
Compressed Learning: Universal Sparse Dimensionality Reduction and Learning in the Measurement Domain

Robert Calderbank
Electrical Engineering and Mathematics
Princeton University
calderbk@princeton.edu

Sina Jafarpour
Computer Science
Princeton University
sina@cs.princeton.edu

Robert Schapire
Computer Science
Princeton University
schapire@cs.princeton.edu

Abstract

In this paper, we provide theoretical results to show that *compressed learning*, learning directly in the compressed domain, is possible. In Particular, we provide tight bounds demonstrating that the linear kernel SVM's classifier in the measurement domain, with high probability, has true accuracy close to the accuracy of the best linear threshold classifier in the data domain. We show that this is beneficial both from the compressed sensing and the machine learning points of view. Furthermore, we indicate that for a family of well-known compressed sensing matrices, compressed learning is universal, in the sense that learning and classification in the measurement domain works provided that the data are sparse in some, even unknown, basis. Moreover, we show that our results are also applicable to a family of smooth manifold-learning tasks. Finally, we support our claims with experimental results.

1 Introduction

A general strategy for tackling many machine learning and signal processing problems is first transforming from the data domain to some appropriate measurement domain, and then perform the filtering or classification task in the measurement domain. Examples are transformation to the Fourier domain followed by low pass filtering, or kernel methods [1, 2] to go to higher domains and then use linear classification techniques. In many applications, the data can be represented as a sparse vector in a much higher dimensional space. Examples are images in the wavelet domain [3, 4], bag of words models for text classification and natural language processing [5, 6], sensor networks and wireless communication [7], and data streaming [8]. In this paper we show that whenever data have sparse representation, even in some unknown basis, compressed sensing can be used as an efficient one-to-one transform, preserving the learnability of the problem while bypassing the computational curse of dimensionality.

Compressed sensing [9, 10] is a novel and efficient data acquisition technique whenever the data are sparse in a high dimensional space. The idea is to replace the traditional

pointwise sampling with linear measurements, and postpone the measurement and compression cost until recovery time. More precisely, define a signal to be k -sparse if it has k non-zero entries. Note that the position of these non-zero entries are totally unknown, and two k -sparse signals can totally differ at the position of their non-zero entries. The only information about a k -sparse signal is that it has only k non-zero entries.

The goal of compressed sensing is then to provide an $m \times n$ measurement matrix A , with the number of measurements m as small as possible, together with a recovery algorithm Δ such that for any k -sparse signal $x \in \mathbb{R}^n$, Δ can recover x from its measurement $b = Ax$.

As a result, compressed sensing provides an efficient transform to the measurement domain, while the recovery algorithm provides a way to return to the data domain;

However, in many applications we would like to manipulate the data in the measurement domain. For instance, in radar applications, reconstruction of a signal from sensor data is not the ultimate goal. Rather one wishes to evaluate a function of the signal that, for example, indicates whether or not the signal is consistent with a target signature. This needs pattern recognition directly in the measurement domain.

Another application of classification in the measurement domain is the data streaming problem in which compressed sensing hardware, such as single pixel cameras [11], send compressed images to a central server with a high domain. The server, on the other hand, is only interested in a few types of signals with special patterns such as alarms and anomalies. The anomaly can be detected easily in the data domain based on changes at the wavelet coefficients; hence, we would like to be able to find these patterns directly in the compressed domain. Other applications are face detection [12, 13], and video streaming [14]. Also, in the bag of words model for natural language processing and text classification [5, 6], data has a sparse representation for which linear kernel SVM's are known to work well [15].

Being able to learn in the compressed domain is beneficial both in the compressed sensing and the machine learning points of view. From the compressed sensing view-point, it eliminates the abundant cost of recovering irrelevant data; in other words, classification in the measurement domain is like a sieve and makes it possible to only recover the desired signals, or even remove the recovery phase totally, if we are only interested in the results of classification. This is like finding a needle in a compressively sampled haystack

without recovering all the hays. From the machine learning view-point, compressed sensing can be regarded as efficient universal sparse dimensionality reduction from the data domain in \mathbb{R}^n to the measurement domain \mathbb{R}^m where $m \ll n$. The curse of dimensionality can be a big obstacle in machine learning threatening the accuracy and computation time of the classification task. Consequently, if the linear projection preserves the structure of the instance space and hence learnability, it can be used as a way to reduce the cost of the learning process perceptibly, while preserving the accuracy of the trained classifier approximately.

In this paper, we show using appropriate compressed sensing matrices, that if the data are approximately linearly separable in a high dimensional space, and the data has sparse representation even in some unknown basis, then compressed sensing approximately preserves the linear separability, and hence learnability. In other words, we provide theoretical bounds guaranteeing that if the data is measured directly in the compressed domain, a soft margin SVM's classifier that is trained based on the compressed data performs almost as well as the best possible classifier in the high domain. This is the first theoretical result for SVM's in measurement domain.

Previously Blum [16], and Balcan, Blum, and Vempala [17] used the idea of data-dependent linear projections for dimensionality reduction via the Johnson-Lindenstrauss lemma [18]; however, our results differ from those due to the following reasons. First, we provide one-to-one measurements, and as a result, data only needs to be measured in the compressed domain. There is no need to measure the data in the high dimensional domain, since we can recover the appropriate data directly later. This also reduces the amount of required storage extensively. The second difference is that we provide tight bounds on the hinge loss of the classifier in the low dimensional domain, in contrast to previous work which provides looser bounds on the performance of a projected classifier from high dimensions and not for the SVM's classifier in the low dimensional domain. Generally, it may be hard to find that classifier directly in the measurement domain without any access to the high dimensional domain. Especially if data is only approximately linearly separable, which is more realistic, those analyses totally ignore the non-separable fraction of the data which may affect the performance of the final SVM's classifier and the corresponding optimization program. Finally, our analysis is different and is based on the convex hinge loss instead of the high level margin loss.

In contrast to the previous work, the number of measurements in our work does not depend on the size of the training examples, it only has dependence on the sparsity k , and has logarithmic dependence on the dimension of the data domain n , which allows our dimensionality reduction techniques to scale up well, independent of the number of the training. Also, it is known that any matrix satisfying the Johnson-Lindenstrauss lemma also satisfies the near isometry property required for compressed learning [19]. However, the reverse is not true: there are some matrices, such as random Fourier matrices, that satisfy the near isometry property of compressed sensing, known as *the restricted isometry property*, but they do not satisfy the Johnson-Lindenstrauss

lemma; consequently, the results provided in this paper are more general.

Most of the previous work on using compressed sensing in machine learning is either experimental [12, 13], or it focuses on clustering and the principle component analysis problem [20, 21], or only deals with lazy classification algorithms such as nearest neighborhood classifiers [22]. In this paper, we provide tight PAC-style bounds relating the performance of the linear-kernel SVM's classifier in the measurement domain to the data domain. Although our analysis is only for linear kernel, it is consistent, and can be adapted with the method of Blum et. al [16] for building explicit feature spaces from the similarity functions. The reason is that for well-designed similarity functions, the explicit feature space of Blum et. al, is only based on the similarity of the new example with the training set, and hence, has sparse representation. Finally, we provide experimental results supporting our main theorems.

As a final remark, we emphasize that although most SVM's optimization algorithms exploit sparsity, in reality, managing very high-dimensional sparse data is a challenging problem, and in fact, the final goal of compressed sensing in general, and compressed learning is to remove this difficulty. Moreover, in Section 6 we show that compressed learning is possible even if the only knowledge we have is that data is sparse even in some unknown basis. Of course, in this case the data domain may not be sparse at all.

The organization of the remaining of the paper is as follows: Section 2 introduces the notations used in the paper. Section 2.1 reviews the SVM's classifiers and their properties need in this paper. Section 3 defines the compressed learning problem and its benefits and advantages in details, especially its universality with respect to unknown bases. Section 4 provides a generalized near isometry property required as a key lemma in this paper. Then Section 5 proves the main theorem of this paper, demonstrating that compressed sensing is theoretically possible, and the SVM's classifier in the measurement domain works approximately as well as the best classifier in the data domain. Section 7 provides experimental results supporting the claims in this paper. Section 8 concludes the paper.

2 Notations

In this paper, we assume that all the data are vectors in \mathbb{R}^n , k -sparse, and their ℓ_2 norm is bounded by some parameter R . Throughout the paper, we show vectors with bold symbols. Hence, the instance space \mathcal{X} is:

$$\mathcal{X} = \{(\mathbf{x}, y) : \mathbf{x} \in \mathbb{R}^n, \|\mathbf{x}\|_0 \leq k, \|\mathbf{x}\|_2 \leq R, y \in \{-1, 1\}\},$$

which we call the *data domain*. Let $A_{m \times n}$ be the linear measurement matrix used in compressed sensing. We define the *measurement domain* \mathcal{M} as:

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

In other words, \mathcal{M} is a compressed representation of \mathcal{X} . We always show data in data domain with bold vectors like \mathbf{x} , and in measurement domain with $A\mathbf{x}$.

We assume \mathcal{D} is some unknown distribution over \mathcal{X} , and

$$S = \langle (\mathbf{x}_1, y_1), \dots, (\mathbf{x}_M, y_M) \rangle$$

is a set of M labeled examples iid from D . Since A is a one-to-one mapping from \mathcal{X} to \mathcal{M} , the measurement domain has the same probability distribution \mathcal{D} as the data domain, and

$$AS = \langle (A\mathbf{x}_1, y_1), \dots, (A\mathbf{x}_M, y_M) \rangle$$

is the compressed representation of

$$S = \langle (\mathbf{x}_1, y_1), \dots, (\mathbf{x}_M, y_M) \rangle.$$

Finally, note that any linear threshold classifier $w(x)$, corresponds to a vector $\mathbf{w} \in \mathbb{R}^n$, such that

$$w(x) = \text{sign}(\mathbf{w}^\top \mathbf{x}).$$

2.1 Soft Margin SVM's Classifiers

Structural risk minimization (SRM) [23, 24] is an inductive principle for model selection used for learning from finite training datasets, in order to prevent the problem of overfitting to the training examples. Support vector machines (SVM's) [25, 2, 23] form a linear threshold classifier with maximum margin and consistent with the training examples. As mentioned before, any linear threshold classifier $w(\mathbf{x})$ corresponds to a vector $\mathbf{w} \in \mathbb{R}^n$ such that $w(\mathbf{x}) = \text{sign}(\mathbf{w}^\top \mathbf{x})$; as a result, in this paper we mention the linear threshold classifiers with their corresponding vectors. Also for simplicity we only focus on classifiers passing through the origin. The results can be simply extended to the general case.

Whenever the training examples are not linearly separable soft margin SVM's are used. The idea is to simultaneously maximize the margin and minimize the empirical hinge loss. More precisely let

$$H(x) = \max\{0, 1 + x\},$$

and $\langle (\mathbf{x}_1, y_1), \dots, (\mathbf{x}_M, y_M) \rangle$ be a set of M labeled instances sampled i.i.d from some distribution \mathcal{D} . For any linear classifier $\mathbf{w} \in \mathbb{R}^n$ we define its true hinge loss as

$$H_D(\mathbf{w}) = \mathbf{E}_{(\mathbf{x}, y) \sim \mathcal{D}}[1 - y\mathbf{w}^\top \mathbf{x}],$$

and its empirical hinge loss

$$\hat{H}_S(\mathbf{w}) = \mathbf{E}_{(\mathbf{x}_i, y_i) \sim S}[1 - y_i \mathbf{w}^\top \mathbf{x}_i].$$

We also define the true regularization loss of a classifier \mathbf{w} as

$$L(\mathbf{w}) = H_D(\mathbf{w}) + \frac{1}{2C} \|\mathbf{w}\|^2,$$

and the empirical regularization loss

$$\hat{L}(\mathbf{w}) = \hat{H}_S(\mathbf{w}) + \frac{1}{2C} \|\mathbf{w}\|^2. \quad (1)$$

Soft margin SVM's then minimize the empirical regularization loss which is a convex optimization program.

The following Theorem is a direct consequence of the convex duality, and we provide its proof for completeness.

Theorem 2.1 *Let*

$$S = \langle (\mathbf{x}_1, y_1), \dots, (\mathbf{x}_M, y_M) \rangle$$

be a set of M examples chosen i.i.d from some distribution \mathcal{D} , and let \mathbf{w} be the SVM's classifier obtained by minimizing Equation (1). Then

$$\mathbf{w} = \sum_{i=1}^M \alpha_i y_i \mathbf{x}_i,$$

where

$$\forall i : 0 \leq \alpha_i \leq \frac{C}{M}$$

and

$$\|\mathbf{w}\|^2 \leq C.$$

Proof: The optimization problem of Equation (1) can be written as

$$\begin{aligned} \text{minimize } & \frac{1}{2} \|\mathbf{w}\|^2 + \frac{C}{M} \sum_{i=1}^M \xi_i \text{ such that:} \\ & \forall i : \xi_i \geq 0 \\ & \forall i : \xi_i \geq 1 - y_i \mathbf{w}^\top \mathbf{x}_i \end{aligned} \quad (2)$$

We form the Lagrangian of the above optimization problem by linearly combining the objective function and the constraints $\mathcal{L}(\mathbf{w}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\eta}) =$

$$\frac{1}{2} \|\mathbf{w}\|^2 + \frac{C}{M} \sum_{i=1}^M \xi_i + \sum_{i=1}^M \alpha_i (1 - y_i \mathbf{w}^\top \mathbf{x}_i - \xi_i) - \sum_{i=1}^M \eta_i \xi_i \quad (3)$$

The KKT conditions are

$$\begin{aligned} \mathbf{w} - \sum_{i=1}^M \alpha_i y_i \mathbf{x}_i &= 0, \\ \forall i : \frac{C}{M} - \alpha_i - \eta_i &= 0, \\ \forall i : \alpha_i \geq 0, \eta_i &\geq 0. \end{aligned}$$

Hence, the dual program is

$$\begin{aligned} \text{maximize } & \sum_{i=1}^M \alpha_i - \frac{1}{2} \sum_{i,j=1}^M \alpha_i \alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j \text{ such that:} \\ & \forall i : 0 \leq \alpha_i \leq \frac{C}{M} \end{aligned}$$

Now the duality theorem implies:

$$\begin{aligned} \frac{1}{2} \|\mathbf{w}\|^2 &\leq \frac{1}{2} \|\mathbf{w}\|^2 + \frac{C}{M} \sum_{i=1}^M \xi_i \\ &= \sum_{i=1}^M \alpha_i - \frac{1}{2} \sum_{i,j=1}^M \alpha_i \alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j \\ &= \sum_{i=1}^M \alpha_i - \frac{1}{2} \|\mathbf{w}\|^2 \leq C - \frac{1}{2} \|\mathbf{w}\|^2. \end{aligned}$$

Hence

$$\|\mathbf{w}\|^2 \leq C. \quad \blacksquare$$

We denote any classifier in high dimensional space by \mathbf{w} , and any classifier in the low dimensional space by \mathbf{z} . Let \mathbf{w}^* be the classifier in the data domain that minimizes the true regularization loss in the data domain, i.e.

$$\mathbf{w}^* \doteq \arg_{\mathbf{w}} \min L(\mathbf{w}), \quad (4)$$

and let z^* be the classifier in the measurement domain that minimizes the true regularization loss in the measurement domain, i.e.

$$z^* \doteq \arg_z \min L(z). \quad (5)$$

Also let \hat{w}_S be the soft margin classifier trained with a training set S in the data domain, i.e.

$$\hat{w}_S \doteq \arg_w \min \hat{L}_S(w),$$

and similarly \hat{z}_{AS} is the soft margin classifier trained with the compressed training set AS in the measurement domain:

$$\hat{z}_{AS} \doteq \arg_z \min \hat{L}_{AS}(z).$$

Finally, let $A\hat{w}_S$ be the classifier in the measurement domain obtained by projecting \hat{w}_S from the data domain to the measurement domain. Although we may not be able to find $A\hat{w}_S$ directly from the data in the measurement domain, we use it in the analysis to show that \hat{z}_{AS} has true loss almost the same as the best classifier in the data domain. Note that $L(Aw^*)$ is :

$$\mathbf{E}_{(x,y) \sim \mathcal{D}} [H(1 - y(Aw^*)^\top(Ax))] + \frac{1}{2C} \|Aw^*\|^2.$$

To provide tight bounds on the performance of the SVM's classifier \hat{z}_{AS} in the measurement domain, we perform the following *oracle* analysis [26]: We assume there is some good low-norm predictor $w_0 \in \mathbb{R}^n$ which achieves a good generalization error (expected hinge loss) of $H_D(w_0)$ and has norm $\|w_0\|_2$. On the other hand, we only receive the examples compressively, i.e. in the measurement domain; We train an SVM's classifier \hat{z}_{AS} by minimizing the empirical regularization loss \hat{L}_{AS} . In the following sections we show that, for a family of well-known matrices used in compressed sensing, this SVM's classifier has true hinge loss close to hinge loss of the good classifier $w_0(x)$ in the high dimensional space.

3 Compressed Learning

Compressed learning is learning directly in the measurement domain. Without paying the cost of recovering back the data to the high dimensional data domain. We show that compressed learning with soft margin SVM's classifiers is possible. This means that if the instances are provided directly in the measurement domain, the soft margin SVM's classifier trained with compressively sampled training data has almost the same performance as the best possible classifier in high dimensional space; therefore, that we do not need to recover the training examples, and since there exist efficient compressed sensing devices capable of sensing data directly in the measurement domain, the total classification task can be done in the measurement domain. In addition to recovery time preserving, compressed learning with soft margin SVM's has another advantage. It eliminates the computational cost of learning in high dimensional domain, which might be costly if the dimension of the data domain is very high, and the data is sparse in some other, even unknown, basis. As a result, the capability of classifying the data in the measurement domain, is equivalent to bypassing the computational curse of dimensionality of learning in the data domain.

The main result of this paper states that, if the dataset is compressively sampled and hence is directly presented in the measurement domain, then with high probability, the SVM's classifier trained over the training set has generalization error close to the generalization error of the best classifier in the data domain. We use a hybrid argument to prove this claim. We show that if we project the SVM's classifier of the data domain to the measurement domain, because of the near-isometry property of the compressed sensing, the generalization error of the resulting classifier is close to the generalization error of the SVM's classifier in the data domain. Of course, it may not be possible to find this projected classifier directly using the compressed data; however, we show that the SVM's classifier trained on the compressed training data in the measurement domain, has also generalization error close to the error of the intermediate, projected classifier. Figure 1 shows the hybrid argument used to prove the accuracy of the SVM's classifier in measurement domain.

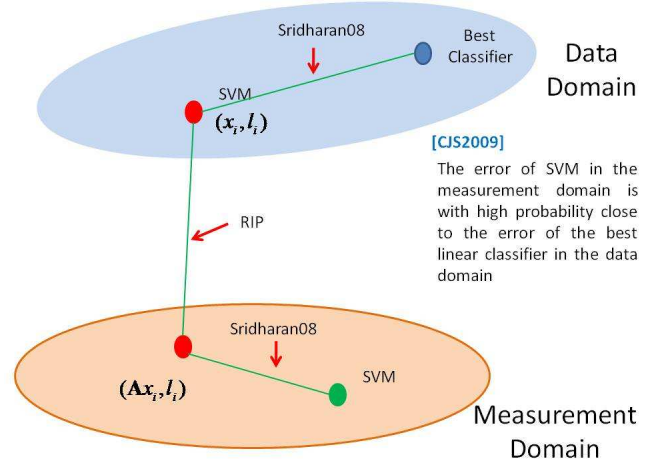


Figure 1: The hybrid argument proves that the SVM's classifier in the measurement domain has generalization regularization loss close to the loss of the best classifier in the data domain.

Note that we only use the classifier $A\hat{w}_S$ in our hybrid analysis. Generally, it may not be easy to find this classifier directly in the measurement domain without first recovering the examples, then finding the SVM's classifier in high domain, and finally projecting it back to the measurement domain.

The following Theorem is the main result of this paper. We prove the theorem in Section 5.

Theorem 3.1 (Compressed Learning) *Let A $m \times n$ be the compressed sensing matrix, which acts as a near-isometry on any $2k$ -sparse vector. That is for any $2k$ -sparse vector $x \in \mathbb{R}^n$:*

$$(1 - \epsilon)\|x\|^2 \leq \|Ax\|^2 \leq (1 + \epsilon)\|x\|^2.$$

Let

$$AS = \langle (Ax_1, y_1), \dots, (Ax_M, y_M) \rangle$$

be i.i.d instances, compressively sampled from some distribution \mathcal{D} , and $\hat{\mathbf{z}}_{AS}$ be the soft-margin SVM's trained on AS. Also let \mathbf{w}_0 be the best linear classifier in the data domain, with low Hinge loss, and large margin (hence small $\|\mathbf{w}_0\|_2$). Then with probability $1 - 2\delta$ over AS:

$$H_{\mathcal{D}}(\hat{\mathbf{z}}_{AS}) \leq H_{\mathcal{D}}(\mathbf{w}_0) + O\left(\sqrt{\|\mathbf{w}_0\|^2 \left(R^2\epsilon + \frac{\log(1/\delta)}{M}\right)}\right).$$

This bound states that the deviation of the hinge loss of $\hat{\mathbf{z}}_{AS}$ from $H(\mathbf{w}_0)$ is at most $O\left(\sqrt{\|\mathbf{w}_0\|^2 \left(R^2\epsilon + \frac{\log(1/\delta)}{M}\right)}\right)$. In case of using projection, and hence complete isometry ($\epsilon = 0$), this bound agrees with the results of the data-laden analysis provided by [26]. On the other hand, if the space dimension n is extremely large, which is the main assumption in compressed sensing, we are efficiently reducing the dimensionality to $O(k \log n)$ while only imposing $O(\sqrt{\epsilon})$ error on the performance of the classifier.

Note that if the data are compressively sampled with very small number of measurements, the distortion factor ϵ becomes large, and hence SVM's classifier becomes a weak learner. The performance of the classifier then can be improved using boosting techniques [27]. This is aligned with the main goal of compressed sensing. We can reduce the number of required measurement, and then compensate this by the computational cost of improving the obtained classifier.

Our results can also be extended to the manifold learning problem. Recently, Baraniuk and Wakin [28], and Hegde, Wakin and Baraniuk [22] showed that random projections also satisfy (k, ϵ) property over k -dimensional smooth manifolds in \mathbb{R}^n . Hence, the results of Theorem (3.1) are also valid if the instances lie on a k -dimensional smooth manifold, instead of just being sparse. This means that the SVM's classifier, obtained by projecting the manifold to the compressed domain, acts almost as well as the SVM's classifier in the data domain.

4 Generalized Restricted Isometry Property

In this section, we show that the regularization loss of the SVM's classifier in the data domain does not change so much by projection to the measurement domain. One class of measurement matrices that are widely used in compressed sensing [29, 10, 30, 31] are the matrices that satisfy the *Restricted Isometry Property* [29, 19]:

Definition 1 (Restricted Isometry Property) *A $m \times n$ sensing matrix A satisfies the restricted isometry property, (k, ϵ) -RIP, if it acts as a near-isometry with distortion factor ϵ , over all k -sparse vectors. In other words, for any k -sparse vector $\mathbf{x} \in \mathbb{R}^n$ the following near-isometry property holds:*

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

The following theorem by Candes and Tao [9], and Baraniuk et. al [19] shows that a large family of random matrices satisfy the restricted isometry property:

Theorem 4.1 *If the entries of $\sqrt{m}A$ are sampled i.i.d from either*

- *Gaussian distribution : $\mathcal{N}(0, 1)$, or*
- *Bernoulli distribution : $U(-1, 1)$,*

and $m = \Omega(k \log(n/k))$ then except with probability $e^{-c(\epsilon)m}$, A satisfies the restricted isometry property with parameters (k, ϵ) .

Throughout this paper, we assume A is a linear measurement matrix satisfying (k, ϵ) -RIP.

We plan to show that the regularization error of the SVM's classifier is preserved by projection. Since SVM's classifier is a linear combination of support vectors, we have to show that a near isometry property also holds for linear combinations of the sparse vectors. Our first theorem shows that if a matrix satisfies RIP, the inner product between linear combinations of arbitrary sparse signals is also approximately preserved by linear projection, up to an additive constant. SVM's classification is based on the inner product of the classifier and the example. At the first step, we show that if a matrix satisfies $(2k, \epsilon)$ -RIP, then not only the ℓ_2 norm, but also the inner product between any two k -sparse vector is approximately preserved. This property is also provided in [32]:

Lemma 4.2 *Let $A_{m \times n}$ be the measurement matrix satisfying $(2k, \epsilon)$ -RIP, and \mathbf{x}, \mathbf{x}' be two k -sparse vectors in \mathbb{R}^n , such that $\|\mathbf{x}\|_2 \leq R, \|\mathbf{x}'\|_2 \leq R$. Then*

$$(1 + \epsilon)\mathbf{x}^T \cdot \mathbf{x}' - 2R^2\epsilon \leq (A\mathbf{x})^T (A\mathbf{x}'). \quad (6)$$

Proof: Since \mathbf{x}, \mathbf{x}' are k -sparse, by triangle inequality their difference is a $2k$ -sparse vector. That is:

$$\|\mathbf{x} - \mathbf{x}'\|_0 \leq \|\mathbf{x}\|_0 + \|\mathbf{x}'\|_0 \leq k + k = 2k.$$

Hence the restricted isometry property applies to $\mathbf{x} - \mathbf{x}'$:

$$\begin{aligned} \|A(\mathbf{x} - \mathbf{x}')\|^2 &\leq (1 + \epsilon)\|\mathbf{x} - \mathbf{x}'\|^2 \quad (7) \\ &= (1 + \epsilon)(\|\mathbf{x}\|^2 + \|\mathbf{x}'\|^2 - 2\mathbf{x}^T \mathbf{x}'). \end{aligned}$$

Also

$$\begin{aligned} (1 - \epsilon)(\|\mathbf{x}\|^2 + \|\mathbf{x}'\|^2) - 2(A\mathbf{x})^T (A\mathbf{x}') \\ \leq \|A\mathbf{x}\|^2 + \|A\mathbf{x}'\|^2 - 2(A\mathbf{x})^T (A\mathbf{x}') \\ = \|A(\mathbf{x} - \mathbf{x}')\|^2. \quad (8) \end{aligned}$$

Putting Equations (7), and (8) together, and noting that $\|\mathbf{x}\| \leq R, \|\mathbf{x}'\| \leq R$ completes the proof. ■

Lemma 4.3 *Let $A_{m \times n}$ be the measurement matrix satisfying $(2k, \epsilon)$ -RIP, and \mathbf{x}, \mathbf{x}' be two k -sparse vectors in \mathbb{R}^n , such that $\|\mathbf{x}\|_2 \leq R, \|\mathbf{x}'\|_2 \leq R$. Then*

$$(A\mathbf{x})^T (A\mathbf{x}') \leq (1 - \epsilon)\mathbf{x}^T \mathbf{x}' + 2R^2\epsilon. \quad (9)$$

Proof: The proof is very similar to the proof of Lemma 4.2. Since \mathbf{x}, \mathbf{x}' are k -sparse $\mathbf{x} - \mathbf{x}'$ is $2k$ -sparse and hence by RIP property

$$\begin{aligned} \|A(\mathbf{x} - \mathbf{x}')\|^2 &\geq (1 - \epsilon)\|\mathbf{x} - \mathbf{x}'\|^2 \quad (10) \\ &= (1 - \epsilon)(\|\mathbf{x}\|^2 + \|\mathbf{x}'\|^2 - 2\mathbf{x}^T \mathbf{x}'). \end{aligned}$$

Also

$$\begin{aligned} & \|A(\mathbf{x} - \mathbf{x}')\|^2 \\ &= \|A\mathbf{x}\|^2 + \|A\mathbf{x}'\|^2 - 2(A\mathbf{x})^\top(A\mathbf{x}') \\ &\leq (1 + \epsilon) (\|\mathbf{x}\|^2 + \|\mathbf{x}'\|^2) - 2(A\mathbf{x})^\top(A\mathbf{x}'). \end{aligned} \quad (11)$$

Putting Equations (10) and (11) together, and noting that $\|\mathbf{x}\| \leq R, \|\mathbf{x}'\| \leq R$ completes the proof. \blacksquare

So we have shown that the restricted isometry property approximately preserves the inner product between any two k -sparse signal. Now we generalize this claim. We show that the restricted isometry property, also approximately preserves the inner product between any two vectors from the convex hull of the set of sparse vectors. This is crucial, because by Theorem 2.1, the normalized SVM's classifier $\frac{1}{\sqrt{C}}\hat{w}_S(x)$ is a member of this convex hull.

The following Theorem is a direct consequence of Lemmas 4.2 and 4.3, and generalizes the generalized RIP of [32] to contain the linear combinations and convex hull of sparse signals:

Theorem 4.4 *Let $A_{m \times n}$ be a matrix satisfying $(2k, \epsilon)$ -RIP. Let M, N be two integers, and*

$$(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_M, y_M), (\mathbf{x}'_1, y'_1), \dots, (\mathbf{x}'_N, y'_N) \in \mathcal{X}.$$

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

$$\begin{aligned} \boldsymbol{\alpha} &= \sum_{i=1}^M \alpha_i y_i \mathbf{x}_i \\ \boldsymbol{\beta} &= \sum_{j=1}^N \beta_j y'_j \mathbf{x}'_j. \end{aligned}$$

Then:

$$|\boldsymbol{\beta}^\top \boldsymbol{\alpha} - (A\boldsymbol{\beta})^\top A\boldsymbol{\alpha}| \leq 3CDR^2\epsilon.$$

Proof: It is easy to see that A is a linear projection operator, and $\boldsymbol{\alpha}, \boldsymbol{\beta}$ are linear combinations of sparse signals. So one can spread the inner product, and then use the RIP property over the inner product of sparse vectors. More precisely, since A is a linear transform, we have:

$$A\boldsymbol{\alpha} = \sum_{i=1}^M \alpha_i y_i (A\mathbf{x}_i),$$

and

$$A\boldsymbol{\beta} = \sum_{j=1}^N \beta_j y'_j (A\mathbf{x}'_j).$$

As a result

$$\begin{aligned} (A\boldsymbol{\alpha})^\top(A\boldsymbol{\beta}) &= \sum_{i=1}^M \sum_{j=1}^N \alpha_i \beta_j y_i y'_j (A\mathbf{x}_i)^\top(A\mathbf{x}'_j) \\ &= \sum_{y_i=y'_j} \alpha_i \beta_j (A\mathbf{x}_i)^\top(A\mathbf{x}'_j) \\ &\quad - \sum_{y_i \neq y'_j} \alpha_i \beta_j (A\mathbf{x}_i)^\top(A\mathbf{x}'_j) \end{aligned} \quad (12)$$

Now since $\alpha_i, \beta_j \geq 0$, using Lemmas 4.2 and 4.3 we get:

$$\alpha_i \beta_j (A\mathbf{x}_i)^\top(A\mathbf{x}'_j) \leq \alpha_i \beta_j ((1 - \epsilon)\mathbf{x}_i^\top \mathbf{x}'_j + 2R^2\epsilon),$$

and

$$\alpha_i \beta_j (A\mathbf{x}_i)^\top(A\mathbf{x}'_j) \geq \alpha_i \beta_j ((1 + \epsilon)\mathbf{x}_i^\top \mathbf{x}'_j - 2R^2\epsilon).$$

Putting these into Equation (12), we get:

$$\begin{aligned} & \sum_{y_i=y'_j} \alpha_i \beta_j (A\mathbf{x}_i)^\top(A\mathbf{x}'_j) \\ & - \sum_{y_i \neq y'_j} \alpha_i \beta_j (A\mathbf{x}_i)^\top(A\mathbf{x}'_j) \\ & \leq \sum_{y_i=y'_j} \alpha_i \beta_j ((1 - \epsilon)\mathbf{x}_i^\top \mathbf{x}'_j + 2R^2\epsilon) \\ & - \sum_{y_i \neq y'_j} \alpha_i \beta_j ((1 + \epsilon)\mathbf{x}_i^\top \mathbf{x}'_j - 2R^2\epsilon) \\ & = \sum_{i,j} \alpha_i \beta_j y_i y'_j (\mathbf{x}_i^\top \mathbf{x}'_j) \\ & + \sum_{i,j} \alpha_i \beta_j \epsilon (2R^2 + \mathbf{x}_i^\top \mathbf{x}'_j) \\ & \leq \boldsymbol{\alpha}^\top \boldsymbol{\beta} + 3R^2\epsilon \sum_{i=1}^M \alpha_i \sum_{j=1}^N \beta_j \\ & \leq \boldsymbol{\alpha}^\top \boldsymbol{\beta} + 3R^2CD\epsilon. \end{aligned}$$

The other side of the absolute value,

$$\boldsymbol{\alpha}^\top \boldsymbol{\beta} - 3R^2CD\epsilon \leq (A\boldsymbol{\alpha})^\top(A\boldsymbol{\beta})$$

can also be proved very similarly. \blacksquare

5 Compressed Learning is Possible

In this section we show that compressed learning is possible. We use a hybrid argument to show that the SVM's classifier in the measurement domain has almost the same performance as the best classifier in data domain. Theorem (2.1) together with the theory of structural risk minimization implies that if the data were represented in high dimensional space, the high dimensional SVM's classifier \hat{w}_S had almost the same performance as the best classifier \mathbf{w}_0 . Then the generalized RIP (Theorem 4.4), implies that if we project the SVM's classifier \hat{w}_S to measurement domain, the true regularization loss of the classifier $A\hat{w}_S$ is almost the same as the true regularization loss of the SVM's classifier \hat{w}_S in high domain. Again, we emphasize that we only use the projected classifier $A\hat{w}_S$ in the analysis. The compressed learning algorithm only uses SVM's classifier in the measurement domain.

The following lemma is the heart of our result, and connects the regularization loss of two classifiers, one in the data domain, and one in the measurement domain. The classifier in the data domain is the SVM's classifier. So this lemma states that if the SVM's classifier in the data domain has good performance, there exists a linear threshold classifier in the measurement domain with high performance. Later we show that this implies that the SVM's classifier in the measurement domain has performance, close to the best classifier in the data domain.

Lemma 5.1 Let $A_{m \times n}$ satisfy $(2k, \epsilon)$ -RIP. Also let

$$S = \langle (\mathbf{x}_1, y_1), \dots, (\mathbf{x}_M, y_M) \rangle$$

be the the training set of size M , where each example is sampled i.i.d from some distribution \mathcal{D} in data domain. Let $\hat{\mathbf{w}}_S$ be the soft-margin SVM's trained on S , and $A\hat{\mathbf{w}}_S$ be the vector in the measurement domain, obtained by projecting down $\hat{\mathbf{w}}_S$. Then

$$L_{\mathcal{D}}(A\hat{\mathbf{w}}_S) \leq L_{\mathcal{D}}(\hat{\mathbf{w}}_S) + O(CR^2\epsilon).$$

Proof: By Theorem 2.1, the soft margin SVM's classifier is a linear combination of the support vectors

$$\hat{\mathbf{w}}_S = \sum_{i=1}^M \alpha_i y_i \mathbf{x}_i,$$

where α_i are positive numbers such that $\sum_{i=1}^M \alpha_i \leq C$. As a result, using Theorem 4.4 with $N = M$, $D = C$, $(\mathbf{x}'_i, y'_i) = (\mathbf{x}_i, y_i)$, we get:

$$(A\mathbf{w})^\top (A\mathbf{w}) \leq \mathbf{w}^\top \mathbf{w} + 3C^2R^2\epsilon. \quad (13)$$

Dividing both sides of Equation (13) by $2C$ implies:

$$\frac{1}{2C} \|A\hat{\mathbf{w}}_S\|^2 \leq \frac{1}{2C} \|\hat{\mathbf{w}}_S\|^2 + O(CR^2\epsilon).$$

Now we show that

$$H_{\mathcal{D}}(A\hat{\mathbf{w}}_S) \leq H_{\mathcal{D}}(\hat{\mathbf{w}}_S) + O(CR^2\epsilon).$$

Fix $(\mathbf{x}, y) \in \mathcal{X}$. Theorem (4.4) with $N = 1$, $D = 1$, $(\mathbf{x}'_1, y'_1) = (\mathbf{x}, y)$ implies that:

$$1 - y(A\hat{\mathbf{w}}_S)^\top (A\mathbf{x}) \leq 1 - y\hat{\mathbf{w}}_S^\top \mathbf{x} + O(CR^2\epsilon) \quad (14)$$

Now since $1 - y\hat{\mathbf{w}}_S^\top \mathbf{x} \leq H(-y\hat{\mathbf{w}}_S^\top \mathbf{x})$, and the right-hand size of Equation (14) is always positive, we get:

$$H_{\mathcal{D}}(-y(A\hat{\mathbf{w}}_S)^\top (A\mathbf{x})) \leq H(-y\hat{\mathbf{w}}_S^\top \mathbf{x}) + O(CR^2\epsilon) \quad (15)$$

Now since the measurement matrix forms a one-to-one mapping from the data domain to the measurement domain, taking the expectation of Equation (15) with respect to \mathcal{D} completes the proof. \blacksquare

Up to now, we have shown how smoothly the regularization loss of the SVM's classifier $\hat{\mathbf{w}}_S$ changes with projection. Next we show that the regularization loss of the SVM's classifier $\hat{\mathbf{z}}_{AS}$ in measurement domain is close to the regularization loss of $A\hat{\mathbf{w}}_S$; and the regularization loss of the SVM's classifier $\hat{\mathbf{w}}_S$ in data domain is close to the regularization loss of the oracle best classifier \mathbf{w}_0 in data domain. This completes our hybrid argument. We show this in the query model, and using a recent result by Sridharan et. al [33]:

Theorem 5.2 (Sridharan 2008 [33, 26]) For all \mathbf{w} with $\|\mathbf{w}\|^2 \leq 2C$, with probability at least $1 - \delta$ over training set:

$$L_{\mathcal{D}}(\mathbf{w}) - L_{\mathcal{D}}(\mathbf{w}^*) \leq 2 \left[\hat{L}_S(\mathbf{w}) - \hat{L}_S(\mathbf{w}^*) \right]_+ + O\left(\frac{C \log(1/\delta)}{M}\right) \quad (16)$$

Corollary 1 Let $\hat{\mathbf{w}}_S$ be the SVM's classifier. Then with probability $1 - \delta$

$$L_{\mathcal{D}}(\hat{\mathbf{w}}_S) \leq L_{\mathcal{D}}(\mathbf{w}^*) + O\left(\frac{C \log(1/\delta)}{M}\right).$$

Proof: Theorem 2.1 implies that $\|\hat{\mathbf{w}}_S\|^2 \leq C$, also the soft margin SVM's classifier minimizes the empirical regularization loss. So

$$\hat{L}_S(\hat{\mathbf{w}}_S) \leq \hat{L}_S(\mathbf{w}^*).$$

Combining these with Theorem 5.2 completes the proof. \blacksquare

Remark 5.3 Although we used Theorem 5.2, we mention that Corollary 1 can also be directly proved using the theory of Rademacher complexity [34].

Now we are ready to prove Theorem 2.1. This Theorem is the main result of this paper, and states that by training the SVM's classifier in the measurement domain, one can obtain accuracy close to the accuracy of the SVM's classifier in the data domain. **Proof:** By definition of the regularization loss we have:

$$H(\hat{\mathbf{z}}_{AS}) \leq H(\hat{\mathbf{z}}_{AS}) + \frac{1}{2C} \|\hat{\mathbf{z}}_{AS}\|^2 = L(\hat{\mathbf{z}}_{AS})$$

Corollary 1 states that the regularization loss of the SVM's classifier in the measurement domain, is close to the regularization loss of the best classifier in the measurement domain \mathbf{z}^* .

$$L_{\mathcal{D}}(\hat{\mathbf{z}}_{AS}) \leq L_{\mathcal{D}}(\mathbf{z}^*) + O\left(\frac{C \log(1/\delta)}{M}\right)$$

By definition of \mathbf{z}^* , Equation (5), \mathbf{z}^* is the best classifier in the measurement domain. So

$$L_{\mathcal{D}}(\mathbf{z}^*) \leq L_{\mathcal{D}}(A\hat{\mathbf{w}}_S)$$

Theorem (5.1) connects the regularization loss of the SVM's classifier in the data domain, $\hat{\mathbf{w}}_S$, to the regularization loss of its projected vector $A\hat{\mathbf{w}}_S$.

$$L_{\mathcal{D}}(A\hat{\mathbf{w}}_S) \leq L_{\mathcal{D}}(\hat{\mathbf{w}}_S) + O(CR^2\epsilon)$$

Corollary (1), now applied in the data domain, connects the regularization loss of the SVM's classifier $\hat{\mathbf{w}}_S$ to the regularization loss of the best classifier \mathbf{w}^* .

$$L_{\mathcal{D}}(\hat{\mathbf{w}}_S) \leq L_{\mathcal{D}}(\mathbf{w}^*) + O\left(\frac{C \log(1/\delta)}{M}\right)$$

By definition of \mathbf{w}^* (Equation (4)), for any classifier \mathbf{w} , in the data domain:

$$L_{\mathcal{D}}(\mathbf{w}^*) \leq L_{\mathcal{D}}(\mathbf{w})$$

In particular, let \mathbf{w}_0 be the good classifier in the query model, with small true Hinge loss and large margin. Then

$$L_{\mathcal{D}}(\mathbf{w}^*) \leq L_{\mathcal{D}}(\mathbf{w}_0)$$

Putting all inequalities together, we get:

$$H_{\mathcal{D}}(\hat{\mathbf{z}}_{AS}) \leq H_{\mathcal{D}}(\mathbf{w}_0) + \frac{1}{2C} \|\mathbf{w}_0\|^2 + O\left(CR^2\epsilon + \frac{C \log(1/\delta)}{M}\right) \quad (17)$$

Equation (17) is valid for any C . By choosing a C which minimizes the Equation (17), we get:

$$H_{\mathcal{D}}(\hat{z}_{AS}) \leq H_{\mathcal{D}}(\mathbf{w}_0) + O\left(\sqrt{\|\mathbf{w}_0\|^2 \left(R^2\epsilon + \frac{\log(1/\delta)}{M}\right)}\right) \quad (18)$$

6 Universality and Unknown Sparse Bases

Now, we focus on the more general case in which the data is not sparse in the observed domain. However, there exists a possibly unknown basis such that data can be represented sparsely in that domain. A well-known example of this case are images. It is known that images have sparse representation in the wavelet domain [3]; however, there exists efficient compressed sensing hardware, single pixel camera [11], which can project the high resolution images from the high dimensional pixel domain to the low dimensional measurement domain. More precisely, now there exists an orthonormal basis Ψ such that each instance can be represented as

$$\mathbf{x} = \Psi \mathbf{s},$$

where \mathbf{s} is k -sparse. Consequently, the instance space \mathcal{X}_{Ψ} can be defined as $\mathcal{X}_{\Psi} = \{(\mathbf{x}, y)\}$ such that

$$\mathbf{x} \in \mathbb{R}^n, \mathbf{x} = \Psi \mathbf{s}, \|\mathbf{s}\|_0 \leq k, \|\mathbf{s}\|_2 \leq R, y \in \{-1, 1\}.$$

Note that since Ψ is orthonormal we have:

$$\|\mathbf{x}\|_2 = \|\mathbf{s}\|_2.$$

Consequently, the measurement domain is

$$\mathcal{M} = \{(A\mathbf{x}, y)\} = \{(A\Psi\mathbf{s}, y)\}.$$

Let $\Phi_{m \times n} = A\Psi$. Then the measurement domain can be written as

$$\mathcal{M} = \{(\Phi\mathbf{x}, y)\}.$$

Hence, Φ maps the data from the sparse domain to the measurement domain. Although, at measurement time, we are not in the sparse domain, and we only deal with sensing matrix Ψ , at recovery time, Φ helps to recover the sparse representation. In order to be able to recover the data successfully, Ψ should be known at recovery time. Generally it is not possible to recover a signal without knowing the sparsity basis Ψ .

As a result, if we only care about classifying instances and not recovering them, there is no need to know the sparsity basis Φ . We still use the Gaussian matrix A for measurement. However, in this case data domain is not sparse. In compressed sensing, we measure the data with the sensing matrix A , then we recover a sparse representation of the data with Φ , and next we transform back the data from the sparse domain to the data domain using Ψ^{-1} . Consequently $\Phi = A\Psi$ maps the sparse domain to the measurement domain; hence, if Φ satisfies RIP, then an SVM's classifier in the measurement domain works well. Since Ψ is orthonormal, and entries of A are sampled iid from the Gaussian distribution, A and $A\Psi$ have the same distribution. Hence if we sample A i.i.d from the normal distribution, with high probability $A\Psi$ satisfies the RIP property. In compressed learning,

Table 1: Parameters involved in experiment

Parameter	Description
n	Data domain dimension.
k	sparsity level.
d	Number of measurements
\mathbf{u}	a random unit vector
err	percentage of labels flipped
\mathcal{D}	A random global distribution
s	Sample size

if we do not want to recover the data and only desire to classify the instance, we do not even need to know Ψ . We just project the data to the measurement domain using the sensing matrix A and then run the SVM's learning algorithm in the measurement domain.

Therefore, compressed learning is universal with respect to bases, the SVM's classifier in the measurement domain works almost as well as the best classifier in the data domain provided that there exists some basis in which the instances have a sparse representation.

7 Experimental Results

In this section, we provide experimental results supporting our theoretical contribution which indicates that compressed learning is possible. In this paper, we just present simple synthesized experiments. A more detailed set of experiments, investigating the impact of different parameters, a comparison with the other dimensionality reduction techniques, empirical results on high dimensional real datasets, and manifold learning results will be provided elsewhere. Table 1 shows the parameters involved in this experiment.

The generalization error of the classifier is measured by sampling 2000 i.i.d instances according to \mathcal{D} . Also, each experiment is repeated 10 times and errors are averaged. Distribution \mathcal{D} is chosen randomly over the set of k -sparse ± 1 vectors in \mathbb{R}^n . vector \mathbf{u} determined the label of each vector \mathbf{x} via $y = \text{sign}(\mathbf{u}^\top \mathbf{x})$ then the label is permuted with some probability according to err .

In this experiment, all of the parameters were fixed and the SVM's classifiers in the data and measurement domains are examined in terms of their training and generalization errors. The experiments were repeated 20 times, with $n = 2000, k = 601, d = 800, s = 100, errRate = 0.0$. Figure 2 shows the histogram of the training error of the measurement domain classifier. We can see that the training error is always very small and on average it has training error 5%.

Figure 3 demonstrates the generalization error of the classifier. The plot is a scatter plot of the generalization error at the data domain versus the generalization error at the measurement domain. Again, we can see that on average the difference is less than 8%.

As we mentioned earlier, due to the lack of space, only an elementary experimental result is provided. A more detailed experimental result, investigating the impact of all compressed learning parameters, comparing the method with the other dimensionality reduction techniques, and more complete ex-

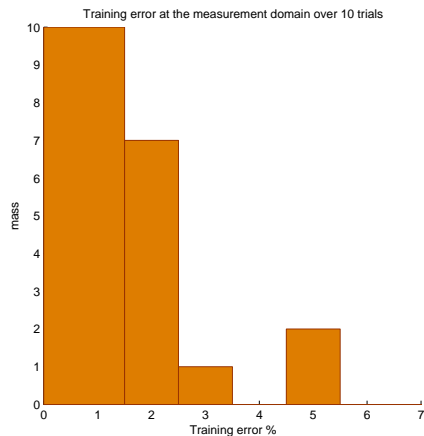


Figure 2: Training error of SVM's classifiers in data and measurement domains versus the number of times the experiment repeated. All of the parameters were fixed at this experiment: $n = 2000$, $k = 601$, $m = 800$, $s = 100$, $errRate = 0.0$, and the experiment repeated 20 times. The data domain classifier remains consistent with data.

planations will be provided elsewhere.

8 Conclusion

In this paper we showed that in compressed sensing framework, learning in the measurement domain is possible. We showed that a family of matrices widely used in compressed sensing, which satisfy near isometry property, preserve the learnability of the data set. We showed that the accuracy of the soft margin SVM's classifier in measurement domain is at most $O(\sqrt{\epsilon})$ worse than the accuracy of the classifier in high dimensional space. However, learning in the compressed domain gains the benefits of compressed sensing, and dimensionality reduction simultaneously. Two other family of matrices are also used in compressed sensing: expander graphs [35, 36] satisfying RIP with respect to ℓ_1 norm, and deterministic sensing matrices [37] satisfying statistical RIP. Finding generalization error bounds when those matrices are used for compressed sensing, and investigating the performance of other learning algorithms like boosting [38] are the two future work in compressed learning. Using machine learning techniques for devising new adaptive sensing methods is the complementary of the compressed learning problem. Finally, performing more learning experiments with novel applications of compressed sensing may be an interesting experimental future work.

Acknowledge

The authors would like to thank Howard Karloff, Patrick Haffner, Indraneel Mukherjee, Avrim Blum, and Aarti Singh for their constructive comments.

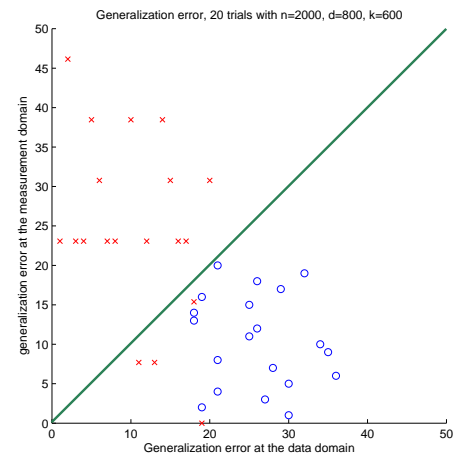


Figure 3: Generalization error of SVM's classifiers in data and measurement domains

References

- [1] N. Cristianini and J. Shawe-Taylor. An introduction to Support Vector Machines and other kernel-based learning methods. *Cambridge University Press*, 2000.
- [2] C. Burges. A tutorial on Support Vector Machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):955-974, 1998.
- [3] I. Daubechies. *Ten Lectures on Wavelets*. Number 61 in CBMS/NSF Series in Applied Math. 1992.
- [4] Stephane Mallat. *A Wavelet Tour of Signal Processing*. AP Professional, London, 1997.
- [5] D. Blei, A. Ng, M. Jordan, and J. Lafferty. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:2003, 2003.
- [6] D. Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. pages 4–15. Springer Verlag, 1998.
- [7] V. Tarokh, N. Seshadri, and R. Calderbank. Space-time codes for high data rate wireless communication: Performance criterion and code construction. *IEEE Trans. Inform. Theory*, 44:744–765, 1998.
- [8] G. Cormode and S. Muthukrishnan. Combinatorial algorithms for Compressed Sensing. *In Proc. 40th Ann. Conf. Information Sciences and Systems, Princeton*, 2006.
- [9] E. Candès and T. Tao. Near optimal signal recovery from random projections: Universal encoding strategies. *IEEE Transactions on Information Theory*, vol. 52, pp.5406-5425, 2006.
- [10] D. Donoho. Compressed Sensing. *IEEE Trans. on Information Theory*, 52(4), pp. 1289 - 1306, April 2006.
- [11] M. Duarte, M. Davenport, D. Takhar, J. Laska, T. Sun, K. Kelly, and R. Baraniuk. Single-pixel imaging via compressive sampling. *IEEE Signal Processing Magazine*, 2008.
- [12] J. Wright, A. Yang, A. Ganesh, S. Shastry, and Y. Ma. Robust face recognition via sparse representation. *To*

- appear in *IEEE Trans. on Pattern Analysis and Machine Intelligence*.
- [13] F. Rodriguez and G. Sapiro. Sparse representations for image classification: Learning discriminative and reconstructive non-parametric dictionaries. *Preprint 2007*.
- [14] I. Drori. Compressed Video Sensing. *Preprint*, 2009.
- [15] F. Colas, P. Paclk, J. Kok, and P. Brazdil. Does SVM Really Scale Up to Large Bag of Words Feature Spaces? In *Advances in Intelligent Data Analysis VII, Proceedings of the 7th International Symposium on Intelligent Data Analysis (IDA 2007)*.
- [16] A. Blum. Random projection, margins, kernels, and feature-selection. *LNCS*, 3940:2006, 2005.
- [17] M. Balcan, A. Blum, and S. Vempala. Kernels as features: On kernels, margins, and low-dimensional mappings. In *15th International Conference on Algorithmic Learning Theory (ALT) 04*, pages 79–94, 2004.
- [18] S. Dasgupta and A. Gupta. An elementary proof of the Johnson-Lindenstrauss lemma. Technical report, 1999.
- [19] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 2007.
- [20] D. Fradkin. Experiments with random projections for machine learning. In *KDD 03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 517–522. ACM Press, 2003.
- [21] S. Dasgupta. Experiments with Random Projection. In *UAI '00: Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 143–151, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.
- [22] M. Wakin and R. Baraniuk. Random projections of smooth manifolds. *ICASSP*, 2006.
- [23] V. Vapnik. *Statistical Learning Theory*. John Wiley, New York, 1995.
- [24] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification, a survey of recent advances. *ESAIM: Probability and Statistics 9*: 323-375, 2005.
- [25] N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines: and other Kernel-based Learning Methods*. Cambridge University Press, 2000.
- [26] S. Shalev-Shwartz and N. Srebro. SVM Optimization: Inverse Dependence on Training set size. *ICML*, 2008.
- [27] R. Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197-227, 1990.
- [28] C. Hegde, M. Wakin, and R. Baraniuk. Random projections for manifold learning. *Neural Information Processing Systems (NIPS), Vancouver, Canada*, 2007.
- [29] E. Candès. The restricted isometry property and its implications for compressed sensing. *Compte Rendus de l'Academie des Sciences, Paris*, 2008.
- [30] E. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59(8):12081223, 2006.
- [31] E. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Inf. Theory*, 52(2):489509, 2006.
- [32] J. Haupt and R. Nowak. A generalized restricted isometry property. *University of Wisconsin Madison Technical Report ECE-07-1*, 2007.
- [33] K. Sridharan, S. Shalev-Shwartz, and N. Srebro. Fast convergence rates for excess regularized risk with application to SVM. <http://ttic.uchicago.edu/~karthik/con.pdf>, 2008.
- [34] P. Bartlett and S. Mendelson. Rademacher and gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Res.*, 2003.
- [35] S. Jafarpour, W. Xu, B. Hassibi, and R. Calderbank. Efficient compressed sensing using high-quality expander graphs. *submitted to the IEEE transactions on Information Theory*.
- [36] P. Indyk and M. Ruzic. Near-optimal sparse recovery in the ℓ_1 norm. *FOCS 2008*.
- [37] R. Calderbank, S. Howard, and S. Jafarpour. Construction of a large class of deterministic sensing matrices that satisfy a statistical isometry property. *Preprint 2008*.
- [38] R. Schapire. The boosting approach to machine learning: An overview. *Nonlinear Estimation and Classification*. Springer, 2003.