Bulletin of Mathematical Sciences and Applications
ISSN: 2278-9634, Vol. 16, pp 1-19
doi:10.18052/www.scipress.com/BMSA.16.1
2016 SciPress Ltd, Switzerland

Submitted: 2016-03-13
Revised: 2016-05-20
Accepted: 2016-05-30
Online: 2016-08-04

Entropies of Overcomplete Kernel Dictionaries

Paul Honeine

LITIS lab, Universit de Rouen, Rouen, France
paul.honeine@univ-rouen.fr

Keywords: Generalized Rényi entropy, Shannon entropy, Tsallis entropy, sparse approximation, dictionary learning, kernel-based methods, Gram matrix, machine learning, pattern recognition.

Abstract. In signal analysis and synthesis, linear approximation theory considers a linear decomposition of any given signal in a set of atoms, collected into a so-called dictionary. Relevant sparse representations are obtained by relaxing the orthogonality condition of the atoms, yielding overcomplete dictionaries with an extended number of atoms. More generally than the linear decomposition, overcomplete kernel dictionaries provide an elegant nonlinear extension by defining the atoms through a mapping kernel function (*e.g.*, the gaussian kernel). Models based on such kernel dictionaries are used in neural networks, gaussian processes and online learning with kernels.

The quality of an overcomplete dictionary is evaluated with a diversity measure the distance, the approximation, the coherence and the Babel measures. In this paper, we develop a framework to examine overcomplete kernel dictionaries with the entropy from information theory. Indeed, a higher value of the entropy is associated to a further uniform spread of the atoms over the space. For each of the aforementioned diversity measures, we derive lower bounds on the entropy. Several definitions of the entropy are examined, with an extensive analysis in both the input space and the mapped feature space.

Introduction

Sparsity in representation has gained increasing popularity in signal and image processing, for pattern recognition, denoising and compression [11]. A sparse representation of a given signal consists in decomposing it on a set of elementary signals, called atoms and collected in a so-called dictionary. In the linear formalism, the signal is written as a linear combination of the dictionary atoms. This decomposition is unique when the latter defines a basis, and in particular with orthogonal dictionaries such as with the Fourier basis. Since the 1960's, there has been much interest in this direction with the use of predefined dictionaries, based on some analytical form, such as with the wavelets in all its forms (curvelets, contourlets, bandelets, shearlets, directionless, grouplets, platelets, surflets, ...) [29]. Predefined dictionaries have been widely investigated in the literature for years, owing to the mathematical simplicity of such structured dictionaries when dealing with orthogonality (as well as bi-orthogonality). When dealing with sparsity, analytical dictionaries perform poorly in general, due to their rigide structure imposed by the orthogonality.

Within the last 15 years, a new class of dictionaries has emerged with dictionaries learned from data, thus with the ability to adapt to the signal under scrutiny. While the Karhunen-Loève transform — also called principal component analysis in advanced statistics [26] — falls in this class, the relaxation of the orthogonality condition delivers an increased flexibility with overcomplete dictionaries, *i.e.*, when the number of atoms (largely) exceeds the signal dimension. Several methods have been proposed to construct oversomplete dictionaries by solving a highly non-convex optimization problem, such as the method of optimal directions [12], its singular-value-decomposition (SVD) counterpart [1], and the "convexification" method [28].

Overcomplete dictionaries are more versatile to provide relevant representations, owing to an increased diversity. Several measures have been proposed to "quantify" the diversity of a given dictionary. The simplest measure of diversity is certainly the cardinality of the dictionary, *i.e.*, the number of atoms. While this measure is too simplistic, several diversity measures have been proposed by examining relations between atoms, either in a pairwise fashion or in a more thorough way. The most used measure to characterize a dictionary is the coherence, which is the largest pairwise correlation between its atoms [46]. By using the largest cumulative correlation between an atom and all the other atoms of the dictionary, this yields the more exhaustive Babel measure [45]. Over the last twenty years or so, the coherence and its variants (such as the Babel measure) have been used for the matching pursuit algorithm [30] and the basis pursuit with arbitrary dictionaries [9], with theoretical results on the approximation quality studied in [15, 45]; see also the extensive literature on compressed sensing [11].

Beyond the literature on linear approximation, several diversity measures for overcomplete dictionary analysis have been investigated separately in the literature, within different frameworks. This is the case of the distance measure, which corresponds to the smallest pairwise distance between all atoms, as often considered in neural networks. Indeed, in resource-allocating networks for function interpolation, the network of gaussian units is assigned a new unit if this unit is *distant enough* to any other unit already in the network [36, 49]. It turns out that these units operate as atoms in the approximation model, with the corresponding dictionary having a small distance measure. While the distance measure of a given dictionary relies only on its nearest pair of atoms, a more thorough measure is the approximation measure, which corresponds to the least error of approximating any atom of the dictionary with a linear combination of its other atoms. This measure of diversity has been investigated in machine learning with gaussian processes [7], online learning with kernels for nonlinear adaptive filtering [37], and more recently kernel principal component analysis [19].

In order to provide a framework that encloses all the aforementioned methods, we consider the *reproducing kernel Hilbert space* formalism. This allows to generalize the well-known linear model used in sparse approximation to a nonlinear one, where each atom is substituted by a nonlinear one given with a kernel function. This yields the so-called kernel dictionaries, where each atom lives in a feature space, the latter being defined with some nonlinear transformation of the input space. While the linear kernel yields the conventional linear model, as given in the literature of linear sparse approximation, the use of nonlinear kernels such as the gaussian kernel, allows to include in our study neural networks with ressource-allocating networks, nonlinear adaptive filtering with kernels and gaussian processes.

All the aforementioned diversity measures allow to quantify the heterogeneity within the dictionary under scrutiny. In this paper, we derive connections between these measures and the entropy in information theory (which is also related to the definition of entropy in other fields, such as thermodynamics and statistical mechanics) [5]. Indeed, the entropy measures the disorder or randomness within a given system. By considering the generalized Rényi entropy, which englobes the definitions given by Shannon, Hartley, as well as the quadratic formulation, we show that any overcomplete kernel dictionary with a given diversity measure has a lower-bounded entropy. These results on the high values of the entropy illustrate that the atoms are favorably spread uniformly over the space. We provide a comprehensive analysis, for any kernel type and any entropy definition, within the Rényi entropy framework as well as the more recent nonadditive entropy proposed by Tsallis [47, 4]. Finally, we provide an entropy analysis in the feature space by deriving lower bounds depending on the diversity measures. As a consequence, we connect the diversity measures between both input and feature spaces.

The remained of this paper is organized as follows. Next section introduces the sparse approximation problem, in its conventional linear model as well as its nonlinear extension with the kernel formalism. Section presents the most used diversity measures for quantifying overcomplete dictionaries, while Section provides a preliminary exploration with results that will be used throughout this paper. Section is the core of this work, where we define the entropy and examine it in the input space, while Section extends this analysis to the feature space. Section concludes this paper.

*Related work

In [17], Girolami considered the estimation of the quadratic entropy with a set of samples, by using the Parzen estimator based on a normalized kernel function. This formulation was investigated in regularization networks, and in particular *least-squares support vector machines* (LS-SVM), in order to reduce the computational complexity by pruning samples that do not contribute sufficiently to the entropy [44]. More recently, an online learning scheme was proposed in [27] for LS-SVM by using the approximation measure as a sparsification criterion. In our paper, we derive the missing connections between this criterion and the entropy maximization.

Richard, Bermudez and Honeine considered in [39] the analysis of the quadratic entropy of a kernel dictionary in terms of its coherence. We provide in our paper a framework to analyse overcomplete dictionaries with a more extensive examination, in both input and feature spaces, and generalizing to other entropy definitions and all types of kernels. The conducted analysis examines several diversity measures, including, but not limited to, the coherence measure.

A primer on overcomplete (kernel) approximation

In this section, we introduce the sparse approximation problem, in its conventional linear model as well as the kernel-based formulation. We conclude this section with an outline of the issues addressed in this paper.

A primer on sparse approximation

Consider a Banach space $\mathbb{X}$ of $\mathbb{R}^d$, denoted input space. The approximation theory studies the representation of a given signal $\boldsymbol{x}$ of $\mathbb{X}$ with a dictionary of atoms (*i.e.*, set of elementary signals), $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{X}$, and estimating their fractions in the signal under scrutiny. In linear approximation, the decomposition takes the form:

$$\boldsymbol{x} \approx \sum_{i=1}^{n} \alpha_i \, \boldsymbol{x}_i. \tag{1}$$

This representation is unique when the atoms form a basis, by approximating the signal with its projection onto the span of the atoms, namely $\alpha_i = \boldsymbol{x}_i^\top \boldsymbol{x}$. Examples that involve orthonormal bases include the Fourier transform and the discrete cosine transform, as well as the data-dependent Karhunen-Loève transform (*i.e.*, the PCA).

Beyond these orthogonal bases, the relaxation of the orthogonality provides more flexibility with the use of overcomplete dictionaries, which allows to investigate different constraints more properly, such as the sparsity of the representation. In this case, the coefficients α_i in (1) are obtained by promoting the sparsity of the representation. This optimization problem is often called sparse coding, assuming that the dictionary is known. In view of the vector $[\alpha_1 \;\; \alpha_2 \;\; \cdots \;\; \alpha_n]^\top$, sparsity can be promoted by minimizing its ℓ_0 pseudo-norm, which counts the number of non-zero entries, or its ℓ_1 norm, which is the closest convex norm to the ℓ_0 pseudo-norm [18].

Since the seminal work [35] where Olshausen and Field considered learning the atoms from a set of available data, data-driven dictionaries have been widely investigated. A large class of approaches have been proposed to solve iteratively the optimization problem by alternating between the dictionary learning (*i.e.*, estimating the atoms $\boldsymbol{x}_i$) and the sparse coding (*i.e.*, estimating the coefficients α_i). The former problem is essentially tackled with the maximum likelihood principle of the data or the maximum *a posteriori* probability of the dictionary. The latter corresponds to the sparse coding problem. The best known methods for solving the optimization problem[1]

$$\arg \min_{\substack{\alpha_i, \boldsymbol{x}_i \\ i=1\cdots n}} \Big\| \boldsymbol{x} - \sum_{i=1}^{n} \alpha_i \, \boldsymbol{x}_i \Big\|^2, \tag{2}$$

subject to some sparsity promoting constraint, are the method of optimal directions [12] and the K-SVD algorithm [1], where the dictionary is determined respectively with the Moore-Penrose pseudo-inverse and the SVD scheme. For more details, see [11] and references therein. It is worth noting that

[1] In practice, one has several signals $\boldsymbol{x}$ in order to construct the dictionary, resulting in a Frobenius norm minimization.

the sparsity constraint yields a difficult optimization problem, even when the model is linear in both coefficients and atoms.

Kernel-based approximation

Nonlinear models provide a more challenging issue. The formalism of *reproducing kernel Hilbert spaces* (RKHS) provides an elegant and efficient framework to tackle nonlinearities. To this end, the signals $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ are mapped with a nonlinear function into some feature space $\mathbb{H}$, as follows:

$$\begin{aligned} \mathbb{X} &\mapsto \mathbb{H} \\ \boldsymbol{x}_i &\rightarrow \kappa(\boldsymbol{x}_i, \cdot) \end{aligned}$$

Here, $\kappa \colon \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$ is a positive definite kernel and the feature space $\mathbb{H}$ is the so-called reproducing kernel Hilbert space. Let $\langle \cdot, \cdot \rangle_{\mathbb{H}}$ and $\|\cdot\|_{\mathbb{H}}$ denote respectively the inner product and norm in the induced space $\mathbb{H}$. This space has some interesting properties, such as the reproducing property which states that any function $\psi(\cdot)$ of $\mathbb{H}$ can be evaluated at any $\boldsymbol{x}_i$ of $\mathbb{X}$ using $\psi(\boldsymbol{x}_i) = \langle \psi(\cdot), \kappa(\boldsymbol{x}_i, \cdot) \rangle_{\mathbb{H}}$. Moreover, we have the kernel trick, that is $\langle \kappa(\boldsymbol{x}_i, \cdot), \kappa(\boldsymbol{x}_j, \cdot) \rangle_{\mathbb{H}} = \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$ for any $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathbb{X}$. In particular, $\|\kappa(\cdot, \boldsymbol{x}_i)\|_{\mathbb{H}}^2 = \kappa(\boldsymbol{x}_i, \boldsymbol{x}_i)$.

Kernels can be roughly divided in two categories, projective kernels as functions of the data inner product (*i.e.*, $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$), and radial kernels as functions of their distance (*i.e.*, $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|$). The most used kernels and there expressions are given in Table 1. From these kernels, only the gaussian and the radial-based exponential kernels are unit-norm, that is $\|\kappa(\boldsymbol{x}, \cdot)\|_{\mathbb{H}} = 1$ for any $\boldsymbol{x} \in \mathbb{X}$. In this paper, we do not restrict ourselves to a particular kernel. We denote

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

where $\kappa(\boldsymbol{x}, \boldsymbol{x}) = \|\kappa(\boldsymbol{x}, \cdot)\|_{\mathbb{H}}^2$. For unit-norm kernels, we get $R = r = 1$.

While the linear kernel yields the conventional model given in (1), nonlinear kernels such as the gaussian kernel provide the models investigated in RBF neural networks, gaussian processes [38] and kernel-based machine learning [42], including the celebrated support vector machines [48]. For a set of data $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{X}$ and a given kernel $\kappa(\cdot, \cdot)$, the induced RKHS $\mathbb{H}$ is defined such as any element $\psi(\cdot)$ of $\mathbb{H}$ takes the form

$$\psi(\cdot) = \sum_{i=1}^{n} \alpha_i \, \kappa(\boldsymbol{x}_i, \cdot). \tag{3}$$

When dealing with an approximation problem in the same spirit of (1)-(2), the element $\psi(\cdot)$ is approximated by $\kappa(\boldsymbol{x}, \cdot)$. Compared to the linear case given in (1), it is easy to see that the above model is still linear in the coefficients α_i, as well as the "atoms" $\kappa(\boldsymbol{x}_i, \cdot)$, while it is nonlinear with respect to $\boldsymbol{x}_i$. Indeed, the resulting optimization problem consists in minimizing the residual in the RKHS, with

$$\arg \min_{\substack{\alpha_i, \boldsymbol{x}_i \\ i = 1 \cdots n}} \Big\| \kappa(\boldsymbol{x}, \cdot) - \sum_{i=1}^{n} \alpha_i \, \kappa(\boldsymbol{x}_i, \cdot) \Big\|_{\mathbb{H}}^2.$$

On the one hand, the estimation of the coefficients is similar to the one given in the linear case with (2); the classical (linear) sparse coders can be investigated for this purpose. On the other hand, the dictionary determination is more difficult, since the model is nonlinear in the $\boldsymbol{x}_i$; thus, conventional techniques such as the K-SVD algorithm can no longer be used. It turns out that the estimation of the elements in the input space is a tough optimization problem, known in the literature as the pre-image problem [22]. More recently, the authors of [40, 41] adjusted the elements $\boldsymbol{x}_i$ in the input space for nonlinear adaptive filtering with kernels. In another context, the authors of [54, 56, 53] estimated these elements for the kernel non-negative matrix factorization. See also [52, 24, 50, 51, 55].

Table 1: The most used kernels with their expressions, including tunable parameters $p, \sigma > 0$ and $c \geq 0$. These kernels are grouped in two categories: projective kernels as functions of $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$, and radial kernels as functions of $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|$.

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

Addressed issues

In either analysis or synthesis of overcomplete (kernel) dictionaries, with the grow in the number of atoms, an increase in the heterogeneity of the atoms is needed. Such diversification requires that the atoms are not too “close” to each other. Depending on the definition of closeness, several diversity measures have been proposed in the literature. This is the case when the closeness is given in terms of the metric, as given with the distance measure for a pairwise measure between atoms, or the approximation measure for a more thorough measure. This is also the case when the collinearity of the atoms is considered, such as with the coherence and the Babel measures. These diversity measures are described in detail in Section , within the formalism for a kernel dictionary $\{\kappa(\boldsymbol{x}_1, \cdot), \kappa(\boldsymbol{x}_2, \cdot), \ldots, \kappa(\boldsymbol{x}_n, \cdot)\}$.

In this paper, we connect these diversity measures to the entropy from information theory [5]. Indeed, from the viewpoint of information theory, the set $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ can be viewed as a finite source alphabet. A fundamental measure of information is the entropy, which quantifies the disorder or randomness of a given system or set. It is also associated to the number of bits needed, in average, to store or communicate the set under investigation. A detailed definition of the entropy is given in Section , with connections between the entropy of the set $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ and the aforementioned diversity measures of the associated kernel dictionary $\{\kappa(\boldsymbol{x}_1, \cdot), \kappa(\boldsymbol{x}_2, \cdot), \ldots, \kappa(\boldsymbol{x}_n, \cdot)\}$. Several entropy definitions are also investigated, including the generalized Rényi entropy and the Tsallis entropy. Finally, Section extends this analysis to the RKHS, by studying the entropy of set of atoms $\{\kappa(\boldsymbol{x}_1, \cdot), \kappa(\boldsymbol{x}_2, \cdot), \ldots, \kappa(\boldsymbol{x}_n, \cdot)\}$.

Diversity measures

In this section, we present measures that quantify the diversity of a given dictionary $\{\kappa(\boldsymbol{x}_1, \cdot), \kappa(\boldsymbol{x}_2, \cdot), \ldots, \kappa(\boldsymbol{x}_n, \cdot)\}$. Each diversity measure is associated to a sparsification criterion for online learning, in order to construct dictionaries with large diversity measures.

Cardinality

The cardinality of the dictionary, namely the number n of atoms, is the simplest measure. However, such measure does not take into account that some atoms can be close to each others, *e.g.*, duplicata.

Distance measure

A simple measure to characterize a dictionary is the smallest distance between all pairs of its atoms, namely

$$\min_{\substack{i,j=1\cdots n\\ i\neq j}} \|\kappa(\boldsymbol{x}_i,\cdot)-\kappa(\boldsymbol{x}_j,\cdot)\|_{\mathbb{H}},$$

where

$$\|\kappa(\boldsymbol{x}_i,\cdot)-\kappa(\boldsymbol{x}_j,\cdot)\|_{\mathbb{H}}^2=\kappa(\boldsymbol{x}_i,\boldsymbol{x}_i)-2\,\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)+\,\kappa(\boldsymbol{x}_j,\boldsymbol{x}_j). \tag{4}$$

In the following, we consider a tighter measure by using the distance between any two atoms, up to a scaling factor, which is a tighter measure since we have

$$\|\kappa(\boldsymbol{x}_i,\cdot)-\kappa(\boldsymbol{x}_j,\cdot)\|_{\mathbb{H}}\geq\min_{\xi}\|\kappa(\boldsymbol{x}_i,\cdot)-\xi\kappa(\boldsymbol{x}_j,\cdot)\|_{\mathbb{H}}. \tag{5}$$

A dictionary is said to be δ-distant when

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

Since the above distance is equivalent to the residual error of approximating any atom by its projection onto another atom, the optimal scaling factor ξ takes the value $\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)/\kappa(\boldsymbol{x}_j,\boldsymbol{x}_j)$, yielding

$$\delta^2=\min_{\substack{i,j=1\cdots n\\ i\neq j}}\left(\kappa(\boldsymbol{x}_i,\boldsymbol{x}_i)-\frac{\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)^2}{\kappa(\boldsymbol{x}_j,\boldsymbol{x}_j)}\right).$$

When dealing with unit-norm atoms, this expression boils down to $\delta^2=1-\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)^2$.

A sparsification criterion for online learning is studied in ressource-allocating networks [36, 3] with the “novelty criterion”, by imposing a lower bound on the distance measure of the dictionary. Thus, any candidate atom is included in the dictionary if the distance measure of the latter does not fall below a given threshold that controls the level of sparseness.

Approximation measure

While the distance measure relies only on the nearest two atoms, the approximation measure provides a more exhaustive analysis by quantifying the capacity of approximating any atom with a linear combination of the other atoms of the dictionary. A dictionary is said to be δ-approximate if the following is satisfied:

$$\delta=\min_{i=1\cdots n}\min_{\xi_1\cdots\xi_n}\Bigg\|\kappa(\boldsymbol{x}_i,\cdot)-\sum_{\substack{j=1\\ j\neq i}}^{n}\xi_j\,\kappa(\boldsymbol{x}_j,\cdot)\Bigg\|_{\mathbb{H}}. \tag{6}$$

This expression corresponds to the residual error of projecting any atom onto the subspace spanned by the others atoms. By nullifying the derivative of the above cost function with respect to each coefficient ξ_j, we get the optimal vector of coefficients

$$\boldsymbol{\xi}=\boldsymbol{K}_{\backslash\{i\}}^{-1}\boldsymbol{\kappa}_{\backslash\{i\}}(\boldsymbol{x}_i), \tag{7}$$

Here, $\boldsymbol{K}_{\backslash\{i\}}$ and $\boldsymbol{\kappa}_{\backslash\{i\}}(\boldsymbol{x}_i)$ are obtained by removing the entries associated to $\boldsymbol{x}_i$ from $\boldsymbol{K}$ and $\boldsymbol{\kappa}(\boldsymbol{x}_i)$, respectively, where $\boldsymbol{K}$ is the Gram matrix of entries $\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)$ and $\boldsymbol{\kappa}(\cdot)$ is the column vector of entries $\kappa(\boldsymbol{x}_j,\cdot)$, for $i,j=1,\ldots,n$. By plugging the above expression in (6), we obtain:

$$\delta^2=\min_{i=1\cdots n}\kappa(\boldsymbol{x}_i,\boldsymbol{x}_i)-\boldsymbol{\kappa}_{\backslash\{i\}}(\boldsymbol{x}_i)^\top\boldsymbol{K}_{\backslash\{i\}}^{-1}\boldsymbol{\kappa}_{\backslash\{i\}}(\boldsymbol{x}_i). \tag{8}$$

The sparsification criterion associated to the approximation measure is studied in [2, 6] and more recently in [13] for system identification and [19] for kernel principal component analysis. This criterion constructs dictionaries with a high approximation measure, thus including any candidate atom in the dictionary if it cannot be well approximated by a linear combination of atoms already in the dictionary, for a given approximation threshold.

Coherence measure

In the literature of sparse linear approximation, the coherence is a fundamental quantity to characterize dictionaries. It corresponds to the largest correlation between atoms of a given dictionary, or mutually between atoms of two dictionaries. While initially introduced for linear matching pursuit in [30], it has been studied for the union of two bases [10], for basis pursuit with arbitrary dictionaries [9], for the analysis of the approximation quality [15, 45]. While most work consider the use of a linear measure, we explore in the following the coherence of a kernel dictionary, as initially studied in [23].

For a given dictionary, the coherence is defined by the largest correlation between all pairs of atoms, namely

$$\max_{\substack{i,j=1\cdots n\\ i\neq j}} \frac{|\langle \kappa(\boldsymbol{x}_i,\cdot),\kappa(\boldsymbol{x}_j,\cdot)\rangle_{\mathbb{H}}|}{\|\kappa(\boldsymbol{x}_i,\cdot)\|_{\mathbb{H}}\|\kappa(\boldsymbol{x}_j,\cdot)\|_{\mathbb{H}}}.$$

It is easy to see that this definition can be written, for a so-called γ-coherent dictionary, as follows:

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

For unit-norm atoms, we get $\max\limits_{\substack{i,j=1\cdots n\\ i\neq j}} |\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)|$.

The coherence criterion for sparsification constructs a “low-coherent” dictionary, thus enforcing an upper bound on the cosine angle between each pair of atoms [39]. In this case, any candidate atom is included in the dictionary if the coherence of the latter does not exceed a given threshold.This threshold controls the level of sparseness of the dictionary, where a null value yields an orthogonal basis.

Babel measure

While the coherence relies only on the most correlated atoms in the dictionary, a more thorough measure is the Babel measure which considers the largest cumulative correlation between an atom and all the other atoms of the dictionary. The Babel measure can be defined in two ways. The first one is by connecting it to the coherence measure, with a definition related to the cumulative coherence, namely

$$\max_{i=1\cdots n} \sum_{\substack{j=1\\ j\neq i}}^{n} \frac{|\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)|}{\sqrt{\kappa(\boldsymbol{x}_i,\boldsymbol{x}_i)\,\kappa(\boldsymbol{x}_j,\boldsymbol{x}_j)}}. \tag{10}$$

The second (and most conventional) way to define the Babel measure is by investigating an analogy with the norm operator [16, 45]. Indeed, while the coherence is the ∞-norm of the Gram matrix when dealing with unit-norm atoms, the Babel measure explores the ℓ_1 matrix-norm, where $\|\boldsymbol{K}\|_1 = \max_i \sum_j |\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)|$. As a consequence, a dictionary is said to be γ-Babel when

$$\gamma = \max_{i=1\cdots n} \sum_{\substack{j=1\\ j\neq i}}^{n} |\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)|. \tag{11}$$

Connecting this definition with (10) — for not necessary unit-norm atoms — is straightforward, since the latter can be box-bounded for any γ-Babel dictionary defined by (11), with

$$\frac{\gamma}{R^2} \leq \max_{i=1\cdots n} \sum_{\substack{j=1\\ j\neq i}}^{n} \frac{|\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)|}{\sqrt{\kappa(\boldsymbol{x}_i,\boldsymbol{x}_i)\,\kappa(\boldsymbol{x}_j,\boldsymbol{x}_j)}} \leq \frac{\gamma}{r^2}.$$

For this reason and for the sake of simplicity, we consider the definition (11) in this paper.

The sparsification criterion associated to the Babel measure constructs dictionaries with a low cumulative coherence [14]. To this end, any candidate atom $\kappa(\boldsymbol{x}_t,\cdot)$ is included in the dictionary if (and only if)

$$\sum_{j=1}^{n} |\kappa(\boldsymbol{x}_t,\boldsymbol{x}_j)| \tag{12}$$

does not exceed a given positive threshold.

Some fundamental results

Before proceeding throughout this paper with a rigorous analysis of any overcomplete dictionary in terms of its diversity measure, we provide in the following some results that are essential to our study. These results provide an attempt to bridge the gap between the different diversity measures.

Coherence versus Babel measure

The following theorems connect the coherence of a dictionary to its Babel measure by quantifying the Babel measure of a γ-coherent dictionary, and vice-versa. The following theorem has been known for a while in the case of unit-norm atoms.

Theorem 1. *A γ-coherent dictionary has a Babel measure that does not exceed $(n-1)\gamma R^2$.*

Proof. Following the definition (11), the Babel of a γ-coherent dictionary is upper-bounded as follows:

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

Furthermore, it is also easy to provide an upper bound on the coherence of a dictionary with a given Babel measure, as given in the following theorem.

Theorem 2. *A γ-Babel dictionary has a coherence that does not exceed γ/r^2.*

Proof. The proof follows from the relation

$$\max_{\substack{i,j=1\cdots n\\ i\neq j}} \frac{|\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)|}{\sqrt{\kappa(\boldsymbol{x}_i,\boldsymbol{x}_i)\,\kappa(\boldsymbol{x}_j,\boldsymbol{x}_j)}} \leq \max_{\substack{i,j=1\cdots n\\ i\neq j}} \frac{|\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)|}{r^2},$$

and the inequality between matrix norms: $\|\cdot\|_{\max} \leq \|\cdot\|_{\infty}$.

Analysis of a δ-approximate dictionary

The following theorem is fundamental in the analysis of a dictionary resulting from the approximation criterion.

Theorem 3. *A δ-approximate dictionary has a Babel measure that does not exceed $R^2 - \delta^2$, and a coherence measure that does not exceed*

$$\frac{R^2 - \delta^2}{r^2}.$$

Proof. For a δ-approximate dictionary, we have from (7): $\boldsymbol{K}_{\backslash\{i\}}\boldsymbol{\xi} = \boldsymbol{\kappa}_{\backslash\{i\}}(\boldsymbol{x}_i)$, for any $i = 1, 2, \ldots, n$. By plugging this relation in (8), we obtain

$$\min_{\boldsymbol{\xi}} \kappa(\boldsymbol{x}_i, \boldsymbol{x}_i) - \boldsymbol{\kappa}_{\backslash\{i\}}(\boldsymbol{x}_i)^\top \boldsymbol{\xi} \geq \delta^2.$$

By considering the special case of the vector $\boldsymbol{\xi}$ with $\xi_j = \operatorname{sign}(\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j))$, for any $j = 1, 2, \ldots, n$ and $j \neq i$, we get

$$\sum_{\substack{j=1 \\ j\neq i}}^{n} |\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)| \leq \kappa(\boldsymbol{x}_i, \boldsymbol{x}_i) - \delta^2,$$

for all $i = 1, 2, \ldots, n$. As a consequence,

$$\max_{i=1\cdots n} \sum_{\substack{j=1 \\ j\neq i}}^{n} |\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)| \leq \max_{i=1\cdots n} \kappa(\boldsymbol{x}_i, \boldsymbol{x}_i) - \delta^2 \leq R^2 - \delta^2.$$

This concludes the proof for the Babel measure, since it is the left-hand-side in the above expression, while the upper bound on the coherence measure is obtained from the aforementioned connection between the coherence and the Babel measures as given in Theorem 2.

Entropy analysis in the input space

The entropy measures the disorder or randomness within a given system. The Rényi entropy provides a generalization of well-known entropy definitions, such as Shannon and Harley entropies as well as the quadratic entropy (see Table 2). It is defined for a given order α by

$$H_\alpha = \frac{1}{1-\alpha} \log \int_{\mathbb{X}} \big(P(\boldsymbol{x})\big)^\alpha \, d\boldsymbol{x}, \tag{13}$$

for the probability distribution P that governs all elements $\boldsymbol{x}$ of $\mathbb{X}$. When dealing with discrete random variables as in source coding, this definition is restricted to the set $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ drawn from the probability distribution P, yielding the expression

$$H_\alpha = \frac{1}{1-\alpha} \log \sum_{j=1}^{n} \big(P(\boldsymbol{x}_j)\big)^\alpha. \tag{14}$$

Large values of the entropy correspond to a more uniform spread of the data[2]. Since this probability distribution is unknown in practice, it is often approximated with a Parzen window estimator (also called kernel density estimator). The estimator takes the form

$$\widehat{P}(\boldsymbol{x}) = \frac{1}{n} \sum_{j=1}^{n} w(\|\boldsymbol{x} - \boldsymbol{x}_j\|), \tag{15}$$

[2] It is well-known for the Shannon entropy (*i.e.*, where $\alpha \to 1$) that the uniform distribution yields the largest entropy. This property seems to extend to the case of any non-zero order, including $\alpha \to \infty$ where we get the min-entropy. See Table 2 for the expressions of well-known entropies.

for a given window function w centered at each $\boldsymbol{x}_j$. For more details, see for instance [25].

In the following, we provide lower bounds on the entropy of an overcomplete dictionary, in terms of its diversity measure. To this end, we initially restrict ourselves to the case of the quadratic entropy (*i.e.*, $\alpha = 2$), first with the gaussian kernel then with any type of kernel, before generalizing these results to any order α of the Rényi entropy as well as the Tsallis entropy.

The quadratic entropy with the gaussian kernel

Before generalizing to any window function in Section and any order in Section , we restrict ourselves first to the case of the gaussian window function with the quadratic entropy. The quadratic entropy is defined by $H_2 = -\log \sum_{j=1}^{n} \big(P(\boldsymbol{x}_j)\big)^2$. Considering the normalized gaussian window

$$w(\|\boldsymbol{x} - \boldsymbol{x}_j\|) = \frac{1}{(\sqrt{\pi}\sigma)^d} \exp\left(-\|\boldsymbol{x} - \boldsymbol{x}_j\|^2/\sigma^2\right),$$

for some bandwidth parameter σ, the Parzen estimator becomes

$$\widehat{P}(\boldsymbol{x}) = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{(\sqrt{\pi}\sigma)^d} \exp\left(-\|\boldsymbol{x} - \boldsymbol{x}_j\|^2/\sigma^2\right).$$

Since the convolution of two gaussian distributions leads to another gaussian distribution, then $H_2 \approx -\log \sum_{j=1}^{n} \big(\widehat{P}(\boldsymbol{x}_j)\big)^2$ becomes

$$H_2 \approx -\log\left(\frac{1}{n^2}\sum_{i,j=1}^{n} \frac{\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)}{(2\pi\sigma^2)^{d/2}}\right) = \frac{d}{2}\log(2\pi\sigma^2) - \log\left(\frac{1}{n^2}\sum_{i,j=1}^{n} \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)\right),$$

where $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(\frac{-1}{2\sigma^2}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2\right)$ is the gaussian kernel. This expression shows that the sum of the entries in the Gram matrix describes the diversity of the dictionary elements, a result corroborated in [17] and more recently in [25]. This property was investigated in [44] for pruning the LS-SVM, by removing samples with the smallest entries in the Gram matrix.

Each diversity measure studied in Section yields a lower bound on the entropy of the dictionary under scrutiny. To shown this, we consider first the Babel measure with a γ-Babel dictionary. Following the Babel definition in (11), the entropy given in is lower-bounded as follows:

$$H_2 \geq \frac{d}{2}\log(2\pi\sigma^2) + \log n - \log(1+\gamma),$$

where we have used the following upper bound on the summation:

$$\sum_{i,j=1}^{n} \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{i=1}^{n} \kappa(\boldsymbol{x}_i, \boldsymbol{x}_i) + \sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) \leq n(1+\gamma).$$

This result provides the core of the proof. Indeed, Theorem 2 shows that this result holds also for a γ-coherent dictionary. Furthermore, we can improve this bound for the coherence measure, since $\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) \leq n(n-1)\gamma$, thus yielding the following lower bound on the entropy

$$H_2 \geq \frac{d}{2}\log(2\pi\sigma^2) + \log n - \log\big(1 + (n-1)\gamma\big).$$

This result is also shared with a δ-distant dictionary, by substituting γ with $\sqrt{1-\delta^2}$, since the distance is equivalent to the coherence when dealing with normalized kernels. Finally, Theorem 3 establishes the connection with a δ-approximate dictionary, where the above upper bound becomes

Table 2: The most known entropies as special cases of the generalized Rényi entropy.

Entropy	order α	H_α
Harley entropy	$\alpha = 0$	$\log n$
Shannon entropy	$\alpha \to 1$	$-\sum_{j=1}^{n} P(\boldsymbol{x}_j) \log P(\boldsymbol{x}_j)$
Quadratic entropy	$\alpha = 2$	$-\log \sum_{j=1}^{n} \big(P(\boldsymbol{x}_j)\big)^2$
Min-entropy	$\alpha \to \infty$	$\min_{j=1\cdots n} - \log P(\boldsymbol{x}_j)$

$$H_2 \geq \frac{d}{2} \log\big(2\pi\sigma^2\big) + \log n - \log\left(2 - \delta^2\right).$$

All these results provide lower bounds on the entropy, with the following observations. These bounds increase with the number of elements in the dictionary, *i.e.*, n, which is obvious as the diversity grows. They decrease when the coherence and the Babel measures increase, while they increase when the distance and the approximation measures increase. These results provide quantitative details that confront the fact that, when using a sparsification criterion for online learning, low values of the coherence and Babel thresholds provide less "correlated" atoms and thus more diversity within the dictionary, as opposed to high values of the distance and approximation thresholds. For online learning with kernel for one-class classification, see [31, 34, 33]

The quadratic entropy with any kernel

The results presented so far can be extended to any kernel, even non-unit-norm kernels. To see this, we define the Parzen estimator in a RKHS, by writing the integral $\int_{\mathbb{X}} \widehat{P}(\boldsymbol{x})^2 \, d\boldsymbol{x}$ as the quadratic norm $\|\widehat{P}\|_{\mathcal{H}}^2$ of

$$\widehat{P}(\cdot) = \frac{1}{n} \sum_{i=1}^{n} \kappa(\boldsymbol{x}_i, \cdot),$$

where the norm is given in the subspace spanned by the kernel functions $\kappa(\boldsymbol{x}_1, \cdot), \kappa(\boldsymbol{x}_2, \cdot), \ldots, \kappa(\boldsymbol{x}_n, \cdot)$. Therefore, we have

$$H_2 \approx -\log \|\widehat{P}\|_{\mathcal{H}}^2 = -\log\left(\frac{1}{n^2} \sum_{i,j=1}^{n} \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)\right).$$

By following the same steps as in Section , we can derive the following lower bounds on the quadratic entropy:

- $\log n - \log\left(R^2 + (n-1)R\sqrt{R^2 - \delta^2}\right)$ for a δ-distant dictionary.
- $\log n - \log\left(2R^2 - \delta^2\right)$ for a δ-approximate dictionary.
- $\log n - \log\left(R^2 + (n-1)\gamma R^2\right)$ for a γ-coherent dictionary.
- $\log n - \log(R^2 + \gamma)$ for a γ-Babel dictionary.

Before providing the proof of these results, it is worth noting that the conclusion and discussion conducted in the case of the gaussian kernel are still satisfied in the general case of any kernel type.

Proof. The bounds for the δ-approximate and γ-Babel dictionaries are straightforward from Theorem 3 and the definition in (11). The lower bounds for γ-coherent and δ-distant dictionaries are a bit trickier to prove. To show this, we use for the former the following relation

$$\begin{aligned} H_2 &\geq -\log\Bigg(\frac{1}{n^2}\sum_{i=1}^{n}\kappa(\boldsymbol{x}_i,\boldsymbol{x}_i)+\frac{\gamma}{n^2}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\sqrt{\kappa(\boldsymbol{x}_i,\boldsymbol{x}_i)\,\kappa(\boldsymbol{x}_j,\boldsymbol{x}_j)}\Bigg)\\ &\geq \log n-\log\big(R^2+(n-1)\gamma R^2\big), \end{aligned}$$

and for the latter the following relation

$$\begin{aligned} H_2 &\geq -\log\Bigg(\frac{1}{n^2}\sum_{i=1}^{n}\kappa(\boldsymbol{x}_i,\boldsymbol{x}_i)+\frac{1}{n^2}\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\sqrt{\big(\kappa(\boldsymbol{x}_i,\boldsymbol{x}_i)-\delta^2\big)\kappa(\boldsymbol{x}_j,\boldsymbol{x}_j)}\Bigg)\\ &\geq \log n-\log\big(R^2+(n-1)R\sqrt{R^2-\delta^2}\big). \end{aligned}$$

Generalization to Rényi and Tsallis entropies

So far, we have investigated the quadratic entropy and derived lower bounds for each diversity measure. It turns out that these results can be extended to the general Rényi entropy and Tsallis entropy, as shown next. Special cases of the former are listed in Table 2, including the Harley or maximum entropy which is associated to the cardinality of the set, the Shannon entropy which is essentially the Gibbs entropy in statistical thermodynamics, the quadratic entropy also called collision entropy, as well as the min-entropy which is the smallest measure in the family of Rényi entropies.

Corollary 4. *Any lower bound ζ on the quadratic entropy provides lower bounds on the Hartley entropy H_0, the Shannon H_1, and the min-entropy H_∞, with*

$$\zeta\leq H_1\leq H_0 \qquad \text{and} \qquad \tfrac{1}{2}\zeta\leq H_\infty.$$

Proof. The proof is due to the Jensen's inequality and the concavity of the Rényi entropy for nonnegative orders. First, the relation of the Shannon entropy is given by exploring the following inequality:

$$\sum_{j=1}^{n}P(\boldsymbol{x}_j)\log P(\boldsymbol{x}_j)\leq\log\sum_{j=1}^{n}\big(P(\boldsymbol{x}_j)\big)^2.$$

The connection to the Hartley entropy is straightforward, with $H_0=\log n$. Finally, it is more trickier to study the min-entropy, since it is the smallest entropy measure in the family of Rényi entropies, as a consequence it is the strongest way to measure the information content. To provide a lower bound on the min-entropy, we use the relations

$$\log\sum_{j=1}^{n}\big(P(\boldsymbol{x}_j)\big)^2\geq\log\max_{j=1\cdots n}\big(P(\boldsymbol{x}_j)\big)^2=2\log\max_{j=1\cdots n}P(\boldsymbol{x}_j),$$

which yields the following inequality: $H_2\leq 2H_\infty$.

Furthermore, one can easily extend these results to the class of the Tsallis entropy, also called nonadditive entropy, defined by the following expression for a given parameter q (called entropic-index) [47, 4]:

$$\frac{1}{q-1}\Big(1-\sum_{j=1}^{n}\big(P(\boldsymbol{x}_j)\big)^q\Big).$$

To this end, the aforementioned lower bounds on the Rényi entropy can be extended to the Tsallis entropy by using for instance the well-known relation $\log u\leq u-1$ for any $u\geq 0$.

As a consequence, the lower bounds on the quadratic entropy given in Sections and can be explored to other orders of Rényi entropy and Tsallis entropy.

Entropy in the feature space

By analogy with the entropy analysis in the input space conducted in Section , we propose to revisit it in the feature space, as given in this section. By examining the pairwise distance between any two atoms of the investigated dictionary, we first establish in Section a topological analysis of overcomplete dictionaries. This analysis is explored in Section with the study of the entropy of the atoms in the feature space. By providing lower bounds in terms of the diversity measures, these results provide connections to the entropy analysis conducted in the previous section.

Fundamental analysis

The following theorem is used in the following section for the analysis of the atoms of a kernel dictionary.

Theorem 5. *For any dictionary with a non-zero approximation measure, or a non-unit coherence measure, or a Babel measure below r^2, we have a low-bounded distance measure.*

Proof. The proof is straightforward for a δ-approximate dictionary, since

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

For the coherence measure, we consider the pairwise distance in terms of kernels as given in (4). Since a γ-coherent dictionary satisfies

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

then we have

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

Thus, $\|\kappa(\boldsymbol{x}_i,\cdot)-\kappa(\boldsymbol{x}_j,\cdot)\|_{\mathbb{H}}^2$ from the right-hand-side of equation (4) is lower-bounded by

$$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_i)-2\,\gamma\sqrt{\kappa(\boldsymbol{x}_i,\boldsymbol{x}_i)\,\kappa(\boldsymbol{x}_j,\boldsymbol{x}_j)}+\kappa(\boldsymbol{x}_j,\boldsymbol{x}_j).$$

Therefore, to complete the proof, it is sufficient to show that this expression is always strictly positive. Indeed, it is a quadratic polynomial of the form $u^2-2\gamma uv+v^2$ where $u=\sqrt{\kappa(\boldsymbol{x}_i,\boldsymbol{x}_i)}$ and $v=\sqrt{\kappa(\boldsymbol{x}_j,\boldsymbol{x}_j)}$ (this form is valid since $\kappa(\boldsymbol{x},\boldsymbol{x})=\|\kappa(\boldsymbol{x},\cdot)\|_{\mathbb{H}}^2>0$ for any $\boldsymbol{x}\in\mathbb{X}$). Considering the roots of this quadratic polynomial with respect to u, its discriminant is $4\,\kappa(\boldsymbol{x}_j,\boldsymbol{x}_j)(\gamma^2-1)$, which is strictly negative since $\gamma\in[\,0\,;1\,[$ and $\kappa(\boldsymbol{x}_j,\boldsymbol{x}_j)$ cannot be zero. Therefore, the polynomial has no real roots, and it is strictly positive.

Finally, for any γ-Babel dictionary, we have

$$\begin{aligned}\min_{\substack{i,j=1\cdots n\\ i\neq j}}\|\kappa(\boldsymbol{x}_i,\cdot)-\kappa(\boldsymbol{x}_j,\cdot)\|_{\mathbb{H}}^2 &= \min_{\substack{i,j=1\cdots n\\ i\neq j}}\kappa(\boldsymbol{x}_i,\boldsymbol{x}_i)-2\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)+\kappa(\boldsymbol{x}_j,\boldsymbol{x}_j)\\ &\geq 2r^2-2\max_{\substack{i,j=1\cdots n\\ i\neq j}}|\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)|,\\ &\geq 2r^2-2\,\gamma,\end{aligned}$$

which is strictly positive when $\gamma<r^2$.

Entropy in the RKHS

The entropy in the feature space provides a measure of diversity of the atoms distribution. In the following, we show that the entropy estimated in the feature space is lower-bounded, with a bound expressed in terms of a diversity measure.

We denote by $P_{\mathrm{H}}(\boldsymbol{x})$ the distribution associated to the kernel functions in the feature space, namely by definition $P_{\mathrm{H}}(\boldsymbol{x}) = P(\kappa(\boldsymbol{x}, \cdot))$. The entropy in the RKHS is given by expression (13) where $P(\boldsymbol{x})$ is substituted with $P_{\mathrm{H}}(\boldsymbol{x})$, yielding[3]

$$\frac{1}{1-\alpha} \log \int_{\mathbb{X}} \big(P_{\mathrm{H}}(\boldsymbol{x})\big)^{\alpha} \, d\boldsymbol{x}. \tag{16}$$

By approximating the integral in this expression with the set $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$, we get

$$\frac{1}{1-\alpha} \log \sum_{j=1}^{n} \big(P_{\mathrm{H}}(\boldsymbol{x}_j)\big)^{\alpha}.$$

The distribution $P_{\mathrm{H}}(\cdot)$ is estimated with the Parzen window estimator. The use of a radial function $w(\cdot)$ defined in the feature space $\mathbb{H}$ yields

$$\widehat{P}_{\mathrm{H}}(\boldsymbol{x}) = \frac{1}{n} \sum_{j=1}^{n} w(\|\kappa(\boldsymbol{x}, \cdot) - \kappa(\boldsymbol{x}_j, \cdot)\|_{\mathbb{H}}).$$

Examples of radial functions are — up to a scaling factor to ensure the integration to one — the gaussian, the radial-based exponential and the inverse mutliquadratic kernels, given in Table 1 and applied here in the feature space. Radial kernels are monotonically decreasing in the distance, namely $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$ grows when $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|$ is decreasing. This statement results from the following lemma; See also [8, Proposition 5].

Lemma 6. *Any kernel κ, of the form $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = g(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2)$ with $g\colon (0, \infty) \to \mathbb{R}$, is positive definite if $g(\cdot)$ is completely monotonic, namely its k-th derivative $g^{(k)}$ satisfies $(-1)^k g^{(k)}(r) \geq 0$ for any $r, k \geq 0$.*

Theorem 7. *Consider an overcomplete kernel dictionary with a lower bound ϵ on its distance measure, or any bounded diversity measure as given in Theorem 5. A Parzen window estimator, estimated over the dictionary atoms in the feature space, is upper-bounded by $w(\epsilon)$, where $w(\cdot)$ is the used window function.*

Proof. The proof is follows from

$$\widehat{P}_{\mathrm{H}}(\boldsymbol{x}) = \frac{1}{n} \sum_{j=1}^{n} w(\|\kappa(\boldsymbol{x}, \cdot) - \kappa(\boldsymbol{x}_j, \cdot)\|_{\mathbb{H}}) < \frac{1}{n} \sum_{j=1}^{n} w(\epsilon) = w(\epsilon),$$

where the inequality is due to the monotonically decreasing property of the window function w and Theorem 5.

This theorem is the main building block of the following corollary that provides lower bounds on the entropy, with the Shannon entropy and generalizing to the Rényi entropy for any order $\alpha > 1$.

[3] The expectation in a RKHS, as in (16), was previously investigated in the literature. The notion of embedding a Borel probability measure P, defined on the topological space $\mathbb{X}$, into a RKHS $\mathbb{H}$ was studied in detail in [43], with $\int_{\mathbb{X}} \kappa(\boldsymbol{x}, \cdot) \, dP(\boldsymbol{x})$. For an algorithmic use, see [32] for a one-class classifier.

Corollary 8. *Consider an overcomplete kernel dictionary with a lower bound ϵ on its distance measure, or any bounded diversity measure as given in Theorem 5. The Shannon entropy and the generalized Rényi entropy for any order $\alpha > 1$ are lower bounded by $-n\, w(\epsilon) \log w(\epsilon)$ and $\frac{1}{1-\alpha} \log \big(n \left(w(\epsilon)\right)^\alpha\big)$, respectively, where $w(\cdot)$ is the used window function.*

Proof. From Theorem 7, we have $\widehat{P}_{\text{H}}(\boldsymbol{x}) < w(\epsilon)$ for any window function $w(\cdot)$. This yields for the Shannon entropy:

$$-\sum_{j=1}^{n} \widehat{P}_{\text{H}}(\boldsymbol{x}_j) \log \widehat{P}_{\text{H}}(\boldsymbol{x}_j) > -n\, w(\epsilon) \log w(\epsilon).$$

More generally, the Rényi entropy for any order α is estimated by

$$\frac{1}{1-\alpha} \log \sum_{j=1}^{n} \big(\widehat{P}_{\text{H}}(\boldsymbol{x}_j)\big)^\alpha > \frac{1}{1-\alpha} \log \Big(n \left(w(\epsilon)\right)^\alpha\Big),$$

where we have used Theorem 7 and $\alpha > 1$.

These results illustrate how the atoms of an overcomplete dictionary are uniformly spread in the feature space.

Final remarks

This paper provided a framework to examine linear and kernel dictionaries with the notion of entropy from information theory. By examining different diversity measures, we showed that overcomplete dictionaries have lower bounds on the entropy. While various definitions were explored here, these results open the door to bridging the gap between information theory and diversity measures for the analysis and synthesis of overcomplete dictionaries, in both input and feature spaces. As of futur works, we are studying connections to the entropy component analysis [25], in order to provide a thorough examination and develop an online learning approach.

The conducted analysis, illustrated here within the framework of kernel-based learning algorithms, can be easily extended to other machines such as gaussian processes and neural networks. It is worth noting that this work does not devise any particular diversity measure for quantifying overcomplete dictionaries, in the same spirit as our recent work [21, 20].

References

[1] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *Signal Processing, IEEE Transactions on*, 54(11):4311–4322, 2006.

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

[3] Guang bin Huang, P. Saratch, Senior Member, and Narashiman Sundararajan. A generalized growing and pruning rbf (ggap-rbf) neural network for function approximation. *IEEE Transactions on Neural Networks*, 16:57–67, 2005.

[4] Jon Cartwright. Roll over, boltzmann. *Physics World*, May 2014.

[5] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley Series in Telecommunications and Signal Processing. Wiley-Interscience, 2nd edition edition, 2006.

[6] Lehel Csató and Manfred Opper. Sparse representation for gaussian process models. In *Advances in Neural Information Processing Systems 13*, pages 444–450. MIT Press, 2001.

[7] Lehel Csató and Manfred Opper. Sparse online gaussian processes. *Neural Computation*, 14:641–668, 2002.

[8] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39:1–49, 2002.

[9] D. L. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *Proceedings - National Academy of Sciences (PNAS)*, 100(5):2197–2202, March 2003.

[10] D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Information Theory*, 47(7):2845–2862, March 2001.

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

[12] K. Engan, S.O. 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*, volume 5, pages 2443–2446 vol.5, 1999.

[13] Y. Engel, S. Mannor, and R. Meir. The kernel recursive least squares algorithm. *IEEE Trans. Signal Processing*, 52(8):2275–2285, 2004.

[14] Haijin Fan, Qing Song, and Sumit Bam Shrestha. Online learning with kernel regularized least mean square algorithms. *Knowledge-Based Systems*, 59:21 – 32, 2014.

[15] 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)*, pages 243–252, Philadelphia, PA, USA, 2003. Society for Industrial and Applied Mathematics.

[16] 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)*, volume 1, pages 37–40, Barcelona, Spain, Sept. 2003.

[17] M. Girolami. Orthogonal series density estimation and the kernel eigenvalue problem. *Neural Computation*, 14:669–688, 2002.

[18] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, second edition edition, 2009.

[19] Paul Honeine. Online kernel principal component analysis: a reduced-order model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1814–1826, September 2012.

[20] Paul Honeine. Analyzing sparse dictionaries for online learning with kernels. *IEEE Transactions on Signal Processing*, 63(23):6343–6353, December 2015.

[21] Paul Honeine. Approximation errors of online sparsification criteria. *IEEE Transactions on Signal Processing*, 63(17):4700–4709, September 2015.

[22] Paul Honeine and Cdric Richard. Preimage problem in kernel-based machine learning. *IEEE Signal Processing Magazine*, 28(2):77–88, March 2011.

[23] Paul Honeine, Cdric Richard, and Jos C. M. Bermudez. On-line nonlinear sparse approximation of functions. In *Proc. IEEE International Symposium on Information Theory*, pages 956–960, Nice, France, June 2007.

[24] Paul Honeine and Fei Zhu. Eviter la maldiction de pr-image : application la factorisation en matrices non ngatives noyaux. In *Actes du 25-me Colloque GRETSI sur le Traitement du Signal et des Images*, Lyon, France, September 2015.

[25] R. Jenssen. Kernel entropy component analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):847–860, 2010.

[26] I.T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, NY, USA, 1986.

[27] Tobias Jung and Daniel Polani. Sequential learning with ls-svm for large-scale data sets. In *Proceedings of the 16th International Conference on Artificial Neural Networks - Volume Part II*, ICANN'06, pages 381–390, Berlin, Heidelberg, 2006. Springer-Verlag.

[28] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 689–696, New York, NY, USA, 2009. ACM.

[29] S. Mallat. *A wavelet tour of signal processing*. Academic Press, 1998.

[30] S. Mallat and Z. Zhang. Matching pursuit with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41:3397–3415, 1993.

[31] Patric Nader, Paul Honeine, and Pierre Beauseroy. Online one-class classification for intrusion detection based on the mahalanobis distance. In *Proc. 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 1–6, Bruges, Belgium, 22–24 April 2015.

[32] Zineb Noumir, Paul Honeine, and Cdric Richard. On simple one-class classification methods. In *Proc. IEEE International Symposium on Information Theory*, pages 2022–2026, MIT, Cambridge (MA), USA, 1–6 July 2012.

[33] Zineb Noumir, Paul Honeine, and Cdric Richard. One-class machines based on the coherence criterion. In *Proc. IEEE workshop on Statistical Signal Processing*, pages 600–603, Ann Arbor, Michigan, USA, 5–8 August 2012.

[34] Zineb Noumir, Paul Honeine, and Cdric Richard. Online one-class machines based on the coherence criterion. In *Proc. 20th European Conference on Signal Processing*, pages 664–668, Bucharest, Romania, 27–31 August 2012.

[35] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

[36] John Platt. A resource-allocating network for function interpolation. *Neural Comput.*, 3(2):213–225, June 1991.

[37] Puskal P. Pokharel, Weifeng Liu, and Jose C. Principe. Kernel least mean square algorithm with constrained growth. *Signal Processing*, 89(3):257 – 265, 2009.

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

[39] Cdric Richard, Jos C. M. Bermudez, and Paul Honeine. Online prediction of time series data with kernels. *IEEE Transactions on Signal Processing*, 57(3):1058–1067, March 2009.

[40] Chafic Said, Rgis Lengell, Paul Honeine, and Roger Achkar. Online kernel adaptive algorithms with dictionary adaptation for mimo models. *IEEE Signal Processing Letters*, 20(5):535–538, May 2013.

[41] Chafic Said, Rgis Lengell, Paul Honeine, Cdric Richard, and Roger Achkar. Nonlinear adaptive filtering using kernel-based algorithms with dictionary adaptation. *International Journal of Adaptive Control and Signal Processing*, 29(11):1391–1410, November 2015.

[42] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, UK, 2004.

[43] Bharath Kumar Sriperumbudur Vangeepuram. *Reproducing Kernel Space Embeddings and Metrics on Probability Measures*. PhD thesis, Electrical Engineering (Signal and Image Processing), University of California at San Diego, La Jolla, CA, USA, 2010. AAI3432386.

[44] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific Pub. Co., Singapore, 2002.

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

[46] J.A. Tropp. Just relax: convex programming methods for identifying sparse signals in noise. *Information Theory, IEEE Transactions on*, 52(3):1030–1051, March 2006.

[47] C. Tsallis. Nonadditive entropy: The concept and its use. *European Physical Journal A*, 40:257–266, June 2009.

[48] V.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, NY, USA, 1995.

[49] Najdan Vuković and Zoran Miljković. A growing and pruning sequential learning algorithm of hyper basis function neural network for function approximation. *Neural Netw.*, 46:210–226, Oct. 2013.

[50] Fei Zhu and Paul Honeine. Online nonnegative matrix factorization based on kernel machines. In *Proc. 23rd European Conference on Signal Processing*, pages 2381–2385, Nice, France, 31 Aug.–4 Sept. 2015.

[51] Fei Zhu and Paul Honeine. Pareto front of bi-objective kernel-based nonnegative matrix factorization. In *Proc. 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 585–590, Bruges, Belgium, 22–24 April 2015.

[52] Fei Zhu and Paul Honeine. Bi-objective nonnegative matrix factorization: Linear versus kernel-based models. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–11, in press 2016.

[53] Fei Zhu and Paul Honeine. Online kernel nonnegative matrix factorization. *Signal Processing*, (in press), 2016.

[54] Fei Zhu, Paul Honeine, and Maya Kallas. Kernel non-negative matrix factorization without the pre-image problem. In *Proc. 24th IEEE workshop on Machine Learning for Signal Processing*, pages 1–6, Reims, France, 21–24 September 2014.

[55] Fei Zhu, Paul Honeine, and Maya Kallas. Kernel nonnegative matrix factorization without the curse of the pre-image. Technical Report arXiv:1407.4420v1, ArXiv, 2015.

[56] Fei Zhu, Paul Honeine, and Maya Kallas. Kernel nonnegative matrix factorization without the curse of the pre-image — application to unmixing hyperspectral images. *http://arxiv.org/abs/1407.4420*, pages 1–13, 2016.