Learning Sparse Representations of High Dimensional Data on Large Scale Dictionaries

Zhen James Xiang, Hao Xu, Peter J. Ramadge
Dept. of EE, Princeton University
1. **The Problem**
   The computational challenge of learning large scale sparse representations.

2. **Lasso Screening**
   Solve lasso problems faster and obtain same solution.

3. **Hierarchical Dictionary**
   Efficient algorithm framework to learn tree structured multi-layer dictionaries.
Sparse Representations: Why?

\[ x \approx Bw \]

Image credit: Tenenbaum et al.

Image credit: He and Niyogi.
Sparse Representations: Why?

\[ x \approx Bw \]

Image credit: He and Niyogi.

Image credit: Tenenbaum et al.
Sparse Representations: Why?

\[ x \approx Bw \]

Image credit: Tenenbaum et al.

Image credit: He and Niyogi.
Sparse Representations: Why?

\[ x \approx Bw \]

Image credit: Tenenbaum et al.

Image credit: He and Niyogi.
Sparse Representations: Why?

$$X \approx BW$$

Image credit: Tenenbaum et al.

Image credit: He and Niyogi.
Sparse Representations: How?

\[ X \approx BW \]
Sparse Representations: How?

\[ X \approx BW \]

\[
\min_{B,W} \frac{1}{2} \|X - BW\|_F^2 + \lambda \|W\|_1 \\
\text{s.t.} \quad \|b_i\|_2^2 \leq 1, \quad 1 \leq i \leq m.
\]
Sparse Representations: How?

\[ \mathbf{X} \approx \mathbf{B}\mathbf{W} \]

\[
\min_{\mathbf{B}, \mathbf{W}} \frac{1}{2} \|\mathbf{X} - \mathbf{B}\mathbf{W}\|_F^2 + \lambda \|\mathbf{W}\|_1 \\
\text{s.t.} \quad \|\mathbf{b}_i\|_2^2 \leq 1, \quad 1 \leq i \leq m.
\]

Updating \( \mathbf{W} \): one lasso problem per data point:

\[
\min_{\mathbf{w}} \frac{1}{2} \|\mathbf{x} - \mathbf{B}\mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_1.
\]

Updating \( \mathbf{B} \): a constrained least square problem.
Examples

Face Recognition: (Wright et al., 2009 PAMI; Wagner et al., 2011 PAMI)

- Dimension: ~32K (192x168 for YALE B Extended);
- Dictionary size: 30~40 times the number of subjects;
- No dictionary iteration yet.
Examples

Face Recognition: (Wright et al., 2009 PAMI; Wagner et al., 2011 PAMI)
• Dimension: ~32K (192x168 for YALE B Extended);
• Dictionary size: 30~40 times the number of subjects;
• No dictionary iteration yet.

Image Restoration: (Mairal et al., 2008 TIP; Mairal et al., 2009 ICCV)
• Dimension: 192 (8*8*3), 1200 (20*20*3), ....
• Dictionary size: 256;
• Nearly millions of data points;
• At least hundreds of iterations.
Examples

Face Recognition: (Wright et al., 2009 PAMI; Wagner et al., 2011 PAMI)
- Dimension: ~32K (192x168 for YALE B Extended);
- Dictionary size: 30~40 times the number of subjects;
- No dictionary iteration yet.

Image Restoration: (Mairal et al., 2008 TIP; Mairal et al., 2009 ICCV)
- Dimension: 192 (8*8*3), 1200 (20*20*3), ....
- Dictionary size: 256;
- Nearly millions of data points;
- At least hundreds of iterations.

This Conference:
- Anna Gilbert, Sparsity: algorithms, approximations, and analysis, Invited Talk;
- Morioka and Satoh, Generalized lasso based approximation of sparse coding for visual recognition, Poster T027;
- Szlam et al., Structured sparse coding via lateral inhibition, Poster W045;
- This talk: Poster T036.
1. **The Problem**
   The computational challenge of learning large scale sparse representations.

2. **Lasso Screening**
   Solve lasso problems faster and obtain same solution.

3. **Hierarchical Dictionary**
   Efficient algorithm framework to learn tree structured multi-layer dictionaries.
Solving One Lasso Problem

- Frontal faces in Yale B Extended data set, dimension = 32256, normalized, $\lambda=0.5$;
- Reporting average time and standard error over 10 randomly chosen lasso problems;
- MatLab implementation, ran on an Intel Xeon X5570 2.93GHz processor.
Screening Tests

Lasso Problem: \[
\min_{w_1, w_2, \ldots, w_m} \frac{1}{2} \| x - \sum_{i=1}^{m} w_i b_i \|_2^2 + \lambda \sum_{i=1}^{m} |w_i|.
\]
Assume \( \|x\|_2 = \|b_i\|_2 = 1 \).

\[ b_i \]

Satisfy the test?

Yes

The optimal \( w_i \) will be 0, reject \( b_i \).

No

Keep \( b_i \)
Screening Tests

Lasso Problem: 
\[
\min_{w_1, w_2, \ldots, w_m} \frac{1}{2} \| x - \sum_{i=1}^{m} w_i b_i \|_2^2 + \lambda \sum_{i=1}^{m} |w_i|.
\]

Assume \( \|x\|_2 = \|b_i\|_2 = 1 \).

The optimal \( w_i \) will be 0, reject \( b_i \).
Screening Tests

Lasso Problem:
\[
\min_{w_1, w_2, \ldots, w_m} \quad \frac{1}{2} \| x - \sum_{i=1}^{m} w_i b_i \|_2^2 + \lambda \sum_{i=1}^{m} |w_i|.
\]

Assume \( \|x\|_2 = \|b_i\|_2 = 1 \).

\[
| x^T b_i | < \lambda - 1 + \lambda / \lambda_{\max} ?
\]

Example:
SAFE/ST1 (Ghaoui et al., 2010 arXiv).

Yes

The optimal \( w_i \) will be 0, reject \( b_i \).

No

Keep \( b_i \)

\[
\lambda_{\max} = \max_i | x^T b_i |.
\]

Online test, two passes through data (1\textsuperscript{st} pass finds \( \lambda_{\max} \), 2\textsuperscript{nd} pass executes the test).
- Frontal faces in Yale B Extended data set, dimension = 32256, normalized, $\lambda=0.5$; dictionary size: 2048;
- Reporting average time and standard error over 10 randomly chosen lasso problems;
- MatLab implementation, ran on an Intel Xeon X5570 2.93GHz processor.
• Frontal faces in Yale B Extended data set, dimension = 32256, normalized, $\lambda=0.5$; dictionary size: 2048;
• Reporting average time and standard error over 10 randomly chosen lasso problems;
• MatLab implementation, ran on an Intel Xeon X5570 2.93GHz processor.
• Frontal faces in Yale B Extended data set, dimension = 32256, normalized, $\lambda=0.5$; dictionary size: 2048;
• Reporting average time and standard error over 10 randomly chosen lasso problems;
• MatLab implementation, ran on an Intel Xeon X5570 2.93GHz processor.
• Frontal faces in Yale B Extended data set, dimension = 32256, normalized, $\lambda=0.5$; dictionary size: 2048;
• Reporting average time and standard error over 10 randomly chosen lasso problems;
• MatLab implementation, ran on an Intel Xeon X5570 2.93GHz processor.
Speed Up Lasso

Time for running test: <1ms

- Frontal faces in Yale B Extended data set, dimension = 32256, normalized, $\lambda=0.5$; dictionary size: 2048;
- Reporting average time and standard error over 10 randomly chosen lasso problems;
- MatLab implementation, ran on an Intel Xeon X5570 2.93GHz processor.
Fast and Online

- **SAFE/ST1** (Ghaoui et al.): $|x^T b_i| < \lambda - 1 + \frac{\lambda}{\lambda_{\text{max}}}$.
- **ST2**: $|x^T b_i| < \lambda_{\text{max}}(1 - 2r_3)$.
- **ST3**: $|x^T b_i - (\lambda_{\text{max}} - \lambda)b^T_\ast b_i| < \lambda(1 - r_3)$.

\[
\lambda_{\text{max}} = \max_i |x^T b_i|, \quad b_\ast = \arg \max_{b \in \{\pm b_i\}} x^T b, \quad r_3 = \sqrt{1/\lambda_{\text{max}}^2 - 1(\lambda_{\text{max}}/\lambda - 1)}.
\]
Fast and Online

- SAFE/ST1 (Ghaoui et al.): $|x^Tb_i| < \lambda - 1 + \lambda/\lambda_{\text{max}}$.
- ST2: $|x^Tb_i| < \lambda_{\text{max}}(1 - 2r_3)$.
- ST3: $|x^Tb_i - (\lambda_{\text{max}} - \lambda)b^*_b b_i| < \lambda(1 - r_3)$.

\[
\lambda_{\text{max}} = \max_i |x^Tb_i|, \quad b_* = \arg \max_{b \in \{\pm b_i\}} x^Tb, \quad r_3 = \sqrt{1/\lambda^2_{\text{max}} - 1(\lambda_{\text{max}}/\lambda - 1)}.
\]

All tests are online tests:
- Two passes;
- Memory footprint = 3 codewords.
Fast and Online

- **SAFE/ST1** (Ghaoui et al.): \[|x^T b_i| < \lambda - 1 + \lambda/\lambda_{\text{max}}.\]
- **ST2**: \[|x^T b_i| < \lambda_{\text{max}}(1 - 2r_3).\]
- **ST3**: \[|x^T b_i - (\lambda_{\text{max}} - \lambda)b^*_T b_i| < \lambda(1 - r_3).\]

\[\lambda_{\text{max}} = \max_i |x^T b_i|, \quad b^* = \arg\max_{b \in \{\pm b_i\}} x^T b, \quad r_3 = \sqrt{1/\lambda_{\text{max}}^2 - 1}(\lambda_{\text{max}}/\lambda - 1).\]

All tests are online tests:
- Two passes;
- Memory footprint = 3 codewords.

Here is the fine print:
- ST2,ST3 and Dome Test are only powerful when \(\lambda_{\text{max}}\) is large.
- Luckily, high correlation is common in real world data sets (Wright et al., 2010).
Fast and Online

- **SAFE/ST1** (Ghaoui et al.): $|x^T b_i| < \lambda - 1 + \lambda/\lambda_{\text{max}}$.
- **ST2**: $|x^T b_i| < \lambda_{\text{max}}(1 - 2r_3)$.
- **ST3**: $|x^T b_i - (\lambda_{\text{max}} - \lambda)b^*_b b_i| < \lambda(1 - r_3)$.

$$
\lambda_{\text{max}} = \max_i |x^T b_i|, \quad b^*_b = \arg \max_{b \in \{\pm b_i\}} x^T b, \quad r_3 = \sqrt{1/\lambda_{\text{max}}^2 - 1}(\lambda_{\text{max}}/\lambda - 1).
$$

All tests are online tests:
- Two passes;
- Memory footprint = 3 codewords.

Here is the fine print:
- ST2,ST3 and Dome Test are only powerful when $\lambda_{\text{max}}$ is large.
- Luckily, high correlation is common in real world data sets (Wright et al., 2010).
The Math Behind the Tests

Dual Problem:

\[
\max_{\theta} \quad \frac{1}{2} \| x \|_2^2 - \frac{\lambda^2}{2} \| \theta - \frac{x}{\lambda} \|_2^2
\]

s.t. \quad |\theta^T b_i| \leq 1 \quad \forall i = 1, 2, \ldots, m.
The Math Behind the Tests

Dual Problem:

$$\max_{\theta} \quad \frac{1}{2} \|x\|^2 - \frac{\lambda^2}{2} \|\theta - \frac{x}{\lambda}\|^2$$

s.t. $$|\theta^T b_i| \leq 1 \quad \forall i = 1, 2, \ldots, m.$$ 

Core Rejection Test:

$$|\tilde{\theta}^T b_i| < 1 \implies \tilde{w}_i = 0.$$
The Math Behind the Tests

Dual Problem:

\[
\begin{align*}
\max_{\theta} & \quad \frac{1}{2} \|x\|^2_2 - \frac{\lambda^2}{2} \|\theta - \frac{x}{\lambda}\|^2_2 \\
\text{s.t.} & \quad |\theta^T b_i| \leq 1 \quad \forall i = 1, 2, \ldots, m.
\end{align*}
\]

Core Rejection Test:

\[|\tilde{\theta}^T b_i| < 1 \Rightarrow \tilde{w}_i = 0.\]

Sphere Test:

If \(\tilde{\theta}\) satisfies \(\|\tilde{\theta} - q\|_2 \leq r\), then \(|q^T b_i| < (1 - r) \Rightarrow \tilde{w}_i = 0\).
1. The Problem
   The computational challenge of learning large scale sparse representations.

2. Lasso Screening
   Solve lasso problems faster and obtain same solution.

3. Hierarchical Dictionary
   Efficient algorithm framework to learn tree structured multi-layer dictionaries.
Tree Structured Dictionaries
Tree Structured Dictionaries

COIL data set, object #80. Train dictionaries using 72 images with rotation angle 5° apart.
Tree Structured Dictionaries

Two layers better than one:
- Deep belief nets (Hinton et al., 2006).
- Deep coding network (Lin et al., 2010 NIPS).
- Proximal methods (Jenatton et al., 2010 ICML).

First Layer

COIL data set, object #80. Train dictionaries using 72 images with rotation angle 5° apart.
Random Projections

First layer uses fewer random projections:

\[ T_1 X \approx B_1 W_1 \]

Second layer uses more random projections:

\[ T_2 X \approx B_2 W_2 \]
Random Projections

First layer uses fewer random projections:

\[ T_1 X \approx B_1 W_1 \]

Use incremental, orthogonal random projections to control the information flow.

Second layer uses more random projections:

\[ T_2 X \approx B_2 W_2 \]
Random Projections

First layer uses fewer random projections:

\[ T_1 X \approx B_1 W_1 \]

Use incremental, orthogonal random projections to control the information flow.

Second layer uses more random projections:

\[ T_2 X \approx B_2 W_2 \]

Previous layers inform the coding of later layers.
Evaluation: Time and Quality

- Hand written digit images in MNIST data set, dimension = 784 (28x28), m is the dictionary size;
- Use the sparse coding solvers in *Lee et al., 2006 NIPS* (MatLab) to perform basic optimizations;
- Use liblinear *Fan et al., 2008 JMLR* (C/C++) classifier on the sparse representation weights;
- Ran on an Intel Xeon X5570 2.93GHz processor.
Conclusion

• Learning sparse representation is computationally challenging, but there is hope.

• Lasso screening test significantly speeds up lasso at virtually no additional cost.

• Hierarchical dictionaries and random projections make learning sparse representation more efficient.

Supplemental material and Matlab toolbox available on my website:

qr.net/sparse
Acknowledgements

Hao Xu
Princeton University

Peter J. Ramadge
Princeton University

Kai Yu
NEC Labs America Inc.

Tong Zhang
Rutgers University

Charlotte Elizabeth Procter Honorific Fellowship
Princeton University

Grant CCF-1116208
National Science Foundation

See you at the poster session!  Poster T036
Solving One Lasso Problem

- Frontal faces in Yale B Extended data set, dimension = 32256, normalized, $\lambda=0.6$;
- Reporting average time and standard error over 10 randomly chosen lasso problems;
- MatLab implementation, ran on an Intel Xeon X5570 2.93GHz processor.
Speed Up Lasso

Time for running test: <1ms

- Frontal faces in Yale B Extended data set, dimension = 32256, normalized, \( \lambda = 0.5 \); dictionary size: 2048;
- Reporting average time and standard error over 10 randomly chosen lasso problems;
- MatLab implementation, ran on an Intel Xeon X5570 2.93GHz processor.
Speed Up Lasso

- Frontal faces in Yale B Extended data set, dimension = 32256, normalized, \( \lambda = 0.6 \); dictionary size: 2048;
- Reporting average time and standard error over 10 randomly chosen lasso problems;
- MatLab implementation, ran on an Intel Xeon X5570 2.93GHz processor.
• Frontal faces in Yale B Extended data set, dimension = 32256, normalized; **dictionary size: 2048**;
• Reporting average time and standard error over 10 randomly chosen lasso problems;
• Averaged over three solvers: Grating, Gauss-Seidel and Feature-sign
• MatLab implementation, ran on an Intel Xeon X5570 2.93GHz processor.
Fast and Online

- **SAFE/ST1 (Ghaoui et al.)**: \( |x^T b_i| < \lambda - 1 + \lambda/\lambda_{\text{max}} \).
- **ST2**: \( |x^T b_i| < \lambda_{\text{max}} (1 - 2r_3) \).
- **ST3**: \( |x^T b_i - (\lambda_{\text{max}} - \lambda) b^*_T b_i| < \lambda (1 - r_3) \).

\[
\lambda_{\text{max}} = \max_i |x^T b_i|, \quad b_* = \arg \max_{b \in \{\pm b_i\}} x^T b, \quad r_3 = \sqrt{1/\lambda^2_{\text{max}} - 1} (\lambda_{\text{max}}/\lambda - 1).
\]

ST3 can replace ST2, ST2 can replace ST1 when \( \lambda_{\text{max}} > \sqrt{3}/2 \approx 0.866 \).
More Organized Dictionaries