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

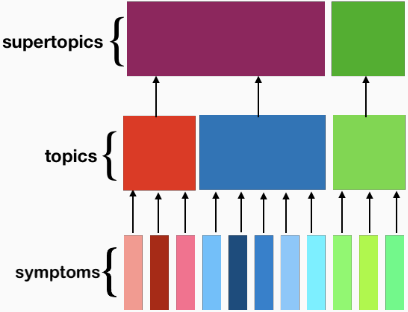
Introduction

Motivation

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

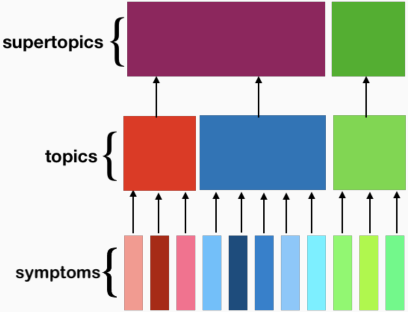
Motivation

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



Motivation

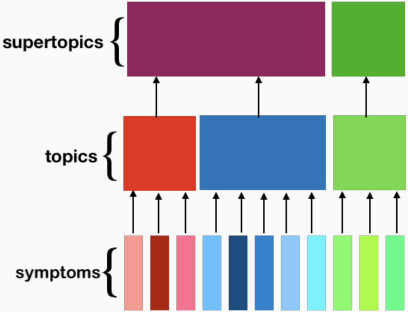
- MyLymeData: large collection of Lyme disease patient survey data collected by LymeDisease.org (∼12,000 patients, 100s of questions)



⇒ hypothesis formation about post-treatment/chronic Lyme disease

Motivation

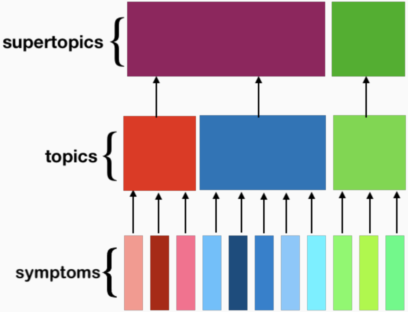
- MyLymeData: large collection of Lyme disease patient survey data collected by LymeDisease.org (~12,000 patients, 100s of questions)



$\Rightarrow$ 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)

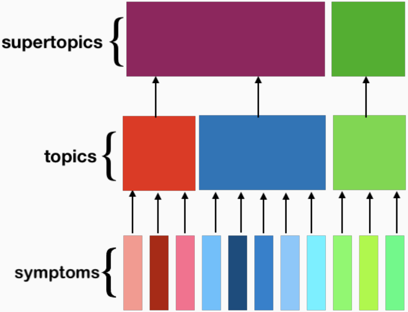


$\Rightarrow$ hypothesis formation about post-treatment/chronic Lyme disease

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

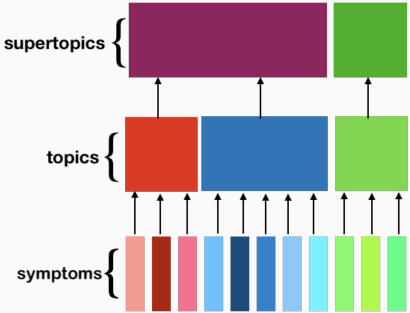
Motivation

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



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]

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]

Motivation

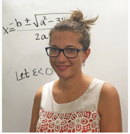
Main question: How can we identify the topic hierarchy of MyLymeData symptom questions?

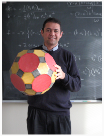
Answer: `Neural Nonnegative Matrix Factorization`

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

`Sampling Kaczmarz-Motzkin Methods`

[H., Ma '19], [De Loera, H., Needell '17]



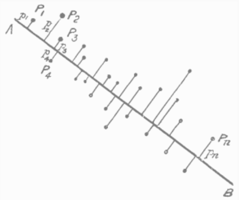


Topic Modeling

- principal component analysis (PCA)

  [Pearson 1901]

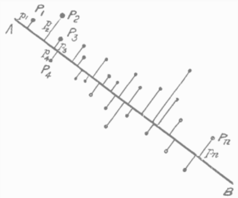
  [Hotelling 1933]



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

Topic Modeling

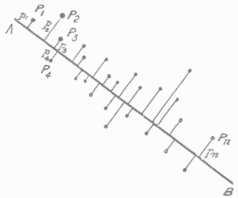
- principal component analysis (PCA)
  - [Pearson 1901]
  - [Hotelling 1933]
- latent dirichlet allocation (LDA)
  - [Pritchard, Stephens, Donnelly 2000]
  - [Blei, Ng, Jordan 2003]



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

Topic Modeling

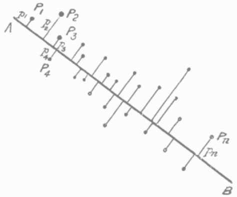
- ▷ principal component analysis (PCA)
  - [Pearson 1901]
  - [Hotelling 1933]
- ▷ latent dirichlet allocation (LDA)
  - [Pritchard, Stephens, Donnelly 2000]
  - [Blei, Ng, Jordan 2003]
- ▷ clustering (*k*-means, Gaussian mixtures)
  - [Lloyd 1957]
  - [Pearson 1894]



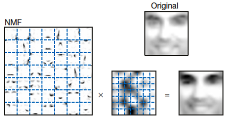
Pearson, K. (1901) *On lines and planes of closest fit to systems of points in space.*

Topic Modeling

- principal component analysis (PCA)
  - [Pearson 1901]
  - [Hotelling 1933]
- latent dirichlet allocation (LDA)
  - [Pritchard, Stephens, Donnelly 2000]
  - [Blei, Ng, Jordan 2003]
- clustering (k-means, Gaussian mixtures)
  - [Lloyd 1957]
  - [Pearson 1894]
- nonnegative matrix factorization (NMF)
  - [Paatero, Tapper 1994]
  - [Lee, Seung 1999]



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

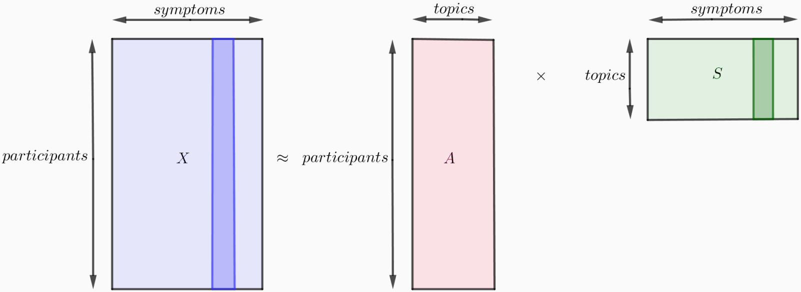


Lee, D., Seung, S. (1999) *Learning the parts of objects by non-negative matrix factorization.*

Nonnegative Matrix Factorization (NMF)

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

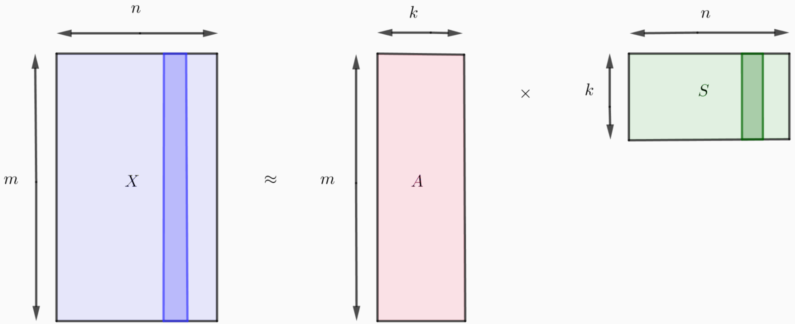
$$X \approx AS.$$



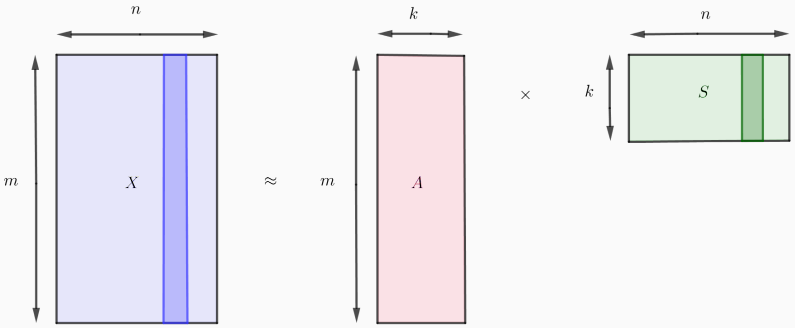
Nonnegative Matrix Factorization (NMF)

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

$$X \approx AS.$$



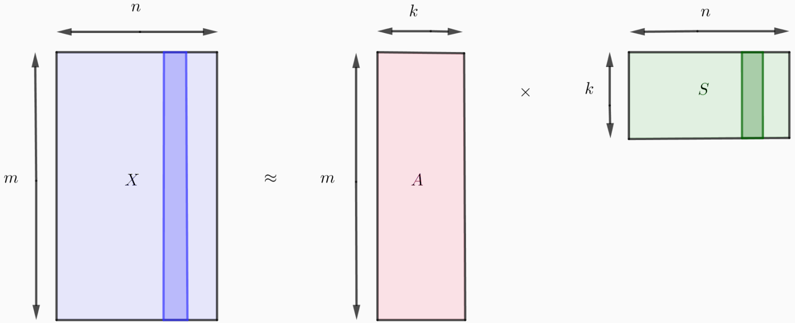
Nonnegative Matrix Factorization (NMF)



- ▷ Often formulated as optimization problem

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

Nonnegative Matrix Factorization (NMF)



- Often formulated as optimization problem

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

- Non-convex optimization problem, NP-hard to compute global optimum for fixed k [Vavasis 2008]

Hierarchical NMF

Model: Sequentially factorize

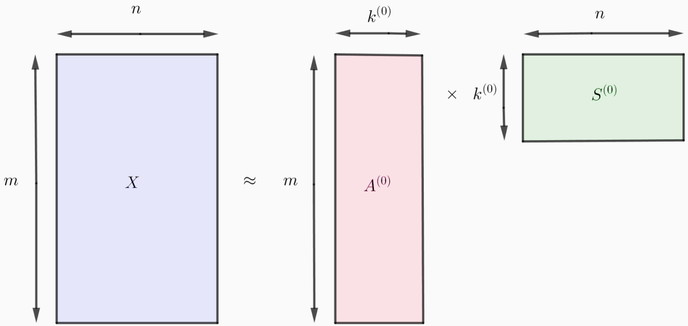
$$X \approx A^{(0)}S^{(0)}, S^{(0)} \approx A^{(1)}S^{(1)}, S^{(1)} \approx A^{(2)}S^{(2)}, ..., S^{(\mathcal{L}-1)} \approx A^{(\mathcal{L})}S^{(\mathcal{L})}.$$

[Cichocki, Zdunek '06]

Hierarchical NMF

Model: Sequentially factorize

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

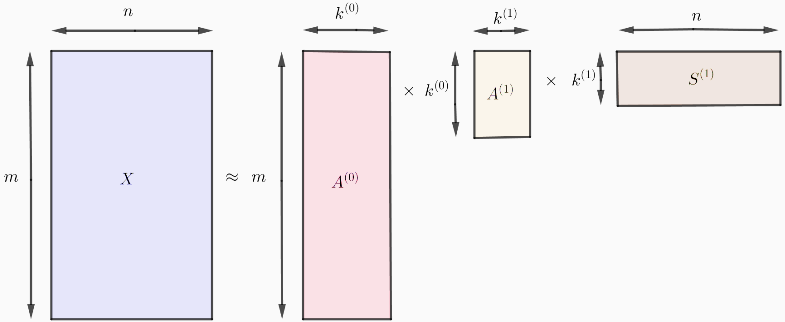


[Cichocki, Zdunek '06]

Hierarchical NMF

Model: Sequentially factorize

$$X \approx A^{(0)}S^{(0)}, S^{(0)} \approx A^{(1)}S^{(1)}, S^{(1)} \approx A^{(2)}S^{(2)}, ..., S^{(\mathcal{L}-1)} \approx A^{(\mathcal{L})}S^{(\mathcal{L})}.$$

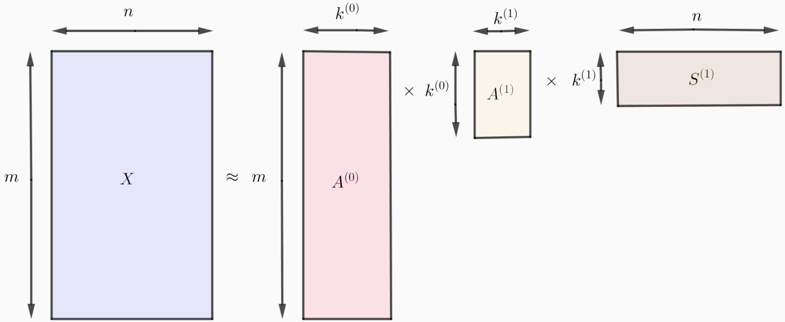


[Cichocki, Zdunek '06]

Hierarchical NMF

Model: Sequentially factorize

$$X \approx A^{(0)}S^{(0)}, S^{(0)} \approx A^{(1)}S^{(1)}, S^{(1)} \approx A^{(2)}S^{(2)}, ..., S^{(\mathcal{L}-1)} \approx A^{(\mathcal{L})}S^{(\mathcal{L})}.$$



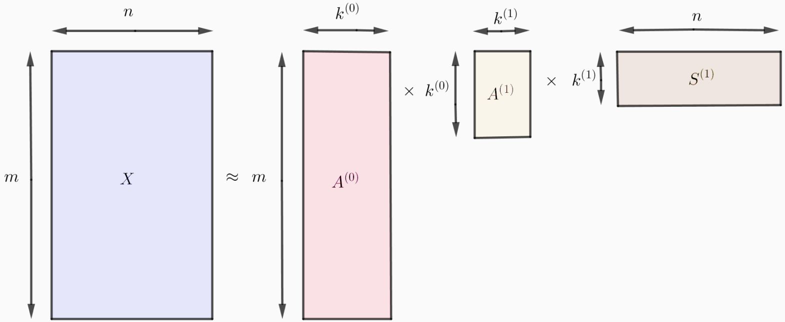
- $k^{(\ell)}$: supertopics collecting $k^{(\ell-1)}$ subtopics

[Cichocki, Zdunek '06]

Hierarchical NMF

Model: Sequentially factorize

$$X \approx A^{(0)}S^{(0)}, S^{(0)} \approx A^{(1)}S^{(1)}, S^{(1)} \approx A^{(2)}S^{(2)}, ..., S^{(\mathcal{L}-1)} \approx A^{(\mathcal{L})}S^{(\mathcal{L})}.$$

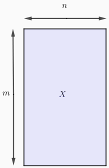


- $k^{(\ell)}$: supertopics collecting $k^{(\ell-1)}$ subtopics
- error propagates through layers

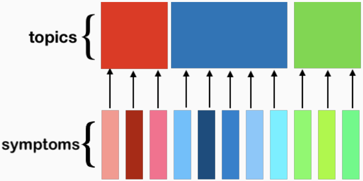
[Cichocki, Zdunek '06]

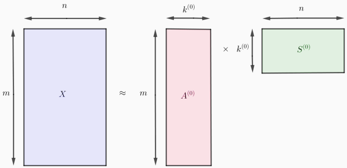
Neural NMF

Hierarchical NMF

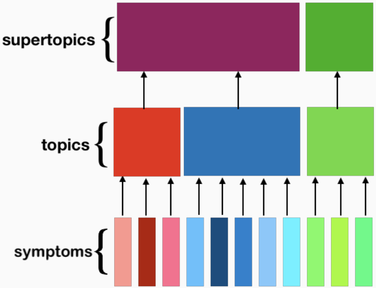


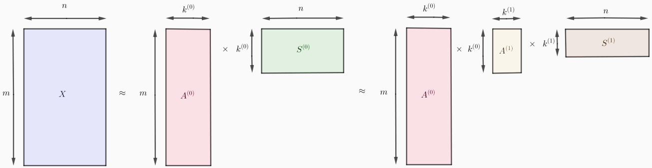
Hierarchical NMF



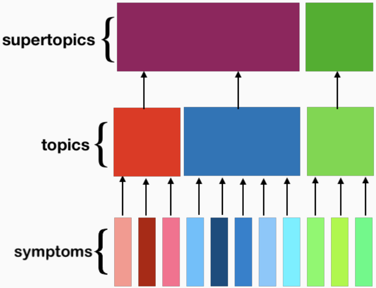


Hierarchical NMF

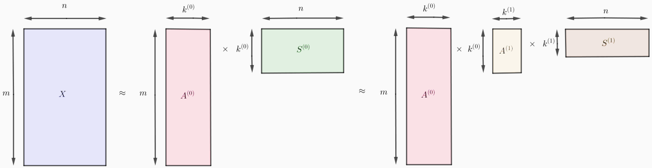




Hierarchical NMF

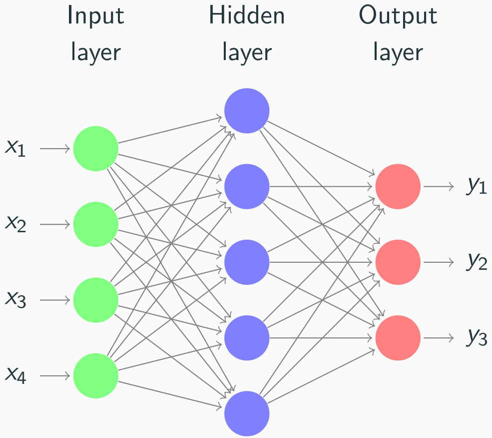


- ▷ hNMF can be implemented in a `feed-forward neural network` structure



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

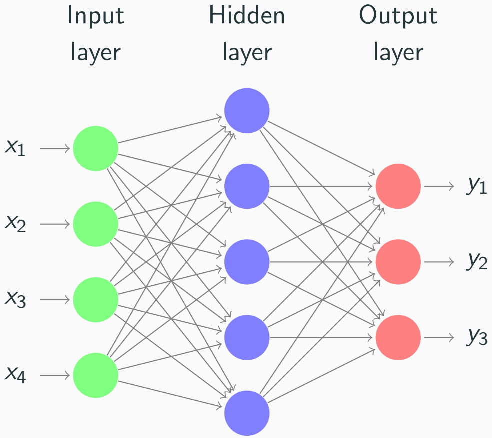
$$\sum_{n=1}^{N} E(\{W_i\}) = f(\mathbf{y}(\mathbf{x}_n, \{W_i\}), \mathbf{x}_n, \mathbf{t}_n).$$



Feed-forward Neural Networks

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

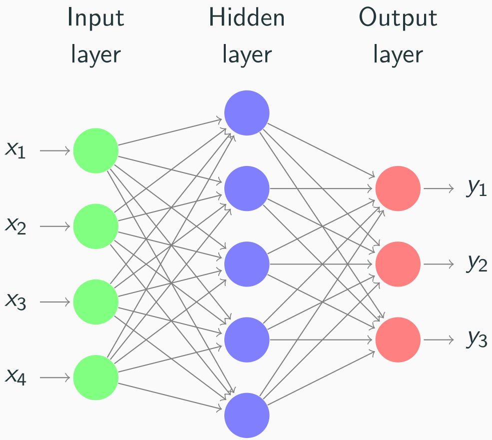
$$E(\{W_i\}) = \sum_{n=1}^{N} \|\mathbf{y}(\mathbf{x}_n, \{W_i\}) - \mathbf{t}_n\|_2^2.$$



Feed-forward Neural Networks

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

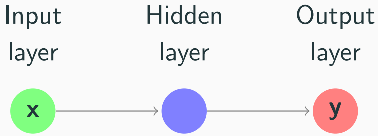
$$E(\{W_i\}) = \sum_{n=1}^{N} f(\mathbf{y}(\mathbf{x}_n, \{W_i\}), \mathbf{x}_n, \mathbf{t}_n).$$



Feed-forward Neural Networks

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

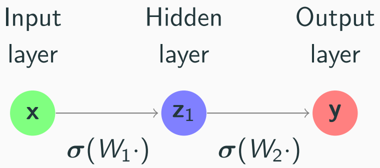
$$E(\{W_i\}) = \sum_{n=1}^{N} f(\mathbf{y}(\mathbf{x}_n, \{W_i\}), \mathbf{x}_n, \mathbf{t}_n).$$



Feed-forward Neural Networks

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

$$E(\{W_i\}) = \sum_{n=1}^{N} f(\mathbf{y}(\mathbf{x}_n, \{W_i\}), \mathbf{x}_n, \mathbf{t}_n).$$



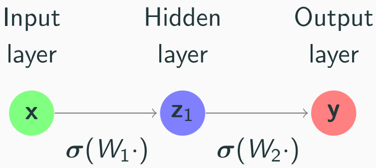
Training:

- forward propagation:
  $\mathbf{z}_1 = \boldsymbol{\sigma}(W_1 \mathbf{x})$,
  $\mathbf{z}_2 = \boldsymbol{\sigma}(W_2 \mathbf{z}_1)$, ...,
  $\mathbf{y} = \boldsymbol{\sigma}(W_L \mathbf{z}_{L-1})$

Feed-forward Neural Networks

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

$$E(\{W_i\}) = \sum_{n=1}^{N} f(\mathbf{y}(\mathbf{x}_n, \{W_i\}), \mathbf{x}_n, \mathbf{t}_n).$$

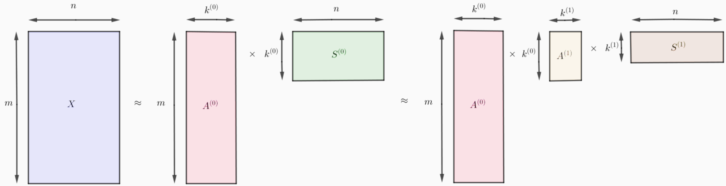


Training:

- forward propagation: $\mathbf{z}_1 = \boldsymbol{\sigma}(W_1\mathbf{x})$, $\mathbf{z}_2 = \boldsymbol{\sigma}(W_2\mathbf{z}_1)$, ..., $\mathbf{y} = \boldsymbol{\sigma}(W_L\mathbf{z}_{L-1})$
- back propagation: update $\{W_i\}$ with $\nabla E(\{W_i\})$

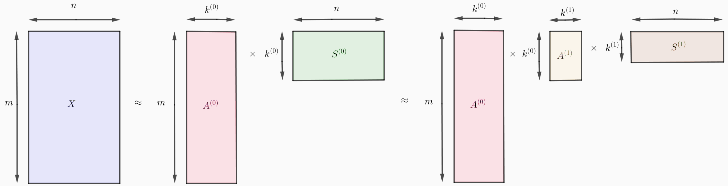
Our method: Neural NMF

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



Our method: Neural NMF

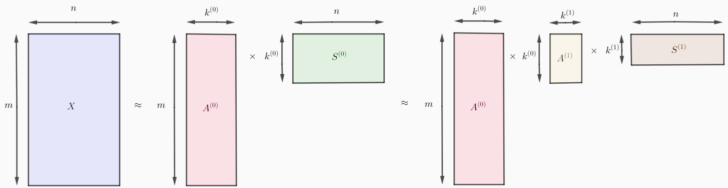
Goal: Develop true forward and back propagation algorithms for hNMF.



▷ Regard the A matrices as independent variables, determine the S matrices from the A matrices.

Our method: Neural NMF

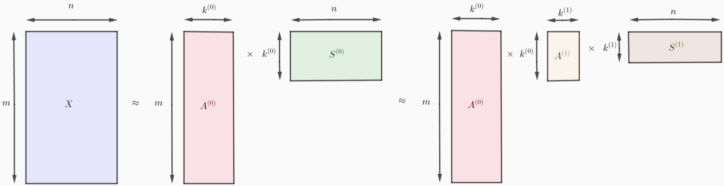
Goal: Develop true forward and back propagation algorithms for hNMF.



- ▷ Regard the A matrices as independent variables, determine the S matrices from the A matrices.
- ▷ Define $q(X, A) := \text{argmin}_{S \geq 0} \|X - AS\|_F^2$.

Our method: Neural NMF

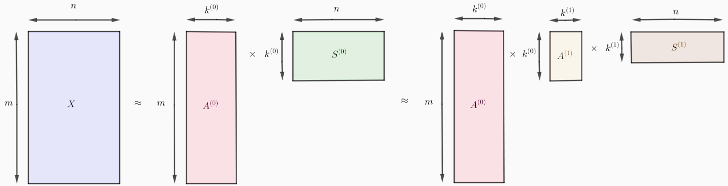
Goal: Develop true forward and back propagation algorithms for hNMF.



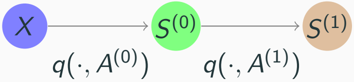
- ▷ Regard the A matrices as independent variables, determine the S matrices from the A matrices.
- ▷ Define $q(X, A) := \text{argmin}_{S \geq 0} \|X - AS\|_F^2$.
- ▷ Pin the values of S to those of A by recursively setting $S^{(\ell)} := q(S^{(\ell-1)}, A^{(\ell)})$.

Our method: Neural NMF

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

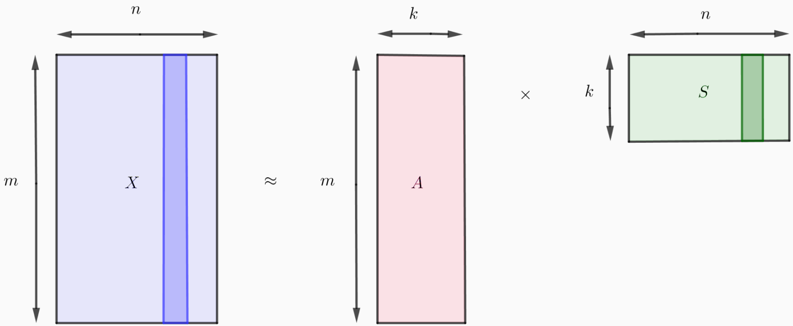


- ▷ Regard the A matrices as independent variables, determine the S matrices from the A matrices.
- ▷ Define $q(X, A) := \text{argmin}_{S \geq 0} \|X - AS\|_F^2$.
- ▷ Pin the values of S to those of A by recursively setting $S^{(\ell)} := q(S^{(\ell-1)}, A^{(\ell)})$.



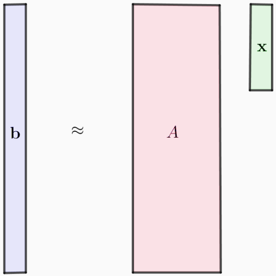
Least-squares Subroutine

- ▷ least-squares is a fundamental subroutine in forward-propagation



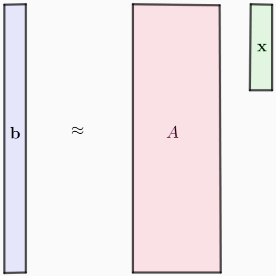
Least-squares Subroutine

- least-squares is a fundamental subroutine in forward-propagation



Least-squares Subroutine

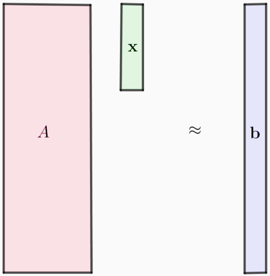
- least-squares is a fundamental subroutine in forward-propagation



- iterative projection methods can solve these problems

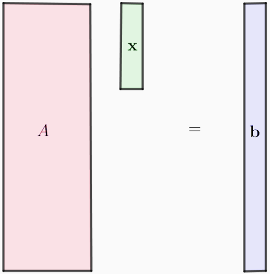
Iterative Projection Methods

General Setup



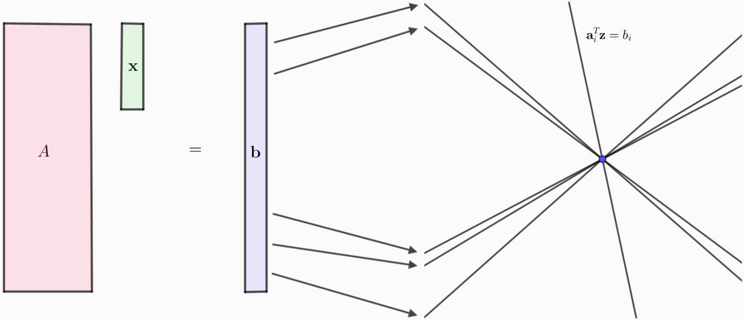
General Setup

We are interested in solving highly overdetermined systems of equations, $A\mathbf{x} = \mathbf{b}$, where $A \in \mathbb{R}^{m \times n}$, $\mathbf{b} \in \mathbb{R}^m$ and $m \gg n$. Rows are denoted $\mathbf{a}_i^T$.



General Setup

We are interested in solving highly overdetermined systems of equations, $A\mathbf{x} = \mathbf{b}$, where $A \in \mathbb{R}^{m \times n}$, $\mathbf{b} \in \mathbb{R}^m$ and $m \gg n$. Rows are denoted $\mathbf{a}_i^T$.



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)

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:

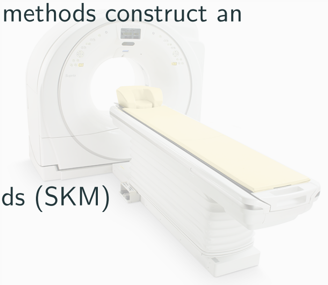
1. Randomized Kaczmarz Method
2. Motzkin's Method

Applications:

1. Tomography (Algebraic Reconstruction Technique)
2. Linear programming

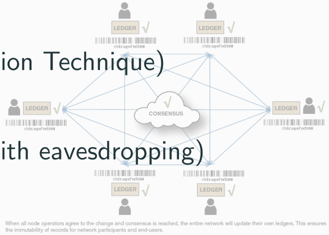
Iterative Projection Methods

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

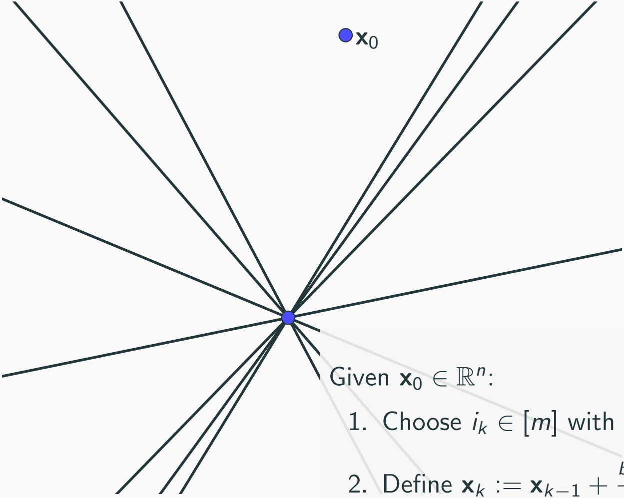


1. Randomized Kaczmarz Method
2. Motzkin's Method
3. Sampling Kaczmarz-Motzkin Methods (SKM)

Applications:



1. Tomography (Algebraic Reconstruction Technique)
2. Linear programming
3. Average consensus (greedy gossip with eavesdropping)

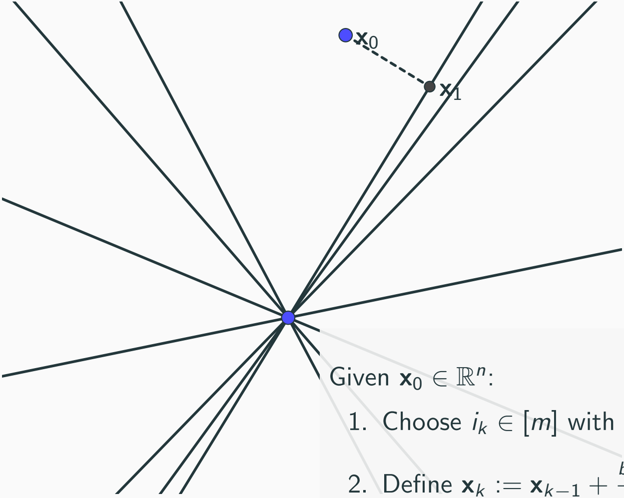


Given $\mathbf{x}_0 \in \mathbb{R}^n$:

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

[Kaczmarz 1937], [Strohmer, Vershynin 2009]

Kaczmarz Method

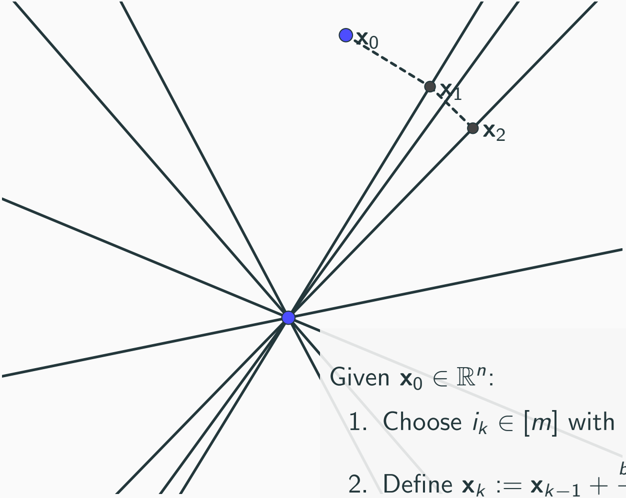


Given $\mathbf{x}_0 \in \mathbb{R}^n$:

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

[Kaczmarz 1937], [Strohmer, Vershynin 2009]

Kaczmarz Method

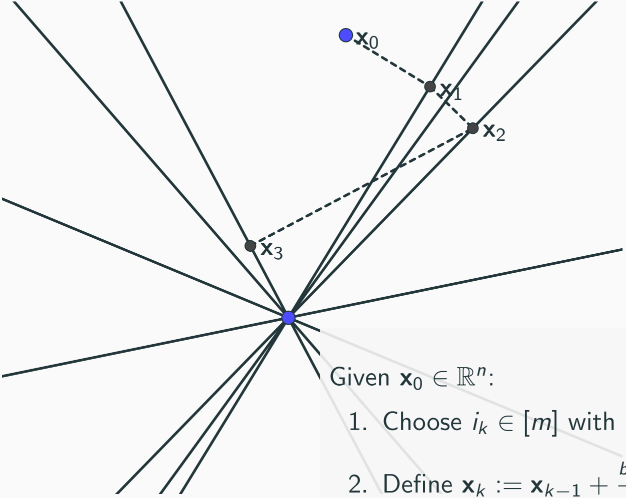


Given $\mathbf{x}_0 \in \mathbb{R}^n$:

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

[Kaczmarz 1937], [Strohmer, Vershynin 2009]

Kaczmarz Method

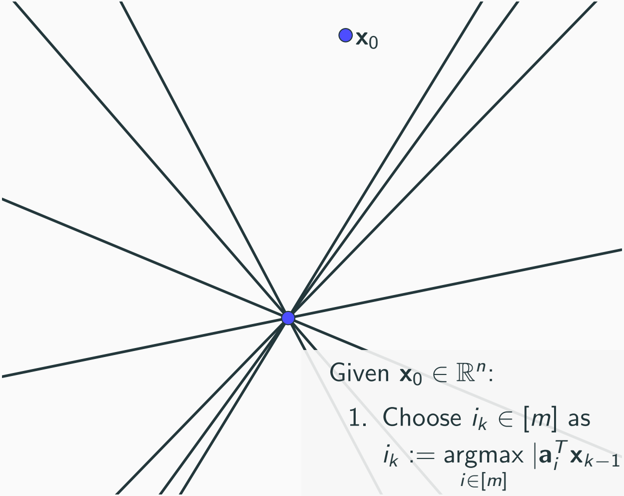


Given $\mathbf{x}_0 \in \mathbb{R}^n$:

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

[Kaczmarz 1937], [Strohmer, Vershynin 2009]

Motzkin's Method

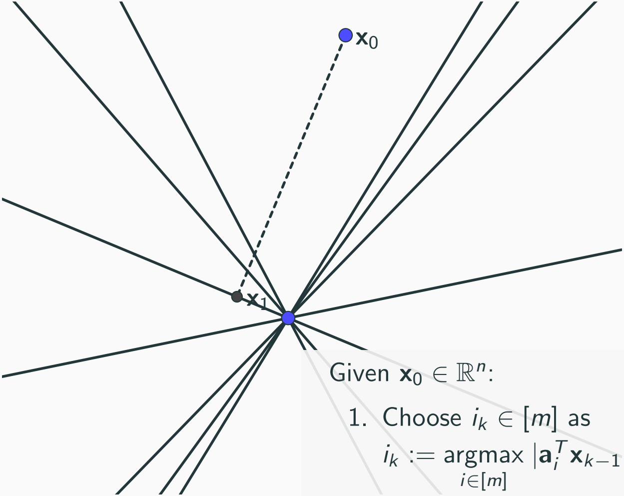


Given $\mathbf{x}_0 \in \mathbb{R}^n$:

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

Motzkin's Method

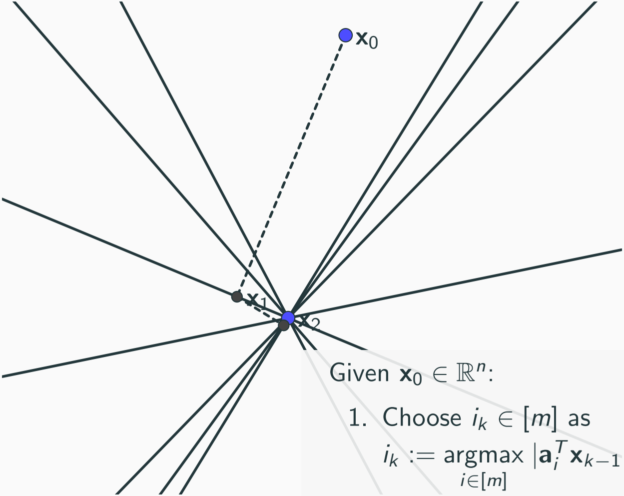


Given $\mathbf{x}_0 \in \mathbb{R}^n$:

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

Motzkin's Method

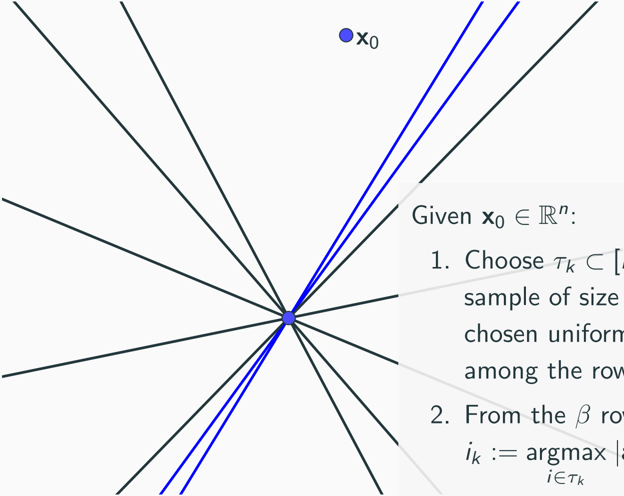


Given $\mathbf{x}_0 \in \mathbb{R}^n$:

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

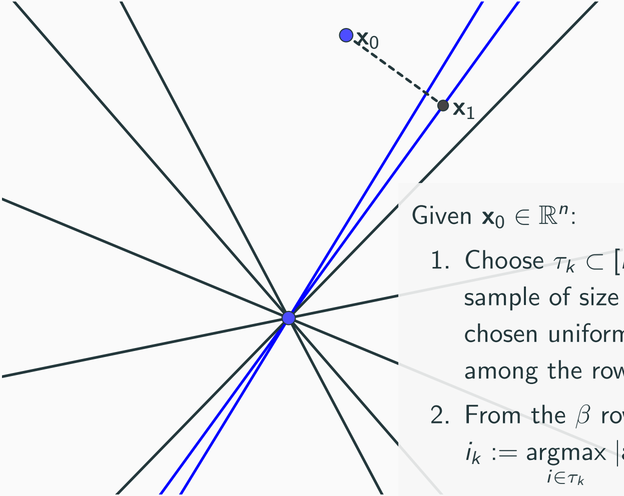


Given $\mathbf{x}_0 \in \mathbb{R}^n$:

1. Choose $\tau_k \subset [m]$ to be a sample of size β constraints chosen uniformly at random among the rows of A.
2. From the β rows, choose $i_k := \underset{i \in \tau_k}{\operatorname{argmax}} |\mathbf{a}_i^T \mathbf{x}_{k-1} - b_i|$.
3. Define $\mathbf{x}_k := \mathbf{x}_{k-1} + \frac{b_{i_k} - \mathbf{a}_{i_k}^T \mathbf{x}_{k-1}}{||\mathbf{a}_{i_k}||^2} \mathbf{a}_{i_k}$.
4. Repeat.

[De Loera, H., Needell '17]

Our Hybrid Method (SKM)

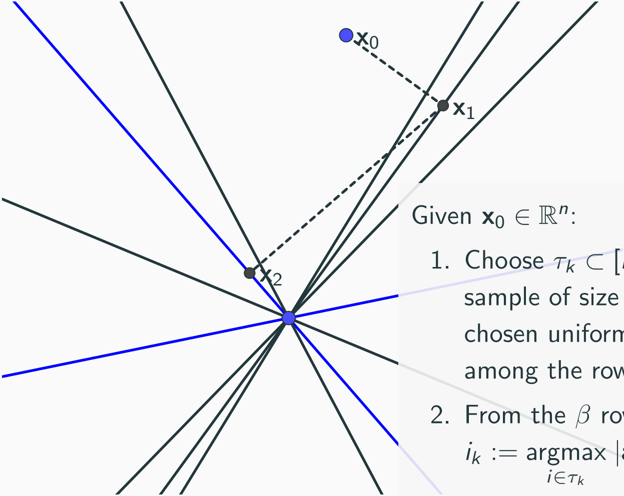


Given $\mathbf{x}_0 \in \mathbb{R}^n$:

1. Choose $\tau_k \subset [m]$ to be a sample of size β constraints chosen uniformly at random among the rows of A.
2. From the β rows, choose $i_k := \underset{i \in \tau_k}{\operatorname{argmax}} |\mathbf{a}_i^T \mathbf{x}_{k-1} - b_i|$.
3. Define $\mathbf{x}_k := \mathbf{x}_{k-1} + \frac{b_{i_k} - \mathbf{a}_{i_k}^T \mathbf{x}_{k-1}}{||\mathbf{a}_{i_k}||^2} \mathbf{a}_{i_k}$.
4. Repeat.

[De Loera, H., Needell '17]

Our Hybrid Method (SKM)

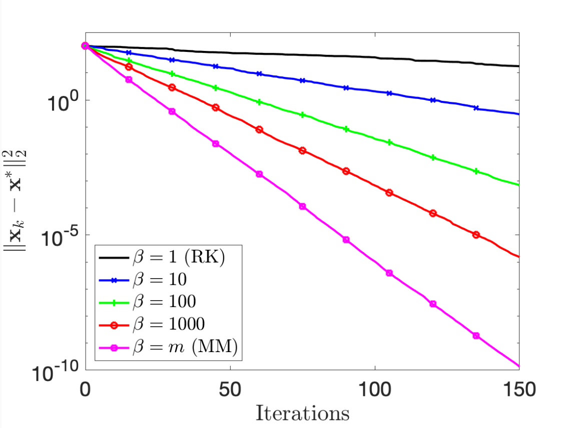


Given $\mathbf{x}_0 \in \mathbb{R}^n$:

1. Choose $\tau_k \subset [m]$ to be a sample of size β constraints chosen uniformly at random among the rows of A.
2. From the β rows, choose $i_k := \underset{i \in \tau_k}{\operatorname{argmax}} |\mathbf{a}_i^T \mathbf{x}_{k-1} - b_i|$.
3. Define $\mathbf{x}_k := \mathbf{x}_{k-1} + \frac{b_{i_k} - \mathbf{a}_{i_k}^T \mathbf{x}_{k-1}}{||\mathbf{a}_{i_k}||^2} \mathbf{a}_{i_k}$.
4. Repeat.

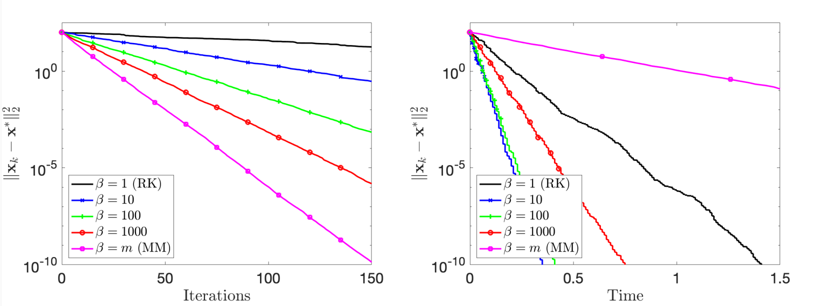
[De Loera, H., Needell '17]

Experimental Convergence



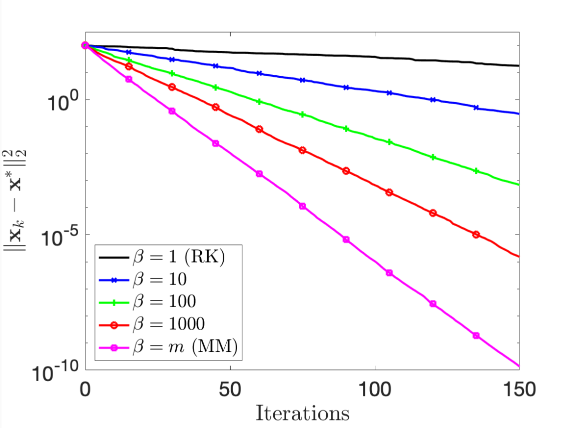
- β: sample size
- A is 50000×100 Gaussian matrix, consistent system
- 'faster' convergence for larger sample size

Experimental Convergence



- β: sample size
- A is 50000×100 Gaussian matrix, consistent system
- 'faster' convergence for larger sample size

Experimental Convergence



- β: sample size
- A is 50000×100 Gaussian matrix, consistent system
- 'faster' convergence for larger sample size

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):

$$\mathbb{E}||\mathbf{x}_k - \mathbf{x}||_2^2 \leq \left(1 - \frac{\sigma_{\min}^2(A)}{m}\right)^k ||\mathbf{x}_0 - \mathbf{x}||_2^2$$

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):

$$\mathbb{E}||\mathbf{x}_k - \mathbf{x}||_2^2 \leq \left(1 - \frac{\sigma_{\min}^2(A)}{m}\right)^k ||\mathbf{x}_0 - \mathbf{x}||_2^2$$

- MM (Agmon '54):

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

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):

$$\mathbb{E}||\mathbf{x}_k - \mathbf{x}||_2^2 \leq \left(1 - \frac{\sigma_{\min}^2(A)}{m}\right)^k ||\mathbf{x}_0 - \mathbf{x}||_2^2$$

- MM (Agmon '54):

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

- SKM (DeLoera, H., Needell '17):

$$\mathbb{E}\|\mathbf{x}_k - \mathbf{x}\|_2^2 \leq \left(1 - \frac{\sigma_{\min}^2(A)}{m}\right)^k \|\mathbf{x}_0 - \mathbf{x}\|_2^2$$

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):

$$\mathbb{E}||\mathbf{x}_k - \mathbf{x}||_2^2 \leq \left(1 - \frac{\sigma_{\min}^2(A)}{m}\right)^k ||\mathbf{x}_0 - \mathbf{x}||_2^2$$

- MM (Agmon '54):

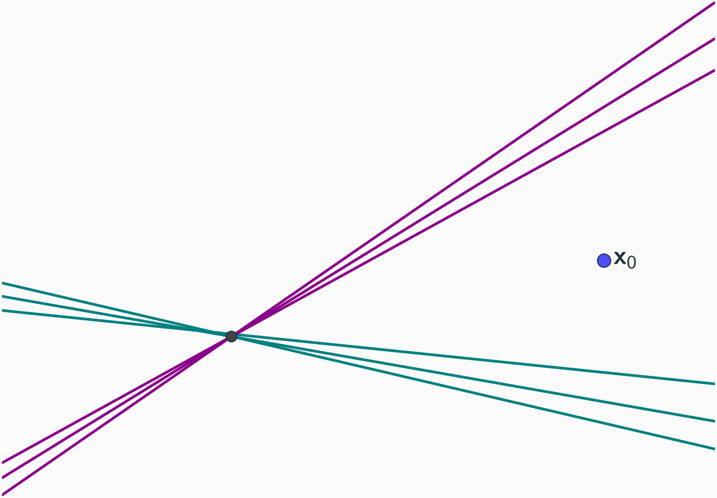
$$\|\mathbf{x}_k - \mathbf{x}\|_2^2 \leq \left(1 - \frac{\sigma_{\min}^2(A)}{m}\right)^k \|\mathbf{x}_0 - \mathbf{x}\|_2^2$$

- SKM (DeLoera, H., Needell '17):

$$\mathbb{E}\|\mathbf{x}_k - \mathbf{x}\|_2^2 \leq \left(1 - \frac{\sigma_{\min}^2(A)}{m}\right)^k \|\mathbf{x}_0 - \mathbf{x}\|_2^2$$

`Why are these all the same?`

A Pathological Example



Structure of the Residual

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]

Structure of the Residual

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.

$$\gamma_k := \frac{\sum_{\tau \in \binom{[m]}{\beta}} \|A_\tau \mathbf{x}_k - \mathbf{b}_\tau\|_2^2}{\sum_{\tau \in \binom{[m]}{\beta}} \|A_\tau \mathbf{x}_k - \mathbf{b}_\tau\|_\infty^2}$$

Accelerated Convergence Rate

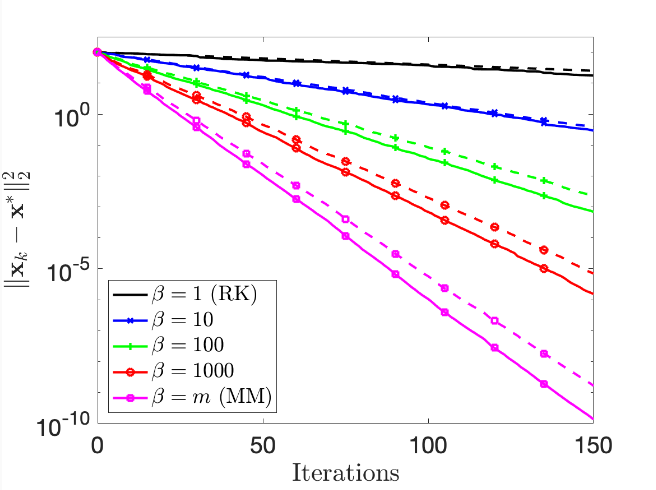
Theorem (H. - Ma 2019)

Let A be normalized so $\|\mathbf{a}_i\|_2 = 1$ for all rows $i = 1, ..., m$. If the system $A\mathbf{x} = \mathbf{b}$ is consistent with the unique solution $\mathbf{x}^*$ 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, τ_j. Precisely, in the $j+1$st iteration of SKM, we have

$$\mathbb{E}_{\tau_j}\|\mathbf{x}_{j+1} - \mathbf{x}^*\|_2^2 \leq \left(1 - \frac{\beta\sigma_{\min}^2(A)}{\gamma_j m}\right)\|\mathbf{x}_j - \mathbf{x}^*\|_2^2,$$

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

Accelerated Convergence Rate



- ▷ A is 50000×100 Gaussian matrix, consistent system
- ▷ bound uses dynamic range of sample of β rows

What can we say about γ_j?

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

$$1 \leq \gamma_j \leq \beta$$

What can we say about γ_j?

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

$$1 \leq \gamma_j \leq \beta$$

What can we say about γ_j?

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

$$1 \leq \gamma_j \leq \beta$$

What can we say about γ_j?

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

$$1 \leq \gamma_j \leq \beta$$

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

	Previous:
RK	$\alpha = 1 - \frac{\sigma_{\min}^2(A)}{m}$
SKM	$\alpha = 1 - \frac{\sigma_{\min}^2(A)}{m}$
MM	$1 - \frac{\sigma_{\min}^2(A)}{4} \leq \alpha \leq 1 - \frac{\sigma_{\min}^2(A)}{m}$

[H., Needell 2019]

What can we say about γ_j?

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

$$1 \leq \gamma_j \leq \beta$$

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

	Previous:	Current:
RK	$\alpha = 1 - \frac{\sigma_{\min}^2(A)}{m}$	$\alpha = 1 - \frac{\sigma_{\min}^2(A)}{m}$
SKM	$\alpha = 1 - \frac{\sigma_{\min}^2(A)}{m}$	$1 - \frac{\beta\sigma_{\min}^2(A)}{m} \leq \alpha \leq 1 - \frac{\sigma_{\min}^2(A)}{m}$
MM	$1 - \frac{\sigma_{\min}^2(A)}{4} \leq \alpha \leq 1 - \frac{\sigma_{\min}^2(A)}{m}$	$1 - \sigma_{\min}^2(A) \leq \alpha \leq 1 - \frac{\sigma_{\min}^2(A)}{m}$

[H., Needell 2019], [H., Ma 2019]

What can we say about γ_j?

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

$$1 \leq \gamma_j \leq \beta$$

▷ nontrivial bounds on γ_k for Gaussian and average consensus systems

Generalizing the Result

- can immediately generalize to varying β (SKM with β_k)

Generalizing the Result

- can immediately generalize to varying β (SKM with β_k)
- to generalize to non-normalized A, we need a sampling distribution that depends upon $\|\mathbf{a}_i\|^2$ and $\mathbf{x}_k$

Generalizing the Result

- can immediately generalize to varying β (SKM with β_k)
- to generalize to non-normalized A, we need a sampling distribution that depends upon $\|\mathbf{a}_i\|^2$ and $\mathbf{x}_k$
- provides RK result that generalizes [Strohmer, Verhsynin '09]

Generalizing the Result

- can immediately generalize to varying β (SKM with β_k)
- to generalize to non-normalized A, we need a sampling distribution that depends upon $\|\mathbf{a}_i\|^2$ and $\mathbf{x}_k$
- 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 β (SKM with β_k)
- ▷ to generalize to non-normalized A, we need a sampling distribution that depends upon $\|\mathbf{a}_i\|^2$ and $\mathbf{x}_k$
- ▷ provides RK result that generalizes [Strohmer, Verhsynin '09]
- ▷ provides simple result for MM in case of non-normalized rows (none previously)
- ▷ now can analyze γ_k for systems with unequal row norms

Sketch of Proof

Pythagorean theorem

$$\begin{aligned} \|\mathbf{x}_k - \mathbf{x}^*\|_2^2 &= \|\mathbf{x}_{k-1} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}_{k-1} - \mathbf{x}_k\|_2^2 \\ &= \|\mathbf{x}_{k-1} - \mathbf{x}^*\|_2^2 - \|A_{\tau_k}\mathbf{x}_{k-1} - \mathbf{b}_{\tau_k}\|_\infty^2 \end{aligned}$$

Sketch of Proof

Pythagorean theorem

$$\begin{aligned}\|\mathbf{x}_k - \mathbf{x}^*\|_2^2 &= \|\mathbf{x}_{k-1} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}_{k-1} - \mathbf{x}_k\|_2^2 \\ &= \|\mathbf{x}_{k-1} - \mathbf{x}^*\|_2^2 - \|A_{\tau_k}\mathbf{x}_{k-1} - \mathbf{b}_{\tau_k}\|_\infty^2\end{aligned}$$

Expected improvement

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

Sketch of Proof

Pythagorean theorem

$$\begin{aligned}\|\mathbf{x}_k - \mathbf{x}^*\|_2^2 &= \|\mathbf{x}_{k-1} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}_{k-1} - \mathbf{x}_k\|_2^2 \\ &= \|\mathbf{x}_{k-1} - \mathbf{x}^*\|_2^2 - \|A_{\tau_k}\mathbf{x}_{k-1} - \mathbf{b}_{\tau_k}\|_\infty^2\end{aligned}$$

Expected improvement

$$\begin{aligned}\mathbb{E}_{\tau_k}\|A_{\tau_k}\mathbf{x}_{k-1} - \mathbf{b}_{\tau_k}\|_\infty^2 &= \sum_{\tau \in \binom{[m]}{\beta}} \frac{1}{\binom{m}{\beta}} \|A_\tau \mathbf{x}_{k-1} - \mathbf{b}_\tau\|_\infty^2 \\ &= \frac{1}{\binom{m}{\beta}} \sum_{\tau \in \binom{[m]}{\beta}} \|A_\tau \mathbf{x}_{k-1} - \mathbf{b}_\tau\|_\infty^2\end{aligned}$$

Sketch of Proof

Pythagorean theorem

$$\begin{aligned}\|\mathbf{x}_k - \mathbf{x}^*\|_2^2 &= \|\mathbf{x}_{k-1} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}_{k-1} - \mathbf{x}_k\|_2^2 \\ &= \|\mathbf{x}_{k-1} - \mathbf{x}^*\|_2^2 - \|A_{\tau_k}\mathbf{x}_{k-1} - \mathbf{b}_{\tau_k}\|_\infty^2\end{aligned}$$

Expected improvement

$$\begin{aligned}\mathbb{E}_{\tau_k}\|A_{\tau_k}\mathbf{x}_{k-1} - \mathbf{b}_{\tau_k}\|_\infty^2 &= \sum_{\tau \in \binom{[m]}{\beta}} \frac{1}{\binom{m}{\beta}} \|A_\tau \mathbf{x}_{k-1} - \mathbf{b}_\tau\|_\infty^2 \\ &= \frac{1}{\binom{m}{\beta}} \sum_{\tau \in \binom{[m]}{\beta}} \|A_\tau \mathbf{x}_{k-1} - \mathbf{b}_\tau\|_\infty^2 \\ &= \frac{1}{\binom{m}{\beta}\gamma_k} \sum_{\tau \in \binom{[m]}{\beta}} \|A_\tau \mathbf{x}_{k-1} - \mathbf{b}_\tau\|_2^2\end{aligned}$$

Sketch of Proof

Pythagorean theorem

$$\begin{aligned}\|\mathbf{x}_k - \mathbf{x}^*\|_2^2 &= \|\mathbf{x}_{k-1} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}_{k-1} - \mathbf{x}_k\|_2^2 \\ &= \|\mathbf{x}_{k-1} - \mathbf{x}^*\|_2^2 - \|A_{\tau_k}\mathbf{x}_{k-1} - \mathbf{b}_{\tau_k}\|_\infty^2\end{aligned}$$

Expected improvement

$$\begin{aligned}\mathbb{E}_{\tau_k}\|A_{\tau_k}\mathbf{x}_{k-1} - \mathbf{b}_{\tau_k}\|_\infty^2 &= \sum_{\tau \in \binom{[m]}{\beta}} \frac{1}{\binom{m}{\beta}} \|A_\tau \mathbf{x}_{k-1} - \mathbf{b}_\tau\|_\infty^2 \\ &= \frac{1}{\binom{m}{\beta}} \sum_{\tau \in \binom{[m]}{\beta}} \|A_\tau \mathbf{x}_{k-1} - \mathbf{b}_\tau\|_\infty^2 \\ &= \frac{1}{\binom{m}{\beta}\gamma_k} \sum_{\tau \in \binom{[m]}{\beta}} \|A_\tau \mathbf{x}_{k-1} - \mathbf{b}_\tau\|_2^2 \\ &= \frac{\beta}{\gamma_k m}\|A\mathbf{x}_{k-1} - \mathbf{b}\|_2^2 \geq \frac{\beta\sigma_{\min}^2(A)}{\gamma_k m}\|\mathbf{x}_{k-1} - \mathbf{x}^*\|_2^2\end{aligned}$$

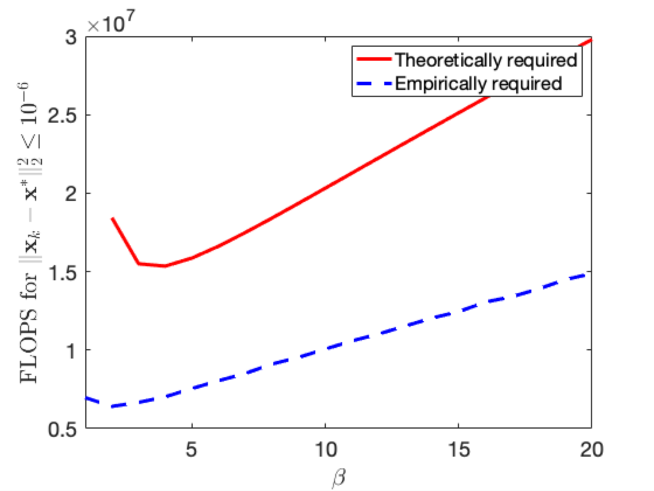
Now can we determine the optimal β?

<u>Now</u> can we determine the optimal β?

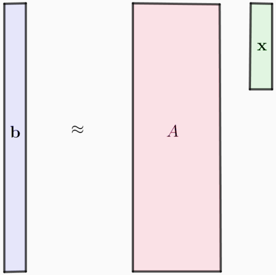
Roughly, if we know the value of γ_j, we can (just) do it.

Now can we determine the optimal β?

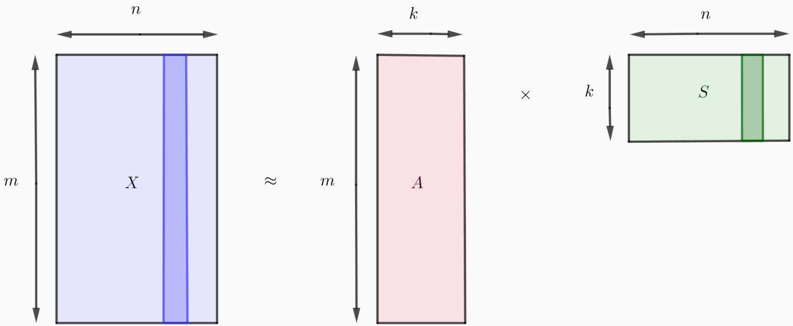
Roughly, if we know the value of γ_j, we can (just) do it.



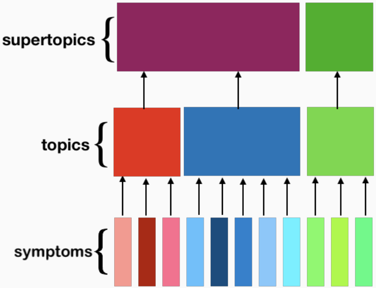
Back to Hierarchical NMF

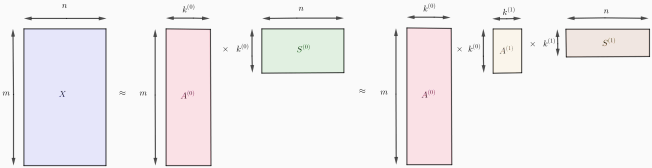


Back to Hierarchical NMF

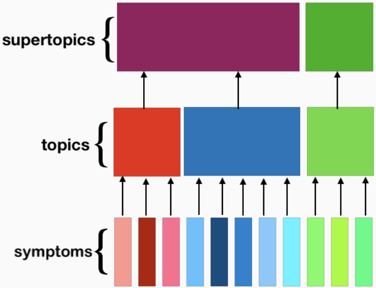


Back to Hierarchical NMF



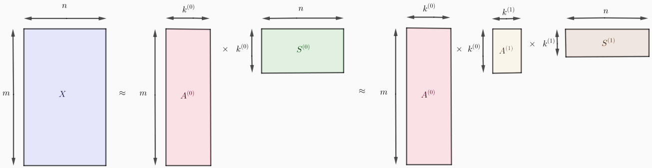


Back to Hierarchical NMF

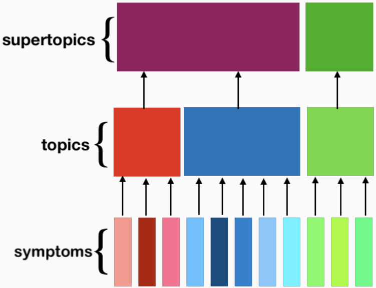


Compare:

- ▷ hNMF (sequential NMF)

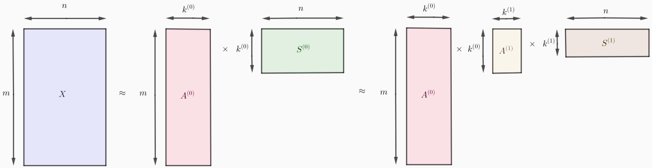


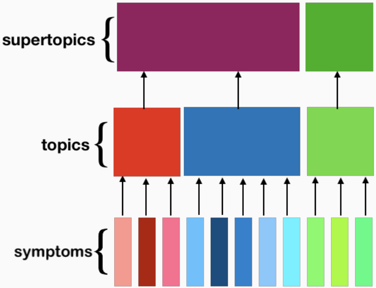
Back to Hierarchical NMF



Compare:

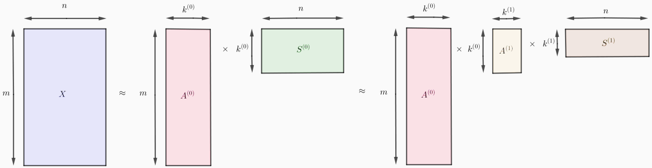
- ▷ hNMF (sequential NMF)
- ▷ Deep NMF [Flenner, Hunter '18]





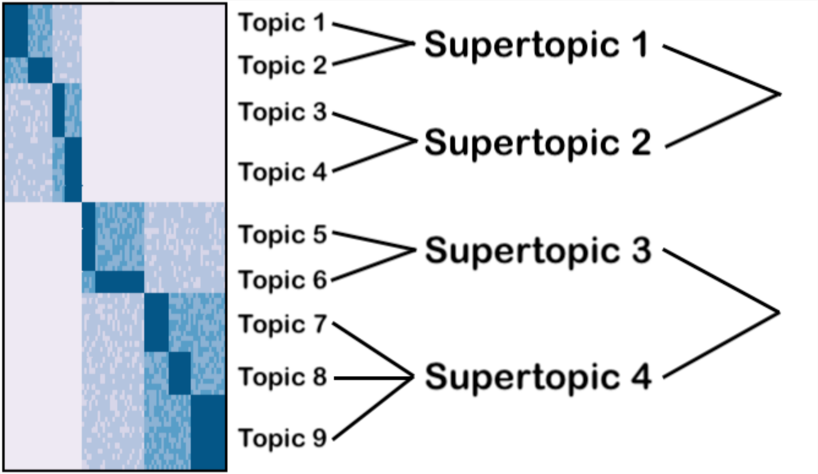
Compare:

- hNMF (sequential NMF)
- Deep NMF [Flenner, Hunter '18]
- Neural NMF

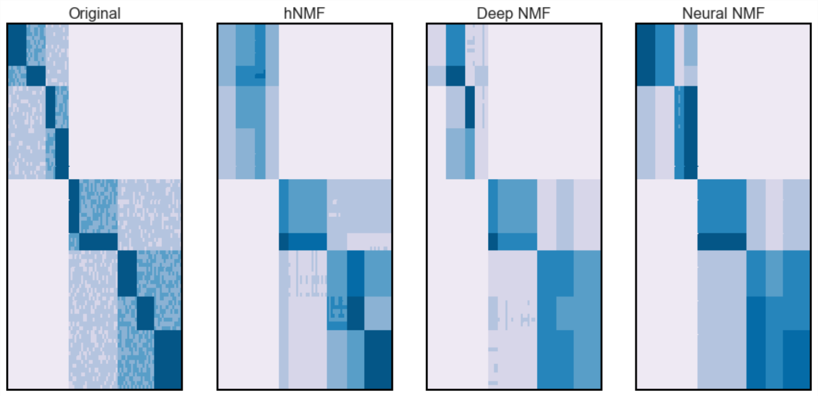


Applications

Experimental results: synthetic data

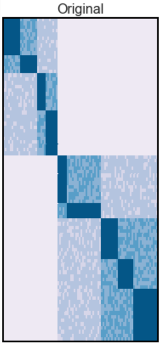


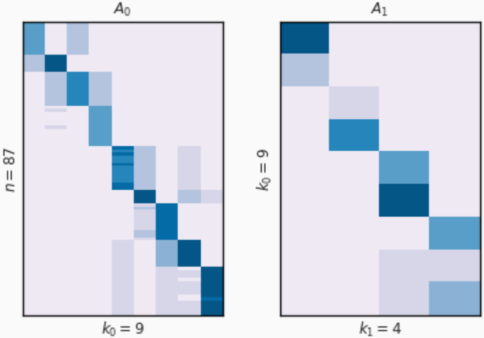
Experimental results: synthetic data



- ▷ unsupervised reconstruction with two-layer structure ($k^{(0)} = 9, k^{(1)} = 4$)

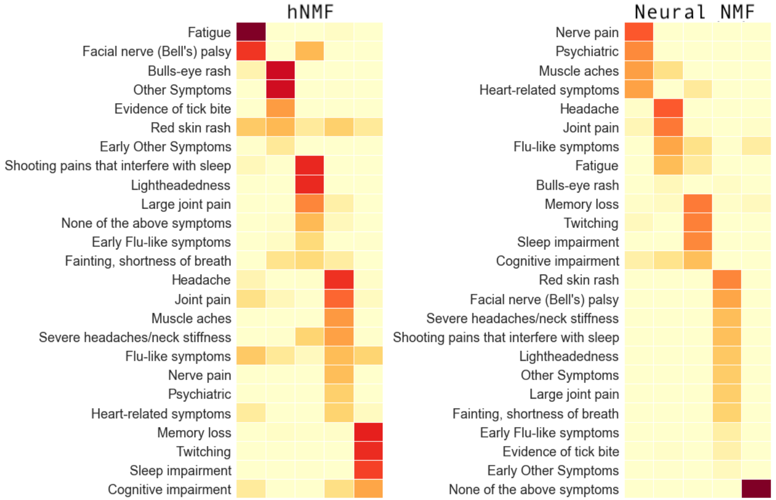
Experimental results: synthetic data



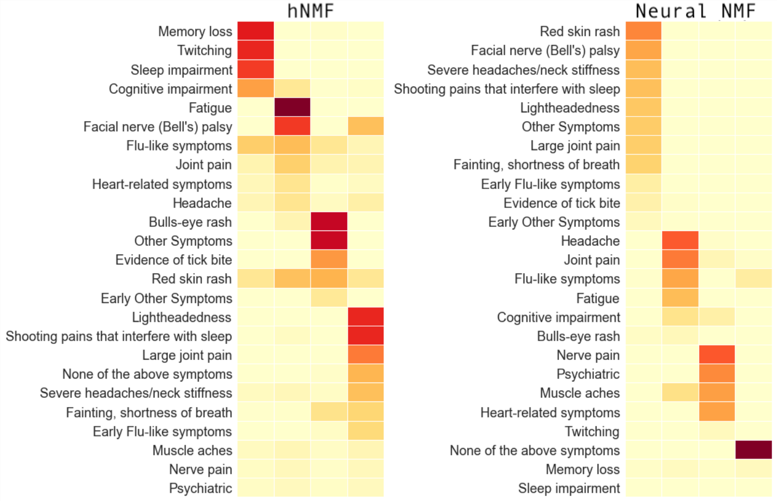


- unsupervised reconstruction with two-layer structure ($k^{(0)} = 9, k^{(1)} = 4$)

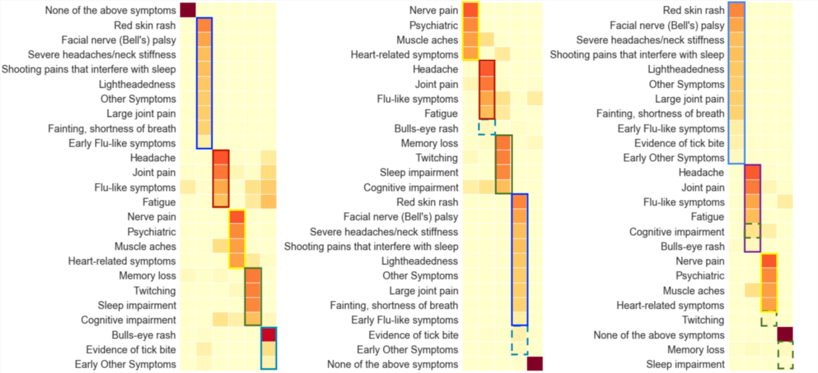
Experimental results: MyLymeData



Experimental results: MyLymeData

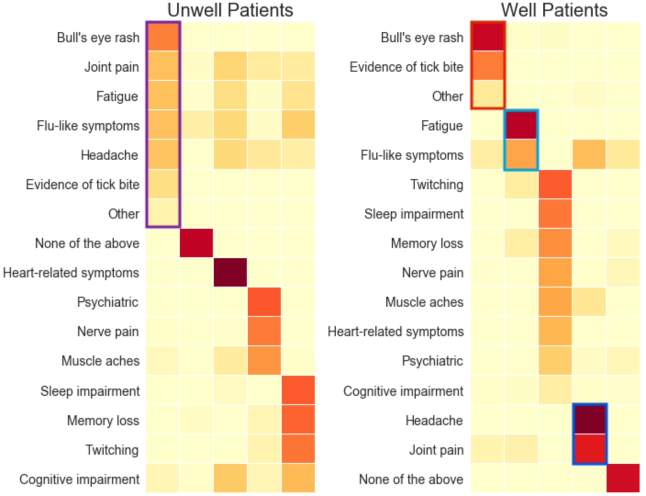


Experimental results: MyLymeData



$k^{(0)} = 6$ $k^{(1)} = 5$ $k^{(2)} = 4$

Experimental results: MyLymeData



MyLymeData Takeaways

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

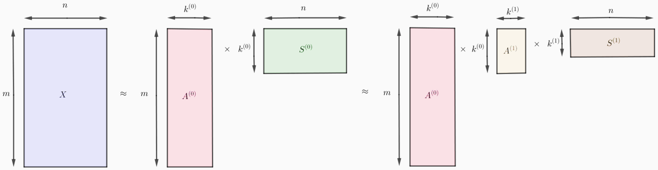
MyLymeData Takeaways

- 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?

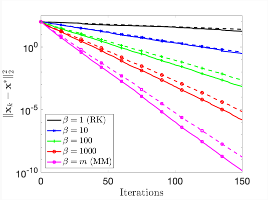
Conclusions

Conclusions

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



- ▷ 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?

[1] S. Agmon. **The relaxation method for linear inequalities.** Canadian J. Math., 6:382–392, 1954.

[2] Z. Bai and W. Wu. **On greedy randomized Kaczmarz method for solving large sparse linear systems.** SIAM J. Sci. Comput., 40(1):A592–A606, 2018.

[3] A. Cichocki and R. Zdunek. **Multilayer nonnegative matrix factorisation.** Electron. Lett., 42(16):947, 2006.

[4] J. A. De Loera, J. Haddock, and D. Needell. **A sampling Kaczmarz-Motzkin algorithm for linear feasibility.** SIAM J. Sci. Comput., 39(5):S66–S87, 2017.

[5] K. Du and H. Gao. **A new theoretical estimate for the convergence rate of the maximal weighted residual Kaczmarz algorithm.** Numer. Math. - Theory Me., 12(2):627–639, 2019.

[6] M. Gao, J. Haddock, D. Molitor, D. Needell, E. Sadovnik, T. Will, and R. Zhang. **Neural nonnegative matrix factorization for hierarchical multilayer topic modeling.** In Proc. Interational Workshop on Computational Advances in Multi-Sensor Adaptive Processing, 2019.

[7] J. Haddock and A. Ma. **Greed works: An improved analysis of sampling Kaczmarz-Motzkin.** 2019. Submitted.

[8] J. Haddock and D. Needell. **On Motzkins method for inconsistent linear systems.** BIT, 59(2):387–401, 2019.

[9] S. Kaczmarz. **Angenäherte auflösung von systemen linearer gleichungen.** Bull. Int. Acad. Polon. Sci. Lett. Ser. A, pages 335–357, 1937.

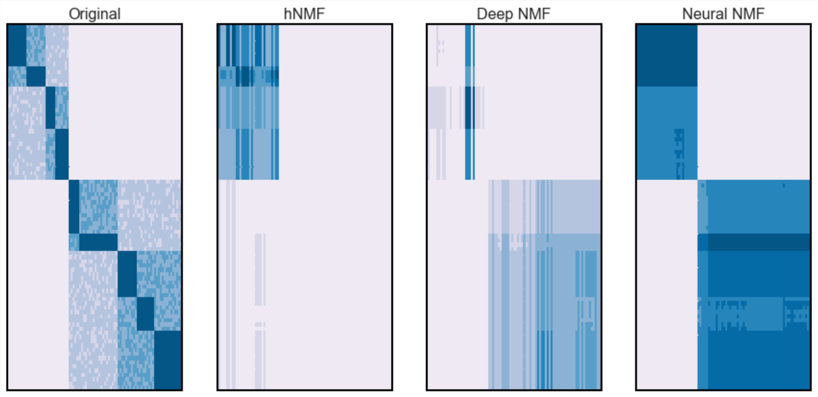
[10] D. D. Lee and H. S. Seung. **Learning the parts of objects by non-negative matrix factorization.** Nature, 401:788–791, 1999.

[11] T. S. Motzkin and I. J. Schoenberg. **The relaxation method for linear inequalities.** Canadian J. Math., 6:393–404, 1954.

[12] P. Paatero and U. Tapper. **Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values.** Environmetrics, 5(2):111–126, 1994.

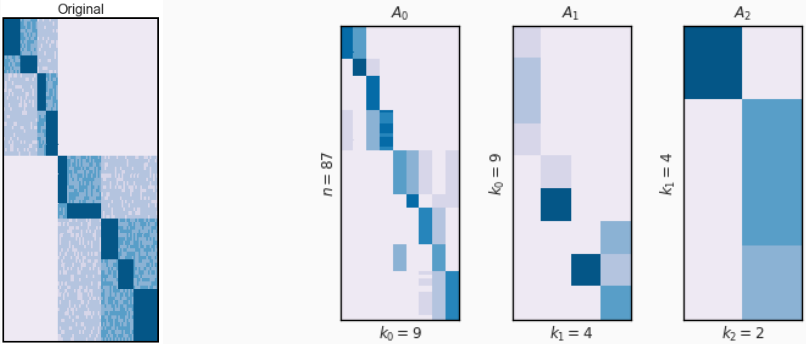
[13] T. Strohmer and R. Vershynin. **A randomized Kaczmarz algorithm with exponential convergence.** J. Fourier Anal. Appl., 15:262–278, 2009.

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



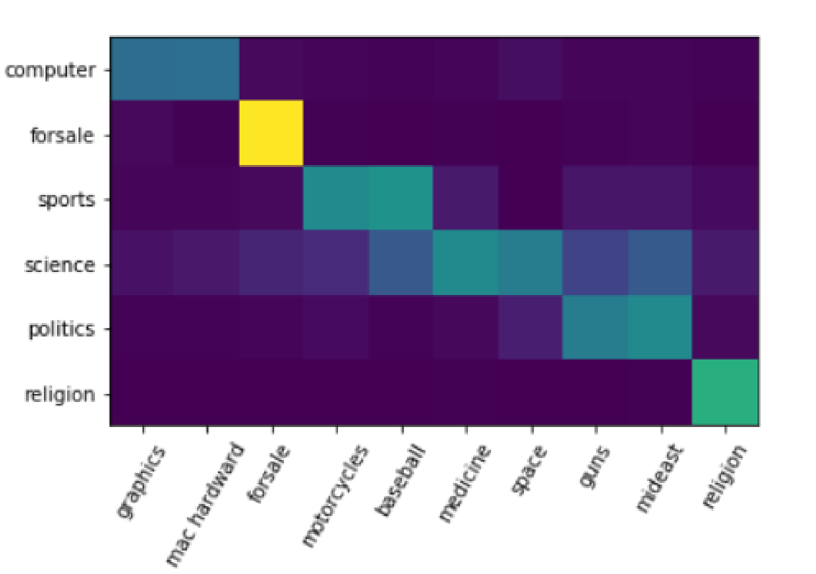
- semisupervised reconstruction (40% labels) with three-layer structure ($k^{(0)} = 9, k^{(1)} = 4, k^{(2)} = 2$)

Experimental results: synthetic data

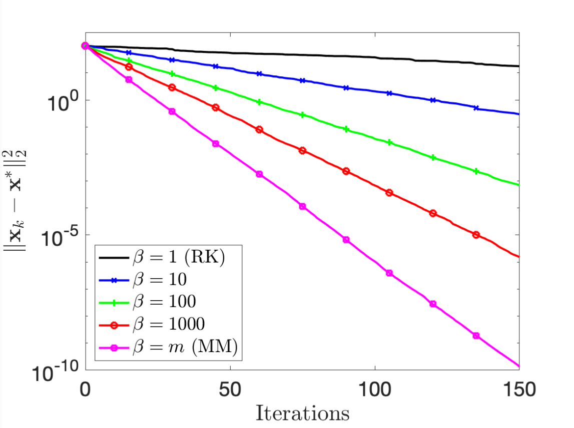
Table 1: Reconstruction error / classification accuracy

	Layers	Hier. NMF	Deep NMF	Neural NMF
Unsuper.	1	0.053	0.031	0.029
	2	0.399	0.414	**0.310**
	3	0.860	0.838	**0.492**
Semisuper.	1	0.049 / 0.933	0.031 / 0.947	0.042 / **1**
	2	0.374 / 0.926	0.394 / 0.911	**0.305** / **1**
	3	0.676 / 0.930	0.733 / 0.930	**0.496** / **0.990**
Supervised	1	0.052 / 0.960	0.042 / 0.962	0.042 / **1**
	2	0.311 / 0.984	0.310 / 0.984	0.307 / 1
	3	0.495 / 1	0.494 / 1	0.498 / 1

Experimental results: 20 Newsgroups data

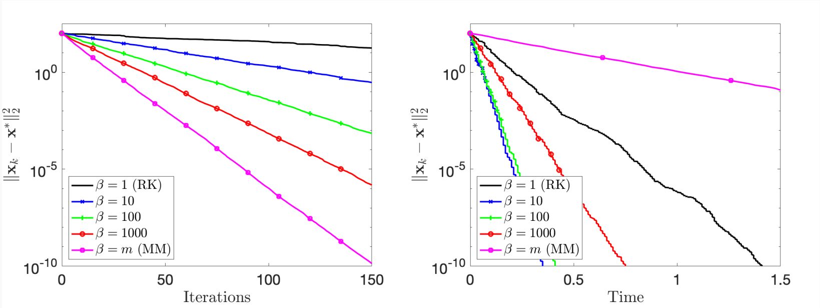


Experimental Convergence



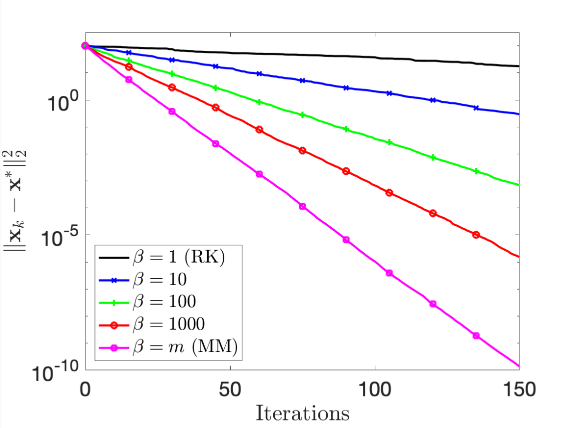
- β: sample size
- A is 50000×100 Gaussian matrix, consistent system
- ‘faster’ convergence for larger sample size

Experimental Convergence



- β: sample size
- A is 50000×100 Gaussian matrix, consistent system
- ‘faster’ convergence for larger sample size

Experimental Convergence



- β: sample size
- A is 50000×100 Gaussian matrix, consistent system
- 'faster' convergence for larger sample size

Deep NMF

Goal: Exploit similarities between neural networks and hierarchical NMF.

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

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

Deep NMF

Goal: Exploit similarities between neural networks and hierarchical NMF.

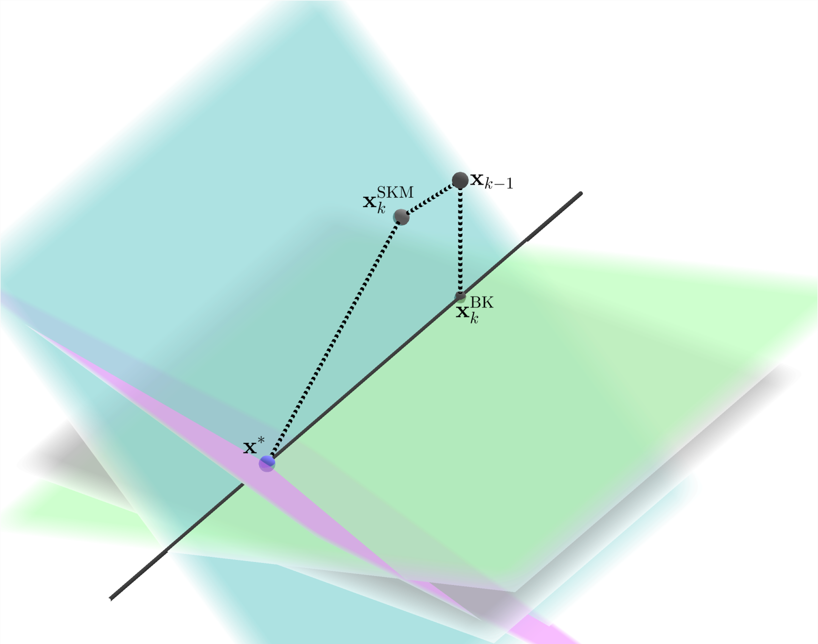
- ▷ [Flenner, Hunter '18]
  - introduces nonlinear pooling operator after each layer
  - introduces multiplicative updates method meant to backpropagate
- ▷ [Trigeorgis, Bousmalis, Zafeiriou, Schuller '16]
  - relaxes some of nonnegativity constraints in hNMF
- ▷ [Le Roux, Hershey, Weninger '15]
  - 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 γ_j

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