Scaling the Hierarchical Topic Modeling Mountain

Neural NMF and Iterative Projection Methods

Jamie Haddock
IPAM,
January 29, 2020

Computational and Applied Mathematics
UCLA
Talk Outline

1. Introduction
2. Neural NMF
3. Iterative Projection Methods
4. Applications
5. Conclusions
Introduction
Motivation

- MyLymeData: large collection of Lyme disease patient survey data collected by LymeDisease.org (~12,000 patients, 100s of questions)
Motivation

- MyLymeData: large collection of Lyme disease patient survey data collected by LymeDisease.org (~12,000 patients, 100s of questions)

Diagram:
- Supertopics
- Topics
- Symptoms
Motivation

- MyLymeData: large collection of Lyme disease patient survey data collected by LymeDisease.org (~12,000 patients, 100s of questions)

⇒ hypothesis formation about post-treatment/chronic Lyme disease
Motivation

- MyLymeData: large collection of Lyme disease patient survey data collected by LymeDisease.org (~12,000 patients, 100s of questions)

\[ \text{supertopics} \}
\[ \text{topics} \}
\[ \text{symptoms} \}

⇒ hypothesis formation
⇒ about post-treatment/chronic Lyme disease
Motivation

- MyLymeData: large collection of Lyme disease patient survey data collected by LymeDisease.org (~12,000 patients, 100s of questions)

Main question: How can we identify the topic hierarchy of MyLymeData symptom questions?
Main question: How can we identify the topic hierarchy of MyLymeData symptom questions?
Motivation

Main question: How can we identify the topic hierarchy of MyLymeData symptom questions?

Answer: Neural Nonnegative Matrix Factorization

[Gao, H., Molitor, Needell, Sadovnik, Will, Zhang '19]
Main question: How can we identify the topic hierarchy of MyLymeData symptom questions?

Answer: Neural Nonnegative Matrix Factorization

[Gao, H., Molitor, Needell, Sadovnik, Will, Zhang ’19]
Main question: How can we identify the topic hierarchy of MyLymeData symptom questions?

Answer: Neural Nonnegative Matrix Factorization

[Gao, H., Molitor, Needell, Sadovnik, Will, Zhang ’19]

Sampling Kaczmarz-Motzkin Methods

[H., Ma ’19], [De Loera, H., Needell ’17]
Topic Modeling

▷ principal component analysis (PCA)
  [Pearson 1901]
  [Hotelling 1933]

Pearson, K. (1901) On lines and planes of closest fit to systems of points in space.
Topic Modeling

➤ principal component analysis (PCA)
  [Pearson 1901]
  [Hotelling 1933]
➤ latent dirichlet allocation (LDA)
  [Pritchard, Stephens, Donnelly 2000]
  [Blei, Ng, Jordan 2003]
Topic Modeling

- principal component analysis (PCA)
  [Pearson 1901]
  [Hotelling 1933]

- latent dirichlet allocation (LDA)
  [Pritchard, Stephens, Donnelly 2000]
  [Blei, Ng, Jordan 2003]

- clustering (k-means, Gaussian mixtures)
  [Lloyd 1957]
  [Pearson 1894]
Topic Modeling

▷ principal component analysis (PCA)
  [Pearson 1901]
  [Hotelling 1933]
▷ latent dirichlet allocation (LDA)
  [Pritchard, Stephens, Donnelly 2000]
  [Blei, Ng, Jordan 2003]
▷ clustering (k-means, Gaussian mixtures)
  [Lloyd 1957]
  [Pearson 1894]
▷ nonnegative matrix factorization (NMF)
  [Paatero, Tapper 1994]
  [Lee, Seung 1999]
**Model**: Given nonnegative data $X$, compute nonnegative $A$ and $S$ of lower rank so that

$$X \approx AS.$$
Nonnegative Matrix Factorization (NMF)

**Model:** Given nonnegative data $X$, compute nonnegative $A$ and $S$ of lower rank so that

$$X \approx AS.$$
Nonnegative Matrix Factorization (NMF)

Often formulated as optimization problem

\[
\min_{A \in \mathbb{R}^{m \times k}_{\geq 0}, S \in \mathbb{R}^{k \times n}_{\geq 0}} \|X - AS\|_F.
\]
Nonnegative Matrix Factorization (NMF)

Often formulated as optimization problem

\[
\min_{A \in \mathbb{R}^{m \times k}_{\geq 0}, S \in \mathbb{R}^{k \times n}_{\geq 0}} \|X - AS\|_F.
\]

Non-convex optimization problem, NP-hard to compute global optimum for fixed \(k\) [Vavasis 2008]
Hierarchical NMF

**Model:** Sequentially factorize

\[ X \approx A^{(0)} S^{(0)}, \quad S^{(0)} \approx A^{(1)} S^{(1)}, \quad S^{(1)} \approx A^{(2)} S^{(2)}, \ldots, \quad S^{(L-1)} \approx A^{(L)} S^{(L)}. \]
Hierarchical NMF

**Model:** Sequentially factorize

\[ X \approx A^{(0)} S^{(0)}, S^{(0)} \approx A^{(1)} S^{(1)}, S^{(1)} \approx A^{(2)} S^{(2)}, \ldots, S^{(L-1)} \approx A^{(L)} S^{(L)}. \]
Hierarchical NMF

**Model:** Sequentially factorize

\[ X \approx A^{(0)} S^{(0)}, S^{(0)} \approx A^{(1)} S^{(1)}, S^{(1)} \approx A^{(2)} S^{(2)}, \ldots, S^{(L-1)} \approx A^{(L)} S^{(L)}. \]
Hierarchical NMF

**Model:** Sequentially factorize

\[ X \approx A^{(0)} S^{(0)}, \ S^{(0)} \approx A^{(1)} S^{(1)}, \ S^{(1)} \approx A^{(2)} S^{(2)}, \ldots, \ S^{(L-1)} \approx A^{(L)} S^{(L)}. \]

\[ \Delta k^{(\ell)}: \text{supertopics collecting } k^{(\ell-1)} \text{ subtopics} \]

[Cichocki, Zdunek '06]
Hierarchical NMF

**Model:** Sequentially factorize

\[ X \approx A^{(0)} S^{(0)}, S^{(0)} \approx A^{(1)} S^{(1)}, S^{(1)} \approx A^{(2)} S^{(2)}, \ldots, S^{(L-1)} \approx A^{(L)} S^{(L)}. \]

\[ \Delta k^{(\ell)}: \text{supertopics collecting } k^{(\ell-1)} \text{ subtopics} \]

\[ \Rightarrow \text{error propagates through layers} \]

[Cichocki, Zdunek '06]
Neural NMF
Hierarchical NMF

\[ \text{symptoms} \{ \]

\[ X \]

\[ m \]

\[ n \]
Hierarchical NMF
Hierarchical NMF can be implemented in a feed-forward neural network structure.

- **Supertopics**
- **Topics**
- **Symptoms**

$$X \approx m \times k^{(0)} \times \mathbf{A}^{(0)} \times k^{(0)} \times S^{(0)} \approx m \times n \times k^{(0)} \times k^{(1)} \times S^{(1)}$$
Hierarchical NMF

- hNMF can be implemented in a feed-forward neural network structure.
**Goal:** Identify weights $W_1, W_2, \ldots, W_L$ to minimize model error

$$\sum_{n=1}^{N} E(\{W_i\}) = f(y(x_n, \{W_i\}), x_n, t_n).$$
**Goal:** Identify weights $W_1, W_2, ..., W_L$ to minimize model error

$$E(\{W_i\}) = \sum_{n=1}^{N} \|y(x_n, \{W_i\}) - t_n\|^2.$$
Feed-forward Neural Networks

Goal: Identify weights $W_1, W_2, ..., W_L$ to minimize model error

$$E(\{W_i\}) = \sum_{n=1}^{N} f(y(x_n, \{W_i\}), x_n, t_n).$$
Goal: Identify weights $W_1, W_2, ..., W_L$ to minimize model error

$$E(\{W_i\}) = \sum_{n=1}^{N} f(y(x_n, \{W_i\}), x_n, t_n).$$
Feed-forward Neural Networks

**Goal:** Identify weights $W_1, W_2, ..., W_L$ to minimize model error

$$E(\{W_i\}) = \sum_{n=1}^{N} f(y(x_n, \{W_i\}), x_n, t_n).$$

**Training:**

- **forward propagation:**
  - $z_1 = \sigma(W_1x)$,
  - $z_2 = \sigma(W_2z_1)$, ...
  - $y = \sigma(W_Lz_{L-1})$
Goal: Identify weights $W_1, W_2, \ldots, W_L$ to minimize model error

$$E(\{W_i\}) = \sum_{n=1}^{N} f(y(x_n, \{W_i\}), x_n, t_n).$$

Training:

▷ forward propagation:

$z_1 = \sigma(W_1x)$,
$z_2 = \sigma(W_2z_1)$, \ldots,
$y = \sigma(W_Lz_{L-1})$

▷ back propagation:
update $\{W_i\}$ with
$$\nabla E(\{W_i\}).$$
Our method: Neural NMF

**Goal:** Develop true forward and back propagation algorithms for hNMF.

\[ \text{Pin the values of } S^{(\ell)} \text{ to those of } A^{(\ell)} \text{ by recursively setting } S^{(\ell)} := q(X, A^{(\ell)}). \]
**Goal:** Develop true forward and back propagation algorithms for hNMF.

- Regard the $A$ matrices as independent variables, determine the $S$ matrices from the $A$ matrices.
Our method: Neural NMF

Goal: Develop true forward and back propagation algorithms for hNMF.

- Regard the $A$ matrices as independent variables, determine the $S$ matrices from the $A$ matrices.
- Define $q(X, A) := \arg\min_{S \geq 0} \|X - AS\|_F^2$. 
Our method: Neural NMF

**Goal:** Develop true forward and back propagation algorithms for hNMF.

- Regard the $A$ matrices as independent variables, determine the $S$ matrices from the $A$ matrices.
- Define $q(X, A) := \arg\min_{S \geq 0} \|X - AS\|_F^2$.
- Pin the values of $S$ to those of $A$ by recursively setting $S^{(\ell)} := q(S^{(\ell-1)}, A^{(\ell)})$. 
Our method: Neural NMF

**Goal:** Develop true forward and back propagation algorithms for hNMF.

- Regard the A matrices as independent variables, determine the S matrices from the A matrices.
- Define \( q(X, A) := \arg\min_{S \geq 0} \|X - AS\|_F^2 \).
- Pin the values of S to those of A by recursively setting \( S^{(\ell)} := q(S^{(\ell-1)}, A^{(\ell)}) \).
Least-squares Subroutine

least-squares is a fundamental subroutine in forward-propagation

\[
\begin{align*}
X & \approx A \times k \\
& \approx n \\
& \approx m \\
& \approx \begin{array}{c}
\text{S}
\end{array}
\end{align*}
\]
least-squares is a fundamental subroutine in forward-propagation
least-squares is a fundamental subroutine in forward-propagation.

iterative projection methods can solve these problems.
Iterative Projection Methods
We are interested in solving highly overdetermined systems of equations, \( Ax = b \), where \( A \in \mathbb{R}^{m \times n} \), \( b \in \mathbb{R}^m \) and \( m \gg n \). Rows are denoted \( a^T_i \).
We are interested in solving **highly overdetermined systems of equations**, $Ax = b$, where $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ and $m \gg n$. Rows are denoted $a_i^T$. 
We are interested in solving highly overdetermined systems of equations, \( Ax = b \), where \( A \in \mathbb{R}^{m \times n} \), \( b \in \mathbb{R}^m \) and \( m \gg n \). Rows are denoted \( a_i^T \).
Iterative Projection Methods

If \( \{ \mathbf{x} \in \mathbb{R}^n : A\mathbf{x} = \mathbf{b} \} \) is nonempty, these methods construct an approximation to a solution:

1. Randomized Kaczmarz Method

Applications:

1. Tomography (Algebraic Reconstruction Technique)
If \( \{ \mathbf{x} \in \mathbb{R}^n : A\mathbf{x} = \mathbf{b} \} \) is nonempty, these methods construct an approximation to a solution:

1. Randomized Kaczmarz Method
2. Motzkin's Method

Applications:

1. Tomography (Algebraic Reconstruction Technique)
2. Linear programming
If \( \{ x \in \mathbb{R}^n : Ax = b \} \) is nonempty, these methods construct an approximation to a solution:

1. Randomized Kaczmarz Method
2. Motzkin's Method
3. Sampling Kaczmarz-Motzkin Methods (SKM)

Applications:

1. Tomography (Algebraic Reconstruction Technique)
2. Linear programming
3. Average consensus (greedy gossip with eavesdropping)
Given $x_0 \in \mathbb{R}^n$:

1. Choose $i_k \in [m]$ with probability $\frac{||a_{ik}||^2}{||A||^2_F}$.

2. Define $x_k := x_{k-1} + \frac{b_{ik} - a_{ik}^T x_{k-1}}{||a_{ik}||^2} a_{ik}$.

3. Repeat.

[Kaczmarz 1937], [Strohmer, Vershynin 2009]
Kaczmarz Method

Given $x_0 \in \mathbb{R}^n$:
1. Choose $i_k \in [m]$ with probability $\frac{\|a_{ik}\|^2}{\|A\|^2_F}$.
2. Define $x_k := x_{k-1} + \frac{b_{ik} - a_{ik}^T x_{k-1}}{\|a_{ik}\|^2} a_{ik}$.
3. Repeat.

[Kaczmarz 1937], [Strohmer, Vershynin 2009]
Kaczmarz Method

Given $x_0 \in \mathbb{R}^n$:

1. Choose $i_k \in [m]$ with probability $\frac{\|a_{i_k}\|^2}{\|A\|^2_F}$.

2. Define $x_k := x_{k-1} + \frac{b_{i_k} - a_{i_k}^T x_{k-1}}{\|a_{i_k}\|^2} a_{i_k}$.

3. Repeat.

[Kaczmarz 1937], [Strohmer, Vershynin 2009]
Kaczmarz Method

Given $\mathbf{x}_0 \in \mathbb{R}^n$:

1. Choose $i_k \in [m]$ with probability $\frac{||a_{ik}||^2}{||A||^2_F}$.
2. Define $\mathbf{x}_k := \mathbf{x}_{k-1} + \frac{b_{ik} - a_{ik}^T \mathbf{x}_{k-1}}{||a_{ik}||^2} a_{ik}$.
3. Repeat.

[Kaczmarz 1937], [Strohmer, Vershynin 2009]
Given $x_0 \in \mathbb{R}^n$:

1. Choose $i_k \in [m]$ as
   
   $$i_k := \text{argmax}_{i \in [m]} |a_i^T x_{k-1} - b_i|.$$ 

2. Define $x_k := x_{k-1} + \frac{b_{i_k} - a_{i_k}^T x_{k-1}}{|a_{i_k}|^2} a_{i_k}.$

3. Repeat.

[Motzkin, Schoenberg 1954]
Motzkin’s Method

Given $x_0 \in \mathbb{R}^n$:

1. Choose $i_k \in [m]$ as
   $$i_k := \text{argmax}_{i \in [m]} |a_i^T x_{k-1} - b_i|.$$

2. Define $x_k := x_{k-1} + \frac{b_{i_k} - a_{i_k}^T x_{k-1}}{|a_{i_k}|^2} a_{i_k}$.

3. Repeat.

[Motzkin, Schoenberg 1954]
Motzkin’s Method

Given $\mathbf{x}_0 \in \mathbb{R}^n$:

1. Choose $i_k \in [m]$ as 
   $$i_k := \arg\max_{i \in [m]} |\mathbf{a}_i^T \mathbf{x}_{k-1} - b_i|.$$

2. Define $\mathbf{x}_k := \mathbf{x}_{k-1} + \frac{b_{i_k} - \mathbf{a}_{i_k}^T \mathbf{x}_{k-1}}{|\mathbf{a}_{i_k}|^2} \mathbf{a}_{i_k}$.

3. Repeat.

[Motzkin, Schoenberg 1954]
Our Hybrid Method (SKM)

Given $\mathbf{x}_0 \in \mathbb{R}^n$:

1. Choose $\tau_k \subseteq [m]$ to be a sample of size $\beta$ constraints chosen uniformly at random among the rows of $A$.

2. From the $\beta$ rows, choose $i_k := \arg\max_{i \in \tau_k} |a_i^T \mathbf{x}_{k-1} - b_i|$.

3. Define $\mathbf{x}_k := \mathbf{x}_{k-1} + \frac{b_{i_k} - a_{i_k}^T \mathbf{x}_{k-1}}{||a_{i_k}||^2} a_{i_k}$.

4. Repeat.
Our Hybrid Method (SKM)

Given \( x_0 \in \mathbb{R}^n \):

1. Choose \( \tau_k \subset [m] \) to be a sample of size \( \beta \) constraints chosen uniformly at random among the rows of \( A \).

2. From the \( \beta \) rows, choose \( i_k := \arg\max_{i \in \tau_k} |a_i^T x_{k-1} - b_i| \).

3. Define \( x_k := x_{k-1} + \frac{b_{i_k} - a_{i_k}^T x_{k-1}}{||a_{i_k}||^2} a_{i_k} \).

4. Repeat.

[De Loera, H., Needell '17]
Our Hybrid Method (SKM)

Given $x_0 \in \mathbb{R}^n$:

1. Choose $\tau_k \subset [m]$ to be a sample of size $\beta$ constraints chosen uniformly at random among the rows of $A$.

2. From the $\beta$ rows, choose $i_k := \arg\max_{i \in \tau_k} |a_i^T x_{k-1} - b_i|$.

3. Define $x_k := x_{k-1} + \frac{b_{i_k} - a_{i_k}^T x_{k-1}}{||a_{i_k}||^2} a_{i_k}$.

4. Repeat.

[De Loera, H., Needell '17]
Experimental Convergence

- $\beta$: sample size
- $A$ is $50000 \times 100$ Gaussian matrix, consistent system
- ‘faster’ convergence for larger sample size
Experimental Convergence

- $\beta$: sample size
- $A$ is $50000 \times 100$ Gaussian matrix, consistent system
- ‘faster’ convergence for larger sample size
Experimental Convergence

- $\beta$: sample size
- $A$ is $50000 \times 100$ Gaussian matrix, consistent system
- ‘faster’ convergence for larger sample size
Convergence Rates

Below are the convergence rates for the methods on a system, $Ax = b$, which is consistent with unique solution $x$, whose rows have been normalized to have unit norm.

- RK (Strohmer, Vershynin ’09):

$$\mathbb{E}\|x_k - x\|_2^2 \leq \left(1 - \frac{\sigma^2_{\min}(A)}{m}\right)^k \|x_0 - x\|_2^2$$

- MM (Agmon ’54):

$$\|x_k - x\|_2^2 \leq \left(1 - \frac{\sigma^2_{\min}(A)}{m}\right)^k \|x_0 - x\|_2^2$$

- SKM (DeLoera, H., Needell ’17):

$$\mathbb{E}\|x_k - x\|_2^2 \leq \left(1 - \frac{\sigma^2_{\min}(A)}{m}\right)^k \|x_0 - x\|_2^2$$

Why are these all the same?
Convergence Rates

Below are the convergence rates for the methods on a system, $Ax = b$, which is consistent with unique solution $x$, whose rows have been normalized to have unit norm.

- RK (Strohmer, Vershynin '09):
  $$\mathbb{E}\|x_k - x\|_2^2 \leq \left(1 - \frac{\sigma_{\min}^2(A)}{m}\right)^k \|x_0 - x\|_2^2$$

- MM (Agmon '54):
  $$\|x_k - x\|_2^2 \leq \left(1 - \frac{\sigma_{\min}^2(A)}{m}\right)^k \|x_0 - x\|_2^2$$

Why are these all the same?
Convergence Rates

Below are the convergence rates for the methods on a system, $Ax = b$, which is consistent with unique solution $x$, whose rows have been normalized to have unit norm.

▶ RK (Strohmer, Vershynin ’09):

$$
\mathbb{E}\|x_k - x\|_2^2 \leq \left(1 - \frac{\sigma_{\min}(A)}{m}\right)^k \|x_0 - x\|_2^2
$$

▶ MM (Agmon ’54):

$$
\|x_k - x\|_2^2 \leq \left(1 - \frac{\sigma_{\min}(A)}{m}\right)^k \|x_0 - x\|_2^2
$$

▶ SKM (DeLoera, H., Needell ’17):

$$
\mathbb{E}\|x_k - x\|_2^2 \leq \left(1 - \frac{\sigma_{\min}(A)}{m}\right)^k \|x_0 - x\|_2^2
$$
Convergence Rates

Below are the convergence rates for the methods on a system, $A\mathbf{x} = \mathbf{b}$, which is consistent with unique solution $\mathbf{x}$, whose rows have been normalized to have unit norm.

- **RK (Strohmer, Vershynin '09):**
  $$\mathbb{E}\|\mathbf{x}_k - \mathbf{x}\|_2^2 \leq \left(1 - \frac{\sigma_{\min}(A)^2}{m}\right)^k \|\mathbf{x}_0 - \mathbf{x}\|_2^2$$

- **MM (Agmon '54):**
  $$\|\mathbf{x}_k - \mathbf{x}\|_2^2 \leq \left(1 - \frac{\sigma_{\min}(A)^2}{m}\right)^k \|\mathbf{x}_0 - \mathbf{x}\|_2^2$$

- **SKM (DeLoera, H., Needell '17):**
  $$\mathbb{E}\|\mathbf{x}_k - \mathbf{x}\|_2^2 \leq \left(1 - \frac{\sigma_{\min}(A)^2}{m}\right)^k \|\mathbf{x}_0 - \mathbf{x}\|_2^2$$

Why are these all the same?
A Pathological Example
Several works have used sparsity of the residual to improve the convergence rate of greedy methods.

[De Loera, H., Needell ’17], [Bai, Wu ’18], [Du, Gao ’19]
Several works have used sparsity of the residual to improve the convergence rate of greedy methods.

[De Loera, H., Needell ’17], [Bai, Wu ’18], [Du, Gao ’19]

However, not much sparsity can be expected in most cases. Instead, we'd like to use dynamic range of the residual to guarantee faster convergence.

$$\gamma_k := \frac{\sum_{\tau \in \binom{[m]}{\beta}} \| A_{\tau} x_k - b_\tau \|_2^2}{\sum_{\tau \in \binom{[m]}{\beta}} \| A_{\tau} x_k - b_\tau \|_\infty^2}$$
Theorem (H. - Ma 2019)

Let $A$ be normalized so $\|a_i\|_2 = 1$ for all rows $i = 1, \ldots, m$. If the system $Ax = b$ is consistent with the unique solution $x^*$ then the SKM method converges at least linearly in expectation and the rate depends on the dynamic range of the random sample of rows of $A$, $\tau_j$. Precisely, in the $j + 1$st iteration of SKM, we have

$$E_{\tau_j} \|x_{j+1} - x^*\|^2_2 \leq \left(1 - \frac{\beta \sigma_{\text{min}}^2(A)}{\gamma_j m}\right) \|x_j - x^*\|^2_2,$$

where $\gamma_j := \frac{\sum_{\tau \in \binom{[m]}{\beta}} \|A_{\tau}x_j - b_{\tau}\|^2_2}{\sum_{\tau \in \binom{[m]}{\beta}} \|A_{\tau}x_j - b_{\tau}\|^2_\infty}$. 


- $A$ is $50000 \times 100$ Gaussian matrix, consistent system
- bound uses dynamic range of sample of $\beta$ rows
What can we say about $\gamma_j$?

Recall $\gamma_j := \frac{\sum_{\tau \in (\lfloor m \rfloor)} \|A_\tau x_j - b_\tau\|_2^2}{\sum_{\tau \in (\lfloor m \rfloor)} \|A_\tau x_j - b_\tau\|_\infty^2}$.

$$1 \leq \gamma_j \leq \beta$$
What can we say about $\gamma_j$?

Recall $\gamma_j := \frac{\sum_{\tau \in \binom{[m]}{\beta}} \|A_{\tau}x_j - b_{\tau}\|_2^2}{\sum_{\tau \in \binom{[m]}{\beta}} \|A_{\tau}x_j - b_{\tau}\|_2\|_\infty^2}$.

$1 \leq \gamma_j \leq \beta$
What can we say about $\gamma_j$?

Recall $\gamma_j := \frac{\sum_{\tau \in \left[ \frac{m}{\beta} \right]} \|A_{\tau}x_j - b_{\tau}\|_2^2}{\sum_{\tau \in \left[ \frac{m}{\beta} \right]} \|A_{\tau}x_j - b_{\tau}\|_\infty^2}$.

$$1 \leq \gamma_j \leq \beta$$
What can we say about $\gamma_j$?

Recall $\gamma_j := \frac{\sum_{\tau \in \binom{[m]}{\beta}} \|A_{\tau}x_j - b_{\tau}\|_2^2}{\sum_{\tau \in \binom{[m]}{\beta}} \|A_{\tau}x_j - b_{\tau}\|_\infty^2}$.

$1 \leq \gamma_j \leq \beta$

$\mathbb{E}_{\tau_k} \|x_k - x^*\|_2^2 \leq \alpha \|x_{k-1} - x^*\|_2^2$

<table>
<thead>
<tr>
<th>Previous:</th>
</tr>
</thead>
<tbody>
<tr>
<td>RK</td>
</tr>
<tr>
<td>$\alpha = 1 - \frac{\sigma_{\min}^2(A)}{m}$</td>
</tr>
<tr>
<td>SKM</td>
</tr>
<tr>
<td>$\alpha = 1 - \frac{\sigma_{\min}^2(A)}{m}$</td>
</tr>
<tr>
<td>MM</td>
</tr>
<tr>
<td>$1 - \frac{\sigma_{\min}^2(A)}{4} \leq \alpha \leq 1 - \frac{\sigma_{\min}^2(A)}{m}$</td>
</tr>
</tbody>
</table>

[H., Needell 2019]
What can we say about $\gamma_j$?

Recall $\gamma_j := \frac{\sum_{\tau \in ([m] \setminus \beta)} \|A_{\tau} x_j - b_{\tau}\|^2_2}{\sum_{\tau \in ([m] \setminus \beta)} \|A_{\tau} x_j - b_{\tau}\|^2_\infty}$.

$1 \leq \gamma_j \leq \beta$

\[ \mathbb{E}_{\tau_k} \|x_k - x^*\|^2_2 \leq \alpha \|x_{k-1} - x^*\|^2_2 \]

<table>
<thead>
<tr>
<th></th>
<th>Previous:</th>
<th>Current:</th>
</tr>
</thead>
<tbody>
<tr>
<td>RK</td>
<td>$\alpha = 1 - \frac{\sigma^2_{\text{min}}(A)}{m}$</td>
<td>$\alpha = 1 - \frac{\sigma^2_{\text{min}}(A)}{m}$</td>
</tr>
<tr>
<td>SKM</td>
<td>$\alpha = 1 - \frac{\sigma^2_{\text{min}}(A)}{m}$</td>
<td>$1 - \frac{\beta\sigma^2_{\text{min}}(A)}{m} \leq \alpha \leq 1 - \frac{\sigma^2_{\text{min}}(A)}{m}$</td>
</tr>
<tr>
<td>MM</td>
<td>$1 - \frac{\sigma^2_{\text{min}}(A)}{4} \leq \alpha \leq 1 - \frac{\sigma^2_{\text{min}}(A)}{m}$</td>
<td>$1 - \sigma^2_{\text{min}}(A) \leq \alpha \leq 1 - \frac{\sigma^2_{\text{min}}(A)}{m}$</td>
</tr>
</tbody>
</table>

[H., Needell 2019], [H., Ma 2019]
What can we say about $\gamma_j$?

Recall $\gamma_j := \frac{\sum_{\tau \in \left[\beta \right]} \| A_{\tau} x_j - b_{\tau} \|_2^2}{\sum_{\tau \in \left[\beta \right]} \| A_{\tau} x_j - b_{\tau} \|_\infty^2}$.

$1 \leq \gamma_j \leq \beta$

▷ nontrivial bounds on $\gamma_k$ for Gaussian and average consensus systems
Generalizing the Result

- can immediately generalize to varying $\beta$ (SKM with $\beta_k$)
Generalizing the Result

- can immediately generalize to varying $\beta$ (SKM with $\beta_k$)
- to generalize to non-normalized $A$, we need a sampling distribution that depends upon $\|a_i\|^2$ and $x_k$
Generalizing the Result

- can immediately generalize to varying $\beta$ (SKM with $\beta_k$)
- to generalize to non-normalized $A$, we need a sampling distribution that depends upon $\|a_i\|^2$ and $x_k$
- provides RK result that generalizes [Strohmer, Verhsynin ’09]
Generalizing the Result

- can immediately generalize to varying $\beta$ (SKM with $\beta_k$)
- to generalize to non-normalized $A$, we need a sampling distribution that depends upon $\|a_i\|^2$ and $x_k$
- provides RK result that generalizes [Strohmer, Verhsynin ’09]
- provides simple result for MM in case of non-normalized rows (none previously)
Generalizing the Result

- can immediately generalize to varying $\beta$ (SKM with $\beta_k$)
- to generalize to non-normalized $A$, we need a sampling distribution that depends upon $\|a_i\|^2$ and $x_k$
- provides RK result that generalizes [Strohmer, Verhsynin ’09]
- provides simple result for MM in case of non-normalized rows (none previously)
- now can analyze $\gamma_k$ for systems with unequal row norms
Sketch of Proof

**Pythagorean theorem**

\[
\|x_k - x^*\|_2^2 = \|x_{k-1} - x^*\|_2^2 - \|x_{k-1} - x_k\|_2^2
\]

\[
= \|x_{k-1} - x^*\|_2^2 - \|A_{\tau_k} x_{k-1} - b_{\tau_k}\|_{\infty}^2
\]
Sketch of Proof

**Pythagorean theorem**

\[
\| \mathbf{x}_k - \mathbf{x}^* \|^2_2 = \| \mathbf{x}_{k-1} - \mathbf{x}^* \|^2_2 - \| \mathbf{x}_{k-1} - \mathbf{x}_k \|^2_2
\]

\[
= \| \mathbf{x}_{k-1} - \mathbf{x}^* \|^2_2 - \| A_{\tau_k} \mathbf{x}_{k-1} - \mathbf{b}_{\tau_k} \|^2_\infty
\]

**Expected improvement**

\[
\mathbb{E}_{\tau_k} \| A_{\tau_k} \mathbf{x}_{k-1} - \mathbf{b}_{\tau_k} \|^2_\infty = \sum_{\tau \in \binom{[m]}{\beta}} \frac{1}{\binom{m}{\beta}} \| A_{\tau} \mathbf{x}_{k-1} - \mathbf{b}_{\tau} \|^2_\infty
\]
Sketch of Proof

Pythagorean theorem

\[ \|x_k - x^*\|_2^2 = \|x_{k-1} - x^*\|_2^2 - \|x_{k-1} - x_k\|_2^2 \]

\[ = \|x_{k-1} - x^*\|_2^2 - \|A_{\tau_k}x_{k-1} - b_{\tau_k}\|_\infty^2 \]

Expected improvement

\[ \mathbb{E}_{\tau_k} \|A_{\tau_k}x_{k-1} - b_{\tau_k}\|_\infty^2 = \sum_{\tau \in \binom{[m]}{\beta}} \frac{1}{\binom{m}{\beta}} \|A_{\tau}x_{k-1} - b_{\tau}\|_\infty^2 \]

\[ = \frac{1}{\binom{m}{\beta}} \sum_{\tau \in \binom{[m]}{\beta}} \|A_{\tau}x_{k-1} - b_{\tau}\|_\infty^2 \]
Sketch of Proof

**Pythagorean theorem**

\[
\|x_k - x^*\|_2^2 = \|x_{k-1} - x^*\|_2^2 - \|x_{k-1} - x_k\|_2^2 = \|x_{k-1} - x^*\|_2^2 - \|A_{\tau_k} x_{k-1} - b_{\tau_k}\|_\infty^2
\]

**Expected improvement**

\[
\mathbb{E}_{\tau_k} \|A_{\tau_k} x_{k-1} - b_{\tau_k}\|_\infty^2 = \sum_{\tau \in (\lceil m \rceil_\beta)} \frac{1}{(m_\beta)} \|A_{\tau} x_{k-1} - b_\tau\|_\infty^2 = \frac{1}{(m_\beta)} \sum_{\tau \in (\lceil m \rceil_\beta)} \|A_{\tau} x_{k-1} - b_\tau\|_\infty^2
\]

\[
= \frac{1}{(m_\beta)} \gamma_k \sum_{\tau \in (\lceil m \rceil_\beta)} \|A_{\tau} x_{k-1} - b_\tau\|_2^2
\]
Sketch of Proof

Pythagorean theorem

$$\|x_k - x^*\|_2^2 = \|x_{k-1} - x^*\|_2^2 - \|x_{k-1} - x_k\|_2^2$$

$$= \|x_{k-1} - x^*\|_2^2 - \|A_{\tau_k} x_{k-1} - b_{\tau_k}\|_\infty^2$$

Expected improvement

$$E_{\tau_k} \|A_{\tau_k} x_{k-1} - b_{\tau_k}\|_\infty^2 = \sum_{\tau \in \binom{m}{\beta}} \frac{1}{\binom{m}{\beta}} \|A_{\tau} x_{k-1} - b_{\tau}\|_\infty^2$$

$$= \frac{1}{\binom{m}{\beta}} \sum_{\tau \in \binom{m}{\beta}} \|A_{\tau} x_{k-1} - b_{\tau}\|_\infty^2$$

$$= \frac{1}{\binom{m}{\beta} \gamma_k} \sum_{\tau \in \binom{m}{\beta}} \|A_{\tau} x_{k-1} - b_{\tau}\|_2^2$$

$$= \frac{\beta}{\gamma_k m} \|Ax_{k-1} - b\|_2^2 \geq \frac{\beta \sigma_{\min}(A)}{\gamma_k m} \|x_{k-1} - x^*\|_2^2$$
Now can we determine the optimal $\beta$?
Now can we determine the optimal $\beta$?

Roughly, if we know the value of $\gamma_j$, we can (just) do it.
Now can we determine the optimal $\beta$?

Roughly, if we know the value of $\gamma_j$, we can (just) do it.
Back to Hierarchical NMF

\[ b \approx A \]
Back to Hierarchical NMF

\[ X \approx \begin{array}{c} m \\ \hline \end{array} \times \begin{array}{c} k \\ \hline \end{array} \begin{array}{c} A \\ \hline \end{array} = \begin{array}{c} m \\ \hline \end{array} \times \begin{array}{c} k \\ \hline \end{array} \begin{array}{c} S \\ \hline \end{array} \]
Back to Hierarchical NMF

$\triangleleft$ hNMF (sequential NMF)

$\triangleleft$ Deep NMF [Flenner, Hunter '18]

$\triangleleft$ Neural NMF
Back to Hierarchical NMF

Compare:

▷ hNMF (sequential NMF)
Back to Hierarchical NMF

Compare:

▷ hNMF (sequential NMF)
▷ Deep NMF [Flenner, Hunter ’18]
Back to Hierarchical NMF

Compare:
- hNMF (sequential NMF)
- Deep NMF [Flenner, Hunter ’18]
- Neural NMF
Applications
Experimental results: synthetic data
Experimental results: synthetic data

- unsupervised reconstruction with two-layer structure
  \( (k^{(0)} = 9, k^{(1)} = 4) \)
Experimental results: synthetic data

- unsupervised reconstruction with two-layer structure
  \((k^{(0)} = 9, k^{(1)} = 4)\)
Experimental results: MyLymeData

- Fatigue
- Facial nerve (Bell's) palsy
- Bulls-eye rash
- Other Symptoms
- Evidence of tick bite
- Red skin rash
- Early Other Symptoms
- Shooting pains that interfere with sleep
- Lightheadedness
- Large joint pain
- None of the above symptoms
- Early Flu-like symptoms
- Fainting, shortness of breath
- Headache
- Joint pain
- Muscle aches
- Severe headaches/neck stiffness
- Flu-like symptoms
- Nerve pain
- Psychiatric
- Heart-related symptoms
- Memory loss
- Twitching
- Sleep impairment
- Cognitive impairment

- Nerve pain
- Psychiatric
- Muscle aches
- Heart-related symptoms
- Headache
- Joint pain
- Flu-like symptoms
- Fatigue
- Bulls-eye rash
- Memory loss
- Twitching
- Sleep impairment
- Cognitive impairment
- Red skin rash
- Facial nerve (Bell's) palsy
- Severe headaches/neck stiffness
- Shooting pains that interfere with sleep
- Lightheadedness
- Other Symptoms
- Large joint pain
- Fainting, shortness of breath
- Early Flu-like symptoms
- Evidence of tick bite
- Early Other Symptoms
- None of the above symptoms
Experimental results: MyLymeData
Experimental results: MyLymeData

\[ k^{(0)} = 6 \] \[ k^{(1)} = 5 \] \[ k^{(2)} = 4 \]
Experimental results: MyLymeData

Unwell Patients

Well Patients

- Bull's eye rash
- Joint pain
- Fatigue
- Flu-like symptoms
- Headache
- Evidence of tick bite
- Other
- None of the above
- Heart-related symptoms
- Psychiatric
- Nerve pain
- Muscle aches
- Sleep impairment
- Memory loss
- Twitching
- Cognitive impairment

- Bull's eye rash
- Evidence of tick bite
- Other
- Fatigue
- Flu-like symptoms
- Twitching
- Sleep impairment
- Memory loss
- Nerve pain
- Muscle aches
- Heart-related symptoms
- Psychiatric
- Cognitive impairment
- Headache
- Joint pain
- None of the above
MyLymeData Takeaways

- bulls-eye rash (diagnosing symptoms) topic does not seem to persist for smaller number of topics
bulls-eye rash (diagnosing symptoms) topic does not seem to persist for smaller number of topics

unwell and well patients have very different presentation of bulls-eye rash symptom in topics
MyLymeData Takeaways

- bulls-eye rash (diagnosing symptoms) topic does not seem to persist for smaller number of topics

- unwell and well patients have very different presentation of bulls-eye rash symptom in topics

- patients unwell because lacking bulls-eye rash for diagnosis or indicative of different disease pathway?
Conclusions
Conclusions

- hNMF model can be implemented as a feed-forward neural network

![Diagram showing neural network structure](image)

- presented our method Neural NMF
- described family of algorithms which can solve fundamental least-squares subroutine
- presented accelerated convergence analysis for SKM

- applied Neural NMF to synthetic data and MyLymeData
Thanks for listening!

Questions?


Experimental results: synthetic data

Semisupervised reconstruction (40% labels) with three-layer structure ($k^{(0)} = 9$, $k^{(1)} = 4$, $k^{(2)} = 2$)
Experimental results: synthetic data

- semisupervised reconstruction (40% labels) with three-layer structure \( (k^{(0)} = 9, k^{(1)} = 4, k^{(2)} = 2) \)
## Experimental results: synthetic data

### Table 1: Reconstruction error / classification accuracy

<table>
<thead>
<tr>
<th>Layers</th>
<th>Hier. NMF</th>
<th>Deep NMF</th>
<th>Neural NMF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsuper.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.053</td>
<td>0.031</td>
<td>0.029</td>
</tr>
<tr>
<td>2</td>
<td>0.399</td>
<td>0.414</td>
<td>0.310</td>
</tr>
<tr>
<td>3</td>
<td>0.860</td>
<td>0.838</td>
<td>0.492</td>
</tr>
<tr>
<td>Semisuper.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.049 / 0.933</td>
<td>0.031 / 0.947</td>
<td>0.042 / 1</td>
</tr>
<tr>
<td>2</td>
<td>0.374 / 0.926</td>
<td>0.394 / 0.911</td>
<td>0.305 / 1</td>
</tr>
<tr>
<td>3</td>
<td>0.676 / 0.930</td>
<td>0.733 / 0.930</td>
<td>0.496 / 0.990</td>
</tr>
<tr>
<td>Supervised</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.052 / 0.960</td>
<td>0.042 / 0.962</td>
<td>0.042 / 1</td>
</tr>
<tr>
<td>2</td>
<td>0.311 / 0.984</td>
<td>0.310 / 0.984</td>
<td>0.307 / 1</td>
</tr>
<tr>
<td>3</td>
<td>0.495 / 1</td>
<td>0.494 / 1</td>
<td>0.498 / 1</td>
</tr>
</tbody>
</table>
Experimental results: 20 Newsgroups data
Experimental Convergence

$\beta$: sample size

$A$ is $50000 \times 100$ Gaussian matrix, consistent system

‘faster’ convergence for larger sample size
Experimental Convergence

- $\beta$: sample size
- $A$ is $50000 \times 100$ Gaussian matrix, consistent system
- ‘faster’ convergence for larger sample size
Experimental Convergence

- $\beta$: sample size
- $A$ is 50000 $\times$ 100 Gaussian matrix, consistent system
- ‘faster’ convergence for larger sample size
**Goal:** Exploit similarities between neural networks and hierarchical NMF.
**Goal:** Exploit similarities between neural networks and hierarchical NMF.

- [Flenner, Hunter ’18]
  - introduces nonlinear pooling operator after each layer
  - introduces multiplicative updates method meant to backpropagate
Goal: Exploit similarities between neural networks and hierarchical NMF.

- [Flenner, Hunter '18]
  - introduces nonlinear pooling operator after each layer
  - introduces multiplicative updates method meant to backpropagate

- [Trigeorgis, Bousmalis, Zafeiriou, Schuller '16]
  - relaxes some of nonnegativity constraints in hNMF
Deep NMF

**Goal:** Exploit similarities between neural networks and hierarchical NMF.

- [Flenner, Hunter ’18]
  - introduces nonlinear pooling operator after each layer
  - introduces multiplicative updates method meant to backpropagate

- [Trigeorgis, Bousmalis, Zafeiriou, Schuller ’16]
  - relaxes some of nonnegativity constraints in hNMF

- [Le Roux, Hershey, Weninger ’15]
  - introduces NMF backpropagation algorithm with “unfolding” (no hierarchy)
Deep NMF

**Goal:** Exploit similarities between neural networks and hierarchical NMF.

- [Flenner, Hunter '18]
  - introduces nonlinear pooling operator after each layer
  - introduces multiplicative updates method meant to backpropagate
- [Trigeorgis, Bousmalis, Zafeiriou, Schuller '16]
  - relaxes some of nonnegativity constraints in hNMF
- [Le Roux, Hershey, Weninger '15]
  - introduces NMF backpropagation algorithm with “unfolding” (no hierarchy)
- [Sun, Nasrabadi, Tran '17]
  - similar method lacking nonnegativity constraints
Block Kaczmarz
Bound on $\gamma_j$

$$\gamma_k \geq \frac{\beta}{m} \sigma^2_{\min}(A) \text{ when } A \text{ is row-normalized}$$