ECE543 Project Report: Learning in Compressed Spaces

Yuqi Li

***Abstract*—Compressed learning is based on the principle of compressed sensing: a high dimensional signal can be recovered from its low dimensional random projection, if the signal itself satisfies the sparsity property and the number of measurement is sufficiently large. This random projection itself serves a universal dimension reduction technique. We investigated learning task in the compressed domain, induced by this random projection. There do exist learning bound in compressed domain, which relies on the properties of the measurement matrix, such as preservation of relative geometry.**

I. Introduction

Learning in high dimension often suffers from curse of dimensionality. A general strategy is to transform the high dimension data to low dimension measurements, or features, which resides in the compressed domain. For example, linear dimension reduction includes principle component analysis (PCA) that preserves the data variance, canonical correlation analysis (CCA) that keeps the correlation between a pair of data [1], and non-linear techniques includes Isomap [2] and t-SNE [3]. In this report, we will consider a universal dimension reduction method used in compressed sensing and its effect in the subsequent classification and regression tasks. The compressed sensing theory [4] proved that the data vector can be recovered exactly if it can be sparsified in another basis and the dimension of compressed domain satisfies certain conditions. This exact recovery provides feasibility to perform high-level tasks like classification and regression directly in the compressed domain. However, recovery the original signal from the lower dimension measurements is expensive as well as the learning in high dimension spaces. If we are only concerned about the result of learning, then the reconstruction is not necessary at all, as the measurements preserves most of the information in measurements, such as relative geometry and learnability. In this report, we will review two papers [5], [6], that gives learning bounds of compressed learning: regression and classification in comrpessed domain.

II. Least Square Regression in Compressed Domain [5]

A. Problem Statement

Given dataset $\mathcal{D}_K = \{x_k, y_k\}_{k=1}^K, x_k \in \mathcal{X}, y_k \in \mathbb{R}$, each pair is i.i.d. samples from distribution P. In detail, $x_k \overset{\text{i.i.d.}}{\sim} P_{\mathcal{X}}$, and $y_k = f^*(x_k) + \eta(x_k), \mathbb{E}[\eta(x_k)] = 0, \text{var}(\eta(x_k)) = \sigma^2(x_k)$, where f^* is the unknown target function and η is the noise. The goal is to recover this f^* from the dataset $\mathcal{D}_K$, the performance of recovery is assessed by the generalization risk on a fresh sample:

$$L(f) \overset{\text{def}}{=} \mathbb{E}_{(X,Y)\sim P}\left[(Y - f(X))^2\right] \tag{1}$$

and the empirical risk is defined as:

$$L_K(f) \overset{\text{def}}{=} \frac{1}{K}\sum_{k=1}^{K}\left[y_k - f(x_k)\right]^2 \tag{2}$$

In the following, we will use

$$\|f - f^*\|_P^2 \overset{\text{def}}{=} \mathbb{E}_{X\sim P_{\mathcal{X}}}[(f(X) - f^*(X))^2] \tag{3}$$

to denote the L_2 norm of difference in f and f^* according to distribution P. This quantity coincides with **excess risk** (see A):

$$\text{excess risk} = L(\widehat{f}) - L(f^*) = \left\|\widehat{f} - f^*\right\|_P^2 \tag{4}$$

If the optimal target funtion is not in the search space $\mathcal{F}$, then this excess risk can be divided into two parts: **estimation error** $L(\widehat{f}) - \inf_{f\in\mathcal{F}} L(f)$ and the **approximation error** $\inf_{f\in\mathcal{F}} L(f) - L(f^*) = \inf_{f\in\mathcal{F}} \|f - f^*\|_P^2$.
In this paper, linear regression in the feature space is considered and the feature map is defined as:

$$\varphi : \mathcal{X} \to \mathbb{R}^N, \quad \varphi_n(x) \in \mathbb{R} \tag{5}$$

The search space of this linear function can be defined as

$$\mathcal{F}_N \overset{\text{def}}{=} \left\{f_\alpha \overset{\text{def}}{=} \alpha^T\varphi, \alpha \in \mathbb{R}^N\right\} \tag{6}$$

and the ordinary least square (OLS) regression in the data domain searches for the best linear coefficients $\alpha \in \mathbb{R}^N$ in terms of empirical risk,

$$\min_{\alpha\in\mathbb{R}^N} \frac{1}{K}\sum_{k=1}^{K}\|y_k - \alpha^T\varphi(x_k)\|^2 = \min_{f_\alpha\in\mathcal{F}_N} L_K(f_\alpha) \tag{7}$$

This paper is interested in the case where N is very large such that the approximation error is small. But in this case, overfitting is likely to occur. There are several approaches to regularize the solution, one is to add L_1 (Lasso) or L_2 (Tikhonov) penalty of the weight, another is to find the minimizer of empirical error with minimal L_2 norm.
The dimension reduction is confined to be a random linear transform $A \in \mathbb{R}^{M\times N}$ and the features is denoted by

$$\psi : \mathcal{X} \to R(A), \quad \psi(x) = A\varphi(x) \in \mathbb{R}^M, M < N \tag{8}$$

This random matrix A can be a Gaussian random matrix, random Hadamard matrix. From Johnson-Lindenstrauss Lemma [7], the norm of the data vectors are approximately preserved

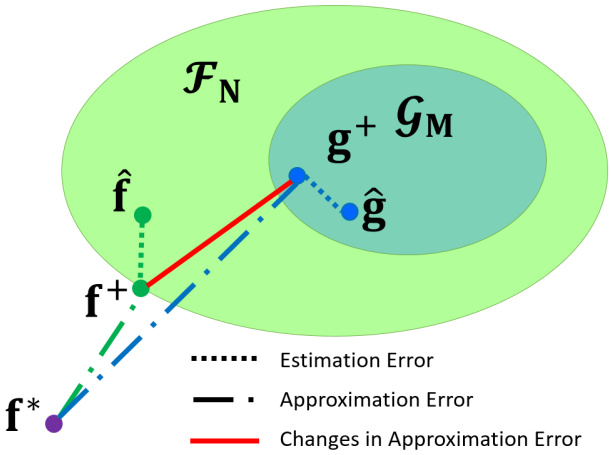

Fig. 1: Regression on the data domain and compressed domain

in random projection, as well as their relative geometry distribution. After the data vectors are projected to compressed domain, the search space is now:

$$\mathcal{G}_M \overset{\text{def}}{=} \left\{ g_\beta = \beta^T \psi, \beta \in \mathbb{R}^M \right\} \tag{9}$$

and $\mathcal{G}_M \subset \mathcal{F}_N$ (see B).

The procedure of bounding the excess risk in the compressed domain is as follows: (see fig. 1):

1) The estimation error (dotted line) is bounded and is decreasing as K increases
2) The expected excess risk (sum of green lines) in data domain is bounded using the approximation error (green dot-dash line)
3) The approximation error is increased when the search space is reduced, but the difference (red solid line) is bounded.
4) The expected excess risk in the compressed domain is bounded (sum of two blue lines).

B. Approximation Error

The following lemma is a variation of Johnson-Lindenstrauss lemma. It states that the inner product of data vectors is also preserved during the projection.

Lemma 1. *Let A be a $M \times N$ random Gaussian matrix, $(u_k)_{k=1}^K$ be one of the K data vectors, and v be any vector from $\mathbb{R}^N$. For any $\varepsilon > 0, \delta > 0$, for $M \geq \frac{1}{\frac{\varepsilon^2}{4} - \frac{\varepsilon^3}{6}} \log \frac{4K}{\delta}$, with probability at least $1 - \delta$, we have*

$$\left| (Au_k)^T Av - u_k^T v \right| \leq \varepsilon \|u_k\| \|v\|, \forall k \in [K] \tag{10}$$

Proof: See C ■

Using lemma 1, the following theorem bounds the changes in approximation error between $\mathcal{G}_M$ and $\mathcal{F}_N$. This applies to linear algorithms like OLS and regularized method like Lasso and ridge regression. The link starts from the optimal linear regressor in $\mathcal{F}_N$.

Theorem 1. *For any $\delta > 0, M \geq 15 \log(8K/\delta)$, let A be a $M \times N$ random Gaussian matrix and $\mathcal{G}_M$ be be the compressed domain resulting from this choice of A, let $\alpha^+ = \arg\min_{\alpha \in \mathbb{R}^N} L(f_\alpha) - L(f^*)$. Then with probability $1 - \delta$, we have*

$$\inf_{g \in \mathcal{G}_M} \|g - f^*\|_P^2 \leq \inf_{f \in \mathcal{F}_N} \|f - f^*\|_P^2 + \frac{8 \log(8K/\delta)}{M} \|\alpha^+\|^2 \left(\mathbb{E}\left[\|\varphi(X)\|^2 \right] + 2 \sup_{x \in \mathcal{X}} \|\varphi(x)\|^2 \sqrt{\frac{\log 4/\delta}{2K}} \right) \tag{11}$$

Proof: See D ■

This bound in approximation error assessed in $\mathcal{G}_M$ increases by at most $O\left(\frac{\log(K/\delta)}{M}\right) \|\alpha^+\|^2 \mathbb{E}\|\varphi(X)\|^2$ compared to that in $\mathcal{F}_N$. If we have infinity number of samples $K \to \infty$, this bound would solely depend on $\|\alpha^+\|^2 \mathbb{E}\left[\|\varphi(X)\|^2\right]$.

C. Expected Excess Risks

In the following, the expected excess loss is analyzed for least-square regression [8]. There is one explicit assumption: $\|f\|_\infty \overset{\text{def}}{=} \max_{x \in \mathcal{X}} |f^*(x)| \leq B < \infty$. The ordinary LS regression with minimum L_2 norm would yield:

$$\widehat{\alpha} = \text{ argmin } \|\alpha\| \text{ s.t. } \alpha \text{ minimizes } \|Y - \Phi\alpha\| \tag{12}$$

Where $Y \in \mathbb{R}^K$ is concatenation of y_k and $\Phi \in \mathbb{R}^{K \times N}$ is the concatenation of $\varphi(x_k)$. From normal equation $\Phi\Phi^T\widehat{\alpha} = \Phi^T Y$, we can write $\widehat{\alpha} = \Phi^\dagger Y$ where $\Phi^\dagger$ is the pseudo inverse of Φ. The truncated regressor can be written as:

$$\widehat{f}_B(x) \overset{\text{def}}{=} T_B\left[f_{\widehat{\alpha}}(x)\right], T_B(u) \overset{\text{def}}{=} \begin{cases} u & \text{if } |u| \leq B \\ L \operatorname{sign}(u) & \text{otherwise} \end{cases} \tag{13}$$

The expected excess risk of $\widehat{f}_B$ is bounded as

$$\mathbb{E}\left(\left\| \widehat{f}_B - f^* \right\|_P^2 \right) \leq c' \max\left\{\sigma^2, B^2\right\} \frac{1 + \log K}{K} N + 8 \inf_{f \in \mathcal{F}_N} \|f - f^*\|_P^2 \tag{14}$$

where a bound on c' is 9216. Another simpler bound is to consider the expectation $\mathbb{E}_Y$ condtionally on input data and resulting empirical distribution P_K:

$$\mathbb{E}_Y\left(\left\| \widehat{f}_B - f^* \right\|_{P_K}^2 \right) \leq \sigma^2 \frac{N}{K} + \inf_{f \in \mathcal{F}} \|f - f^*\|_{P_K}^2 \tag{15}$$

These two expected excess risk bounds are built on the basis of known approximation error. In case of $N \to \infty$, the approximation would approach zero. In the following, the expected excess risk in compressed domain can be bounded combining II-B.

D. Excess Risk of Compressed Least-Square Regression (CLSR) - Main Result

In compressed domain, we aim to search for the best linear regressor in $\mathcal{G}_M$. Same as before, the optimal trucated regressor is:

$$\widehat{g}_B(x) \overset{\text{def}}{=} T_B\left[g_{\widehat{\beta}}(x)\right], \widehat{\beta} = \Psi^\dagger Y \tag{16}$$

where $\Psi \in \mathbb{K} \times \mathbb{M}$ is defined similar to Φ, the concatenation of $\psi(x_k)$. From theorem 1, we can obtain the following corollary of CLSR

Corollary 1. *For any $\delta > 0$, set $M = 8\frac{\|\alpha^+\|\sqrt{\mathbb{E}\|\varphi(X)\|^2}}{\max(\sigma,B)}\sqrt{\frac{K\log(8K/\delta)}{c'(1+\log K)}}$, then if $M \geq 15\log(8K/\delta)$, with probability $1-\delta$, the expected excess risk of the CLSR estimate is bounded as:*

$$\begin{aligned}\mathbb{E}\left(\|\widehat{g}_B - f^*\|_P^2\right) \leq & 16\sqrt{c'}\max\{\sigma, B\}\|\alpha^+\|\sqrt{\mathbb{E}\|\varphi(X)\|^2} \\ & \times\sqrt{\frac{(1+\log K)\log(8K/\delta)}{K}} \\ & \times\left(1+\frac{\sup_x\|\varphi(x)\|^2}{\mathbb{E}\|\varphi(X)\|^2}\sqrt{\frac{\log 4/\delta}{2K}}\right) \\ & + 8\inf_{f\in\mathcal{F}_N}\|f-f^*\|_P^2\end{aligned} \tag{17}$$

Set $M = \frac{\|\alpha^+\|\sqrt{\mathbb{E}\|\varphi(X)\|^2}}{\sigma}\sqrt{8K\log(8K/\delta)}$. Assume $N > K$ and that the features $(\varphi_k)_{1\leq k\leq K}$ are linearly independent, then if $M \geq 15\log(8K/\delta)$, with probability $1-\delta$, the expected excess risk of the CLSR estimate conditionally on the input samples is upper bounded as:

$$\begin{aligned}\mathbb{E}_Y\left(\|\widehat{g}_B - f^*\|_{P_K}^2\right) \leq & 4\sigma\|\alpha^+\|\sqrt{\mathbb{E}\|\varphi(X)\|^2}\sqrt{\frac{2\log(8K/\delta)}{K}} \\ & \times\left(1+\frac{\sup_x\|\varphi(x)\|^2}{\mathbb{E}\|\varphi(X)\|^2}\sqrt{\frac{\log 4/\delta}{2K}}\right)\end{aligned} \tag{18}$$

Proof: See E ∎

In case of $K \gg \log 1/\delta$, the expected excess risk reduces to

$$O\left(\|\alpha^+\|\sqrt{\mathbb{E}\|\varphi(X)\|^2}\frac{\log K/\delta}{\sqrt{K}} + \inf_{f\in\mathcal{F}_N}\|f-f^*\|_P^2\right) \tag{19}$$

The factor $\|\alpha^+\|\sqrt{\mathbb{E}\|\varphi(X)\|^2}$ determines the generalization error of CLSR. If it is a constant that does not depend on N, the estimation error bound of CLSR reduces to $O(\log K/\sqrt{K})$. While in the OLS in the data domain, this estimation error is $O(N\log K/K)$. It is clear in the case when $N > \sqrt{K}$, CLSR is better than the OLS in terms of estimation error.

In summary, CLSR, which operates in a random subspace of lower dimension, provides an alternative to usual penalization techniques. It has smaller estimation error bound when the term $\|\alpha^+\|\sqrt{\mathbb{E}\|\varphi(X)\|^2}$ has a mild dependency on N. By theorem 1, it also has a controlled changes in approximation error compared to OLS.

III. SVM in Compressed Domain [6]

A. Problem Statement

The problem setting is very similar to the previous regression problem. Given dataset $\mathcal{D}_K = \{x_k, y_k\}_{k=1}^K, x_k \in \mathcal{X}, y_k \in \{-1, 1\}$ We assume that data vector x is s-sparse and has finite ℓ_2 norm. Hence, the original data domain is:

$$\mathcal{X} = \left\{x \in \mathbb{R}^N : \|x\|_0 \leq s, \|x\|_2 \leq R\right\} \tag{20}$$

The data distribution is again denoted by P, and (x_k, y_k) are i.i.d. samples from P. The empirical distribution is denoted by P_K. The goal of linear SVM is reduce the misclassification rate $\mathbb{E}_P[Y \neq f(X)]$, where f is the linear classifier function. A surrogate objective, hinge loss, is introduced in the soft margin SVM for dataset that are not linearly separable and the hinge loss is defined as:

$$H(t) = \max(0, 1+t) \geq \mathbf{1}\{t \geq 0\} \tag{21}$$

For a linear SVM classifier $f_w \overset{\text{def}}{=} \text{sgn}(w^T x)$, the hinge loss is defined as:

$$H_P(w) \overset{\text{def}}{=} \mathbb{E}_P[1 - Yw^T X] \tag{22}$$

and the empirical hinge loss is defined as:

$$H_{P_K}(w) \overset{\text{def}}{=} \frac{1}{K}\sum_{k=1}^K 1 - y_k w^T x_k \tag{23}$$

The true regularization loss of a classifier is:

$$L_P(w) \overset{\text{def}}{=} H_P(w) + \frac{1}{2C}\|w\|^2 \tag{24}$$

and empirical regularization loss:

$$L_{P_K}(w) \overset{\text{def}}{=} H_{P_K}(w) + \frac{1}{2C}\|w\|^2 \tag{25}$$

The following lemma states the property of optimal W that minimizes empirical regularization loss. It is direct result of convex duality and the proof is omitted here.

Lemma 2. *Given dataset $\mathcal{D}_K$, let f_w be the linear SVM classifier obtained by minimizing $L_{P_K}(w)$, then*

$$w = \sum_{k=1}^K \lambda_k y_k x_k \tag{26}$$

where

$$\forall k : 0 \leq \lambda_K \leq \frac{C}{K} \text{ and } \|w\|^2 \leq C \tag{27}$$

Similar to the previous regression problem, the compressed sensing is achieved by random matrix $A \in \mathbb{R}^{M\times N}$. We will denote data domain dataset $\mathcal{D}_K$ and compressed domain dataset $A\mathcal{D}_K \overset{\text{def}}{=} \{Ax_k, y_k\}_{k=1}^K, Ax_k \in \mathbb{R}^M$. The corresponding search space of linear classifier's weights is denoted as $\mathcal{W}_N \overset{\text{def}}{=} \{w \in \mathbb{R}^N\}$ and $\mathcal{Z}_M \overset{\text{def}}{=} \{z \in \mathbb{R}^M\}$. The following three classifier weights are used to bound the regularization error.

$$\begin{aligned} w^* &\overset{\text{def}}{=} \arg\min_{w\in\mathcal{W}_N} L_P(w) \\ z^* &\overset{\text{def}}{=} \arg\min_{z\in\mathcal{Z}_M} L_P(z) \\ \widehat{w} &\overset{\text{def}}{=} \arg\min_{w\in\mathcal{W}_N} L_{P_K}(w) \\ \widehat{z} &\overset{\text{def}}{=} \arg\min_{z\in\mathcal{Z}_M} L_{P_K}(z) \end{aligned} \tag{28}$$

Note here we are overloading L_P, L_{P_K}, H_P and H_{P_K}, when the input is of compressed space, they are defined as:

$$\begin{aligned} H_P(z) &\overset{\text{def}}{=} \mathbb{E}_P[1 - Yz^T(AX)] \\ H_{P_K}(z) &\overset{\text{def}}{=} \frac{1}{K}\sum_{k=1}^K 1 - y_k z^T A x_k \\ L_P(z) &\overset{\text{def}}{=} H_P(z) + \frac{1}{2C}\|z\|^2 \\ L_{P_K}(z) &\overset{\text{def}}{=} H_{P_K}(z) + \frac{1}{2C}\|z\|^2 \end{aligned} \tag{29}$$

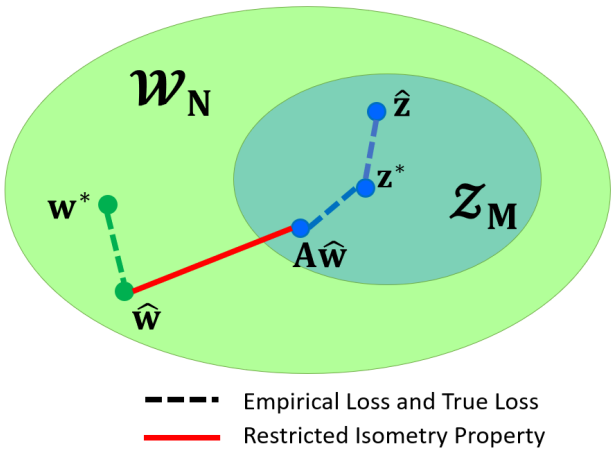

Fig. 2: Relative relationship of $z^*, w^*, \widehat{z}, \widehat{w}$

B. Compressed Learning - Main Result

In this section, learning in compressed space with soft margin SVM classifier will be discussed. If the dataset is compressed via random matrix A and hence is directly presented in the compressed domain, then with high probability, the SVM classifier trained over the training set has generalization error close to the generalization error of the best classifier in the data domain.
The proof is based only on the classifier $\text{sgn}(A\widehat{w})$. Generally, it is not available from compressed training dataset.

Theorem 2. *Let $A \in \mathbb{R}^{M\times N}$ be the compressed sensing matrix, which acts as a near-isometry on any $2s$-sparse vector. That is for any $2s$-sparse vector $x \in \mathbb{R}^N$:*

$$(1-\epsilon)\|x\|^2 \le \|Ax\|^2 \le (1+\epsilon)\|x\|^2 \tag{30}$$

Let $\widehat{z}$ be the soft margin SVM classifier weight trained on $A\mathcal{D}_K$ and w^* be the best linear classifier in the data domain, with low Hinge loss, and large margin (hence small $\|w^*\|^2$). Then with probability over $1-2\delta$,

$$H_P(\widehat{z}) \le H_P(w^*) + O\left(\sqrt{\|w^*\|^2\left(R^2\epsilon + \frac{\log(1/\delta)}{K}\right)}\right) \tag{31}$$

In case the data dimension N is very large, the dimension is efficiently reduced to $O(s\log N)$ while imposing $O(\epsilon)$ error on the performance of the classifier.
When the number of measurements M is sufficiently small, the distortion factor ϵ would be large, yielding the compressed SVM a weak learner. The performance could be improved using boosting techniques like Ada-Boost. In this way, the number of required measurement is reduced and the burden is transferred to computational cost.

C. Compressed Learning - Proof

Definition 1 (Restricted Isometry Property)**.** *$A \in \mathbb{R}^{M\times N}$ satisfies restricted isometry property, (s,ϵ)-RIP, if $\forall x \in \mathbb{R}^N, \|x\|_0 \le s$ (s-sparse vector), the following near-isometry property holds:*

$$(1-\epsilon)\|x\|^2 \le \|Ax\|^2 \le (1+\epsilon)\|x\|^2 \tag{32}$$

From [4], a random Gaussian matrix A, whose entry follows $\mathcal{N}(0, \frac{1}{M})$, satisfies the (s,ϵ)-RIP with probability $e^{-c(\epsilon)M}$ when $\Omega(s\log(M/s))$. This RIP property is a weaker condition compared to aforementioned the Johnson-Lindenstrauss lemma. And it is not constrained by the number of sample points K. But similar to the lemma 1, we also need bounds on the inner product between linear combinations of arbitrary sparse signals, as SVM is based on inner product of the classifier and the samples.

Lemma 3. *Let $A \in \mathbb{R}^{M\times N}$ satisfies $(2s,\epsilon)$-RIP, and $x, x' \in \mathcal{X}$ be two s-sparse vector and their ℓ_2 norm is not larger than R, then*

$$(1+\epsilon)x^Tx' - 2R^2\epsilon \le (Ax)^\top(Ax') \le (1+\epsilon)x^Tx' + 2R^2\epsilon \tag{33}$$

Proof: See F ■

The following lemma generalizes the preservation of inner product of sparse vectors to the inner product between any two vectors from the convex hull of the set of sparse vectors. By lemma 2, the SVM classifier $\widehat{w}$ is a member of this convex hull.

Lemma 4. *Let $A \in \mathbb{R}^{M\times N}$ satisfies $(2s,\epsilon)$-RIP and let two dataset be $\mathcal{D}_K = \{x_k, y_k\}_{k=1}^K, \mathcal{D}'_{K'} = \{x'_k, y'_k\}_{k=1}^{K'}$ where $x_k, x'_k \in \mathcal{X}, y_k, y'_k \in \{-1,1\}$. Let $\lambda_1,\dots,\lambda_K,\lambda'_1,\dots,\lambda'_{K'}$ be non-negative numbers such that $\sum_{k=1}^K \lambda_k \le C, \sum_{k=1}^{K'} \lambda'_k \le C'$ for some $C, C' \ge 0$. Let*

$$w = \sum_{k=1}^K \lambda_k y_k x_k, \quad w' = \sum_{k=1}^{K'} \lambda'_k y'_k x'_k \tag{34}$$

Then

$$|w^Tw' - (Aw)^T(Aw')| \le 3CC'R^2\epsilon \tag{35}$$

Proof: See G ■

This lemma states that implies that if the SVM classifier weight w is projected to the compressed domain, the regularization loss of Aw is almost the same as the regularization loss in high dimensional data domain.
The following lemma connects the regularization loss of the SVM classifiers in data and compressed domain.

Lemma 5. *Let $A \in \mathbb{R}^{M\times N}$ satisfies $(2s,\epsilon)$-RIP. Let $\widehat{w}$ be the soft-margin SVM trained on $\mathcal{D}_K$ with K samples, and let $A\widehat{w}$ be the classifier weight in compressed domain, then*

$$L_P(A\widehat{w}) \le L_P(\widehat{w}) + O(CR^2\epsilon) \tag{36}$$

Proof: See H ■

The following lemma states that the empirical regularization loss is close to the true loss when K is large. The proof is omitted here.

Lemma 6. *Let $\widehat{w}$ be the SVM classifier trained on $\mathcal{D}_K$, then with probability $1-\delta$,*

$$L_P(\widehat{w}) \le L_P(w^*) + O\left(\frac{C\log(1/\delta)}{K}\right) \tag{37}$$

Finally, we can prove the main result of compressed learning (theorem 2). By the definition of regularization loss,

$$\begin{aligned}
H_P(\widehat{z}) &\leq H_P(\widehat{z}) + \frac{1}{2C}\|\widehat{z}\|^2 = L_P(\widehat{z}) \\
&\quad \text{(from definition of regularization loss)} \\
&\leq L_P(z^*) + O\left(\frac{C\log(1/\delta)}{K}\right) \\
&\quad \text{(from lemma 6)} \\
&\leq L_P(A\widehat{w}) + O\left(\frac{C\log(1/\delta)}{K}\right) \\
&\quad \text{(from definition of } z^*\text{)} \\
&\leq L_P(\widehat{w}) + O(CR^2\epsilon) + O\left(\frac{C\log(1/\delta)}{K}\right) \\
&\quad \text{(from lemma 5)} \\
&\leq L_P(w^*) + +O(CR^2\epsilon) + O\left(\frac{C\log(1/\delta)}{K}\right) \\
&\quad \text{(from lemma 6)} \\
&= H_P(w^*) + \frac{1}{2C}\|w^*\|^2 + O\left(CR^2\epsilon + \frac{C\log(1/\delta)}{K}\right) \\
&\quad \text{(from definition of } L_P\text{)}
\end{aligned} \tag{38}$$

Since the above inequality holds for all C, we can pick the best C and the bound on the hinge loss becomes:

$$H_P(\widehat{z}) \leq H_P(w^*) + O\left(\sqrt{\|w^*\|^2\left(R^2\epsilon + \frac{\log(1/\delta)}{K}\right)}\right) \tag{39}$$

From this, we can see that when the number of samples K goes to infinity, there still is a ineligible difference of hinge loss between data domain and compressed domain.

IV. Conclusion

In this report, we reviewed two examples of compressed learning: compressed least square regression and compressed SVM classification. These high-level task is performed in the compressed domain, induced by a random projection to a lower dimension. The learning bound depends largely on the properties of the measurement matrix, such as preservation of relative geometry. The excess risk in compressed regression can be bounded through, and it approaches the best regressor in data domain with infinite number of samples. The hinge loss in compressed SVM can also be bounded, but there is a gap between the best classifier in data domain and compressed domain, even if the number of samples goes to infinity.

References

[1] I. Jolliffe, *Principal component analysis*. Springer, 2011. 1

[2] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *science*, vol. 290, no. 5500, pp. 2319–2323, 2000. 1

[3] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008. 1

[4] E. J. Candes, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, vol. 59, no. 8, pp. 1207–1223, 2006. 1, 4

[5] O. Maillard and R. Munos, "Compressed least-squares regression," in *Advances in Neural Information Processing Systems*, 2009, pp. 1213–1221. 1

[6] R. Calderbank, S. Jafarpour, and R. Schapire, "Compressed learning: Universal sparse dimensionality reduction and learning in the measurement domain," *preprint*, 2009. 1, 3

[7] S. Dasgupta and A. Gupta, "An elementary proof of the johnson-lindenstrauss lemma," *International Computer Science Institute, Technical Report*, vol. 22, no. 1, pp. 1–5, 1999. 1

[8] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk, *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006. 2

Appendix A
Excess risk

$$\begin{aligned}
L(\widehat{f}) - L(f^*) &= \mathbb{E}\left[(Y - \widehat{f}(X))^2\right] - \mathbb{E}\left[(Y - f^*(X))^2\right] \\
&= \mathbb{E}\left[(\widehat{f}(X))^2 - (f^*(X))^2\right] \\
&\quad - \mathbb{E}\left[2(f^*(X) + \eta(X))(\widehat{f}(X) - f^*(X))\right] \\
&= \mathbb{E}\left[(\widehat{f}(X) - f^*(X))^2\right]
\end{aligned} \tag{40}$$

the last equality is because the noise $\eta(X)$ is independent of the target function f^* as well as the estimate $\widehat{f}$.

Appendix B
Search spaces relationship

For every function $g_\beta \in \mathcal{G}_M, \forall x \in \mathcal{X}$, we have

$$g_\beta(x) = \beta^T\psi(x) = (A^T\beta)^T\varphi(x) \Longrightarrow g_\beta \in \mathcal{F}_N \tag{41}$$

Appendix C
Variation of Johnson-Lindenstrauss Lemma

Proof:

Given a random Gaussian matrix $A \in \mathbb{R}^{M\times N}, A_{ij} \sim \mathcal{N}(0, \frac{1}{M})$, then for any u in $\mathbb{R}^N$ and any $\varepsilon \in (0,1)$, we have

$$\begin{aligned}
\mathbb{P}\left(\|Au\|^2 \geq (1+\varepsilon)\|u\|^2\right) &\leq e^{-M\left(\varepsilon^2/4 - \varepsilon^3/6\right)} \\
\mathbb{P}\left(\|Au\|^2 \leq (1-\varepsilon)\|u\|^2\right) &\leq e^{-M\left(\varepsilon^2/4 - \varepsilon^3/6\right)}
\end{aligned} \tag{42}$$

We apply the above lemma to any vector $u - w$ and $u + w$, where $u = \frac{u_k}{\|u_k\|}, w = \frac{v_k}{\|v_k\|}$. From the parallelogram law, we have the following event

$$\begin{aligned}
4(Au)^T Aw &= \|Au + Aw\|^2 - \|Au - Aw\|^2 \\
&\leq (1+\varepsilon)\|u+w\|^2 - (1-\varepsilon)\|u-w\|^2 \\
&= 4u^Tw + \varepsilon\left(\|u+w\|^2 + \|u-w\|^2\right) \\
&= 4u^Tw + 2\varepsilon\left(\|u\|^2 + \|w\|^2\right) \\
&= 4u^Tw + 4\varepsilon
\end{aligned} \tag{43}$$

happens with probability larger than $1 - 2e^{-M\left(\varepsilon^2/4 - \varepsilon^3/6\right)}$. Thus for each $k \in [K]$, with probability $1 - 4e^{-M\left(\varepsilon^2/4 - \varepsilon^3/6\right)}$, we have

$$|(Au_k)^T Av - u_k^T v| \leq \varepsilon \|u_k\| \|v\| \tag{44}$$

using union bound considering all u_k, the above inequality holds for all $k \in [K]$ with probability $1 - 4Ke^{-M\left(\varepsilon^2/4 - \varepsilon^3/6\right)}$. Taking $M \geq \frac{1}{\frac{\varepsilon^2}{4} - \frac{\varepsilon^3}{6}} \log\frac{4K}{\delta}$ suffices. ■

APPENDIX D
PROOF ON BOUND OF APPROXIMATION ERROR

Proof: Let $f^+ \overset{\text{def}}{=} f_{\alpha^+} = \arg\min_{f\in\mathcal{F}_N} \|f - f^*\|_P$ denote the best linear regressor in $\mathcal{F}_N$, and $g^+ \overset{\text{def}}{=} g_{A\alpha^+}$. The approximation error of $\mathcal{G}_M$ is bounded as:

$$\begin{aligned}\inf_{g\in\mathcal{G}_M} \|g - f^*\|_P^2 &\le \left\|g^+ - f^*\right\|_P^2 \\ &= L(g^+) - L(f^*) \\ &= L(g^+) - \inf_{f\in\mathcal{F}_N} L(f) + \inf_{f\in\mathcal{F}_N} L(f) - L(f^*) \\ &= \left\|g^+ - f^+\right\|_P^2 + \inf_{f\in\mathcal{F}_N} \|f - f^*\|_P^2\end{aligned} \tag{45}$$

Since

$$\begin{aligned}\alpha^+ &= \arg\min_{\alpha\in\mathbb{R}^N} L\left(f_\alpha\right) - L\left(f^*\right) \\ &= \arg\min_{\alpha\in\mathbb{R}^N} \|f_\alpha - f^*\|_P^2 \\ &= \mathbb{E}_P\left[(\alpha^T\varphi(x) - f^*(x))^2\right]\end{aligned} \tag{46}$$

The optimal f^+ within $\mathcal{F}_N$ is a orthoprojection of f^* onto $\mathcal{F}_N$. Also, $g^+ \in \mathcal{G}_M \in \mathcal{F}_N$, we can bound $\|g^+ - f^+\|_P^2$ using concentration inequality.

$$\begin{aligned}\left\|g^+ - f^+\right\|_P^2 &= \|g_{A\alpha^+} - f_{\alpha^+}\|_P^2 \\ &= \mathbb{E}_P\left[(g_{A\alpha^+}(x) - f_{\alpha^+}(x))^2\right] \\ &= \mathbb{E}_P\left[((A\alpha^+)^T(A\varphi)(x) - f_{\alpha^+}(x))^2\right]\end{aligned} \tag{47}$$

Let $Z(x) \overset{\text{def}}{=} (A\alpha^+)^T(A\varphi)(x) - f_{\alpha^+}(x)$. Since $M \ge 15\log(8K/\delta)$, we have $\varepsilon^2 \overset{\text{def}}{=} \frac{8}{M}\log(8K/\delta) < \frac{3}{4}$, therefore $M \ge \frac{\log(8K/\delta)}{\varepsilon^2/4 - \varepsilon^3/6}$. From C we know with probability $> 1 - \frac{\delta}{2}$, the following holds for all $k \in [K]$.

$$|Z\left(x_k\right)| \le \varepsilon\left\|\alpha^+\right\|\|\varphi\left(x_k\right)\| \le \varepsilon\left\|\alpha^+\right\| \sup_{x\in\mathcal{X}}\|\varphi(x)\| \overset{\text{def}}{=} C \tag{48}$$

Let $g(x^K) = \sum_{k=1}^K |Z(x_k)|^2$, we have

$$\begin{aligned}C^2 &= \sup_x g(x_k = x) - \inf_{x'} g(x_k = x') \\ \mathbb{E}_P\left[g(x^K)\right] &= \mathbb{E}_P\left[|Z(x^K)|^2\right]\end{aligned} \tag{49}$$

from McDiarmid inequality, we have

$$\begin{aligned}&\mathbb{P}\left\{\mathbb{E}_{X\sim P_\mathcal{X}}|Z(X)|^2 - \frac{1}{K}\sum_{k=1}^K |Z\left(x_k\right)|^2 \ge C^2\sqrt{\frac{\log\left(2/\delta'\right)}{2K}}\right\} \\ &= \mathbb{P}\left\{\mathbb{E}_P\left[g(x^K)\right] - g(x^K) \ge C^2\sqrt{\log(2/\delta')K/2}\right\} \\ &\le \exp(-\frac{KC^4\log(2/\delta')}{KC^4}) = \frac{\delta'}{2}\end{aligned} \tag{50}$$

Now suppose $\frac{1}{K}\sum_{k=1}^K |Z\left(x_k\right)|^2$ is concentrated around its mean with probability $1 - \frac{\delta'}{2}$ and $Z(x_K)$ is bounded by C for all k with probability $1 - \frac{\delta}{2}$, we have

$$\begin{aligned}&\left\|g^+ - f^+\right\|_P^2 = \mathbb{E}_P\left[|Z(x)|^2\right] \\ &\le \frac{1}{K}\sum_{k=1}^K |Z\left(x_k\right)|^2 + C^2\sqrt{\frac{\log\left(2/\delta'\right)}{2K}} \\ &\le \varepsilon^2\left\|\alpha^+\right\|^2\left(\frac{1}{K}\sum_{k=1}^K \|\varphi\left(x_k\right)\|^2 + \sup_{x\in\mathcal{X}}\|\varphi(x)\|^2\sqrt{\frac{\log\left(2/\delta'\right)}{2K}}\right) \\ &\le \varepsilon^2\left\|\alpha^+\right\|^2\left(\mathbb{E}\left[\|\varphi(X)\|^2\right] + 2\sup_{x\in\mathcal{X}}\|\varphi(x)\|^2\sqrt{\frac{\log\left(2/\delta'\right)}{2K}}\right)\end{aligned} \tag{51}$$

Setting $\delta' = \delta/2$, this inequality holds with probability at least $(1 - \delta/2)(1 - \delta') \ge 1 - \delta$.

■

APPENDIX E
PROOF ON EXCESS RISK BOUND OF CLSR

Proof: From (14), replace $\widehat{f}_L$ with $\widehat{g}_L$, we have

$$\begin{aligned}\mathbb{E}\left(\|\widehat{g}_B - f^*\|_P^2\right) \le& c'\max\left\{\sigma^2, B^2\right\}\frac{1 + \log K}{K}N \\ &+ 8\inf_{f\in\mathcal{G}_M}\|g - f^*\|_P^2\end{aligned} \tag{52}$$

Since we assume that $M \ge 15\log(8K/\delta)$, from theorem 1, with probability $1 - \delta$, we have

$$\begin{aligned}&\mathbb{E}\left(\|\widehat{g}_B - f^*\|_P^2\right) \le \\ &c'\max\left\{\sigma^2, B^2\right\}\frac{1 + \log K}{K}M + 8\inf_{f\in\mathcal{F}_N}\|f - f^*\|_P^2 \\ &+ \frac{64\log(8K/\delta)}{M}\left\|\alpha^+\right\|^2\left(\mathbb{E}\|\varphi(X)\|^2 + 2\sup_x\|\varphi(x)\|^2\sqrt{\frac{\log 4/\delta}{2K}}\right)\end{aligned} \tag{53}$$

Minimizing the RHS of (53) with respect to M would yield $M^* = 8\frac{\|\alpha^+\|\sqrt{\mathbb{E}\|\varphi(X)\|^2}}{\max(\sigma, B)}\sqrt{\frac{K\log(8K/\delta)}{c'(1+\log K)}}$. Thus, we have the first part of corollary 1. Similarly, using (15), we have

$$\begin{aligned}&\mathbb{E}_Y\left(\|\widehat{g}_B - f^*\|_{P_K}^2\right) \le \sigma^2\frac{M}{K} + \inf_{g\in\mathcal{G}_M}\|g - f^*\|_{P_K}^2 \\ &\le \sigma^2\frac{M}{K} + \inf_{f\in\mathcal{F}_N}\|f - f^*\|_{P_K}^2 \\ &+ \frac{8}{M}\log(8K/\delta)\left\|\alpha^+\right\|^2\left(\mathbb{E}\|\varphi(X)\|^2 + 2\sup_x\|\varphi(x)\|^2\sqrt{\frac{\log 4/\delta}{2K}}\right)\end{aligned} \tag{54}$$

Notice that in case of $N > K$ and $\varphi(x_k)$ linear independent, the $\inf_{f\in\mathcal{F}_N}\|f - f^*\|_{P_K}^2$ reduces to zero. By setting $M = \frac{\|\alpha^+\|\sqrt{\mathbb{E}\|\varphi(X)\|^2}}{\sigma}\sqrt{8K\log(8K/\delta)}$, we obtain the second half of the corollary. ■

APPENDIX F
PROOF ON THE PRESERVATION OF INNER PRODUCT IN COMPRESSION

Proof: It is obvious that $x - x'$ is at most $2s$-sparse, therefore,

$$\begin{aligned}&(1-\epsilon)\left(\|x\|^2+\|x'\|^2\right)-2(Ax)^\top(Ax')\\&\le\|Ax\|^2+\|Ax'\|^2-2(Ax)^\top(Ax')\\&=\|A(x-x')\|^2\end{aligned}\tag{55}$$

and

$$\begin{aligned}\|A(x-x')\|^2&\le(1+\epsilon)\|x-x'\|^2\\&=(1+\epsilon)\left(\|x\|^2+\|x'\|^2-2x^\top x'\right)\end{aligned}\tag{56}$$

Putting these two equations together and use $\|x\|_2 \le R, \|x'\|_2 \le R$ completes the proof of $(1+\epsilon)x^T x' - 2R^2\epsilon \le (Ax)^\top (Ax')$. And the other side of inequality can be proved similarly and omitted here. ■

APPENDIX G
PROOF ON PRESERVATION OF INNER PRODUCT OF CLASSIFIER WEIGHT

Proof: From the definition of w, w', we have

$$\begin{aligned}(Aw)^\top(Aw')&=\sum_{i=1}^{K}\sum_{j=1}^{K}\lambda_i\lambda'_j y_i y'_j\left(Ax_i\right)^\top\left(Ax'_j\right)\\&=\sum_{y_i=y'_j}\lambda_i\lambda'_j\left(Ax_i\right)^\top\left(Ax'_j\right)-\sum_{y_i\neq y'_j}\lambda_i\lambda'_j\left(Ax_i\right)^\top\left(Ax'_j\right)\end{aligned}\tag{57}$$

Now since $\lambda_i, \lambda'_j \ge 0$, using lemma 3, we have

$$\begin{aligned}\lambda_i\lambda'_j\left(Ax_i\right)^\top\left(Ax'_j\right)&\le\lambda_i\lambda'_j\left((1-\epsilon)x^\top x'+2R^2\epsilon\right)\\\lambda_i\lambda'_j\left(Ax_i\right)^\top\left(Ax'_j\right)&\ge\lambda_i\lambda'_j\left((1+\epsilon)x_i^\top x'_j-2R^2\epsilon\right)\end{aligned}\tag{58}$$

Therefore,

$$\begin{aligned}&\sum_{y_i=y'_j}\lambda_i\lambda'_j\left(Ax_i\right)^\top\left(Ax'_j\right)-\sum_{y_i\neq y'_j}\lambda_i\lambda'_j\left(Ax_i\right)^\top\left(Ax'_j\right)\\&\le\sum_{y_i=y'_j}\lambda_i\lambda'_j\left((1-\epsilon)\boldsymbol{x}_i^\top\boldsymbol{x}'_j+2R^2\epsilon\right)\\&\quad-\sum_{y_i\neq y'_j}\lambda_i\lambda'_j\left((1+\epsilon)x_i^\top x'_j-2R^2\epsilon\right)\\&=\sum_{i,j}\lambda_i\lambda'_j y_i y'_j\left(\boldsymbol{x}_i^\top\boldsymbol{x}'_j\right)+\sum_{i,j}\lambda_i\lambda'_j\epsilon\left(2R^2+\boldsymbol{x}_i^\top\boldsymbol{x}'_j\right)\\&\le\quad w^\top w'+3R^2\epsilon\sum_{i=1}^{M}\lambda_i\sum_{j=1}^{N}\lambda'_j\\&\le w^\top w'+3R^2CC'\epsilon\end{aligned}\tag{59}$$

The other side of absolute value:

$$w^\top w'-3R^2CC'\epsilon\le(Aw)^\top(Aw')\tag{60}$$

can also be proved similarly. ■

APPENDIX H
PROOF ON REGULARIZATION LOSS OF $\widehat{w}, A\widehat{w}$

Proof: From lemma 2, we can write $\widehat{w}=\sum_{k=1}^{K}\lambda_k y_k x_k$, with $\lambda_K \ge 0, \sum_{k=1}^{K} \le C$. Apply lemma 4 with $K = K', x'_k = x_k, y'_k = y_k$ and $D = C$, we have

$$(A\widehat{w})^T(A\widehat{w})\le\widehat{w}^T\widehat{w}+3C^2R^2\epsilon\tag{61}$$

this gives $\frac{1}{2C}\|A\widehat{w}\|^2 \le \frac{1}{2C}\|\widehat{w}\|^2 + O(CR^2\epsilon)$. Now if we fix $x \in \mathcal{X}, y$, apply lemma 4 again with $K' = 1, D = 1$ and $(x'_1, y'_1) = (x, y)$, we have

$$1-y(A\widehat{w}^T)Ax\le1-y\widehat{w}^Tx+O(CR^2\epsilon)\tag{62}$$

Since $1 - y\widehat{w}^T x \le H(-y\widehat{w}^T x)$, and the RHS of the above equation is always positive, we have

$$H(y(A\widehat{w})^T(Ax)\le H(-y\widehat{w}^Tx)+O(CR^2\epsilon)\tag{63}$$

it follows that $H_P(A\widehat{w}) \le H_P(\widehat{w}) + O(CR^2\epsilon)$. Combine the difference in norms, the proof is complete. ■