Affine Rank Minimization Solution to Sparse Modeling

Iman Abbasnejad[1,2], Sridha Sridharan[1], Simon Denman[1], Clinton Fookes[1], Simon Lucey[2]
[1]Image and Video Laboratory, Queensland University of Technology (QUT), Brisbane, QLD, Australia
[2]The Robotics Institute, Carnegie Mellon University, 5000 Forbes Ave, PA, USA
Email:{i.abbasnejad, s.sridharan, s.denman, c.fookes}@qut.edu.au, slucey@cs.cmu.edu

Abstract

Compressed sensing is a simple and efficient technique that has a number of applications in signal processing and machine learning. In machine learning it provides answers to questions such as: "Under what conditions is the sparse representation of data efficient?", "When is learning a large margin classifier directly on the compressed domain possible?" and "Why does a large margin classifier learn easier if the data is sparse?". This work tackles the problem of feature representation from the context of sparsity and affine rank minimization by leveraging compressed sensing from learning perspective in order to provide answers to the aforementioned questions. We show, for a full-rank signal the high dimensional sparse representation of data is efficient, because from the classifier viewpoint such a representation is in fact a low dimensional problem. We provide practical bounds on the linear classifier to investigate the relationship between the SVM classifier in the high dimensional and compressed domains and show for the high dimensional sparse signals, when the bounds are tight directly learning in the compressed domain is possible.

1. Introduction

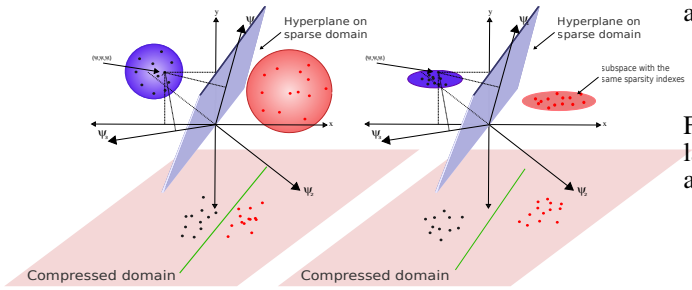
Classification is one of the most fundamental problems in machine learning and has a number of applications in signal and image processing. The problem of classification can be categorized into first, transforming the input examples into an appropriate feature space and second, applying a classification algorithm on the transformed features. Generally two approaches can be introduced for rich feature representation: (i) *sparse representation* of the input signals with respect to only a few high dimensional basis and (ii) *low-rank* structure of the input examples. Sparse representation has a number of advantages in the context of machine learning, for instance representing the input examples with only a few active elements makes learning a classifier easier, or a sparse signal can leverage the benefits of a low dimensional data by applying a linear compressing method on it. On the other hand, the low-rank structure helps a classifier to simply measure the similarity between the low-rank patterns of the input examples.

Recently Candès and Donoho [16, 20] introduced a novel sampling theory called compressed sensing (CS) and showed that one can recover the signal from far fewer samples than the Nyquist-Shanon sampling rate under two conditions: first, the signal be sparse and second, the restricted isometric property holds. We shall discuss these properties more in the subsequent sections. CS enables a potentially large reduction in the sampling and computation costs for sensing signals and has a number of applications in signal processing and machine learning. From the context of machine learning, recently Calderbank [15] show learning directly on the compressed domain is possible. They provide theoretical bounds on the SVM classifier and show with high probability, a random projection of linear SVM classifier to low dimensional domain has true accuracy close to the best SVM classifier in the data domain:

$$\|A\boldsymbol{\omega}\|_2^2 \leq \|\boldsymbol{\omega}\|_2^2 + e \tag{1}$$

where A is a linear projection from high dimensional to low dimensional feature space, $\boldsymbol{\omega}$ is a linear classifier trained on the high dimensional feature space and e is the error (see Fig. 1 for the visualization). There are two central advantages in the context of compressed learning from machine learning viewpoint. Firstly learning classifiers on the compressed domain enables a large reduction in computational cost. Secondly, it eliminates the cost of recovering the signals if we are only interested in the classification task.

One drawback with Calderbank's bounds is that, they are highly sensitive to the degree of sparsity of the input signals. As is shown in Section 4, for sparser signals Calderbank's bounds are tight, however by increasing the sparsity the bounds become looser. In general introducing an algorithm that controls the level of sparsity is an NP-hard problem. Furthermore, in many applications we cannot control the degree of sparsity. In order to make compressed learning more efficient and applicable we present new bounds on the margin of linear classifier that are derived by anneal-



(a) Full-rank representation (b) Low-rank representation

Figure 1: Compressed learning visualization. a) For the full-rank representation the data is represented as a cloud of points within the three dimensions (sphere). b) when the data are in the low-rank representation the cloud of points are in 1-dimension or at most 2 dimensions (here we assume disk).

ing the entropy of the sparsity configuration and the rank of input signals. As a result we will be able to investigate the problem of feature representation and its impact on the performance of the classifier.

1.1. Paper Contributions

In this paper, we study the problem of compressed sensing in the context of linear kernel SVM classifier and feature representation. The underlying idea behind our work is that each linear SVM is a linear combination of the training examples. Such a presentation is not unique and can be expressed as a mixture of examples which are formed in different structures i.e. sparse and low-rank. Based on this assumption we construct the new bounds on the SVM classifier to draw the relationship between the classifier in the high dimensional domain and the compressed domain. More specifically, we build our assumption based on two principles of sparsity and affine rank minimization to show when tighter bounds are provided. As a result we can demonstrate in which conditions sparse and low-rank representation of the input examples are efficient from learning viewpoint. Unlike Calderbank's model [15], where the bounds are only depend on the sparsity, we present more practical bounds that make compressed learning applicable. Finally, through experiments we support our claim.

Notations and Definitions

In this paper we adopt the same notation as [2, 8, 15]. We assume $\mathbf{x} \in \mathbb{R}^n$ is a $k-$sparse vector and its ℓ_2-norm is bounded by some parameter R, $\|\mathbf{x}\|_2 \leq R$. $A_{m\times n}$ is the linear measurement matrix used in compressed sensing. We define the data domain as:

$$\mathcal{X} = \{(\mathbf{x}, y) : \mathbf{x} \in \mathbb{R}^n, y \in \{-1, 1\}\}$$

and the measurement domain $\mathcal{M}$ as:

$$\mathcal{M} = \{(A\mathbf{x}, y) : (\mathbf{x}, y) \in \mathcal{X}\}$$

For $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\} \in \mathbb{R}^{n\times m}$, σ_i denotes the i-th largest singular value of $\mathbf{X}$. We define the Frobenius, ℓ_2 and nuclear norm as:

$$\|\mathbf{X}\|_F = \sqrt{\sum_{i=1}^{\min\{n,m\}} \sigma_i^2}, \|\mathbf{X}\|_2 = \sigma_1, \|\mathbf{X}\|_* = \sum_{i=1}^{\min\{n,m\}} \sigma_i \tag{2}$$

For $\alpha \in \mathbb{R}$ and for any arbitrary vector $\mathbf{x}$:

$$\|\alpha\mathbf{x}\|_2 = |\alpha|\|\mathbf{x}\|_2, \quad \|\mathbf{X}\|_F^2 = \sum_{i=1}^{m} \|\mathbf{x}_i\|_2^2 \tag{3}$$

Similar to [15] we assume $\mathcal{D}$ is some unknown distribution over $\mathcal{X}$, and S has M labeled examples i.i.d from $\mathcal{D}$:

$$S = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_M, y_M)\}$$

and AS is the compressed representation of S:

$$AS = \{(A\mathbf{x}_1, y_1), (A\mathbf{x}_2, y_2), \ldots, (A\mathbf{x}_M, y_M)\}$$

Definition 1.1. *Support Vector Machine: For M examples sampled i.i.d from distribution $\mathcal{D}$ the SVM's classifier $\boldsymbol{\omega}$ is obtained as a linear combination of the training vectors:*

$$\boldsymbol{\omega} = \sum_{i=1}^{M} \alpha_i y_i \mathbf{x}_i, \quad \forall i : 0 \leq \alpha_i \leq \frac{C}{M}, \quad \|\boldsymbol{\omega}\|_2 \leq C$$

For the linear classifier $\boldsymbol{\omega}$ we define its true hinge loss as:

$$H_{\mathcal{D}}(\boldsymbol{\omega}) = \mathbf{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[1 - y\boldsymbol{\omega}^T\mathbf{x}]$$

and the true regularization loss of $\boldsymbol{\omega}$ as:

$$L(\boldsymbol{\omega}) = H_{\mathcal{D}}(\boldsymbol{\omega}) + \frac{1}{2C}\|\boldsymbol{\omega}\|_2 \tag{4}$$

$\boldsymbol{\omega}$ can be found by minimizing the empirical loss over $\mathcal{X}$:

$$\hat{L}(\boldsymbol{\omega}) = \hat{H}_S(\boldsymbol{\omega}) + \frac{1}{2C}\|\boldsymbol{\omega}\|_2$$

where $\hat{H}_S(\boldsymbol{\omega})$ is the empirical hinge loss:

$$\hat{H}_S(\boldsymbol{\omega}) = \mathbf{E}_{(\mathbf{x}_i,y_i)\sim S}[1 - y_i\boldsymbol{\omega}^T\mathbf{x}_i] \tag{5}$$

For the proof, the readers are referred to [15, 19].

Definition 3.2. *Convex hull is defined by a unique bounded polyhedron, that vertices constitute a set of points:*

$$\mathbf{s} = \sum_{i=1}^{M} \alpha_i \mathbf{x}_i, \forall\, i : \ \alpha_i \geq 0 \ , \ \sum_{i=1}^{M} \alpha_i = 1$$

2. Related work

Previously Johnson and Lindenstrauss [23] demonstrated that projecting high dimensional data onto a random low dimensional subspace, with high probability only changes the pairwise distances between all the points by $(1 \pm \epsilon)$. The problem of random projection has widely studied in the literature in the context of dimensionality reduction, clustering [12, 14] and nearest neighbor algorithms [11]. The use of Johnson and Lindenstrauss's lemma in classification is first introduced in [9]. They theoretically showed that the random projection of high dimensional data onto a low dimension subspace correctly classifies. This work is followed by Blum et al. [13] and Balcan et al. [10]. They demonstrated that if high dimensional data is separable by a large margin, then a random projection to a low dimensional subspace will with high probability preserve separability.

More recently Calderbank et al. [15] used the idea of CS and introduced an algorithm to show learning on the compressed domain is possible. They provide the theoretical bounds on the regularization loss of the linear SVM (Eq. 4) in the low dimensional domain, and show that the performance of the best classifier in the high dimensional domain, is approximately preserved by random projection to a low dimensional subspace.

An initial question one might ask is whether compressed learning can be generalized to any arbitrary input example. As is shown in Section 4, there are two important parameters that fundamentally affect the bounds presented by Calderbank: (i) the level of sparsity and (ii) the entropy of sparsity indexes (i.e. rank) of the input examples. One limitation with such a presentation is that, in order to the bounds be tight the degree of sparsity has to be fixed or limited by some threshold. However, in many cases, it is challenging to control the level of sparsity. In this work to address this problem we build new bounds on the classifier based on the notion of affine rank minimization and sparsity patterns in order to present more practical bounds. We should note that in this presentation we are not trying to present tighter bounds however we are looking to build a practical bounds that make compressed learning more applicable and generalizable.

3. An Introduction to Compressed Sensing

Compressed sensing (CS), is based on the idea that signals that can be represented in their proper basis can have a concise representation. From signal processing viewpoint, for the sparse signals, information rate is much smaller than bandwidth and the sparse signals can be recovered from far fewer samples than required by Nyquist-Shanon sampling theorem. CS is governed by two principles; sparsity and incoherence. Mathematically,

$$\mathbf{z} = A_{m\times n}\mathbf{x}$$

where $m \ll n$, $\mathbf{x}$ is a $k-$sparse signal, $\mathbf{z} \in \mathbb{R}^m$ is the compressed vector and $A_{m\times n}$ is the measurement matrix. In [17] they show that in order $A_{m\times n}$ act as a compressed matrix it should satisfy the Restricted Isometric Property (RIP).

Definition 4.1. *Restricted Isometric Property: $A_{m\times n}$ satisfies $(2k,\epsilon)-$RIP if it acts as a near-isometric with distortion factor ϵ, over all $2k-$sparse vectors $\mathbf{x}$.*

$$(1-\epsilon)\|\mathbf{x}\|_2 \leq \|A\mathbf{x}\|_2 \leq (1+\epsilon)\|\mathbf{x}\|_2$$

In [17] they show that if the entries of $A_{m\times n}$ sampled i.i.d. from a Gaussian random distribution then with high probability this matrix satisfies the RIP:

$$A_{m\times n} \sim \frac{1}{\sqrt{m}}\mathcal{N}(0,1), \qquad m = \Omega(k\log(n/k))$$

4. Compressed Learning

Recently Calderbank and his colleagues demonstrate, with high probability, SVM classifier on the compressed domain performs almost as accurate as the best classifier on the data domain (see Fig. 1). To establish the bounds they first expand RIP and observe that it preserves the inner product between the two $k-$sparse vectors $\mathbf{x}, \acute{\mathbf{x}}$.

Lemma 4.1 [15] *Let $A_{m\times n}$ the measurement matrix satisfying $(2k,\epsilon)-$RIP, and $\mathbf{x}, \acute{\mathbf{x}}$ be two $k-$sparse vectors in $\mathbb{R}^n$, with $\|\mathbf{x}\|_2 \leq R, \|\acute{\mathbf{x}}\|_2 \leq R$. Then:*

$$(1+\epsilon)\mathbf{x}^T.\acute{\mathbf{x}} - 2R^2\epsilon \leq (A\mathbf{x})^T(A\acute{\mathbf{x}})$$

and

$$(A\mathbf{x})^T(A\acute{\mathbf{x}}) \leq (1-\epsilon)\mathbf{x}^T.\acute{\mathbf{x}} + 2R^2\epsilon$$

To see the proof, we refer the readers to [15]. Since the SVM classifier is a linear combination of training examples (Definition 1.1) they have generalized Lemma 4.1 on the combination of the sparse vectors and show that the RIP also approximately preserves the inner product between any two vectors from the convex hull of the set of sparse vectors.

Theorem 4.2 [15] *Let $A_{m\times n}$ be a matrix satisfying $(2k,\epsilon)-$RIP. Let M, N be two integers, and*

$$\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_M, y_M), (\acute{\mathbf{x}}_1, \acute{y}_1), \ldots, (\acute{\mathbf{x}}_N, \acute{y}_N)\} \in \mathcal{X}$$

Let $\alpha_1, \ldots, \alpha_M, \beta_1, \ldots, \beta_N$ be non-negative numbers such that $\sum_{i=1}^{M}\alpha_i \leq C$ and $\sum_{j=1}^{N}\acute{\alpha}_j \leq D$ for some $C, D \geq 0$.

$$\boldsymbol{\omega} = \sum_{i=1}^{M}\alpha_i y_i \mathbf{x}_i, \quad \acute{\boldsymbol{\omega}} = \sum_{j=1}^{N}\acute{\alpha}_j \acute{y}_i \acute{\mathbf{x}}_i$$

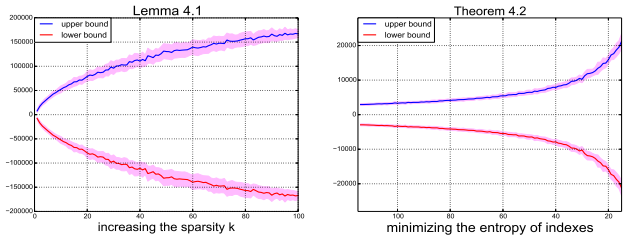
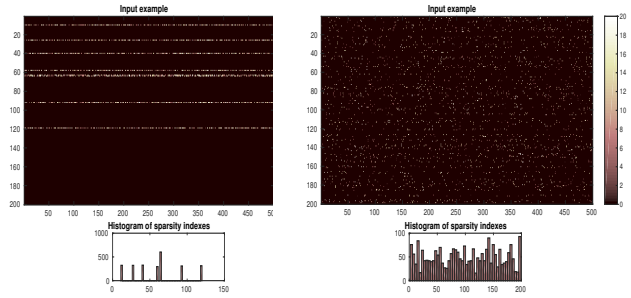


Figure 2: Lemma 4.1 for different sparsity values, Theorem 4.2 for different inputs with different histogram of indexes.



(a) (b)

Figure 3: Two sets of examples with their corresponding histogram of indexes. In this figure, (a) has a low rank structure and (b) has a full rank structure.

Then:

$$|\boldsymbol{\omega}^T \acute{\boldsymbol{\omega}} - (A\boldsymbol{\omega})^T (A\acute{\boldsymbol{\omega}})| \leq 3CDR^2\epsilon$$

The details of the proof can be found in [15]. Finally Definition 1.1, Lemma 4.1 and Theorem 4.2 imply that the true regularization term, Eq. 4, of the projected classifier on the compressed domain $A\boldsymbol{\omega}_S$, performs as accurate as the best classifier on the data domain $\boldsymbol{\omega}_S$. For more details we strongly encourage the readers to read [15].

4.1. Setting the Problem

Let us begin with a review of Lemma 4.1. This lemma is crucial because it shows the distance between the inner product of any two $k-$sparse vectors under a linear projection. One parameter that is very important and plays a significant role in the theory of compressed learning is the level of sparsity: k . We investigate how the bounds vary with respect to k. To do so we generate $n = 5000$ dimensional $k-$sparse vectors ($\mathbf{x} \in \mathbb{R}^n$), where k changes between $[10 : 200]$ and $\|\mathbf{x}\|_2 \leq R = 500$. Fig. 2 shows the upper and lower bounds obtained from Lemma 4.1 for different sparsity values.

We also examine Theorem 4.2, which implies that the linear projection, approximately preserves the inner product between the combination of input examples. For simulation in this case we synthetically examine the other important factor that could affect the bounds. Since $\boldsymbol{\omega}$ is a linear combination of the training examples, such a presentation is not unique and can be expressed as a combination of examples that are formed in different patterns and structures. In this simulation we consider $\boldsymbol{\omega}$ as a linear combination of the input examples that belong to different sets with different entropy of indexes (Fig. 3 shows two different sets of examples with their corresponding histogram of indexes). Fig. 2 shows the upper and lower bounds obtained from Theorem 4.2 for this experiment.

From Fig. 2 we can see how the bounds derived in [15] turn out in practical cases. As can be seen the tighter bounds are provided only for some limited sparsity values. In this paper we construct the new bounds that show where the bounds presented in Lemma 4.1 and Theorem 4.2 are tighter and compressed learning is possible. This presentation enables us to study the problem of feature representation to see when the sparse representation of data is efficient from the classification viewpoint.

5. Compressed Learning via Rank Minimization

In this section, we investigate the problem of compressed learning in order to provide more practical bounds on the margin of SVM classifier.

5.1. From CS to Rank Minimization

The affine rank minimization problem can be described as:

$$\min \ \textit{rank}(\mathbf{X}), \quad \text{s.t. } \mathcal{A}(\mathbf{X}) = \mathcal{A}(\mathbf{X}_0) \tag{6}$$

where $\mathbf{X} \in \mathbb{R}^{n\times m}$ is the optimization variable, $\mathcal{A} : \mathbb{R}^{n\times m} \rightarrow \mathbb{R}^p$ is a linear operator and $\mathbf{X}_0$ denotes the $r-$rank solution ($r \ll \min(n, m)$). Thus we are interested:

$$\min \ \|\mathcal{A}(\mathbf{X}) - \mathcal{A}(\mathbf{X}_0)\|_2, \quad \text{s.t. } rank(\mathbf{X}_0) \ll r \tag{7}$$

To guarantee uniqueness, $\mathcal{A}(.)$ is assumed to satisfy the corresponding RIP for affine transformation as:

$$(1-\epsilon)\|\mathbf{X}\|_F \leq \|\mathcal{A}(\mathbf{X})\| \leq (1+\epsilon)\|\mathbf{X}\|_F$$

In general Eq. 7 is known to be an NP-hard problem and also hard to approximate [24] due to the non-convexity of $rank(\mathbf{X})$. One popular heuristic that recently has been proposed in the literature [21] replaces the *rank* function with the summation of singular values of the decision variable. The heuristic is to solve:

$$\min \ \|\mathbf{X}\|_*, \quad \text{s.t. } \mathbf{X} \in \mathcal{C} \tag{8}$$

where $\|.\|_*$ is the nuclear norm and is defined in Eq. 2 and $\mathcal{C}$ is belong to the set of low rank matrices. This optimization is convex, and can be cast as a semidefinite program and solved efficiently. For a survey we refer the readers to [27].

5.2. Rank Minimization and SVM Classifiers

As is shown in Section 4, previous work on compressed learning is hard to generalize to any arbitrary input examples, because the bounds are highly sensitive to the degree of sparsity and for some sparsity values the bounds are loose. In this section we study the theoretical features of our model based on two principles of sparsity and affine rank minimization in order to demonstrate the practical bounds on the margin of linear classifier . To do so, we begin with some results from [15].

The results of Definition 1.1, Lemma 4.1 and Theorem 4.2 suggest that the inner product between the linear combination of two sets of examples under random projection is bounded by a lower bound:

$$(1+\epsilon)\sum_{i=1}^{M}\sum_{i=1}^{N}\alpha_i\acute{\alpha}_j y_i\acute{y}_j\mathbf{x}_i^T\acute{\mathbf{x}}_j - \sum_{i=1}^{M}\sum_{j=1}^{N}(\|\alpha_i y_i\mathbf{x}_i\|_2^2 + \|\acute{\alpha}_j\acute{y}_j\acute{\mathbf{x}}_j\|_2^2)\epsilon \leq (A\sum_{i=1}^{M}\alpha_i y_i\mathbf{x}_i)^T(A\sum_{j=1}^{N}\acute{\alpha}_j\acute{y}_j\acute{\mathbf{x}}_j) \tag{9}$$

and an upper bound:

$$(A\sum_{i=1}^{M}\alpha_i y_i\mathbf{x}_i)^T(A\sum_{j=1}^{N}\acute{\alpha}_j\acute{y}_j\acute{\mathbf{x}}_j) \leq \tag{10}$$
$$(1-\epsilon)\sum_{i=1}^{M}\sum_{i=1}^{N}\alpha_i y_i\mathbf{x}_i^T\acute{\alpha}_j\acute{y}_j\acute{\mathbf{x}}_j + \sum_{i=1}^{M}\sum_{j=1}^{N}(\|\alpha_i y_i\mathbf{x}_i\|_2^2 + \|\acute{\alpha}_j\acute{y}_j\acute{\mathbf{x}}_j\|_2^2)\epsilon$$

by putting $\boldsymbol{\omega} = \sum_{i=1}^{M}\alpha_i y_i\mathbf{x}_i, \acute{\boldsymbol{\omega}} = \sum_{j=1}^{N}\acute{\alpha}_j\acute{y}_j\acute{\mathbf{x}}_j$ and $\mathbf{x} = \acute{\mathbf{x}}, N = M$ in the above equations and using Eq. 3 we have[1]:

$$(A\boldsymbol{\omega})^T(A\boldsymbol{\omega}) \leq (1-\epsilon)\boldsymbol{\omega}^T\boldsymbol{\omega} + 2K^2\epsilon\sum_{i=1}^{M}\|\alpha_i\mathbf{x}_i\|_2\sum_{i=1}^{M}\|\alpha_i\mathbf{x}_i\|_2 \tag{11}$$

where $\sum_{i=1}^{M}|y_i| \leq K$. This bound is the main result of [15] and they use it to construct the bounds on the true regularization loss of SVM as presented in Eq. 4.

In this paper in order to study the problem of compressed learning we begin with the above presentation to introduce our main results on the linear kernel SVM classifier. Our main assumption in this paper is that, since the SVM classifier is a linear combination of training examples, such a presentation can be defined as a linear summation of the training examples in terms of different structures. In other words, we re-parametrize $\boldsymbol{\omega}$ presented in Eq. 11 as:

$$\boldsymbol{\omega} = \sum_{i=1}^{M}\beta_i y_i\mathbf{v}_i, = [\beta_1 y_1(\mathbf{x}_1\circ\tau_1)|\ldots|\beta_M y_M(\mathbf{x}_M\circ\tau_M)]$$

where $\tau_i = \{\tau_1,\ldots,\tau_M\}$ is a set of transformations, which are used to manipulate the input examples. In this presentation we consider such a transformation only applies on the indexes of sparse input examples and makes the rank of the transformed examples as small as possible (see Fig 3). Rewriting Eq. 10 and Eq. 11 and substituting $\boldsymbol{\omega} = \sum_{i=1}^{M}\beta_i y_i\mathbf{v}_i$ in the equations, yields:

$$(A\boldsymbol{\omega})^T(A\boldsymbol{\omega}) \leq (1-\epsilon)\boldsymbol{\omega}^T\boldsymbol{\omega} + 2K^2\epsilon\sum_{i=1}^{M}\|\beta_i\mathbf{v}_i\|_2\sum_{i=1}^{M}\|\beta_i\mathbf{v}_i\|_2 \tag{12}$$

In order to compare our proposed bounds with the previous work [15] we only need to compare the right-hand sides of the inequalities in Eq. 11 and Eq. 12 ($\sum_{i=1}^{M}\|\beta_i\mathbf{v}_i\|_2$ and $\sum_{i=1}^{M}\|\alpha_i\mathbf{x}_i\|_2$ terms). From the definitions and Eq. 3 we have:

$$\sum_{i=1}^{M}\|\alpha_i\mathbf{x}_i\|_2 = \sum_{i=1}^{M}|\alpha_i|\|\mathbf{x}_i\|_2, \quad \sum_{i=1}^{M}\|\beta_i\mathbf{v}_i\|_2 = \sum_{i=1}^{M}|\beta_i|\|\mathbf{v}_i\|_2$$

and since each $\mathbf{x}_i, \mathbf{v}_i$ belongs to the convex hull, $\sum_{i=1}^{M}\alpha_i = 1, \sum_{i=1}^{M}\beta_i = 1$, therefore we have:

$$\sum_{i=1}^{M}|\alpha_i|\|\mathbf{x}_i\|_2 \leq \sum_{i=1}^{M}\|\mathbf{x}_i\|_2, \quad \sum_{i=1}^{M}|\beta_i|\|\mathbf{v}_i\|_2 \leq \sum_{i=1}^{M}\|\mathbf{v}_i\|_2$$

so the only parameters we need to compare are $\sum_{i=1}^{M}\|\mathbf{v}_i\|_2$ and $\sum_{i=1}^{M}\|\mathbf{x}_i\|_2$ which are equal to the Frobenius norm of the sets of examples, $\mathbf{V}$ and $\mathbf{X}$.

5.3. Norm Effects and SVM Bounds

In Section 5.2 we demonstrated the effects of Frobenius norm on the presented bounds in Eq. 11 and Eq. 12. In this section we investigate the bounds behavior based on the principle of affine rank minimization in the context of feature representation. In order to investigate the effects of low-rank structure we use the notion of alignment [18, 26] which can be efficiently computed through the rank minimization algorithm. As is shown in [26], the matrix of aligned examples (images) will have low-rank, ideally rank one, however for noisy signals, the aligned examples might have an unknown rank higher than one. For comparison we consider two different cases: $\|\mathbf{V}\|_F \leq \|\mathbf{X}\|_F$ and $\|\mathbf{V}\|_F \geq \|\mathbf{X}\|_F$. We define $\mathbf{V} = \mathbf{X}\circ\tau = [\mathbf{v}_1|\ldots|\mathbf{v}_M]$, where τ belongs to the set of transformations that minimizes the rank of the input matrix $\mathbf{X}$ and makes it well aligned.

Unfortunately, in this case the value of Frobenius norm is not straightforward to calculate as it depends on the problem of alignment and sets of transformations such as warping, but instead from the norm definitions, we know for any arbitrary matrix $\mathbf{X}$ of rank r the Frobenius norm is bounded by:

$$\|\mathbf{X}\|_2 \leq \|\mathbf{X}\|_F \leq r\|\mathbf{X}\|_2$$

Therefore in order to compare the Frobenius norm we can instead compare the ℓ_2-norm of $\mathbf{X}$ and $\mathbf{V}$ which is the dual norm of the nuclear norm [21].

[1] In this work due to the lack of space we only provide the upper bound, however we have the same procedure for the lower bound.

We start with the case when the signal is well-aligned. In other words the entropy of indexes is minimized and all the $\mathbf{v}_i$ vectors have almost a similar sparsity structure (Fig. 3a). As explained in [26] this situation is equivalent to $rank(\mathbf{V}) \leq rank(\mathbf{X})$. Follow from Eq. 8 this state can be expressed as:

$$\|\mathbf{V}\|_* \leq \|\mathbf{X}\|_* \Rightarrow \sum_{i=1}^{\acute{r}} \acute{\sigma_i} \leq \sum_{i=1}^{r} \sigma_i$$

where $rank(\mathbf{V}) = \acute{r}, rank(\mathbf{X}) = r$ and $\acute{\sigma_i}, \sigma_i$ are the $i-$th singular values of $\mathbf{V}$ and $\mathbf{X}$ respectively. We can show that if:

$$\|\mathbf{X}\|_* - \|\mathbf{V}\|_* \leq (r - \acute{r})\sigma_1 \tag{13}$$

then the following inequality holds:

$$\|\mathbf{V}\|_2 \geq \|\mathbf{X}\|_2$$

In other words the nuclear norm minimization is equivalent to maximizing the ℓ_2-norm of a matrix, if Eq. 13 holds. Since in this paper we assume τ is a set of transformations that only applies on the input examples and makes them low-rank and well-aligned, therefore Eq.13 holds (otherwise the critical information of the input examples is lost). Therefore when the input examples are well-aligned the ℓ_2-norm increases and the bounds become loose. On the other hand in the case of $rank(\mathbf{V}) \geq rank(\mathbf{X})$ (see Fig 3b), following the same procedure it implies that:

$$\|\mathbf{V}\|_2 \leq \|\mathbf{X}\|_2$$

As a result, as long as $rank(\mathbf{V}) > rank(\mathbf{X})$ the bounds become tighter.

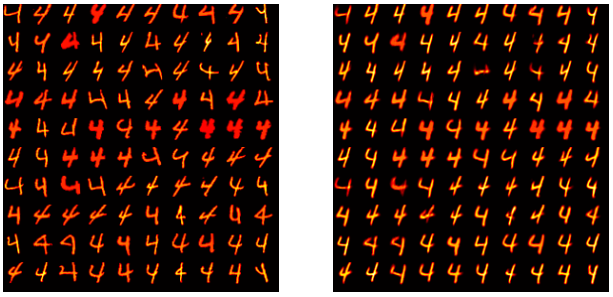
5.4. CS and Feature Representation

Up to now we have dealt with the problem of compressed learning from the compressed sensing perspective and theoretically showed that when the presented bounds in Eq. 11 and Eq. 12 are tight, the performance of the classifier on the compressed domain $\boldsymbol{\omega}_{AS}$ is similar to the best classifier on the data domain $\boldsymbol{\omega}_S$:

$$\boldsymbol{\omega}_S \approx \boldsymbol{\omega}_{AS} \Longrightarrow \boldsymbol{\omega}_S^T \boldsymbol{\omega}_S \approx (A\boldsymbol{\omega}_S)^T (A\boldsymbol{\omega}_S)$$

we have also shown for the input examples $\mathbf{X} \in \mathbb{R}^{m \times n}$ the tighter bounds are obtained when $rank(\mathbf{X}) > 1$.

Remark 5.1 *When the input features are not well-aligned the sparse representation of the signal is efficient and it is advantageous to learn the SVM classifier $\boldsymbol{\omega}_S$ on the sparse domain. Because such data can be randomly projected to a low dimensional feature space, and it is in fact a low-dimensional problem from the classification viewpoint. On the other hand, when the input examples are low-rank and well-aligned, dictionary learning and sparse coding do not have remarkable effects on the classifiers from the learning aspect. Because such a problem is not a low dimensional problem from the classifiers perspective anymore.*



Input Image Aligned Image

(a) MNIST dataset

Input Image Aligned Image

(b) LFW dataset

Figure 4: Examples of the MNIST and LFW databases. "Input Images" refers to the original images from the dataset, "Aligned Images" refers to the images aligned using RASL [26].

The following remark is the consequence of Section 5.3 and can be generalized on any arbitrary input examples $\mathbf{X}$.

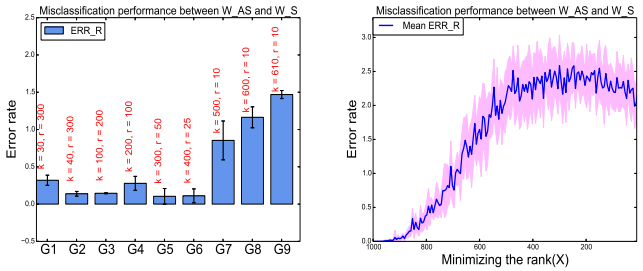
6. Experiments

In this section, we introduce our experiments on synthetic and real data to complement our theoretical study.

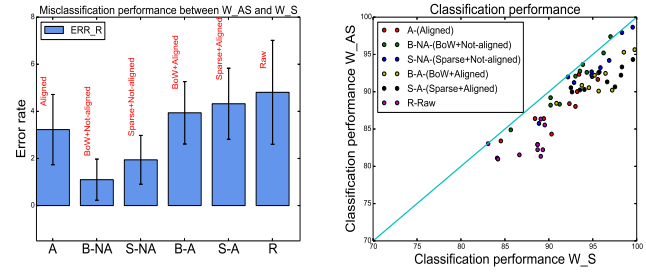
6.1. Experimental Setup

For comparison we consider two sets of features: (i) sparse and, (ii) low-rank. For sparse representations, we consider two well-known approaches, Bag-of-Words (BoW) and sparse coding and for low-rank feature templates we use the RASL model presented in [26].

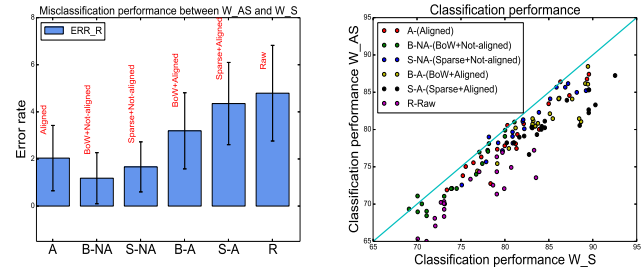
BoW: BoW representations can be viewed as the sparse encoding of the i-th input vector using the codebook matrix $\mathbf{D} \in \mathbb{R}^{D \times T}$ where T is the number of vocabular-



(a) Synthetic experiments



(b) MNIST experiments



(c) LFW experiments

Figure 5: This figure shows the results on the proposed datasets. The bars and the plot in Fig. 5a show the error rates between the performance of the classifier on the data domain $\boldsymbol{\omega}_S$ and the compressed domain $\boldsymbol{\omega}_{AS}$ and the plots in Fig. 5b and Fig.5c show the classification performance of $\boldsymbol{\omega}_S$ vs $\boldsymbol{\omega}_{AS}$ for different feature representations.

ies [4, 5]:

$$\eta\{\mathbf{x}\} = \arg\min_{\mathbf{b}} \|\mathbf{x} - \mathbf{D}\mathbf{b}\|_2, \quad \text{s.t. } \mathbf{b} \in \mathbb{B} \tag{14}$$

$\mathbb{B} = \{\mathbf{e}_t\}_{t=1}^{T}$ is the set of all T dimensional vectors $\mathbf{e}_t$ containing all zeros except for one. In this work, the codebook is learned through k-means clustering.

Sparse coding: Sparse coding aims to factorize an ensemble of input vectors $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_M]$ into a linear combination of some basis under sparsity constraints:

$$\eta\{\mathbf{x}\} = \arg\min_{\mathbf{b}} \|\mathbf{x} - \mathbf{D}\mathbf{b}\|_2 + \lambda\|\mathbf{b}\|_1 \tag{15}$$

where λ is a parameter controlling the sparsity penalty.

RASL model: For alignment, we use the RASL model presented in [26]. RASL is presented by solving the following rank-minimization problem:

$$\min \|\mathbf{X}_i\|_* + \lambda\|E_i\|_1, \quad \text{s.t. } I_i \circ (\tau_0, \tau_i)^{-1} = \mathbf{X}_i + E_i \tag{16}$$

where $\mathbf{X}$ is the low-rank representation of image I, E is the error and τ is the set of transformations. We should note that, RASL model only rotate, translate and warp the images (see Fig 4), therefore Eq. 13 holds.

Evaluation metrics: To evaluate the performance, we report the error between the area under the ROC curve for the classifiers on the data domain $\boldsymbol{\omega}_S$ and compressed domain $\boldsymbol{\omega}_{AS}$.

SVM: For the linear SVM we use the implementation [25]. We do the standard grid-search on cross-validation to tune parameters including C.

6.2. Databases

Synthetic Data: To conduct controlled experiments with known ground truth, we synthetically generate two classes of $k-$sparse data $\mathbf{X}_1 \in \mathbb{R}^{n\times M}, \mathbf{X}_2 \in \mathbb{R}^{n\times M}$ with known $r = rank(\mathbf{X})$ that are sampled i.i.d. from two sets of Gaussian Random distributions:$\mathbf{X}_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathbf{X}_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$. In this work we generate nine sets of data (G1-G9) with various values of r and k. For this experiment we set $n = 1000, M = 300, \mu_1 = 0, \mu_2 = 3, \sigma_1^2 = \sigma_2^2 = 0.1$.

MNIST: This dataset is a large handwritten digits dataset and contains a training set of 60,000, and a test set of 10,000 examples from 10 digits, zero to nine. The images are centered in 28 grey level images. Fig. 4a shows some examples [1].

Labeled Faces in the Wild: LFW [22] contains $13,233$ face images of $5,749$ different persons with different gender, ages and etc. under different constraints [7]. In this experiment we choose 700 images from 20 distinguished classes for evaluation. Fig 4b displays examples.

6.3. Results

Fig. 5 shows the results on the proposed datasets. We report the performance of the SVM on the data domain $\boldsymbol{\omega}_S$ vs the compressed domain $\boldsymbol{\omega}_{AS}$ (the plots in Fig. 5b and 5c), and the error rate between the performance of $\boldsymbol{\omega}_S$ and $\boldsymbol{\omega}_{AS}$ (the plot in Fig. 5a and the bars). Fig. 5a visualizes the results on the synthetic data for various sparsity and rank values. As is shown, when the data is sparse and the signal is full-rank, "G1"-"G6" the bounds are tight (the error rate is small), however by minimizing the rank, "G7", "G8" and "G9" the bounds become looser (the error rate increases). The plot in Fig. 5a displays the error rate between $\boldsymbol{\omega}_S$ and $\boldsymbol{\omega}_{AS}$ for the synthetic data with the fix sparsity $k = 10$ and different ranks values. As can be seen by minimizing the rank the error rate increases and the bounds become looser.

Fig. 5b and Fig. 5c show the results on the MNIST and LFW datasets respectively. In these figures, "A" refers to the low-rank representation of the input signals using RASL model in Eq. 16, "B-NA" and "S-NA" refer to the sparse representation of the examples using BoW in Eq. 14 and Sparse coding in Eq. 15, "B-A" and "S-A" correspond to the low-rank representation of the BoW and the Sparse coding features using RASL model, and "R" refers to the raw representation of the input examples. We should note that for the feature representation using BoW, the sparsity is one, $k = 1$ and for the sparse coding, $k > 1$. As can be seen from the figures, the sparse and high rank representation of the examples, "B-NA" and "S-NA", provide tighter bounds in Eq. 12 (error rate $\approx 1\%$), however for the low-rank representation the bounds become looser (error rate $\approx 5\%$). On the basis of our experiments, we make the following conclusions:

– Generally for the full-rank examples, the sparse representation of the signal provides tighter bounds. As a consequence, from the linear classifiers viewpoint, the full-rank and high dimensional representation of the input examples is similar to the random projection of such data to a low dimensional subspace.

– The sparse and low-rank data, obtain large error rate between the performance of the classifier on the data domain and the compressed domain. Thus, such representation provides looser bounds. As a result, from the classifier aspect, such data is not in fact a low dimensional problem.

7. Discussion and Conclusion

By drawing the connections between the sparsity and affine rank minimization, we are able to determine the bounds on the linear SVM classifier to show in which conditions the alignment and sparsity are efficient from classifier viewpoint. We, theoretically and empirically demonstrate that for the full-rank signal the sparsity is efficient, because such a representation can randomly project to a low dimensional subspace and is in fact a low dimensional problem and classifier is learning easier in such data. On the other hand, for well-aligned and low-rank data, sparsity does not have advantages. Because, such a representation cannot randomly project to a low dimensional subspace anymore. Finally through experiments we supported our claim.

Acknowledgment

To see the proof of Section 5.3 please see [3]. Also the published version of this paper can be found in [6].

References

[1] E. Abbasnejad, A. Dick, and A. v. d. Hengel. Infinite variational autoencoder for semi-supervised learning. *arXiv preprint arXiv:1611.07800*, 2016.

[2] E. Abbasnejad, J. Domke, S. Sanner, et al. Loss-calibrated monte carlo action selection. In *AAAI*, pages 3447–3453, 2015.

[3] I. Abbasnejad. Rank minimization and sparse modeling. 2017.

[4] I. Abbasnejad, S. Sridharan, S. Denman, C. Fookes, and S. Lucey. Learning temporal alignment uncertainty for efficient event detection. In *Digital Image Computing: Techniques and Applications (DICTA), 2015 International Conference on*, pages 1–8. IEEE, 2015.

[5] I. Abbasnejad, S. Sridharan, S. Denman, C. Fookes, and S. Lucey. Complex event detection using joint max margin and semantic features. In *Digital Image Computing: Techniques and Applications (DICTA)*, Gold Coast, QLD, December 2016.

[6] I. Abbasnejad, S. Sridharan, S. Denman, C. Fookes, and S. Lucey. From affine rank minimization solution to sparse modeling. In *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, pages 501–509. IEEE, 2017.

[7] I. Abbasnejad and D. Teney. A hierarchical bayesian network for face recognition using 2d and 3d facial data. In *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2015.

[8] M. E. Abbasnejad, E. V. Bonilla, and S. Sanner. Decision-theoretic sparsification for gaussian process preference learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 515–530. Springer, 2013.

[9] R. Arriaga, S. Vempala, et al. An algorithmic theory of learning: Robust concepts and random projection. In *Foundations of Computer Science, 1999. 40th Annual Symposium on*, pages 616–623. IEEE, 1999.

[10] M.-F. Balcan, A. Blum, and S. Vempala. Kernels as features: On kernels, margins, and low-dimensional mappings. *Machine Learning*, 65(1):79–94, 2006.

[11] R. G. Baraniuk and M. B. Wakin. Random projections of smooth manifolds. *Foundations of computational mathematics*, 9(1):51–77, 2009.

[12] E. Bingham and H. Mannila. Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 245–250. ACM, 2001.

[13] A. Blum. Random projection, margins, kernels, and feature-selection. In *Subspace, Latent Structure and Feature Selection*, pages 52–68. Springer, 2006.

[14] C. Boutsidis, A. Zouzias, and P. Drineas. Random projections for k-means clustering. In *Advances in Neural Information Processing Systems*, pages 298–306, 2010.

[15] R. Calderbank, S. Jafarpour, and R. Schapire. Compressed learning: Universal sparse dimensionality reduction and learning in the measurement domain. *preprint*, 2009.

[16] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *Information Theory, IEEE Transactions on*, 52(2):489–509, 2006.

[17] E. J. Candes and T. Tao. Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51(12):4203–4215, 2005.

[18] X. Cheng, C. Fookes, S. Sridharan, J. Saragih, and S. Lucey. Deformable face ensemble alignment with robust grouped-l1 anchors. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–7. IEEE, 2013.

[19] C. Cortes and V. Vapnik. Soft margin classifier, June 17 1997. US Patent 5,640,492.

[20] D. L. Donoho. Compressed sensing. *Information Theory, IEEE Transactions on*, 52(4):1289–1306, 2006.

[21] M. Fazel, H. Hindi, and S. P. Boyd. A rank minimization heuristic with application to minimum order system approximation. In *American Control Conference, 2001. Proceedings of the 2001*, volume 6, pages 4734–4739. IEEE, 2001.

[22] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[23] W. B. Johnson and J. Lindenstrauss. Extensions of lipschitz mappings into a hilbert space, 1984.

[24] R. Meka, P. Jain, C. Caramanis, and I. S. Dhillon. Rank minimization via online learning. In *Proceedings of the 25th International Conference on Machine learning*, pages 656–663. ACM, 2008.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[26] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma. Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(11):2233–2246, 2012.

[27] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.