit-norm kernels, (19) becomes $\max_{\substack{i,j=1\cdots m \\ i\neq j}} |\kappa(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_j)| \leq \gamma$.

The coherence criterion constructs a low-coherent dictionary [34], [35]. It includes the candidate kernel function $\kappa(\boldsymbol{x}_t, \cdot)$ in the dictionary if the coherence of the latter does not exceed a given threshold $\gamma \in \,]0\,;1]$, namely

$$\max_{j=1\cdots m} \frac{|\kappa(\boldsymbol{x}_t, \grave{\boldsymbol{x}}_j)|}{\sqrt{\kappa(\boldsymbol{x}_t, \boldsymbol{x}_t) \, \kappa(\grave{\boldsymbol{x}}_j, \grave{\boldsymbol{x}}_j)}} \leq \gamma. \qquad (20)$$

This condition enforces an upper bound on the cosine of the angle between each pair of kernel functions. The threshold γ controls the level of sparseness of the dictionary, where a null value yields an orthogonal basis. This criterion is computationally efficient as given in expression (20), where the denominator reduces to 1 for unit-norm atoms, thus becomes in this case $\max_{j=1\cdots m} |\kappa(\boldsymbol{x}_t, \grave{\boldsymbol{x}}_j)| \leq \gamma$.

D. The Babel measure

From a norm perspective, the coherence is essentially the ∞-norm when dealing with unit-norm atoms. The Babel notion explores such analogy with the norm operator, thus providing a more complete description of the dictionary structure [58], [33]. The Babel is related to the 1-norm of the Gram matrix, with the definition

$$\text{Babel} = \max_{i=1\cdots m} \sum_{\substack{j=1 \\ j\neq i}}^{m} |\kappa(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_j)|. \qquad (21)$$

It corresponds to the maximum cumulative correlation between an atom and all the other atoms of the dictionary. It is easy to see that, when dealing with unit-norm atoms, the coherence of the dictionary cannot exceed its Babel measure.

The Babel criterion is defined as follows. A candidate kernel function $\kappa(\boldsymbol{x}_t, \cdot)$ is included in the dictionary if $\max_{i=1\cdots m} \sum_{j=1, j\neq i}^{m} |\kappa(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_j)| + |\kappa(\grave{\boldsymbol{x}}_i, \boldsymbol{x}_t)| \leq \gamma$, for a given positive threshold γ. This definition can be viewed as an extension of the coherence criterion in the same sense as the approximation is an extension of the distance criterion. See [44] for the use of the Babel measure for sparsification.

IV. An eigenvalue analysis

Since the Gram matrix is fundamental in the analysis of the dictionary, we study in the following its eigenvalues, and provide theoretical bounds. These results provide an analysis of the span defined by a sparse dictionary, given in terms of the

sparsity measure under scrutiny. Lower bounds are used in the forthcoming linear independence analysis (cf. Section IV-B), while lower and upper bounds are investigated in the forthcoming study of the condition number (cf. Section IV-C) and in the main results derived in next section (cf. Section V). Let $\grave{\lambda}_1, \grave{\lambda}_2, \ldots, \grave{\lambda}_m$ be the eigenvalues of the matrix $\grave{\boldsymbol{K}}$, given in non-increasing order, namely $\grave{\lambda}_1 \geq \grave{\lambda}_2 \geq \ldots \geq \grave{\lambda}_m$.

A. Bounds on the eigenvalues

Before proceeding, we bring to mind the well-known Geršgorin Discs Theorem [59, Chapter 6], revisited here for the Gram matrix of a sparse dictionary. It is also well known that the trace of a matrix equals the sum of its eigenvalues. We get for unit-norm atoms: $\sum_{j=1}^m \grave{\lambda}_j = \mathrm{Trace}(\grave{\boldsymbol{K}}) = \sum_{j=1}^m \kappa(\grave{\boldsymbol{x}}_j, \grave{\boldsymbol{x}}_j) = m$, thus $1 \leq \grave{\lambda}_1$ and $\grave{\lambda}_m \leq 1$.

Theorem 1 (Geršgorin Discs Theorem): Every eigenvalue of an m-by-m matrix $\grave{\boldsymbol{K}}$ 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

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

This theorem is a cornerstone in our study, as described next by providing upper and lower bounds on the eigenvalues of the Gram matrix associated to a sparse dictionary, by investigating its sparsity measure.

Distance measure

When the distance measure of a given sparse dictionary is known, namely δ, we have from (12)-(13) that any pair $(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_j)$ satisfies

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

Therefore, we have

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

By applying the Geršgorin Discs Theorem (Theorem 1) with the above relation in mind, we get that, for each eigenvalue $\grave{\lambda}_k$, there exists at least one i such that

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

By exploring these results, the proof of the following theorem is straightforward.

Theorem 2: The eigenvalues of the Gram matrix associated to a δ-distant dictionary are bounded as follows:

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

where $r^2 = \inf_{\boldsymbol{x}} \kappa(\boldsymbol{x}, \boldsymbol{x})$ and $R^2 = \sup_{\boldsymbol{x}} \kappa(\boldsymbol{x}, \boldsymbol{x})$. For unit-norm atoms, we get

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

Approximation measure

The following theorem follows from the previous theorem.

Theorem 3: The eigenvalues of the Gram matrix associated to a δ-approximate dictionary are bounded as follows:

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

where $r^2 = \inf_{\boldsymbol{x}} \kappa(\boldsymbol{x}, \boldsymbol{x})$ and $R^2 = \sup_{\boldsymbol{x}} \kappa(\boldsymbol{x}, \boldsymbol{x})$. For unit-norm atoms, we get

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

Proof: To prove this theorem, we show that a δ-approximate dictionary is also δ-distant. Indeed, a δ-approximate dictionary (*i.e.*, satisfying definition (15)) verifies

$$\begin{aligned} \delta^2 &\leq \min_{i=1\cdots m} \min_{\xi_1 \cdots \xi_m} \left\| \kappa(\grave{\boldsymbol{x}}_i, \cdot) - \sum_{\substack{j=1 \\ j \neq i}}^{m} \xi_j \, \kappa(\grave{\boldsymbol{x}}_j, \cdot) \right\|_{\mathbb{H}}^2 \\ &\leq \min_{\substack{i,j=1\cdots m \\ i \neq j}} \min_{\xi_j} \|\kappa(\grave{\boldsymbol{x}}_i, \cdot) - \xi_j \, \kappa(\grave{\boldsymbol{x}}_j, \cdot)\|_{\mathbb{H}}, \end{aligned}$$

where the special case for the $\xi_1, \xi_2, \ldots \xi_m$ is considered, with all entries null except a single one to be optimized, the one denoted ξ_j. This corresponds to a δ-distant dictionary, as given in (12). ∎

Due to the rough approximation of the approximate measure with a distance measure, the bounds given in Theorem 3 are not tight. Still, this theorem allows to extend any result, from the δ-distant to the δ-approximate dictionaries.

The following theorem illustrates that the largest eigenvalue

Theorem 4: The largest eigenvalue of the Gram matrix associated to a δ-approximate dictionary is lower-bounded as follows:

$$\max_{i=1\cdots m} \frac{\grave{\boldsymbol{\kappa}}_{\setminus\{i\}}(\grave{\boldsymbol{x}}_i)^\top \grave{\boldsymbol{\kappa}}_{\setminus\{i\}}(\grave{\boldsymbol{x}}_i)}{\kappa(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_i) - \delta^2} \leq \grave{\lambda}_1$$

Proof: On the first hand, by applying the Courant-Fischer Minimax Theorem [52, Theorem 8.1.2] to the inverse of the matrix $\grave{\boldsymbol{K}}_{\setminus\{i\}}$, we get

$$\frac{1}{\grave{\lambda}_1} \leq \grave{\boldsymbol{\kappa}}_{\setminus\{i\}}(\grave{\boldsymbol{x}}_i)^\top \, \grave{\boldsymbol{K}}_{\setminus\{i\}}^{-1} \, \grave{\boldsymbol{\kappa}}_{\setminus\{i\}}(\grave{\boldsymbol{x}}_i) \leq \frac{1}{\grave{\lambda}_{m-1}},$$

where we have used the fact that, for any eigenvalue $\grave{\lambda}_j$ of a given matrix, the inverse of the matrix has an eigenvalue $1/\grave{\lambda}_j$. On the second hand, since a δ-approximate dictionary satisfies (17), we have for any $i = 1, 2, \ldots, m$:

$$\grave{\boldsymbol{\kappa}}_{\setminus\{i\}}(\grave{\boldsymbol{x}}_i)^\top \, \grave{\boldsymbol{K}}_{\setminus\{i\}}^{-1} \, \grave{\boldsymbol{\kappa}}_{\setminus\{i\}}(\grave{\boldsymbol{x}}_i) \leq \kappa(\grave{\boldsymbol{x}}_i, \grave{\boldsymbol{x}}_i) - \delta^2.$$

The proof of the theorem is obtained by combining these two results. ∎

Coherence measure

When measuring the sparsity of the dictionary with the coherence measure, we have the following theorem. Only the lower bound has been previously investigated in the literature when dealing with unit-norm atoms; see [34].

Theorem 5: The eigenvalues of the Gram matrix associated to a γ-coherent dictionary of m atoms are bounded as follows:

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

where $R^2 = \sup_{\boldsymbol{x}} \kappa(\boldsymbol{x},\boldsymbol{x})$ and $r^2 = \inf_{\boldsymbol{x}} \kappa(\boldsymbol{x},\boldsymbol{x})$. For unit-norm atoms, we get

$$1 - (m-1)\,\gamma \leq \grave{\lambda}_m \leq \cdots \leq \grave{\lambda}_1 \leq 1 + (m-1)\,\gamma.$$

Proof: A γ-coherent dictionary satisfies

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

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

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

Finally, the proof results from applying the Geršgorin Discs Theorem (Theorem 1), since

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

■

Babel measure

When dealing with the Babel measure as a sparsity measure, the eigenvalues of the Gram matrix associated to the dictionary are bounded as given in the following theorem.

Theorem 6: The eigenvalues of the Gram matrix associated to a γ-Babel dictionary are bounded as follows:

$$r^2 - \gamma \leq \grave{\lambda}_m \leq \cdots \leq \grave{\lambda}_1 \leq R^2 + \gamma,$$

where $R^2 = \sup_{\boldsymbol{x}} \kappa(\boldsymbol{x},\boldsymbol{x})$ and $r^2 = \inf_{\boldsymbol{x}} \kappa(\boldsymbol{x},\boldsymbol{x})$. For unit-norm atoms, we get $1 - \gamma \leq \grave{\lambda}_m \leq \cdots \leq \grave{\lambda}_1 \leq 1 + \gamma$.

Proof: The proof follows from the Geršgorin Discs Theorem (Theorem 1) since, for any eigenvalue $\grave{\lambda}_k$, there exists an $i \in \{1, 2, \ldots, m\}$ with

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

■

B. Linear independence

It is relevant to construct a dictionary with linearly independent atoms, a condition that allows to represent any feature of $\mathcal{D}_{\mathbb{H}}$ in a unique linear way. For a dictionary of m kernel functions, the atoms are linearly independent if the following is satisfied: any linear combination $\sum_{j=1}^{m} \xi_j\, \kappa(\grave{\boldsymbol{x}}_j, \cdot)$ is the zero element if and only if all the weighting coefficients ξ_j are null.

It is trivial that a dictionary with a nonzero approximation measure has linear independent atoms, since we have

$$\begin{aligned}\Big\|\sum_{j=1}^{m} \xi_j\, \kappa(\grave{\boldsymbol{x}}_j, \cdot)\Big\|_{\mathbb{H}} &= \Big\|\xi_i \kappa(\grave{\boldsymbol{x}}_i, \cdot) - \sum_{\substack{j=1 \\ j\neq i}}^{m} \xi_j\, \kappa(\grave{\boldsymbol{x}}_j, \cdot)\Big\|_{\mathbb{H}} \text{ (for any } i) \\ &= |\xi_i| \,\Big\|\kappa(\grave{\boldsymbol{x}}_i, \cdot) - \sum_{\substack{j=1 \\ j\neq i}}^{m} \frac{\xi_j}{\xi_i}\, \kappa(\grave{\boldsymbol{x}}_j, \cdot)\Big\|_{\mathbb{H}} \\ &\geq |\xi_i| \min_{\xi_1 \cdots \xi_m} \Big\|\kappa(\grave{\boldsymbol{x}}_i, \cdot) - \sum_{\substack{j=1 \\ j\neq i}}^{m} \xi_j\, \kappa(\grave{\boldsymbol{x}}_j, \cdot)\Big\|_{\mathbb{H}} \\ &\geq |\xi_i|\, \delta,\end{aligned}$$

for any decomposition, *i.e.*, $i \in \{1, 2, \ldots, m\}$. Thus, the linear combination is the zero element only when all coefficients ξ_i are null or when the threshold δ is null.

In the following, we show that all the sparsity measures provide sufficient conditions for linear independence of the dictionary's atoms. To this end, we investigate the duality between linear independence and the non singularity of the associated Gram matrix, which is essentially considered in [57] for the coherence of a linear dictionary with unit-norm atoms and extended in [35] for kernel-based dictionaries. Indeed, we have

$$\Big\|\sum_{j=1}^{m} \xi_j\, \kappa(\grave{\boldsymbol{x}}_j, \cdot)\Big\|_{\mathbb{H}}^2 = \boldsymbol{\xi}^\top \grave{\boldsymbol{K}} \boldsymbol{\xi} \geq \grave{\lambda}_m \|\boldsymbol{\xi}\|^2,$$

where the Courant-Fischer Minimax Theorem is used [52, Theorem 8.1.2]. As a consequence, we prove the linear independence of a atoms by providing a lower bound on the eigenvalues of the associated Gram matrix. The following theorem summarizes this property for different sparsity measures.

Theorem 7 (Linear independence): A sufficient condition for the linear independence of the m atoms is:

- $(m-1)R\sqrt{R^2 - \delta^2} < r^2$ for a δ-distant dictionary.
- $\delta > 0$ for a δ-approximate dictionary.
- $(m-1)\gamma R^2 < r^2$ for a γ-coherent dictionary.
- $\gamma < r^2$ for a γ-Babel dictionary.

These results generalize the bounds given for only unit-norm atoms, in [34] for the coherence measure with $(m-1)\gamma < 1$ and in [35] for the Babel measure with $\gamma < 1$.

C. Condition number

The condition number of a matrix $\grave{\boldsymbol{K}}$, for a given matrix norm, is defined by $\mathrm{cond}(\grave{\boldsymbol{K}}) = \|\grave{\boldsymbol{K}}\| \|\grave{\boldsymbol{K}}^{-1}\|$, which reduces for the ℓ_2-norm to:

$$\mathrm{cond}(\grave{\boldsymbol{K}}) = \frac{|\grave{\lambda}_1|}{|\grave{\lambda}_m|}. \tag{22}$$

It is an important measure of the sensitivity, with respect to variations within the matrix $\grave{\boldsymbol{K}}$, of the resolution of a problem of the form $\grave{\boldsymbol{K}}\boldsymbol{\alpha} = \boldsymbol{y}$, $\boldsymbol{\alpha}$ being the unknown. It gives a bound on how inaccurate the solution $\boldsymbol{\alpha}$ will be after approximation. When its value is small, *i.e.*, close to 1, the solution is robust to perturbations, as opposed to large values that lead to ill-conditioned problems, if not even ill-posed.

For instance, consider a gradient descent procedure to solve the linear system $\grave{\boldsymbol{K}}\boldsymbol{\alpha} = \boldsymbol{y}$. It is shown in [60] that the error reduction at each iteration is bounded by an upper bound that is proportional to the condition number of the matrix $\grave{\boldsymbol{K}}$. The condition number has been studied more recently in kernel-based machine learning; see for instance [61]. Next, we provide an upper bound on the condition number, in terms of the sparsity measure of the dictionary. The proof of the following theorem is straightforward from the definition of the condition number (22) and the aforementioned theorems on lower and upper bounds on the eigenvalues.

Theorem 8 (Condition number): The condition number of the Gram matrix associated to a sparse dictionary is upper-bounded by:

- $\dfrac{R^2 + (m-1)R\sqrt{R^2-\delta^2}}{r^2 - (m-1)R\sqrt{R^2-\delta^2}}$ for a δ-distant dictionary and a δ-approximate dictionary.
- $\dfrac{R^2 + (m-1)\gamma R^2}{r^2 - (m-1)\gamma R^2}$ for a γ-coherent dictionary.
- $\dfrac{R^2 + \gamma}{r^2 - \gamma}$ for a γ-Babel dictionary.

The case of unit-norm atoms is obtained from the relation $r = R = 1$, which yields for instance the upper bound $\frac{1+(m-1)\gamma}{1-(m-1)\gamma}$ for a γ-coherent dictionary. These results demonstrate how the choice of the threshold value in the sparsification criterion impacts on the conditioning of the system, towards a well-posed optimization problem.

V. Connecting the dictionary's induced feature space and the dual space

In this section, we show that both feature subspace and the dual space are intimately related in their topologies, when the feature subspace is spanned by the atoms from a sparse dictionary. To this end, we show in Section V-A that the pairwise distances in both spaces are almost preserved. This quasi-isometry property associated to a given sparse dictionary is quantified in terms of each of the sparsity measures presented in Section III, namely the distance, approximation, coherence, and Babel measures. These results on the isometry are extended in Section V-B to the issue of preserving the pairwise inner-products in both spaces. All these results establish the structural-preserving map that connects both spaces, namely the map $\Theta_{\mathcal{D}}$ defined as follows

$$\begin{array}{rccc} \Theta_{\mathcal{D}}\colon \mathbb{R}^m & \longmapsto & \grave{\mathbb{H}} \subset \mathbb{H} \\ \boldsymbol{\alpha} & \longrightarrow & \psi(\cdot) = \boldsymbol{\alpha}^\top \grave{\boldsymbol{\kappa}}(\cdot) \end{array}$$

It is worth noting that these results require that the atoms of the dictionary are linear independent, since this condition guarantees that any feature $\psi(\cdot)$ of $\grave{\mathbb{H}}$ can be uniquely represented by atoms of the dictionary. See Section IV-B and in particular Theorem 7 which provides weak conditions in terms of the sparsity measure of the dictionary.

A. Isometry property

Without limiting ourselves to online learning by comparing $\psi_t(\cdot)$ with $\psi_{t-1}(\cdot)$, we consider here any two features from the feature space $\grave{\mathbb{H}}$, denoted $\psi'(\cdot) = \sum_{j=1}^m \alpha'_j\, \kappa(\grave{\boldsymbol{x}}_j, \cdot)$ and $\psi''(\cdot) = \sum_{j=1}^m \alpha''_j\, \kappa(\grave{\boldsymbol{x}}_j, \cdot)$. Their representations in the dual space $\mathbb{R}^m$ are denoted $\boldsymbol{\alpha}'$ and $\boldsymbol{\alpha}''$, respectively. There exists an isometry between these two spaces if the distance between any pair of features corresponds to the distance between their parameter vectors, namely $\|\psi'(\cdot) - \psi''(\cdot)\|_{\mathbb{H}} = \|\boldsymbol{\alpha}' - \boldsymbol{\alpha}''\|$. While the isometry property is too restrictive, we relax it with the following definition of quasi-isometry, by showing that the quotient of these two distances is close to unity. We denote $\psi(\cdot) = \psi'(\cdot) - \psi''(\cdot)$, then its parameter vector is $\boldsymbol{\alpha} = \boldsymbol{\alpha}' - \boldsymbol{\alpha}''$.

Definition 9 (Quasi-isometry): Given a dictionary of kernel functions $\{\kappa(\grave{\boldsymbol{x}}_1, \cdot), \kappa(\grave{\boldsymbol{x}}_2, \cdot), \ldots, \kappa(\grave{\boldsymbol{x}}_m, \cdot)\}$, and $\grave{\mathbb{H}}$ the space spanned by its atoms, we say that the spaces $\mathbb{R}^m$ and $\grave{\mathbb{H}}$ are quasi-isometric if there exists an isometry constant ν (the smallest number) such that, for any vector $\boldsymbol{\alpha}$ of entries α_j, the feature $\psi(\cdot) = \boldsymbol{\alpha}^\top \grave{\boldsymbol{\kappa}}(\cdot)$ satisfies

$$1 - \nu \leq \frac{\|\psi(\cdot)\|^2_{\mathbb{H}}}{\|\boldsymbol{\alpha}\|_2^2} \leq 1 + \nu. \tag{23}$$

This means that the map $\Theta_{\mathcal{D}}\colon \boldsymbol{\alpha} \to \boldsymbol{\alpha}^\top \grave{\boldsymbol{\kappa}}(\cdot)$ approximately preserves the distances in both spaces $\mathbb{R}^m$ and $\grave{\mathbb{H}}$. It is easy to see that a dictionary with an isometry constant $\nu = 0$ provides a "total" isometry between these spaces.

In the following, we show that the quasi-isometry property is satisfied for sparse dictionaries, by relying on the investigated sparsity measure. Before generalizing with Theorem 11, we restrict ourselves in Theorem 10 to the case of unit-norm atoms, which is often sufficient in most work in the literature of sparse approximation, *e.g.*, when using the Gaussian kernel.

Theorem 10 (Isometry property –unit-norm atoms–): A dictionary of unit-norm atoms has an isometry constant ν defined as follows:

- $\nu = (m-1)\sqrt{1-\delta^2}$ for a δ-distant dictionary and a δ-approximate dictionary.
- $\nu = (m-1)\gamma$ for a γ-coherent dictionary.
- $\nu = \gamma$ for a γ-Babel dictionary.

Proof: For any $\psi(\cdot)$ with its parameter vector $\boldsymbol{\alpha}$ we have $\|\psi(\cdot)\|^2_{\mathbb{H}} = \|\sum_{j=1}^m \alpha_j\, \kappa(\grave{\boldsymbol{x}}_j, \cdot)\|^2_{\mathbb{H}} = \boldsymbol{\alpha}^\top \grave{\boldsymbol{K}} \boldsymbol{\alpha}$, then the quotient in (23) is the Rayleigh-Ritz quotient of the Gram matrix $\grave{\boldsymbol{K}}$. By applying the Courant-Fischer Minimax Theorem, we get

$$\grave{\lambda}_m \leq \frac{\|\psi(\cdot)\|^2_{\mathbb{H}}}{\|\boldsymbol{\alpha}\|_2^2} \leq \grave{\lambda}_1,$$

where $\grave{\lambda}_m$ and $\grave{\lambda}_1$ and the smallest and largest eigenvalues of the matrix $\grave{\boldsymbol{K}}$. We can easily identify from (23) the following pair of inequalities:

$$1 - \nu \leq \grave{\lambda}_m \qquad \text{and} \qquad \grave{\lambda}_1 \leq 1 + \nu.$$

By exploring the results derived in Section IV, we can identify the isometry constants of the dictionary in terms of its distance, approximation, coherence and Babel measures. All

these expressions are straightforward from Theorems 2, 3, 5, 6, owing to the bounds on the eigenvalues that are symmetric about 1. ∎

When dealing with non-unit-norm atoms, expressions are a bit more difficult to derive, due to the asymmetry of the bounds on the eigenvalues, as shown by the following theorem.

Theorem 11 (Isometry property): A dictionary has an isometry constant ν defined as follows:

- $\nu = \dfrac{R^2 - r^2 + 2(k-1)R\sqrt{R^2-\delta^2}}{R^2+r^2}$ for a δ-distant dictionary and a δ-approximate dictionary.
- $\nu = \dfrac{R^2 - r^2 + 2(k-1)\gamma R^2}{R^2+r^2}$ for a γ-coherent dictionary.
- $\nu = \dfrac{R^2 - r^2 + 2\gamma}{R^2+r^2}$ for a γ-Babel dictionary.

In these expressions, $R^2 = \sup_{\boldsymbol{x}} \kappa(\boldsymbol{x},\boldsymbol{x})$ and $r^2 = \inf_{\boldsymbol{x}} \kappa(\boldsymbol{x},\boldsymbol{x})$.

Proof: Consider the general asymmetric bounds

$$l_k \leq \grave{\lambda}_m \leq \frac{\|\psi(\cdot)\|^2_{\mathbb{H}}}{\|\boldsymbol{\alpha}\|_2^2} \leq \grave{\lambda}_1 \leq u_k,$$

for some lower bound l_k and upper bound u_k, such that $0 < l_k \leq u_k < \infty$. In order to get bounds that are symmetric about 1, as in Definition 9, we divide each term by $(u_k + l_k)/2$. This yields the isometry constant $\nu = (u_k - l_k)/(u_k + l_k)$ for the rescaled atoms of the dictionary, where each atom is divided by $\sqrt{(u_k+l_k)/2}$. Finally, the proof of the theorem follows the same steps as in the proof of Theorem 10. ∎

It is easy to see that Theorem 10 is a special case of this theorem when dealing with unit-norm atoms, *i.e.*, $R = r = 1$.

B. Preserving inner products

Theorems 10 and 11 show that a sparse dictionary provides a quasi-isometry, with respect to the distances, between the dual space and the subspace spanned by its atoms. In the following, we show that this property of quasi-isometry extends to inner products. It is worth noting that, when dealing with a "total" isometry, the isometry with respect to inner products extends naturally to the isometry with respect to distances, and vice versa[6]. This is not the case when using the quasi-isometry definition. We aim to bridge this gap in the following.

Definition 12 (Quasi-isometry w.r.t. inner products): Given a dictionary of kernel functions $\{\kappa(\grave{\boldsymbol{x}}_1,\cdot), \kappa(\grave{\boldsymbol{x}}_2,\cdot), \ldots, \kappa(\grave{\boldsymbol{x}}_m,\cdot)\}$, and $\grave{\mathbb{H}}$ the space spanned by its atoms, we say that the spaces $\mathbb{R}^m$ and $\grave{\mathbb{H}}$ are quasi-isometric with respect to inner products if there exists an isometry constant ν (the smallest number) such that, for any pair of vectors $(\boldsymbol{\alpha}', \boldsymbol{\alpha}'')$, we have

$$\frac{\left| \left\langle \sum_{j=1}^m \alpha'_j \, \kappa(\grave{\boldsymbol{x}}_j,\cdot), \sum_{j=1}^m \alpha''_j \, \kappa(\grave{\boldsymbol{x}}_j,\cdot) \right\rangle_{\mathbb{H}} - \boldsymbol{\alpha}'^{\top} \boldsymbol{\alpha}'' \right|}{\|\boldsymbol{\alpha}'\|_2 \, \|\boldsymbol{\alpha}''\|_2} \leq \nu. \tag{24}$$

[6]For any linear operator $\boldsymbol{A}$ from an inner product space to another inner product space, there exists an equivalence between $\langle \boldsymbol{A}\boldsymbol{u}, \boldsymbol{A}\boldsymbol{v}\rangle = \langle \boldsymbol{u}, \boldsymbol{v}\rangle$ for any $(\boldsymbol{u},\boldsymbol{v})$ and $\|\boldsymbol{A}\boldsymbol{u}\| = \|\boldsymbol{u}\|$ for any $\boldsymbol{u}$.This equivalence is less obvious when dealing with quasi-isometry.

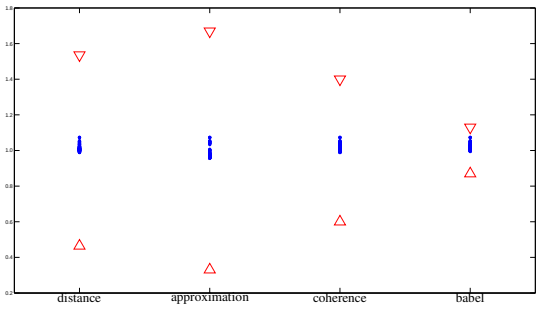

Fig. 1. Illustration of Theorem 10 for the four studied sparsification criteria. The corresponding upper and lower bounds are illustrated with "▽" and "△", respectively, and the isometry ratio in (23) is evaluated on a set of 200 samples, illustrated with "·".

It is easy to see that the "total" isometry with respect to inner products corresponds to $\nu = 0$ in (24). This expression becomes $\left\langle \sum_{j=1}^m \alpha'_j \, \kappa(\grave{\boldsymbol{x}}_j,\cdot), \sum_{j=1}^m \alpha''_j \, \kappa(\grave{\boldsymbol{x}}_j,\cdot) \right\rangle_{\mathbb{H}} = \boldsymbol{\alpha}'^{\top}\boldsymbol{\alpha}''$, and as a consequence the condition (23) is satisfied as a special case where $\boldsymbol{\alpha}' = \boldsymbol{\alpha}''$.

In the general case, the quotient in (24) can be written as

$$\frac{\left|\boldsymbol{\alpha}'^{\top}\grave{\boldsymbol{K}}\boldsymbol{\alpha}'' - \boldsymbol{\alpha}'^{\top}\boldsymbol{\alpha}''\right|}{\|\boldsymbol{\alpha}'\|_2 \, \|\boldsymbol{\alpha}''\|_2} = \frac{\left|\boldsymbol{\alpha}'^{\top}(\grave{\boldsymbol{K}} - \mathbf{I})\boldsymbol{\alpha}''\right|}{\|\boldsymbol{\alpha}'\|_2 \, \|\boldsymbol{\alpha}''\|_2},$$

and therefore the inequality (24) becomes

$$-\nu \leq \frac{\boldsymbol{\alpha}'^{\top}(\grave{\boldsymbol{K}} - \mathbf{I})\,\boldsymbol{\alpha}''}{\|\boldsymbol{\alpha}'\|_2 \, \|\boldsymbol{\alpha}''\|_2} \leq \nu. \tag{25}$$

To tackle this expression, several issues need to be addressed. First of all, the above quotient needs to be connected to the Rayleigh-Ritz quotient of the matrix $\grave{\boldsymbol{K}} - \mathbf{I}$, in order to apply the Courant-Fischer Minimax Theorem. Indeed, this theorem can be also applied to study a quotient of the form

$$\frac{\boldsymbol{u}^{\top}\boldsymbol{A}\,\boldsymbol{v}}{\|\boldsymbol{u}\|_2 \, \|\boldsymbol{v}\|_2},$$

for any pair $(\boldsymbol{u},\boldsymbol{v})$, as shown in [52, Theorem 8.6.1]; see also [62, Theorem 3] for a detailed proof. As a consequence, the quotient in (25) is bounded by the extreme eigenvalues of the matrix $\grave{\boldsymbol{K}} - \mathbf{I}$. Second, it is easy to see that both matrices $\grave{\boldsymbol{K}}$ and $\grave{\boldsymbol{K}} - \mathbf{I}$ share the same eigenvectors, while for any eigenvalue $\grave{\lambda}_j$ of $\grave{\boldsymbol{K}}$ corresponds the eigenvalue $\grave{\lambda}_j - 1$ of $\grave{\boldsymbol{K}} - \mathbf{I}$. Indeed, any eigenpair $(\grave{\boldsymbol{v}}, \grave{\lambda}_j)$ of $\grave{\boldsymbol{K}}$ satisfies $\big(\grave{\boldsymbol{K}} - \mathbf{I}\big)\grave{\boldsymbol{v}} = \grave{\boldsymbol{K}}\grave{\boldsymbol{v}} - \mathbf{I}\grave{\boldsymbol{v}} = \grave{\lambda}_j\grave{\boldsymbol{v}} - \mathbf{I}\grave{\boldsymbol{v}} = \big(\grave{\lambda}_j - 1\big)\grave{\boldsymbol{v}}$, therefore $(\grave{\boldsymbol{v}}, \grave{\lambda}_j - 1)$ is an eigenpair of the matrix $\grave{\boldsymbol{K}} - \mathbf{I}$.

As a consequence, one can take advantage of bounds on the eigenvalues from Theorems 2, 3, 5 and 6 to provide expressions for the isometry constant w.r.t. inner products, as detailed in Theorems 10 and 11.

VI. Experimental results

In order to illustrate the relevance of these results, we consider the Henon map given in 2D with $\boldsymbol{x}_t = [x_t \;\; x_{t-1}]^{\top}$ where $x_t = 1 - a_1 x_{t-1}^2 + a_2 x_{t-2}$. In the following, we set $a_1 = 1.4$ and $a_2 = 0.3$, and use the initialization $x_0 = -0.3$

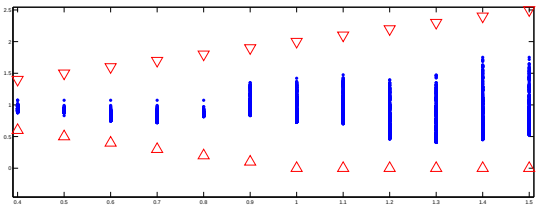

Fig. 2. The evolution of the isometry ratio in (23), as well as the upper and lower bounds in Theorem 10 for different threshold values of the Babel criterion. Same legend as Fig. 1.

and $x_1 = 0$, which yields a chaotic time-series [49]. A set of 200 samples is generated, the Gaussian kernel is used with the bandwidth $\sigma = 0.35$, and the stochastic gradient descent rule (9) with $\epsilon = 0$ is used to generate 200 values of the isometry ratio given in (23). Fig. 1 shows that these values are close to 1, and bounded as demonstrated in Theorem 10. To provide a comparative analysis, we have fixed the size of the dictionaries to $m = 5$ for all the sparsification criteria. The impact of the threshold value of the sparsification criterion is shown in Fig. 2 for the Babel criterion.

VII. Final remarks

This paper provided a framework, based on an eigenvalue analysis, to study sparsity measures and sparsification criteria. We proposed a unified study for the well-conditioning of the optimization problem and for the condition on the uniqueness of the solution. We established a quasi-isometry between the dual space and the dictionary's induced feature space, thus connecting the functional to the dual frameworks and illustrating the impact of the sparsity measures on the topologies. As for future work, we are extending this framework to include new insights on sparse dictionary analysis.

Appendix

The projection of any kernel function $\kappa(\boldsymbol{x}, \cdot)$ onto the subspace spanned by a dictionary of kernel functions $\kappa(\grave{\boldsymbol{x}}_j, \cdot)$, for $j = 1, 2, \ldots, m$, takes the form

$$\grave{\kappa}_{\boldsymbol{x}}(\cdot) = \sum_{j=1}^{m} \xi_j \, \kappa(\grave{\boldsymbol{x}}_j, \cdot),$$

or equivalently $\grave{\kappa}_{\boldsymbol{x}}(\cdot) = \boldsymbol{\xi}^\top \grave{\boldsymbol{\kappa}}(\cdot)$, where $\boldsymbol{\xi}$ is obtained by minimizing the quadratic reconstruction error

$$\|\kappa(\boldsymbol{x}, \cdot) - \boldsymbol{\xi}^\top \grave{\boldsymbol{\kappa}}(\cdot)\|^2_{\mathrm{H}}. \tag{26}$$

The expansion of this norm is given by $\kappa(\boldsymbol{x}, \boldsymbol{x}) - 2\,\boldsymbol{\xi}^\top \grave{\boldsymbol{\kappa}}(\boldsymbol{x}) + \boldsymbol{\xi}^\top \grave{\boldsymbol{K}} \boldsymbol{\xi}$. By taking its derivative with respect to $\boldsymbol{\xi}$ and nullifying it, we get

$$\grave{\boldsymbol{K}} \boldsymbol{\xi} = \grave{\boldsymbol{\kappa}}(\boldsymbol{x}).$$

Therefore, the projection is given by

$$\grave{\kappa}_{\boldsymbol{x}}(\cdot) = \grave{\boldsymbol{\kappa}}(\boldsymbol{x})^\top \grave{\boldsymbol{K}}^{-1} \grave{\boldsymbol{\kappa}}(\cdot).$$

The quadratic reconstruction error of such approximation is obtained by substituting this expression into (26), yielding

$$\kappa(\boldsymbol{x}, \boldsymbol{x}) - \grave{\boldsymbol{\kappa}}(\boldsymbol{x})^\top \grave{\boldsymbol{K}}^{-1} \grave{\boldsymbol{\kappa}}(\boldsymbol{x}).$$

References

[1] R. Baraniuk, E. Candes, R. Nowak, and M. Vetterli, eds., *IEEE signal processing magazine, special issue on "Sensing, Sampling, and Compression"*, vol. 25 (2). IEEE Signal Processing Society, March 2008.

[2] 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.

[3] M. Wu, B. Schölkopf, and G. Bakır, "A direct method for building sparse kernel learning algorithms," *Journal of Machine Learning Research*, vol. 7, pp. 603–624, 2006.

[4] D. Wipf, J. Palmer, and B. Rao, "Perspectives on sparse bayesian learning," in *Advances in Neural Information Processing Systems 16* (S. Thrun, L. Saul, and B. Schölkopf, eds.), MIT Press, 2004.

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

[6] Y. L. Cun, J. S. Denker, and S. A. Solla, "Optimal brain damage," in *Advances in Neural Information Processing Systems 2* (D. S. Touretzky, ed.), pp. 598–605, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1990.

[7] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

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

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

[10] 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.

[11] T. Poggio and S. Smale, "The mathematics of learning: Dealing with data," *Notices of the American Mathematical Society*, 2003.

[12] 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*, COLT/EuroCOLT, (London, UK), pp. 416–426, Springer-Verlag, 2001.

[13] P. Vincent and Y. Bengio, "Kernel matching pursuit," *Mach. Learn.*, vol. 48, pp. 165–187, Sept. 2002.

[14] V. Guigue, A. Rakotomamonjy, and S. Canu, "Kernel basis pursuit," in *Proc. 16th European Conference on Machine Learning* (J. Gama, R. Camacho, P. Brazdil, A. Jorge, and L. Torgo, eds.), Lecture Notes in Computer Science, pp. 146–157, Springer, 2005.

[15] H. V. Nguyen, V. M. Patel, N. M. Nasrabadi, and R. Chellappa, "Kernel dictionary learning.," in *ICASSP*, pp. 2021–2024, IEEE, 2012.

[16] L. Zhang, W.-D. Zhou, P.-C. Chang, J. Liu, Z. Yan, T. Wang, and F.-Z. Li, "Kernel sparse representation-based classifier," *Signal Processing, IEEE Transactions on*, vol. 60, pp. 1684–1695, April 2012.

[17] M. Harandi, C. Sanderson, R. Hartley, and B. Lovell, "Sparse coding and dictionary learning for symmetric positive definite matrices: A kernel approach," in *Computer Vision ECCV 2012* (A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, eds.), Lecture Notes in Computer Science, pp. 216–229, Springer Berlin Heidelberg, 2012.

[18] J. Weston, A. Elisseeff, B. Schölkopf, and M. Tipping, "Use of the zero norm with linear models and kernel methods," *J. Mach. Learn. Res.*, vol. 3, pp. 1439–1461, Mar. 2003.

[19] S. Gao, I. Tsang, and L.-T. Chia, "Sparse representation with kernels," *Image Processing, IEEE Transactions on*, vol. 22, pp. 423–434, Feb 2013.

[20] F. Zhu, P. Honeine, and M. Kallas, "Kernel non-negative matrix factorization without the pre-image problem," in *Proc. 24th IEEE workshop on Machine Learning for Signal Processing*, (Reims, France), 21–24 September 2014.

[21] K. Engan, S. Aase, and J. Hakon Husoy, "Method of optimal directions for frame design," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, vol. 5, pp. 2443–2446 vol.5, 1999.

[22] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *Signal Processing, IEEE Transactions on*, vol. 54, no. 11, pp. 4311–4322, 2006.

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

[24] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998.

[25] 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.
[26] 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.
[27] 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.
[28] 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.
[29] N. Vuković and Z. Miljković, "A growing and pruning sequential learning algorithm of hyper basis function neural network for function approximation," *Neural Netw.*, vol. 46, pp. 210–226, Oct. 2013.
[30] L. Csató and M. Opper, "Sparse online gaussian processes," *Neural Computation*, vol. 14, pp. 641–668, 2002.
[31] 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.
[32] 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.
[33] J. A. Tropp, "Greed is good: algorithmic results for sparse approximation," *IEEE Trans. Information Theory*, vol. 50, pp. 2231–2242, 2004.
[34] 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.
[35] 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.
[36] P. Honeine, M. Essoloh, C. Richard, and H. Snoussi, "Distributed regression in sensor networks with a reduced-order kernel model," in *Proc. 51st IEEE GLOBECOM Global Communications Conference*, (New Orleans, LA, USA), pp. 1–5, 2008.
[37] P. Honeine, C. Richard, H. Snoussi, J. C. M. Bermudez, and J. Chen, "A decentralized approach for non-linear prediction of time series data in sensor networks," *Journal on Wireless Communications and Networking*, vol. Special issue on theoretical and algorithmic foundations of wireless ad hoc and sensor networks, pp. 12:1–12:12, Jan. 2010.
[38] 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.
[39] 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, USA), pp. 600–603, 5–8 August 2012.
[40] M. Yukawa, "Multikernel adaptive filtering," *Signal Processing, IEEE Transactions on*, vol. 60, pp. 4672–4682, Sept 2012.
[41] F. Tobar, S.-Y. Kung, and D. Mandic, "Multikernel least mean square algorithm," *Neural Networks and Learning Systems, IEEE Transactions on*, vol. 25, pp. 265–277, Feb 2014.
[42] T. Ishida and T. Tanaka, "Multikernel adaptive filters with multiple dictionaries and regularization," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific*, pp. 1–6, Oct 2013.
[43] 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.
[44] 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.
[45] A. Sayed, *Fundamentals of adaptive filtering*. NY, USA: Wiley-IEEE Press, June 2003.
[46] J. Kivinen, A. J. Smola, and R. C. Williamson, "Online learning with kernels," *IEEE Transactions on Signal Processing*, vol. 52, Aug 2004.
[47] S. Smale and Y. Yao, "Online learning algorithms," *Found. Comput. Math.*, vol. 6, no. 2, pp. 145–170, 2006.
[48] W. Liu, J. C. Principe, and S. Haykin, *Kernel Adaptive Filtering: A Comprehensive Introduction*. Wiley Publishing, 1st ed., 2010.
[49] P. Honeine, C. Richard, and J. C. M. Bermudez, "Modélisation parcimonieuse non linéaire en ligne par une méthode à noyau reproduisant et un critère de cohérence," in *Actes du XXI-ème Colloque GRETSI sur le Traitement du Signal et des Images*, (Troyes, France), September 2007.
[50] P. Honeine, *Méthodes à noyau pour l'analyse et la décision en environnement non-stationnaire*. PhD thesis, mémoire de thèse de doctorat en Optimisation et Sûreté des Systèmes, Ecole doctoral SSTO - UTT, Troyes, France, 2007.
[51] M. Yukawa, "Adaptive filtering based on projection method." Lecture Notes, December 2010.
[52] G. Golub and C. Van Loan, *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press, 2013.
[53] 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.
[54] S. Van Vaerenbergh, I. Santamaria, W. Liu, and J. Principe, "Fixed-budget kernel recursive least-squares," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 1882–1885, March 2010.
[55] D. Rzepka, "Fixed-budget kernel least mean squares," in *Emerging Technologies Factory Automation (ETFA), IEEE 17th Conference on*, pp. 1–4, Sept 2012.
[56] 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.
[57] A. C. Gilbert, S. Muthukrishnan, and M. J. Strauss, "Approximation of functions over redundant dictionaries using coherence," in *Proc. 14-th annual ACM-SIAM symposium on Discrete algorithms (SODA)*, (Philadelphia, PA, USA), pp. 243–252, Society for Industrial and Applied Mathematics, 2003.
[58] A. C. Gilbert, S. Muthukrishnan, M. J. Strauss, and J. Tropp, "Improved sparse approximation over quasi-incoherent dictionaries," in *International Conference on Image Processing (ICIP)*, vol. 1, (Barcelona, Spain), pp. 37–40, Sept. 2003.
[59] R. A. Horn and C. R. Johnson, *Matrix analysis*. New York, NY, USA: Cambridge University Press, 2nd edition ed., December 2012.
[60] D. Luenberger, *Introduction to Linear and Nonlinear Programming*. Addison-Wesley, second ed., 1989.
[61] V. Kurková and M. Sanguineti, "Learning with generalization capability by kernel methods of bounded complexity," *J. Complex.*, vol. 21, no. 3, pp. 350–367, 2005.
[62] H. Xiang, "A note on the minimax representation for the subspace distance and singular values," *Linear Algebra and its Applications*, vol. 414, no. 2–3, pp. 470 – 473, 2006.

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.