Machine Learning (2019) 108:2035–2060
https://doi.org/10.1007/s10994-019-05804-3

2D compressed learning: support matrix machine with bilinear random projections

Di Ma[1] · **Songcan Chen**[1]

Received: 16 December 2016 / Accepted: 29 April 2019 / Published online: 23 May 2019

Abstract
Support matrix machine (SMM) is an efficient matrix classification method that can leverage the structure information within the matrix to improve the classification performance. However, its computational and storage costs are still expensive for high-dimensional data. To address these problems, in this paper, we consider a 2D compressed learning paradigm to learn the SMM classifier in some compressed data domain. Specifically, we use the Kronecker compressed sensing (KCS) to obtain the compressive measurements and learn the SMM classifier. We show that the Kronecker product measurement matrices used by KCS satisfies the restricted isometry property (RIP), which is a property to ensure the learnability of the compressed data. We further give a lower bound on the number of measurements required for KCS. Though this lower bound shows that KCS requires more measurements than the regular CS to satisfy the same RIP condition, KCS itself still enjoys lower computational and storage complexities. Then, using the RIP condition, we verify that the learned SMM classifier in the compressed domain can perform almost as well as the best linear classifier in the original uncompressed domain. Finally, our experimental results also demonstrate the feasibility of 2D compressed learning.

Keywords 2D compressed learning · Bilinear random projection · Dimension reduction · Support matrix machine · Kronecker compressed learning

1 Introduction

Classification is a fundamental problem in machine learning and statistics. Conventional methods such as support vector machines (SVMs) (Cortes and Vapnik 1995) and logistic regression (Friedman et al. 2001) are originally designed for vector data while the real-world

Editor: Maria-Florina Balcan.

✉ Songcan Chen
s.chen@nuaa.edu.cn

Di Ma
madi_nuaa@163.com

[1] Department of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China

data tends to have data of other forms, such as matrix(image) or tensor(video). In conventional classification methods that deal with the latter form, we often reshape such data into vectors, which breaks down the structure relationship of the data, e.g., the correlation between different channels for EEG data (Zhou and Li 2014) or the spatial relationship between the nearby pixels of an image (Wolf et al. 2007). Support matrix machine (SMM) (Luo et al. 2015) is proposed for exploiting the relationship among the rows and columns of the matrix data. To this end, it imposes a spectral elastic net constraint to capture the structure among the matrix data for obtaining the desired solution. Experiment results verify that the SMM outperforms the conventional SVM on matrix data.

Though SMM realizes effective and efficient processing for matrix data compared to the vector-based counterpart, its storage and computation costs are still expensive for large-scale and high-dimensional data, such as high-resolution images. To address these challenges, one of the commonly used methods is to first compress the data (e.g. project the high-dimensional data into a low-dimensional subspace) and then learn directly in the compressed domain.

Compressed sensing (CS) (Candes and Tao 2006; Donoho 2006) is an efficient method to simultaneously realize data acquisition and compression, and is able to recover the data from far fewer measurements than required by the Shannon–Nyquist sampling theorem (Rish and Grabarnik 2014). It has widely applied in both reconstruction problems, e.g., MRI (Lustig et al. 2008), Single Pixel Camera (Duarte et al. 2008), and compressive learning problems, e.g., compressive classification problems (Reboredo et al. 2013), compressive regression problems (Maillard and Munos 2009). However, the regular CS essentially performs on the vectorized data. That is, when handling matrix data, we have to firstly convert it to a vector. Such vectorization unavoidably destroys the inherent structure of the matrix, making the regular CS not quite suitable for matrix classification problem. To preserve the structure, Duarte and Baraniuk (2012) proposes the Kronecker compressed sensing (KCS) using the measurement matrices formed by the Kronecker product. KCS can be implemented by performing independent linear projection on each dimension to reflect the structure presented in that dimension.

Motivated by the above works, in this paper, we consider to learn the SMM classifier using the KCS measurements realized by a bilinear projection. The latter involves two measurement matrices respectively for row and column of the matrix. The choice for them will influence the classification accuracy in the compressed domain. One commonly chosen class of measurement matrices in CS is that satisfying the restricted isometry property (RIP) (Baraniuk et al. 2008; Recht et al. 2010). It is a property that ensures this class of matrices can approximately preserve the structure of the original instance space and in turn approximately preserve the classification accuracy in the compressed domain (Calderbank et al. 2009). For this reason, we expect that the Kronecker product measurement matrix can also satisfy the RIP. Fortunately, our theoretical analysis shows that as long as the two measurement matrices both satisfy the RIP, their Kronecker product likewise satisfies the RIP. Moreover, we further give a lower bound on the number of measurements required for KCS, which is larger than that for regular CS under the same RIP condition. Nonetheless, we show that KCS enjoys lower computational and space complexities. Afterwards, using the RIP condition, we verify that the learned SMM classifier in the Kronecker compressed domain performs almost as well as the best linear classifier in the original data domain. Furthermore, our experimental results also show that with the increasing number of the measurements (but still smaller than the original dimensionalites), the classification accuracy in the compressed domain gets as close as to that in the original data domain.

Our work can be regarded as a generalization of Calderbank et al. (2009) from both the CS and machine learning points of view. From the CS perspective, the KCS generalizes the

regular CS for the matrix data. From the machine learning viewpoint, the SMM classifier generalizes the SVM classifier. We conduct experiments to confirm the effectiveness of these generalizations and the results exactly show (1) KCS is more suitable for matrix data than the regular CS; (2) the SMM classifier is more suitable for the KCS measurements than the SVM classifier.

The remainder of the paper is as follows: The notations and a review of SMM are presented in Sect. 2. In Sect. 3 we introduce the Kronecker compressed sensing (KCS) and the generalized restricted isometry property (RIP) for the Kronecker product measurement matrices. Section 4 provides the theoretical results and corresponding proofs to verify the feasibility for learning SMM classifier in the compressed domain. Section 5 presents a series of experiments to support our theorems. We give a conclusion in Sect. 6.

2 Preliminaries

2.1 Notation

We assume all data are matrices with rank at most r, the Frobenius norm of X is bounded by R, the data domain is:

$$\mathcal{X} = \left\{(X, y) : X \in \mathbb{R}^{d_1 \times d_2}, rank(X) \leq r, \|X\|_F \leq R, y \in \{-1, 1\}\right\}$$

The sample set of N i.i.d. labeled samples is:

$$S = \{(X_1, y_1), \ldots, (X_N, y_N)\}$$

The matrix I_d is the $d \times d$ identity matrix. For a vector $x \in \mathbb{R}^d$, the Euclidean norm is denoted as $\|x\| = \sqrt{\sum_{i=1}^{d} x_i^2}$. For a matrix $X \in \mathbb{R}^{d_1 \times d_2}$ of rank r where $r \leq \min(d_1, d_2)$, the truncated singular value decomposition (truncated SVD) of X is $X = U\Sigma V^T$ where $U \in \mathbb{R}^{d_1 \times r}$ and $V \in \mathbb{R}^{d_2 \times r}$ satisfy $U^T U = I_r$ and $V^T V = I_r$, $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_r)$ with $\sigma_1 \geq \ldots \geq \sigma_r > 0$. Let $\|X\|_F = \sqrt{\sum_{i,j} X_{ij}^2} = \sqrt{\sum_{i=1}^{r} \sigma_i^2}$ be the Frobenius norm, $\|X\|_* = \sum_{i=1}^{r} \sigma_i$ be the nuclear norm, and $\|X\|_{spec} = \sigma_1$ be the spectral norm.

For any $\tau > 0$, the singular value thresholding (SVT) of matrix X is defined as $\mathcal{D}_\tau(X) = U\mathcal{D}_\tau(\Sigma)V^T$, where $\mathcal{D}_\tau(\Sigma) = \text{diag}\left((\sigma_1 - \tau)_+, \ldots, (\sigma_r - \tau)_+\right)$, $(\sigma_i - \tau)_+ = \max(\sigma_i - \tau, 0)$.

Since the nuclear norm $\|X\|_*$ is not differentiable, one considers the subdifferential of $\|X\|_*$, which is the set of subgradients denoted by $\partial\|X\|_*$ as

$$\partial\|X\|_* = \left\{UV^T + Z : Z \in R^{d_1 \times d_2}, U^T Z = 0, ZV = 0, \|Z\|_{spec} \leq 1\right\} \tag{1}$$

2.2 Support matrix machine

The support matrix machine (SMM) is a classification method proposed for matrix data classification problems. Concretely, given a set of training samples $S = \{X_i, y_i\}_{i=1}^N$, where $X_i \in \mathbb{R}^{d_1 \times d_2}$ is the ith sample, $y_i \in \{-1, 1\}$ is the corresponding label. SMM considers to exploit the structure information among the rows or columns in the matrix samples for improving the classification performance. To this end, SMM imposes a low-rank constraint on its weight matrix W. Furthermore, to avoid the NP-hard problem brought by the matrix rank

minimization, (Luo et al. 2015) uses the nuclear norm $\|W\|_*$ as a best convex approximation of rank(W). As a result, the approximated optimization problem can be cast as follows:

$$\min_W \ \frac{1}{2}\|W\|_F^2 + \tau\|W\|_* + \frac{C}{N}\sum_{i=1}^{N}\{1 - y_i[\mathrm{tr}(W^T X_i)]\}_+ \tag{2}$$

where $W \in \mathbb{R}^{d_1 \times d_2}$ is the matrix of the weight coefficients, the nuclear norm enforces low-rank property on W and the Frobenius norm induces a stable solution, parameter τ controls the trade-off between the nuclear norm and the Frobenius norm. Since $\|W\|_F^2 = \sum_{i=1}^{\min(d_1,d_2)} \sigma_i^2(W)$, $\|W\|_* = \sum_{i=1}^{\min(d_1,d_2)} \sigma_i(W)$, the combination of the above two norms $\frac{1}{2}\|W\|_F^2 + \tau\|W\|_*$ is also called the spectral elastic net, which can be interpreted as a elastic net penalty (Zou and Hastie 2005) on the eigenvalues for incorporating the sparsity property and the grouping property into the eigenvalues to capture the latent structure among matrix samples. Recall that $\mathrm{tr}(W^T W) = \mathrm{vec}(W)^T \mathrm{vec}(W) = w^T w$, $\mathrm{tr}(W^T X) = \mathrm{vec}(W)^T \mathrm{vec}(X) = w^T x$, hence SMM would degenerate to the classical soft margin SVM when $\tau = 0$.

The following theorem is a consequence of SMM optimization problem, which is vital in the proof of our main theorem.

Theorem 1 *Suppose that the optimal solution of problem (2) is $\tilde{W}$, then*

$$\begin{aligned} \tilde{W} &= \mathcal{D}_\tau\left(\sum_{i=1}^{N} \tilde{\alpha}_i y_i X_i\right) \\ \|\tilde{W}\|_F^2 + \tau\|\tilde{W}\|_* &\le C \end{aligned} \tag{3}$$

where $0 \le \tilde{\alpha}_i \le \dfrac{C}{N}$.

Proof See "Appendix 1". □

Although SMM has achieved great success in the classification problem on matrix data, it suffers from the storage and computation burden when dealing with large-scale and high-dimensional data. In the next section, we introduce a universal data compression method and then directly perform SMM in the compressed domain.

3 2D compressed learning

3.1 Kronecker compressed sensing

Compressed sensing (CS) is an efficient method to obtain the compressed data. The regular CS model is originally proposed for acquiring sparse signal $x \in \mathbb{R}^d$ through

$$x_A = Ax$$

where $x_A \in \mathbb{R}^p$ is the CS measurements, $A \in \mathbb{R}^{p \times d}$ represents the measurement matrix. Recht et al. (2010) then generalizes the regular CS model to low-rank matrix

$$x_{\mathcal{M}} = \mathcal{M}(X)$$

where $x_{\mathcal{M}} \in \mathbb{R}^k$ is the CS measurements, $X \in \mathbb{R}^{d_1 \times d_2}$ is the original matrix data, $\mathcal{M} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^k$ is a linear map and always written in terms of a linear projection as

$$\mathcal{M}(X) = \Phi \mathrm{vec}(X) \tag{4}$$

where $\Phi \in \mathbb{R}^{k \times d_1 d_2}$ is the measurement matrix, $\text{vec}(X)$ denotes the vectorized X with its columns stacked in order on top of one another.

However, the regular CS acquisition procedure (4) is not quite suitable for classification problem on matrix data since the structure among rows and columns in the matrix would be destroyed by the vectorization. To preserve the structure, Duarte and Baraniuk (2012) proposes the Kronecker compressed sensing (KCS) using the measurement matrices formed by the Kronecker product, i.e.,

$$\text{vec}(X_\Phi) = (\Phi_2 \otimes \Phi_1)\text{vec}(X)$$

where $X_\Phi \in \mathbb{R}^{k_1 \times k_2}$ is the KCS measurements, $\Phi_2 \otimes \Phi_1$ is the Kronecker product of Φ_1 and Φ_2, $\Phi_1 \in \mathbb{R}^{k_1 \times d_1}$ and $\Phi_2 \in \mathbb{R}^{k_2 \times d_2}$ are the measurement matrices for row and column separately. According to the property of Kronecker product, the KCS can be realized by a bilinear projection as follows:

$$X_\Phi = \Phi_1 X \Phi_2^T$$

where independent linear projection on each dimension reflects the structure presented in that dimension (Duarte and Baraniuk 2012).

The problem now is to choose appropriate measurement matrices. One commonly chosen class of measurement matrices in CS is that satisfying the restricted isometry property (RIP). The definitions of the RIP conditions for sparse signal and low-rank matrix are given by Candes and Tao (2006) and Recht et al. (2010) respectively and we restate them together as follows:

Definition 1 Let $A \in \mathbb{R}^{p \times d}$ be a matrix and $\mathcal{M} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^k$ be a linear map. For integers $1 \leq s \leq p$ and $1 \leq r \leq \min(d_1, d_2)$, define the restricted isometry constants (RIC) $\delta_s(A)$ and $\delta_r(\mathcal{M})$ to be the smallest numbers such that for all s-sparse vectors x and all matrices X of rank at most r

$$(1 - \delta_s(A))\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \delta_s(A))\|x\|_2^2$$
$$(1 - \delta_r(\mathcal{M}))\|X\|_F^2 \leq \|\mathcal{M}(X)\|_2^2 \leq (1 + \delta_r(\mathcal{M}))\|X\|_F^2$$

then the matrix A and the linear map $\mathcal{M}$ are said to satisfy RIP with RIC $\delta_s(A)$ and $\delta_r(\mathcal{M})$.

The RIP condition ensures this class of matrices can approximately preserve the structure of the original instance space and in turn approximately preserve the classification accuracy in the compressed domain (Calderbank et al. 2009). For the above reason, we expect the Kronecker product measurement matrices can also satisfy the RIP condition. In the following subsection, we give an analysis on the RIP condition of the Kronecker product measurement matrix $\Phi_2 \otimes \Phi_1$ for all matrices of rank at most r.

3.2 Generalized restricted isometry property for Kronecker product measurement matrices

In this subsection, we firstly study the RIC of the Kronecker product $\Phi_2 \otimes \Phi_1$, denoted by $\delta_r(\Phi_2 \otimes \Phi_1)$, for all matrices of rank at most r. We have the following theorem as a generalization of Lemma 3.2 in Duarte and Baraniuk (2012).

Theorem 2 *Let $\Phi_1 \in \mathbb{R}^{k_1 \times d_1}$, $\Phi_2 \in \mathbb{R}^{k_2 \times d_2}$ be matrices with RIC $\delta_r(\Phi_1)$, $\delta_r(\Phi_2)$ respectively, then,*

$$\delta_r(\Phi_2 \otimes \Phi_1) \leq \prod_{i=1}^{2} (1 + \delta_r(\Phi_i)) - 1$$

Proof According to the definition of RIC for the low-rank matrix, $\delta_r(\Phi_2 \otimes \Phi_1)$ is the smallest number such that for all matrices of rank at most r, the following inequality holds

$$(1 - \delta_r(\Phi_2 \otimes \Phi_1))\|X\|_F^2 \leq \|(\Phi_2 \otimes \Phi_1)\mathrm{vec}(X)\|_2^2 \leq (1 + \delta_r(\Phi_2 \otimes \Phi_1))\|X\|_F^2$$

Thus the eigenvalues of $(\Phi_2 \otimes \Phi_1)(\Phi_2 \otimes \Phi_1)^T$ obey

$$\begin{aligned} 1 - \delta_r(\Phi_2 \otimes \Phi_1) &\leq \sigma_{\min}((\Phi_2 \otimes \Phi_1)(\Phi_2 \otimes \Phi_1)^T) \\ &\leq \sigma_{\max}((\Phi_2 \otimes \Phi_1)(\Phi_2 \otimes \Phi_1)^T) \leq 1 + \delta_r(\Phi_2 \otimes \Phi_1) \end{aligned}$$

where $\sigma_{\min}((\Phi_2 \otimes \Phi_1)(\Phi_2 \otimes \Phi_1)^T)$ and $\sigma_{\max}((\Phi_2 \otimes \Phi_1)(\Phi_2 \otimes \Phi_1)^T)$ denote the minimal and maximal eigenvalues of $(\Phi_2 \otimes \Phi_1)(\Phi_2 \otimes \Phi_1)^T$, respectively. Furthermore, it is well known that $\sigma_{\min}((\Phi_2 \otimes \Phi_1)(\Phi_2 \otimes \Phi_1)^T) = \sigma_{\min}((\Phi_2 \Phi_2^T) \otimes (\Phi_1 \Phi_1^T)) = \sigma_{\min}(\Phi_2 \Phi_2^T)\sigma_{\min}(\Phi_1 \Phi_1^T)$, $\sigma_{\max}((\Phi_2 \otimes \Phi_1)(\Phi_2 \otimes \Phi_1)^T) = \sigma_{\max}((\Phi_2 \Phi_2^T) \otimes (\Phi_1 \Phi_1^T)) = \sigma_{\max}(\Phi_2 \Phi_2^T)\sigma_{\max}(\Phi_1 \Phi_1^T)$. By using the RIC of Φ_1 and Φ_2, we have

$$\begin{aligned} (1 - \delta_r(\Phi_1))(1 - \delta_r(\Phi_2)) &\leq \sigma_{\min}((\Phi_2 \otimes \Phi_1)(\Phi_2 \otimes \Phi_1)^T) \\ &\leq \sigma_{\max}((\Phi_2 \otimes \Phi_1)(\Phi_2 \otimes \Phi_1)^T) \leq (1 + \delta_r(\Phi_1))(1 + \delta_r(\Phi_2)) \end{aligned}$$

Hence we must have

$$\delta_r(\Phi_2 \otimes \Phi_1) \leq \prod_{i=1}^{2} (1 + \delta_r(\Phi_i)) - 1$$

□

The following theorem gives a lower bound on the RIC $\delta_r(\Phi_2 \otimes \Phi_1)$, which is a generalization of Theorem 3.7 in Jokar and Mehrmann (2012) from vectors to low-rank matrices.

Theorem 3 *Let $\Phi_1 \in \mathbb{R}^{k_1 \times d_1}$, $\Phi_2 \in \mathbb{R}^{k_2 \times d_2}$ have normalized rows with RIC $\delta_r(\Phi_1)$, $\delta_r(\Phi_2)$. Then*

$$\delta_r(\Phi_2 \otimes \Phi_1) \geq \max(\delta_r(\Phi_1), \delta_r(\Phi_2)) \tag{5}$$

Proof We prove that $\delta_r(\Phi_2 \otimes \Phi_1) \geq \delta_r(\Phi_1)$, the proof that $\delta_r(\Phi_2 \otimes \Phi_1) \geq \delta_r(\Phi_2)$ follows analogously, thus is omitted. We know that $\delta_r(\Phi_1)$ is the smallest constant such that for all matrices $X \in \mathbb{R}^{p \times q}$ $(pq = d_1)$ with $\mathrm{rank}(X) \leq r$, we have

$$(1 - \delta_r(\Phi_1))\|X\|_F^2 \leq \|\Phi_1 \mathrm{vec}(X)\|_2^2 \leq (1 + \delta_r(\Phi_1))\|X\|_F^2$$

For any $\mathrm{rank}(X) \leq r$, we construct the matrix $X_L = \begin{pmatrix} \mathrm{vec}(X) & 0 & \dots & 0 \end{pmatrix} \in \mathbb{R}^{d_1 \times d_2}$ with $\mathrm{rank}(X_L) = 1 \leq r$ and $\|X_L\|_F^2 = \|X\|_F^2$. Since Φ_2 has normalized rows, we have

$$\|(\Phi_2 \otimes \Phi_1)\mathrm{vec}(X_L)\|_2^2 = \sum_{i=1}^{k_1} \phi_{1,k_1}^2 \|\Phi_1 \mathrm{vec}(X)\|_2^2 = \|\Phi_1 \mathrm{vec}(X)\|_2^2 \tag{6}$$

On the other hand, $\delta_r(\Phi_2 \otimes \Phi_1)$ is the smallest constant such that

$$(1 - \delta_r(\Phi_2 \otimes \Phi_1))\|X\|_F^2 \leq \|(\Phi_2 \otimes \Phi_1)\mathrm{vec}(X_L)\|_2^2 \leq (1 + \delta_r(\Phi_1))\|X\|_F^2$$

for the special case of X_L from (6) we have

$$(1-\delta_r(\Phi_1))\|X\|_F^2 \leq \|(\Phi_2 \otimes \Phi_1)\text{vec}(X_L)\|_2^2 \leq (1+\delta_r(\Phi_1))\|X\|_F^2$$

where $\delta_r(\Phi_1)$ is the smallest constant for this special class of matrices. Therefore, for general matrices of rank at most r, we have

$$\delta_r(\Phi_2 \otimes \Phi_1) \geq \delta_r(\Phi_1)$$

□

From Theorems 2 and 3, we see that the pair of bounds on $\delta_r(\Phi_2 \otimes \Phi_1)$ becomes tight if there is a measurement matrix with dominant RIC. Obviously, when one of the measurement matrices is identity matrix, the pair of bounds is tightest since $\delta_r(\mathbf{I}) = 0$. Now without loss of generality, let $\Phi_2 = \mathbf{I}_{d_2 \times d_2}$, then we have

$$\delta_r(\mathbf{I}_{d_2 \times d_2} \otimes \Phi_1) = \delta_r(\Phi_1) \tag{7}$$

Recht et al. (2010) gives the following theorem to demonstrate that when the linear map $\mathcal{M} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^k$ is nearly isometric random variable, it will obey the RIP with a small RIC under appropriate k, d_1, d_2.

Theorem 4 *Fix* $0 < \delta < 1$. *If* $\mathcal{M} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^k$ *is a nearly isometric random variable, then for every* $1 \leq r \leq \min(d_1, d_2)$, *there exist constants* c_0, c_1 *depending only on* δ *such that, with probability at least* $1 - \exp(-c_1 k)$, $\delta_r(\mathcal{M}) \leq \delta$ *whenever* $k \geq c_0 r(d_1 + d_2)\log(d_1 d_2)$.

Let $\Phi_1 \in \mathbb{R}^{k_1 \times d_1}$ be a nearly isometric random variable corresponding to the linear map $\mathcal{M} : \mathbb{R}^{p \times q} \to \mathbb{R}^{k_1}$ with $d_1 = pq$. According to Theorem 4, if we wish $\delta_r(\Phi_1) \leq \delta$ with probability at least $1 - \exp(-c_1 k_1)$, the number of measurements needs to satisfy

$$k_1 \geq c_0 r(p+q)\log(pq) \tag{8}$$

The lower bound reaches the maximum value when $p = q = \sqrt{d_1}$ and we have $k_1 \geq 2c_0 r\sqrt{d_1}\log(d_1)$. To sum up, if we wish $\delta_r(\mathbf{I}_{d_2 \times d_2} \otimes \Phi_1) \leq \delta$, the overall number of measurements required for the KCS for the special case (7) is

$$k_{\text{kron}} = k_1 d_2 \geq 2c_0 r\sqrt{d_1} d_2 \log(d_1) \tag{9}$$

On the other hand, immediately using Theorem 4, the number of measurements required for the regular CS (4) with $\delta_r(\Phi) \leq \delta$ is

$$k_{\text{stan}} \geq c_0 r(d_1 + d_2)\log(d_1 d_2) \tag{10}$$

Considering the row and column dimensionalities d_1 and d_2 are of the same order $O(d)$, we see that the lower bound in (9) is larger than the lower bound in (10). This implies that to guarantee the same RIC, the KCS requires more measurements than the regular CS. Nevertheless, the computational and space complexities for KCS are $O(c_0 r d^{5/2}\log(d))$ and $O(c_0 r d^{3/2}\log(d))$ respectively, which are lower than those of the regular CS with both $O\left(c_0 r d^3 \log(d)\right)$.

In the following, we will make a further generalization of the RIP condition. Since we plan to bound the regularization loss of SMM classifier in the compressed domain, we need to show that the near isometry property holds for the terms in the SMM's objective function. Different from traditional SVM, the objective function of SMM has an additional nuclear norm constraint on the weight matrix W. Besides, the weight vector of SVM is a linear combination of support vectors, while the weight matrix W is a SVT of the linear combination

of support matrices, which makes it more complicated. Due to the differences between SVM and SMM, we plan to show that the near isometry property holds for the Frobenius norm and the nuclear norm jointly, which is equivalent to show that the spectral elastic net of the weight matrix $\tilde{W}$ can be approximately preserved after the bilinear projection. Then we show that the inner product between the $\tilde{W}$ and arbitrary sample X can also be approximately preserved. At the first step, we show that the inner product between any two low-rank matrices is approximately preserved.

Lemma 1 *Let $\Phi_1 \in \mathbb{R}^{k_1 \times d_1}$, $\Phi_2 \in \mathbb{R}^{k_2 \times d_2}$ be the measurement matrices satisfying $2r$-RIP with RIC $\delta_{2r}(\Phi_1)$ and $\delta_{2r}(\Phi_2)$, and X, X' be any two matrices in sample set S. Then,*

$$\begin{aligned} \operatorname{tr}\left(X^T X'\right) - 3R^2\delta_{2r}(\Phi_2 \otimes \Phi_1) \le \operatorname{tr}\left[\left(\Phi_1 X \Phi_2^T\right)^T \left(\Phi_1 X' \Phi_2^T\right)\right] \\ \le \operatorname{tr}\left(X^T X'\right) + 3R^2\delta_{2r}(\Phi_2 \otimes \Phi_1), \end{aligned} \tag{11}$$

where $\delta_{2r}(\Phi_2 \otimes \Phi_1) \le \prod_{p=1}^{2}\left(1+\delta_{2r}(\Phi_p)\right) - 1$.

Proof Since X, X' are matrices with rank at most r, according to the subadditivity of the rank (Recht et al. 2010), $X - X'$ is a matrix with rank at most $2r$,

$$rank\left(X - X'\right) \le rank(X) + rank\left(X'\right) \le 2r.$$

According to Eq. (1) and Theorem 2,

$$\begin{aligned} \|\Phi_1\left(X - X'\right)\Phi_2^T\|_F^2 &\le (1+\delta_{2r}(\Phi_2 \otimes \Phi_1))\|X - X'\|_F^2 \\ &= (1+\delta_{2r}(\Phi_2 \otimes \Phi_1))\left(\|X\|_F^2 + \|X'\|_F^2 - 2\operatorname{tr}\left(X^T X'\right)\right). \end{aligned} \tag{12}$$

where $\delta_{2r}(\Phi_2 \otimes \Phi_1) \le \prod_{p=1}^{2}\left(1+\delta_{2r}(\Phi_p)\right) - 1$. Also,

$$\begin{aligned} &(1-\delta_{2r}(\Phi_2 \otimes \Phi_1))\left(\|X\|_F^2 + \|X'\|_F^2\right) - 2\operatorname{tr}\left(\left(\Phi_1 X \Phi_2^T\right)^T \left(\Phi_1 X' \Phi_2^T\right)\right) \\ &\le \|\Phi_1 X \Phi_2^T\|_F^2 + \|\Phi_1 X' {\Phi_2}^T\|_F^2 - 2\operatorname{tr}\left(\left(\Phi_1 X \Phi_2^T\right)^T \left(\Phi_1 X' {\Phi_2}^T\right)\right) \\ &= \|\Phi_1^T\left(X - X'\right)\Phi_2\|_F^2 \end{aligned} \tag{13}$$

Putting (12) and (13) together, and noting $\|X\|_F \le R$, $\|X'\|_F \le R$, then

$$\operatorname{tr}\left(X^T X'\right) - 3R^2\delta_{2r}(\Phi_2 \otimes \Phi_1) \le \operatorname{tr}\left[\left(\Phi_1 X \Phi_2^T\right)^T \left(\Phi_1 X' {\Phi_2}^T\right)\right]$$

It's similar to prove the right side of (11). □

Lemma 2 *Let $\Phi_1 \in \mathbb{R}^{k_1 \times d_1}$, $\Phi_2 \in \mathbb{R}^{k_2 \times d_2}$ be the measurement matrices satisfying $2r$-RIP with RIC $\delta_{2r}(\Phi_1)$ and $\delta_{2r}(\Phi_2)$, and $\tilde{W}$ be the SMM's classifier trained on sample set S. Then,*

$$\begin{aligned} &\frac{1}{2}\|\tilde{W}\|_F^2 + \tau\|\tilde{W}\|_* - O\left(\left(R^2C^2 + \tau^2\tilde{r}\right)\delta_{2r}(\Phi_2 \otimes \Phi_1)\right) \\ &\le \frac{1}{2}\|\Phi_1 \tilde{W} \Phi_2^T\|_F^2 + \tau\|\Phi_1 \tilde{W} \Phi_2^T\|_* \\ &\le \frac{1}{2}\|\tilde{W}\|_F^2 + \tau\|\tilde{W}\|_* + O\left(\left(R^2C^2 + \tau^2\tilde{r}\right)\delta_{2r}(\Phi_2 \otimes \Phi_1)\right). \end{aligned} \tag{14}$$

where $\tilde{r} = \min(d_1, d_2)$, $\delta_{2r}(\Phi_2 \otimes \Phi_1) \le \prod_{p=1}^{2}\left(1+\delta_{2r}(\Phi_p)\right) - 1$.

Proof According to Eq. (3), we have,

$$\|\tilde{W}\|_F^2 = \|\sum_{i=1}^{N}\tilde{\alpha}_i y_i X_i\|_F^2 + \|\tilde{\Lambda}\|_F^2 + 2\mathrm{tr}\left(\tilde{\Lambda}^T\sum_{i=1}^{N}\tilde{\alpha}_i y_i X_i\right) \tag{15}$$

Hence we need to prove that the near isometry property holds for each term in (15). We firstly prove that $\|\sum_{i=1}^{N}\tilde{\alpha}_i y_i X_i\|_F^2$ can be approximately preserved after bilinear projection.

$$\begin{aligned}\left\|\Phi_1\sum_{i=1}^{N}\tilde{\alpha}_i y_i X_i \Phi_2^T\right\|_F^2 &= \left\|\sum_{i=1}^{N}\tilde{\alpha}_i y_i \Phi_1 X_i \Phi_2^T\right\|_F^2 \\ &= \mathrm{tr}\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\tilde{\alpha}_i\tilde{\alpha}_j y_i y_j \left(\Phi_1 X_i \Phi_2^T\right)^T \Phi_1 X_i \Phi_2^T\right) \\ &= \sum_{y_i=y_j}\tilde{\alpha}_i\tilde{\alpha}_j \mathrm{tr}\left(\left(\Phi_1 X_i \Phi_2^T\right)^T \Phi_1 X_i \Phi_2^T\right) \\ &\quad - \sum_{y_i\neq y_j}\tilde{\alpha}_i\tilde{\alpha}_j \mathrm{tr}\left(\left(\Phi_1 X_i \Phi_2^T\right)^T \Phi_1 X_i \Phi_2^T\right)\end{aligned}$$

According to Lemma 1, we have,

$$\begin{aligned}&\sum_{y_i=y_j}\tilde{\alpha}_i\tilde{\alpha}_j \mathrm{tr}\left(\left(\Phi_1 X_i \Phi_2^T\right)^T \Phi_1 X_i \Phi_2^T\right) \\ &\quad - \sum_{y_i\neq y_j}\tilde{\alpha}_i\tilde{\alpha}_j \mathrm{tr}\left(\left(\Phi_1 X_i \Phi_2^T\right)^T \Phi_1 X_i \Phi_2^T\right) \\ &\leq \sum_{y_i=y_j}\tilde{\alpha}_i\tilde{\alpha}_j\left(\mathrm{tr}\left(X_i^T X_j\right) + 3R^2\delta_{2r}\left(\Phi_2\otimes\Phi_1\right)\right) \\ &\quad - \sum_{y_i\neq y_j}\tilde{\alpha}_i\tilde{\alpha}_j\left(\mathrm{tr}\left(X_i^T X_j\right) - 3R^2\delta_{2r}\left(\Phi_2\otimes\Phi_1\right)\right) \\ &\leq \sum_{i=1}^{N}\sum_{j=1}^{N}\tilde{\alpha}_i\tilde{\alpha}_j y_i y_j \mathrm{tr}\left(X_i^T X_j\right) + 3R^2\delta_{2r}\left(\Phi_2\otimes\Phi_1\right)\sum_{i=1}^{N}\tilde{\alpha}_i\sum_{j=1}^{N}\tilde{\alpha}_j \\ &\leq \left\|\sum_{i=1}^{N}\tilde{\alpha}_i y_i X_i\right\|_F^2 + 3R^2C^2\delta_{2r}\left(\Phi_2\otimes\Phi_1\right).\end{aligned}$$

Hence,

$$\left\|\Phi_1\sum_{i=1}^{N}\tilde{\alpha}_i y_i X_i \Phi_2^T\right\|_F^2 \leq \left\|\sum_{i=1}^{N}\tilde{\alpha}_i y_i X_i\right\|_F^2 + 3R^2C^2\delta_{2r}\left(\Phi_2\otimes\Phi_1\right). \tag{16}$$

Then, we prove that $\|\tilde{\Lambda}\|_F^2$ can be approximately preserved after bilinear projection. Considering $\tilde{\Lambda} = -\tau\left(\tilde{U}_0\tilde{V}_0^T + \frac{1}{\tau}\tilde{U}_1\tilde{\Sigma}_1\tilde{V}_1^T\right)$ rewritten as

$$\tilde{\Lambda} = -\tau\sum_{i=1}^{\tilde{r}}\tilde{\sigma}_i\tilde{u}_i\tilde{v}_i^T,$$

where $\tilde{r}=\min(d_1,d_2)$ is the worst case rank of matrix $\tilde{\Lambda}$, $0<\tilde{\sigma}_i\leq 1$, $\tilde{u}_i$ and $\tilde{v}_i$ corresponds to the colums in $\left[\tilde{U}_0,\tilde{U}_1\right]$ and $\left[\tilde{V}_0,\tilde{U}_1\right]$. Then,

$$\begin{aligned}
\|\Phi_1\tilde{\Lambda}\Phi_2^T\|_F^2 &= \tau^2\mathrm{tr}\left(\left(\sum_{i=1}^{\tilde{r}}\tilde{\sigma}_i\Phi_1\tilde{u}_i\tilde{v}_i^T\Phi_2^T\right)^T\sum_{j=1}^{\tilde{r}}\tilde{\sigma}_j\Phi_1\tilde{u}_j\tilde{v}_j^T\Phi_2^T\right)\\
&= \tau^2\sum_{i=1}^{\tilde{r}}\sum_{j=1}^{\tilde{r}}\tilde{\sigma}_i\tilde{\sigma}_j\mathrm{tr}\left(\left(\Phi_1\tilde{u}_i\tilde{v}_i^T\Phi_2^T\right)^T\Phi_1\tilde{u}_j\tilde{v}_j^T\Phi_2^T\right)\\
&\leq \tau^2\sum_{i=1}^{\tilde{r}}\sum_{j=1}^{\tilde{r}}\tilde{\sigma}_i\tilde{\sigma}_j\left(\mathrm{tr}\left(\left(\tilde{u}_i\tilde{v}_i^T\right)^T\tilde{u}_j\tilde{v}_j^T\right)+3\delta_{2r}\left(\Phi_2\otimes\Phi_1\right)\right)\\
&\leq \tau^2\sum_{i=1}^{\tilde{r}}\sum_{j=1}^{\tilde{r}}\tilde{\sigma}_i\tilde{\sigma}_j\mathrm{tr}\left(\left(\tilde{u}_i\tilde{v}_i^T\right)^T\tilde{u}_j\tilde{v}_j^T\right)+3\tau^2\delta_{2r}\left(\Phi_2\otimes\Phi_1\right)\sum_{i=1}^{\tilde{r}}\sum_{j=1}^{\tilde{r}}\tilde{\sigma}_i\tilde{\sigma}_j\\
&\leq \|\tilde{\Lambda}\|_F^2+3\tau^2\tilde{r}\delta_{2r}\left(\Phi_2\otimes\Phi_1\right).
\end{aligned}\tag{17}$$

For the third term in (15), we have

$$\begin{aligned}
&\mathrm{tr}\left(\left(\Phi_1\tilde{\Lambda}\Phi_2^T\right)^T\Phi_1\sum_{j=1}^{N}\tilde{\alpha}_j y_j X_j\Phi_2^T\right)\\
&= -\tau\sum_{i=1}^{\tilde{r}}\sum_{j=1}^{N}\tilde{\sigma}_i\tilde{\alpha}_j y_j\mathrm{tr}\left(\left(\Phi_1\tilde{u}_i\tilde{v}_i^T\Phi_2^T\right)^T\Phi_1 X_i\Phi_2^T\right)\\
&= \tau\sum_{i=1}^{\tilde{r}}\sum_{y_j=-1}\tilde{\sigma}_i\tilde{\alpha}_j\mathrm{tr}\left(\left(\Phi_1\tilde{u}_i\tilde{v}_i^T\Phi_2^T\right)^T\Phi_1 X_i\Phi_2^T\right)\\
&\quad -\tau\sum_{i=1}^{\tilde{r}}\sum_{y_j=1}\tilde{\sigma}_i\tilde{\alpha}_j\mathrm{tr}\left(\left(\Phi_1\tilde{u}_i\tilde{v}_i^T\Phi_2^T\right)^T\Phi_1 X_i\Phi_2^T\right)\\
&\leq \tau\sum_{i=1}^{\tilde{r}}\sum_{y_j=-1}\tilde{\sigma}_i\tilde{\alpha}_j\left(\mathrm{tr}\left(\left(\tilde{u}_i\tilde{v}_i^T\right)^T X_j\right)+O\left(R\delta_{2r}\left(\Phi_2\otimes\Phi_1\right)\right)\right)\\
&\quad -\tau\sum_{i=1}^{\tilde{r}}\sum_{y_j=1}\tilde{\sigma}_i\tilde{\alpha}_j\left(\mathrm{tr}\left(\left(\tilde{u}_i\tilde{v}_i^T\right)^T X_j\right)-O\left(R\delta_{2r}\left(\Phi_2\otimes\Phi_1\right)\right)\right)\\
&\leq -\tau\mathrm{tr}\left(\sum_{i=1}^{\tilde{r}}\tilde{\sigma}_i\tilde{u}_i\tilde{v}_i^T\sum_{j=1}^{N}\tilde{\alpha}_j y_j X_j\right)+\tau\sum_{i=1}^{\tilde{r}}\tilde{\sigma}_i\sum_{j=1}^{N}\tilde{\alpha}_j O\left(R\delta_{2r}\left(\Phi_2\otimes\Phi_1\right)\right)\\
&\leq \mathrm{tr}\left(\tilde{\Lambda}^T\sum_{j=1}^{N}\tilde{\alpha}_j y_j X_j\right)+O\left(\tau\tilde{r}RC\delta_{2r}\left(\Phi_2\otimes\Phi_1\right)\right)
\end{aligned}\tag{18}$$

Putting Eqs. (15)–(18) together, we have,

$$\|\Phi_1\tilde{W}\Phi_2^T\|_F^2\leq\|\tilde{W}\|_F^2+O\left((R^2C^2+\tau^2\tilde{r})\delta_{2r}(\Phi_2\otimes\Phi_1)\right)\tag{19}$$

Then, using (19) and the following inequalities which hold for any matrix W of rank at most r

$$\|W\|_F \leq \|W\|_* \leq \sqrt{r}\|W\|_F \tag{20}$$

we have

$$\begin{aligned}\|\Phi_1 W \Phi_2^T\|_* &\leq \sqrt{\tilde{r}}\|\Phi_1 W \Phi_2^T\|_F \\ &\leq \sqrt{\tilde{r}}\sqrt{\|W\|_F^2 + O((R^2C^2+\tau^2\tilde{r})\delta_{2r}(\Phi_2\times\Phi_1))} \\ &\leq \sqrt{\tilde{r}}\|W\|_F + O\left((\sqrt{\tilde{r}}RC+\tau\tilde{r})\sqrt{\delta_{2r}(\Phi_2\times\Phi_1)}\right) \\ &\leq \sqrt{\tilde{r}}\|W\|_* + O\left((\sqrt{\tilde{r}}RC+\tau\tilde{r})\sqrt{\delta_{2r}(\Phi_2\times\Phi_1)}\right)\end{aligned} \tag{21}$$

Noting that $\|W\|_* \leq \frac{C}{\tau}$, then

$$\|\Phi_1 W \Phi_2^T\|_* \leq \|W\|_* + O\left((\sqrt{\tilde{r}}RC+\tau\tilde{r})\sqrt{\delta_{2r}(\Phi_2\times\Phi_1)}\right) \tag{22}$$

Combining (19) and (22) we finally obtain

$$\begin{aligned}&\frac{1}{2}\|\Phi_1\tilde{W}\Phi_2^T\|_F^2 + \tau\|\Phi_1\tilde{W}\Phi_2^T\|_* \\ &\leq \frac{1}{2}\|\tilde{W}\|_F^2 + \tau\|\tilde{W}\|_* + O\left(\left(R^2C^2+\tau^2\tilde{r}\right)\delta_{2r}\left(\Phi_2\otimes\Phi_1\right)\right)\end{aligned} \tag{23}$$

It is similar to prove the left side of (14). □

So far, we have shown that the restricted isometry property approximately preserves the spectral elastic net of SMM's classifier $\tilde{W}$. Next, we will show that the inner product between SMM's classifier $\tilde{W}$ and arbitrary low-rank sample matrix is also approximately preserved after the bilinear projection.

Lemma 3 *Let $\Phi_1 \in \mathbb{R}^{k_1\times d_1}$, $\Phi_2 \in \mathbb{R}^{k_2\times d_2}$ be the measurement matrices satisfying $2r$-RIP with RIC $\delta_{2r}(\Phi_1)$ and $\delta_{2r}(\Phi_2)$, and $\tilde{W}$ be the SMM's classifier trained on sample set S, X be arbitrary low-rank sample matrix from data domain. Then,*

$$\begin{aligned}&\mathrm{tr}\left(\tilde{W}^T X\right) - O\left(\left(R^2C+\tau\tilde{r}R\right)\delta_{2r}\left(\Phi_2\otimes\Phi_1\right)\right) \\ &\leq \mathrm{tr}\left(\left(\Phi_1\tilde{W}\Phi_2^T\right)^T \Phi_1 X \Phi_2^T\right) \\ &\leq \mathrm{tr}\left(\tilde{W}^T X\right) + O\left(\left(R^2C+\tau\tilde{r}R\right)\delta_{2r}\left(\Phi_2\otimes\Phi_1\right)\right),\end{aligned}$$

where $\tilde{r} = \min(d_1, d_2)$, $\delta_{2r}(\Phi_2\otimes\Phi_1) \leq \prod_{p=1}^{2}\left(1+\delta_{2r}\left(\Phi_p\right)\right) - 1$.

Proof According to Eq. (36),

$$\mathrm{tr}\left(\tilde{W}^T X\right) = \mathrm{tr}\left(\tilde{\Lambda}^T X\right) + \mathrm{tr}\left(\left(\sum_{i=1}^{N}\tilde{\alpha}\, y_i X_i\right)^T X\right) \tag{24}$$

From Lemmas 1 and 2, we can easily prove,

$$\begin{aligned}\operatorname{tr}\left(\left(\sum_{i=1}^{N}\tilde{\alpha}y_iX_i\right)^T X\right) - 3R^2C\delta_{2r}\left(\Phi_2\otimes\Phi_1\right) &\le \operatorname{tr}\left(\left(\sum_{i=1}^{N}\tilde{\alpha}y_i\Phi_1X_i\Phi_2^T\right)^T \Phi_1X\Phi_2\right)\\ &\le \operatorname{tr}\left(\left(\sum_{i=1}^{N}\tilde{\alpha}y_iX_i\right)^T X\right) + 3R^2C\delta_{2r}\left(\Phi_2\otimes\Phi_1\right)\end{aligned} \tag{25}$$

Besides,

$$\begin{aligned}\operatorname{tr}\left(\left(\Phi_1\tilde{\Lambda}\Phi_2^T\right)^T\Phi_1X\Phi_2^T\right) &= -\tau\operatorname{tr}\left(\left(\sum_{i=1}^{\tilde{r}}\tilde{\sigma}_i\Phi_1\tilde{u}_i\tilde{v}_i^T\Phi_2^T\right)^T\Phi_1X\Phi_2^T\right)\\ &= -\tau\sum_{i=1}^{\tilde{r}}\tilde{\sigma}_i\operatorname{tr}\left(\left(\Phi_1\tilde{u}_i\tilde{v}_i^T\Phi_2^T\right)^T\Phi_1X\Phi_2^T\right)\\ &\le -\tau\sum_{i=1}^{\tilde{r}}\tilde{\sigma}_i\left(\operatorname{tr}\left(\left(\tilde{u}_i\tilde{v}_i^T\right)^TX\right) - O\left(R\delta_{2r}\left(\Phi_2\otimes\Phi_1\right)\right)\right)\\ &\le \operatorname{tr}\left(\tilde{\Lambda}^TX\right) + O\left(\tau\tilde{r}R\delta_{2r}\left(\Phi_2\otimes\Phi_1\right)\right)\end{aligned} \tag{26}$$

where $\tilde{r}\le\min\left(d_1,d_2\right)$ is the rank of matrix $\tilde{\Lambda}$. Similarly, we can prove that,

$$\operatorname{tr}\left(\left(\Phi_1\tilde{\Lambda}\Phi_2^T\right)^T\Phi_1X\Phi_2^T\right) \ge \operatorname{tr}\left(\tilde{\Lambda}^TX\right) - O\left(\tau\tilde{r}R\delta_{2r}\left(\Phi_2\otimes\Phi_1\right)\right) \tag{27}$$

Putting Eqs. (24)–(27) together, we obtain

$$\begin{aligned}&\operatorname{tr}\left(\tilde{W}^TX\right) - O\left(\left(R^2C+\tau\tilde{r}R\right)\delta_{2r}\left(\Phi_2\otimes\Phi_1\right)\right)\\ &\le \operatorname{tr}\left(\left(\Phi_1\tilde{W}\Phi_2^T\right)^T\Phi_1X\Phi_2^T\right)\\ &\le \operatorname{tr}\left(\tilde{W}^TX\right) + O\left(\left(R^2C+\tau\tilde{r}R\right)\delta_{2r}\left(\Phi_2\otimes\Phi_1\right)\right)\end{aligned}$$

□

4 Theoretical results

In this section, we present the theoretical analysis of 2D compressed learning. We still employ the two-step strategy used by Calderbank et al. (2009). Consider the SMM classifier trained on S as $\tilde{W}$ and trained on S_Φ as $\tilde{W}_\Phi$. The first step is to investigate the relationship between the generalization performance of $\tilde{W}$ and the generalization performance of the intermediate, projected classifier $\Phi_1\tilde{W}\Phi_2^T$ according to the generalized RIP we introduced in previous section. The second step is to study the relationship between the generalization performance of $\tilde{W}_\Phi$ and the generalization performance of the projected classifier $\Phi_1\tilde{W}\Phi_2^T$. Then we can build a bridge between the generalization performance of $\tilde{W}$ and $\tilde{W}_\Phi$ via $\Phi_1\tilde{W}\Phi_2^T$.

For simplicity of subsequent expression, we rewrite the optimization problem (2) as minimizing the empirical regularization loss:

$$\hat{L}(W) = \frac{1}{N}\sum_{i=1}^{N} \ell(W; X_i, y_i) \tag{28}$$

where $\ell(W; X, y) = h(\langle W, X\rangle, y) + r(W)$ and $h(\langle W, X\rangle, y) = \{1 - y\mathrm{tr}(W^T X)\}_+$ is the hinge loss, $r(W) = \frac{1}{C}(\frac{1}{2}\|W\|_F^2 + \tau\|W\|_*)$ is the spectral elastic net penalty. The corresponding true regularization loss is

$$L(W) = \mathbb{E}_{(X,y)\sim\mathcal{X}}[\ell(W; X, y)] \tag{29}$$

The empirical and true hinge loss are defined respectively as

$$\hat{H}(W) = \frac{1}{N}\sum_{i=1}^{N} h(\langle W, X_i\rangle, y_i)$$

and

$$H(W) = \mathbb{E}_{(X,y)\sim\mathcal{X}}[h(\langle W, X\rangle, y)]$$

The true minimizer is

$$W^\star = \arg\min_W L(W) \tag{30}$$

The following theorem states the relationship between the regularization loss of SMM classifier in data domain and projected classifier in compressed domain.

Theorem 5 *Let $\Phi_1 \in \mathbb{R}^{k_1\times d_1}$, $\Phi_2 \in \mathbb{R}^{k_2\times d_2}$ be the measurement matrices satisfying $2r$-RIP with RIC $\delta_{2r}(\Phi_1)$ and $\delta_{2r}(\Phi_2)$, $\tilde{W}$ be the SMM classifier trained on training set S, then*

$$L_\Phi\left(\Phi_1\tilde{W}\Phi_2^T\right) \le L\left(\tilde{W}\right) + O\left(\left(R^2C + \frac{\tau^2\tilde{r}}{C}\right)\delta_{2r}(\Phi_2\otimes\Phi_1)\right)$$

where $\tilde{r} = \min(d_1, d_2)$, $\delta_{2r}(\Phi_2\otimes\Phi_1) \le \prod_{p=1}^{2}\left(1 + \delta_{2r}\left(\Phi_p\right)\right) - 1$.

Proof See "Appendix 2". □

We have already demonstrated that the regularization loss of the projected SMM classifier is close to the regularization loss of SMM classifier in original data domain. In below, we are going to investigate the relationship between the regularization loss of the projected SMM classifier and SMM classifier in the measurement domain.

Theorem 6 *Let $W^\star$, $L(W)$, $\hat{L}(W)$ be as defined in (28)–(30), where $\|X\|_F \le R$. Then for any $\delta > 0$, with probability at least $1-\delta$ over a sample set of size N, for all $W \in \mathcal{W} := \{W : \|W\|_F^2 + \tau\|W\|_* \le C\}$ we have*

$$L(W) - L(W^\star) \le \hat{L}(W) - \hat{L}(W^\star) + O\left(\left(\sqrt{2C+\tau^2} - \tau\right)\sqrt{\frac{R^2\log(1/\delta)}{N}}\right)$$

Proof See "Appendix 3". □

Up to now, we have accomplished the preparations for our main theorem. We synthesize Theorems 5 and 6 and obtain the main result of our paper as follows:

Theorem 7 *Let $\Phi_1 \in \mathbb{R}^{k_1 \times d_1}$, $\Phi_2 \in \mathbb{R}^{k_2 \times d_2}$ be the measurement matrices satisfying $2r$-RIP with RIC $\delta_{2r}(\Phi_1)$ and $\delta_{2r}(\Phi_2)$. Let $\tilde{W}$ and $\tilde{W}_\Phi$ be the SMM classifier trained on S and S_Φ respectively, W_0 be a good SMM classifier in the data domain with small spectral elastic net penalty which attains low generalization error. Then with probability $1-2\delta$:*

$$H_\Phi(\tilde{W}_\Phi) \leq H(W_0) + O\left(\sqrt{\left(\frac{1}{2}\|W_0\|_F^2 + \tau\|W_0\|_* + \tau^2\delta_{2r}(\Phi_2 \otimes \Phi_1)\right)\left(\delta_{2r}(\Phi_2 \otimes \Phi_1) + \sqrt{\frac{\log(1/\delta)}{(\tau+a)^2 N}}\right)}\right)$$

where a is some small constant that ensures non-zero dominant, $\tilde{r} = \min(d_1, d_2)$, $\delta_{2r}(\Phi_2 \otimes \Phi_1) \leq \prod_{p=1}^{2}\left(1+\delta_{2r}(\Phi_p)\right) - 1$.

Proof See "Appendix 4". □

Note that this result is a weak upper bound due to the relaxation of the upper bound on the regularization loss of SMM. According to Theorem 7, the deviation of $H_\Phi\left(\tilde{W}_\Phi\right)$ from $H(W_0)$ will converges to $O(\delta_{2r}(\Phi_2 \otimes \Phi_1))$ as the number of samples increases. When SMM reduces to SVM (removing the nuclear norm term in the objective function with $\tau = 0$), the deviation of $H_\Phi\left(\tilde{W}_\Phi\right)$ from $H(W_0)$ will converges to $O\left(\sqrt{\delta_{2r}(\Phi_2 \otimes \Phi_1)}\right)$ as the number of samples increases, which is consistent to the result given by Calderbank et al. (2009).

5 Experiments

In this section, we investigate the learning performance of 2D compressed learning for classification problem on the real-world data sets including those originally in matrix representation: (1) the EEG alcoholism database (Luo et al. 2015); (2) the FEI face database (Thomaz and Giraldi 2010) and those originally in vector representation from UCI data sets: (1) the DBWorld e-mails data set (Filannino 2011); (2) the p53 Mutants data set (Danziger et al. 2006) (Although our framework is proposed to adapt to matrix data, we still perform experiments on data originally in vector representation to explore the validity of 2D compressed learning on general data).

We compare 2D compressed learning with conventional compressed learning using SMM with bilinear projection and SVM with single linear projection, referred as SMM-BP and SVM-SP. Besides, to see the influence of bilinear projection on the performance compared with single linear projection, we also perform SMM with single linear projection and SVM with bilinear projection, referred as SMM-SP and SVM-BP. The measurement matrices are generated with i.i.d. Gaussian entries $\Phi_{ij} \sim \mathcal{N}\left(0, \frac{1}{k}\right)$, where k is the dimension of the compressed data.

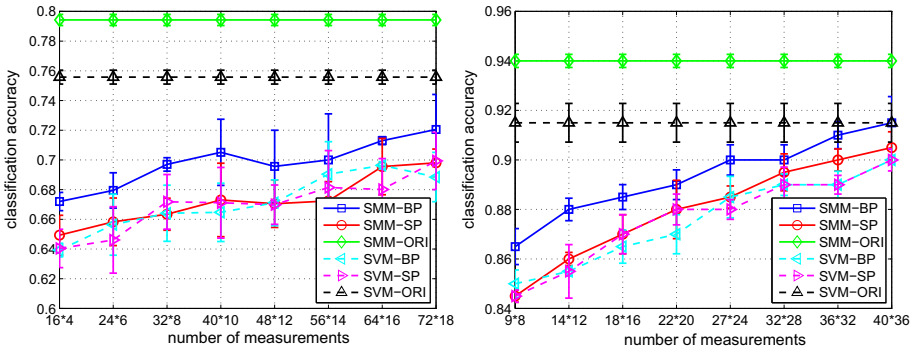
We use the 10-fold cross-validation to evaluate the learning performance. The hyperparameters are also selected via cross-validation. More specifically, we select C and τ from $[10^{-3}, 10^{-2}, \ldots, 10^2, 10^3]$ and $[10^{-5}, 10^{-4}, \ldots, 10^4, 10^5]$.

5.1 Experiments on matrix representation data sets

The EEG alcoholism data set arises to examine EEG correlates of genetic predisposition to alcoholism. It contains two groups of subjects: alcoholic and control. For each subject, 64 channels of electrodes are placed and the voltage values are recorded at 256 time points.

Table 1 Summary of matrix representation data sets

Data sets	Positive	Negative	Matrix dimension	Vector dimension
EEG alcoholism	77	45	256×64	16,384
FEI face	100	100	216×192	41,472



(a) EEG alcoholism **(b)** FEI face

Fig. 1 Classification accuracy for SMM and SVM with a single linear projection and a bilinear projection on EEG alcoholism and FEI face with respect to the different number of measurements

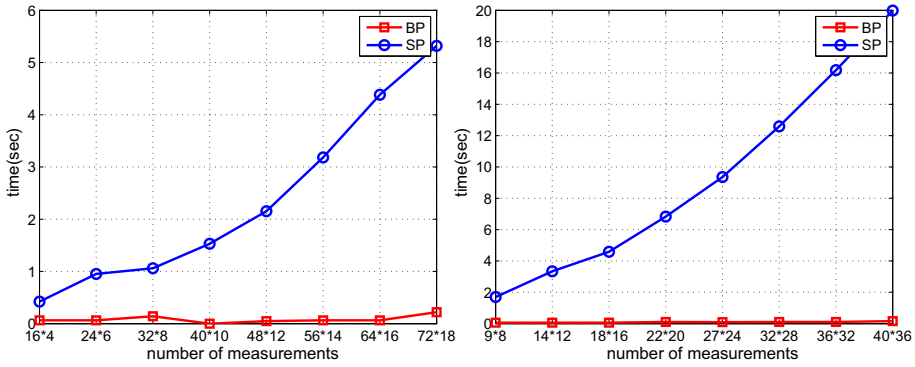
The FEI face database contains the face images of 100 females and 100 males, 14 images for each at various angles. The image size is 640×480. We cropped the images into 216×192 gray images to retain the frontal images of each person.

Table 1 summarizes the characteristics of matrix representation data sets.

Figure 1 presents the classification accuracy for SMM and SVM with bilinear projection and single linear projection on matrix data sets (EEG alcoholism and FEI face) according to different number of measurements. We can see that SMM-BP achieves good performances on the compressed measurements and the performance gets closer to that of the original data as the number of the compressed measurements increases, which verifies the feasibility of 2D compressed learning. In addition, the performances of SMM-BP on both original data and compressed data outperforms other three compressed algorithms, shows the superiority of 2D compressed learning than conventional compressed learning. In more details, the performance of SMM-BP is better than SMM-SP while SVM-BP gets a similar performance with SVM-SP. The performance improvement of SMM-BP may attribute to two aspects, (1) the bilinear projection could preserve the structure information while single linear projection can not; (2) SMM can leverage the structure information while SVM can not take advantages from it.

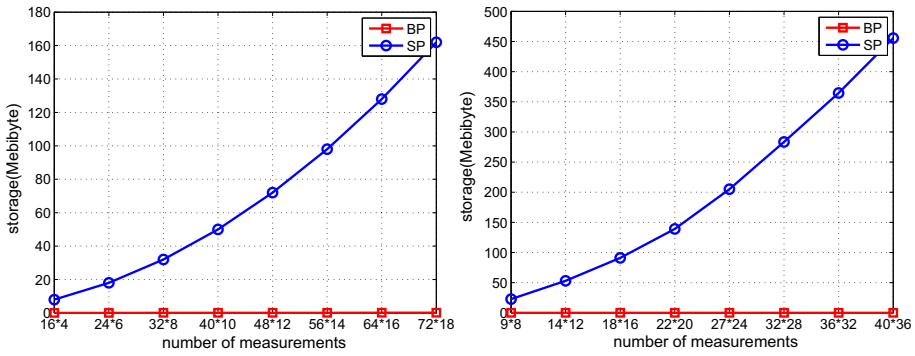
Figures 2 and 3 present the comparison between SMM with bilinear projection and single linear projection in terms of the computational time and storage cost. In case of bilinear projection, the computational time and storage cost for generating compressed measurements are significantly reduced compared to a single linear projection.

As shown in Fig. 3, the storage space required by single linear projection is much more than bilinear projection under the same number of measurements. Thus, we consider to increase the number of measurements for bilinear projection, whose storage requirement is still smaller than single linear projection. Fig. 4 shows the classification accuracy for SMM-BP, SMM-SP, SVM-BP and SVM-SP, where the top abscissa axis describes the number of



(a) EEG alcoholism **(b)** FEI face

Fig. 2 The comparison of the computation time between the single linear projection and the bilinear projection



(a) EEG alcoholism **(b)** FEI face

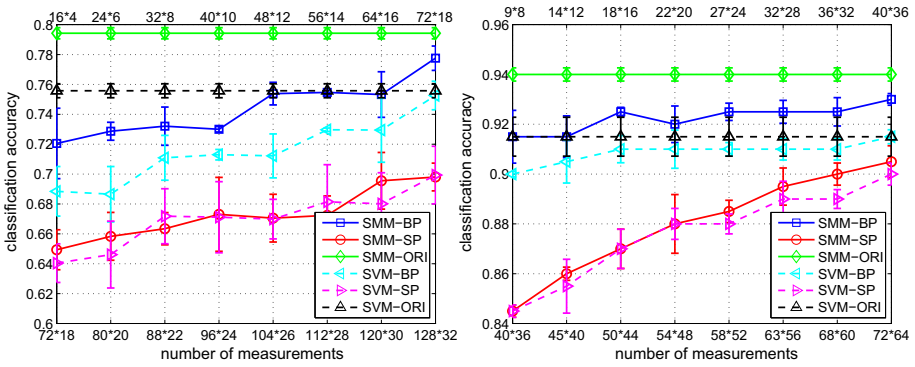
Fig. 3 The comparison of the storage cost between the single linear projection and the bilinear projection

measurements for single linear projection and the bottom abscissa axis describes the number of measurements for bilinear projection. We can see that SMM and SVM can achieve higher accuracy on the bilinear projected measurements while still retain a smaller storage cost compared with the single projected measurements. Thus from the view of storage cost, it's a good choice to use bilinear projection. Besides, SMM-BP can reach an even higher accuracy than SVM-BP, also reflects that SMM can leverage the structure information while SVM can not.

5.2 Experiments on vector representation data sets

The DBWorld e-mails data set consists of 64 e-mails from DBWorld newsletter, announces conferences, jobs, books, software and grants. Every e-mail is represented as a vector containing 5704 binary values, and separated into two classes, for 1 if the sample is an announcement of conference, 0 otherwise. For the convenience of the data matrixing, we drop the last four features, and then reshape the data to a 94×50 matrix data without overlapping.

The p53 Mutants data set is utilized as the benchmark data set to predict the transcriptional activity (active vs inactive) based on data extracted from biophysical simulations. There are



(a) EEG alcoholism **(b)** FEI face

Fig. 4 Classification accuracy for SMM-BP, SVM-BP and SMM-SP, SVM-BP, SVM-SP on EEG alcoholism and FEI face with increased number of measurements for BP

Table 2 Summary of vector representation data sets

Data sets	Positive	Negative	Matrix dimension	Vector dimension
DBWorld e-mails	29	35	94×50	4700
p53 Mutants	151	349	90×60	5400

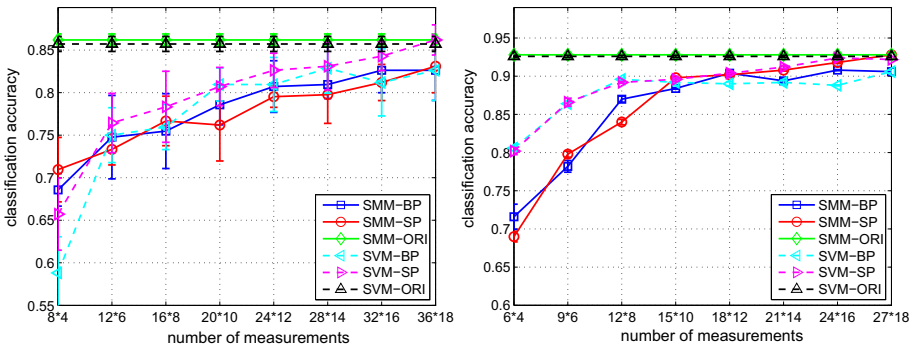
a total of 31,420 instances, each instance contains 5408 attributes. We randomly selected 500 instances for our experiment. For the convenience of the data matrixing, we drop the last nine features, and then reshape the data to a 90×60 matrix data without overlapping.

Table 2 summarizes the characteristics of vector representation data sets.

Figure 5 presents the classification accuracy for SMM and SVM with bilinear projection and single linear projection on vector data sets (DBworld e-mails and p53 mutants) according to different number of measurements. The Experiment results of Wang and Chen (2007) and Wang et al. (2013) have shown that different matrix sizes would lead to different classification results. In this paper, we don't concern about matrixing of the vector data, thus we fix the matrix size in our experiments. Although SMM-BP can not outperform other three algorithms, it can achieve comparable results with others while obtain a less storage burden. Thus, we can also perform 2D compressed learning on vector data from the perspective of storage saving.

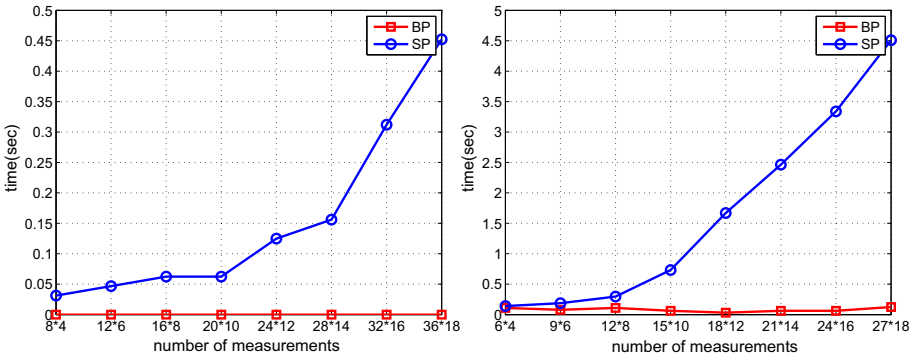
Figures 6 and 7 present the comparison between SMM with bilinear projection and single linear projection in terms of the computational time and storage cost. In case of bilinear projection, the computational time and storage cost for generating compressed measurements are significantly reduced compared to the single linear projection.

We also consider to increase the number of measurements for bilinear projection on vector representation data sets, whose storage requirement is still smaller than single linear projection. Fig. 8 shows the classification accuracy for SMM-BP, SMM-SP, SVM-BP and SVM-SP, where the top abscissa axis describes the number of measurements for single linear projection and the bottom abscissa axis describes the number of measurements for bilinear projection. The results also demonstrate the storage saving of bilinear projection. Although



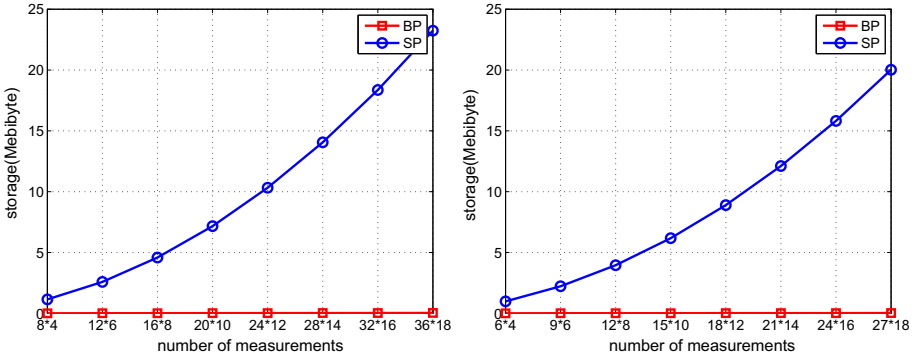
(a) DBWorld e-mails **(b)** p53 mutants

Fig. 5 Classification accuracy for SMM and SVM with a single linear projection and a bilinear projection on DBworld e-mails and p53 mutants with respect to the different number of measurements



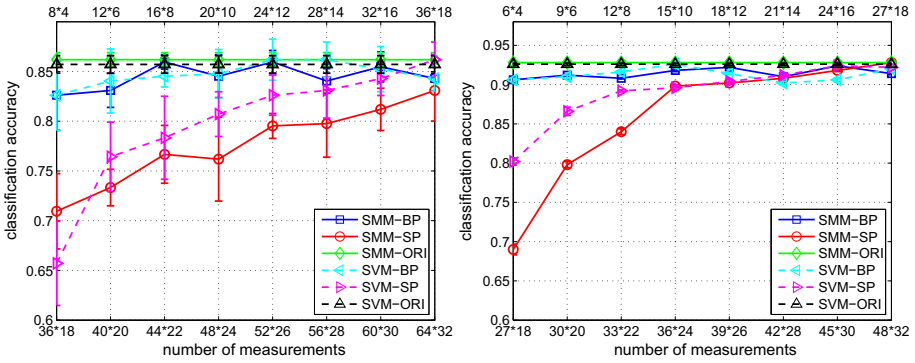
(a) DBworld e-mails **(b)** p53 mutants

Fig. 6 The comparison of the computation time between the single linear projection and the bilinear projection



(a) DBWorld e-mails **(b)** p53 mutants

Fig. 7 The comparison of the storage cost between the single linear projection and the bilinear projection



(a) DBWorld e-mails **(b)** p53 mutants

Fig. 8 Classification accuracy for SMM-BP, SVM-BP and SMM-SP, SVM-BP, SVM-SP on DBworld e-mails and p53 mutants with increased number of measurements for BP

SMM-BP can't reach a higher accuracy than SVM-BP on vector data, it can obtain comparable results.

6 Conclusion

In this paper, we have considered a 2D compressive classification problem that learns the SMM classifier using the KCS measurements realized by a bilinear projection. KCS can preserve the structure presented in each dimension and SMM can leverage the structure of the KCS measurements for improving classification accuracy. We have provided theoretical analysis to show the feasibility of our method and demonstrated that: (1) The RIP condition holds for the bilinear projection; (2) The computational and space complexities of KCS are lower than the regular CS; (3) The performance of the SMM in the Kronecker compressed domain is close to that in the original domain. Experiments on real-world datasets also showed the feasibility of our method. Future directions include incorporating the nonlinear technique to handle the linearly non-separable problems and the sketching technique to handle the large scale problem.

Acknowledgements We would like to express our appreciation to the editors and the reviewers, who have greatly helped us in improving the quality of the paper. This work is supported by the National Natural Science Foundation of China (NSFC) under Grant No. 61672281 and the Key Program of NSFC under Grant No. 61732006.

Appendices

Appendix 1: The Proof of Theorem 1

Proof The optimization problem (2) can be rewritten as

$$\begin{aligned}\min_{W,\xi} \quad & \frac{1}{2}\|W\|_F^2 + \tau\|W\|_* + \frac{C}{N}\sum_{i=1}^{N}\xi_i \\ \text{s.t.} \quad & y_i \text{tr}\left(W^T X_i\right) \geq 1 - \xi_i \quad i = 1, \ldots, N \\ & \xi_i \geq 0 \quad i = 1, \ldots, N\end{aligned}$$

The Lagrangian function is as follows:

$$\begin{aligned}L(W, \xi, \alpha, \gamma) = & \frac{1}{2}\|W\|_F^2 + \tau\|W\|_* + \frac{C}{N}\sum_{i=1}^{N}\xi_i \\ & - \sum_{i=1}^{N}\alpha_i\left[y_i \text{tr}\left(W^T X_i\right) - 1 + \xi_i\right] - \sum_{i=1}^{N}\gamma_i\xi_i\end{aligned} \tag{31}$$

Setting the derivative of L with respect to ξ to be 0, we have

$$\gamma_i = \frac{C}{N} - \alpha_i \geq 0, i = 1, \ldots, N. \tag{32}$$

Substituting (32) into (31) to eliminate ξ_i and γ_i, we obtain the dual problem as

$$L(W, \alpha) = \frac{1}{2}\|W\|_F^2 + \tau\|W\|_* - \sum_{i=1}^{N}\alpha_i\left[y_i \text{tr}\left(W^T X_i\right) - 1\right] \tag{33}$$

where $0 \leq \tilde{\alpha}_i \leq \frac{C}{N}$. The optimal solution of problem (33) is given by Cai et al. (2010) as,

$$\tilde{W} = \mathcal{D}_\tau\left(\sum_{i=1}^{N}\tilde{\alpha}_i y_i X_i\right),$$

where $\tilde{\alpha}$ is the corresponding value of Lagrangian multiplier when $\tilde{W}$ is the optimal solution. According to the dual theorem,

$$\begin{aligned}\frac{1}{2}\|\tilde{W}\|_F^2 + \tau\|\tilde{W}\|_* &\leq \frac{1}{2}\|\tilde{W}\|_F^2 + \tau\|\tilde{W}\|_* + \frac{C}{N}\sum_{i=1}^{N}\xi_i \\ &= \frac{1}{2}\|\tilde{W}\|_F^2 + \tau\|\tilde{W}\|_* - \sum_{i=1}^{N}\tilde{\alpha}_i\left[y_i \text{tr}\left(\tilde{W}^T X_i\right) - 1\right]\end{aligned}$$

Hence,

$$\text{tr}\left(\tilde{W}^T\sum_{i=1}^{N}\tilde{\alpha}_i y_i X_i\right) \leq \sum_{i=1}^{N}\tilde{\alpha}_i \leq C \tag{34}$$

Let the linear combination $\sum_{i=1}^{N}\tilde{\alpha}_i y_i X_i$ have the condensed SVD of the following form,

$$\sum_{i=1}^{N}\tilde{\alpha}_i y_i X_i = \tilde{U}_0\tilde{\Sigma}_0\tilde{V}_0^T + \tilde{U}_1\tilde{\Sigma}_1\tilde{V}_1^T, \tag{35}$$

where $\tilde{\Sigma}_0$ is the diagonal matrix whose diagonal entries are greater than τ, $\tilde{U}_0$ and $\tilde{U}_0$ are matrices of the corresponding left and right singular vectors; $\tilde{\Sigma}_1$, $\tilde{U}_1$ and $\tilde{U}_1$ correspond the rest parts of the SVD whose singular values $0 < \sigma \leq \tau$. Define $\tilde{\Lambda} =$

$-\tau\left(\tilde{U}_0\tilde{V}_0^T + \frac{1}{\tau}\tilde{U}_1\tilde{\Sigma}_1\tilde{V}_1^T\right)$ and substituting $\tilde{\Lambda}$ into (35)

$$\tilde{W} = \tilde{U}_0\left(\tilde{\Sigma}_0 - \tau\right)\tilde{V}_0^T = \tilde{\Lambda} + \sum_{i=1}^{N}\tilde{\alpha}_i y_i X_i. \tag{36}$$

Substituting (36) into (34), we have

$$\|\tilde{W}\|_F^2 - \mathrm{tr}\left(\tilde{\Lambda}^T\tilde{W}\right) = \mathrm{tr}\left(\tilde{W}^T\sum_{i=1}^{N}\tilde{\alpha}_i y_i X_i\right) \le C. \tag{37}$$

Furthermore, using Eq. (1) we have,

$$-\frac{1}{\tau}\tilde{\Lambda} \in \partial\|W\|_*\big|_{W=\tilde{W}},$$

and,

$$\left\|\tilde{W}\right\|_* = -\frac{1}{\tau}\mathrm{tr}\left(\tilde{\Lambda}^T\tilde{W}\right). \tag{38}$$

Substituting (38) into (37), we have

$$\|\tilde{W}\|_F^2 + \tau\|\tilde{W}\|_* \le C.$$

□

Appendix 2: The Proof of Theorem 5

Proof According to Lemma 3, we have

$$\begin{aligned}&1 - y\mathrm{tr}\left(\left(\Phi_1\tilde{W}\Phi_2^T\right)^T\left(\Phi_1 X\Phi_2^T\right)\right)\\ &\le 1 - y\mathrm{tr}\left(\tilde{W}^T X\right) + O\left(\left(R^2C + \tau\tilde{r}R\right)\delta_{2r}\left(\Phi_2\otimes\Phi_1\right)\right).\end{aligned}$$

Since the measurement matrix forms a one-to-one mapping from the data domain to measurement domain, we can take the expectation of the hinge loss as:

$$H_\Phi\left(\Phi_1\tilde{W}\Phi_2^T\right) \le H\left(\tilde{W}\right) + O\left(\left(R^2C + \tau\tilde{r}R\right)\delta_{2r}\left(\Phi_2\otimes\Phi_1\right)\right).$$

Thus the hinge loss $H\left(\tilde{W}\right)$ can be preserved after bilinear random projection. According to Lemma 2, the near isometry property holds for the spectral elastic net, thus

$$\frac{1}{2C}\|\Phi_1\tilde{W}\Phi_2^T\|_F^2+\frac{\tau}{C}\|\Phi_1\tilde{W}\Phi_2^T\|_* \leq \frac{1}{2C}\|\tilde{W}\|_F^2+\frac{\tau}{C}\|\tilde{W}\|_*+O\left(\left(R^2C+\frac{\tau^2\tilde{r}}{C}\right)\delta_{2r}\left(\Phi_2\otimes\Phi_1\right)\right)$$

Then we can complete the proof,

$$\begin{aligned}L_\Phi\left(\Phi_1\tilde{W}\Phi_2^T\right)&=H_\Phi\left(\Phi_1\tilde{W}\Phi_2^T\right)+\frac{1}{2C}\|\Phi_1\tilde{W}\Phi_2^T\|_F^2+\frac{\tau}{C}\|\Phi_1\tilde{W}\Phi_2^T\|_*\\&\leq H\left(\tilde{W}\right)+\frac{1}{2C}\|\tilde{W}\|_F^2+\frac{\tau}{C}\|\tilde{W}\|_*\\&\quad+O\left(\left(R^2C+\frac{\tau^2\tilde{r}}{C}\right)\delta_{2r}\left(\Phi_2\otimes\Phi_1\right)\right)\\&=L\left(\tilde{W}\right)+O\left(\left(R^2C+\frac{\tau^2\tilde{r}}{C}\right)\delta_{2r}\left(\Phi_2\otimes\Phi_1\right)\right)\end{aligned}$$

□

Appendix 3: The Proof of Theorem 6

Proof For each W, we define $g_W(X,y)=\ell(W;X,y)-\ell(W^\star;X,y)$, our goal is to bound the expectation of g_W in terms of its empirical average. We denote $\mathcal{G}=\{g_W|W\in\mathcal{W}\}$. Instead of bounding the variation between the expected and the empirical values of $g_W\in\mathcal{G}$ in terms of the complexity of $\mathcal{G}$, we use the complexity of an alternative class of functions, which ignores the spectral elastic net penalty $r(W)$. Define

$$\mathcal{H}=\left\{h_W=g_W-(r(W)-r(W^\star)):g_W\in\mathcal{G}\right\}$$

With this definition, we have

$$\mathbb{E}[g_W]-\hat{\mathbb{E}}[g_W]=\mathbb{E}[h_W]-\hat{\mathbb{E}}[h_W] \tag{39}$$

hence it is enough to bound the right hand side of (39), which can be done by the Rademacher complexity of the class $\mathcal{R}(\mathcal{H})$ (Bartlett and Mendelson 2002), i.e., for any $\delta>0$, with probability $1-\delta$,

$$\sup_{h_W\in\mathcal{H}}\mathbb{E}[h_W]-\hat{\mathbb{E}}[h_W]\leq 2\mathcal{R}(\mathcal{H})+\left(\sup_{h_W\in\mathcal{H},X,y}|h_W(X,y)|\right)\sqrt{\frac{\log(1/\delta)}{2N}}$$

From the definition of h_W, the Lipschitz continuity of the hinge loss, and the bound $\|X\|_F\leq R$, we have

$$|h_W(X,y)|=|h(\langle W,X\rangle,y)-h(\langle W^\star,X\rangle,y)|\leq R\|W-W^\star\|_F \tag{40}$$

Recall that we have restricted our analysis in the hypothesis space $\mathcal{W}$, then using the following inequalities which hold for any matrix X,

$$\|X\|_F\leq\|X\|_*$$

we have

$$\|W\|_F\leq\sqrt{C+\frac{\tau^2}{4}}-\frac{\tau}{2}$$

For the true minimizer $W^\star$, we have

$$\mathbb{E}[\ell(\langle W^\star, X\rangle, y)] = \mathbb{E}[h(\langle W^\star, X\rangle, y)] + \frac{1}{2C}\|W^\star\|_F^2 + \frac{\tau}{C}\|W^\star\|_* \le L(0) \le 1$$

hence we can conclude

$$\|W^\star\|_F \le \sqrt{2C+\tau^2} - \tau,\ \|W\|_F \le \sqrt{C + \frac{\tau^2}{4}} - \frac{\tau}{2} \le \sqrt{2C+\tau^2} - \tau \tag{41}$$

Substituting (41) into (40) yields

$$|h_W(X, y)| \le \sqrt{2}R\left(\sqrt{2C+\tau^2} - \tau\right) \tag{42}$$

The Rademacher complexity can be upper bounded by

$$\begin{aligned}
\mathcal{R}(\mathcal{H}) &= \frac{1}{N}\mathbb{E}\left[\sup_{W\in\mathcal{W}} \sum_{i=1}^N \sigma_i h_W(X_i, y_i)\right] \\
&\le \frac{1}{N}\mathbb{E}\left[\sup_{W\in\mathcal{W}} \sum_{i=1}^N \sigma_i \mathrm{tr}\left((W - W^\star)^T X_i\right)\right] \\
&\le \frac{1}{N}\mathbb{E}\left[\sup_{W\in\mathcal{W}} \|W - W^\star\|_F \left\|\sum_{i=1}^N \sigma_i X_i\right\|_F\right] \\
&\le \frac{\sqrt{2}R\left(\sqrt{2C+\tau^2} - \tau\right)}{\sqrt{N}}
\end{aligned}$$

From the above, for any $\delta > 0$ with probability at least $1 - \delta$ we have

$$\mathbb{E}[g_W] - \hat{\mathbb{E}}[g_W] = \mathbb{E}[h_W] - \hat{\mathbb{E}}[h_W] \le O\left(\left(\sqrt{2C+\tau^2} - \tau\right)\sqrt{\frac{R^2\log(1/\delta)}{N}}\right) \tag{43}$$

□

Appendix 4: The Proof of Theorem 7

Proof By definition of the true regularization loss we have,

$$H_\Phi\left(\tilde{W}_\Phi\right) \le H_\Phi\left(\tilde{W}_\Phi\right) + \frac{1}{2C}\left\|\tilde{W}_\Phi\right\|_F^2 + \frac{\tau}{C}\left\|\tilde{W}_\Phi\right\|_* = L_\Phi\left(\tilde{W}_\Phi\right)$$

According to Theorem 6,

$$L(\tilde{W}) - L(W^\star) \le \hat{L}(\tilde{W}) - \hat{L}(W^\star) + O\left(\left(\sqrt{2C+\tau^2} - \tau\right)\sqrt{\frac{R^2\log(1/\delta)}{N}}\right)$$

$$L_\Phi(\tilde{W}_\Phi) - L_\Phi(W_\Phi^\star) \le \hat{L}_\Phi(\tilde{W}_\Phi) - \hat{L}_\Phi(W_\Phi^\star) + O\left(\left(\sqrt{2C+\tau^2} - \tau\right)\sqrt{\frac{R^2\log(1/\delta)}{N}}\right)$$

Besides, since the SMM classifier $\tilde{W}$ minimizes the empirical regularization loss,

$$\hat{L}(\tilde{W}) \leq \hat{L}(W^{\star})$$
$$\hat{L}_{\Phi}(\tilde{W}_{\Phi}) \leq \hat{L}_{\Phi}(W_{\Phi}^{\star})$$

we have,

$$L(\tilde{W}) \leq L(W^{\star}) + O\left(\left(\sqrt{2C+\tau^2}-\tau\right)\sqrt{\frac{R^2\log(1/\delta)}{N}}\right)$$
$$L_{\Phi}(\tilde{W}_{\Phi}) \leq L_{\Phi}(W_{\Phi}^{\star}) + O\left(\left(\sqrt{2C+\tau^2}-\tau\right)\sqrt{\frac{R^2\log(1/\delta)}{N}}\right)$$

As W_{Φ}^{*} is the best SMM classifier in measurement domain, then

$$L_{\Phi}\left(W_{\Phi}^{*}\right) \leq L_{\Phi}\left(\Phi_1\tilde{W}\Phi_2^T\right)$$

Theorem 5 connects the regularization loss of the SMM classifier $\tilde{W}$ in data domain and the regularization loss of the projected classifier $\Phi_1\tilde{W}\Phi_2^T$

$$L_{\Phi}\left(\Phi_1\tilde{W}\Phi_2^T\right) \leq L\left(\tilde{W}\right) + O\left(\left(R^2C + \frac{\tau^2\tilde{r}}{C}\right)\delta_{2r}\left(\Phi_2\otimes\Phi_1\right)\right)$$

In particular, let W_0 be a good SMM classifier with small true spectral elastic net penalty. By the definition of W^*

$$L(W^*) \leq L(W_0)$$

Putting above inequalities together, we get

$$\begin{aligned} H_{\Phi}\left(\tilde{W}_{\Phi}\right) &\leq H(W_0) + \frac{1}{2C}\|W_0\|_F^2 + \frac{\tau}{C}\|W_0\|_* \\ &+ O\left(\left(R^2C + \frac{\tau^2\tilde{r}}{C}\right)\delta_{2r}\left(\Phi_2\otimes\Phi_1\right) + \left(\sqrt{2C+\tau^2}-\tau\right)\sqrt{\frac{R^2\log(1/\delta)}{N}}\right) \end{aligned} \tag{44}$$

To balance the terms, we need to choose an appropriate C. It is difficult to find the optimal C for Eq. (44) directly, we in turn relax the right hand side of it and find the optimal C for the relaxed upper bound. Noting that $\sqrt{2C+\tau^2}-\tau \leq \frac{C}{\tau+a}$ for some small constant a, thus we obtain the relaxed upper bound as

$$\begin{aligned} H_{\Phi}\left(\tilde{W}_{\Phi}\right) &\leq H(W_0) + \frac{1}{2C}\|W_0\|_F^2 + \frac{\tau}{C}\|W_0\|_* \\ &+ O\left(\left(R^2C + \frac{\tau^2\tilde{r}}{C}\right)\delta_{2r}\left(\Phi_2\otimes\Phi_1\right) + \sqrt{\frac{R^2C^2\log(1/\delta)}{(\tau+a)^2N}}\right) \end{aligned} \tag{45}$$

Considering R and $\tilde{r}$ as fixed constants and choose a C which minimizes the relaxed upper bound (45) we get

$$H_{\Phi}(\tilde{W}_{\Phi}) \leq H(W_0) + O\left(\sqrt{\left(\frac{1}{2}\|W_0\|_F^2 + \tau\|W_0\|_* + \tau^2\delta_{2r}(\Phi_2 \otimes \Phi_1)\right)\left(\delta_{2r}(\Phi_2 \otimes \Phi_1) + \sqrt{\frac{\log(1/\delta)}{(\tau+a)^2 N}}\right)}\right)$$

□

References

Baraniuk, R., Davenport, M., DeVore, R., & Wakin, M. (2008). A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, *28*(3), 253–263.

Bartlett, P. L., & Mendelson, S. (2002). Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, *3*, 463–482.

Cai, J. F., Candès, E. J., & Shen, Z. (2010). A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, *20*(4), 1956–1982.

Calderbank, R., Jafarpour, S., & Schapire, R. (2009). Compressed learning: Universal sparse dimensionality reduction and learning in the measurement domain. Technical report, Rice University

Candes, E. J., & Tao, T. (2006). Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, *52*(12), 5406–5425.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Danziger, S. A., Swamidass, S. J., Zeng, J., Dearth, L. R., Lu, Q., Chen, J. H., et al. (2006). Functional census of mutation sequence spaces: The example of p53 cancer rescue mutants. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, *3*(2), 114–125.

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

Duarte, M. F., & Baraniuk, R. G. (2012). Kronecker compressive sensing. *IEEE Transactions on Image Processing*, *21*(2), 494–504.

Duarte, M. F., Davenport, M. A., Takhar, D., Laska, J. N., Sun, T., Kelly, K. E., et al. (2008). Single-pixel imaging via compressive sampling. *IEEE Signal Processing Magazine*, *25*(2), 83.

Filannino, M. (2011). Dbworld e-mail classification using a very small corpus. The University of Manchester.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning. Springer series in statistics* (Vol. 1). New York: Springer.

Jokar, S., & Mehrmann, V. (2012). Sparse representation of solutions of kronecker product systems. Mathematics

Luo, L., Xie, Y., Zhang, Z., & Li, W. J. (2015). Support matrix machines. In *Proceedings of the 32nd international conference on machine learning (ICML-15)* (pp. 938–947).

Lustig, M., Donoho, D. L., Santos, J. M., & Pauly, J. M. (2008). Compressed sensing MRI. *IEEE Signal Processing Magazine*, *25*(2), 72–82.

Maillard, O., & Munos, R. (2009). Compressed least-squares regression. In *Advances in neural information processing systems* (pp. 1213–1221).

Reboredo, H., Renna, F., Calderbank, R., & Rodrigues, M. R. (2013). Compressive classification. In *2013 IEEE international symposium on information theory proceedings (ISIT)* (pp. 674–678). IEEE

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

Rish, I., & Grabarnik, G. (2014). *Sparse modeling: Theory, algorithms, and applications*. Boca Raton: CRC Press.

Thomaz, C. E., & Giraldi, G. A. (2010). A new ranking method for principal components analysis and its application to face image analysis. *Image and Vision Computing*, *28*(6), 902–913.

Wang, Z., & Chen, S. (2007). New least squares support vector machines based on matrix patterns. *Neural Processing Letters*, *26*(1), 41–56.

Wang, Z., Zhu, C., Gao, D., & Chen, S. (2013). Three-fold structured classifier design based on matrix pattern. *Pattern Recognition*, *46*(6), 1532–1555.

Wolf, L., Jhuang, H., & Hazan, T. (2007). Modeling appearances with low-rank SVM. In *IEEE conference on computer vision and pattern recognition, 2007. CVPR'07* (pp. 1–6). IEEE

Zhou, H., & Li, L. (2014). Regularized matrix regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(2), 463–483.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320.