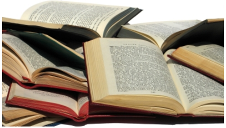
Learning sparsely used overcomplete dictionaries

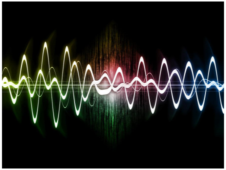
Alekh Agarwal
Microsoft Research

Joint work with Anima Anandkumar, Prateek Jain, Praneeth Netrapalli and Rashish Tandon

Motivation I: Feature learning

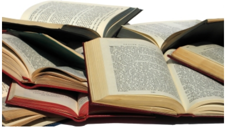
Practice

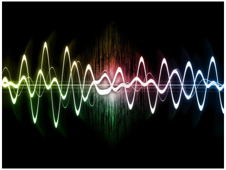




Motivation I: Feature learning

Practice





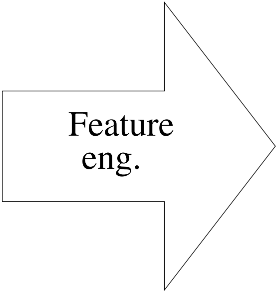
Papers

Features

$$\begin{pmatrix} 1.2 & 0.8 & 0 & 1.5 & \cdots \\ 0 & 3.5 & 2 & 0.1 & \cdots \\ 0.7 & 0 & 0.3 & 0.8 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Motivation I: Feature learning

Practice



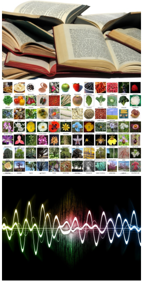
Papers

Features

$$\begin{pmatrix} 1.2 & 0.8 & 0 & 1.5 & \cdots \\ 0 & 3.5 & 2 & 0.1 & \cdots \\ 0.7 & 0 & 0.3 & 0.8 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Motivation I: Feature learning

Practice



Feature eng.

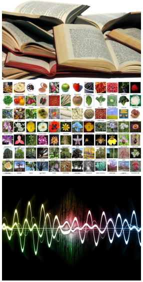
Papers
Features

$$\begin{pmatrix} 1.2 & 0.8 & 0 & 1.5 & \cdots \\ 0 & 3.5 & 2 & 0.1 & \cdots \\ 0.7 & 0 & 0.3 & 0.8 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

- Feature engineering takes considerable time and skill
- Typically critical to good performance

Motivation I: Feature learning

Practice



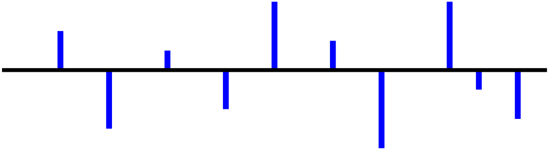
Feature eng.

Papers
Features

$$\begin{pmatrix} 1.2 & 0.8 & 0 & 1.5 & \cdots \\ 0 & 3.5 & 2 & 0.1 & \cdots \\ 0.7 & 0 & 0.3 & 0.8 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

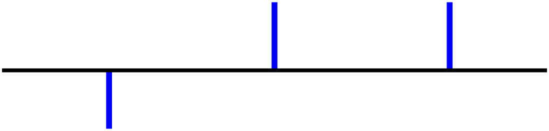
- Feature engineering takes considerable time and skill
- Typically critical to good performance
- **Can we learn good features from data?**

Motivation II: Signal compression



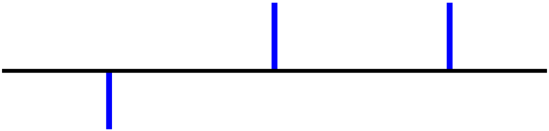
- Expensive to store high-dimensional signals

Motivation II: Signal compression



- Expensive to store high-dimensional signals
- Sparse signals have compact representation

Motivation II: Signal compression



- Expensive to store high-dimensional signals
- Sparse signals have compact representation
- **Can we learn a representation where signals of interest are sparse?**

Dictionary learning in practice

Original JPEG (26.59dB) JPEG-2000 (27.81dB) PCA (29.27dB) K-SVD (33.26dB)

Image compression (Bruckstein et al., 2009)

Dictionary learning in practice

Original JPEG (26.59dB) JPEG-2000 (27.81dB) PCA (29.27dB) K-SVD (33.26dB)

Image compression (Bruckstein et al., 2009)

- Similar successes in image denoising, inpainting, superresolution, . . .

Dictionary learning in practice

Image compression (Bruckstein et al., 2009)

- Similar successes in image denoising, inpainting, superresolution, ...
- **Non-convex optimization, limited theoretical understanding**

Dictionary learning setup

Goal

Find a dictionary with r elements such that each data point is a combination of only s dictionary elements.

Dictionary learning setup

Goal

Find a dictionary with r elements such that each data point is a combination of only s dictionary elements.

Dictionary learning setup

Goal

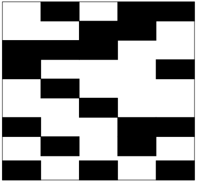
Find a dictionary with r elements such that each data point is a combination of only s dictionary elements.

Y

$=$

A^*

X^*



Examples

Dictionary

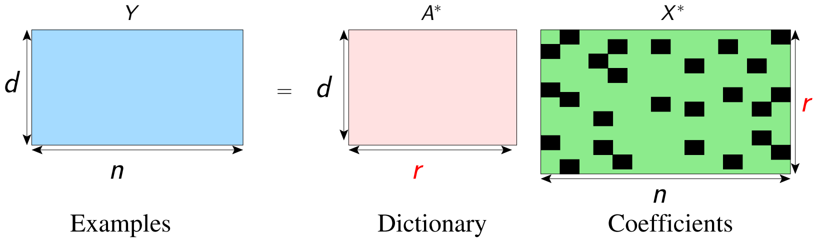
Coefficients

- Encode faces using dictionary rather than pixel values
- Sparsity for compression, signal processing ...

Dictionary learning setup

Goal

Find a dictionary with r elements such that each data point is a combination of only s dictionary elements.



- Topic models, overlapping clustering, image representation
- **Overcomplete setting**, $r \gg d$ relevant in practice

Alternating minimization

Objective

$$\min_{A,X} \underbrace{\|X\|_1}_{\sum_{i,j}|X_{ij}|} \quad \textit{subject to} \quad Y = AX$$

- Dominant approach in practice
- Start with initial dictionary $A(0)$
- **Sparse regression** for coefficients given dictionary

$$X(t+1)_i = \arg\min_{x\in\mathbb{R}^r} \|x\|_1 \;\; \text{s.t.} \quad \|Y_i - A(t)x\|_2 \leq \epsilon_t$$

Alternating minimization

Objective

$$\min_{A,X} \underbrace{\|X\|_1}_{\sum_{i,j}|X_{ij}|} \quad \textit{subject to} \quad Y = AX$$

- Dominant approach in practice
- Start with initial dictionary $A(0)$
- **Sparse regression** for coefficients given dictionary
- **Least squares** for dictionary given coefficients

$$A(t+1) = YX(t+1)^{+} \quad \text{i.e. } Y \approx A(t+1)X(t+1)$$

Alternating minimization

Objective

$$\min_{A,X} \underbrace{\|X\|_1}_{\sum_{i,j} |X_{ij}|} \quad \textit{subject to} \quad Y = AX$$

- Dominant approach in practice
- Start with initial dictionary $A(0)$
- **Sparse regression** for coefficients given dictionary
- **Least squares** for dictionary given coefficients

$$A(t+1) = YX(t+1)^{+} \quad \text{i.e. } Y \approx A(t+1)X(t+1)$$

- Similar to EM for this problem

Alternating minimization

Objective

$$\min_{A,X} \underbrace{\|X\|_1}_{\sum_{i,j}|X_{ij}|} \quad \textit{subject to} \quad Y = AX$$

- Dominant approach in practice
- Start with initial dictionary $A(0)$
- **Sparse regression** for coefficients given dictionary
- **Least squares** for dictionary given coefficients

$$A(t+1) = YX(t+1)^{+} \quad \text{i.e. } Y \approx A(t+1)X(t+1)$$

- Similar to EM for this problem
- Does *not* converge to global optimum from arbitrary $A(0)$

Alternating minimization goal

$$(\widehat{A}, \widehat{X}) = \min_{A,X} \|X\|_1 \quad \text{subject to} \quad Y = AX$$

- $Y = AX$ is a non-convex constraint
- Average of solutions is not a solution!

$$Y = AX, \quad Y = (-A)(-X),$$

Alternating minimization goal

$$(\widehat{A}, \widehat{X}) = \min_{A,X} \|X\|_1 \quad \text{subject to} \quad Y = AX$$

- $Y = AX$ is a non-convex constraint
- Average of solutions is not a solution!

$$Y = AX, \quad Y = (-A)(-X), \quad \mathbf{Y} \neq \left(\frac{\mathbf{A} + (-\mathbf{A})}{\mathbf{2}}\right)\left(\frac{\mathbf{X} + (-\mathbf{X})}{\mathbf{2}}\right)$$

- Non-convex optimization, NP-hard in general

Previous theory work

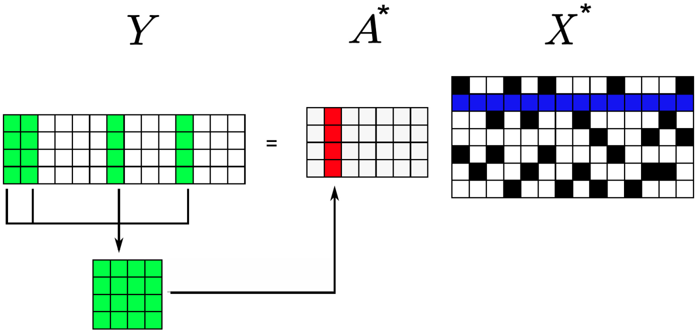
- Exact recovery in undercomplete setting by Spielman et al. via linear programming

Previous theory work

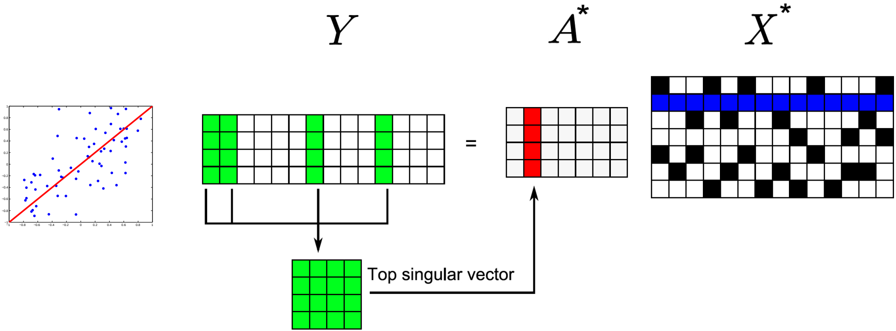
- Exact recovery in undercomplete setting by Spielman et al. via linear programming

- We combine alternating minimization with a *novel initialization*
- **Global optimum despite non-convexity in overcomplete setting**

Initialization: Key ideas



- Find several samples with a common dictionary element

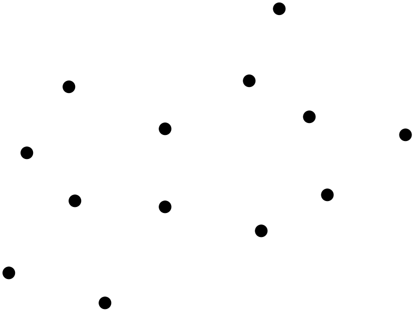


- Find several samples with a common dictionary element
- Top singular vector of these samples is an estimate of this element

Correlation graph

Definition (Correlation graph)

- One node for each example

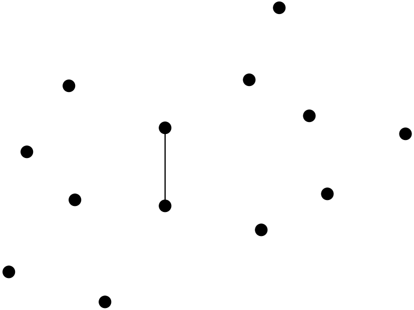


- Large correlation $\Rightarrow$ common dictionary element

Correlation graph

Definition (Correlation graph)

- One node for each example
- Edge $\{Y_i, Y_j\}$ if $|\langle Y_i, Y_j \rangle| \geq \rho$

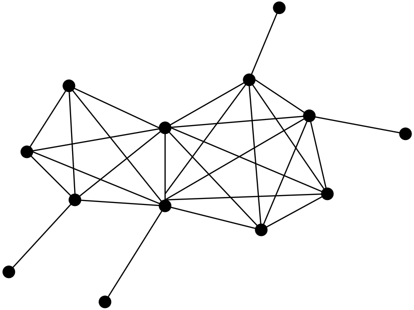


- Large correlation $\Rightarrow$ common dictionary element

Correlation graph

Definition (Correlation graph)

- One node for each example
- Edge $\{Y_i, Y_j\}$ if $|\langle Y_i, Y_j \rangle| \geq \rho$

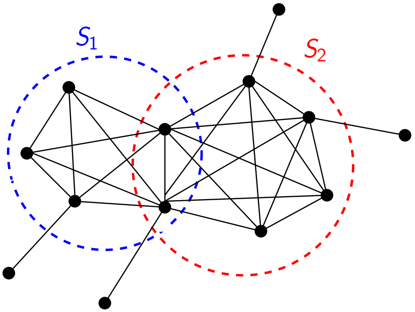


- Large correlation $\Rightarrow$ common dictionary element

Correlation graph

Definition (Correlation graph)

- One node for each example
- Edge $\{Y_i, Y_j\}$ if $|\langle Y_i, Y_j \rangle| \geq \rho$

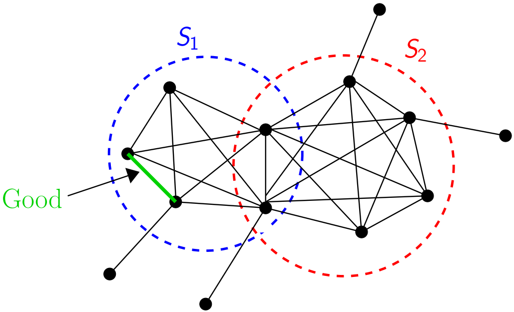


- Large correlation $\Rightarrow$ common dictionary element
- Samples in a clique contain a common dictionary element

Correlation graph

Definition (Correlation graph)

- One node for each example
- Edge $\{Y_i, Y_j\}$ if $|\langle Y_i, Y_j \rangle| \geq \rho$

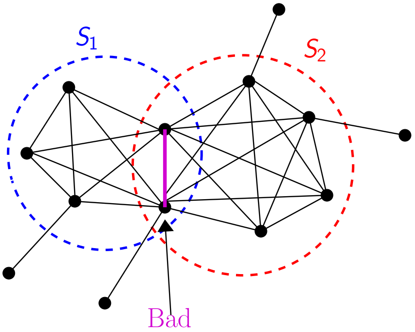


- Large correlation $\Rightarrow$ common dictionary element
- Samples in a clique contain a common dictionary element
- Easy to construct cliques

Correlation graph

Definition (Correlation graph)

- One node for each example
- Edge $\{Y_i, Y_j\}$ if $|\langle Y_i, Y_j \rangle| \geq \rho$



- Large correlation $\Rightarrow$ common dictionary element
- Samples in a clique contain a common dictionary element
- Easy to construct cliques

Initialization algorithm

1. Construct correlation graph G_ρ given a threshold ρ
2. For each edge (Y_i, Y_j) in G_ρ

Initialization algorithm

1. Construct correlation graph G_ρ given a threshold ρ
2. For each edge (Y_i, Y_j) in G_ρ
   - If (Y_i, Y_j) is good
     (a) Let S be all common neighbors of Y_i and Y_j

Initialization algorithm

1. Construct correlation graph G_ρ given a threshold ρ
2. For each edge (Y_i, Y_j) in G_ρ
   - If (Y_i, Y_j) is good
     (a) Let S be all common neighbors of Y_i and Y_j
     (b) Let M be the covariance matrix of S: $\sum_{i \in S} Y_i Y_i^T$
     (c) Set $\widehat{a}$ to the **top singular vector** of M
3. Each vector $\widehat{a}$ is estimate of some A_i^*

- Similar algorithm developed simultaneously and independently in Arora et al. (2013)

Assumptions

- **Incoherent dictionary:** $|\langle A_i^*, A_j^* \rangle| \leq \mu_0/\sqrt{d}$

Assumptions

- **Incoherent dictionary:** $|\langle A_i^*, A_j^* \rangle| \leq \mu_0 / \sqrt{d}$
- **Sparse coefficients:** Each sample has at most s non-zero X_{ij}^* with random sparsity pattern

Exact recovery

Theorem (AAJNT'13)

Suppose we have $\mathcal{O}(r^2)$ examples. Use graph clustering algorithm to initialize alternating minimization. With high probability, for all $t \geq 1$ and $i = 1, 2, \ldots, r$

$$\|A(t)_i - \mathbf{A_i^*}\|_2 \leq \|A(0)_i - \mathbf{A_i^*}\|_2 \, 2^{-t}$$

- Exact recovery from $\mathcal{O}(r^2)$ samples
- **Global optimum through novel initialization**

Exact recovery

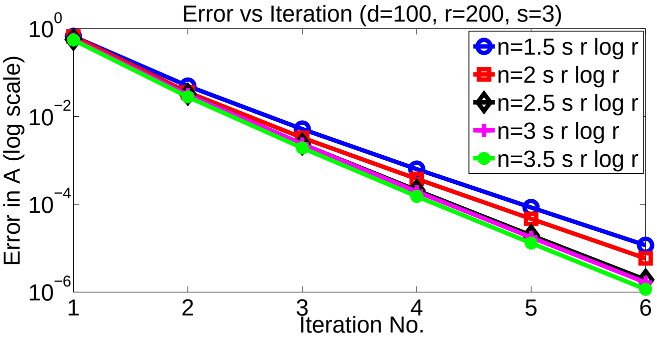
Theorem (AAJNT'13)

Suppose we have $\mathcal{O}(r^2)$ examples. Use graph clustering algorithm to initialize alternating minimization. With high probability, for all $t \geq 1$ and $i = 1, 2, \ldots, r$

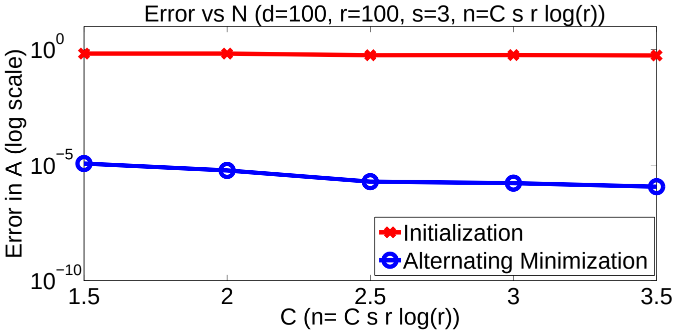
$$\|A(t)_i - \mathbf{A_i^*}\|_2 \leq \|A(0)_i - \mathbf{A_i^*}\|_2 \, 2^{-t}$$

- Exact recovery from $\mathcal{O}(r^2)$ samples
- **Global optimum through novel initialization**
- Approximate recovery in initialization step
- Local linear convergence of alternating minimization

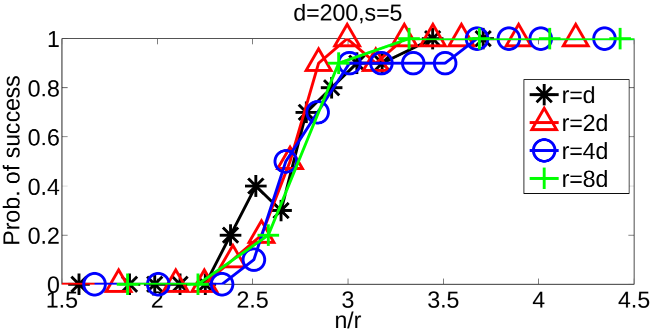
Local linear convergence



One-shot vs alternating



Sample complexity



Alternating minimization proof sketch

- Ideally want

$A(0)$ A^*

X^*

Alternating minimization proof sketch

- Ideally want

$$\begin{tikzcd}
A(0) \arrow[d] & & A^* \\
X(1) & & X^*
\end{tikzcd}$$

Alternating minimization proof sketch

- Ideally want

$$
\begin{tikzcd}
A(0) \arrow[dr] & & A(1) & & A^* \\
& X(1) \arrow[ur] & & & X^*
\end{tikzcd}
$$

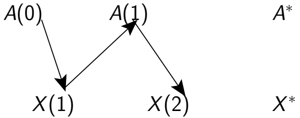
Alternating minimization proof sketch

- Ideally want

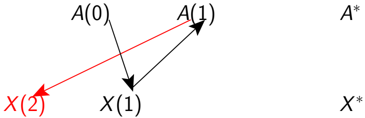
$$\begin{tikzcd}
A(0) \arrow[dr] & & A(1) \arrow[dr] & & A^* \\
& X(1) \arrow[ur] & & X(2) & X^*
\end{tikzcd}$$

Alternating minimization proof sketch

- Ideally want



- But what about



Alternating minimization proof sketch (contd.)

Lemma

Suppose $\|X(t+1) - X^*\|_\infty = \mathcal{O}(1/s)$. *Then*

$$\|A(t+1)_i - A_i^*\|_2 = \mathcal{O}\left(\frac{s^2}{\sqrt{d}}\|X(t+1) - X^*\|_\infty\right)$$

.

- $s^2 \leq \sqrt{d}$ ensures error decreases
- Contraction by relating $\|X(t+1) - X^*\|_\infty$ to $\|A(t)_i - A_i^*\|_2$
- Good initialization ensures precondition

Conclusions

- Provable recovery of overcomplete dictionaries
- Global optimality through novel initialization
- Local linear convergence of alternating minimization
- Local convexity under same initialization
- General theory for latent variable models

- A Clustering Approach to Learn Sparsely-Used Overcomplete Dictionaries, `arxiv:1309.1952`
- Learning Sparsely Used Overcomplete Dictionaries via Alternating Minimization, `arxiv:1310.7991`

Questions?