Learning Sparse Representations of High Dimensional Data on Large Scale Dictionaries

Paper Review By: David Lue

Zhen James Xiang, Hao Xu, and Peter J. Ramadge

Outline

1. Introduction
   - Problem Outline
   - Challenge
2. Reducing Dictionary By Screening
   - Overview
   - Sphere Tests
   - Comparison
3. Random Projections of Data
   - Scale Indifference
   - Preservation of Pairwise Distances
4. Learning Hierarchical Dictionary
   - Weights
   - Dictionary
5. Experiments

Problem Outline

Consider approximating $\mathbf{x} \in \mathbb{R}^p$ by $\mathbf{x} \approx \mathbf{Bw}$ with the constraint that $\mathbf{w}$ is sparse

$$\begin{aligned} \min_{\mathbf{B},\mathbf{W}} \quad & \frac{1}{2} \parallel \mathbf{X} - \mathbf{BW} \parallel_F^2 + \lambda \parallel \mathbf{W} \parallel_1 \\ \text{s.t.} \quad & \parallel \mathbf{b}_i \parallel_2^2 \leq 1, \quad \forall i = 1, 2, \ldots, m \end{aligned} \tag{1}$$

$\mathbf{X} \in \mathbb{R}^{p \times n}$ $\mathbf{B} \in \mathbb{R}^{p \times m}$ $\mathbf{W} \in \mathbb{R}^{m \times n}$

Dictionary $\mathbf{B}$ is adapted to the data

Challenge

- Solving the non-convex optimization problem (1) is computationally expensive - state of the art algorithms solve by iteratively optimizing **W** and **B**
- Fixed **B**
  - Optimize **W** $\Rightarrow$ solve n, p-dimensional, LASSO problems of size m
  - Each lasso costs: $O(mp\kappa + m\kappa^2)$
- Fixed **W**
  - Optimze **B** $\Rightarrow$ least squares problem of pm variables and m constraints
  - Requires inverting $m \times m$ matrices: $O(m^3)$ complexity

Overview

Assume all data points and codewords are normalized

$$\| \mathbf{x}_j \|_2 = \| \mathbf{b}_i \|_2 = 1, \quad 1 \leq j \leq n, 1 \leq i \leq m$$

Primal subproblem for j^{th} user for fixed $\mathbf{B}$

$$\min_{w_1, w_2, \ldots, w_m} \quad \frac{1}{2} \| \mathbf{x}_j - \sum_{i=1}^{m} w_{ji} \mathbf{b}_i \|_2^2 + \lambda \sum_{i=1}^{m} |w_{ji}|. \tag{2}$$

Dual problem of (2)

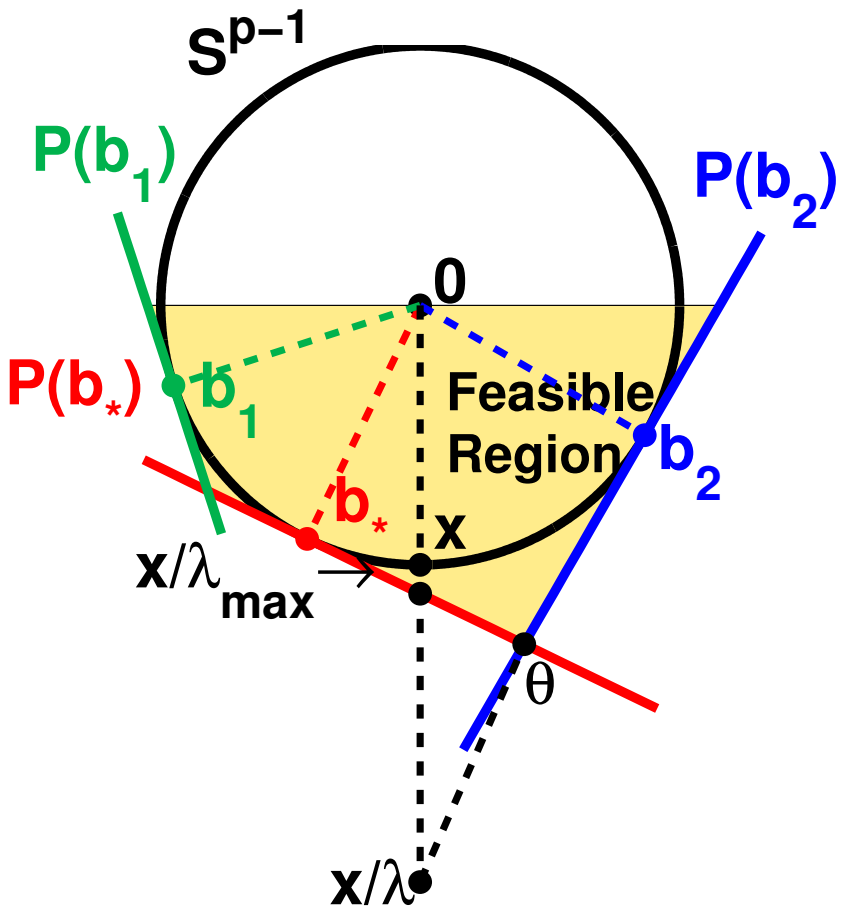
$$\begin{aligned} \max_{\boldsymbol{\theta}} \quad & \frac{1}{2} \| \mathbf{x}_j \|_2^2 - \frac{\lambda^2}{2} \| \boldsymbol{\theta} - \frac{\mathbf{x}_j}{\lambda} \|_2^2 \\ \text{s.t.} \quad & |\boldsymbol{\theta}^T \mathbf{b}_i| \leq 1 \quad \forall i = 1, 2, \ldots, m \end{aligned} \tag{3}$$

Optimal solutions of (2) $\tilde{\mathbf{w}}_j$ and (3) $\tilde{\theta}$ are related via

$$\mathbf{x}_j = \sum_{i=1}^{m} \tilde{w}_{ji} \mathbf{b}_i + \lambda \tilde{\theta}, \qquad \boldsymbol{\theta}^T \mathbf{b}_i \in \begin{cases} \{\text{sign } \tilde{w}_{ji}\} & \text{if } \tilde{w}_{ji} \neq 0, \\ [-1, 1] & \text{if } \tilde{w}_{ji} = 0 \end{cases} \tag{4}$$

Geometric Intuition

- $\mathbf{x}_j$ and $\mathbf{b}_i$ lie on unit sphere S^{p-1}
- Define:
  - $P(\mathbf{y}) = \{\mathbf{z} : \mathbf{z}^T\mathbf{y} = 1\}$
  - $H(\mathbf{y}) = \{\mathbf{z} : \mathbf{z}^T\mathbf{y} \leq 1\}$
  - $\lambda_{\max} = \max_i |\mathbf{x}_j^T\mathbf{b}_i|$
  - $\mathbf{b}_* \in \{\pm\mathbf{b}_i\}_{i=1}^m \quad \Rightarrow \quad \lambda_{\max} = \mathbf{x}_j^T\mathbf{b}_*$
- $|\theta^T\mathbf{b}_i| \leq 1 : \theta$ must be in $H(\mathbf{b}_i)$ and $H(-\mathbf{b}_i)$ for all i
- If $\tilde{\theta}$ is not on $P(\mathbf{b}_i)$ or $P(-\mathbf{b}_i)$
  - $\tilde{w}_{ji}$=0 $\Rightarrow$ safely discard $\mathbf{b}_i$ from (2)

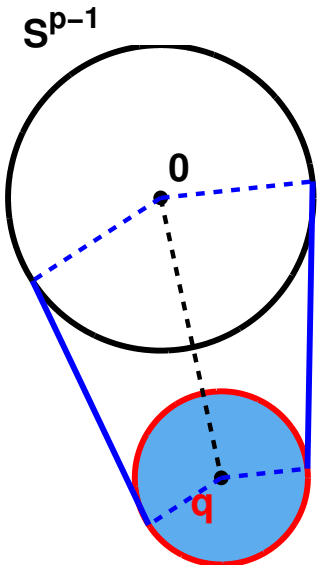


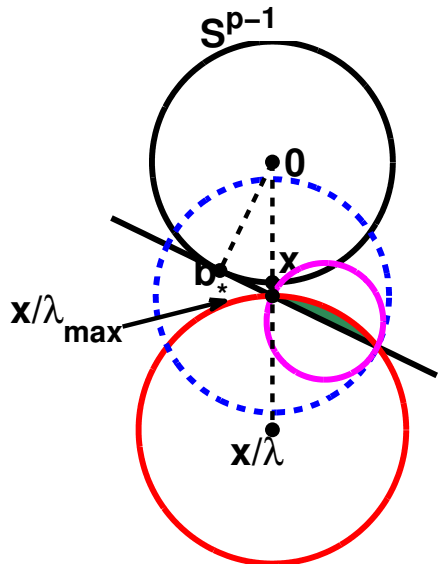
Sphere Test #1

Lemma 1: $\tilde{\theta}$ satisfies $\| \tilde{\theta} - \mathbf{q} \|_2 \leq r$, then $|\mathbf{q}^T \mathbf{b}_i| < (1 - r) \Rightarrow \tilde{w}_{ji} = 0$

Sphere Test # 1 (El Ghaoui's SAFE Rule)

$$\text{If } |\mathbf{x}_j^T \mathbf{b}_i| < \lambda - 1 + \lambda/\lambda_{\max}, \text{ then } \tilde{w}_{ji} = 0$$
$$\mathbf{q} = \mathbf{x}_j/\lambda \qquad r = 1/\lambda - 1/\lambda_{\max}$$





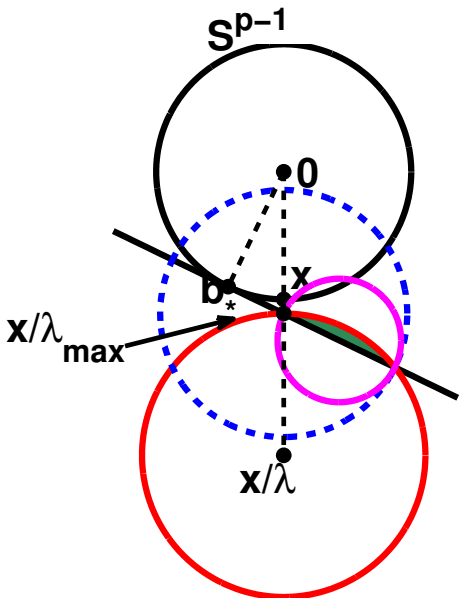
Sphere Test #2

Lemma 2: If $(a) \| \theta - \mathbf{x}_j/\lambda \|_2 \leq 1/\lambda - 1/\lambda_{\max}$ *and* $(b)\ \theta^T \mathbf{b}_* \leq 1$

Then $\| \theta - \mathbf{x}_j/\lambda_{\max} \|_2 \leq 2\sqrt{1/\lambda_{\max}^2 - 1}\,(\lambda_{\max}/\lambda - 1)$

Sphere Test # 2

If $|\mathbf{x}_j^T \mathbf{b}_i| \leq \lambda_{\max}\left(1 - 2\sqrt{1/\lambda_{\max}^2 - 1}\,(\lambda_{\max}/\lambda - 1)\right)$, then $\tilde{w}_{ji} = 0$

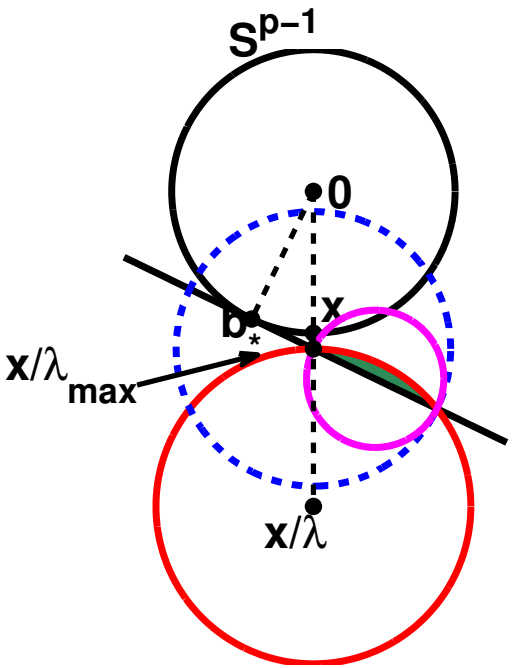


Sphere Test # 3

Lemma 3: When $\lambda_{max} > \sqrt{3}/2$, if ST1 discards $\mathbf{b}_i$, then so does ST2

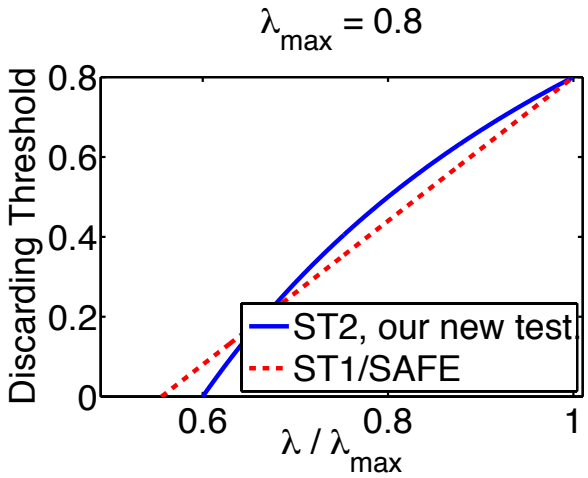
Sphere Test # 3

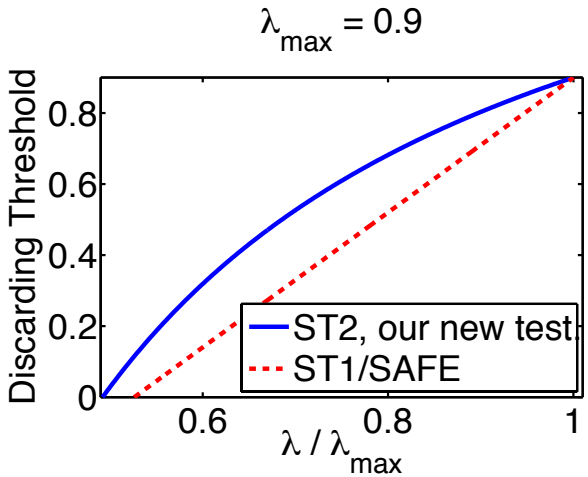
$$|\mathbf{x}_j^T \mathbf{b}_i - (\lambda_{max} - \lambda)\mathbf{b}_*^T \mathbf{b}_i| < \lambda(1 - \sqrt{1/\lambda_{max}^2 - 1}(\lambda_{max}/\lambda - 1)) \Rightarrow \tilde{w}_{ji} = 0$$



Performance Comparison

- ST3 completely outperforms ST2
- When $\lambda_{\max} > \sqrt{3}/2 \approx 0.866$, ST2 completely outperforms ST1/SAFE
- Two passes through dictionary
  1. Holds $\mathbf{x}_j, \mathbf{u}, \mathbf{b}_i \in \mathbb{R}^p$ in memory and compute $\mathbf{u}_i = \mathbf{x}_j^T \mathbf{b}_i$
  2. Holds $\mathbf{u}, \mathbf{b}_*, \mathbf{b}_i$ in memory and computes $\mathbf{b}_*^T \mathbf{b}_i$ and executes test





Scale Indifference

Definition 1: $\mathcal{X}$ satisfies the scale indifference (SI) property if $\forall \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$, with $\mathbf{x}_1 \neq \mathbf{x}_2$, and $\forall \gamma \neq 0, \mathbf{x}_1 \neq \gamma \mathbf{x}_2$

- Important for random projection to preserve the *SI* property so it is reasonable to renormalize projected data
- Random projection matrix $\mathbf{T} \in \mathbb{R}^{d \times p}$ $\quad (d < p)$
- Use $\mathbf{TX} \in \mathbb{R}^{d \times n}$ as the new data
- With high probability, random projection preserves pairwise distances

$$(1 - \epsilon)\sqrt{d/p} \leq \frac{\| \mathbf{Tx}_1 - \mathbf{Tx}_2 \|_2}{\| \mathbf{x}_1 - \mathbf{x}_2 \|_2} \leq (1 + \epsilon)\sqrt{d/p} \qquad (5)$$

Preservation of Pairwise Distances

Theorem 1

Define: $S(\mathcal{X}) = \{\mathbf{z} : \mathbf{z} = \gamma\mathbf{x}, \mathbf{x} \in \mathcal{X}, |\gamma| \leq 1\}$.
If $\mathcal{X}$ satisfies *SI* and $\forall(\mathbf{x}_1, \mathbf{x}_2) \in S(\mathcal{X}) \times S(\mathcal{X})$ (5) is satisfied,
Then $T(\mathcal{X}) = \{\mathbf{z} : \mathbf{z} : \mathbf{T}\mathbf{x}, \mathbf{x} \in \mathcal{X}\}$ also satisfies *SI*

Theorem 2

If $\mathcal{X}$ satisfies *SI* and has a κ-sparse representation using dictionary $\mathbf{B}$, then projected data $T(\mathcal{X})$ satisifies *SI* if
$(2\kappa - 1)M(\mathbf{TB}) < 1$ $M(\cdot)$: matrix mutual coherence

Key Insight: Projected data **TX** contains rough information about the original data $\mathcal{X}$ and we can continue to use formulation (1) on **TX**

Weights

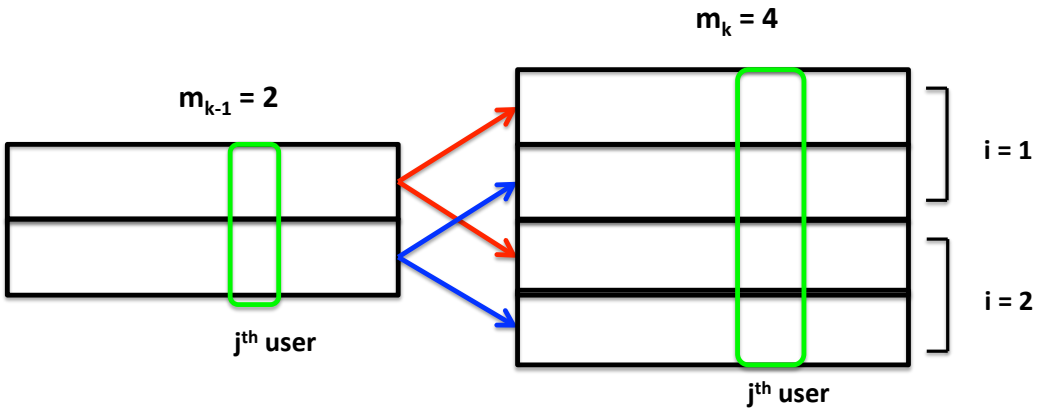
Use l random projections $\mathbf{T}_k \in \mathbb{R}^{d_k \times p} (1 \leq k \leq l)$ to extract information incrementally from data in l stages $\quad (\mathbf{B}_k \in \mathbb{R}^{d_k \times m_k} \quad \mathbf{W}_k \in \mathbb{R}^{m_k \times n})$

$$\min_{\mathbf{B}_k \mathbf{W}_k} \quad \frac{1}{2} \parallel \mathbf{T}_k \mathbf{X} - \mathbf{B}_k \mathbf{W}_k \quad \parallel_F^2 + \lambda_k \parallel \mathbf{W}_k \parallel_1 \tag{6}$$

$$\text{s.t.} \quad \parallel \mathbf{b}_i^{(k)} \parallel_2^2 \leq 1, \quad \forall i = 1, 2, \ldots, m_k$$

Enforced tree structure $(0 \leq r < m_k / m_{k-1})$, $(1 \leq i \leq m_{k-1})$

$$\mathbf{w}_j^{(k-1)}(i) = 0 \Rightarrow \mathbf{w}_j^{(k)}(r m_{k=1} + i) = 0 \tag{7}$$



Solving for $\mathbf{w}_j^k$, (7) reduces number of codewords from m_k to $(m_k / m_{k-1}) \parallel \mathbf{w}_j^{(k-1)} \parallel_0$

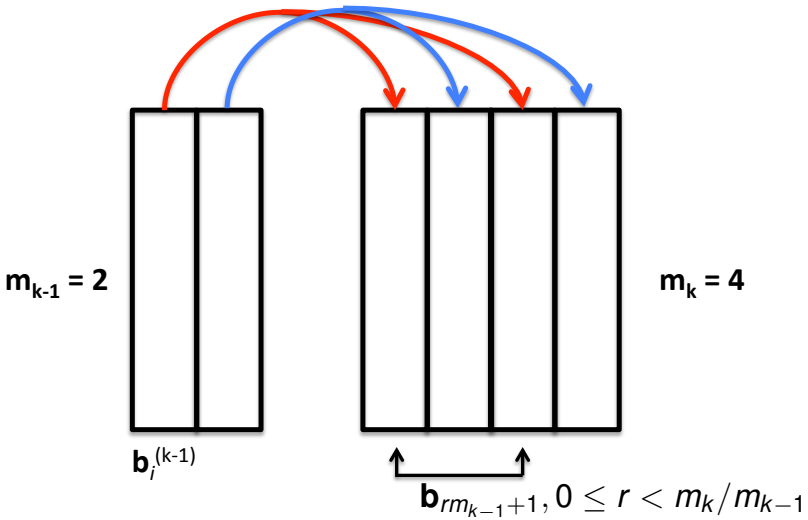
Dictionary

Divide m_k codewords into m_{k-1} groups

Define: $\mathbf{B}' = [\mathbf{b}_{rm_{k-1}+i}]_{r=0}^{m_k/m_{k-1}-1}$

$\mathbf{W}' = (rm_{k-1}+i)^{th}$ rows of $\mathbf{W}$, $r = 0, 1, \ldots, m_k/m_{k=1} - 1$

$\mathbf{B}''$ and $\mathbf{W}''$ remaining codewords and weights respectively



Fix $\mathbf{B}''$ and update $\mathbf{B}'$ by:

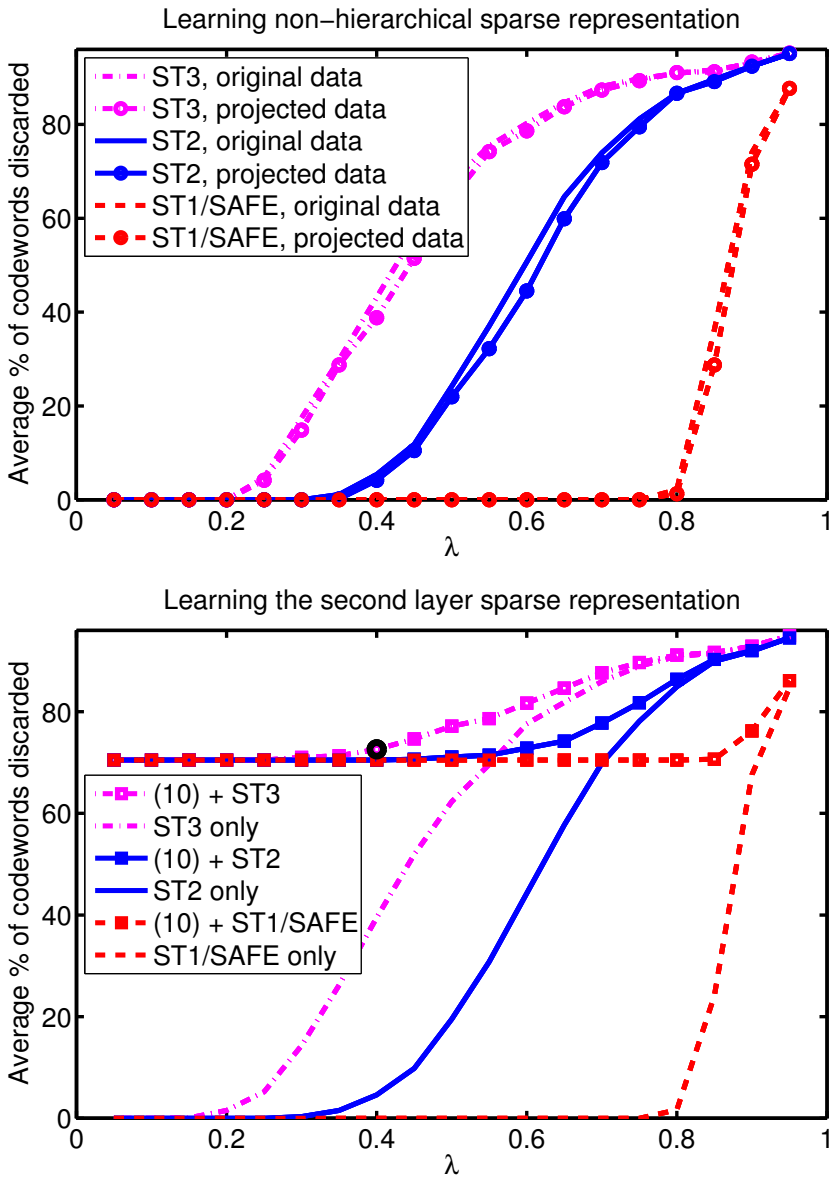
$$\min_{\mathbf{B}'} \frac{1}{2} \parallel \mathbf{z} - \mathbf{B}'\mathbf{W}' \parallel_F^2 + \lambda_k \parallel \mathbf{W}' \parallel_1$$

$$\mathbf{z} = \mathbf{T}_k\mathbf{X} - \mathbf{B}''\mathbf{W}''$$

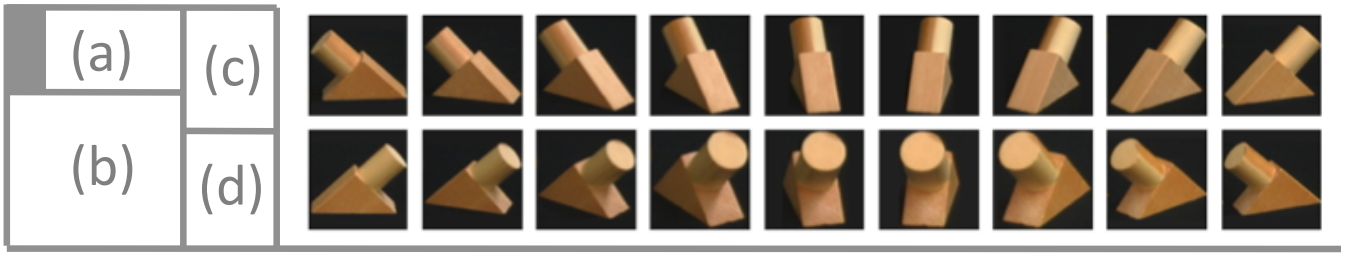
Complexity: $O(m_k^q) \Rightarrow O(m_k^q/m_{k-1}^{q-1})$

Finalizing $\mathbf{W}_k$ and $\mathbf{B}_k$, solve unconstrained QP

$$\mathbf{C}_k = \text{argmin}_{\mathbf{C}} \parallel \mathbf{X} - \mathbf{C}\mathbf{W}_k \parallel_F^2$$



- COIL Rotational Image Data
- 72 - 128x128 color images
- $(d_1, m_1, \lambda_1) = (100, 4, 0.5)$
- $(d_2, m_2) = (200, 16)$
- Result of 1^{st} layer helps 2^{nd} layer discard more codewords when tree constraint (7) is imposed


(a)
(b)
(c)
(d)

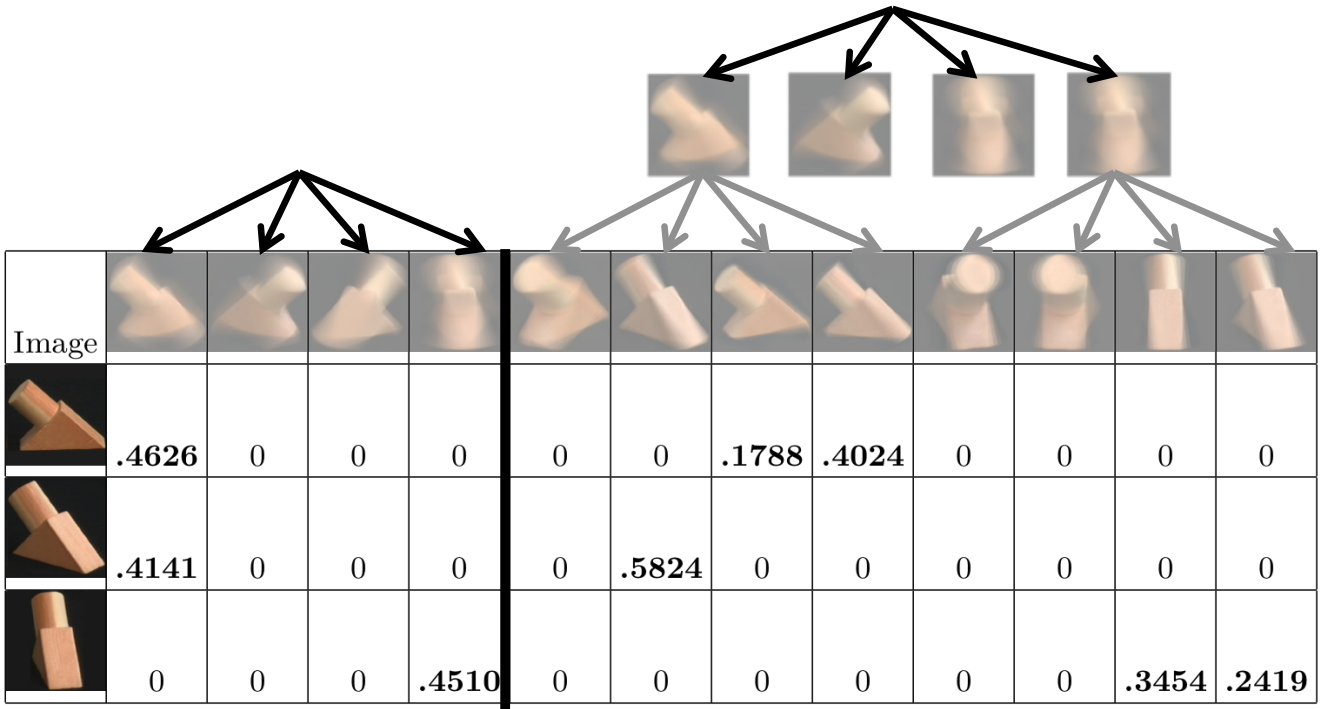
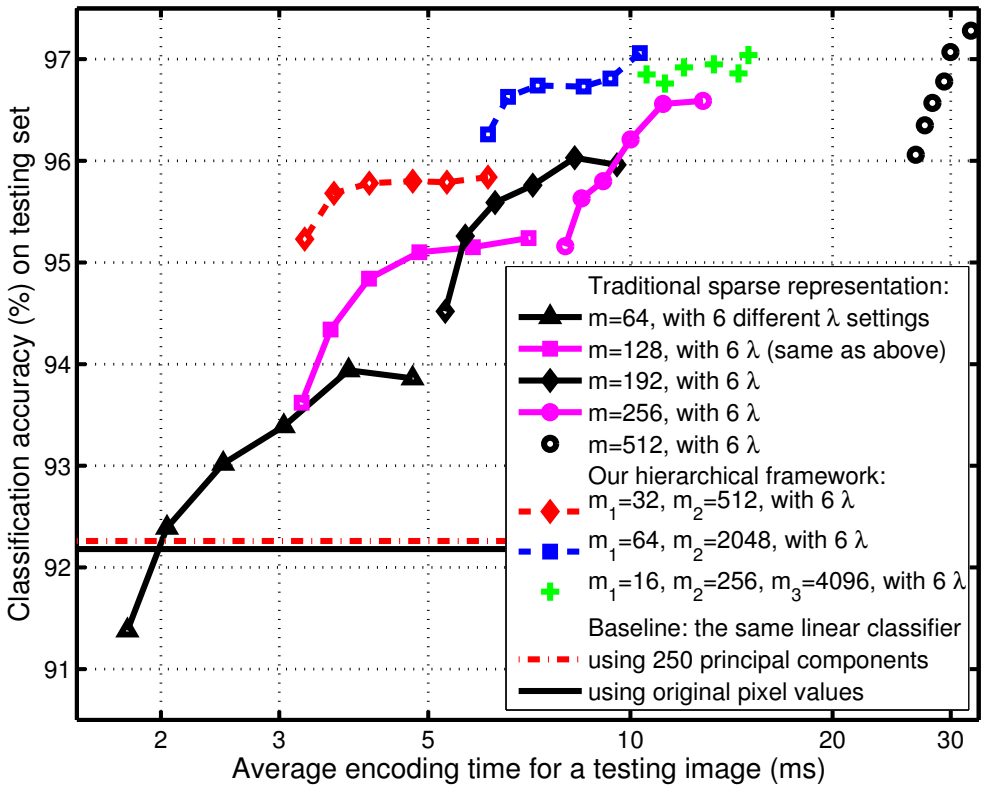
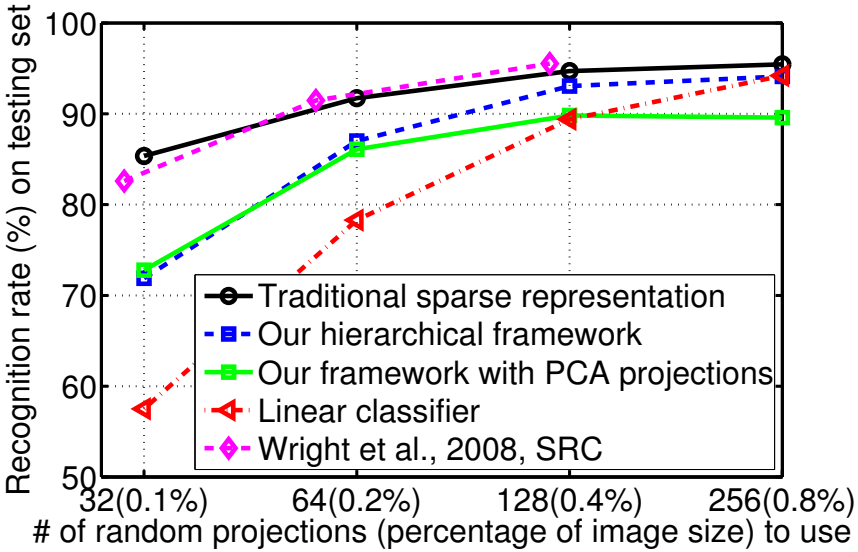
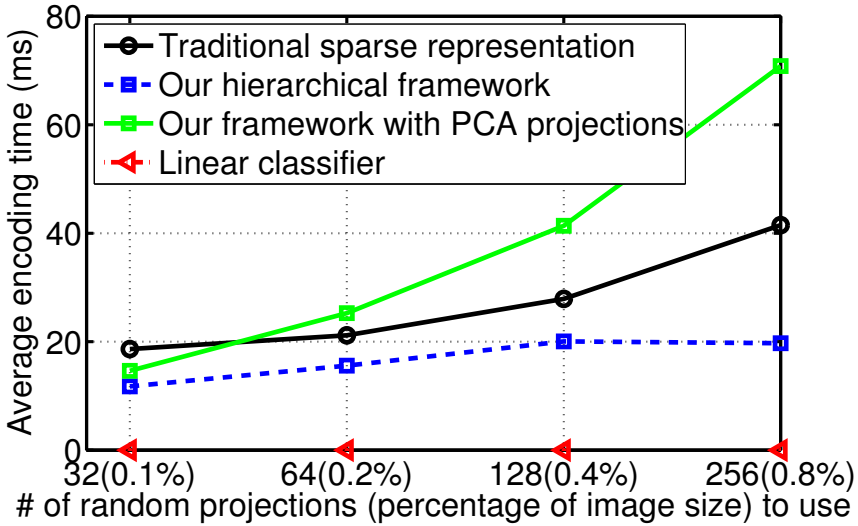


Image
.4626 0 0 0 0 0 .1788 .4024 0 0 0 0
.4141 0 0 0 0 .5824 0 0 0 0 0 0
0 0 0 .4510 0 0 0 0 0 0 .3454 .2419
First Layer
Second Layer



- MNIST Digit Classification
- Classification using **W** as features
- $\lambda \in \{.06, 0.08, 0.11, 0.16, 0.23, 0.32\}$
- Curve: same m but varying λ
- Points to the right correspond to smaller λ values
- 1% accuracy improvement given same encoding time
- Roughly $2\times$ speedup given same accuracy





- 64 cropped frontal face views
- $(m_1, m_2) = (32, 1024)$
- $(d_1, d_2) = (\frac{3}{8}p, \frac{5}{8}p)(p = 128, 256)$
- $\lambda_1 = 0.02$ and $\lambda_2 = 0.029$
- Good balance between speed and accuracy