Matrix and Tensor Factorization Models: Applications, Algorithms, and Theory

Hanbaek Lyu

Department of Mathematics, IFDS
University of Wisconsin - Madison

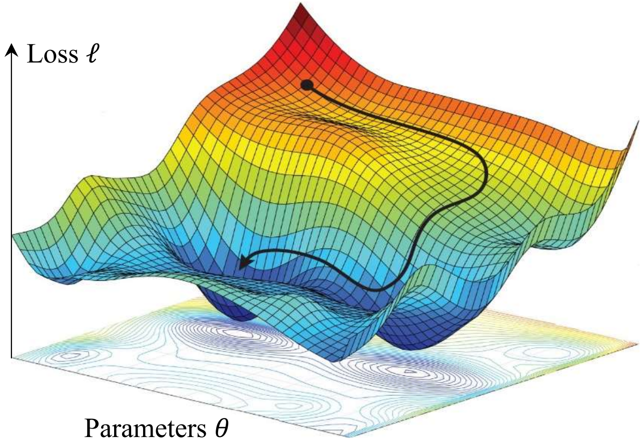
Partially supported by NSF DMS #2206296 and #2010035

POSTECH MINDS seminar

May 31, 2022

Outline

- Optimization is a fundamental task whenever there is data to be explained by a model with parameters
- Data ≈ Model(θ)
  - e.g., Regression models (linear, logistic,..), latent variable models (matrix/tensor factorization,..), deep neural networks (CNN, RNN, GNN,..)



- How to chose optimal parameter θ^*?

$$\theta^* = \underset{\theta \in \Theta}{\operatorname{argmin}}\ \ell(\text{Data}, \theta)$$

ℓ = Loss function

Θ = Parameter space

- In this talk:
  - **Data** : images, texts, graphs, video frames
  - **Models** : matrix/tensor factorization (latent variable models)
  - **Optimization** : block coordinate descent, SGD, SMM (stochastic majorization-minimization)
  - **Theory** : Convergence to stationary points, non-unique global min, rate of convergecne
- Models:
  - Nonnegative Matrix Factorization — (Dictionary learning for vector signals)

$$\min_{\mathbf{W}\in\mathbb{R}^{p\times r}_{\geq 0},\mathbf{H}\in\mathbb{R}^{r\times n}_{\geq 0}} \|\mathbf{X}-\mathbf{WH}\|_F^2$$

- In this talk:
  - **Data** : images, texts, graphs, video frames
  - **Models** : matrix/tensor factorization (latent variable models)
  - **Optimization** : block coordinate descent, SGD, SMM (stochastic majorization-minimization)
  - **Theory** : Convergence to stationary points, non-unique global min, rate of convergecne
- Models:
  - Nonnegative Matrix Factorization — (Dictionary learning for vector signals)

$$\min_{\mathbf{W}\in\mathbb{R}^{p\times r}_{\geq 0},\,\mathbf{H}\in\mathbb{R}^{r\times n}_{\geq 0}} \|\mathbf{X}-\mathbf{W}\mathbf{H}\|_F^2$$

  - Nonnegative CP Decomposition — (Dictionary learning for multimodal signals)

$$\min_{\mathbf{U}^{(1)}\in\mathbb{R}^{a\times r}_{\geq 0},\,\mathbf{U}^{(2)}\in\mathbb{R}^{b\times r}_{\geq 0},\,\mathbf{U}^{(3)}\in\mathbb{R}^{c\times r}_{\geq 0}} \|\mathbf{X}-\mathtt{Out}(\mathbf{U}^{(1)},\mathbf{U}^{(2)},\mathbf{U}^{(3)})\|_F^2$$

- In this talk:
  - **Data** : images, texts, graphs, video frames
  - **Models** : matrix/tensor factorization (latent variable models)
  - **Optimization** : block coordinate descent, SGD, SMM (stochastic majorization-minimization)
  - **Theory** : Convergence to stationary points, non-unique global min, rate of convergecne
- Models:
  - Nonnegative Matrix Factorization — (Dictionary learning for vector signals)

$$\min_{\mathbf{W}\in\mathbb{R}^{p\times r}_{\geq 0},\mathbf{H}\in\mathbb{R}^{r\times n}_{\geq 0}} \|\mathbf{X}-\mathbf{W}\mathbf{H}\|_F^2$$

  - Nonnegative CP Decomposition — (Dictionary learning for multimodal signals)

$$\min_{\mathbf{U}^{(1)}\in\mathbb{R}^{a\times r}_{\geq 0},\mathbf{U}^{(2)}\in\mathbb{R}^{b\times r}_{\geq 0},\mathbf{U}^{(3)}\in\mathbb{R}^{c\times r}_{\geq 0}} \|\mathbf{X}-\mathtt{Out}(\mathbf{U}^{(1)},\mathbf{U}^{(2)},\mathbf{U}^{(3)})\|_F^2$$

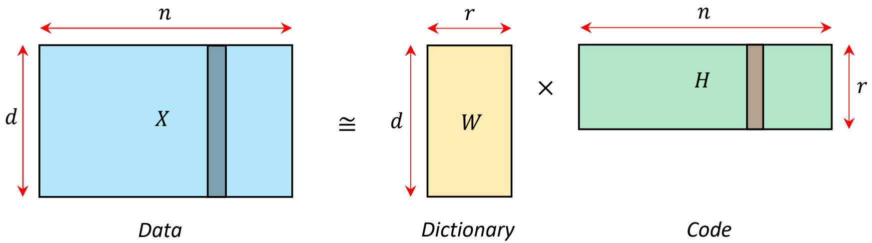
  - Supervised Dictionary Learning — (Learning class-discriminating dictionary)

$$\min_{\mathbf{W}\in\mathbb{R}^{p\times r}_{\geq 0},\mathbf{H}\in\mathbb{R}^{r\times n}_{\geq 0},\boldsymbol{\beta}\in\mathbb{R}^r} NLL(\mathbf{Y},\text{logistic}(\mathbf{W}^T\mathbf{X},\boldsymbol{\beta}))+\xi\|\mathbf{X}-\mathbf{W}\mathbf{H}\|_F^2$$

Outline

Matrix Factorization

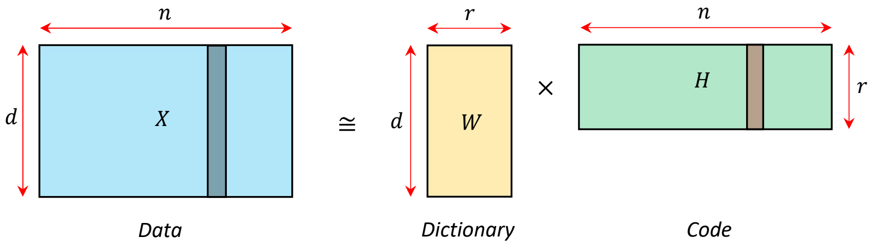
- Matrix factorization is a fundamental tool in dictionary learning problems.



$$\text{Data} \approx \text{Linear combination of } \overbrace{\text{latent features}}^{\text{cols. of } W}$$

Matrix Factorization

- Matrix factorization is a fundamental tool in dictionary learning problems.



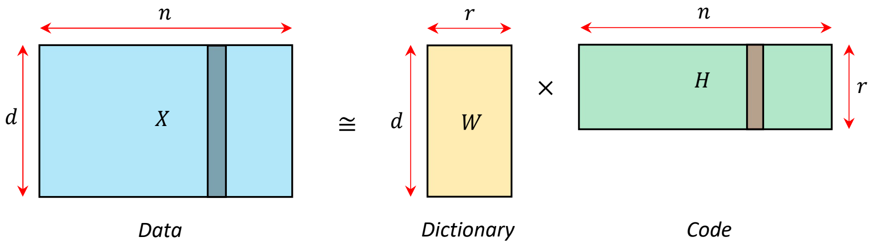
$$\text{Data} \approx \text{Linear combination of } \overbrace{\text{latent features}}^{\text{cols. of } W}$$

- Formulated as a non-convex optimization problem:

$$\begin{cases} \text{minimize} \quad \|X - WH\|_F^2 + \lambda \|H\|_1 & \text{(Reconstruction error)} \\ \text{subject to} \quad W \in \mathscr{C},\, H \in \mathscr{C}' & \text{(Constraints)} \end{cases}$$

Matrix Factorization

- Matrix factorization is a fundamental tool in dictionary learning problems.



$$\text{Data} \approx \text{Linear combination of } \overbrace{\text{latent features}}^{\text{cols. of } W}$$

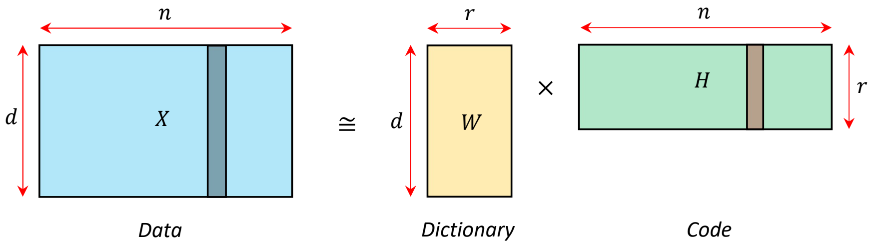
- Formulated as a non-convex optimization problem:

$$\begin{cases} \text{minimize} & \|X - WH\|_F^2 + \lambda \|H\|_1 & \text{(Reconstruction error)} \\ \text{subject to} & W \in \mathscr{C},\ H \in \mathscr{C}' & \text{(Constraints)} \end{cases}$$

  - Unconstrained MF ($\mathscr{C} = \mathbb{R}^{d \times r}$, $\mathscr{C}' = \mathbb{R}^{r \times n}$, $\lambda = 0$): Solved by SVD

Matrix Factorization

- Matrix factorization is a fundamental tool in dictionary learning problems.



$$\text{Data} \approx \text{Linear combination of } \overbrace{\text{latent features}}^{\text{cols. of } W}$$

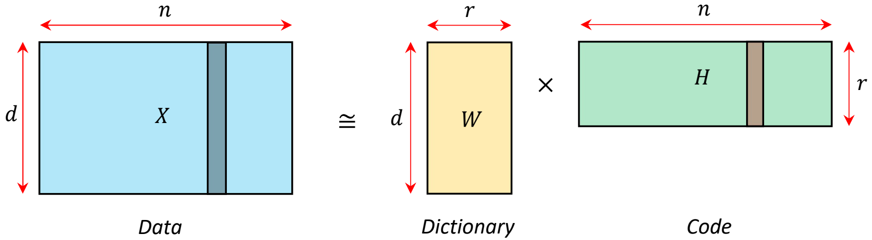
- Formulated as a non-convex optimization problem:

$$\begin{cases} \text{minimize} & \|X - WH\|_F^2 + \lambda \|H\|_1 & \text{(Reconstruction error)} \\ \text{subject to} & W \in \mathscr{C},\, H \in \mathscr{C}' & \text{(Constraints)} \end{cases}$$

- Unconstrained MF ($\mathscr{C} = \mathbb{R}^{d\times r}$, $\mathscr{C}' = \mathbb{R}^{r\times n}$, $\lambda = 0$): Solved by SVD
- Nonnegative Matrix Factorization (NMF): $\mathscr{C} = \mathbb{R}_{\geq 0}^{d\times r}$, $\mathscr{C}' = \mathbb{R}_{\geq 0}^{r\times n}$, $\lambda = 0$,

Matrix Factorization

- Matrix factorization is a fundamental tool in dictionary learning problems.



$$\text{Data} \approx \text{Linear combination of } \overbrace{\text{latent features}}^{\text{cols. of } W}$$

- Formulated as a non-convex optimization problem:

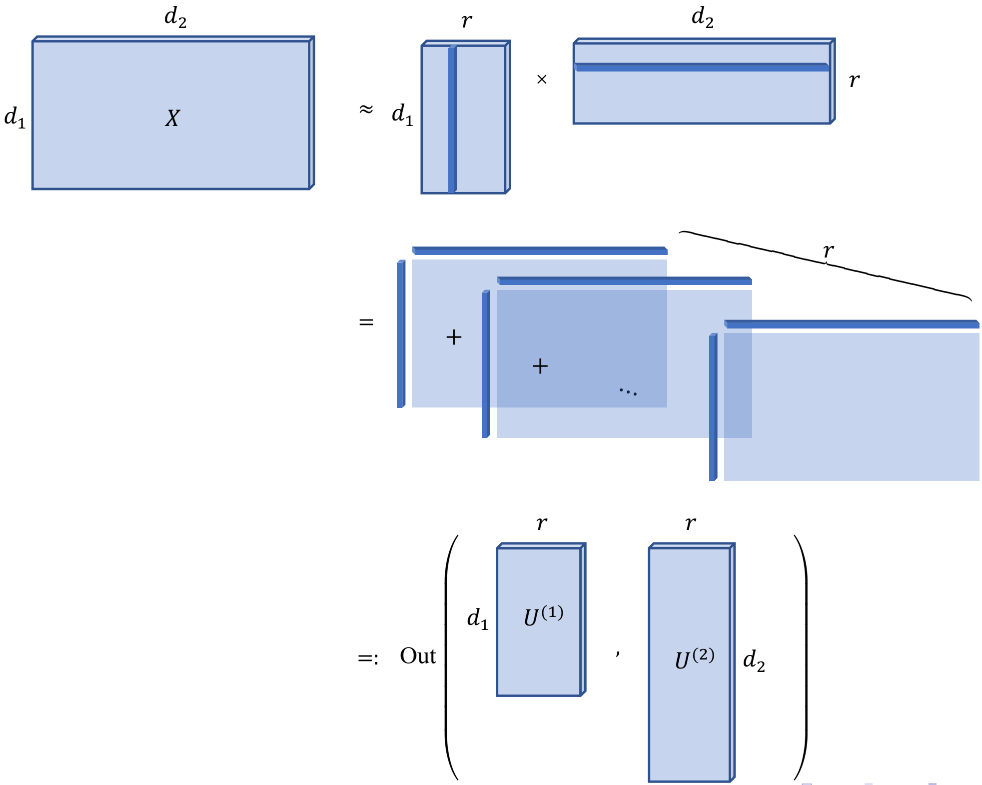
$$\begin{cases} \text{minimize} & \|X - WH\|_F^2 + \lambda \|H\|_1 \qquad \text{(Reconstruction error)} \\ \text{subject to} & W \in \mathscr{C},\ H \in \mathscr{C}' \qquad \text{(Constraints)} \end{cases}$$

  - Unconstrained MF ($\mathscr{C} = \mathbb{R}^{d\times r}$, $\mathscr{C}' = \mathbb{R}^{r\times n}$, $\lambda = 0$): Solved by SVD
  - Nonnegative Matrix Factorization (NMF): $\mathscr{C} = \mathbb{R}_{\geq 0}^{d\times r}$, $\mathscr{C}' = \mathbb{R}_{\geq 0}^{r\times n}$, $\lambda = 0$,

- Applications in text analysis, image reconstruction, medical imaging, bioinformatics, etc.

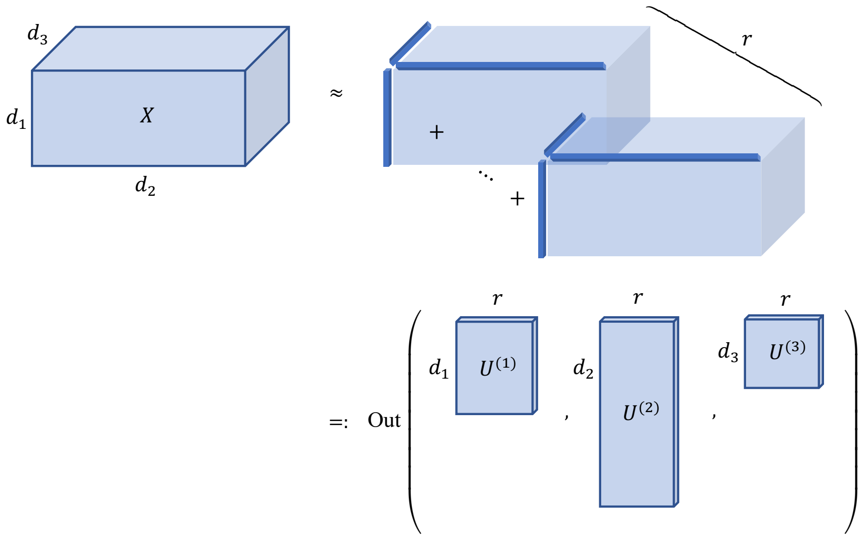
An alternative view of Matrix Factorization

- $\mathbf{X} \approx \mathrm{Out}(\mathbf{U}^{(1)}, \mathbf{U}^{(2)})$



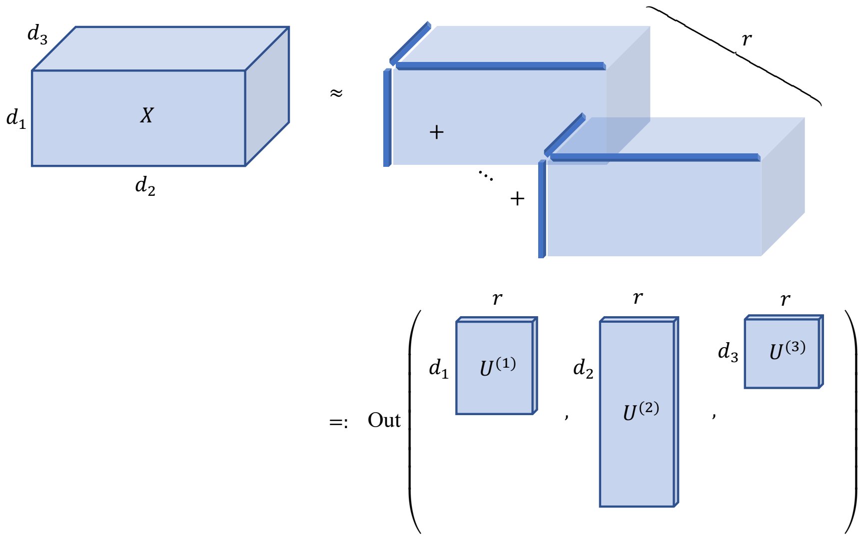
Tensor Factorization (CP decomposition)

- $\mathbf{X} \approx \text{Out}(\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \mathbf{U}^{(3)})$



Tensor Factorization (CP decomposition)

- $\mathbf{X} \approx \text{Out}(\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \mathbf{U}^{(3)})$



- Nonnegative CP Decomposition

$$\min_{U^{(1)} \in \mathbb{R}_{\geq 0}^{d_1 \times r}, U^{(2)} \in \mathbb{R}_{\geq 0}^{d_2 \times r}, U^{(3)} \in \mathbb{R}_{\geq 0}^{d_3 \times r}} \|\mathbf{X} - \text{Out}(\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \mathbf{U}^{(3)})\|_F^2$$

Learning parts of images (MNIST handwritten digits)

- **Dictionary Learning**: Learn r basis vectors from a given data set of 'vectors'
  - 'vectors' may represent images, texts, time-serieses, graphs, etc.
  - Provides a compressed representation of complex objects using a few dictionary elements.

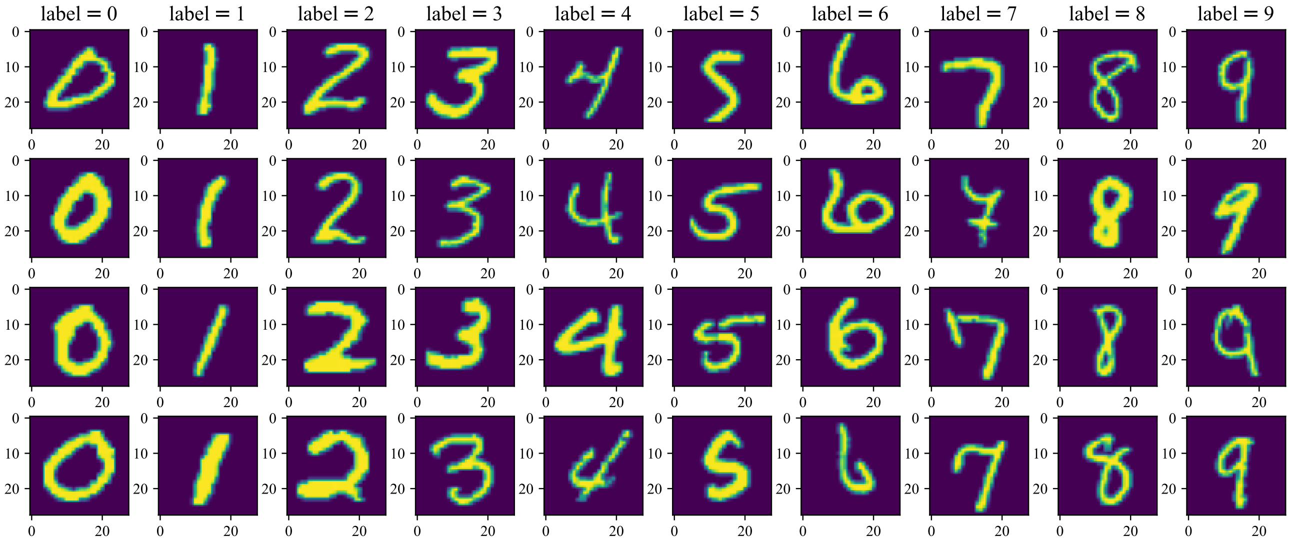


Figure: Sample MNIST images (total 70000 images of size 28×28)

Learning parts of images (MNIST handwritten digits)

- **Dictionary Learning**: Learn r basis vectors from a given data set of 'vectors'
  - 'vectors' may represent images, texts, time-serieses, graphs, etc.
  - Provides a compressed representation of complex objects using a few dictionary elements.

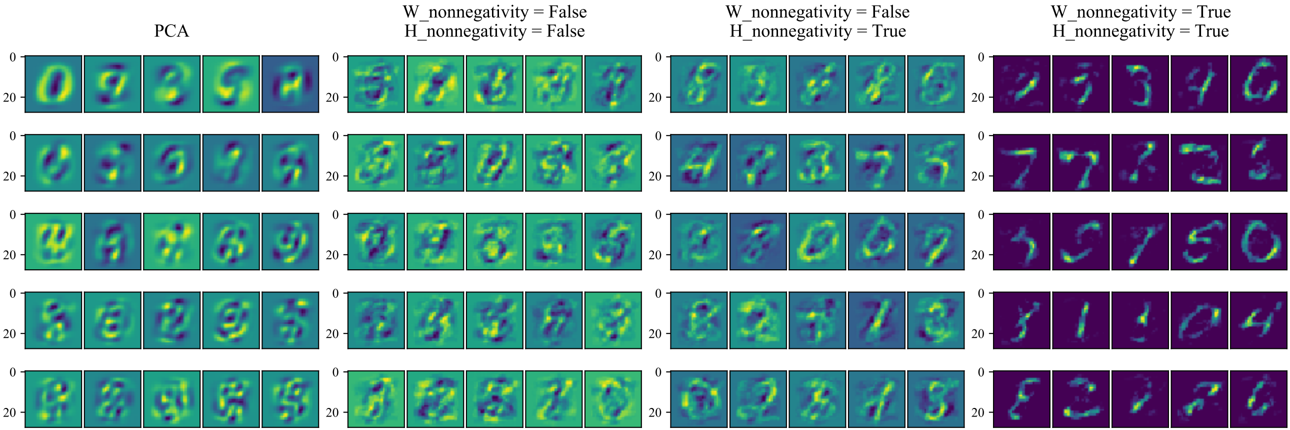


Figure: Example dictionaries learned by PCA and matrix factorization

Topic modeling (20 News Grpups)

- **Dictionary Learning**: Learn r basis vectors from a given data set of 'vectors'
  - 'vectors' may represent images, texts, time-serieses, graphs, etc.
  - Provides a compressed representation of complex objects using a few dictionary elements.

```
>>>> data_cleaned[i] Anyone know what would cause my IIcx to not turn on when I hit the keyboard
switch?  The one in the back of the machine doesn't work either...
The only way I can turn it on is to unplug the machine for a few minutes,
then plug it back in and hit the power switch in the back immediately...
Sometimes this doesn't even work for a long time...

I remember hearing about this problem a long time ago, and that a logic
board failure was mentioned as the source of the problem...is this true?
```

Figure: Example of text data from the 20 News Groups (20 categories, 5616 articles)

Topic modeling (20 News Grpups)

- **Dictionary Learning**: Learn r basis vectors from a given data set of 'vectors'
  - 'vectors' may represent images, texts, time-serieses, graphs, etc.
  - Provides a compressed representation of complex objects using a few dictionary elements.

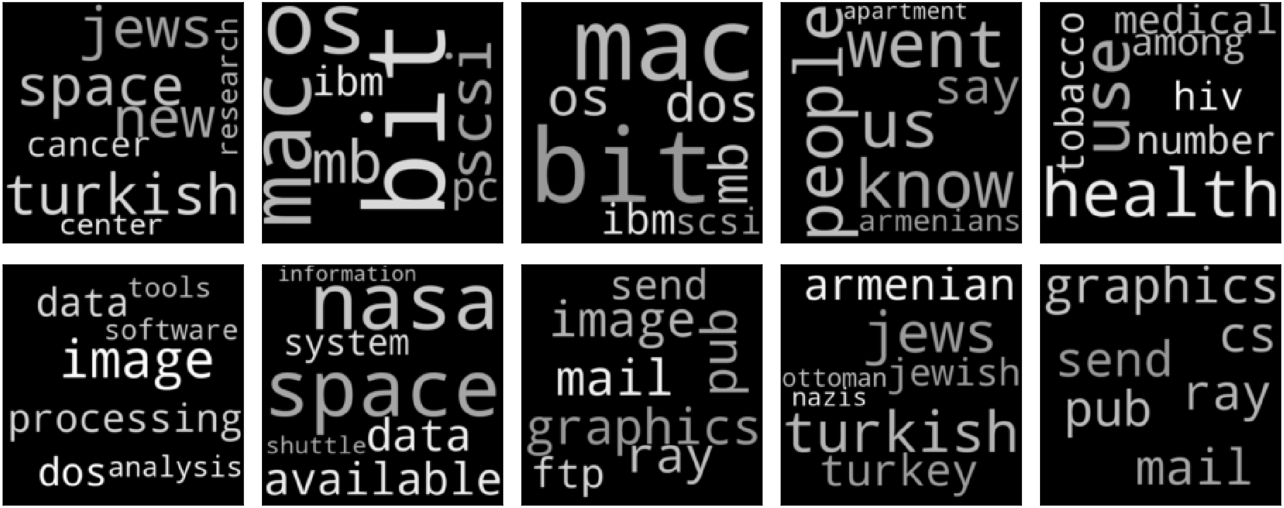
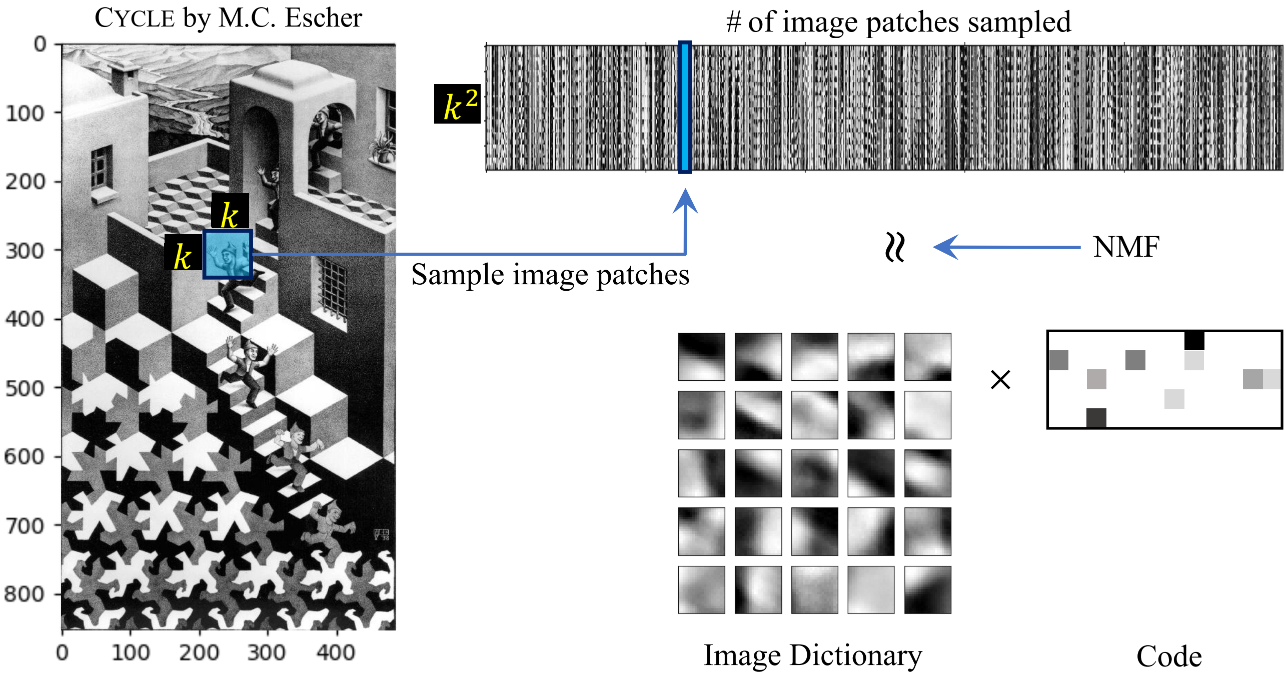


Figure: Example dictionaries (topics) learned by nonnegative matrix factorization from 20 News Groups

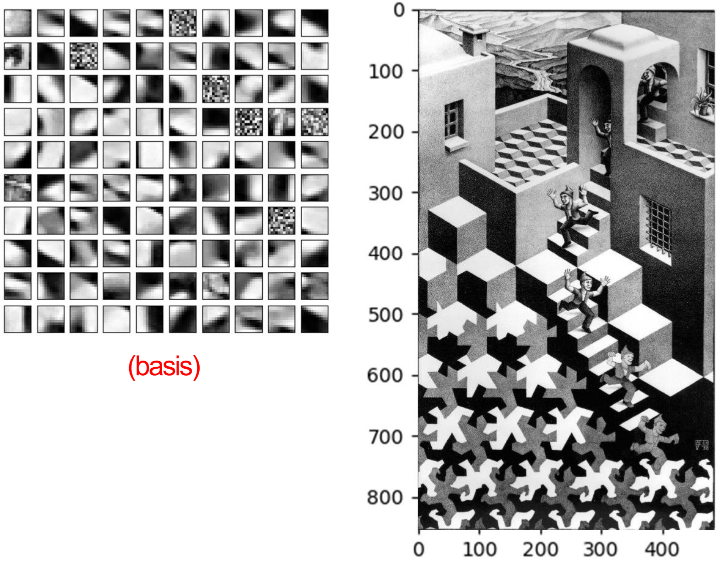
Example of NMF for Image dictionary learning



Learning parts of images – Image reconstruction

Dictionary learned from
Cycle by M. C. Escher

Original image -
Cycle by M. C. Escher (1928)

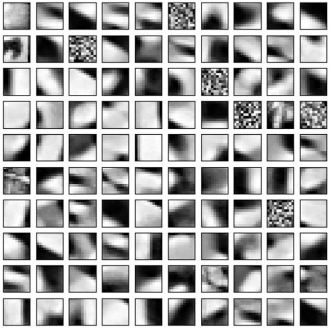


(basis)

- Dictionary learning enables a compressed representation of complex objects using a few dictionary elements.
- Used in data compression, reconstruction, denoising, transfer learning, etc.

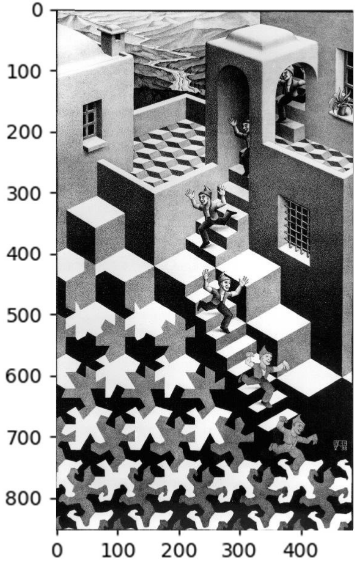
Learning parts of images – Image reconstruction

Dictionary learned from Cycle by M. C. Escher

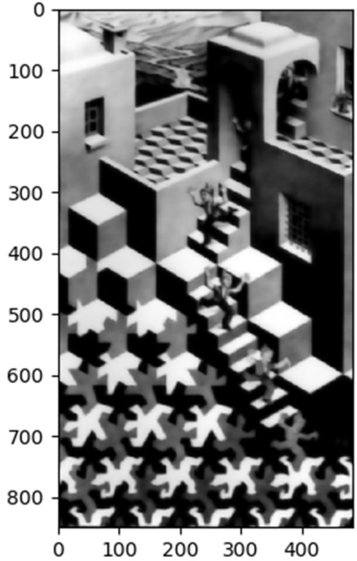


(basis)

Original image - Cycle by M. C. Escher (1928)

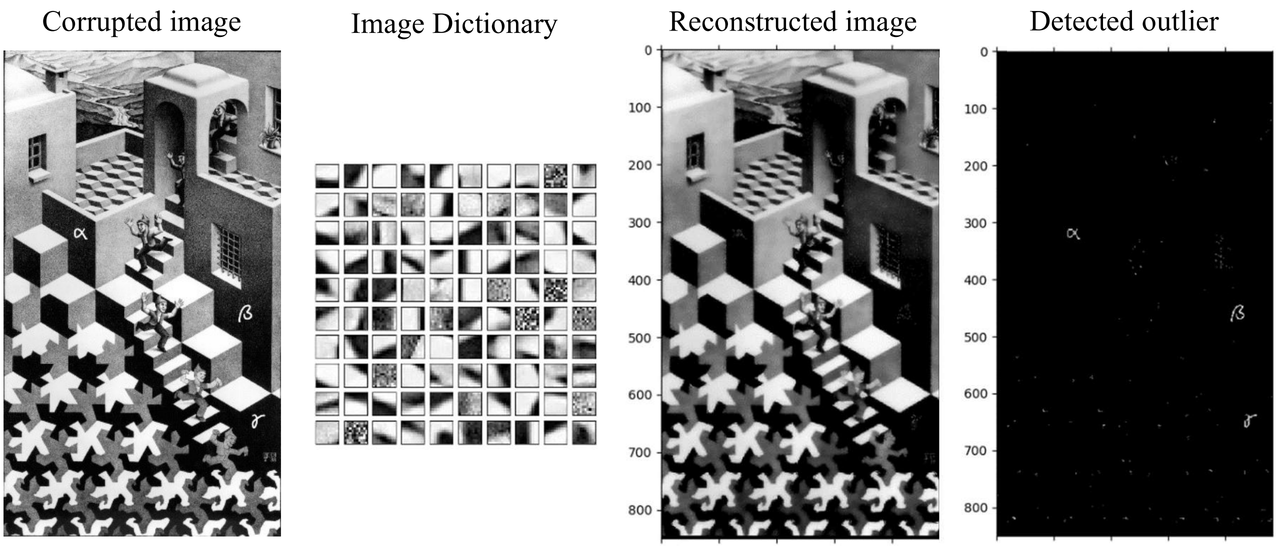


Reconstructed image using learned dictionary



- Dictionary learning enables a compressed representation of complex objects using a few dictionary elements.
- Used in data compression, reconstruction, denoising, transfer learning, etc.
- Img recons. = (local approx. by dict.) + (Averaging)

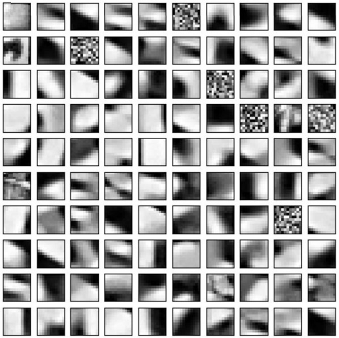
Learning parts of images – Image denoising



- Dictionary learning enables a compressed representation of complex objects using a few dictionary elements.
- Used in data compression, reconstruction, denoising [1, 7], transfer learning, etc.
- Img recons. = (local approx. by dict.) + (Averaging)

Learning parts of images – Transfer learning

Dictionary learned from
Cycle by M. C. Escher



(basis)

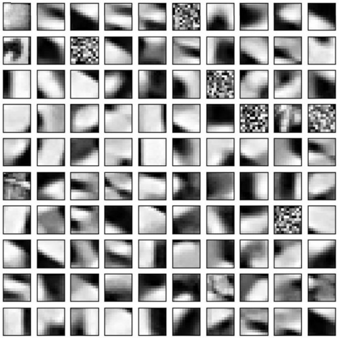
Original image -
Two Sisters by A. Renoir (1882)

(New data)

- Dictionary learning enables a compressed representation of complex objects using a few dictionary elements.
- Used in data compression, reconstruction, transfer learning, etc.
- Img recons. = (local approx. by dict.) + (Averaging)

Learning parts of images – Transfer learning

Dictionary learned from
Cycle by M. C. Escher

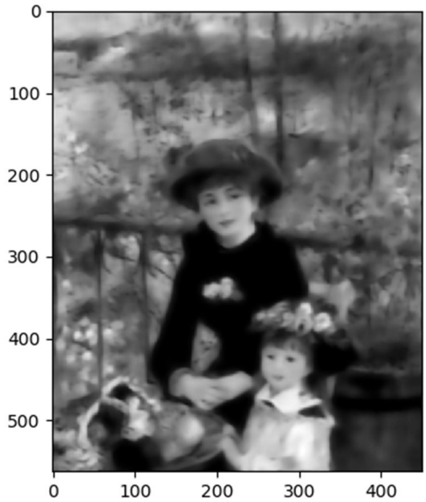


(basis)

Original image -
Two Sisters by A. Renoir (1882)

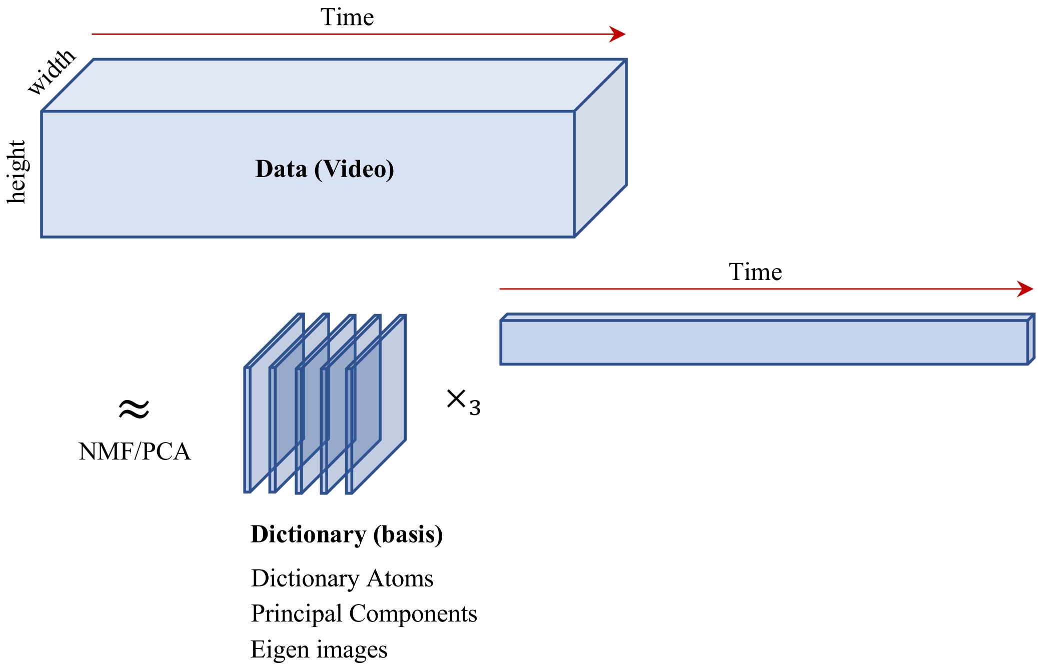
(New data)

Reconstructed image using Dict. learned from Escher's Cycle

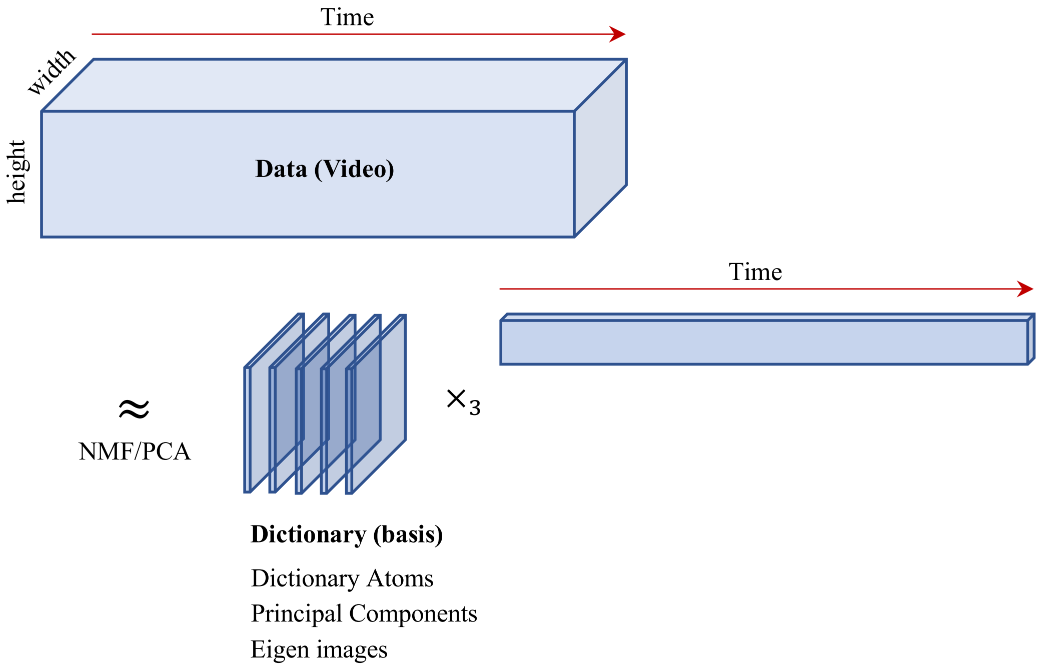


- Dictionary learning enables a compressed representation of complex objects using a few dictionary elements.
- Used in data compression, reconstruction, transfer learning, etc.
- Img recons. = (local approx. by dict.) + (Averaging)

Dictionary Learning from Video Frames



Dictionary Learning from Video Frames

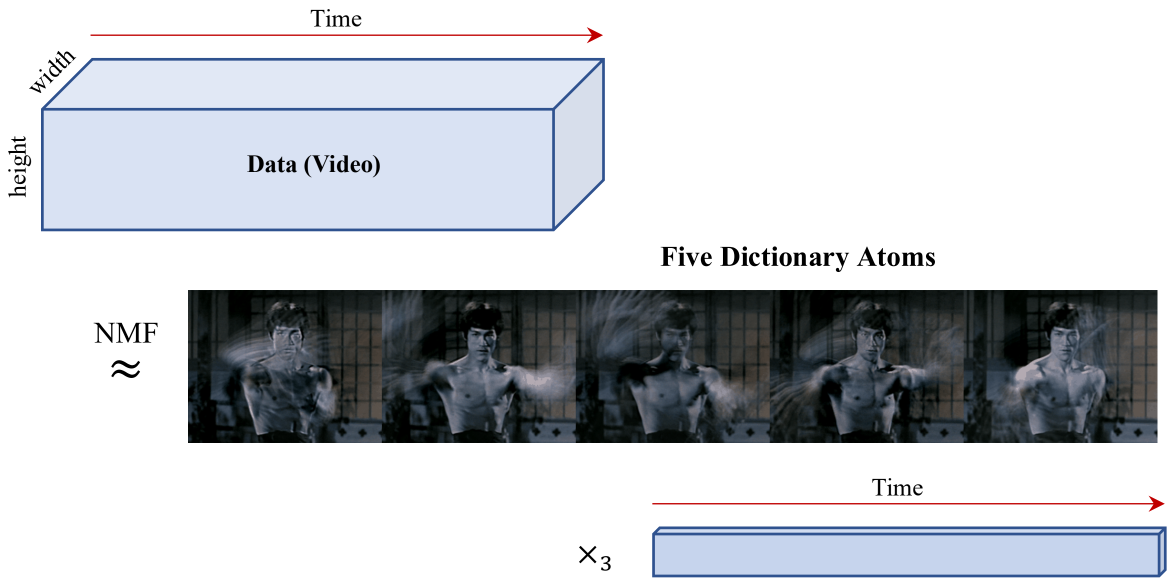


- Entire video frames are processed at once (batch processing)

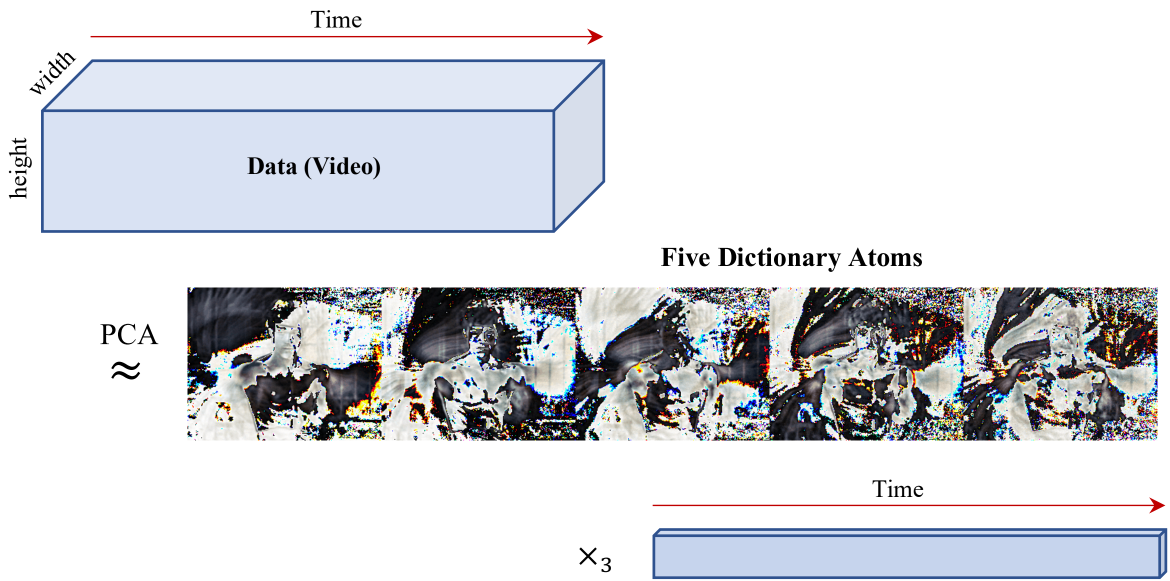
A Toy Example Video

Figure: Bruce Lee (doing his stuff)

Dictionary Learning from Video Frames



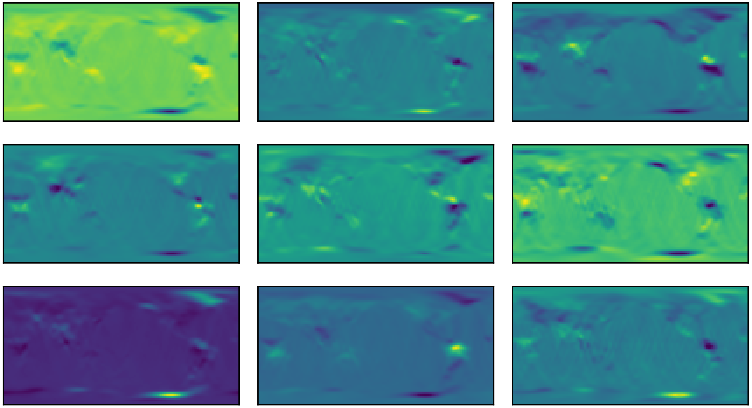
Dictionary Learning from Video Frames



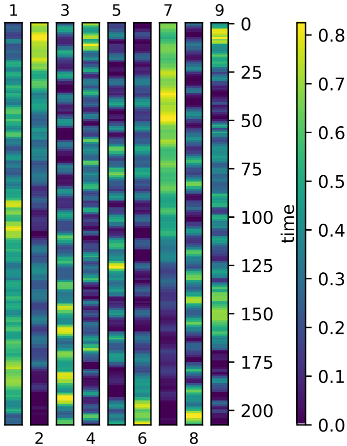
Dictionary Learning from Video Frames

- Denoising and predicting GRACE satellite data (with Keunsu Kim, Jinsu Kim, Jae-Hun Jung)
- $\mathbf{X} = (x \times y \times \text{month}) = (181 \times 361 \times 208)$
- Each time slice gives a heat-map of Earth's average monthly gravity potential measured by satellites

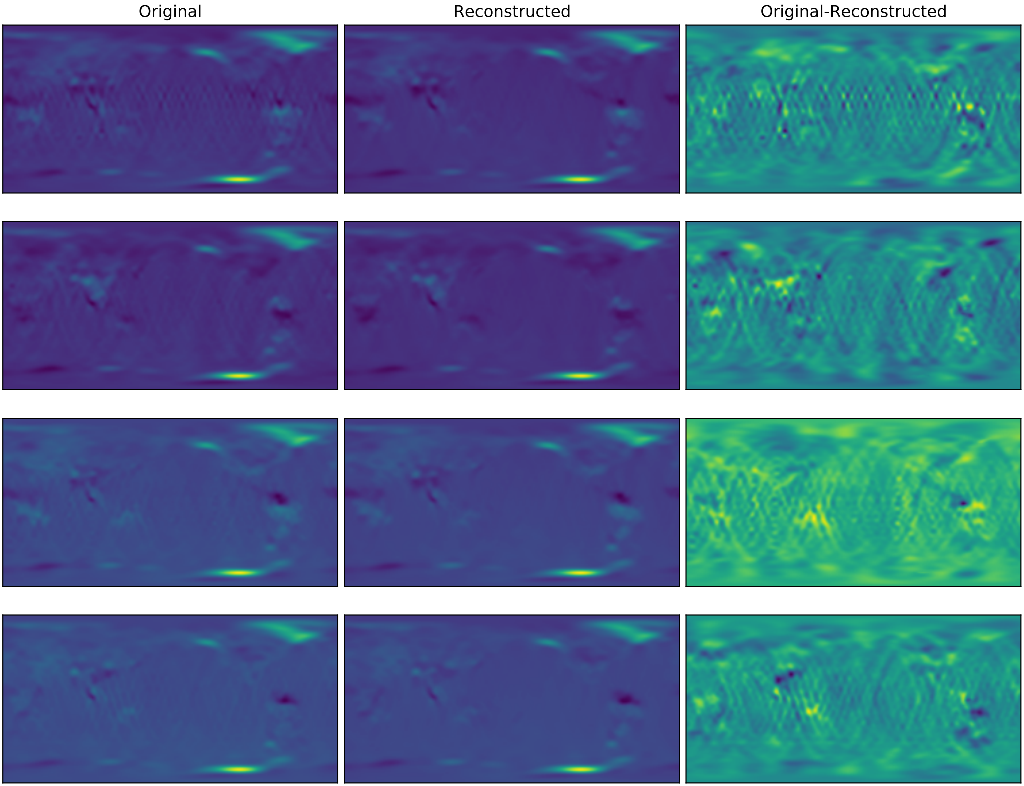
Spatial atoms



Time spectrum



Dictionary Learning from Video Frames



Dynamic topic modeling using NCPD for News Headlines

- $\mathbf{X} = \text{words} \times \text{time} \times \text{docs}$
- $\mathbf{U}^{(1)} = \text{words} \times \text{topic}$, $\mathbf{U}^{(2)} = \text{time} \times \text{topic}$, $\mathbf{U}^{(3)} = \text{docs} \times \text{topic}$

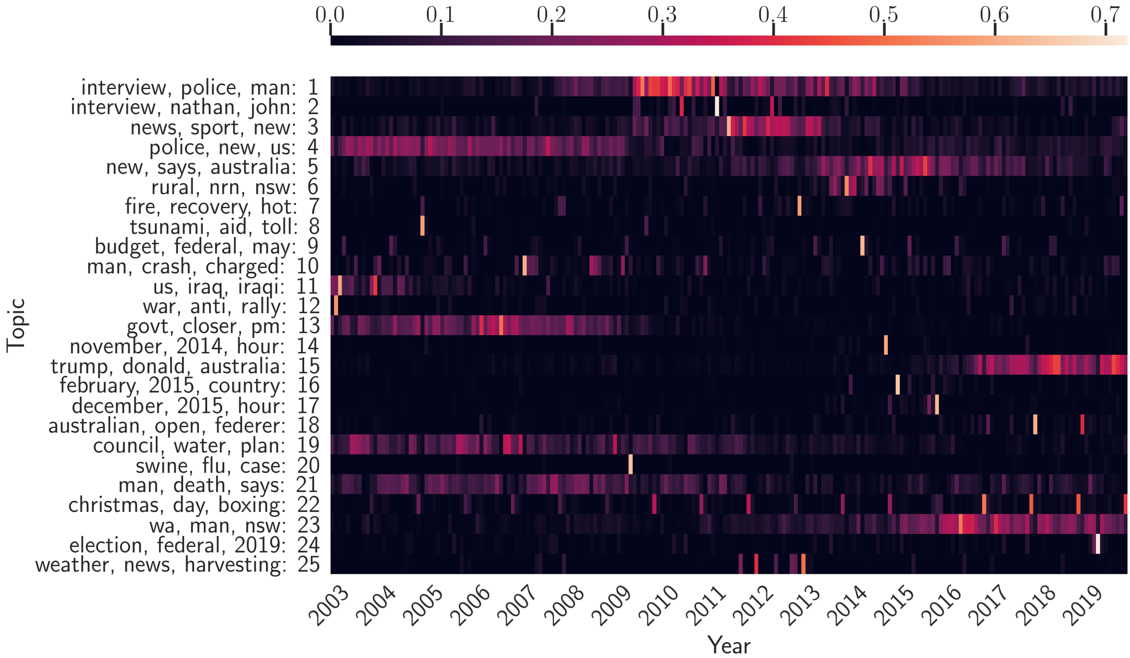


Figure: From (Kassab, Kryshchenko, L., Molitor, Needell, and Rebrova '21)

Dynamic topic modeling using NCPD for News Headlines

- $\mathbf{X}$ = words × time × docs
- $\mathbf{U}^{(1)}$ = words × topic, $\mathbf{U}^{(2)}$ = time × topic, $\mathbf{U}^{(3)}$ = docs × topic

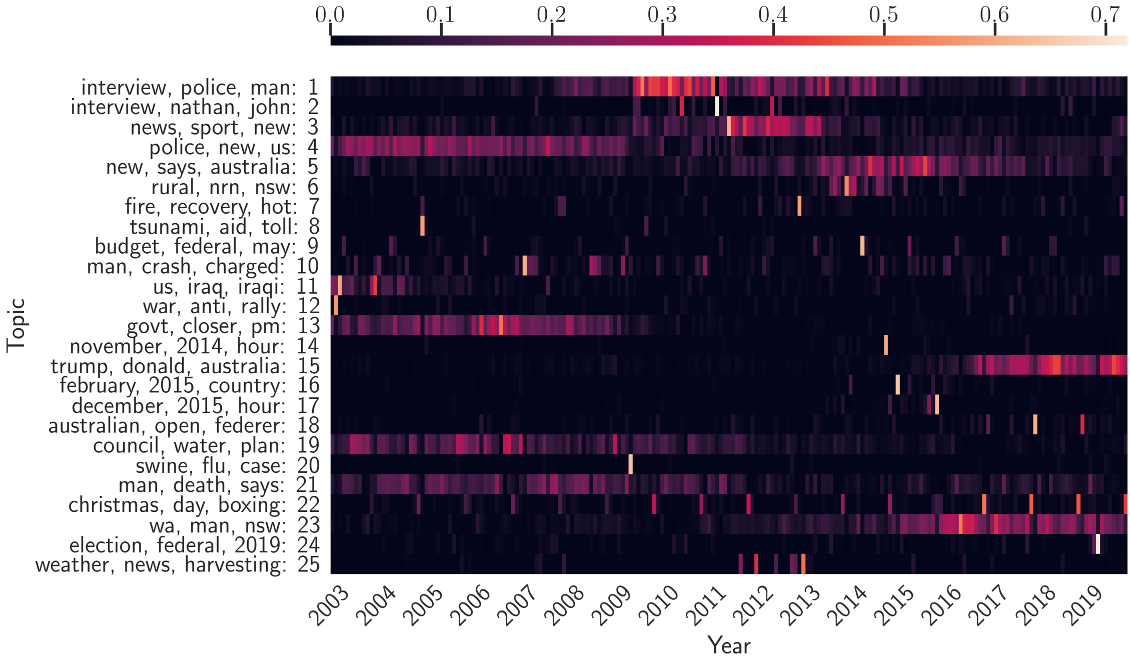


Figure: From (Kassab, Kryshchenko, L., Molitor, Needell, and Rebrova '21)

Outline

Supervised Dictionary Learning

- Given feature vectors $\mathbf{X}_{\text{data}} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ and binary labels $\mathbf{Y}_{\text{labels}} = [y_1, \ldots, y_n]$

Supervised Dictionary Learning

- Given feature vectors $\mathbf{X}_{\text{data}} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ and binary labels $\mathbf{Y}_{\text{labels}} = [y_1, \ldots, y_n]$
- Solve Classification and Dictionary learning (dimension reduction) at the same time

Supervised Dictionary Learning

- Given feature vectors $\mathbf{X}_{\text{data}} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ and binary labels $\mathbf{Y}_{\text{labels}} = [y_1, \ldots, y_n]$
- Solve Classification and Dictionary learning (dimension reduction) at the same time

$$\min_{\mathbf{W}, \mathbf{H}, \boldsymbol{\beta}} \quad L(\mathbf{W}, \mathbf{H}, \boldsymbol{\beta}) := \underbrace{\left(- \sum_{i=1}^{n} \sum_{j=0}^{1} \mathbf{1}(y_i = j) \log g_j(\langle \boldsymbol{\beta}, \mathbf{h}_i \rangle) \right)}_{\text{NLL of logistic regression}} + \overbrace{\xi}^{\text{tuning param.}} \underbrace{\|\mathbf{X}_{\text{data}} - \mathbf{WH}\|_F^2}_{\text{Reconstruction error}}$$

Supervised Dictionary Learning

- Given feature vectors $\mathbf{X}_{\text{data}} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ and binary labels $\mathbf{Y}_{\text{labels}} = [y_1, \ldots, y_n]$
- Solve Classification and Dictionary learning (dimension reduction) at the same time

$$\min_{\mathbf{W},\mathbf{H},\boldsymbol{\beta}} \quad L(\mathbf{W},\mathbf{H},\boldsymbol{\beta}) := \underbrace{\left(-\sum_{i=1}^{n}\sum_{j=0}^{1} \mathbf{1}(y_i = j) \log g_j(\langle \boldsymbol{\beta}, \mathbf{h}_i \rangle) \right)}_{\text{NLL of logistic regression}} + \overbrace{\xi}^{\text{tuning param.}} \underbrace{\|\mathbf{X}_{\text{data}} - \mathbf{WH}\|_F^2}_{\text{Reconstruction error}}$$

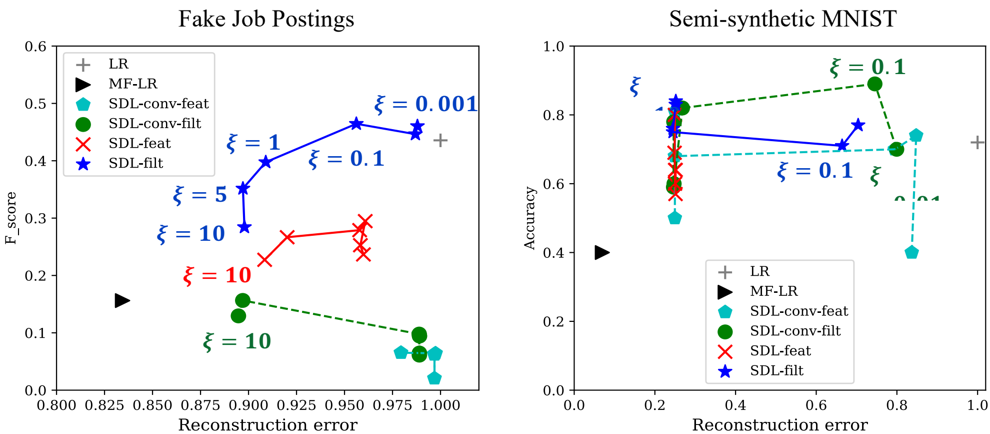


Figure: From Lee, L., Yao 2022+

Supervised Topic Modeling for imbalanced document classification

- Fake job postings dataset
  - $\mathbf{X}_{\text{data}} = \text{words} \times \text{postings} = (2{,}480 \times 17{,}880)$, $\mathbf{Y}_{\text{label}} \in \{0,1\}^{17{,}880}$
  - 95% are true, and 5% are fake postings (highly imbalanced)

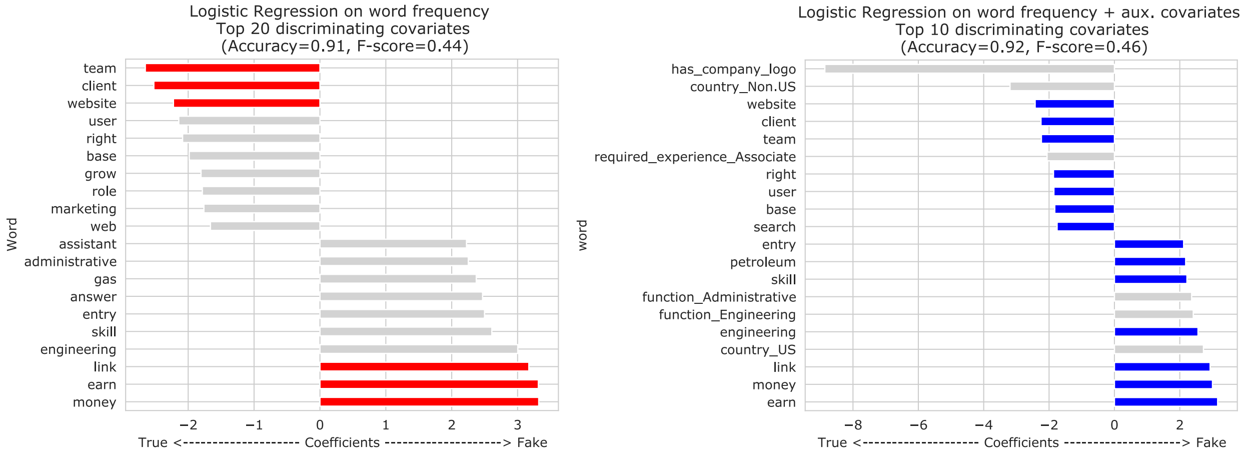


Figure: From Lee, L., Yao 2022+

Supervised Topic Modeling for imbalanced document classification

- Fake job postings dataset
  - $\mathbf{X}_{\text{data}} = \text{words} \times \text{postings} = (2,480 \times 17,880)$, $\mathbf{Y}_{\text{label}} \in \{0,1\}^{17,880}$
  - 95% are true, and 5% are fake postings (highly imbalanced)

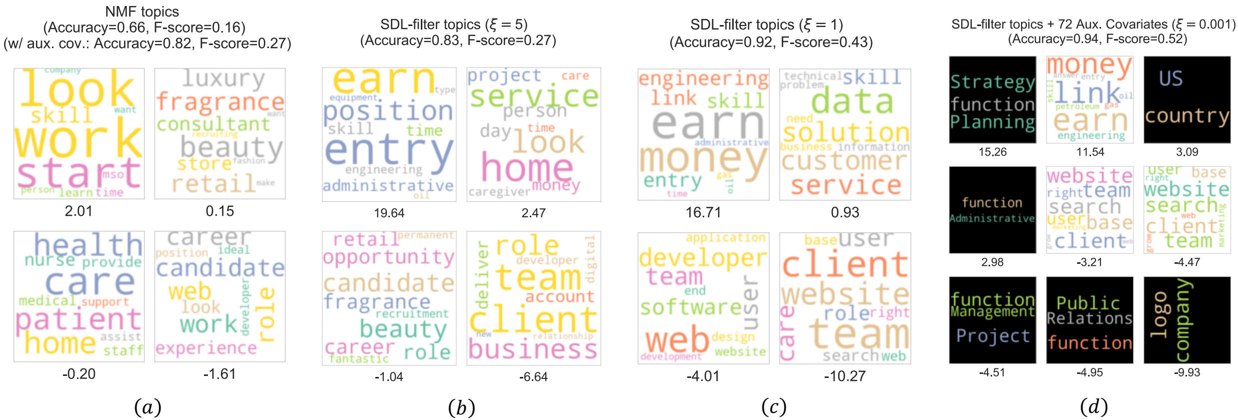


Figure: From Lee, L., Yao 2022+

Supervised Topic Modeling for imbalanced document classification

- Chest X-ray pneumonia dataset
  - $\mathbf{X}_{\text{data}} = \text{width} \times \text{height} \times \text{subjects} = (180 \times 180 \times 5,863)$, $\mathbf{Y}_{\text{label}} \in \{0,1\}^{5,863}$

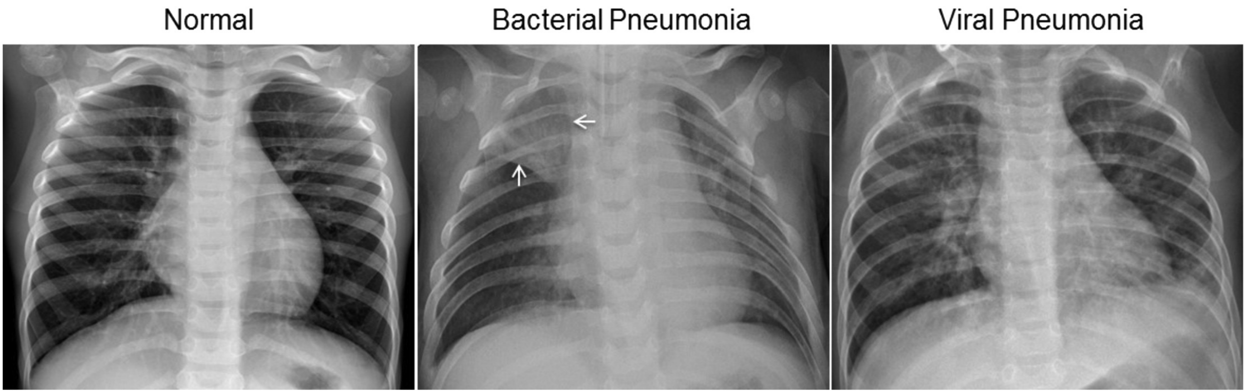


Figure: From Kermany et al. '18

Supervised Image Dictionary Learning for pneumonia detection

- Chest X-ray pneumonia dataset
  - $\mathbf{X}_{\text{data}} = \text{width} \times \text{height} \times \text{subjects} = (180 \times 180 \times 5{,}863)$, $\mathbf{Y}_{\text{label}} \in \{0,1\}^{5{,}863}$
  - Atoms with positive regression coefficient — Latent feature associated with pneumonia

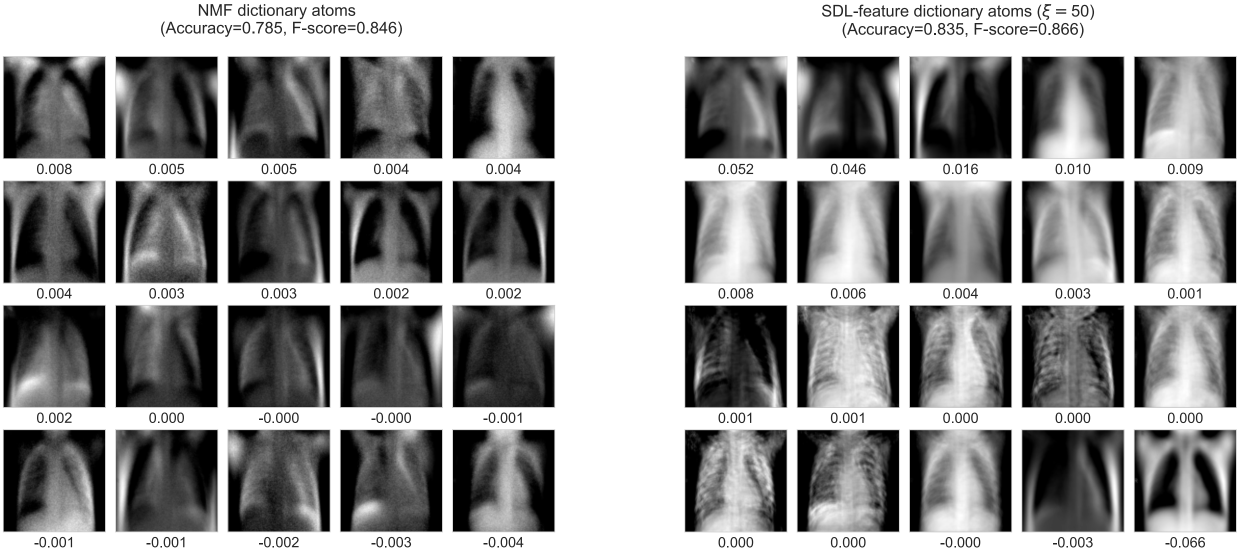


Figure: From Lee, L., Yao 2022+

Outline

Dictionary Learning with Subgraphs

- Given a large sparse network (e.g., Facebook social network), analyze the structure of random subgraphs

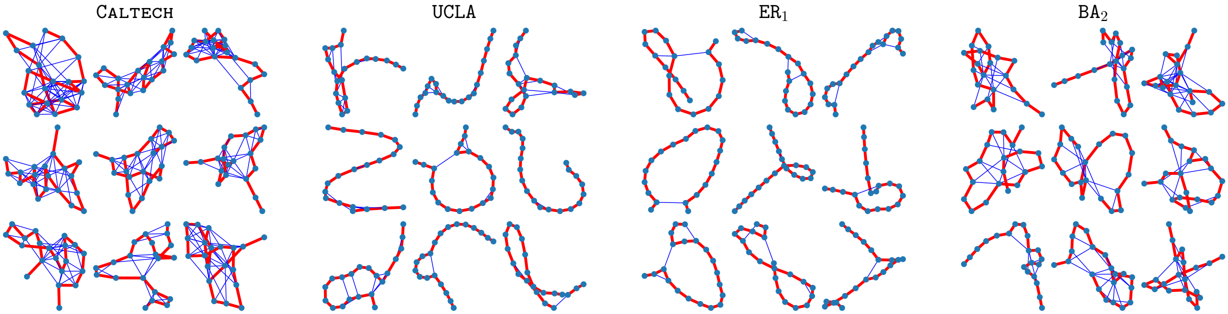


Figure: From L., Kureh, Vendrow, Porter '22+

Dictionary Learning with Subgraphs

- Given a large sparse network (e.g., Facebook social network), analyze the structure of random subgraphs

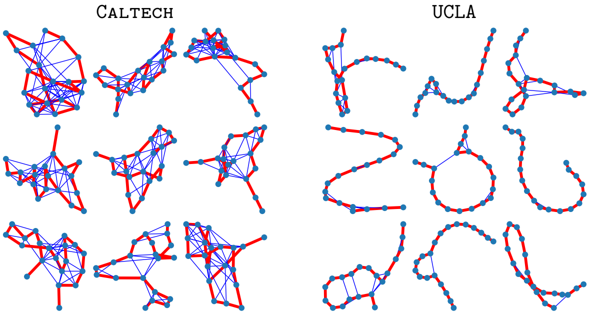


Figure: From L., Kureh, Vendrow, Porter '22+

- How do we sample subgraphs?

Dictionary Learning with Subgraphs

- Given a large sparse network (e.g., Facebook social network), analyze the structure of random subgraphs

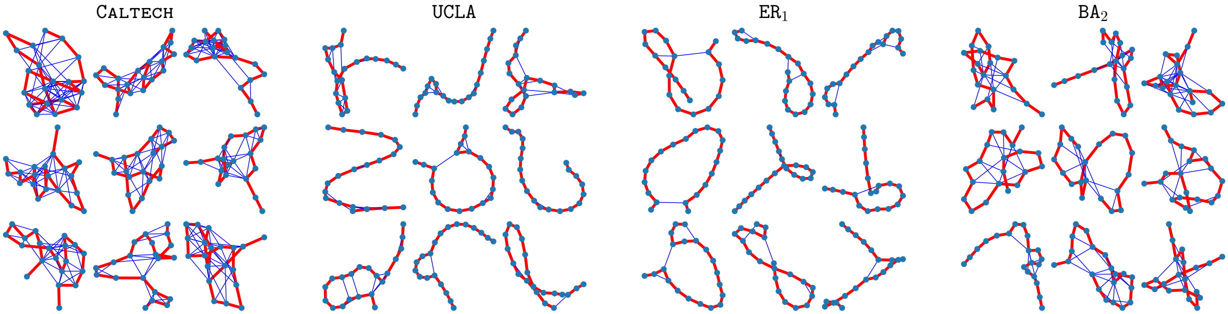


Figure: From L., Kureh, Vendrow, Porter '22+

- How do we sample subgraphs?
  - Sample a uniformly random k-path (red edges)

Dictionary Learning with Subgraphs

- Given a large sparse network (e.g., Facebook social network), analyze the structure of random subgraphs

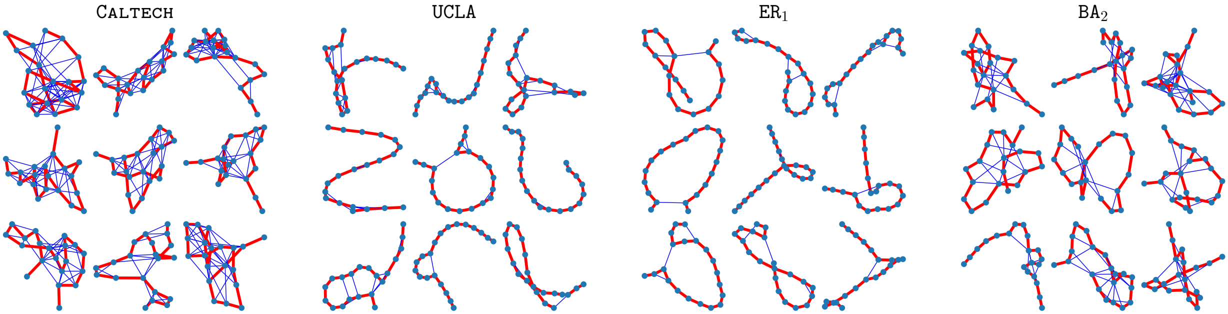


Figure: From L., Kureh, Vendrow, Porter '22+

- How do we sample subgraphs?
  - Sample a uniformly random k-path (red edges)
  — Use MCMC motif sampling by L. Memoli, Sivakoff '22

Dictionary Learning with Subgraphs

- Given a large sparse network (e.g., Facebook social network), analyze the structure of random subgraphs

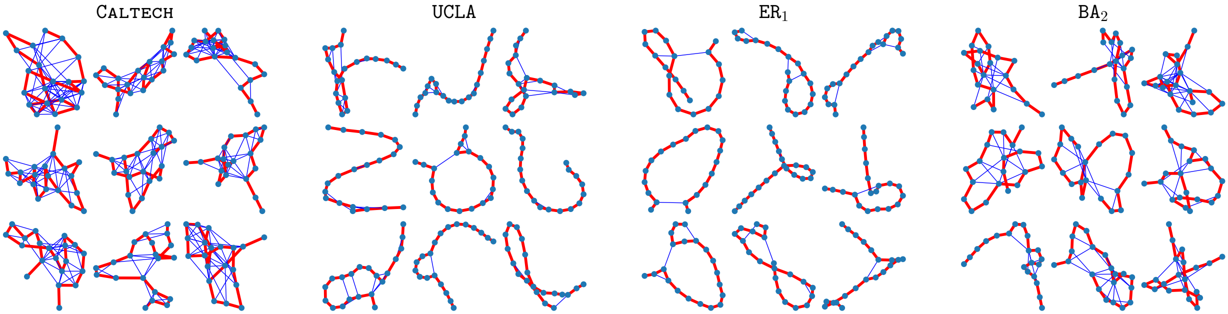


Figure: From L., Kureh, Vendrow, Porter '22+

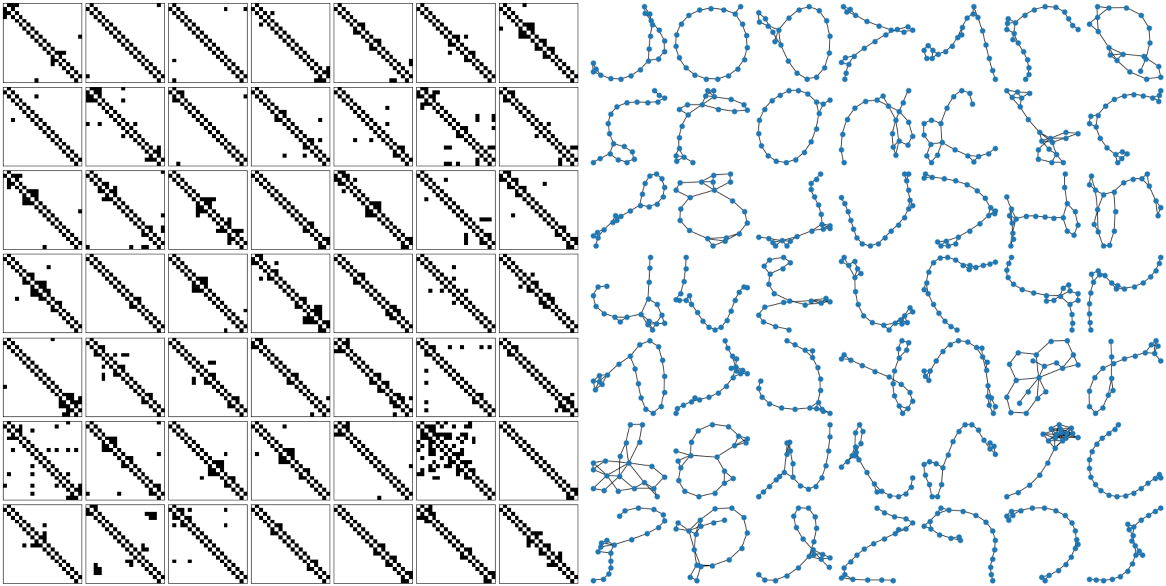
- How do we sample subgraphs?
  - Sample a uniformly random k-path (red edges)

    — Use MCMC motif sampling by L. Memoli, Sivakoff '22
  - Take the induced subgraph (blue edges)

Dictionary Learning with Network Subgraphs

- Sample 20-node subgraphs induced on 20-paths (seq. of 20 adjacent & distinct nodes)

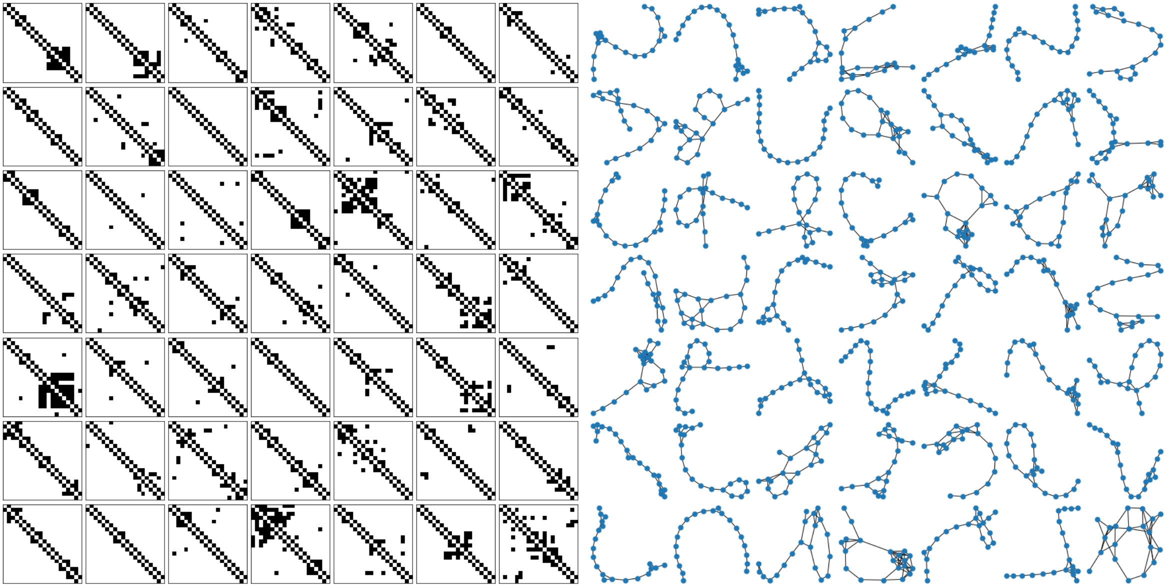
Induced subgraphs on 20-paths in Wisconsin



Dictionary Learning with Network Subgraphs

- Sample 20-node subgraphs induced on 20-paths (seq. of 20 adjacent & distinct nodes)

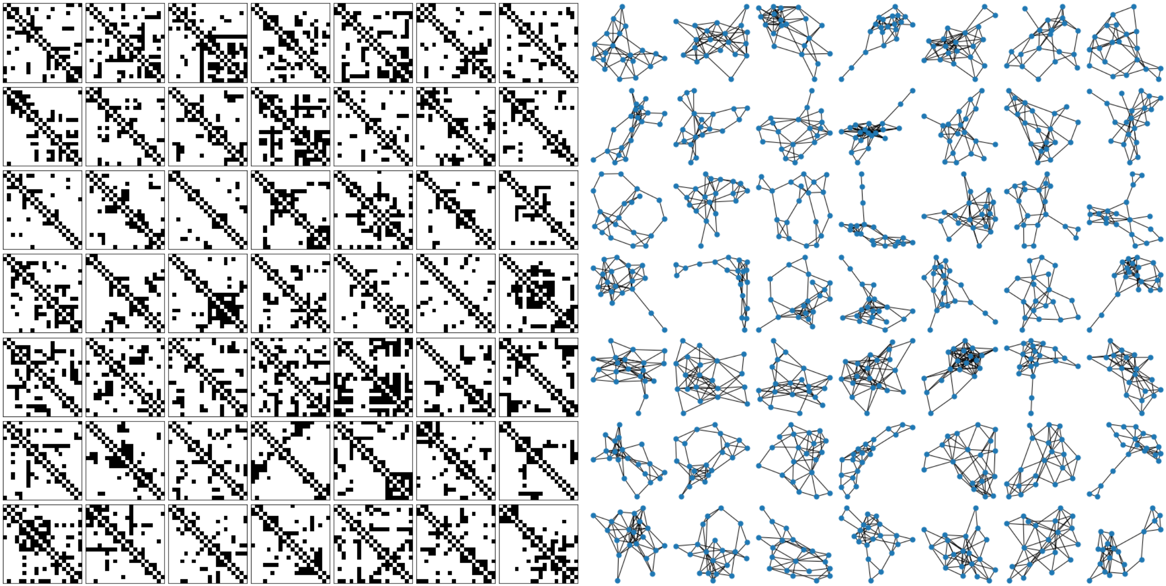
Induced subgraphs on 20-paths in UCLA



Dictionary Learning with Network Subgraphs

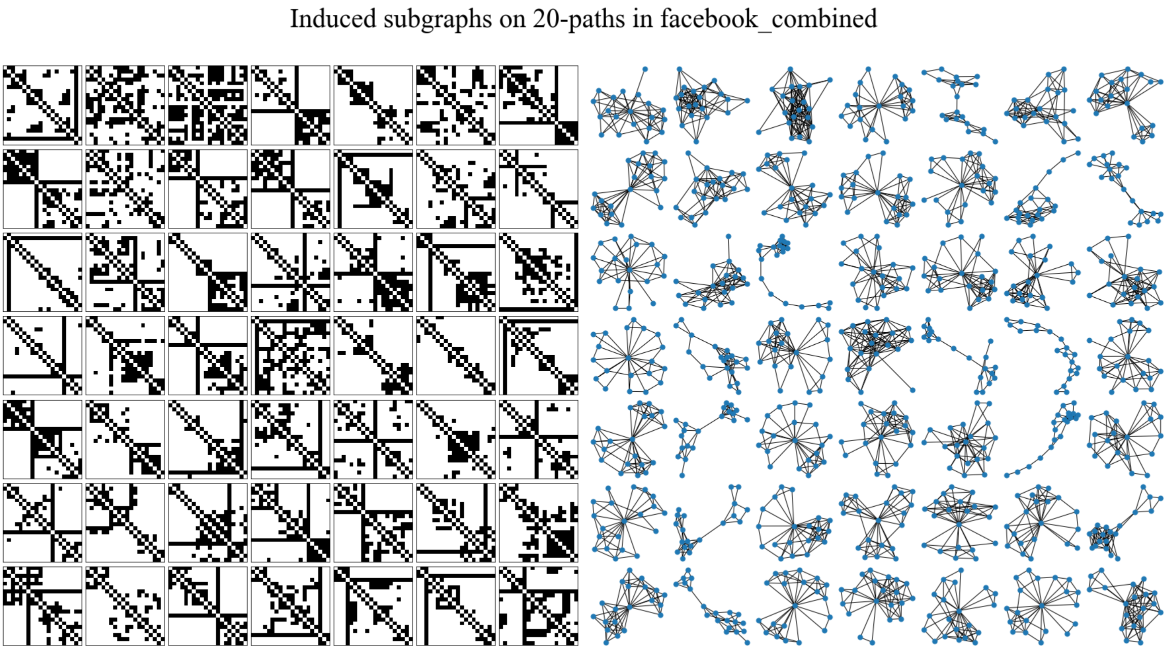
- Sample 20-node subgraphs induced on 20-paths (seq. of 20 adjacent & distinct nodes)

Induced subgraphs on 20-paths in Caltech



Dictionary Learning with Network Subgraphs

- Sample 20-node subgraphs induced on 20-paths (seq. of 20 adjacent & distinct nodes)



(a) arXiv

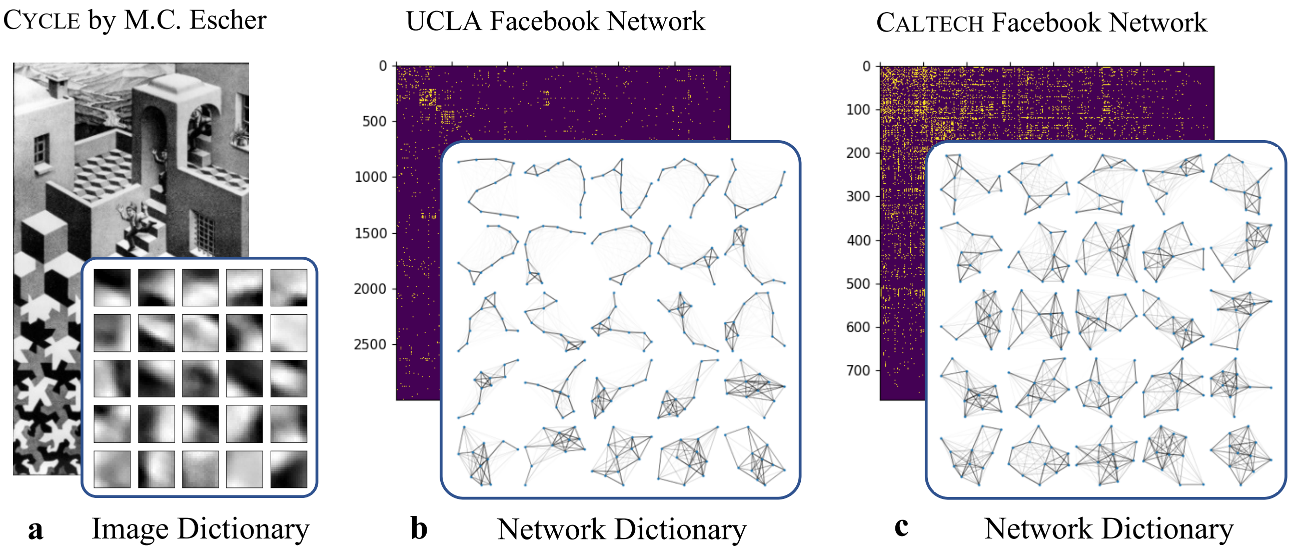
(b) Facebook

(c) Caltech

(d) UCLA

(e) UW-Madison

Network Dictionary Learning (NDL)



- NDL: Network data ⟶ Latent motifs (nonnegative basis for subgraphs)
  - First introduced in L., Needell, Balzano [4]
  - Further developed in L., Kureh, Vendrow, Porter [6]

Outline

Problem setup and BCD

- Problem setup:
  - (Multi-convex objective) $f : \mathbb{R}^{I_1} \times \cdots \times R^{I_m} \to [0, \infty)$ — Convex in each block
  - (Parameter space) $\boldsymbol{\Theta} := \Theta^{(1)} \times \cdots \times \Theta^{(m)} \subseteq \mathbb{R}^{I_1} \times \cdots \times \mathbb{R}^{I_m}$ — Product of convex sets
  - (Constrained nonconvex problem):

$$\min_{\boldsymbol{\theta}=[\theta_1,\ldots,\theta_m]\in\boldsymbol{\Theta}} f(\theta_1,\ldots,\theta_m).$$

  - Ex: NMF, NCPD, SDL, skip-gram, etc.

$$\text{(NMF)} \qquad \min_{\mathbf{W}\in\mathbb{R}^{p\times r}_{\geq 0},\mathbf{H}\in\mathbb{R}^{r\times n}_{\geq 0}} \left(f(\mathbf{W},\mathbf{H}) := \|\mathbf{X}-\mathbf{W}\mathbf{H}\|_F^2 \right)$$

Problem setup and BCD

- Problem setup:
  - (Multi-convex objective) $f:\mathbb{R}^{I_1}\times\cdots\times R^{I_m}\to[0,\infty)$ — Convex in each block
  - (Parameter space) $\boldsymbol{\Theta}:=\Theta^{(1)}\times\cdots\times\Theta^{(m)}\subseteq\mathbb{R}^{I_1}\times\cdots\times\mathbb{R}^{I_m}$ — Product of convex sets
  - (Constrained nonconvex problem):

$$\min_{\boldsymbol{\theta}=[\theta_1,\dots,\theta_m]\in\boldsymbol{\Theta}} f(\theta_1,\dots,\theta_m).$$

  - Ex: NMF, NCPD, SDL, skip-gram, etc.

$$\text{(NMF)}\qquad \min_{\mathbf{W}\in\mathbb{R}^{p\times r}_{\geq 0},\mathbf{H}\in\mathbb{R}^{r\times n}_{\geq 0}}\left(f(\mathbf{W},\mathbf{H}):=\|\mathbf{X}-\mathbf{W}\mathbf{H}\|_F^2\right)$$

- Block Coordinate Descent (BCD): For $n=1,\dots,N$ and for $i=1,\dots,m$:

$$\theta_n^{(i)}\in\operatorname*{argmin}_{\theta\in\Theta^{(i)}} f\left(\theta_n^{(1)},\cdots,\theta_n^{(i-1)},\theta,\theta_{n-1}^{(i+1)},\cdots,\theta_{n-1}^{(m)}\right).$$

Problem setup and BCD

- Problem setup:
  - (Multi-convex objective) $f : \mathbb{R}^{I_1} \times \cdots \times R^{I_m} \to [0,\infty)$ — Convex in each block
  - (Parameter space) $\mathbf{\Theta} := \Theta^{(1)} \times \cdots \times \Theta^{(m)} \subseteq \mathbb{R}^{I_1} \times \cdots \times \mathbb{R}^{I_m}$ — Product of convex sets
  - (Constrained nonconvex problem):

$$\min_{\boldsymbol{\theta}=[\theta_1,\ldots,\theta_m]\in\mathbf{\Theta}} f(\theta_1,\ldots,\theta_m).$$

  - Ex: NMF, NCPD, SDL, skip-gram, etc.

$$\text{(NMF)} \qquad \min_{\mathbf{W}\in\mathbb{R}^{p\times r}_{\geq 0},\, \mathbf{H}\in\mathbb{R}^{r\times n}_{\geq 0}} \left(f(\mathbf{W},\mathbf{H}) := \|\mathbf{X}-\mathbf{WH}\|_F^2 \right)$$

- Block Coordinate Descent (BCD): For $n = 1,\ldots,N$ and for $i = 1,\ldots,m$:

$$\theta_n^{(i)} \in \operatorname*{argmin}_{\theta\in\Theta^{(i)}} f\left(\theta_n^{(1)},\cdots,\theta_n^{(i-1)},\theta,\theta_{n-1}^{(i+1)},\cdots,\theta_{n-1}^{(m)}\right).$$

  - Sequentially update each block coordinate (by PGD) while fixing the rest

Block Coordinate Descent for Matrix/Tensor Factorization

- Nonnegative CP Decomposition (NCPD)

$$\min_{\mathbf{U}^{(1)}\in\mathbb{R}_{\geq 0}^{d_1\times r},\mathbf{U}^{(2)}\in\mathbb{R}_{\geq 0}^{d_2\times r},\mathbf{U}^{(3)}\in\mathbb{R}_{\geq 0}^{d_3\times r}} \|\mathbf{X}-\mathrm{Out}(\mathbf{U}^{(1)},\mathbf{U}^{(2)},\mathbf{U}^{(3)})\|_F^2$$

Block Coordinate Descent for Matrix/Tensor Factorization

- Nonnegative CP Decomposition (NCPD)

$$\min_{\mathbf{U}^{(1)}\in\mathbb{R}^{d_1\times r}_{\geq 0},\mathbf{U}^{(2)}\in\mathbb{R}^{d_2\times r}_{\geq 0},\mathbf{U}^{(3)}\in\mathbb{R}^{d_3\times r}_{\geq 0}} \|\mathbf{X}-\text{Out}(\mathbf{U}^{(1)},\mathbf{U}^{(2)},\mathbf{U}^{(3)})\|_F^2$$

  - Block Coordinate Descent (BCD) for NCPD (=Alternating Least Sqaures)

$$\begin{cases} \mathbf{U}_t^{(1)} \leftarrow \operatorname{argmin}_{\mathbf{U}\in\mathbb{R}^{d_1\times r}_{\geq 0}} \|\mathbf{X}-\text{Out}(\mathbf{U}, \mathbf{U}_{t-1}^{(2)}, \mathbf{U}_{t-1}^{(3)})\|_F^2 \\ \mathbf{U}_t^{(2)} \leftarrow \operatorname{argmin}_{\mathbf{U}\in\mathbb{R}^{d_2\times r}_{\geq 0}} \|\mathbf{X}-\text{Out}(\mathbf{U}_t^{(1)}, \mathbf{U}, \mathbf{U}_{t-1}^{(3)})\|_F^2 \\ \mathbf{U}_t^{(3)} \leftarrow \operatorname{argmin}_{\mathbf{U}\in\mathbb{R}^{d_3\times r}_{\geq 0}} \|\mathbf{X}-\text{Out}(\mathbf{U}_t^{(1)}, \mathbf{U}_t^{(2)}, \mathbf{U})\|_F^2 \end{cases}$$

Block Coordinate Descent for Matrix/Tensor Factorization

- Nonnegative CP Decomposition (NCPD)

$$\min_{\mathbf{U}^{(1)}\in\mathbb{R}_{\geq 0}^{d_1\times r},\mathbf{U}^{(2)}\in\mathbb{R}_{\geq 0}^{d_2\times r},\mathbf{U}^{(3)}\in\mathbb{R}_{\geq 0}^{d_3\times r}} \|\mathbf{X}-\text{Out}(\mathbf{U}^{(1)},\mathbf{U}^{(2)},\mathbf{U}^{(3)})\|_F^2$$

  - Block Coordinate Descent (BCD) for NCPD (=Alternating Least Sqaures)

$$\begin{cases} \mathbf{U}_t^{(1)} \leftarrow \operatorname{argmin}_{\mathbf{U}\in\mathbb{R}_{\geq 0}^{d_1\times r}} \|\mathbf{X}-\text{Out}(\mathbf{U},\mathbf{U}_{t-1}^{(2)},\mathbf{U}_{t-1}^{(3)})\|_F^2 \\ \mathbf{U}_t^{(2)} \leftarrow \operatorname{argmin}_{\mathbf{U}\in\mathbb{R}_{\geq 0}^{d_2\times r}} \|\mathbf{X}-\text{Out}(\mathbf{U}_t^{(1)},\mathbf{U},\mathbf{U}_{t-1}^{(3)})\|_F^2 \\ \mathbf{U}_t^{(3)} \leftarrow \operatorname{argmin}_{\mathbf{U}\in\mathbb{R}_{\geq 0}^{d_3\times r}} \|\mathbf{X}-\text{Out}(\mathbf{U}_t^{(1)},\mathbf{U}_t^{(2)},\mathbf{U})\|_F^2 \end{cases}$$

- BCD for multi-convex minimization problems:

Block Coordinate Descent for Matrix/Tensor Factorization

- Nonnegative CP Decomposition (NCPD)

$$\min_{\mathbf{U}^{(1)}\in\mathbb{R}_{\geq 0}^{d_1\times r},\mathbf{U}^{(2)}\in\mathbb{R}_{\geq 0}^{d_2\times r},\mathbf{U}^{(3)}\in\mathbb{R}_{\geq 0}^{d_3\times r}} \|\mathbf{X}-\text{Out}(\mathbf{U}^{(1)},\mathbf{U}^{(2)},\mathbf{U}^{(3)})\|_F^2$$

  - Block Coordinate Descent (BCD) for NCPD (=Alternating Least Sqaures)

$$\begin{cases} \mathbf{U}_t^{(1)} \leftarrow \operatorname{argmin}_{\mathbf{U}\in\mathbb{R}_{\geq 0}^{d_1\times r}} \|\mathbf{X}-\text{Out}(\mathbf{U},\mathbf{U}_{t-1}^{(2)},\mathbf{U}_{t-1}^{(3)})\|_F^2 \\ \mathbf{U}_t^{(2)} \leftarrow \operatorname{argmin}_{\mathbf{U}\in\mathbb{R}_{\geq 0}^{d_2\times r}} \|\mathbf{X}-\text{Out}(\mathbf{U}_t^{(1)},\mathbf{U},\mathbf{U}_{t-1}^{(3)})\|_F^2 \\ \mathbf{U}_t^{(3)} \leftarrow \operatorname{argmin}_{\mathbf{U}\in\mathbb{R}_{\geq 0}^{d_3\times r}} \|\mathbf{X}-\text{Out}(\mathbf{U}_t^{(1)},\mathbf{U}_t^{(2)},\mathbf{U})\|_F^2 \end{cases}$$

- BCD for multi-convex minimization problems:
  - No guarantee to find global minimizer

Block Coordinate Descent for Matrix/Tensor Factorization

- Nonnegative CP Decomposition (NCPD)

$$\min_{\mathbf{U}^{(1)}\in\mathbb{R}_{\geq 0}^{d_1\times r},\mathbf{U}^{(2)}\in\mathbb{R}_{\geq 0}^{d_2\times r},\mathbf{U}^{(3)}\in\mathbb{R}_{\geq 0}^{d_3\times r}} \|\mathbf{X}-\text{Out}(\mathbf{U}^{(1)},\mathbf{U}^{(2)},\mathbf{U}^{(3)})\|_F^2$$

  - Block Coordinate Descent (BCD) for NCPD (=Alternating Least Sqaures)

$$\begin{cases} \mathbf{U}_t^{(1)} \leftarrow \operatorname{argmin}_{\mathbf{U}\in\mathbb{R}_{\geq 0}^{d_1\times r}} \|\mathbf{X}-\text{Out}(\mathbf{U}, \mathbf{U}_{t-1}^{(2)}, \mathbf{U}_{t-1}^{(3)})\|_F^2 \\ \mathbf{U}_t^{(2)} \leftarrow \operatorname{argmin}_{\mathbf{U}\in\mathbb{R}_{\geq 0}^{d_2\times r}} \|\mathbf{X}-\text{Out}(\mathbf{U}_t^{(1)}, \mathbf{U}, \mathbf{U}_{t-1}^{(3)})\|_F^2 \\ \mathbf{U}_t^{(3)} \leftarrow \operatorname{argmin}_{\mathbf{U}\in\mathbb{R}_{\geq 0}^{d_3\times r}} \|\mathbf{X}-\text{Out}(\mathbf{U}_t^{(1)}, \mathbf{U}_t^{(2)}, \mathbf{U})\|_F^2 \end{cases}$$

- BCD for multi-convex minimization problems:
  - No guarantee to find global minimizer
  - Finds stationary points (local minimizers) under some condition

Block Coordinate Descent for Matrix/Tensor Factorization

- Nonnegative CP Decomposition (NCPD)

$$\min_{\mathbf{U}^{(1)}\in\mathbb{R}_{\geq 0}^{d_1\times r},\mathbf{U}^{(2)}\in\mathbb{R}_{\geq 0}^{d_2\times r},\mathbf{U}^{(3)}\in\mathbb{R}_{\geq 0}^{d_3\times r}} \|\mathbf{X}-\mathrm{Out}(\mathbf{U}^{(1)},\mathbf{U}^{(2)},\mathbf{U}^{(3)})\|_F^2$$

  - Block Coordinate Descent (BCD) for NCPD (=Alternating Least Sqaures)

$$\begin{cases} \mathbf{U}_t^{(1)} \leftarrow \operatorname{argmin}_{\mathbf{U}\in\mathbb{R}_{\geq 0}^{d_1\times r}} \|\mathbf{X}-\mathrm{Out}(\mathbf{U},\mathbf{U}_{t-1}^{(2)},\mathbf{U}_{t-1}^{(3)})\|_F^2 \\ \mathbf{U}_t^{(2)} \leftarrow \operatorname{argmin}_{\mathbf{U}\in\mathbb{R}_{\geq 0}^{d_2\times r}} \|\mathbf{X}-\mathrm{Out}(\mathbf{U}_t^{(1)},\mathbf{U},\mathbf{U}_{t-1}^{(3)})\|_F^2 \\ \mathbf{U}_t^{(3)} \leftarrow \operatorname{argmin}_{\mathbf{U}\in\mathbb{R}_{\geq 0}^{d_3\times r}} \|\mathbf{X}-\mathrm{Out}(\mathbf{U}_t^{(1)},\mathbf{U}_t^{(2)},\mathbf{U})\|_F^2 \end{cases}$$

- BCD for multi-convex minimization problems:
  - No guarantee to find global minimizer
  - Finds stationary points (local minimizers) under some condition
  - No known rate of convergence to stationary points

BCD with Proximal Regularization and Diminishing Radius

- BCD-PR (Proximal Regularization) : For $n = 1, \ldots, N$ and for $i = 1, \ldots, m$:

$$\theta_n^{(i)} \in \operatorname*{argmin}_{\theta \in \Theta^{(i)}} f\left(\theta_n^{(1)}, \cdots, \theta_n^{(i-1)}, \theta, \theta_{n-1}^{(i+1)}, \cdots, \theta_{n-1}^{(m)}\right) + \lambda_n \|\theta - \theta_{n-1}^{(i)}\|^2$$

BCD with Proximal Regularization and Diminishing Radius

- BCD-PR (Proximal Regularization) : For $n = 1, \dots, N$ and for $i = 1, \dots, m$:

$$\theta_n^{(i)} \in \operatorname*{argmin}_{\theta \in \Theta^{(i)}} f\left(\theta_n^{(1)}, \cdots, \theta_n^{(i-1)}, \theta, \theta_{n-1}^{(i+1)}, \cdots, \theta_{n-1}^{(m)}\right) + \lambda_n \|\theta - \theta_{n-1}^{(i)}\|^2$$

- BCD-DR (Diminishing Radius) : For $n = 1, \dots, N$ and for $i = 1, \dots, m$:

$$\theta_n^{(i)} \in \operatorname*{argmin}_{\theta \in \Theta^{(i)}, \, \|\theta - \theta_{n-1}^{(i)}\| \le r_n} f\left(\theta_n^{(1)}, \cdots, \theta_n^{(i-1)}, \theta, \theta_{n-1}^{(i+1)}, \cdots, \theta_{n-1}^{(m)}\right)$$

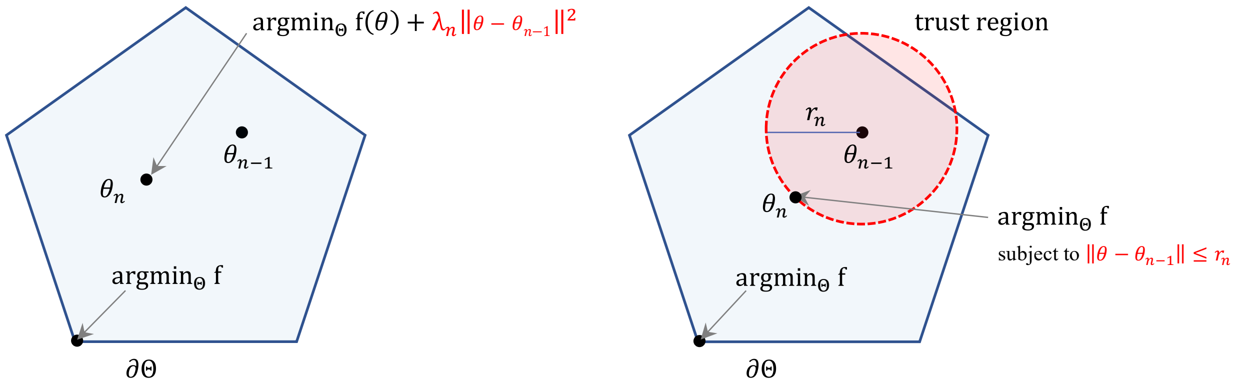
BCD with Proximal Regularization and Diminishing Radius

- BCD-PR (Proximal Regularization) : For $n = 1, \ldots, N$ and for $i = 1, \ldots, m$:

$$\theta_n^{(i)} \in \operatorname*{argmin}_{\theta \in \Theta^{(i)}} f\left(\theta_n^{(1)}, \cdots, \theta_n^{(i-1)}, \theta, \theta_{n-1}^{(i+1)}, \cdots, \theta_{n-1}^{(m)}\right) + \lambda_n \|\theta - \theta_{n-1}^{(i)}\|^2$$

- BCD-DR (Diminishing Radius) : For $n = 1, \ldots, N$ and for $i = 1, \ldots, m$:

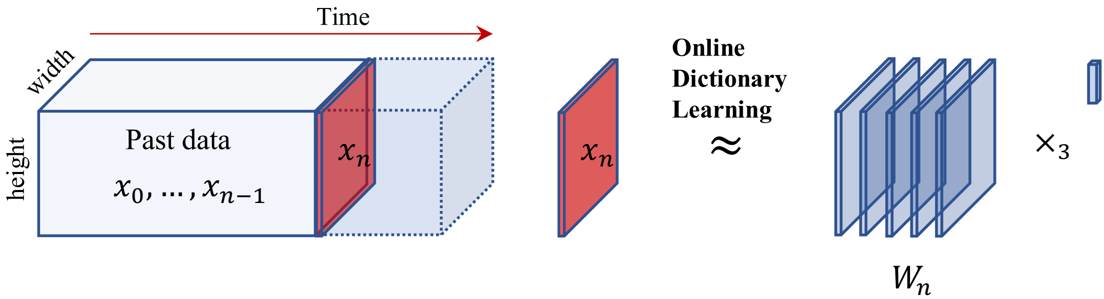
$$\theta_n^{(i)} \in \operatorname*{argmin}_{\theta \in \Theta^{(i)},\, \|\theta - \theta_{n-1}^{(i)}\| \le r_n} f\left(\theta_n^{(1)}, \cdots, \theta_n^{(i-1)}, \theta, \theta_{n-1}^{(i+1)}, \cdots, \theta_{n-1}^{(m)}\right)$$



Outline

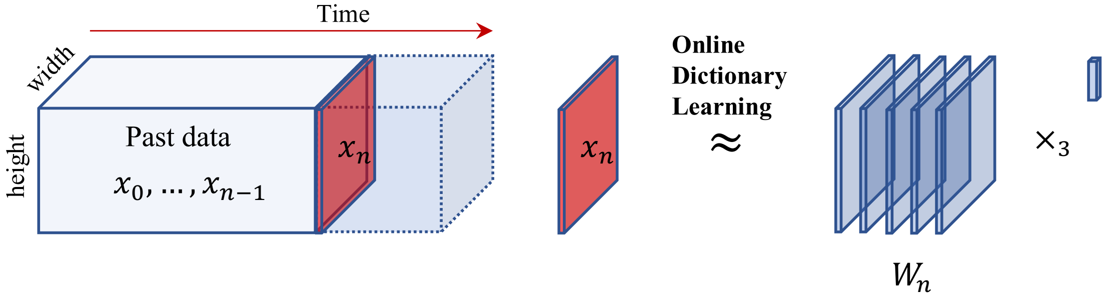
Online Dictionary Learning

- Instead of processing the entire frames at once, can we process one image at a time to learn the dictionary? (mini-batch processing)



Online Dictionary Learning

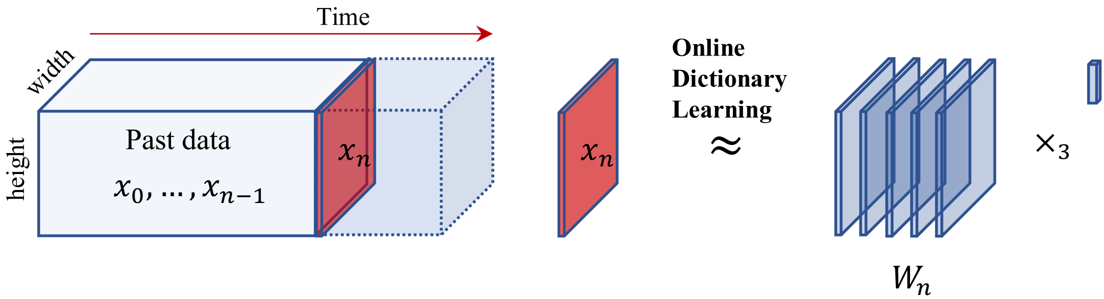
- Instead of processing the entire frames at once, can we process one image at a time to learn the dictionary? (mini-batch processing)



- Why do 'online learning'?

Online Dictionary Learning

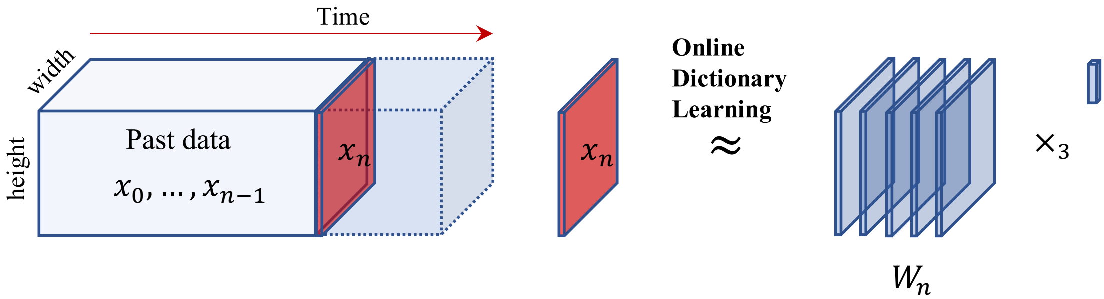
- Instead of processing the entire frames at once, can we process one image at a time to learn the dictionary? (mini-batch processing)



- Why do 'online learning'?
  - Reduced per-iteration computational cost

Online Dictionary Learning

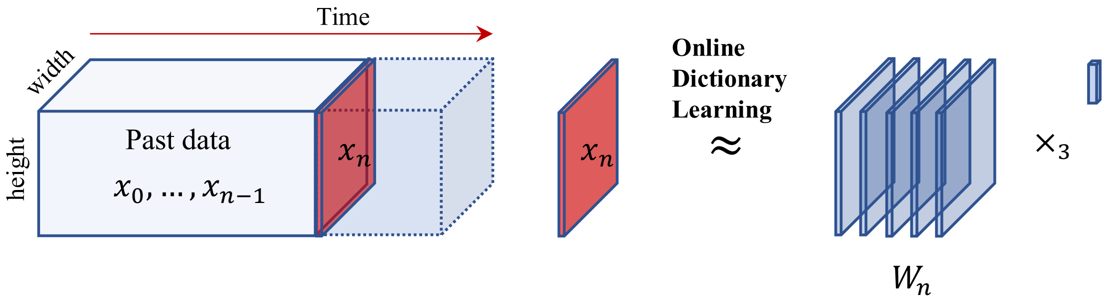
- Instead of processing the entire frames at once, can we process one image at a time to learn the dictionary? (mini-batch processing)



- Why do 'online learning'?
  - Reduced per-iteration computational cost
  - Reduced memory requirement (no need to hold the entire data)

Online Dictionary Learning

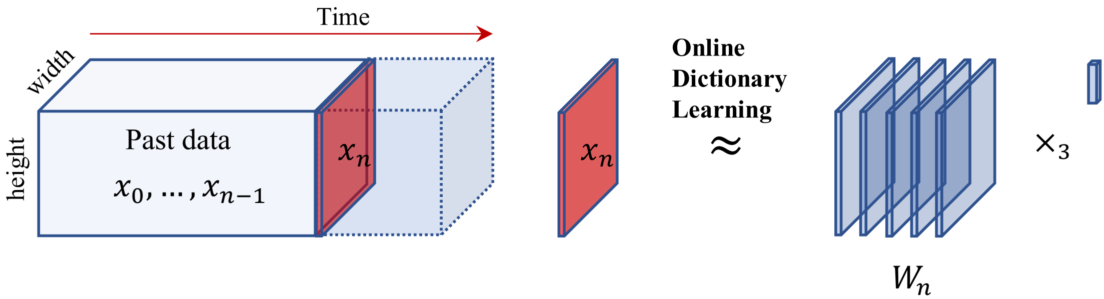
- Instead of processing the entire frames at once, can we process one image at a time to learn the dictionary? (mini-batch processing)



- Why do 'online learning'?
  - Reduced per-iteration computational cost
  - Reduced memory requirement (no need to hold the entire data)
  - Full data may not be available

Online Dictionary Learning

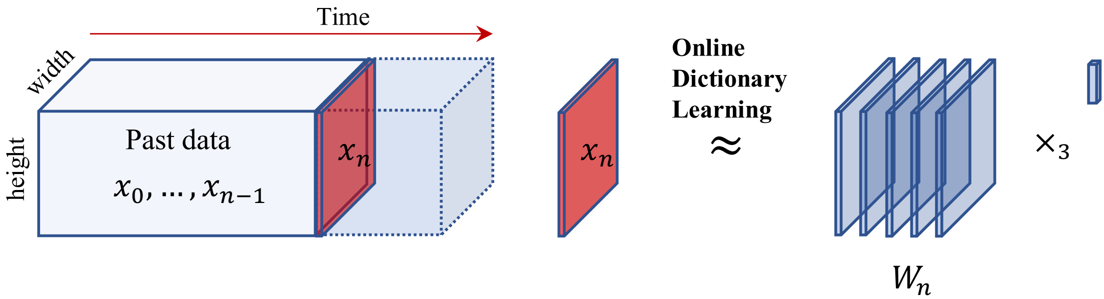
- Instead of processing the entire frames at once, can we process one image at a time to learn the dictionary? (mini-batch processing)



- Why do 'online learning'?
  - Reduced per-iteration computational cost
  - Reduced memory requirement (no need to hold the entire data)
  - Full data may not be available
  - May learn additional temporal features

Online Dictionary Learning

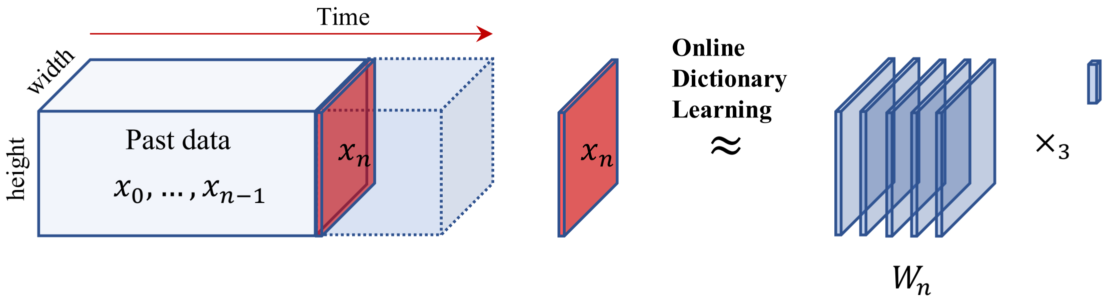
- Instead of processing the entire frames at once, can we process one image at a time to learn the dictionary? (mini-batch processing)



- Why do 'online learning'?
  - Reduced per-iteration computational cost
  - Reduced memory requirement (no need to hold the entire data)
  - Full data may not be available
  - May learn additional temporal features
  - May learn new trending features

Online Dictionary Learning

- Instead of processing the entire frames at once, can we process one image at a time to learn the dictionary? (mini-batch processing)



- Why do 'online learning'?
  - Reduced per-iteration computational cost
  - Reduced memory requirement (no need to hold the entire data)
  - Full data may not be available
  - May learn additional temporal features
  - May learn new trending features
- Algorithms: Stochastic GD, Stochastic PGD, Stochastic MM, etc.

Empirical Loss Minimization

- Empirical Loss Minimization

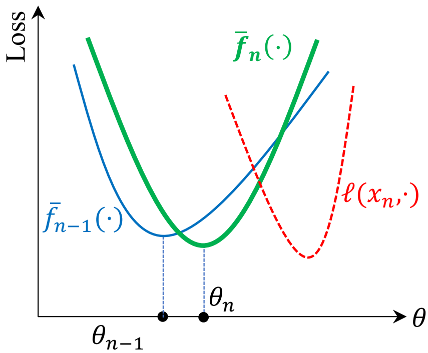
$$\text{Upon arrival of } \mathbf{x}_n: \quad \boldsymbol{\theta}_n \in \operatorname*{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \Big(\bar{f}_n(\boldsymbol{\theta}) := (1-w_n)\underbrace{\bar{f}_{n-1}(\boldsymbol{\theta})}_{\text{old loss}} + w_n \underbrace{\ell(\mathbf{x}_n,\boldsymbol{\theta})}_{\text{new loss}} \Big),$$

Empirical Loss Minimization

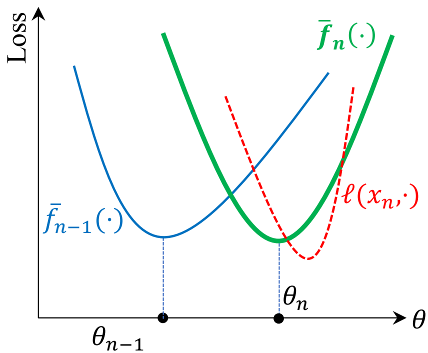
- Empirical Loss Minimization

$$\text{Upon arrival of } \mathbf{x}_n: \quad \boldsymbol{\theta}_n \in \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmin}} \Big(\bar{f}_n(\boldsymbol{\theta}) := (1-w_n) \underbrace{\bar{f}_{n-1}(\boldsymbol{\theta})}_{\text{old loss}} + w_n \underbrace{\ell(\mathbf{x}_n, \boldsymbol{\theta})}_{\text{new loss}} \Big),$$

- Depending on the data sequence $(\mathbf{x}_n)_{n\geq 1}$ and adaptivity weights $(w_n)_{n\geq 1}$, the optimization landscape $\bar{f}_n$ changes over time



Slow adaptation



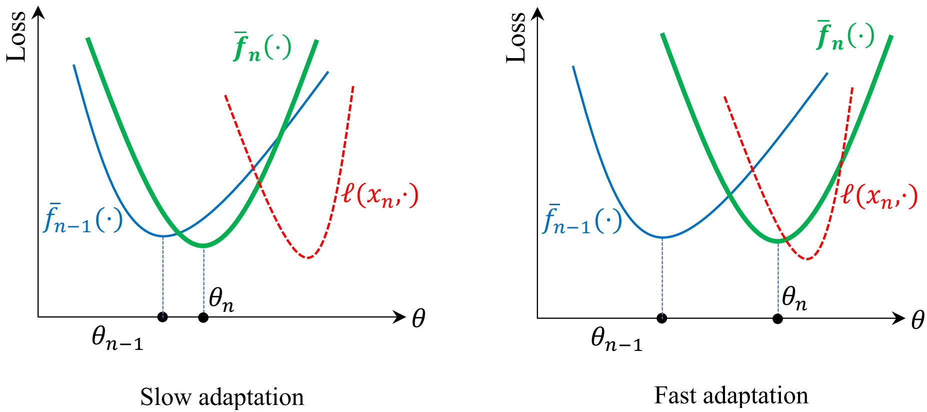
Fast adaptation

Empiricial Loss Minimization

- Empirical Loss Minimization

$$\text{Upon arrival of } \mathbf{x}_n: \quad \boldsymbol{\theta}_n \in \operatorname*{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \Big(\bar{f}_n(\boldsymbol{\theta}) := (1-w_n) \underbrace{\bar{f}_{n-1}(\boldsymbol{\theta})}_{\text{old loss}} + w_n \underbrace{\ell(\mathbf{x}_n, \boldsymbol{\theta})}_{\text{new loss}} \Big),$$

- Depending on the data sequence $(\mathbf{x}_n)_{n\geq 1}$ and adaptivity weights $(w_n)_{n\geq 1}$, the optimization landscape $\bar{f}_n$ changes over time
  - Fast-adapting $w_n \Rightarrow$ learn short-time scale features (could be noisy)

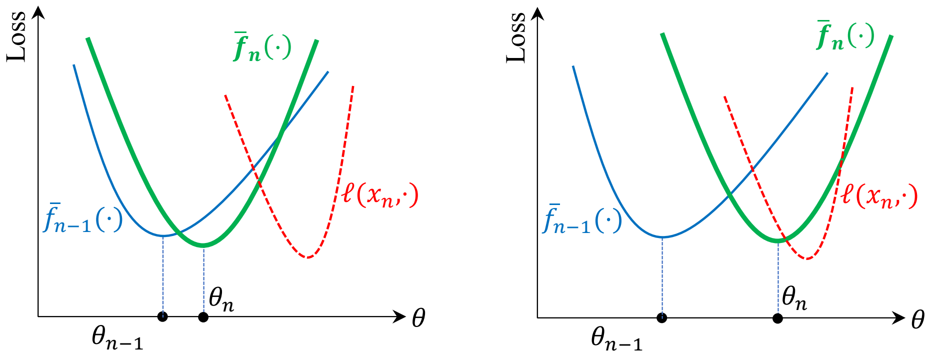


Empirical Loss Minimization

- Empirical Loss Minimization

$$\text{Upon arrival of } \mathbf{x}_n: \quad \boldsymbol{\theta}_n \in \operatorname*{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \Big(\bar{f}_n(\boldsymbol{\theta}) := (1-w_n)\underbrace{\bar{f}_{n-1}(\boldsymbol{\theta})}_{\text{old loss}} + w_n \underbrace{\ell(\mathbf{x}_n,\boldsymbol{\theta})}_{\text{new loss}} \Big),$$

- Depending on the data sequence $(\mathbf{x}_n)_{n\geq 1}$ and adaptivity weights $(w_n)_{n\geq 1}$, the optimization landscape $\bar{f}_n$ changes over time
  - Fast-adapting w_n ⇒ learn short-time scale features (could be noisy)
  - Slow-adapting w_n ⇒ learn long-time scale features (could be smoothed out too much)



Slow adaptation   Fast adaptation

(a) past2future + fast adaptation

(b) past2future + slow adaptation

So how do we solve empirical loss minimization?

- When $\boldsymbol{\theta} \mapsto \ell(\mathbf{x}_n, \boldsymbol{\theta})$ is convex, empirical loss $\bar{f}_n$ is convex for $n \geq 1$.

So how do we solve empirical loss minimization?

- When $\boldsymbol{\theta} \mapsto \ell(\mathbf{x}_n, \boldsymbol{\theta})$ is convex, empirical loss $\bar{f}_n$ is convex for $n \geq 1$.
- But many interesting problems assume nonconvex loss ℓ:

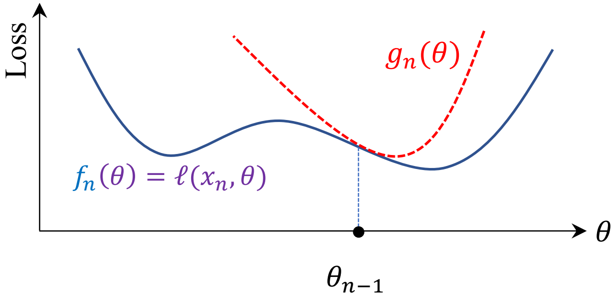
$$\text{(Dictionary Learning)} \qquad \ell(\mathbf{x}_n, \boldsymbol{\theta}) = \inf_{H} \|\mathbf{x}_n - \boldsymbol{\theta} H\|^2$$

So how do we solve empirical loss minimization?

- When $\boldsymbol{\theta} \mapsto \ell(\mathbf{x}_n, \boldsymbol{\theta})$ is convex, empirical loss $\bar{f}_n$ is convex for $n \geq 1$.
- But many interesting problems assume nonconvex loss ℓ:

$$\text{(Dictionary Learning)} \qquad \ell(\mathbf{x}_n, \boldsymbol{\theta}) = \inf_H \|\mathbf{x}_n - \boldsymbol{\theta} H\|^2$$

- Majorization-Minimization: Minimize a majorizing surrogate g_n of $\boldsymbol{\theta} \mapsto \ell(\mathbf{x}_n, \boldsymbol{\theta})$:

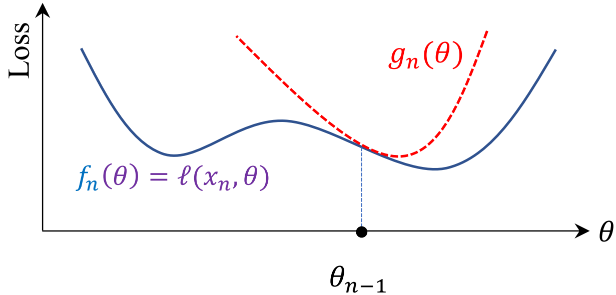


So how do we solve empirical loss minimization?

- When $\boldsymbol{\theta} \mapsto \ell(\mathbf{x}_n, \boldsymbol{\theta})$ is convex, empirical loss $\bar{f}_n$ is convex for $n \geq 1$.
- But many interesting problems assume nonconvex loss ℓ:

$$\text{(Dictionary Learning)} \qquad \ell(\mathbf{x}_n, \boldsymbol{\theta}) = \inf_{H} \|\mathbf{x}_n - \boldsymbol{\theta} H\|^2$$

- Majorization-Minimization: Minimize a majorizing surrogate g_n of $\boldsymbol{\theta} \mapsto \ell(\mathbf{x}_n, \boldsymbol{\theta})$:



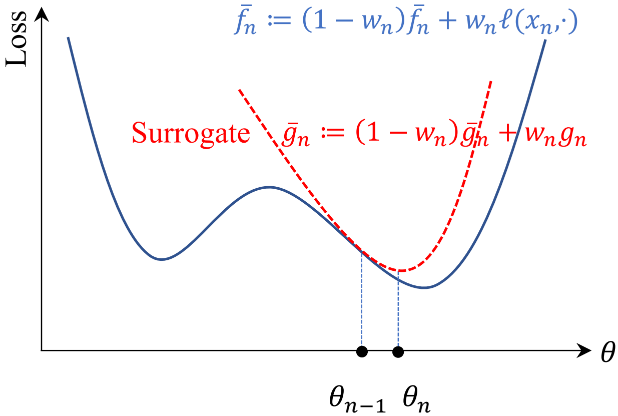
  - Ex: Gradient descent — Assuming ∇f_n is L-Lipschitz,

$$\boldsymbol{\theta}_n \in \operatorname*{argmin}_{\boldsymbol{\theta}} \Big(\underbrace{f_n(\boldsymbol{\theta}) + \langle \nabla f_n(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta} - \boldsymbol{\theta}_{n-1} \rangle + \frac{L}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\|^2}_{\text{quadratic surrogate of } f_n \text{ at } \boldsymbol{\theta}_{n-1}} \Big) \iff \boldsymbol{\theta}_n \leftarrow \boldsymbol{\theta}_{n-1} - \frac{1}{L}\nabla f_n(\boldsymbol{\theta}_{n-1})$$

Stochastic Majorization-Minimization

- Stochastic MM (SMM) — Sampling + MM + Recursive averaging

$$
(\textbf{SMM}) \quad \begin{cases} \text{Sample } \mathbf{x}_n \sim \pi(\cdot \,|\, \mathbf{x}_1, \dots, \mathbf{x}_{n-1}) \; ; \\ g_n \leftarrow \text{Strongly convex majorizing surrogate of } f_n(\cdot) = \ell(\mathbf{x}_n, \cdot); \\ \boldsymbol{\theta}_n \in \operatorname{argmin}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left(\bar{g}_n(\boldsymbol{\theta}) := (1 - w_n) \underbrace{\bar{g}_{n-1}(\boldsymbol{\theta})}_{\text{old avgd surr.}} + w_n \underbrace{g_n(\boldsymbol{\theta})}_{\text{new surr.}} \right). \end{cases}
$$



Examples of SMM

- Stochastic Gradient Descent (Proximal Gradient Mapping in the constrained case)

$$
\textbf{(SGD)} \quad \begin{cases} \text{Sample } \mathbf{x}_n \sim \pi(\cdot \,|\, \mathbf{x}_1, \ldots, \mathbf{x}_{n-1}) \ ; \\ g_n(\boldsymbol{\theta}) \leftarrow f_n(\boldsymbol{\theta}) + \langle \nabla f_n(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta} - \boldsymbol{\theta}_{n-1} \rangle + \frac{L}{2} \|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\|^2 \\ \boldsymbol{\theta}_n \in \operatorname{argmin}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left(\bar{g}_n(\boldsymbol{\theta}) := (1 - 1) \bar{g}_{n-1}(\boldsymbol{\theta}) + 1 \, g_n(\boldsymbol{\theta}) \right). \end{cases}
$$

Examples of SMM

- Stochastic Gradient Descent (Proximal Gradient Mapping in the constrained case)

$$
\textbf{(SGD)} \quad \begin{cases} \text{Sample } \mathbf{x}_n \sim \pi(\cdot \,|\, \mathbf{x}_1, \dots, \mathbf{x}_{n-1}) \; ; \\ g_n(\boldsymbol{\theta}) \leftarrow f_n(\boldsymbol{\theta}) + \langle \nabla f_n(\boldsymbol{\theta}_{n-1}), \boldsymbol{\theta} - \boldsymbol{\theta}_{n-1} \rangle + \frac{L}{2} \|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\|^2 \\ \boldsymbol{\theta}_n \in \operatorname{argmin}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left(\bar{g}_n(\boldsymbol{\theta}) := (1 - 1) \bar{g}_{n-1}(\boldsymbol{\theta}) + 1 \, g_n(\boldsymbol{\theta}) \right). \end{cases}
$$

- Online Matrix Factorization in Mairal et al. (2010), Mensch et al. (2017), Lyu et al. (2020):

$$
\textbf{(OMF)} \quad \begin{cases} \text{Sample } \mathbf{x}_n \in \mathbb{R}^d \sim \pi(\cdot \,|\, \mathbf{x}_1, \dots, \mathbf{x}_{n-1}) \; ; \\ H_n \leftarrow \operatorname{argmin}_H \|\mathbf{x}_n - \underbrace{\boldsymbol{\theta}_{n-1}}_{\text{old dict.}} H\|_F^2 \\ \boldsymbol{\theta}_n \in \operatorname{argmin}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left(\bar{g}_n(\boldsymbol{\theta}) := (1 - w_n) \underbrace{\bar{g}_{n-1}(\boldsymbol{\theta})}_{\text{old avgd surr.}} + w_n \underbrace{\|\mathbf{x}_n - \boldsymbol{\theta} H_n\|_F^2}_{\text{new surr.}} \right). \end{cases}
$$

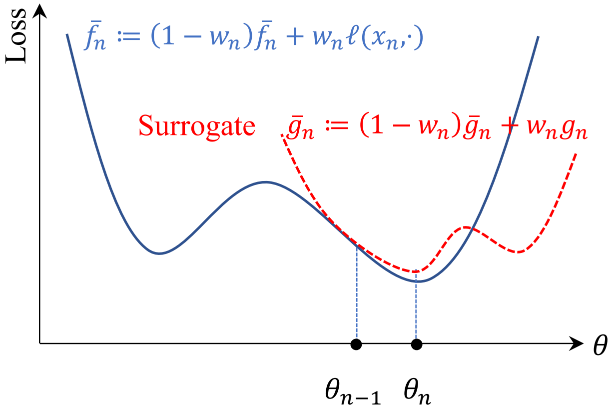
Stochastic (Block) Majorization-Minimization

- What if we can't find strongly convex surrogate? — e.g., Online CP Tensor Decomposition

Stochastic (Block) Majorization-Minimization

- What if we can't find strongly convex surrogate? — e.g., Online CP Tensor Decomposition
- Stochastic Block MM — SMM + block multi-convex surrogates + Diminishing Radius (Lyu '22 [3], '20 [2])

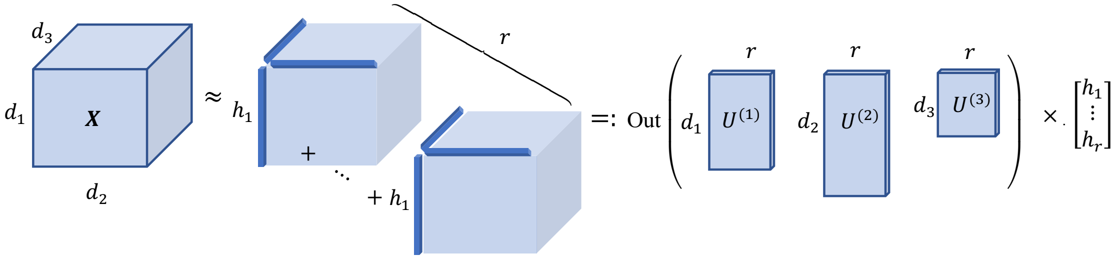
$$
(\textbf{SBMM}) \quad \begin{cases} \text{Sample } \mathbf{x}_n \sim \pi(\cdot \,|\, \mathbf{x}_1, \dots, \mathbf{x}_{n-1}) \; ; \\ g_n \leftarrow \text{Block multi-convex majorizing surrogate of } f_n(\cdot) = \ell(\mathbf{x}_n, \cdot); \\ \boldsymbol{\theta}_n \approx \operatorname{argmin}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left(\bar{g}_n(\boldsymbol{\theta}) := (1 - w_n)\bar{g}_{n-1}(\boldsymbol{\theta}) + w_n \, g_n(\boldsymbol{\theta}) \right) \\ \qquad \text{subject to } \|\boldsymbol{\theta} - \boldsymbol{\theta}_{n-1}\| \le c' w_n, \end{cases}
$$



Stochastic (Block) Majorization-Minimization

- Online CP-dictionary Learning (L., Strohmeier, Needell '20 [5]):

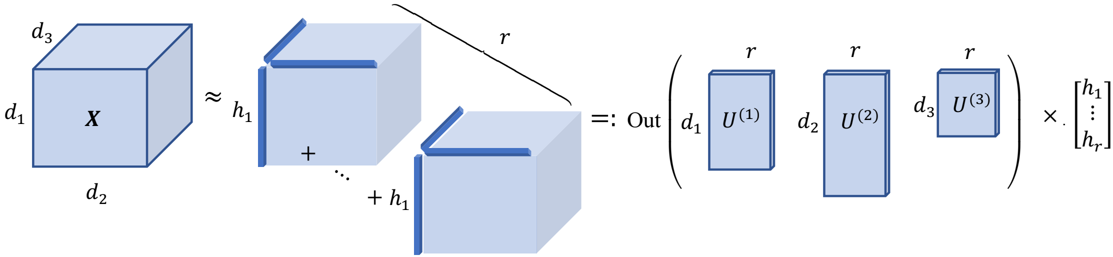
$$\text{(CP-recons. error)} \qquad \ell(\underbrace{\mathbf{X}}_{m\text{-tensor}}, \mathbf{U} = \underbrace{[U^{(1)}, \dots, U^{(m)}]}_{\text{factor matrices}}, H) := \|\mathbf{X} - \underbrace{\text{Out}(\mathbf{U})}_{\text{CP-dict.}} \times_{m+1} H\|_F^2$$



Stochastic (Block) Majorization-Minimization

- Online CP-dictionary Learning (L., Strohmeier, Needell '20 [5]):

$$\text{(CP-recons. error)} \qquad \ell(\underbrace{\mathbf{X}}_{m\text{-tensor}}, \mathbf{U} = \underbrace{[U^{(1)}, \dots, U^{(m)}]}_{\text{factor matrices}}, H) := \|\mathbf{X} - \underbrace{\mathrm{Out}(\mathbf{U})}_{\text{CP-dict.}} \times_{m+1} H\|_F^2$$



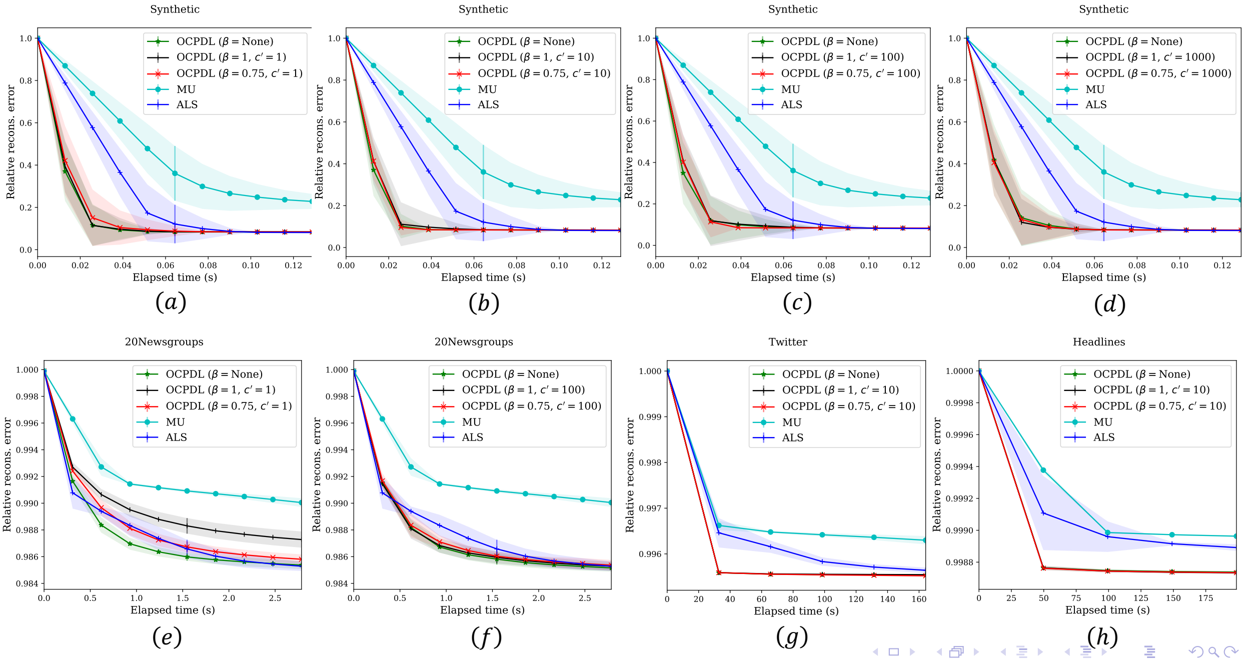
- Upon arrival of $\mathbf{X}_n \in \mathbb{R}^{d_1 \times \cdots \times d_m}$:

$$\begin{cases} H_n = \operatorname{argmin}_{H \in \subseteq \mathbb{R}_{\geq 0}^{r \times 1}} \ell(\mathbf{X}_n, \mathbf{U}_{n-1}, H) \\ \bar{g}_n(\mathbf{U}) = (1 - w_n)\bar{g}_{n-1}(\mathbf{U}) + w_n \ell(\mathbf{X}_n, \mathbf{U}, H_n) \\ \text{for } i = 1, \dots, m: \\ \quad U_n^{(i)} \in \operatorname{argmin}_{\substack{U \in \mathbb{R}_{\geq 0}^{d_i \times r} \\ \|U - U_{n-1}^{(i)}\| \leq c' w_n}} \bar{g}_n(U_n^{(1)}, \dots, U_n^{(i-1)}, U, U_{n-1}^{(i+1)}, \dots, U_{n-1}^{(m)}). \end{cases}$$

(m-block multi-convex)

Stochastic (Block) Majorization-Minimization

- Online CP-dictionary Learning (L., Strohmeier, Needell '22 [5]):
  - Only bounded memory to learn from infinitely many samples
  - Cheaper per-iteration cost than offline methods
  - Converges faster than offline methods (empirically)



Outline

BCD with Proximal Regularization and Diminishing Radius

- BCD-PR (Proximal Regularization) : For $n = 1, \ldots, N$ and for $i = 1, \ldots, m$:

$$\theta_n^{(i)} \in \operatorname*{argmin}_{\theta \in \Theta^{(i)}} f\left(\theta_n^{(1)}, \cdots, \theta_n^{(i-1)}, \theta, \theta_{n-1}^{(i+1)}, \cdots, \theta_{n-1}^{(m)}\right) + \lambda_n \|\theta - \theta_{n-1}^{(i)}\|^2$$

- BCD-DR (Diminishing Radius) : For $n = 1, \ldots, N$ and for $i = 1, \ldots, m$:

$$\theta_n^{(i)} \in \operatorname*{argmin}_{\theta \in \Theta^{(i)},\, \|\theta - \theta_{n-1}^{(i)}\| \le r_n} f\left(\theta_n^{(1)}, \cdots, \theta_n^{(i-1)}, \theta, \theta_{n-1}^{(i+1)}, \cdots, \theta_{n-1}^{(m)}\right)$$

BCD with Proximal Regularization and Diminishing Radius

- BCD-PR (Proximal Regularization) : For $n = 1, \dots, N$ and for $i = 1, \dots, m$:

$$\theta_n^{(i)} \in \operatorname*{argmin}_{\theta \in \Theta^{(i)}} f\left(\theta_n^{(1)}, \cdots, \theta_n^{(i-1)}, \theta, \theta_{n-1}^{(i+1)}, \cdots, \theta_{n-1}^{(m)}\right) + \lambda_n \|\theta - \theta_{n-1}^{(i)}\|^2$$

- BCD-DR (Diminishing Radius) : For $n = 1, \dots, N$ and for $i = 1, \dots, m$:

$$\theta_n^{(i)} \in \operatorname*{argmin}_{\theta \in \Theta^{(i)}, \|\theta - \theta_{n-1}^{(i)}\| \le r_n} f\left(\theta_n^{(1)}, \cdots, \theta_n^{(i-1)}, \theta, \theta_{n-1}^{(i+1)}, \cdots, \theta_{n-1}^{(m)}\right)$$

Def. $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$ is an *ε-approxiate stationary point* of f over $\boldsymbol{\Theta}$ if

$$-\inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\langle \nabla f(\boldsymbol{\theta}^*), \frac{(\boldsymbol{\theta} - \boldsymbol{\theta}^*)}{\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|} \right\rangle \le \sqrt{\varepsilon}$$

BCD with Proximal Regularization and Diminishing Radius

- BCD-PR (Proximal Regularization) : For $n = 1, \ldots, N$ and for $i = 1, \ldots, m$:

$$\theta_n^{(i)} \in \operatorname*{argmin}_{\theta \in \Theta^{(i)}} f\left(\theta_n^{(1)}, \cdots, \theta_n^{(i-1)}, \theta, \theta_{n-1}^{(i+1)}, \cdots, \theta_{n-1}^{(m)}\right) + \lambda_n \|\theta - \theta_{n-1}^{(i)}\|^2$$

- BCD-DR (Diminishing Radius) : For $n = 1, \ldots, N$ and for $i = 1, \ldots, m$:

$$\theta_n^{(i)} \in \operatorname*{argmin}_{\theta \in \Theta^{(i)},\, \|\theta - \theta_{n-1}^{(i)}\| \le r_n} f\left(\theta_n^{(1)}, \cdots, \theta_n^{(i-1)}, \theta, \theta_{n-1}^{(i+1)}, \cdots, \theta_{n-1}^{(m)}\right)$$

Def. $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$ is an *ε-approxiate stationary point* of f over $\boldsymbol{\Theta}$ if

$$-\inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\langle \nabla f(\boldsymbol{\theta}^*), \frac{(\boldsymbol{\theta} - \boldsymbol{\theta}^*)}{\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|} \right\rangle \le \sqrt{\varepsilon}$$

Theorem (L. '21+, L. and Kwon '22+)

Under mild conditions, BCD-DR and BCD-PR converges to the set of stationary points of f at rate $O(1/n)$; They find ε-approx. stationary point within $O(\varepsilon^{-1}(\log \varepsilon^{-1})^2)$ iterations.

Known results for SMM

- When $\boldsymbol{\theta} \mapsto \ell(\mathbf{x}, \boldsymbol{\theta})$ is convex, $\boldsymbol{\theta}_n \to$ global minimum at rate $O(\log n / \sqrt{n})$ for i.i.d. data samples $\mathbf{x}_n$ (Mairal 2013)

Known results for SMM

- When $\boldsymbol{\theta} \mapsto \ell(\mathbf{x}, \boldsymbol{\theta})$ is convex, $\boldsymbol{\theta}_n \to$ global minimum at rate $O(\log n/\sqrt{n})$ for i.i.d. data samples $\mathbf{x}_n$ (Mairal 2013)
- When $\boldsymbol{\theta} \mapsto \ell(\mathbf{x}, \boldsymbol{\theta})$ is non-convex, $\boldsymbol{\theta}_n \to$ {stationary pts. of expected loss} for i.i.d. data samples $\mathbf{x}_n$ (Mairal et al. 2010, Mairal 2013, Mensch et al. 2017)

Known results for SMM

- When $\boldsymbol{\theta} \mapsto \ell(\mathbf{x}, \boldsymbol{\theta})$ is convex, $\boldsymbol{\theta}_n \to$ global minimum at rate $O(\log n/\sqrt{n})$ for i.i.d. data samples $\mathbf{x}_n$ (Mairal 2013)
- When $\boldsymbol{\theta} \mapsto \ell(\mathbf{x}, \boldsymbol{\theta})$ is non-convex, $\boldsymbol{\theta}_n \to$ {stationary pts. of expected loss} for i.i.d. data samples $\mathbf{x}_n$ (Mairal et al. 2010, Mairal 2013, Mensch et al. 2017)
  - Holds for Online NMF loss with Markovian data samples (L., Balzano, Needell '20)

Known results for SMM

- When $\boldsymbol{\theta} \mapsto \ell(\mathbf{x}, \boldsymbol{\theta})$ is convex, $\boldsymbol{\theta}_n \to$ global minimum at rate $O(\log n/\sqrt{n})$ for i.i.d. data samples $\mathbf{x}_n$ (Mairal 2013)
- When $\boldsymbol{\theta} \mapsto \ell(\mathbf{x}, \boldsymbol{\theta})$ is non-convex, $\boldsymbol{\theta}_n \to$ {stationary pts. of expected loss} for i.i.d. data samples $\mathbf{x}_n$ (Mairal et al. 2010, Mairal 2013, Mensch et al. 2017)
  - Holds for Online NMF loss with Markovian data samples (L., Balzano, Needell '20)
  - Holds for Online CP-dictionary learning loss with Markovian data samples (L., Strohmeier, Needell 22)

Known results for SMM

- When $\boldsymbol{\theta} \mapsto \ell(\mathbf{x}, \boldsymbol{\theta})$ is convex, $\boldsymbol{\theta}_n \to$ global minimum at rate $O(\log n/\sqrt{n})$ for i.i.d. data samples $\mathbf{x}_n$ (Mairal 2013)
- When $\boldsymbol{\theta} \mapsto \ell(\mathbf{x}, \boldsymbol{\theta})$ is non-convex, $\boldsymbol{\theta}_n \to$ {stationary pts. of expected loss} for i.i.d. data samples $\mathbf{x}_n$ (Mairal et al. 2010, Mairal 2013, Mensch et al. 2017)
  - Holds for Online NMF loss with Markovian data samples (L., Balzano, Needell '20)
  - Holds for Online CP-dictionary learning loss with Markovian data samples (L., Strohmeier, Needell 22)
- No **rate of convergence** known for SMM in the nonconvex case (even with i.i.d. input + strongly cvx surrogates)

Known results for SMM

- When $\boldsymbol{\theta} \mapsto \ell(\mathbf{x}, \boldsymbol{\theta})$ is convex, $\boldsymbol{\theta}_n \to$ global minimum at rate $O(\log n/\sqrt{n})$ for i.i.d. data samples $\mathbf{x}_n$ (Mairal 2013)
- When $\boldsymbol{\theta} \mapsto \ell(\mathbf{x}, \boldsymbol{\theta})$ is non-convex, $\boldsymbol{\theta}_n \to$ {stationary pts. of expected loss} for i.i.d. data samples $\mathbf{x}_n$ (Mairal et al. 2010, Mairal 2013, Mensch et al. 2017)
  - Holds for Online NMF loss with Markovian data samples (L., Balzano, Needell '20)
  - Holds for Online CP-dictionary learning loss with Markovian data samples (L., Strohmeier, Needell 22)
- No **rate of convergence** known for SMM in the nonconvex case (even with i.i.d. input + strongly cvx surrogates)
  - For unconstrained nonconvex SGD, $O(\log n/\sqrt{n})$ rate to stationary pts. known for Markovian input (Sun et al. '18)

Known results for SMM

- When $\boldsymbol{\theta} \mapsto \ell(\mathbf{x}, \boldsymbol{\theta})$ is convex, $\boldsymbol{\theta}_n \to$ global minimum at rate $O(\log n/\sqrt{n})$ for i.i.d. data samples $\mathbf{x}_n$ (Mairal 2013)
- When $\boldsymbol{\theta} \mapsto \ell(\mathbf{x}, \boldsymbol{\theta})$ is non-convex, $\boldsymbol{\theta}_n \to$ {stationary pts. of expected loss} for i.i.d. data samples $\mathbf{x}_n$ (Mairal et al. 2010, Mairal 2013, Mensch et al. 2017)
  - Holds for Online NMF loss with Markovian data samples (L., Balzano, Needell '20)
  - Holds for Online CP-dictionary learning loss with Markovian data samples (L., Strohmeier, Needell 22)
- No **rate of convergence** known for SMM in the nonconvex case (even with i.i.d. input + strongly cvx surrogates)
  - For unconstrained nonconvex SGD, $O(\log n/\sqrt{n})$ rate to stationary pts. known for Markovian input (Sun et al. '18)
  - For constrained nonconvex PSGD, $O(\log n/\sqrt{n})$ rate to stationary pts. known for i.i.d. input (Davis, Drusvyatskiy '20)

Known results for SMM

- When $\boldsymbol{\theta} \mapsto \ell(\mathbf{x}, \boldsymbol{\theta})$ is convex, $\boldsymbol{\theta}_n \to$ global minimum at rate $O(\log n/\sqrt{n})$ for i.i.d. data samples $\mathbf{x}_n$ (Mairal 2013)
- When $\boldsymbol{\theta} \mapsto \ell(\mathbf{x}, \boldsymbol{\theta})$ is non-convex, $\boldsymbol{\theta}_n \to$ {stationary pts. of expected loss} for i.i.d. data samples $\mathbf{x}_n$ (Mairal et al. 2010, Mairal 2013, Mensch et al. 2017)
  - Holds for Online NMF loss with Markovian data samples (L., Balzano, Needell '20)
  - Holds for Online CP-dictionary learning loss with Markovian data samples (L., Strohmeier, Needell 22)
- No **rate of convergence** known for SMM in the nonconvex case (even with i.i.d. input + strongly cvx surrogates)
  - For unconstrained nonconvex SGD, $O(\log n/\sqrt{n})$ rate to stationary pts. known for Markovian input (Sun et al. '18)
  - For constrained nonconvex PSGD, $O(\log n/\sqrt{n})$ rate to stationary pts. known for i.i.d. input (Davis, Drusvyatskiy '20)
    - Recently extended to the Markovian case (L., Alacaoglu '22+)

Rate of Convergence of SBMM

Corollary (L. '22+)

$(\boldsymbol{\theta}_n)_{n\geq 0}$ = *output of SBMM*, $(\mathbf{x}_n)_{n\geq 1}$: *exponentially mixing data samples.*
If $\boldsymbol{\theta}_n \in \mathrm{interior}(\boldsymbol{\Theta})$ *for* $n \geq 1$ *and* $w_n = n^{-1/2}(\log n)^{1+\varepsilon}$,

$$\min_{1\leq k\leq n} \left\| \nabla \bar{g}_k(\boldsymbol{\theta}_k) \right\|^2 = O\left(\frac{(\log n)^{2+2\varepsilon}}{n} \right), \quad \min_{1\leq k\leq n} \left\| \nabla \bar{f}_k(\boldsymbol{\theta}_k) \right\|^2 = O\left(\frac{(\log n)^{1+\varepsilon}}{\sqrt{n}} \right),$$

$$\min_{1\leq k\leq n} \left\| \nabla f(\boldsymbol{\theta}_k) \right\|^2 = O\left(\frac{(\log n)^{1+\varepsilon}}{\sqrt{n}} \right).$$

Rate of Convergence of SBMM

Corollary (L. '22+)

$(\boldsymbol{\theta}_n)_{n\geq 0}$ = *output of SBMM,* $(\mathbf{x}_n)_{n\geq 1}$*: exponentially mixing data samples.*
If $\boldsymbol{\theta}_n \in \text{interior}(\boldsymbol{\Theta})$ *for* $n \geq 1$ *and* $w_n = n^{-1/2}(\log n)^{1+\varepsilon}$,

$$\min_{1\leq k\leq n} \left\|\nabla \bar{g}_k(\boldsymbol{\theta}_k)\right\|^2 = O\left(\frac{(\log n)^{2+2\varepsilon}}{n}\right), \quad \min_{1\leq k\leq n} \left\|\nabla \bar{f}_k(\boldsymbol{\theta}_k)\right\|^2 = O\left(\frac{(\log n)^{1+\varepsilon}}{\sqrt{n}}\right),$$

$$\min_{1\leq k\leq n} \left\|\nabla f(\boldsymbol{\theta}_k)\right\|^2 = O\left(\frac{(\log n)^{1+\varepsilon}}{\sqrt{n}}\right).$$

- Provides first convergence rate bound for Online NMF, Online CPDL, SMM, and SBMM in the general Markovian data case

Rate of Convergence of SBMM

Corollary (L. '22+)

$(\boldsymbol{\theta}_n)_{n\geq 0}$ = *output of SBMM,* $(\mathbf{x}_n)_{n\geq 1}$*: exponentially mixing data samples.*
If $\boldsymbol{\theta}_n \in \text{interior}(\boldsymbol{\Theta})$ *for* $n \geq 1$ *and* $w_n = n^{-1/2}(\log n)^{1+\varepsilon}$,

$$\min_{1\leq k\leq n} \left\| \nabla \bar{g}_k(\boldsymbol{\theta}_k) \right\|^2 = O\left(\frac{(\log n)^{2+2\varepsilon}}{n} \right), \quad \min_{1\leq k\leq n} \left\| \nabla \bar{f}_k(\boldsymbol{\theta}_k) \right\|^2 = O\left(\frac{(\log n)^{1+\varepsilon}}{\sqrt{n}} \right),$$

$$\min_{1\leq k\leq n} \left\| \nabla f(\boldsymbol{\theta}_k) \right\|^2 = O\left(\frac{(\log n)^{1+\varepsilon}}{\sqrt{n}} \right).$$

- Provides first convergence rate bound for Online NMF, Online CPDL, SMM, and SBMM in the general Markovian data case
- Matches with optimal SGD/PSGD rate of convergence $O((\log n)/\sqrt{n})$ up to a log factor

Rate of Convergence of SBMM

Corollary (L. '22+)

$(\boldsymbol{\theta}_n)_{n\geq 0}$ = *output of SBMM*, $(\mathbf{x}_n)_{n\geq 1}$: *exponentially mixing data samples.*
If $\boldsymbol{\theta}_n \in \mathrm{interior}(\boldsymbol{\Theta})$ *for* $n \geq 1$ *and* $w_n = n^{-1/2}(\log n)^{1+\varepsilon}$,

$$\min_{1\leq k\leq n} \left\|\nabla \bar{g}_k(\boldsymbol{\theta}_k)\right\|^2 = O\left(\frac{(\log n)^{2+2\varepsilon}}{n}\right), \quad \min_{1\leq k\leq n} \left\|\nabla \bar{f}_k(\boldsymbol{\theta}_k)\right\|^2 = O\left(\frac{(\log n)^{1+\varepsilon}}{\sqrt{n}}\right),$$

$$\min_{1\leq k\leq n} \left\|\nabla f(\boldsymbol{\theta}_k)\right\|^2 = O\left(\frac{(\log n)^{1+\varepsilon}}{\sqrt{n}}\right).$$

- Provides first convergence rate bound for Online NMF, Online CPDL, SMM, and SBMM in the general Markovian data case
- Matches with optimal SGD/PSGD rate of convergence $O((\log n)/\sqrt{n})$ up to a log factor
- Best known rate of convergence of SGD/PSGD for the empirical loss $\bar{f}_n$ is $O(1/n^{1/4})$.

Rate of Convergence of SBMM

Corollary (L. '22+)

$(\boldsymbol{\theta}_n)_{n\geq 0}$ = *output of SBMM*, $(\mathbf{x}_n)_{n\geq 1}$: *exponentially mixing data samples.*
If $\boldsymbol{\theta}_n \in \text{interior}(\boldsymbol{\Theta})$ *for* $n \geq 1$ *and* $w_n = n^{-1/2}(\log n)^{1+\varepsilon}$,

$$\min_{1\leq k\leq n} \left\|\nabla \bar{g}_k(\boldsymbol{\theta}_k)\right\|^2 = O\left(\frac{(\log n)^{2+2\varepsilon}}{n}\right), \quad \min_{1\leq k\leq n} \left\|\nabla \bar{f}_k(\boldsymbol{\theta}_k)\right\|^2 = O\left(\frac{(\log n)^{1+\varepsilon}}{\sqrt{n}}\right),$$

$$\min_{1\leq k\leq n} \left\|\nabla f(\boldsymbol{\theta}_k)\right\|^2 = O\left(\frac{(\log n)^{1+\varepsilon}}{\sqrt{n}}\right).$$

- Provides first convergence rate bound for Online NMF, Online CPDL, SMM, and SBMM in the general Markovian data case
- Matches with optimal SGD/PSGD rate of convergence $O((\log n)/\sqrt{n})$ up to a log factor
- Best known rate of convergence of SGD/PSGD for the empirical loss $\bar{f}_n$ is $O(1/n^{1/4})$.
  - SGD/PSGD solves for f and indirectly solves for $\bar{f}_n$;

Rate of Convergence of SBMM

Corollary (L. '22+)

$(\boldsymbol{\theta}_n)_{n\geq 0}$ = *output of SBMM,* $(\mathbf{x}_n)_{n\geq 1}$: *exponentially mixing data samples.*
If $\boldsymbol{\theta}_n \in \text{interior}(\boldsymbol{\Theta})$ *for* $n \geq 1$ *and* $w_n = n^{-1/2}(\log n)^{1+\varepsilon}$,

$$\min_{1\leq k\leq n} \left\| \nabla \bar{g}_k(\boldsymbol{\theta}_k) \right\|^2 = O\left(\frac{(\log n)^{2+2\varepsilon}}{n} \right), \quad \min_{1\leq k\leq n} \left\| \nabla \bar{f}_k(\boldsymbol{\theta}_k) \right\|^2 = O\left(\frac{(\log n)^{1+\varepsilon}}{\sqrt{n}} \right),$$

$$\min_{1\leq k\leq n} \left\| \nabla f(\boldsymbol{\theta}_k) \right\|^2 = O\left(\frac{(\log n)^{1+\varepsilon}}{\sqrt{n}} \right).$$

- Provides first convergence rate bound for Online NMF, Online CPDL, SMM, and SBMM in the general Markovian data case
- Matches with optimal SGD/PSGD rate of convergence $O((\log n)/\sqrt{n})$ up to a log factor
- Best known rate of convergence of SGD/PSGD for the empirical loss $\bar{f}_n$ is $O(1/n^{1/4})$.
  - SGD/PSGD solves for f and indirectly solves for $\bar{f}_n$;
  - SBMM solves for $\bar{f}_n$ and indirectly solves for f

Rate of Convergence of SBMM

Theorem (L. '22+)

$(\boldsymbol{\theta}_n)_{n\geq 0}$ = *output of SBMM*, $(\mathbf{x}_n)_{n\geq 1}$: *exponentially mixing data samples*,

Slow adaptation regime: $\frac{1}{n} \leq w_n \ll \frac{1}{\sqrt{n}}$

(i) *(Surrogate and Empirical Loss Stationarity) Asymtotically almost surely,*

$$\min_{1\leq k\leq n} \left[-\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\langle \nabla \bar{g}_k(\boldsymbol{\theta}_k), \frac{(\boldsymbol{\theta}-\boldsymbol{\theta}_k)}{\|\boldsymbol{\theta}-\boldsymbol{\theta}_k\|} \right\rangle \right]^2 = O\left(\left(\sum_{k=1}^{n} w_k \right)^{-2} \right).$$

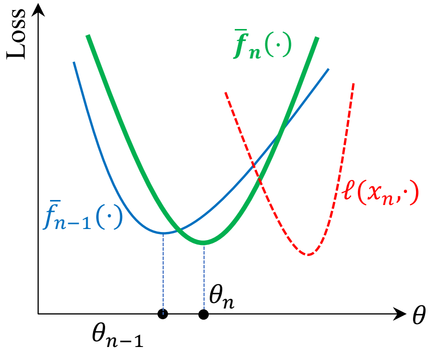
$$\min_{1\leq k\leq n} \left[-\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\langle \nabla \bar{f}_k(\boldsymbol{\theta}_k), \frac{(\boldsymbol{\theta}-\boldsymbol{\theta}_k)}{\|\boldsymbol{\theta}-\boldsymbol{\theta}_k\|} \right\rangle \right]^2 = O\Bigg(\underbrace{\left(\sum_{k=1}^{n} w_k \right)^{-1}}_{\text{best case: } 1/\sqrt{n}} \Bigg).$$

(ii) *(Expected Loss Stationarity) If* $\overbrace{w_n = o(1/n^{3/4})}^{\text{Slower adaptation}}$, *asymtotically almost surely,*

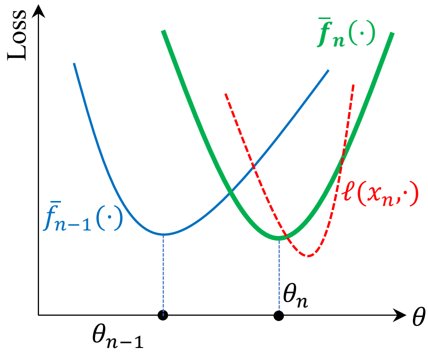
$$\min_{1\leq k\leq n} \left[\underbrace{-\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\langle \nabla f(\boldsymbol{\theta}_k), \frac{(\boldsymbol{\theta}-\boldsymbol{\theta}_k)}{\|\boldsymbol{\theta}-\boldsymbol{\theta}_k\|} \right\rangle}_{\text{optimality gap for constrained case}} \right]^2 = O\Bigg(\underbrace{\left(\sum_{k=1}^{n} w_k \right)^{-1}}_{\text{best case: } 1/n^{1/4}} \Bigg).$$

Open questions

- What happens in the fast adaptation regime $w_n = \Omega(1/\sqrt{n})$?



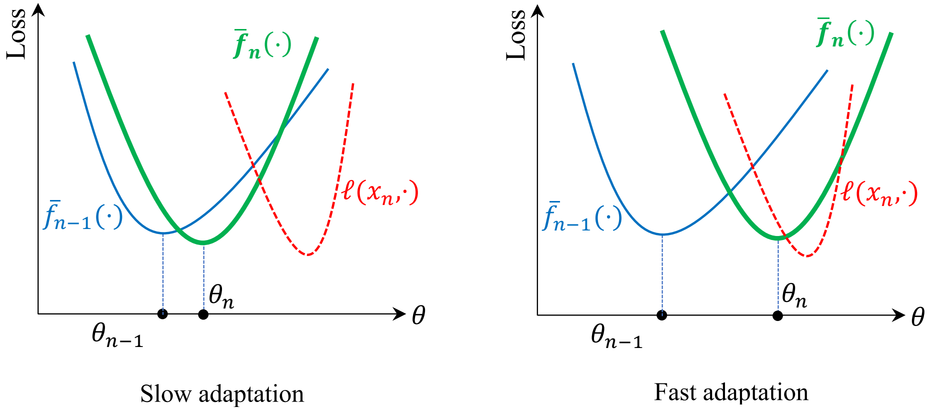
Slow adaptation



Fast adaptation

Open questions

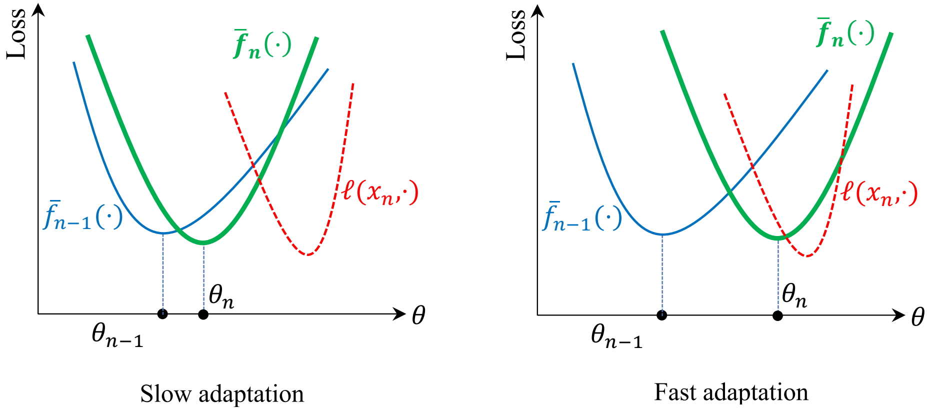
- What happens in the fast adaptation regime $w_n = \Omega(1/\sqrt{n})$?



  - All theoretical analysis assumes slow adaptation regime $\frac{1}{n} \leq w_n \ll \frac{1}{\sqrt{n}}$:

Open questions

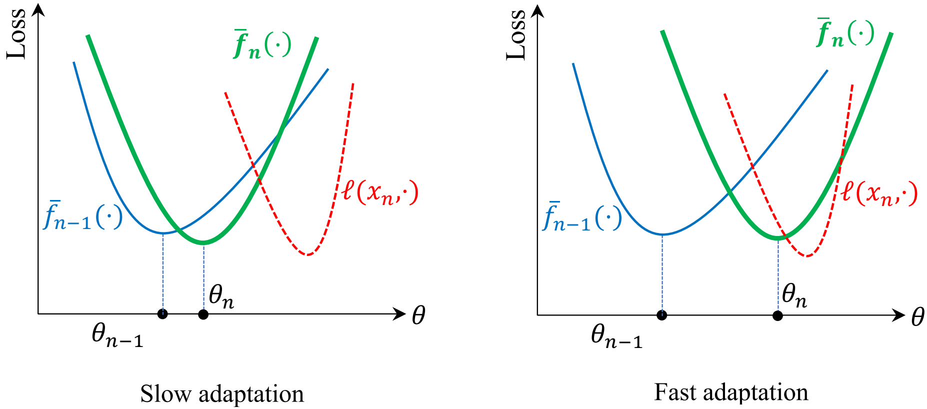
- What happens in the fast adaptation regime $w_n = \Omega(1/\sqrt{n})$?



  - All theoretical analysis assumes slow adaptation regime $\frac{1}{n} \leq w_n \ll \frac{1}{\sqrt{n}}$:
  - It allows to use weighted versions of SLLN and CLT: $\bar{f}_n \approx f$, $\nabla \bar{f}_n \approx \nabla f$

Open questions

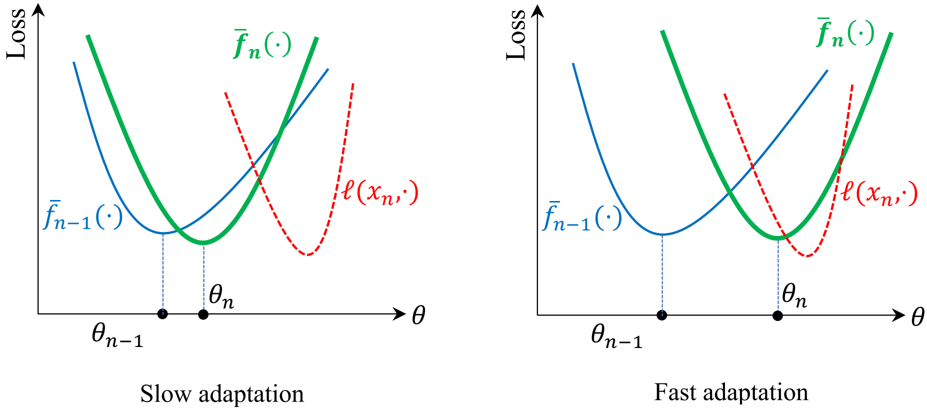
- What happens in the fast adaptation regime $w_n = \Omega(1/\sqrt{n})$?



  - All theoretical analysis assumes slow adaptation regime $\frac{1}{n} \le w_n \ll \frac{1}{\sqrt{n}}$:
  - It allows to use weighted versions of SLLN and CLT: $\bar{f}_n \approx f$, $\nabla \bar{f}_n \approx \nabla f$
  - Formulate the goal of learning non-stationary (short-time scale) features

Open questions

- What happens in the fast adaptation regime $w_n = \Omega(1/\sqrt{n})$?

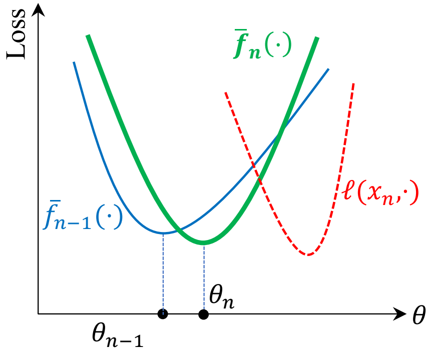


  - All theoretical analysis assumes slow adaptation regime $\frac{1}{n} \le w_n \ll \frac{1}{\sqrt{n}}$:
  - It allows to use weighted versions of SLLN and CLT: $\bar{f}_n \approx f$, $\nabla \bar{f}_n \approx \nabla f$
  - Formulate the goal of learning non-stationary (short-time scale) features

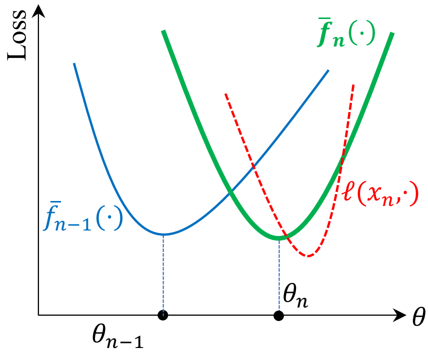
- Finding global minimizer for some online nonconvex problems?

Open questions

- What happens in the fast adaptation regime $w_n = \Omega(1/\sqrt{n})$?



Slow adaptation



Fast adaptation

  - All theoretical analysis assumes slow adaptation regime $\frac{1}{n} \le w_n \ll \frac{1}{\sqrt{n}}$:
  - It allows to use weighted versions of SLLN and CLT: $\bar{f}_n \approx f$, $\nabla \bar{f}_n \approx \nabla f$
  - Formulate the goal of learning non-stationary (short-time scale) features

- Finding global minimizer for some online nonconvex problems?
  - Many recent developments on global landscape analysis on low-rank problems / Tucker decomposition

Thanks!

Outline

Proposition (Finite first-order variation)

For BCD-DR with $\sum_{n=1}^{\infty} r_n^2 < \infty$,

$$\sum_{n=1}^{\infty} \left|\left\langle \nabla f(\boldsymbol{\theta}_{n+1}), \boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1} \right\rangle\right| \leq \frac{L}{2}\left(\sum_{n=1}^{\infty} \underbrace{\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}\|^2}_{\leq r_n^2} \right) + f(\boldsymbol{\theta}_1) < \infty.$$

Proposition (Finite first-order variation)

For BCD-DR with $\sum_{n=1}^{\infty} r_n^2 < \infty$,

$$\sum_{n=1}^{\infty} \left| \langle \nabla f(\boldsymbol{\theta}_{n+1}), \boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1} \rangle \right| \le \frac{L}{2} \left(\sum_{n=1}^{\infty} \underbrace{\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}\|^2}_{\le r_n^2} \right) + f(\boldsymbol{\theta}_1) < \infty.$$

Proposition (Asymptotic first-order optimality)

Fix a sequence $(b_n)_{n\ge1}$ *such that* $0 < b_n \le r_n$ *for all* $n \ge 1$. *Then*

$$-b_{n+1} \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\langle \nabla f(\boldsymbol{\theta}_n), \frac{\boldsymbol{\theta} - \boldsymbol{\theta}_n}{\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\|} \right\rangle \le \left| \langle \nabla f(\boldsymbol{\theta}_{n+1}), \boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n \rangle \right| + c_1 \left(b_{n+1}^2 + \|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n\|^2 \right)$$

Proposition (Finite first-order variation)

For BCD-DR with $\sum_{n=1}^{\infty} r_n^2 < \infty$,

$$\sum_{n=1}^{\infty} \left| \langle \nabla f(\boldsymbol{\theta}_{n+1}), \boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1} \rangle \right| \le \frac{L}{2} \left(\sum_{n=1}^{\infty} \underbrace{\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}\|^2}_{\le r_n^2} \right) + f(\boldsymbol{\theta}_1) < \infty.$$

Proposition (Asymptotic first-order optimality)

Fix a sequence $(b_n)_{n\ge 1}$ such that $0 < b_n \le r_n$ for all $n \ge 1$. Then

$$-b_{n+1} \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\langle \nabla f(\boldsymbol{\theta}_n), \frac{\boldsymbol{\theta} - \boldsymbol{\theta}_n}{\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\|} \right\rangle \le \left| \langle \nabla f(\boldsymbol{\theta}_{n+1}), \boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n \rangle \right| + c_1 \left(b_{n+1}^2 + \|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n\|^2 \right)$$

- By adding up the previous inequality:

$$\sum_{n=1}^{\infty} r_n \left[-\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\langle \nabla f(\boldsymbol{\theta}_n), \frac{\boldsymbol{\theta} - \boldsymbol{\theta}_n}{\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\|} \right\rangle \right] < M < \infty.$$

- By adding up the previous inequality:

$$\sum_{n=1}^{\infty} r_n \left[-\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\langle \nabla f(\boldsymbol{\theta}_n), \frac{\boldsymbol{\theta}-\boldsymbol{\theta}_n}{\|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|} \right\rangle \right] < M < \infty.$$

- By adding up the previous inequality:

$$\sum_{n=1}^{\infty} r_n \left[- \inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\langle \nabla f(\boldsymbol{\theta}_n), \frac{\boldsymbol{\theta} - \boldsymbol{\theta}_n}{\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\|} \right\rangle \right] < M < \infty.$$

- This easily gives

$$\min_{1 \le k \le n} \left[- \inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\langle \nabla f(\boldsymbol{\theta}_k), \frac{\boldsymbol{\theta} - \boldsymbol{\theta}_k}{\|\boldsymbol{\theta} - \boldsymbol{\theta}_k\|} \right\rangle \right] \le \frac{M}{\sum_{k=1}^{n} r_k}.$$

- By adding up the previous inequality:

$$\sum_{n=1}^{\infty} r_n \left[-\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\langle \nabla f(\boldsymbol{\theta}_n), \frac{\boldsymbol{\theta}-\boldsymbol{\theta}_n}{\|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|} \right\rangle \right] < M < \infty.$$

- This easily gives

$$\min_{1\le k\le n} \left[-\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\langle \nabla f(\boldsymbol{\theta}_k), \frac{\boldsymbol{\theta}-\boldsymbol{\theta}_k}{\|\boldsymbol{\theta}-\boldsymbol{\theta}_k\|} \right\rangle \right] \le \frac{M}{\sum_{k=1}^{n} r_k}.$$

- Do a bookeeping for M and show it does not depend on the initialization $\boldsymbol{\theta}_0$:

$$\min_{1\le k\le n} \sup_{\boldsymbol{\theta}_0\in\boldsymbol{\Theta}} \left[-\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\langle \nabla f(\boldsymbol{\theta}_k), \frac{\boldsymbol{\theta}-\boldsymbol{\theta}_k}{\|\boldsymbol{\theta}-\boldsymbol{\theta}_k\|} \right\rangle \right] \le \frac{M}{\sum_{k=1}^{n} r_k}.$$

- By adding up the previous inequality:

$$\sum_{n=1}^{\infty} r_n \left[-\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\langle \nabla f(\boldsymbol{\theta}_n), \frac{\boldsymbol{\theta}-\boldsymbol{\theta}_n}{\|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|} \right\rangle \right] < M < \infty.$$

- This easily gives

$$\min_{1\le k\le n} \left[-\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\langle \nabla f(\boldsymbol{\theta}_k), \frac{\boldsymbol{\theta}-\boldsymbol{\theta}_k}{\|\boldsymbol{\theta}-\boldsymbol{\theta}_k\|} \right\rangle \right] \le \frac{M}{\sum_{k=1}^{n} r_k}.$$

- Do a bookeeping for M and show it does not depend on the initialization $\boldsymbol{\theta}_0$:

$$\min_{1\le k\le n} \sup_{\boldsymbol{\theta}_0\in\boldsymbol{\Theta}} \left[-\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\langle \nabla f(\boldsymbol{\theta}_k), \frac{\boldsymbol{\theta}-\boldsymbol{\theta}_k}{\|\boldsymbol{\theta}-\boldsymbol{\theta}_k\|} \right\rangle \right] \le \frac{M}{\sum_{k=1}^{n} r_k}.$$

- Is that it? Not quite, this only gives a subsequencial convergence and its rate. (Though it does imply iteration complexity bound.)

- By adding up the previous inequality:

$$\sum_{n=1}^{\infty} r_n \left[-\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\langle \nabla f(\boldsymbol{\theta}_n), \frac{\boldsymbol{\theta}-\boldsymbol{\theta}_n}{\|\boldsymbol{\theta}-\boldsymbol{\theta}_n\|} \right\rangle \right] < M < \infty.$$

- This easily gives

$$\min_{1\le k\le n} \left[-\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\langle \nabla f(\boldsymbol{\theta}_k), \frac{\boldsymbol{\theta}-\boldsymbol{\theta}_k}{\|\boldsymbol{\theta}-\boldsymbol{\theta}_k\|} \right\rangle \right] \le \frac{M}{\sum_{k=1}^{n} r_k}.$$

- Do a bookeeping for M and show it does not depend on the initialization $\boldsymbol{\theta}_0$:

$$\min_{1\le k\le n} \sup_{\boldsymbol{\theta}_0\in\boldsymbol{\Theta}} \left[-\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\langle \nabla f(\boldsymbol{\theta}_k), \frac{\boldsymbol{\theta}-\boldsymbol{\theta}_k}{\|\boldsymbol{\theta}-\boldsymbol{\theta}_k\|} \right\rangle \right] \le \frac{M}{\sum_{k=1}^{n} r_k}.$$

- Is that it? Not quite, this only gives a subsequencial convergence and its rate. (Though it does imply iteration complexity bound.)
  - How do we know if every convergent subsequence of $(\boldsymbol{\theta}_n)_{n\ge 1}$ converges to a stationary point?

- Suppose W.L.O.G. $(\boldsymbol{\theta}_n)_{n\geq 1}$ (from BCD-DR) converges to a limit point $\boldsymbol{\theta}_\infty \in \boldsymbol{\Theta}$.

- Suppose W.L.O.G. $(\boldsymbol{\theta}_n)_{n\geq 1}$ (from BCD-DR) converges to a limit point $\boldsymbol{\theta}_\infty \in \boldsymbol{\Theta}$.
- WTS: $\boldsymbol{\theta}_\infty$ is stationary for f over $\boldsymbol{\Theta}$:

$$\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \langle \nabla f(\boldsymbol{\theta}_\infty), \boldsymbol{\theta} - \boldsymbol{\theta}_\infty \rangle \geq 0$$

- Suppose W.L.O.G. $(\boldsymbol{\theta}_n)_{n\geq 1}$ (from BCD-DR) converges to a limit point $\boldsymbol{\theta}_\infty \in \boldsymbol{\Theta}$.
- WTS: $\boldsymbol{\theta}_\infty$ is stationary for f over $\boldsymbol{\Theta}$:

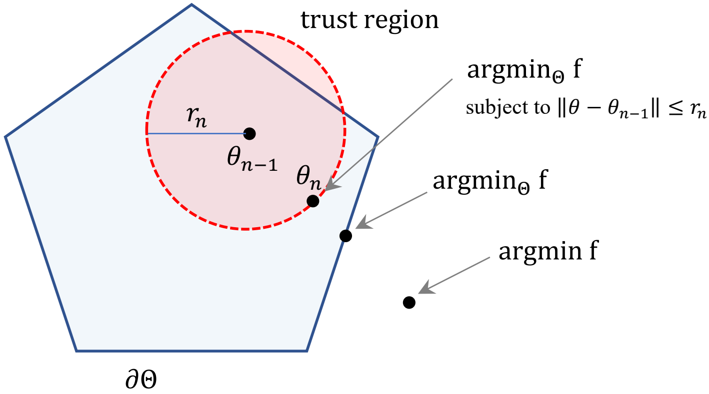
$$\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \langle \nabla f(\boldsymbol{\theta}_\infty), \boldsymbol{\theta} - \boldsymbol{\theta}_\infty \rangle \geq 0$$

- Main difficulty: Show that the DR (also the PR) modification of BCD does not affect the asymptotic property of iterates

- Suppose W.L.O.G. $(\boldsymbol{\theta}_n)_{n\geq 1}$ (from BCD-DR) converges to a limit point $\boldsymbol{\theta}_\infty \in \boldsymbol{\Theta}$.
- WTS: $\boldsymbol{\theta}_\infty$ is stationary for f over $\boldsymbol{\Theta}$:

$$\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \langle \nabla f(\boldsymbol{\theta}_\infty), \boldsymbol{\theta} - \boldsymbol{\theta}_\infty \rangle \geq 0$$

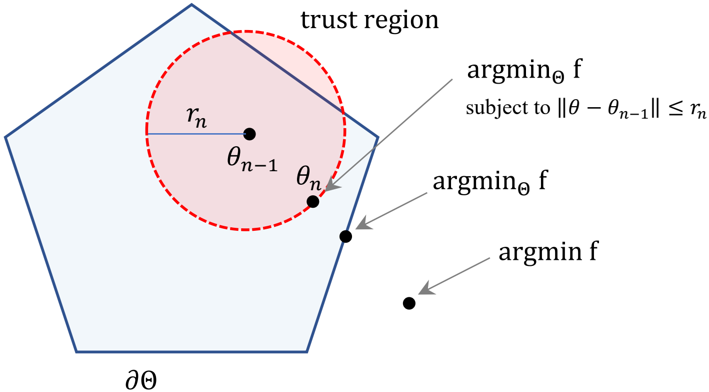
- Main difficulty: Show that the DR (also the PR) modification of BCD does not affect the asymptotic property of iterates
  - For BCD-DR: What if $\boldsymbol{\theta}_n$ touches the trust region boundary $\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\| \leq r_n$ infintely often?



- Suppose W.L.O.G. $(\boldsymbol{\theta}_n)_{n\geq 1}$ (from BCD-DR) converges to a limit point $\boldsymbol{\theta}_\infty \in \boldsymbol{\Theta}$.
- WTS: $\boldsymbol{\theta}_\infty$ is stationary for f over $\boldsymbol{\Theta}$:

$$\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \langle \nabla f(\boldsymbol{\theta}_\infty), \boldsymbol{\theta} - \boldsymbol{\theta}_\infty \rangle \geq 0$$

- Main difficulty: Show that the DR (also the PR) modification of BCD does not affect the asymptotic property of iterates
  - For BCD-DR: What if $\boldsymbol{\theta}_n$ touches the trust region boundary $\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\| \leq r_n$ infintely often?



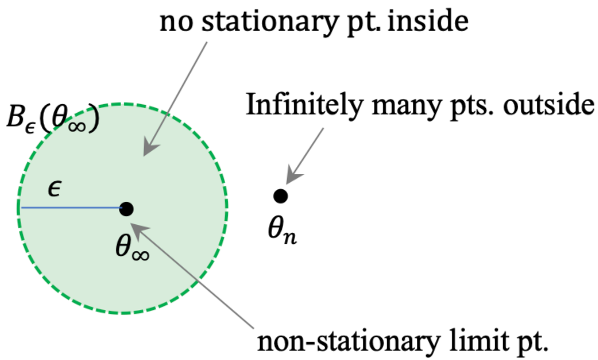
  - For BCD-PR: What if the PR term tilts the true gradient asymptotically?

Proposition (Local structure of a non-stationary limit point)

Assume $\sum_{n=1}^{\infty} r_n = \infty$, and $\sum_{n=1}^{\infty} r_n^2 < \infty$. Suppose there exists a non-stationary limit point $\boldsymbol{\theta}_\infty$ of $(\boldsymbol{\theta}_n)_{n\geq 1}$. Then there exists $\varepsilon > 0$ such that the ε-neighborhood $B_\varepsilon(\boldsymbol{\theta}_\infty) := \{\boldsymbol{\theta} \in \boldsymbol{\Theta} \,|\, \|\boldsymbol{\theta} - \boldsymbol{\theta}_\infty\| < \varepsilon\}$ s.t.

(a) *$B_\varepsilon(\boldsymbol{\theta}_\infty)$ does not contain any stationary points of f over $\boldsymbol{\Theta}$*

(b) *There exists infinitely many $\boldsymbol{\theta}_n$'s outside of $B_\varepsilon(\boldsymbol{\theta}_\infty)$.*

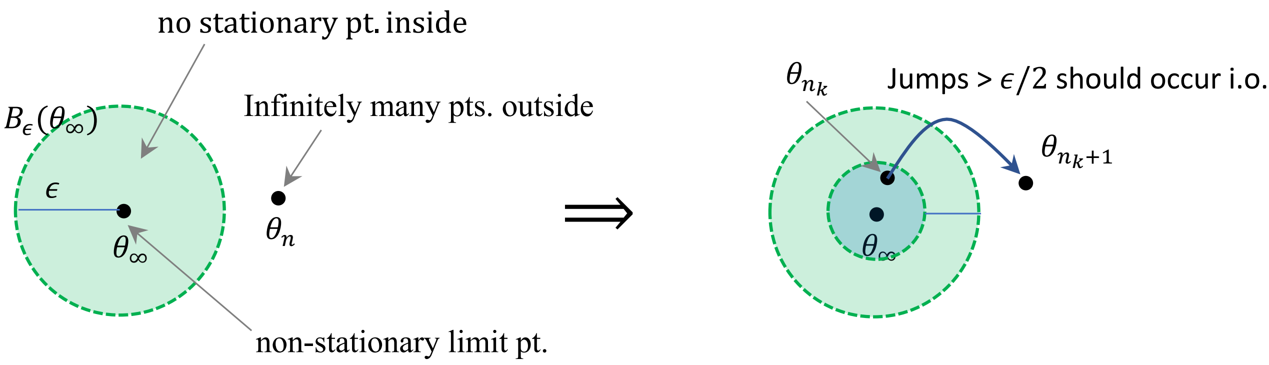


Proposition (Local structure of a non-stationary limit point)

Assume $\sum_{n=1}^{\infty} r_n = \infty$, and $\sum_{n=1}^{\infty} r_n^2 < \infty$. Suppose there exists a non-stationary limit point $\boldsymbol{\theta}_\infty$ of $(\boldsymbol{\theta}_n)_{n\geq 1}$. Then there exists $\varepsilon > 0$ such that the ε-neighborhood $B_\varepsilon(\boldsymbol{\theta}_\infty) := \{\boldsymbol{\theta} \in \boldsymbol{\Theta} \,|\, \|\boldsymbol{\theta} - \boldsymbol{\theta}_\infty\| < \varepsilon\}$ s.t.

(a) *$B_\varepsilon(\boldsymbol{\theta}_\infty)$ does not contain any stationary points of f over $\boldsymbol{\Theta}$*

(b) *There exists infinitely many $\boldsymbol{\theta}_n$'s outside of $B_\varepsilon(\boldsymbol{\theta}_\infty)$.*

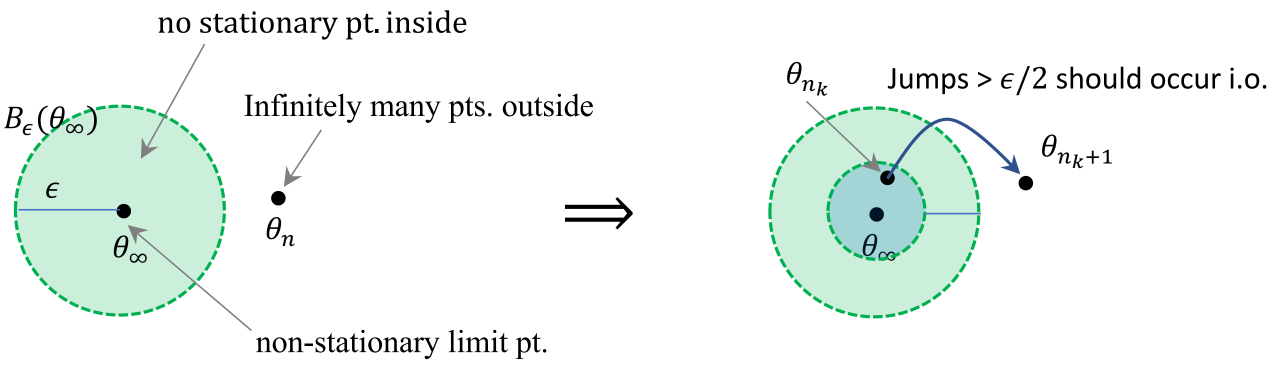


Proposition (Local structure of a non-stationary limit point)

Assume $\sum_{n=1}^{\infty} r_n = \infty$, and $\sum_{n=1}^{\infty} r_n^2 < \infty$. Suppose there exists a non-stationary limit point $\boldsymbol{\theta}_\infty$ of $(\boldsymbol{\theta}_n)_{n\geq 1}$. Then there exists $\varepsilon > 0$ such that the ε-neighborhood $B_\varepsilon(\boldsymbol{\theta}_\infty) := \{\boldsymbol{\theta} \in \boldsymbol{\Theta} \,|\, \|\boldsymbol{\theta} - \boldsymbol{\theta}_\infty\| < \varepsilon\}$ s.t.

(a) *$B_\varepsilon(\boldsymbol{\theta}_\infty)$ does not contain any stationary points of f over $\boldsymbol{\Theta}$*

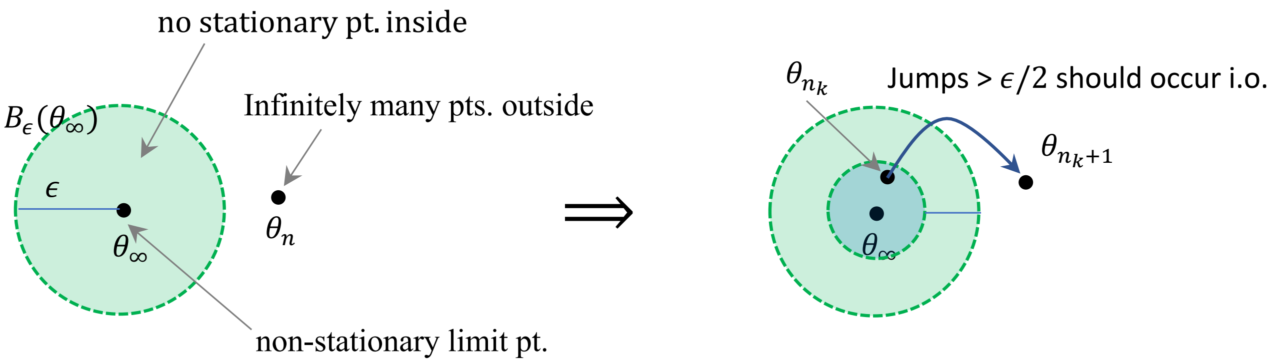
(b) *There exists infinitely many $\boldsymbol{\theta}_n$'s outside of $B_\varepsilon(\boldsymbol{\theta}_\infty)$.*



- So one can deduce $\sum_{n=1}^{\infty} \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| = \infty$.

Proposition (Sufficient condition for stationarity II)

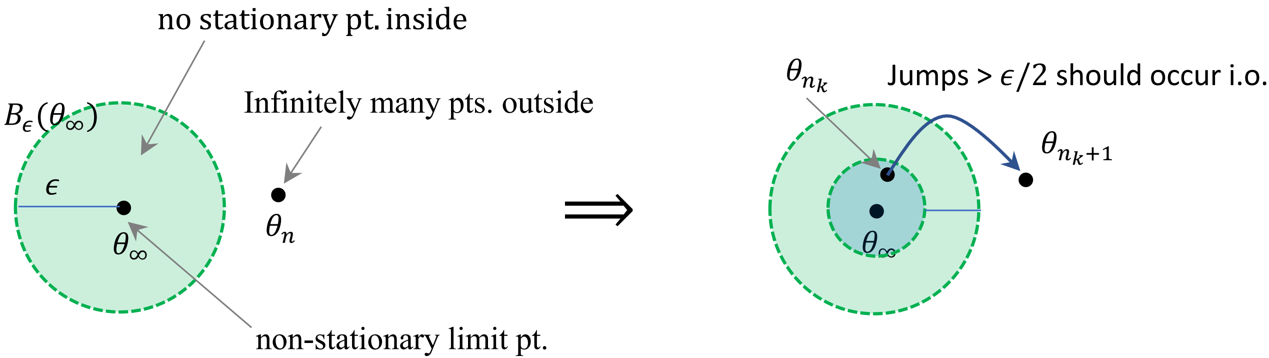
Suppose there exists a subsequence $(\boldsymbol{\theta}_{n_k})_{k\geq 1}$ such that $\sum_{k=1}^{\infty} \|\boldsymbol{\theta}_{n_k} - \boldsymbol{\theta}_{n_k+1}\| = \infty$. There exists a further subsequence $(s_k)_{k\geq 1}$ of $(n_k)_{k\geq 1}$ such that $\boldsymbol{\theta}_\infty := \lim_{k\to\infty} \boldsymbol{\theta}_{s_k}$ exists and is stationary.



- So one can deduce $\sum_{n=1}^{\infty} \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| = \infty$.

Proposition (Sufficient condition for stationarity II)

Suppose there exists a subsequence $(\boldsymbol{\theta}_{n_k})_{k\geq 1}$ *such that* $\sum_{k=1}^{\infty} \|\boldsymbol{\theta}_{n_k} - \boldsymbol{\theta}_{n_k+1}\| = \infty$. *There exists a further subsequence* $(s_k)_{k\geq 1}$ *of* $(n_k)_{k\geq 1}$ *such that* $\boldsymbol{\theta}_\infty := \lim_{k\to\infty} \boldsymbol{\theta}_{s_k}$ *exists and is stationary.*



- So one can deduce $\sum_{n=1}^{\infty} \|\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}\| = \infty$.
- This implies $(\boldsymbol{\theta}_n)_{n\geq 1}$ has a subsequence that converges to a stationary point, which should be inside $B_\varepsilon(\boldsymbol{\theta}_\infty)$, $\Rightarrow\Leftarrow$.

References I

[1] Michael Elad and Michal Aharon. "Image denoising via sparse and redundant representations over learned dictionaries". In: *IEEE Transactions on Image processing* 15.12 (2006), pp. 3736–3745.

[2] Hanbaek Lyu. "Convergence and complexity of block coordinate descent with diminishing radius for nonconvex optimization". In: *arXiv preprint arXiv:2012.03503* (2020).

[3] Hanbaek Lyu. "Convergence and Complexity of Stochastic Block Majorization-Minimization". In: *arXiv preprint arXiv:2201.01652* (2022).

[4] Hanbaek Lyu, Deanna Needell, and Laura Balzano. "Online matrix factorization for Markovian data and applications to network dictionary learning". In: *Journal of Machine Learning Research 21* 21 (2021), pp. 1–49.

[5] Hanbaek Lyu, Christopher Strohmeier, and Deanna Needell. "Online nonnegative CP-dictionary learning for Markovian data". In: *To appear in JMLR. arXiv:2009.07612* (2020).

[6] Hanbaek Lyu et al. "Learning low-rank latent mesoscale structures in networks". In: *arXiv preprint arXiv:2102.06984* (2021).

[7] Julien Mairal et al. "Non-local sparse models for image restoration". In: *2009 IEEE 12th international conference on computer vision*. IEEE. 2009, pp. 2272–2279.