Classification of unions of subspaces with sparse representations

Alhussein Fawzi and Pascal Frossard
École Polytechnique Fédérale de Lausanne (EPFL)
Signal Processing Laboratory (LTS4)
E-mail: alhussein.fawzi@epfl.ch, pascal.frossard@epfl.ch.

***Abstract*—We propose a preliminary investigation on the benefits and limitations of classifiers based on sparse representations. We specifically focus on the union of subspaces data model and examine binary classifiers built on a sparse non linear mapping (in a redundant dictionary) followed by a linear classifier. We study two common sparse non linear mappings (namely ℓ_0 and ℓ_1) and show that, in both cases, there exists a finite dictionary such that the classifier discriminates the two classes correctly. This result paves the way towards a better understanding of the increasingly popular classifiers based on sparse representations, and provides initial insights on appropriate dictionary design.**

I. Introduction

The past decade has witnessed an explosion of high-dimensional data. Leveraging the geometric properties of data has therefore become crucial in order to develop efficient data processing techniques. In many practical problems, the data points lie on low dimensional subspaces instead of being uniformly distributed across the ambient space. Sparse representations in redundant dictionaries have attracted much attention recently, as they permit to build representations that adapt to the data geometry. Sparse representations have for example led to state-of-the art results in many recognition and classification tasks [1], [2], [3], [4]. In this context, sparse representations are viewed as feature sets that are fed to a classifier, and sparse coding is seen as a nonlinear feature extraction mapping. Even though the relevance of sparse coding in classification is now well-established from an experimental point of view, there are, to the best of our knowledge, no theoretical results that explain the benefits of sparse representations in classification. Interestingly, the authors in [3], [5], [6] conjecture that sparse codes are advantageous for classification since features are more likely to be linearly separable in the high dimensional feature space, but no theoretical results are shown.

In this paper, we propose to tackle the problem of classification in *Unions of Subspaces* (UoS) with classifiers based on sparse representations. The UoS model, where datapoints belong to a union of usually low-dimensional subspaces, has been shown to be applicable in many computer vision problems ([7] and references therein). For example, face images of one subject under different illumination conditions lie approximately on a low dimensional subspace in the ambient space [1]. Face images of different subjects therefore lie on a *union* of low dimensional subspaces. The UoS model has also received attention in areas such as biomedical data processing [8] and system control theory [9]. We assume in this paper that datapoints of class 1 and class 2 lie on two different *unions* of *unknown* low dimensional subspaces. The goal is then to correctly classify a new unlabeled datapoint (or equivalently, determine to which union of subspaces the datapoint belongs). Using the example of *face* images classification, this could correspond to separating between *male* and *female* face images, as male and female face images lie on different unions of low dimensional subspaces. To tackle this classification problem, we consider a simple classification architecture where a non linear sparse coding mapping is followed by a linear classifier. We show the existence of a classifier following this architecture that succeeds in classifying the unions of subspaces. Specifically, we show the existence of a redundant dictionary for which the images of the two unions of subspaces by the sparse representation mapping are linearly separable in the feature space. This result is valid even when the angle between subspaces becomes *arbitrarily small*, as long as the subspaces intersect only at the origin.

The paper is organized as follows. In Section 2, we formally define the problem considered in this paper. We analyze theoretically the introduced classification architecture for the ℓ_0 and ℓ_1 case respectively in Section 3 and 4. We then review related work in Section 5 and finally conclude with open questions in Section 6.

II. Problem formulation

Let $\{\mathcal{S}_i\}_{i=1}^{L_\mathcal{S}}$ and $\{\mathcal{T}_i\}_{i=1}^{L_\mathcal{T}}$ denote two sets of subspaces in $\mathbb{R}^n$ that define respectively the classes 1 and 2. In other words, any point $\mathbf{x} \in \cup_{i=1}^{L_\mathcal{S}} \mathcal{S}_i$ belongs to class 1 whereas any point $\mathbf{x} \in \cup_{i=1}^{L_\mathcal{T}} \mathcal{T}_i$ belongs to class 2. We assume that $\mathcal{S}_i \cap \mathcal{T}_j = \{0\}$ for all i, j. The classification problem consists in finding a mapping $\mathcal{C}$ that verifies:

$$\begin{aligned}&\mathcal{C} : \cup_{i=1}^{L_\mathcal{S}} \mathcal{S}_i \cup_{i=1}^{L_\mathcal{T}} \mathcal{T}_i \longrightarrow \{1, 2\}\\ &\mathcal{C}\left(\cup_{i=1}^{L_\mathcal{S}} \mathcal{S}_i\right) = 1,\\ &\mathcal{C}\left(\cup_{i=1}^{L_\mathcal{T}} \mathcal{T}_i\right) = 2.\end{aligned}$$

If $\mathcal{C}$ verifies the above conditions, we say that $\mathcal{C}$ succeeds in separating $\cup_{i=1}^{L_\mathcal{S}} \mathcal{S}_i$ and $\cup_{i=1}^{L_\mathcal{T}} \mathcal{T}_i$. When the subspaces are known, the construction of such a classifier is trivial. We study in this paper a classification architecture based on sparse representations that *does not require the knowledge of the subspaces*. Inspired by kernel methods for non linear classification, we focus on a simple classification architecture based on a sparse representation nonlinear mapping followed by a linear classifier. It is important to note that, while in kernel methods the non linear feature mapping is defined implicitly by a so called kernel function, the sparse coding mapping considered here is explicitly computed. In particular, for a fixed

dictionary $\mathbf{D} \in \mathbb{R}^{n \times N}$ and $p \in \{0, 1\}$, we define the ℓ_p sparse coding mapping by

$$f_{\ell_p}(\mathbf{x}|\mathbf{D}) = \arg\min_{\mathbf{c} \in \mathbb{R}^N} \|\mathbf{c}\|_p \text{ subject to } \mathbf{x} = \mathbf{D}\mathbf{c}.$$

While the ℓ_0 "norm" measures the sparsity of a vector with the number of nonzero elements, the ℓ_1 norm is a tractable convex relaxation. In this work, we further constrain the coefficients of the sparse codes to be nonnegative. That is, we augment the dictionary $\mathbf{D}$ with the negative of each atom and add a nonnegative constraint to the coefficients. The nonnegative ℓ_p sparse coding mapping is therefore defined as follows:

$$f^+_{\ell_p}(\mathbf{x}|\mathbf{D}) = \arg\min_{\mathbf{c} \in \mathbb{R}^{2N}} \|\mathbf{c}\|_p \text{ subject to } \mathbf{x} = [\mathbf{D}, -\mathbf{D}]\mathbf{c} \text{ and } \mathbf{c} \geq 0.$$

To illustrate the benefits of sparse representation mappings $f_{\ell_p}(\cdot|\mathbf{D})$ and $f^+_{\ell_p}(\cdot|\mathbf{D})$ in the classification of data lying on unions of subspaces, let us first consider the two-dimensional toy example of Fig. 1 where datapoints of class 1 live on subspaces $\mathcal{S}_1 \cup \mathcal{S}_2$, whereas datapoints of class 2 live on subspace $\mathcal{T}_1$. Clearly, the two classes cannot be separated by a line in $\mathbb{R}^2$. However, the feature vectors (in $\mathbb{R}^3$) obtained by applying $f_{\ell_p}(\cdot|\mathbf{D})$ (for $p = 0$ or 1) with $\mathbf{D} = [\mathbf{d_1}|\mathbf{d_2}|\mathbf{d_3}]$ are separable by a plane. We stress here that the choice of the dictionary is crucial: another choice of the dictionary may not have led to linearly separable features. The performance of a classifier based on sparse representations therefore strongly depends on the dictionary. When datapoints are not constrained to lie in the positive quadrant, the features obtained with $f_{\ell_p}(\cdot|\mathbf{D})$ are no more linearly separable. Consequently, $f_{\ell_p}(\mathcal{S}_1 \cup \mathcal{S}_2|\mathbf{D})$ and $f_{\ell_p}(\mathcal{T}_1|\mathbf{D})$ are *not* linearly separable. On the other hand, it is easy to see that $f^+_{\ell_p}(\mathcal{S}_1 \cup \mathcal{S}_2|\mathbf{D})$ and $f^+_{\ell_p}(\mathcal{T}_1|\mathbf{D})$ are separable by a hyperplane. This motivates the use of nonnegative sparse coding dictionary over $f_{\ell_p}(\cdot|\mathbf{D})$. On an empirical level, Coates and Ng [10] have observed that replicating the dictionary and constraining the coefficients to be nonnegative induce a significant gain in classification performance.

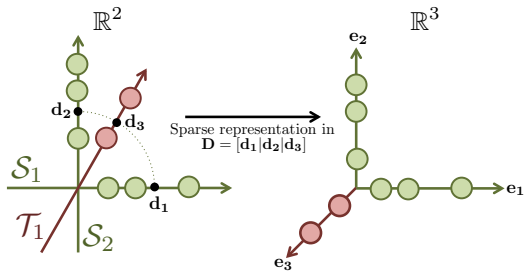

Fig. 1. Toy example in $\mathbb{R}^2$ that illustrates the benefits of the non linear sparse coding mapping for classification.

We therefore focus on the set of classifiers $\mathscr{C}$ built by applying a non negative sparse coding followed by a linear classifier. Specifically, for $\mathbf{D} \in \mathbb{R}^{n \times N}$ and $(\mathbf{w}, b) \in \mathbb{R}^N \times \mathbb{R}$, we denote by $\mathcal{C}_{(\mathbf{D},\mathbf{w},b)} \in \mathscr{C}$ the classifier defined using the following two-step procedure:

1) **Feature extraction:** For an input test point $\mathbf{x} \in \mathbb{R}^n$, compute $f^+_{\ell_p}(\mathbf{x}|\mathbf{D})$.
2) **Linear classification:** If $\mathbf{w}^{\mathbf{T}} f^+_{\ell_p}(\mathbf{x}|\mathbf{D}) \leq b$, then assign label 1 to x. Otherwise, assign label 2.

We examine in this paper the relevance of the set of classifiers $\mathscr{C}$ in separating unions of subspaces. Specifically, we study the *existence* of a classifier $\mathcal{C}_{(\mathbf{D},\mathbf{w},b)} \in \mathscr{C}$ that *separates* $\cup_{i=1}^{L_\mathcal{S}} \mathcal{S}_i$ and $\cup_{i=1}^{L_\mathcal{T}} \mathcal{T}_i$.

Problem 1. *Let $\{\mathcal{S}_i\}_{i=1}^{L_\mathcal{S}}$ and $\{\mathcal{T}_i\}_{i=1}^{L_\mathcal{T}}$ be two sets of subspaces such that $\mathcal{S}_i \cap \mathcal{T}_j = \{0\}$ for all i, j. Does there exist a finite dictionary $\mathbf{D}$ and a linear classifier $(\mathbf{w}, b)$ such that $\mathcal{C}_{(\mathbf{D},\mathbf{w},b)} \in \mathscr{C}$ separates the two unions of subspaces?*

Clearly, the above problem formulation is equivalent to asking whether there exists a finite dictionary $\mathbf{D}$ such that $f^+_{\ell_p}\left(\cup_{i=1}^{L_\mathcal{S}} \mathcal{S}_i|\mathbf{D}\right)$ and $f^+_{\ell_p}\left(\cup_{i=1}^{L_\mathcal{T}} \mathcal{T}_i|\mathbf{D}\right)$ are linearly separable. Problem 1 involves studying the notion of linear separability between subspaces that might be difficult to study theoretically. Hence, we focus instead on a sufficient condition that involves the notion of ***D*-subspace detection property** defined as follows:

Property 1. *(**D**-Subspace detection property (**D**-SDP)) Let $\{\mathcal{W}_i\}_{i=1}^L$ be a set of subspaces and $\mathbf{D}$ be a dictionary where the atoms live in $\cup_{i=1}^L \mathcal{W}_i$. We say that $\{\mathcal{W}_i\}_{i=1}^L$ satisfies the $\mathbf{D}$-subspace detection property if for any $\mathbf{x} \in \mathcal{W}_i$, the sparsest representation of $\mathbf{x}$ (that is, $f_{\ell_p}(\mathbf{x}|\mathbf{D})$) activates only atoms in the subspace $\mathcal{W}_i$.*

We are now ready to define Problem 2 as follows:

Problem 2. *Let $\{\mathcal{W}_i\}_{i=1}^L$ be a set of subspaces in $\mathbb{R}^n$. Does there exist a finite dictionary $\mathbf{D}$ such that the $\mathbf{D}$-subspace detection property is satisfied for subspaces $\{\mathcal{W}_i\}_{i=1}^L$?*

Clearly, if the answer to Problem 2 is "yes" for $\{\mathcal{W}_i\}_{i=1}^L = \{\mathcal{S}_i\}_{i=1}^{L_\mathcal{S}} \cup \{\mathcal{T}_i\}_{i=1}^{\mathcal{L}_\mathcal{T}}$, then so is the answer to Problem 1. Indeed, if there exists a dictionary $\mathbf{D}$ such that the $\mathbf{D}$-SDP is satisfied, the two unions of subspaces are linearly separable as it suffices to take a normal vector $\mathbf{w}$ with positive entries for atoms in $\mathcal{S}_i$ and negative entries otherwise, and a bias $b = 0$. This hyperplane separates the features of both classes since subspaces $\{\mathcal{S}_i\}_i$ and $\{\mathcal{T}_i\}_i$ are assumed to be disjoint. In the rest of this paper, we therefore focus on the analysis of Problem 2 for two different sparse coding mappings.

III. The ℓ_0 sparse coding mapping

The goal of this section is to address Problem 2 when $p = 0$. We define $\mathbf{D}$ to be a dictionary made up by concatenating arbitrary basis of each subspace. That is, $\mathbf{D} = [\mathcal{B}_\mathbf{1}| \ldots |\mathcal{B}_\mathbf{L}]$ where $\mathcal{B}_i$ is an arbitrary basis of subspace $\mathcal{W}_i$. We then consider the following probability distribution defined on $\cup_{i=1}^L \mathcal{W}_i$:

$$\forall i \in \{1, \ldots, L\}, \mathbb{P}(\mathbf{x} \in \mathcal{W}_i) = \frac{1}{L}$$
$$\mathbb{P}(\mathbf{x}|\mathbf{x} \in \mathcal{W}_i) = \mathcal{U}_{\mathcal{W}_i},$$

where $\mathcal{U}_{\mathcal{W}_i}$ denotes the uniform distribution over the unit sphere of $\mathcal{W}_i$. In other words, the choice of the subspace is equiprobable, and the distribution is uniform on the unit sphere of the subspace. We have the following result:

Theorem 1. *The $\mathbf{D}$-subspace detection property holds almost everywhere. That is, when $\mathbf{x}$ is chosen according to the*

above mentioned distribution, the ℓ_0 sparsest representation of $\mathbf{x}$ activates only atoms in the same subspace as $\mathbf{x}$ with probability 1.

Proof: Remember that $\mathbf{D}$ is constructed by appending arbitrary bases of each subspace. For any $\mathbf{x}$, we consider the error event $E(\mathbf{x})$:

"$f_{\ell_0}(\mathbf{x}|\mathbf{D})$ activates at least one atom *not* in the subspace of $\mathbf{x}$"

The error probability is therefore:

$$\begin{aligned}&\mathbb{P}_{\mathbf{x}}\left(E(\mathbf{x})\right)\\&=\frac{1}{L}\sum_{i=1}^{L}\mathbb{P}_{\mathbf{x}}\left(E(\mathbf{x})|\mathbf{x}\in\mathcal{W}_i\right).\end{aligned}\tag{1}$$

Let $i \in \{1,\dots,L\}$, and denote by δ_i the dimension of subspace $\mathcal{W}_i$. We define $\mathcal{A}_i$ to be the set of subsets of $\mathbf{D}$ of cardinality at most δ_i and that contain at least one atom *not* in $\mathcal{W}_i$. That is, we have

$$\mathcal{A}_i = \{\tilde{\mathbf{D}} \subset \mathbf{D} : |\tilde{\mathbf{D}}| \leq \delta_i \text{ and } \exists \tilde{\mathbf{d}} \in \tilde{\mathbf{D}}, \tilde{\mathbf{d}} \notin \mathcal{W}_i\}.$$

We have:

$$\begin{aligned}&\mathbb{P}_{\mathbf{x}\sim\mathcal{U}_{\mathcal{W}_i}}(E(\mathbf{x}))\\&\leq \mathbb{P}_{\mathbf{x}\sim\mathcal{U}_{\mathcal{W}_i}}\Big(\bigcup_{\tilde{\mathbf{D}}\in\mathcal{A}_i}\mathbf{x}\in\text{span}(\tilde{\mathbf{D}})\Big)\\&\leq\sum_{\tilde{\mathbf{D}}\in\mathcal{A}_i}\mathbb{P}_{\mathbf{x}\sim\mathcal{U}_{\mathcal{W}_i}}(\mathbf{x}\in\text{span}(\tilde{\mathbf{D}}))\\&=\sum_{\tilde{\mathbf{D}}\in\mathcal{A}_i}\mathbb{P}_{\mathbf{x}\sim\mathcal{U}_{\mathcal{W}_i}}(\mathbf{x}\in\text{span}(\tilde{\mathbf{D}})\cap\mathcal{W}_i).\end{aligned}\tag{2}$$

Note moreover that, for $\tilde{\mathbf{D}} \in \mathcal{A}_i$, we have $\text{span}(\tilde{\mathbf{D}}) \neq \mathcal{W}_i$. Therefore, $\dim(\mathcal{W}_i \cap \text{span}(\tilde{\mathbf{D}})) \leq \delta_i - 1$. We obtain

$$\mathbb{P}_{\mathbf{x}\sim\mathcal{U}_{\mathcal{W}_i}}(\mathbf{x}\in\text{span}(\tilde{\mathbf{D}})\cap\mathcal{W}_i)=0.$$

By injecting this equality into the upper bound on the error probability in Eq. (2), we conclude that

$$\mathbb{P}_{\mathbf{x}\sim\mathcal{U}_{\mathcal{W}_i}}(E(\mathbf{x}))=0.$$

Eq. (1) then concludes the proof of Theorem 1. ∎

Theorem 1 outlines an important property of sparse representations. That is, the sparsest representation of a point living in a particular subspace tends to activate atoms living in the same subspace, when we consider a simple dictionary made up of the union of different subspace bases. We believe that this result partially explains the discriminative power of sparse representations and therefore the empirical success of sparsity-based classifiers.

We further remark that the proof of Theorem 1 uses the knowledge of the subspaces $\mathcal{W}_i$ in order to construct the dictionary (i.e., we need to have a basis per subspace). Theorem 1 is an *existence* result and has no intention of providing a practical way to construct the dictionary (as the subspaces are unknown). However, we note that sampling uniformly at random δ_i points per subspace gives us a basis with probability 1. Therefore, as long as we have access to δ_i points drawn uniformly at random for each subspace, this gives a way to construct a dictionary that satisfies the D-SDP with probability 1.

Finally, it is known that the ℓ_0 sparse representation mapping is NP-hard to compute (for a particular test point $\mathbf{x}$). Moreover, this mapping is not stable, such that small perturbations of the input vector might induce significant changes in the sparsest representation. Both reasons motivate the study of another sparse coding scheme, namely the mapping $f_{\ell_1}(\cdot|\mathbf{D})$.

IV. The ℓ_1 sparse coding mapping

We now study Problem 2 in the case of ℓ_1 sparse coding. Unlike the ℓ_0 mapping, the ℓ_1 sparse representation mapping is tractable to compute. We assume in this section that *all* the subspaces $\{\mathcal{W}_i\}_{i=1}^{L} = \{\mathcal{S}_i\}_{i=1}^{L_S} \cup \{\mathcal{T}_i\}_{i=1}^{L_T}$ are disjoint.

A. Counterexample

We first show with a simple counterexample in $\mathbb{R}^3$ (Fig. 2) that the dictionary construction of Section III does not satisfy the $\mathbf{D}$-SDP in the ℓ_1 case. Specifically, we consider three subspaces $\{\mathcal{W}_i\}_{i=1}^{3}$ defined as follows:

$$\begin{aligned}\mathcal{W}_1 &= \text{span}(\mathbf{e_1}, \mathbf{e_2}),\\ \mathcal{W}_2 &= \text{span}(\mathbf{f_2}),\\ \mathcal{W}_3 &= \text{span}(\mathbf{f_3}),\end{aligned}$$

where $\mathbf{e_1}, \mathbf{e_2}, \mathbf{f_2}, \mathbf{f_3}$ have all a unit Euclidean norm. Clearly, $(\mathbf{e_1}, \mathbf{e_2})$, $\mathbf{f_2}$ and $\mathbf{f_3}$ define respectively a basis for subspaces $\mathcal{W}_1$, $\mathcal{W}_2$ and $\mathcal{W}_3$. We consider a dictionary $\mathbf{D}$ defined by concatenating those bases:

$$\mathbf{D} = [\mathbf{e_1}, \mathbf{e_2}, \mathbf{f_2}, \mathbf{f_3}].$$

It is not hard to see that for values of η that are small enough, the ℓ_1 sparsest representation of a point outside the cone defined by $(\mathbf{e_1}, \mathbf{e_2})$ involves $\mathbf{f_2}$ and $\mathbf{f_3}$. Unlike the ℓ_0 sparse coding case where the D-SDP would fail only for $\mathbf{x}$ that lies on the line $\text{span}(\Pi_{\mathcal{W}_1}(\mathbf{f_2}))$ (where $\Pi_{\mathcal{W}_1}$ defines the orthogonal projection operator onto the plane $\mathcal{W}_1$), the ℓ_1 sparse coding mapping fails to select the atoms of the right subspace for a set of nonzero measure, when η is small.

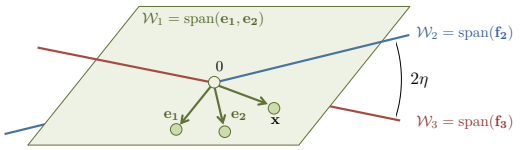

Fig. 2. Example where a set of non-zero measure is wrongly classified, when the ℓ_1 norm is used in the sparse coding step.

We now show in the next section that there exists a dictionary, which satisfies the $\mathbf{D}$-SDP when the classifier uses an ℓ_1 sparse coding non linear mapping.

B. Existence of a dictionary for ℓ_1 sparse coding mapping

The main message one can take from the counter example in Fig. 2 is that when the atoms are not well spread across all the subspace, the ℓ_1 norm fails to guarantee the subspace detection property. This was already observed in [11] and

[12]. In the following, we show the existence of a dictionary that is "sufficiently spread" in order to satisfy the D-SDP on all subspaces. More precisely, the dictionary we consider is constructed by concatenating two components:

- An ϵ-net [13] on the unit sphere of each subspace[1], where ϵ is carefully chosen depending on the principal angles between the subspaces.
- An orthonormal basis of each subspace. We add this component to the dictionary for purely technical reasons.

We let κ be the cosine of the maximum principal angle between the subspaces. In other words:

$$\kappa = \max_{\mathbf{x_i} \in \mathcal{W}_i} \max_{\mathbf{y_j} \in \mathcal{W}_j} \frac{|\langle \mathbf{x_i}, \mathbf{y_j} \rangle|}{\|\mathbf{x_i}\|_2 \|\mathbf{y_j}\|_2}.$$

Since the subspaces are disjoint, we have $\kappa < 1$. With the above dictionary construction $\mathbf{D}$, we have the following result:

Theorem 2. *The subspaces $\{\mathcal{W}_i\}_{i=1}^L$ satisfy the $\mathbf{D}$-subspace detection property. This dictionary contains at most N atoms, with*

$$N = \left(\left(1 + \frac{2(\sqrt{\delta_{max}} + 1)}{1 - \kappa} \right)^{\delta_{max}} + \delta_{\max} \right) L,$$

and δ_{max} being the maximum dimension of the subspaces.

Proof: Let $\epsilon = \frac{1-\kappa}{\sqrt{\delta_{\max}}+1}$. The following lemma featured in [13] gives an upper bound on the number of points needed for constructing an ϵ-net on the unit sphere, when the ambient space is δ-dimensional:

Lemma 1. *For any $\epsilon > 0$, there exists an ϵ-net $\mathcal{N}_\epsilon$ on the unit sphere $\mathbb{S}^{\delta-1}$ with size*

$$|\mathcal{N}_\epsilon| \leq \left(1 + \frac{2}{\epsilon} \right)^{\delta}.$$

Recall that our dictionary is obtained by concatenating an ϵ-net on the unit sphere of each subspace with an orthonormal basis of each subspace. Using Lemma 1, the size of our dictionary therefore satisfies

$$|\mathbf{D}| \leq \left(1 + \frac{2(\sqrt{\delta_{\max}} + 1)}{1 - \kappa} \right)^{\delta_{\max}} L + L\delta_{\max}.$$

We now prove that the proposed dictionary $\mathbf{D} = [\mathbf{d_1} | \dots | \mathbf{d_N}]$ satisfies D-SDP. Suppose, by contradiction, that the ℓ_1 sparsest representation of $\mathbf{x} \in \mathcal{W}_i$ in $\mathbf{D}$ activates at least one atom *not* in $\mathcal{W}_i$. We write le ℓ_1 sparsest representation of $\mathbf{x}$ as follows:

$$\mathbf{x} = \sum_{j=1}^{N} c_j \mathbf{d_j} = \sum_{j:\mathbf{d_j} \in \mathcal{W}_i} c_j \mathbf{d_j} + \sum_{j:\mathbf{d_j} \notin \mathcal{W}_i} c_j \mathbf{d_j}.$$

Since $\mathbf{x} \in \mathcal{W}_i$ and $\sum_{j:d_j \in \mathcal{W}_i} c_j \mathbf{d_j} \in \mathcal{W}_i$, we have $\mathbf{x}' \stackrel{def}{=} \mathbf{x} - \sum_{j:\mathbf{d_j} \in \mathcal{W}_i} c_j \mathbf{d_j} \in \mathcal{W}_i$. Moreover, it is not hard to see that $\sum_{j:\mathbf{d_j} \notin \mathcal{W}_i} c_j \mathbf{d_j}$ is the sparsest (in the ℓ_1 sense) representation of $\mathbf{x}'$ in the dictionary. By orthogonal projection of the equality $\mathbf{x}' = \sum_{j:\mathbf{d_j} \notin \mathcal{W}_i} c_j \mathbf{d_j}$ into $\mathcal{W}_i$, we have:

$$\mathbf{x}' = \sum_{j:\mathbf{d_j} \notin \mathcal{W}_i} c_j \Pi_{\mathcal{W}_i}(\mathbf{d_j}), \tag{3}$$

where $\Pi_{\mathcal{W}_i}$ is the orthogonal projection operator onto $\mathcal{W}_i$. We now exhibit a representation of $\mathbf{x}'$ in $\mathbf{D}$ whose ℓ_1 norm is strictly smaller than $\sum_{j:\mathbf{d_j} \notin \mathcal{W}_i} |c_j|$ and whose atoms all belong to $\mathcal{W}_i$. To do so, we represent each vector $\Pi_{\mathcal{W}_i}(\mathbf{d_j})$ in Eq.(3) with its *nearest neighbor atom* and a residual vector that accounts for the approximation error. Let $\mathbf{p_j} = \frac{\Pi_{\mathcal{W}_i}(\mathbf{d_j})}{\|\Pi_{\mathcal{W}_i}(\mathbf{d_j})\|_2}$. There exists a dictionary atom $\tilde{\mathbf{d}}_\mathbf{j}$ in $\mathcal{W}_i$ such that:

$$\left\| \mathbf{p_j} - \tilde{\mathbf{d}}_\mathbf{j} \right\|_2 \leq \epsilon, \tag{4}$$

due to the ϵ-net dictionary construction. Then, for each $\mathbf{d_j} \notin \mathcal{W}_i$, we compute the orthogonal projection of $\mathbf{p_j}$ onto $\tilde{\mathbf{d}}_\mathbf{j}$:

$$\mathbf{p_j} = \left\langle \mathbf{p_j}, \tilde{\mathbf{d}}_\mathbf{j} \right\rangle \tilde{\mathbf{d}}_\mathbf{j} + \tilde{\mathbf{d}}_\mathbf{j}^{\perp}. \tag{5}$$

Taking the norms of the previous equality, we have

$$\begin{aligned} \|\tilde{\mathbf{d}}_\mathbf{j}^{\perp}\|_2^2 &= 1 - \left\langle \mathbf{p_j}, \tilde{\mathbf{d}}_\mathbf{j} \right\rangle^2 \\ &\stackrel{(*)}{\leq} 1 - \left(1 - \frac{\epsilon^2}{2} \right)^2 \\ &\leq \epsilon^2 \end{aligned} \tag{6}$$

where the inequality (*) is due to Eq. (4). Let further $(\mathbf{e_k})_k$ be an orthonormal basis of $\mathcal{W}_i$ in the dictionary. We have $\mathbf{d_j^\perp} = \sum_k r_k^j \mathbf{e_k}$ for some vector $\mathbf{r^j}$ with $\|\mathbf{d_j^\perp}\|_2 = \|\mathbf{r^j}\|_2$. We therefore conclude that $\|\mathbf{r^j}\|_1 \leq \sqrt{\delta_{\max}} \|\mathbf{d_j}^{\perp}\|_2 = \sqrt{\delta_{\max}}\epsilon$.

We now upper bound $\|\Pi_{\mathcal{W}_i}(\mathbf{d_j})\|_2$. By noting that $\left\langle \Pi_{\mathcal{W}_i}(\mathbf{d_j}), \tilde{\mathbf{d}}_\mathbf{j} \right\rangle = \left\langle \mathbf{d_j}, \tilde{\mathbf{d}}_\mathbf{j} \right\rangle$ and rewriting Eq. (5), we have:

$$\Pi_{\mathcal{W}_i}(\mathbf{d_j}) = \left\langle \mathbf{d_j}, \tilde{\mathbf{d}}_\mathbf{j} \right\rangle \tilde{\mathbf{d}}_\mathbf{j} + \tilde{\mathbf{d}}_\mathbf{j}^{\perp} \|\Pi_{\mathcal{W}_i}(\mathbf{d_j})\|_2.$$

Therefore, by taking the norm of the previous equality, and making use of Eq. (6), we obtain:

$$\|\Pi_{\mathcal{W}_i}(\mathbf{d_j})\|_2 \leq \frac{\left\langle \mathbf{d_j}, \tilde{\mathbf{d}}_\mathbf{j} \right\rangle}{\sqrt{1 - \epsilon^2}} \leq \frac{\kappa}{\sqrt{1 - \epsilon^2}}. \tag{7}$$

We rewrite the representation in Eq. (3) using dictionary elements:

$$\begin{aligned} \mathbf{x}' &= \sum_{j:\mathbf{d_j} \notin \mathcal{W}_i} \|\Pi_{\mathcal{W}_i}(\mathbf{d_j})\|_2 c_j \mathbf{p_j} \\ &= \sum_{j:\mathbf{d_j} \notin \mathcal{W}_i} \|\Pi_{\mathcal{W}_i}(\mathbf{d_j})\|_2 c_j \left(\left\langle \mathbf{p_j}, \tilde{\mathbf{d}}_\mathbf{j} \right\rangle \tilde{\mathbf{d}}_\mathbf{j} + \tilde{\mathbf{d}}_\mathbf{j}^{\perp} \right) \\ &= \sum_{j:\mathbf{d_j} \notin \mathcal{W}_i} \|\Pi_{\mathcal{W}_i}(\mathbf{d_j})\|_2 c_j \left\langle \mathbf{p_j}, \tilde{\mathbf{d}}_\mathbf{j} \right\rangle \tilde{\mathbf{d}}_\mathbf{j} \\ &+ \sum_k \mathbf{e_k} \sum_{j:\mathbf{d_j} \notin \mathcal{W}_i} r_k^j \|\Pi_{\mathcal{W}_i}(\mathbf{d_j})\|_2 c_j. \end{aligned}$$

[1] We recall that an ϵ-net on the unit sphere $\mathbb{S}^{\delta-1}$ is a finite set, denoted by $\mathcal{N}_\epsilon$, that satisfies: $\forall \mathbf{x} \in \mathbb{S}^{\delta-1}, \exists \mathbf{d} \in \mathcal{N}_\epsilon$ such that $\|\mathbf{x} - \mathbf{d}\|_2 \leq \epsilon$.

The ℓ_1 norm of the above representation is therefore upper bounded by:

$$\begin{aligned}
&\sum_{j:\mathbf{d_j}\notin\mathcal{W}_i} \|\Pi_{\mathcal{W}_i}(\mathbf{d_j})\|_2 \left|c_j \left\langle \mathbf{p_j}, \tilde{\mathbf{d}}_\mathbf{j} \right\rangle\right| \\
&+ \sum_k \left| \sum_{j:\mathbf{d_j}\notin\mathcal{W}_i} r_k^j \|\Pi_{\mathcal{W}_i}(\mathbf{d_j})\|_2 c_j \right| \\
&\leq \frac{\kappa}{\sqrt{1-\epsilon^2}} \sum_{j:\mathbf{d_j}\notin\mathcal{W}_i} |c_j| + \sum_{j:\mathbf{d_j}\notin\mathcal{W}_i} |c_j| \|\mathbf{r^j}\|_1 \\
&= \left(\frac{\kappa}{\sqrt{1-\epsilon^2}} + \sqrt{\delta_{\max}}\epsilon \right) \sum_{j:\mathbf{d_j}\notin\mathcal{W}_i} |c_j|
\end{aligned}$$

Finally, it is easy to show that for $\epsilon = \frac{1-\kappa}{\sqrt{\delta_{\max}}+1}$, we have:

$$\frac{\kappa}{\sqrt{1-\epsilon^2}} + \sqrt{\delta_{\max}}\epsilon < 1,$$

hence the new representation of $\mathbf{x}'$ where atoms live exclusively in $\mathcal{W}_i$ has an ℓ_1 norm that is strictly smaller than $\sum_{j:\mathbf{d_j}\notin\mathcal{W}_i} |c_j|$. This contradicts our initial assumption and therefore concludes the proof of Theorem 2. ■

Theorem 2 shows that, if the dictionary is made of atoms that cover *enough* directions, the D-SDP holds. Unlike the ℓ_0 case, it is not sufficient to take a dictionary made of a union of bases, when ℓ_1 sparse coding is used. In this case, the number of atoms is exponential in the dimension of the subspaces (as the covering number on the sphere is exponential), if one wants to satisfy the sufficient D-SDP for proper linear classification of unions of subspaces.

V. Related work

In [12], Elhamifar and Vidal consider the problem of *subspace clustering*, where unlabeled data belongs to a union of subspaces and the goal is then to cluster points into their corresponding subspaces. The authors approached this problem through a sparse representation based method (*sparse subspace clustering*) that precisely use the property that sparse representations tend to activate atoms of the same subspace of the datapoint (similar to the $\mathbf{D}$-SDP). A theoretical analysis is conducted and shows roughly that, as the maximum principal angle between subspaces κ satisfies $\kappa < \frac{1}{\sqrt{\delta_{\max}}}$, the subspace detection property holds. Unfortunately, in many practical scenarios, this constraint does not hold. Soltanolkotabi and Candes in [11] improved this theoretical analysis, with a different notion of subspace detection property, where it is only asked that *the sparse representations of the specific points we wish to cluster activate other points in the same subspace*. This differs from our study, where we ask that *any point belonging to the union of subspaces activates atoms of the same subspace*. Note finally that in these works ([12], [11]), the authors assume that the datapoints are *fixed* and the goal is to find conditions on the subspaces and datapoints for which sparse subspace clustering works.

Our perspective through this paper is different. Motivated by a classification problem, we study whether we can *find* a dictionary for which the $\mathbf{D}$-SDP holds for *any given* unions of subspaces (with no particular assumption on the principal angle between subspaces[2]). We show that this is true for the ℓ_0 and ℓ_1 sparse representation mappings, as long as the subspaces are disjoint.

VI. Discussion and open questions

This paper represents what we believe to be one of the first investigations on the benefits and limitations of classifiers based on sparse representations. Many questions remain unanswered and will be the focus of future work. For example, do we really need an *exponential* number of atoms (with respect to the dimension) in order to satisfy D-SDP, in the ℓ_1 sparse coding case? Also, what is the behavior of *practical* dictionaries vis-a-vis the D-SDP, when the number of atoms is limited? Finally, we are developping discriminative dictionary learning methods that use the insights provided above to guarantee good classification accuracy.

References

[1] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[2] R. Raina, A. Battle, H. Lee, B. Packer, and A. Ng, "Self-taught learning: transfer learning from unlabeled data," in *Proceedings of the 24th International Conference on Machine Learning*, 2007, pp. 759–766.

[3] M. Ranzato, C. Poultney, S. Chopra, and Y. LeCun, "Efficient learning of sparse representations with an energy-based model," *Advances in neural information processing systems*, vol. 19, pp. 1137–1144, 2006.

[4] I. Ramirez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3501–3508.

[5] K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "Fast inference in sparse coding algorithms with applications to object recognition," *arXiv:1010.3467*, 2010.

[6] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1794–1801, 2009.

[7] René Vidal, "Subspace clustering," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 52–68, 2011.

[8] H-P. Kriegel, P. Kröger, and A. Zimek, "Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 3, no. 1, pp. 1, 2009.

[9] R. Vidal, S. Soatto, Y. Ma, and S. Sastry, "An algebraic geometric approach to the identification of a class of linear hybrid systems," in *IEEE Conference on Decision and Control*, 2003, vol. 1, pp. 167–172.

[10] Adam Coates and Andrew Y Ng, "The importance of encoding versus training with sparse coding and vector quantization," in *International Conference on Machine Learning*, 2011, vol. 8, p. 10.

[11] M. Soltanolkotabi and E. Candes, "A geometric analysis of subspace clustering with outliers," *The Annals of Statistics*, vol. 40, no. 4, pp. 2195–2238, 2012.

[12] Ehsan Elhamifar and René Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

[13] Roman Vershynin, "Introduction to the non-asymptotic analysis of random matrices," *arXiv:1011.3027*, 2010.

[2] In fact, the angle between subspaces controls the size of the dictionary that satisfies D-SDP in the ℓ_1 case.