Expanding Compressed Sensing and learning

Kevin Shi   Kui Tang

Columbia University

11 Dec. 2015

Outline

Compressed Learning
- Review of Calderblank et. al. 09
- Extension to Regression
- Other Attempted Generalizations

Explicit RIP Constructions
- Bipartite graph model of measurement
- Poisson Random Matrices

Learning Compressively-Sensed Data

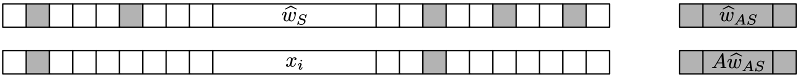


Figure: Sample space S and measurement (compressed) space AS.

- There exist training data $(x_i, y_i) \subset \mathbb{R}^n \times -1, 1$ where the x_i are k-sparse, learn binary classifier $f : \mathcal{X} \rightarrow -1, 1$.
- We observe compressively-sensed measurements $(Ax_i, y_i) \subset \mathbb{R}^M \times -1, 1$ for a $(2k, \epsilon)$ RIP $m \times n$ matrix A.
- Two options
  - Recover n-dimensional sparse vectors, learn classifier in the high dimensional space.
  - **Learn classifier directly in the compressed space!**

Support Vector Machine Review

- We minimize the hinge loss, which on one example is $H(x, y; w) = \max 0, 1 - yw^\top x$
- The true hinge loss on distribution $\mathcal{D}$ is $H_\mathcal{D}(w) = E_{(x_i,y_i)\sim\mathcal{D}}[1 - y_i w^\top x_i]$
- The true regularization loss is $L(w) = H_\mathcal{D}(w) + \frac{1}{2C}\|w\|$.
- The trained SVM classifier $\widehat{w}_S$ can be written as

  $$\widehat{w}_S = \sum_i \alpha_i y_i x_i$$

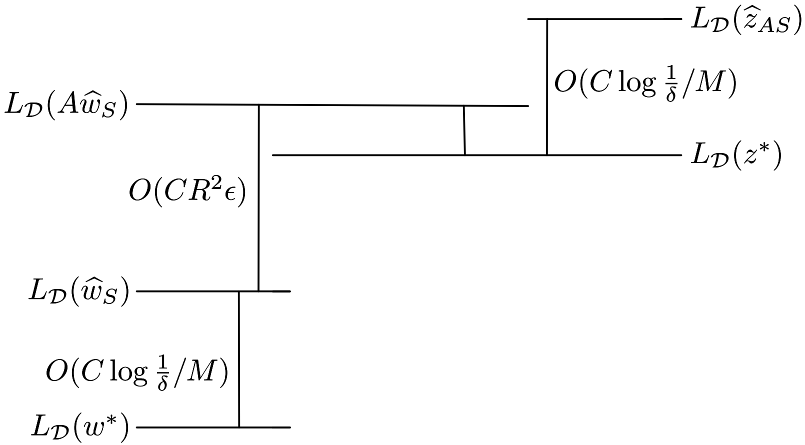
  where $0 \leq \alpha_i \leq C/M$ and $\|\widehat{w}_S\| \leq C$.
- If w^* is the best SVM classifier over $\mathcal{D}$, then with probability $1 - \delta$, we have (Sridharan, Srebro, and Shalev-Shwartz, 2008)

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

Compressed Learning Bound

Main result is

$$L_{\mathcal{D}}(\widehat{z}_{AS}) \leq L_{\mathcal{D}}(w^*) + O(CR^2\epsilon + C\log\frac{1}{\delta}/M)$$



RIP for Dot Products

Theorem (Calderbank, Jafarpour, and Schapire (2009))

Let $A_{m \times n}$ be $(2k, \epsilon)$-RIP, x, x' two k-sparse vectors in $\mathbb{R}^n$ with $\|x\|, \|x\| \leq R$. Then

$$(1+\epsilon)x^\top x' - 2R^2\epsilon \leq (Ax)^\top \left(Ax'\right) \leq (1-\epsilon)x^\top x' + 2R^2\epsilon$$

Applying Dot-RIP to SVM Loss

- Suppose we train a classifier $\widehat{w}_S$ in the high-dimensional space.
- Project to low dimensional space, getting classifier $A\widehat{w}_S$.
- A key result is to show that projection does not increase the loss too much:

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

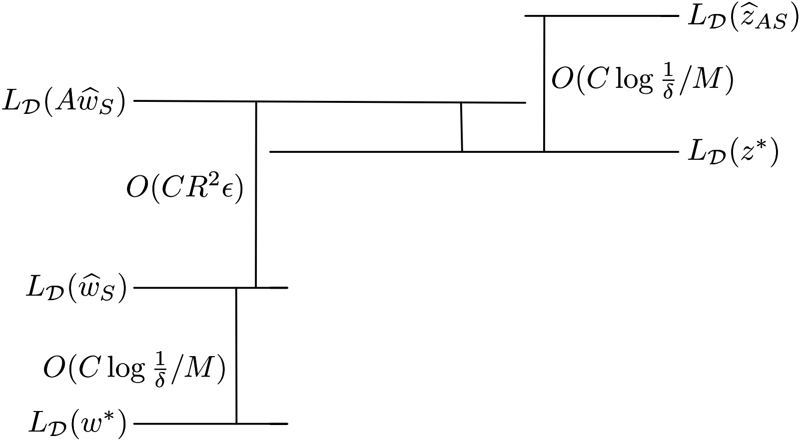
- L contains terms of the form $y_i w_i^\top x_i$ and $\|w\|$.
- Use kernel representation

$$\widehat{w}_S = \sum_i \alpha_i y_i x_i$$

  to write (7) in terms of $(A\widehat{w}_S)^\top(Ax)$ and $\widehat{w}_S^\top x$, and use Theorem to get result.
- Technical issues with signs and cases... tedious but it works out.

Putting it Together

$$L_{\mathcal{D}}(\widehat{z}_{AS}) \leq L_{\mathcal{D}}(w^*) + O(CR^2\epsilon + C \log \frac{1}{\delta}/M)$$



Support Vector Regression

- We have continuous y and use the ρ-insensitive *tube loss*

$$T(x, y; w) = \max\left\{y - w^\top x - \rho, w^\top x - y - \rho, 0\right\}$$

- The dual (kernel) representation of the learned classifier is

$$w = \sum_i (\alpha_i - \alpha_i^*)\, x_i$$

- (Almost) the same projection bound holds! Need another term in ρ:

$$T_{\mathcal{D}}(A\widehat{w}_S) \leq T_{\mathcal{D}}(\widehat{w}_S) + O(CR^2\epsilon + \rho)$$

Compressed Learning for Support Vector Machines

- The loss function has 3 cases

$$T(x, y; w) = \begin{cases} y - w^\top x - \rho & (+) \quad y - w^\top x - \rho > 0 \\ w^\top x - y - \rho & (-) \quad w^\top x - y - \rho > 0 \\ 0 & (0) \quad \left| y - w^\top x \right| \leq \rho \end{cases}$$

- The difference $T_{\mathcal{D}}(A\widehat{w}_S) - T_{\mathcal{D}}(\widehat{w}_S)$ needs to be evaluated for 9 cases (some trivial).
- Supporting lemmas also need to be upgraded to handle negative cases.

Attempts to Generalize to other Kernels

- Recall the RIP for linear kernels:

$$(1+\epsilon)x^\top x' - 2R^2\epsilon \leq (Ax)^\top \left(Ax'\right) \leq (1-\epsilon)x^\top x' + 2R^2\epsilon$$

- The squared exponential kernel has with variance σ^2 and length scale ℓ is

$$k(x,x') = \sigma^2 \exp\left(-\frac{\|x-x'\|_2^2}{2\ell^2}\right) \tag{1}$$

- We have obtained

$$\exp(-2\epsilon R - \epsilon^2 R)k(x,x') \leq k(Ax,Ax') \leq \exp(2\epsilon R)k(x,x') \tag{2}$$

- Both the $(1 \pm \epsilon)$ and $2R^2\epsilon$ terms are exponentiated.
- Tried Matern and rational quadratic kernels... no luck.

Attempts to Generalize to Linear Regression

- Classical analysis: suppose $Y = X\beta + \epsilon$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.
- Y random, X and β fixed.
- The measure of generalization error is *risk*, e.g. expected squared error under the true distribution:

$$\begin{aligned} E_Y\left[\frac{1}{n}\left\|Y - X\widehat{\beta}\right\|^2\right] &= E_Y\left[\left\|\widehat{\beta} - \beta\right\|_\Sigma^2\right] \\ &= E_Y\left[\left\|\widehat{\beta} - \overline{\beta}\right\|_\Sigma^2\right] + E_Y\left[\left\|\overline{\beta} - \widehat{\beta}\right\|_\Sigma^2\right] \end{aligned}$$

  where $\Sigma = \frac{1}{n}X\top X$ and $\widehat{\beta} = E_Y[\widehat{\beta}]$.
- For both linear and ridge (ℓ_2 regularized) regression, $\widehat{\beta} = \widehat{\beta}(X, Y, \lambda)$ available in closed form.
- Note: For positive semi-definite matrix C, $\|x\|_C = x^\top Cx$.

Problem with Linear Regression

- Again, let's compute the risk of the projected model $A\widehat{\beta}_{AS}$.
- The *variance* of the this estimator is

$$E\left[\left\|A\widehat{\beta}_S - A\overline{\beta}_S\right\|^2_{A\Sigma A^\top}\right] = E\left[\left\|\widehat{\beta}_S - \overline{\beta}_S\right\|^2_{A^\top A\Sigma A^\top A}\right]$$

- The term $E\left[\left\|\widehat{\beta}_S - \overline{\beta}_S\right\|^2_{\Sigma}\right]$ is the variance of the estimator in the high-dimensional space $\widehat{\beta}_S$.
- The norm $\|x\|_{A^\top A\Sigma A^\top A}$ contains the factor $A^\top Ax$.
- But RIP doesn't apply here because Ax is no longer sparse, and $A^\top$ does not have interesting properties.
- **Or am I missing something here?**

Outline

Compressed Learning
- Review of Calderblank et. al. 09
- Extension to Regression
- Other Attempted Generalizations

Explicit RIP Constructions
- Bipartite graph model of measurement
- Poisson Random Matrices

RIP constructions with $m = \tilde{O}(k)$

- Draw a random matrix with entries sampled from $\mathcal{N}(0, 1/m)$.
- Draw a random matrix with entries sampled from $\left\{+\frac{1}{\sqrt{m}}, -\frac{1}{\sqrt{m}}\right\}$ with Bernoulli parameter 0.5.

Theorem

With high probability the random matrix Φ sampled from either distribution above satisfies

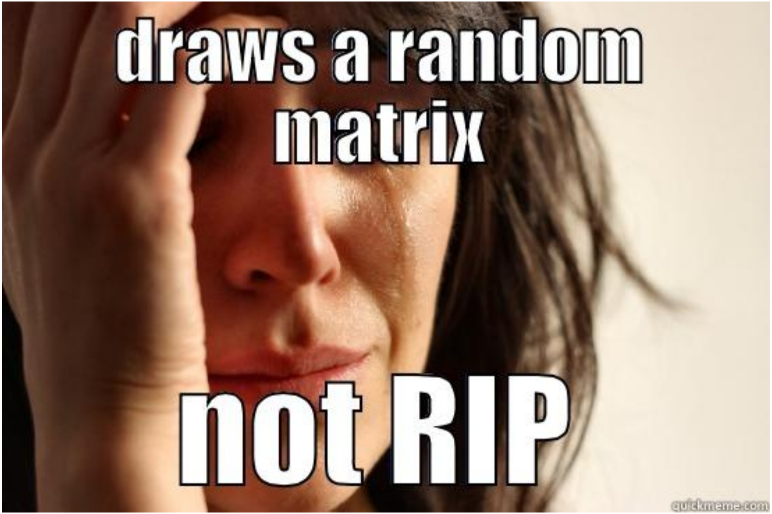
$$(1-\epsilon)\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1+\epsilon)\|x\|_2^2$$

for all k-sparse x.

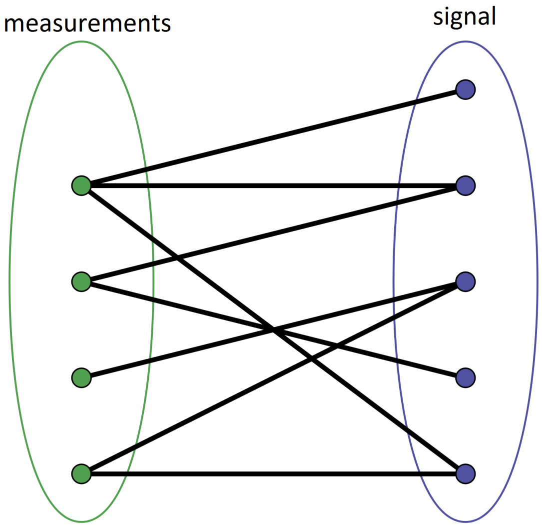
But...

Theorem (Bandeira, Dobriban, Mixon, and Sawin (2012))

Given a matrix Φ and parameters (k, ϵ), certifying whether Φ is (k, ϵ)-RIP is NP-hard.


draws a random matrix
not RIP
quickmeme.com

Bipartite graph model of measurement



Expander graphs

- Expander graphs capture many properties of random graphs, but can be constructed deterministically.
- Expander graphs can also be constructed probabilistically, and the graph expansion property can be certified.

Definition (Vertex expansion)

Let $G = (A, B, E)$ be a bipartite graph with left degree d. G has (k, ϵ)-vertex expansion if for every subset $X \subset A, |X| \leq k$, the set of neighbors $N(A) = \{j \in B \mid \exists\, i \in X, (i, j) \in E\}$ has size at least $|N(A)| \geq (1 - \epsilon)d|X|$.

Explicit RIP for ℓ_1-norm

Theorem (Berinde, Gilbert, Indyk, Karloff, and Strauss (2008))

Let (A, B, E) be a bipartite expander graph with left degree d and with (k, ϵ)-vertex expansion. That is, for all $X \subset A, |X| < k$, then $|N(X)| \leq (1 - \epsilon)d|X|$. Then the scaled adjacency matrix $\frac{1}{d^{1/p}}\Phi$ satisfies the $(p, k, \epsilon)-$RIP property

$$(1 - \epsilon)\|x\|_p^2 \leq \|\Phi x\|_p^2 \leq (1 + \epsilon)\|x\|_p^2$$

for all k-sparse x and for p close to 1.

The paper goes on to show this RIP-1 property gives the same sparse recovery bound for basis pursuit but with the ℓ_1 norm of the error vector instead of the ℓ_2 norm.

Extension to ℓ_2-norm

Theorem (Chandar 08)

A matrix $\Phi \in \{0,1\}^{m \times n}$ which satisfies the $(2, k, \epsilon)$-RIP property must have $m = \Omega(k^2)$.

- The technique of Berinde doesn't work directly.
- Possible way around the lower bound: use multigraphs.
  $\Phi_{ij} = \#$ of edges between i and j.

Possible generalization of lower bound?

- The lower bound uses some techniques very specific to the $\{0, 1\}$ assumption.
- Want to see if more entries helps, such as using $\{0, \cdots, d\}$ in the case of a degree$-d$ multigraph.
- Tried to get a lower bound in terms of the ratio of ℓ_1 to ℓ_2 norms of the columns, which should be smaller when larger entries are used... no luck.
- Found out about the Bernoulli$\left(\left\{-\frac{1}{\sqrt{m}}, +\frac{1}{\sqrt{m}}\right\}\right)$ construction of JL, which has an even worse ℓ_1 to ℓ_2 ratio on the columns.

A more modest question

Let's ignore derandomization for a moment.

- Are cancellations in sign necessary for the ℓ_2 norm?
- Can RIP matrices with $m = \tilde{O}(k)$ can be constructed using only nonnegative entries?

Poisson Random Matrices

- Given $a, b \overset{i.i.d.}{\sim} \text{Pois}(\lambda)$, $\mathbb{E}[ab] = \lambda^2$ and $\mathbb{E}[a^2] = \lambda + \lambda^2$.
- For $\lambda << 1$, this property seems almost as good as $\mathbb{E}[ab] = 0$ for Gaussian random variables.

Lemma

Given a matrix $\Phi \in \mathbb{Z}_{\geq 0}^{m \times n}$ where $\Phi_{ij} \sim Pois(\lambda)$, for k-sparse x

$$\|x\|_2^2 \leq \mathbb{E}\left[\left\|\frac{1}{m\lambda}\Phi x\right\|_2^2\right] \leq (1 + \lambda k)\|x\|_2^2$$

- Doesn't quite work for the full JL statement - need the k-sparsity or else λ has to be inversely proportional to the raw signal dimension n.
- Concentration bounds being worked out...

References I

A. S. Bandeira, E. Dobriban, D. G. Mixon, and W. F. Sawin. Certifying the restricted isometry property is hard. *ArXiv e-prints*, April 2012.

R. Berinde, A.C. Gilbert, P. Indyk, H. Karloff, and M.J. Strauss. Combining geometry and combinatorics: A unified approach to sparse signal recovery. In *Communication, Control, and Computing, 2008 46th Annual Allerton Conference on*, pages 798–805, Sept 2008. 10.1109/ALLERTON.2008.4797639.

Robert Calderbank, Sina Jafarpour, and Robert Schapire. Compressed learning: Universal sparse dimensionality reduction and learning in the measurement domain. Technical report, Princeton University, 2009.

Karthik Sridharan, Nathan Srebro, and Shai Shalev-Shwartz. Fast convergence rates for excess regularized risk with application to svm. Technical report, TTIC, 2008.