Convergence and complextiy of stochastic block majorization-minimization

Hanbaek Lyu

Department of Mathematics, IFDS
University of Wisconsin - Madison

Partially supported by NSF DMS #2010035

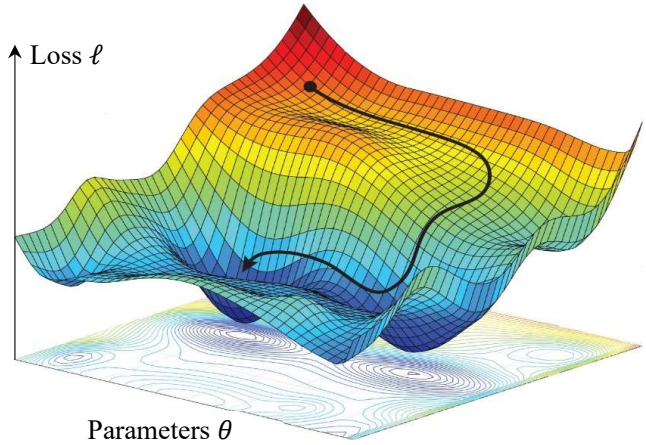
KIAS

Jun. 13, 2022

Outline

Why Optimization?

- Optimization is a fundamental task whenever there is data to be explained by a model with parameters
- Data $\approx$ 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

Methods of Least Squares

- Least Squares: Classical setting for linear regression

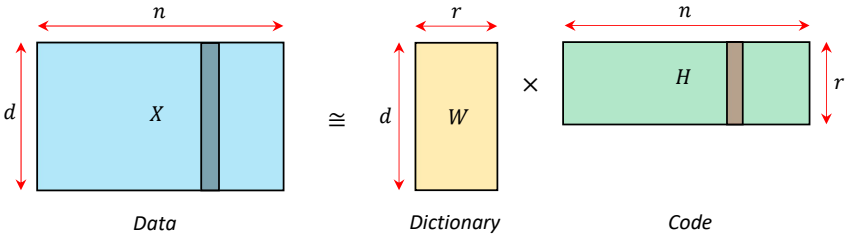
$$\min_{\mathbf{H}} \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2$$

Methods of Least Squares

- Least Squares: Classical setting for linear regression

$$\min_{\mathbf{H}} \|\mathbf{X} - \mathbf{WH}\|_F^2$$

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

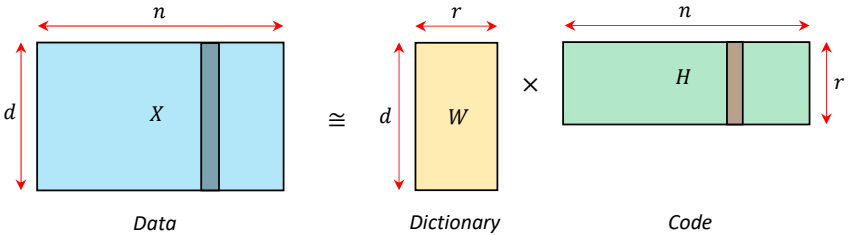


Methods of Least Squares

- Least Squares: Classical setting for linear regression

$$\min_{\mathbf{H}} \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2$$

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



- Convex optimization problem with closed-form solution (when $\mathbf{W}$ has full-rank):

$$\hat{\mathbf{H}} = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{X}$$

Methods of Least Squares

- Nonnegative Least Squares: Require nonnegative linear representation over the basis

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

Methods of Least Squares

- Nonnegative Least Squares: Require nonnegative linear representation over the basis

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

  - Convex optimization problem withj convex constraint ($\mathbf{\Theta} = \mathbb{R}^{r\times n}_{\geq 0}$)

Methods of Least Squares

- Nonnegative Least Squares: Require nonnegative linear representation over the basis

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

  - Convex optimization problem withj convex constraint ($\mathbf{\Theta} = \mathbb{R}^{r\times n}_{\geq 0}$)
  - Can be solved iteratively by Projected Gradient Descent (PGD):

$$\begin{aligned}\mathbf{H}_{t+1} &\leftarrow \mathrm{Proj}_{\mathbf{\Theta}}\left(\mathbf{H}_t - \eta_t \nabla f(\mathbf{H}_t)\right)\\ &= \max\left(\mathbf{0}, \mathbf{H}_t - \eta_t \mathbf{W}^T(\mathbf{W}\mathbf{H}_n - \mathbf{X})\right)\end{aligned}$$

Methods of Least Squares

- Nonnegative Least Squares: Require nonnegative linear representation over the basis

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

  - Convex optimization problem withj convex constraint ($\mathbf{\Theta} = \mathbb{R}^{r\times n}_{\geq 0}$)
  - Can be solved iteratively by Projected Gradient Descent (PGD):

$$\begin{aligned}\mathbf{H}_{t+1} &\leftarrow \mathrm{Proj}_{\mathbf{\Theta}}\left(\mathbf{H}_t - \eta_t \nabla f(\mathbf{H}_t)\right)\\ &= \max\left(\mathbf{0}, \mathbf{H}_t - \eta_t \mathbf{W}^T(\mathbf{W}\mathbf{H}_n - \mathbf{X})\right)\end{aligned}$$

  - PGD finds 'ε-accuracte' global minimizer within $O(\varepsilon^{-1})$ iterations

Matrix Factorization

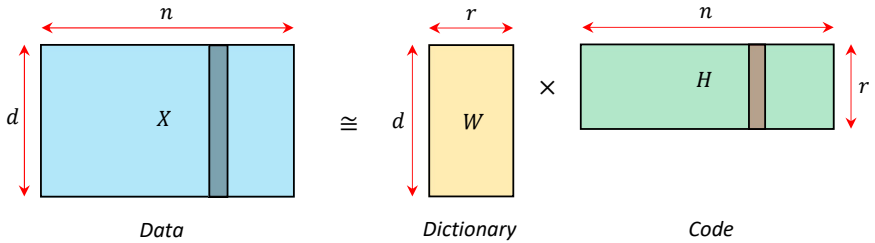
- Q: What if we don't know what basis features $\mathbf{W}$ to use?

Matrix Factorization

- Q: What if we don't know what basis features $\mathbf{W}$ to use?
  - Simultaneously find the basis $\mathbf{W}$ and the linear representation $\mathbf{H}$ for the data $\mathbf{X}$?

Matrix Factorization

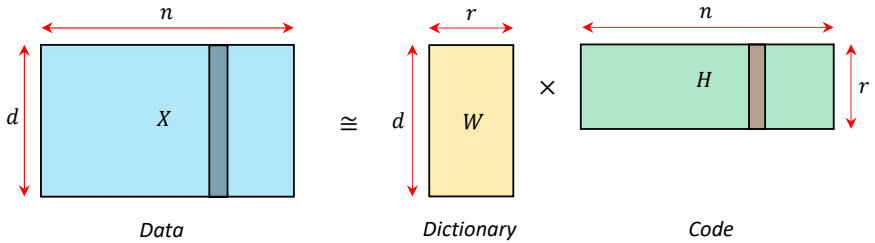
- Q: What if we don't know what basis features $\mathbf{W}$ to use?
  - Simultaneously find the basis $\mathbf{W}$ and the linear representation $\mathbf{H}$ for the data $\mathbf{X}$?
- Matrix factorization is a fundamental tool in dictionary learning problems.



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

Matrix Factorization

- Q: What if we don't know what basis features $\mathbf{W}$ to use?
  - Simultaneously find the basis $\mathbf{W}$ and the linear representation $\mathbf{H}$ for the data $\mathbf{X}$?
- Matrix factorization is a fundamental tool in dictionary learning problems.



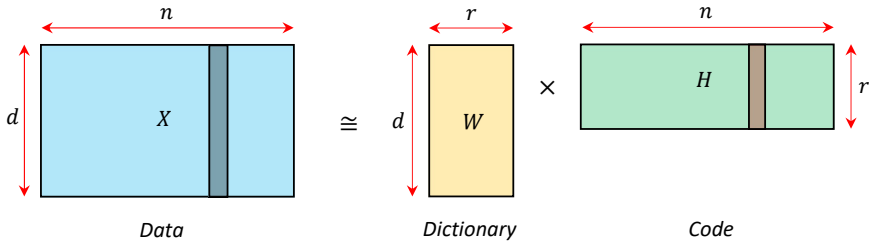
Data ≈ Linear combination of $\overbrace{\text{latent features}}^{\text{cols. of } W}$

- Formulated as a nonconvex optimization problem:

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

Matrix Factorization

- Q: What if we don't know what basis features $\mathbf{W}$ to use?
  - Simultaneously find the basis $\mathbf{W}$ and the linear representation $\mathbf{H}$ for the data $\mathbf{X}$?
- Matrix factorization is a fundamental tool in dictionary learning problems.



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

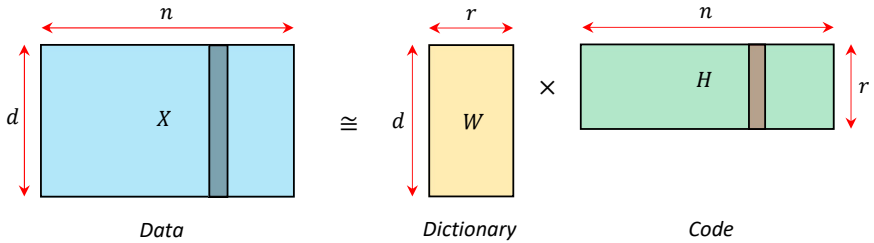
- Formulated as a nonconvex optimization problem:

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

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

Matrix Factorization

- Q: What if we don't know what basis features $\mathbf{W}$ to use?
  - Simultaneously find the basis $\mathbf{W}$ and the linear representation $\mathbf{H}$ for the data $\mathbf{X}$?
- Matrix factorization is a fundamental tool in dictionary learning problems.



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

- Formulated as a nonconvex optimization problem:

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

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

Matrix Factorization

- How do we solve NMF?

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

Matrix Factorization

- How do we solve NMF?

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

  - Can't find both $\mathbf{W}$ and $\mathbf{H}$ at the same time, so alternate!

$$\mathbf{H}_{t+1} \leftarrow \operatorname*{argmin}_{\mathbf{H}\in\mathbb{R}^{r\times n}_{\geq 0}} f(\mathbf{W}_t,\mathbf{H}) \qquad (NLS)$$

$$\mathbf{W}_{t+1} \leftarrow \operatorname*{argmin}_{\mathbf{W}\in\mathbb{R}^{d\times r}_{\geq 0}} f(\mathbf{W},\mathbf{H}_{t+1}) \qquad (NLS)$$

Matrix Factorization

- How do we solve NMF?

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

  - Can't find both $\mathbf{W}$ and $\mathbf{H}$ at the same time, so alternate!

$$\mathbf{H}_{t+1} \leftarrow \operatorname*{argmin}_{\mathbf{H}\in\mathbb{R}^{r\times n}_{\geq 0}} f(\mathbf{W}_t,\mathbf{H}) \qquad (NLS)$$

$$\mathbf{W}_{t+1} \leftarrow \operatorname*{argmin}_{\mathbf{W}\in\mathbb{R}^{d\times r}_{\geq 0}} f(\mathbf{W},\mathbf{H}_{t+1}) \qquad (NLS)$$

  - Block Coordinate Descent for NMF (a.k.a. Alternating Least Squares)

Matrix Factorization

- How do we solve NMF?

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

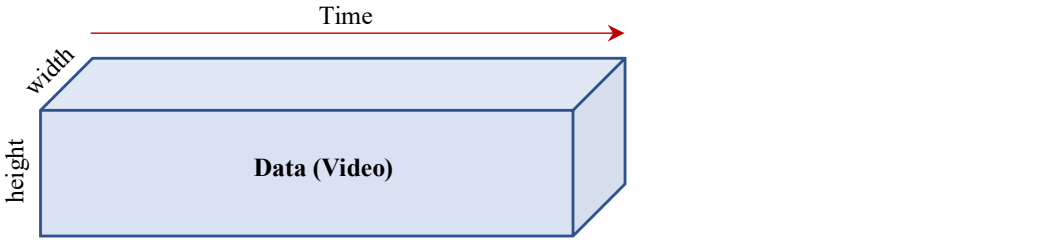
  - Can't find both $\mathbf{W}$ and $\mathbf{H}$ at the same time, so alternate!

$$\mathbf{H}_{t+1} \leftarrow \operatorname*{argmin}_{\mathbf{H}\in\mathbb{R}^{r\times n}_{\geq 0}} f(\mathbf{W}_t,\mathbf{H}) \qquad (NLS)$$

$$\mathbf{W}_{t+1} \leftarrow \operatorname*{argmin}_{\mathbf{W}\in\mathbb{R}^{d\times r}_{\geq 0}} f(\mathbf{W},\mathbf{H}_{t+1}) \qquad (NLS)$$

  - Block Coordinate Descent for NMF (a.k.a. Alternating Least Squares)
  - NOT guaranteed to converge to global optimum (will come back to this point later)

Dictionary Learning from Video Frames



$\approx$

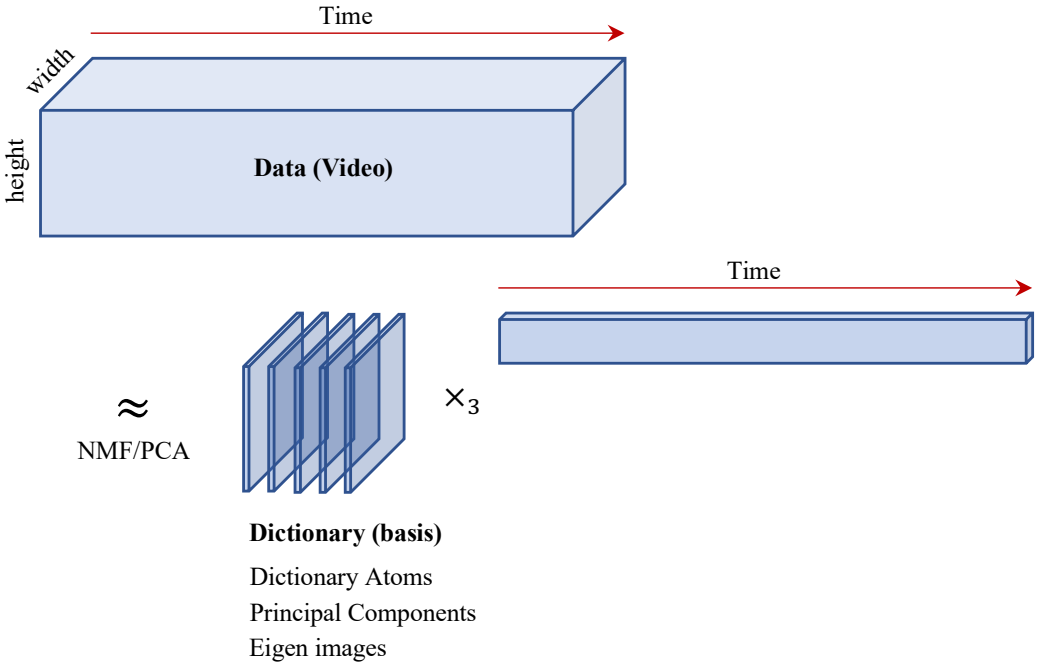
NMF/PCA

$\times_3$

Dictionary (basis)

Dictionary Atoms
Principal Components
Eigen images

Dictionary Learning from Video Frames

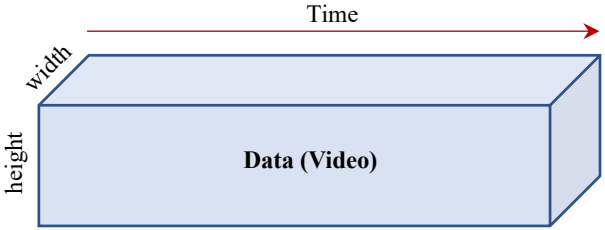


- Entire video frames are processed at once (batch processing)

A Toy Example Video

Figure: Bruce Lee (doing his stuff)

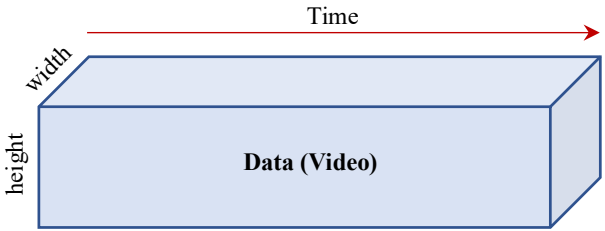
Dictionary Learning from Video Frames



Five Dictionary Atoms

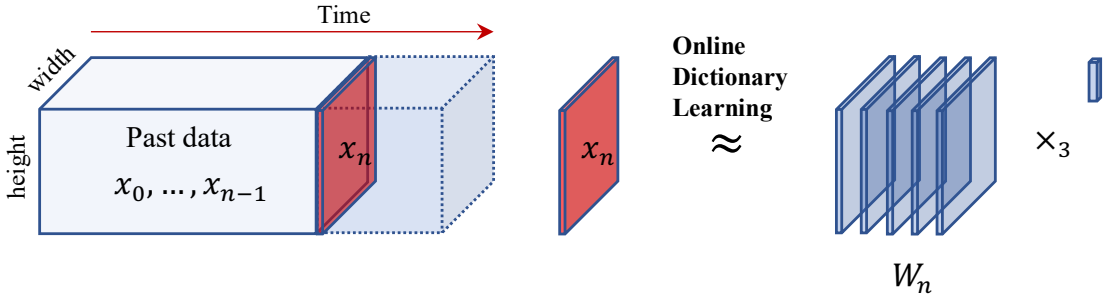
NMF
≈

Dictionary Learning from Video Frames



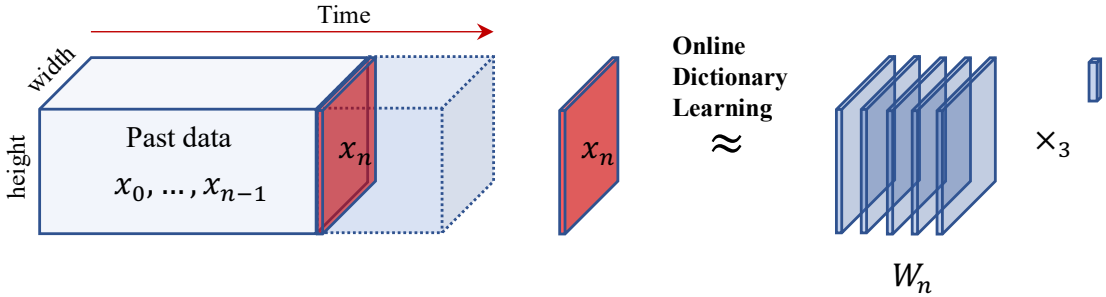
Online Dictionary Learning

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



Online Dictionary Learning

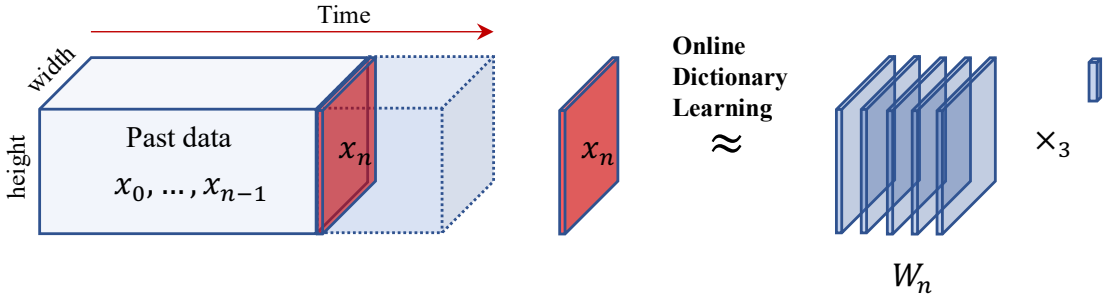
- Instead processing the entire frames at once, can we process one image at a time to learn the dictionary? (mini-batch processing)



- Why do 'online learning'?

Online Dictionary Learning

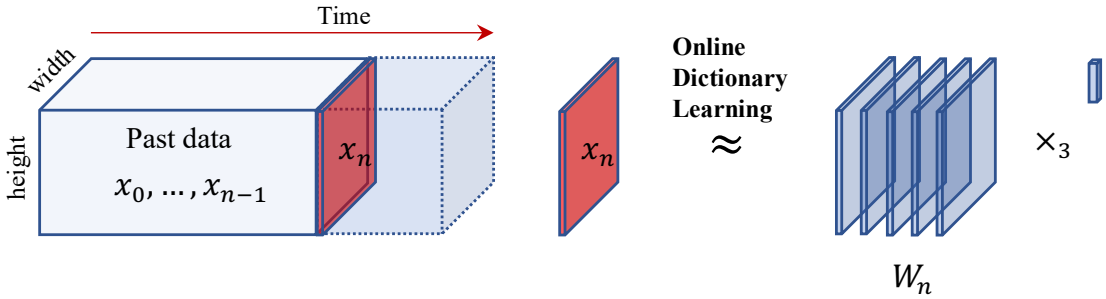
- Instead processing the entire frames at once, can we process one image at a time to learn the dictionary? (mini-batch processing)



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

Online Dictionary Learning

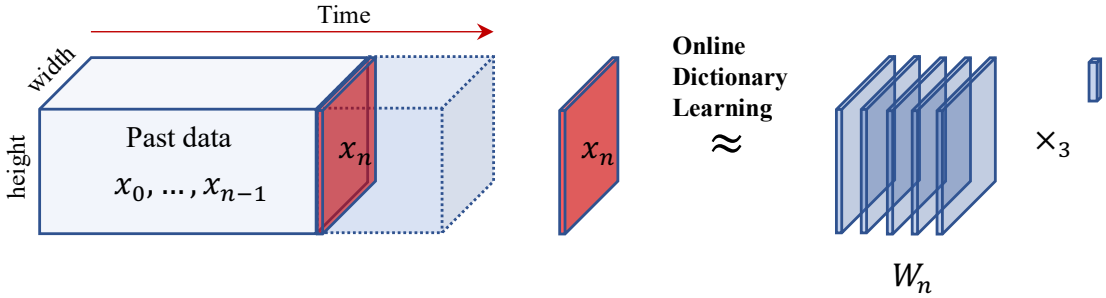
- Instead processing the entire frames at once, can we process one image at a time to learn the dictionary? (mini-batch processing)



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

Online Dictionary Learning

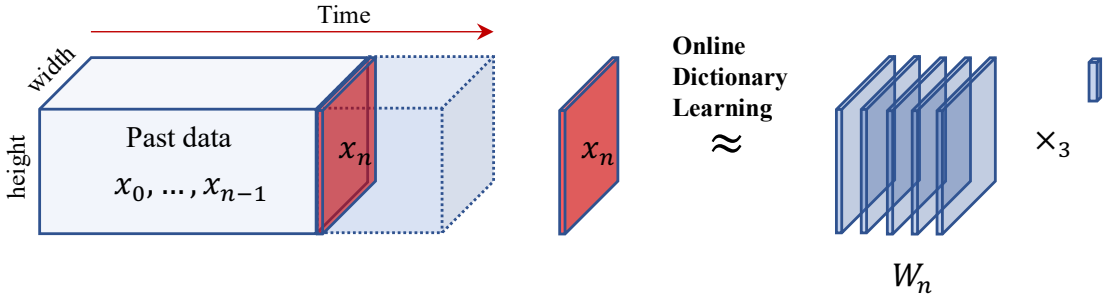
- Instead processing the entire frames at once, can we process one image at a time to learn the dictionary? (mini-batch processing)



- 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

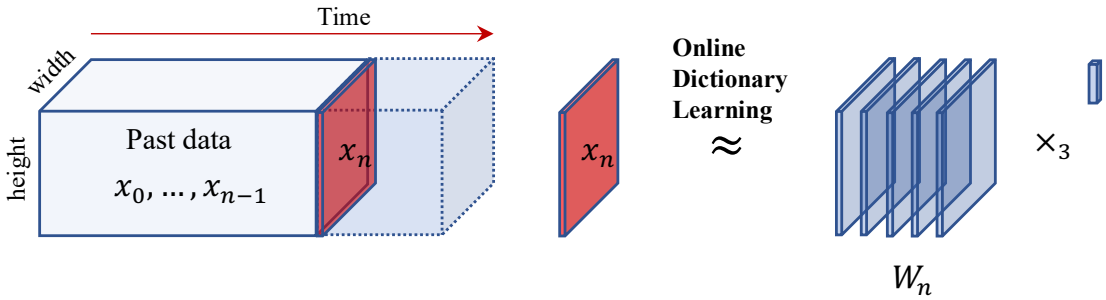
- Instead processing the entire frames at once, can we process one image at a time to learn the dictionary? (mini-batch processing)



- 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 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

Empirical Loss Minimization

- Empirical Loss Minimization

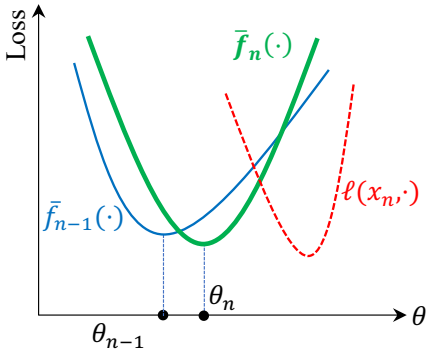
$$\text{Upon arrival of } \mathbf{x}_n\text{:} \quad \boldsymbol{\theta}_n \in \operatorname*{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \Big(\bar{f}_n(\boldsymbol{\theta}) := (1-w_n) \underbrace{\bar{f}_{n-1}(\boldsymbol{\theta})}_{\text{old loss}} + w_n \underbrace{\ell(\mathbf{x}_n, \boldsymbol{\theta})}_{\text{new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

Empirical Loss Minimization

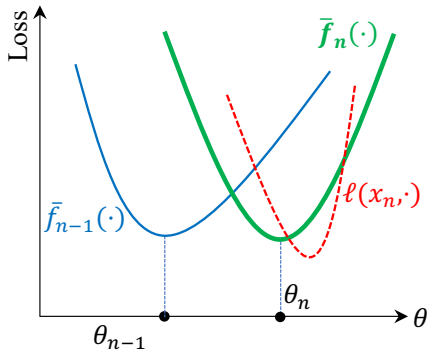
- Empirical Loss Minimization

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

- 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



Slow adaptation



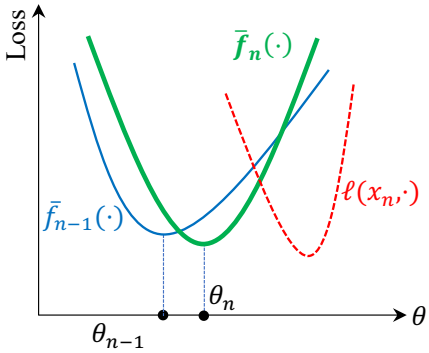
Fast adaptation

Empirical Loss Minimization

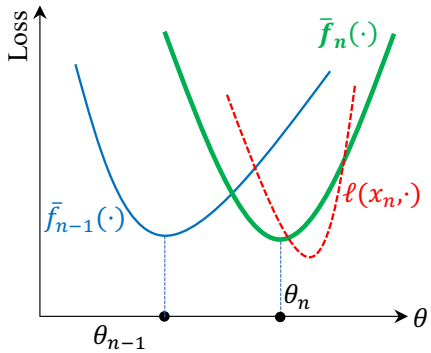
- Empirical Loss Minimization

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

- 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)



Slow adaptation



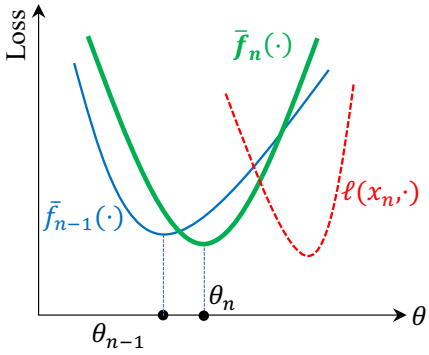
Fast adaptation

Empirical Loss Minimization

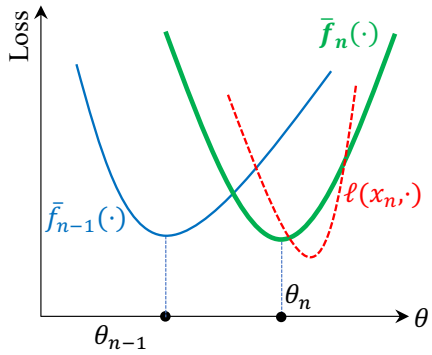
- Empirical Loss Minimization

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

- 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)
  - Slow-adapting $w_n \Rightarrow$ learn long-time scale features (could be smoothed out too much)



Slow adaptation



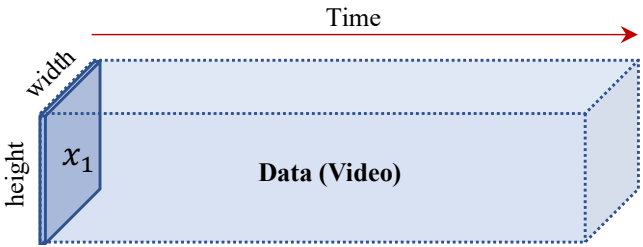
Fast adaptation

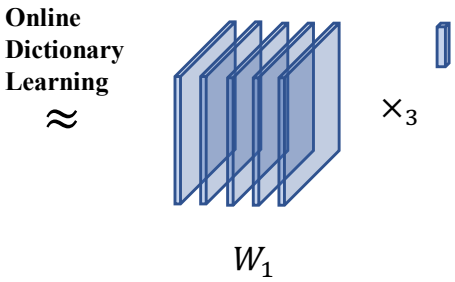
(a) past2future + fast adaptation

(b) past2future + slow adaptation

Online Dictionary Learning (past2future sampling)

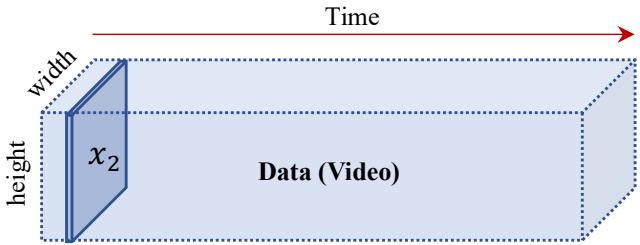
- How do we sample image frames $\mathbf{x}_n$? — Past2future
  - In the truely online setting, one needs to process the data as they arrive
  - Hard to analyze theoretically

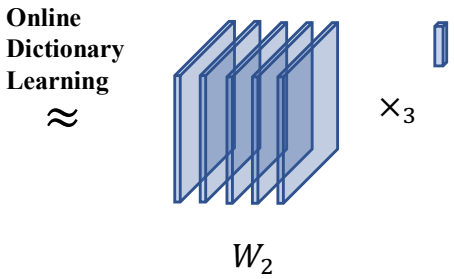




Online Dictionary Learning (past2future sampling)

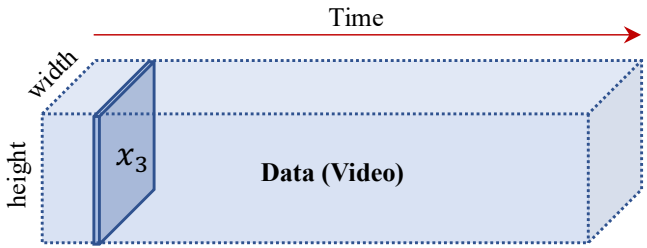
- How do we sample image frames $\mathbf{x}_n$? — Past2future
  - In the truely online setting, one needs to process the data as they arrive
  - Hard to analyze theoretically

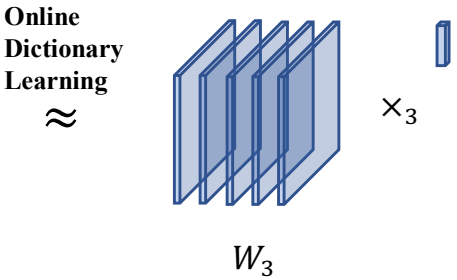




Online Dictionary Learning (past2future sampling)

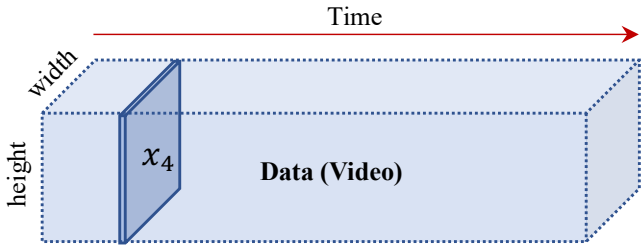
- How do we sample image frames $\mathbf{x}_n$? — Past2future
  - In the truely online setting, one needs to process the data as they arrive
  - Hard to analyze theoretically

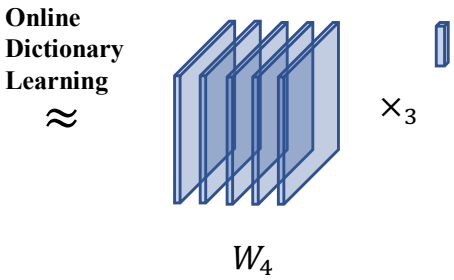




Online Dictionary Learning (past2future sampling)

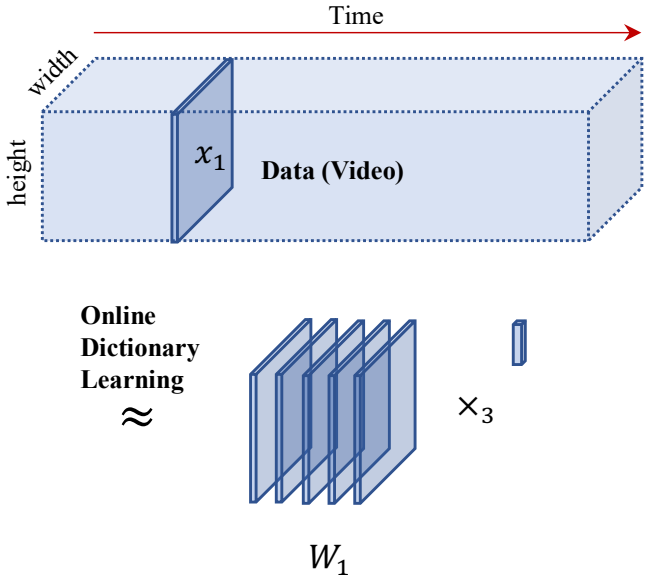
- How do we sample image frames $\mathbf{x}_n$? — Past2future
  - In the truely online setting, one needs to process the data as they arrive
  - Hard to analyze theoretically





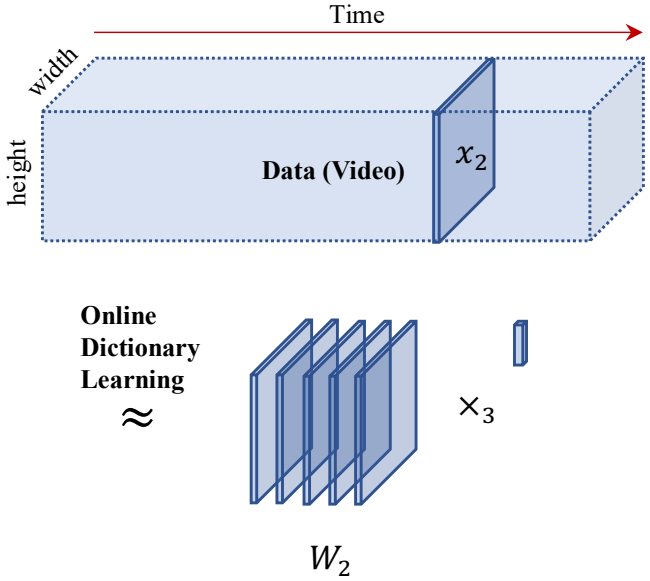
Online Dictionary Learning (uniform i.i.d. sampling)

- How do we sample image frames $\mathbf{x}_n$? — Uniformly i.i.d. among all frames
  - Requires a priori access to the full data;
  - (Relatively) easy to analyze theoretically;
  - Hard to guarantee/computationally expensive



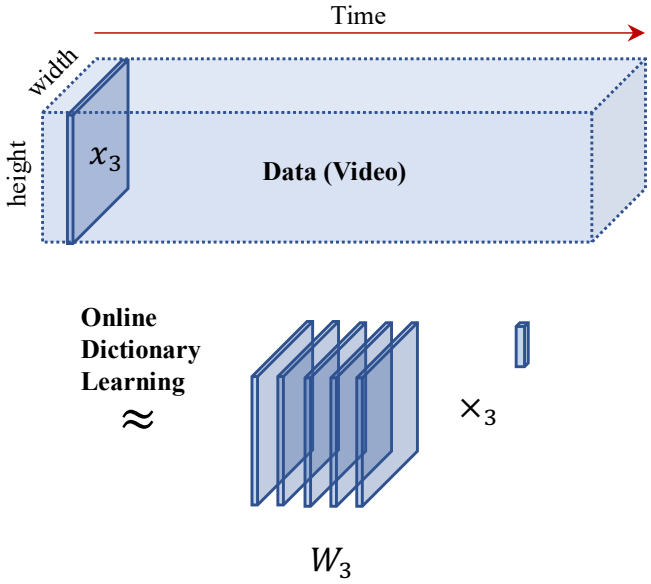
Online Dictionary Learning (uniform i.i.d. sampling)

- How do we sample image frames $\mathbf{x}_n$? — Uniformly i.i.d. among all frames
  - Requires a priori access to the full data;
  - (Relatively) easy to analyze theoretically;
  - Hard to guarantee/computationally expensive



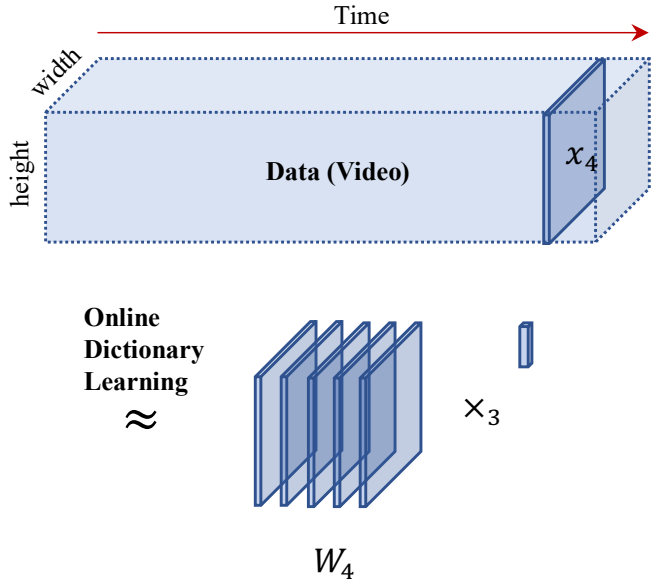
Online Dictionary Learning (uniform i.i.d. sampling)

- How do we sample image frames $\mathbf{x}_n$? — Uniformly i.i.d. among all frames
  - Requires a priori access to the full data;
  - (Relatively) easy to analyze theoretically;
  - Hard to guarantee/computationally expensive



Online Dictionary Learning (uniform i.i.d. sampling)

- How do we sample image frames $\mathbf{x}_n$? — Uniformly i.i.d. among all frames
  - Requires a priori access to the full data;
  - (Relatively) easy to analyze theoretically;
  - Hard to guarantee/computationally expensive



(a) past2future + fast adaptation

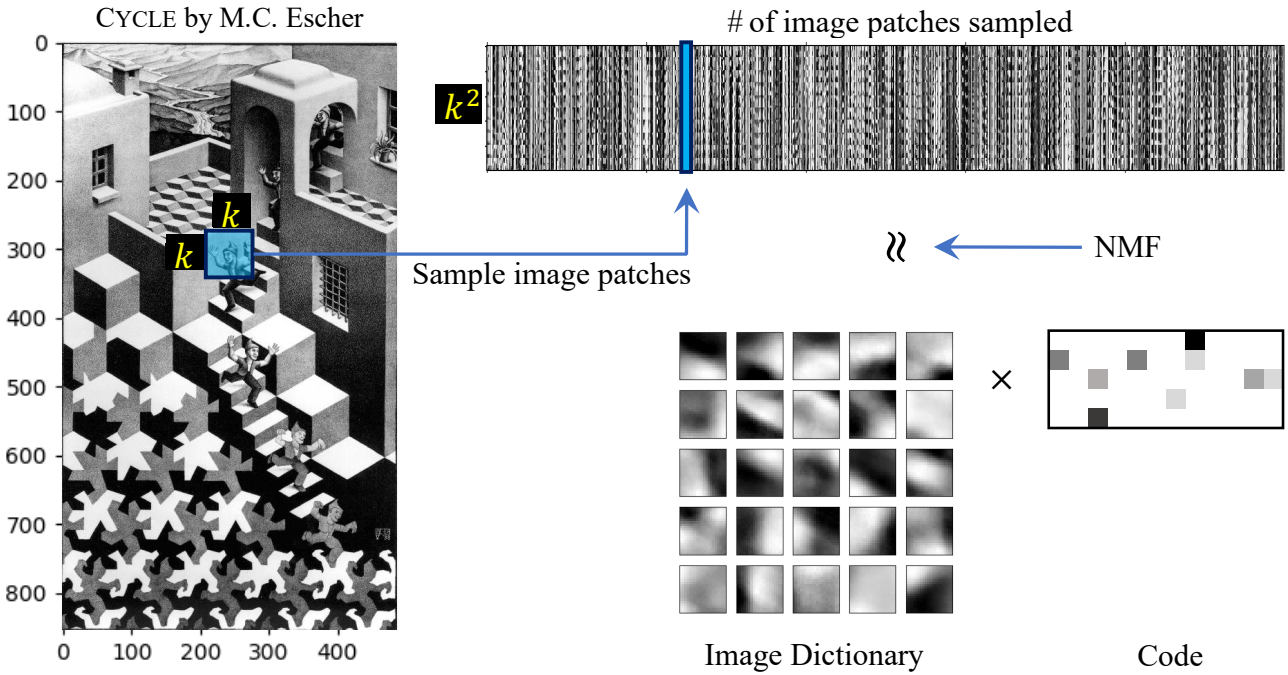
(b) past2future + slow adaptation

(c) i.i.d. + fast adaptation

(d) i.i.d. + slow adaptation

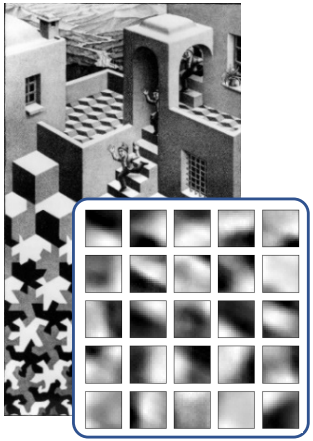
Outline

Example of NMF for Image dictionary learning



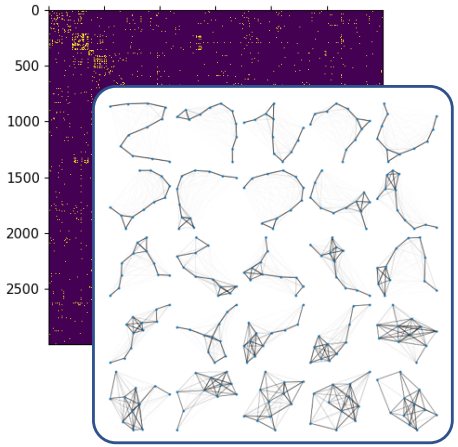
Network Dictionary Learning (NDL)

CYCLE by M.C. Escher



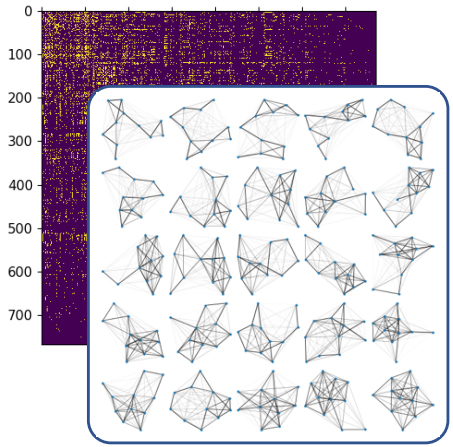
a Image Dictionary

UCLA Facebook Network



b Network Dictionary

CALTECH Facebook Network



c Network Dictionary

- NDL: Network data $\longrightarrow$ Latent motifs (nonnegative basis for subgraphs)
  - First introduced in L., Needell, Balzano [3]
  - Further developed in L., Kureh, Vendrow, Porter [5]

Dictionary Learning with Subgraphs

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

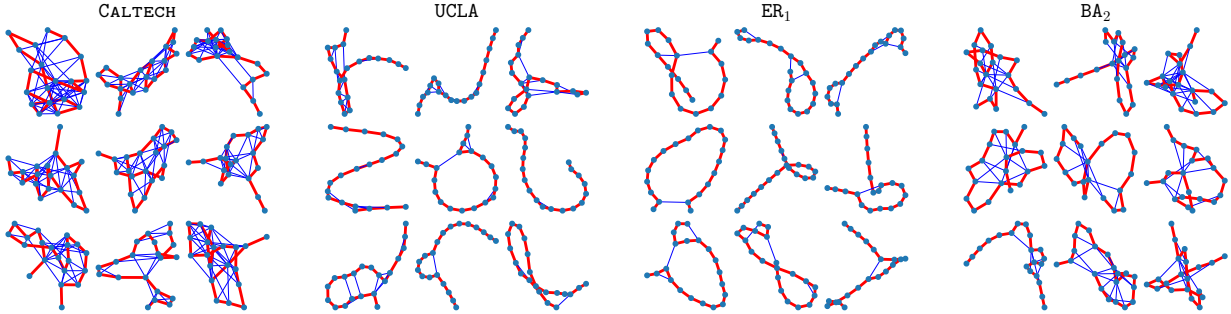


Figure: From L., Kureh, Vendrow, Porter '22+

Dictionary Learning with Subgraphs

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

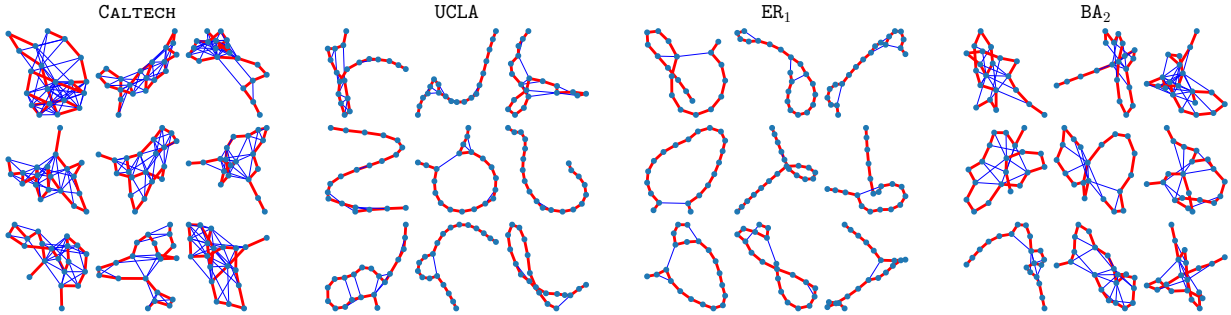


Figure: From L., Kureh, Vendrow, Porter '22+

- 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

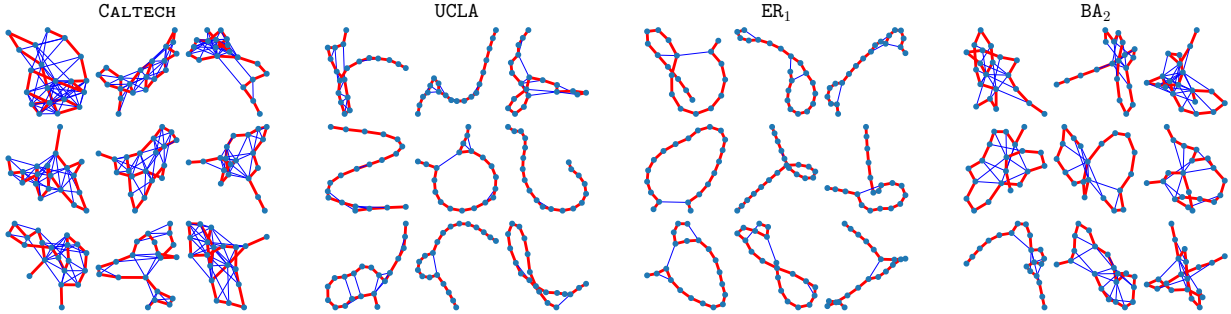


Figure: From L., Kureh, Vendrow, Porter '22+

- 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

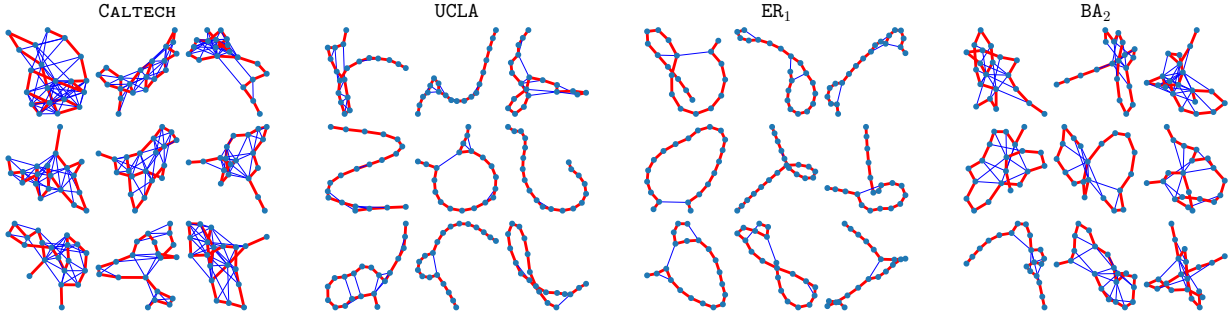


Figure: From L., Kureh, Vendrow, Porter '22+

- 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

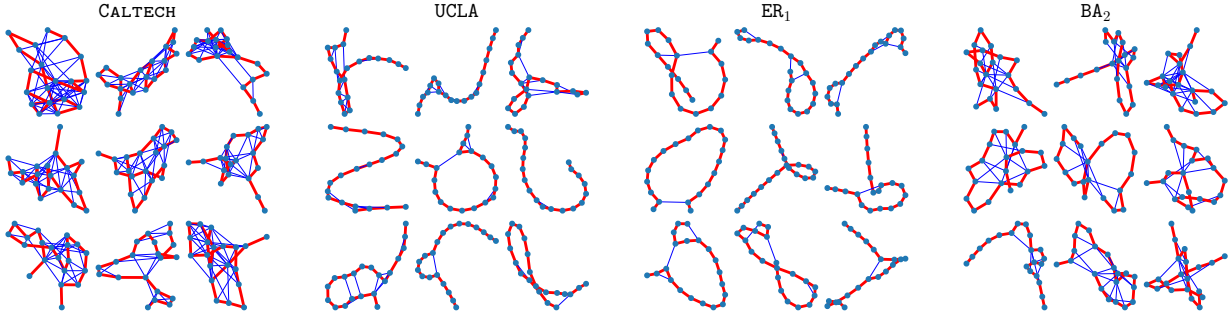
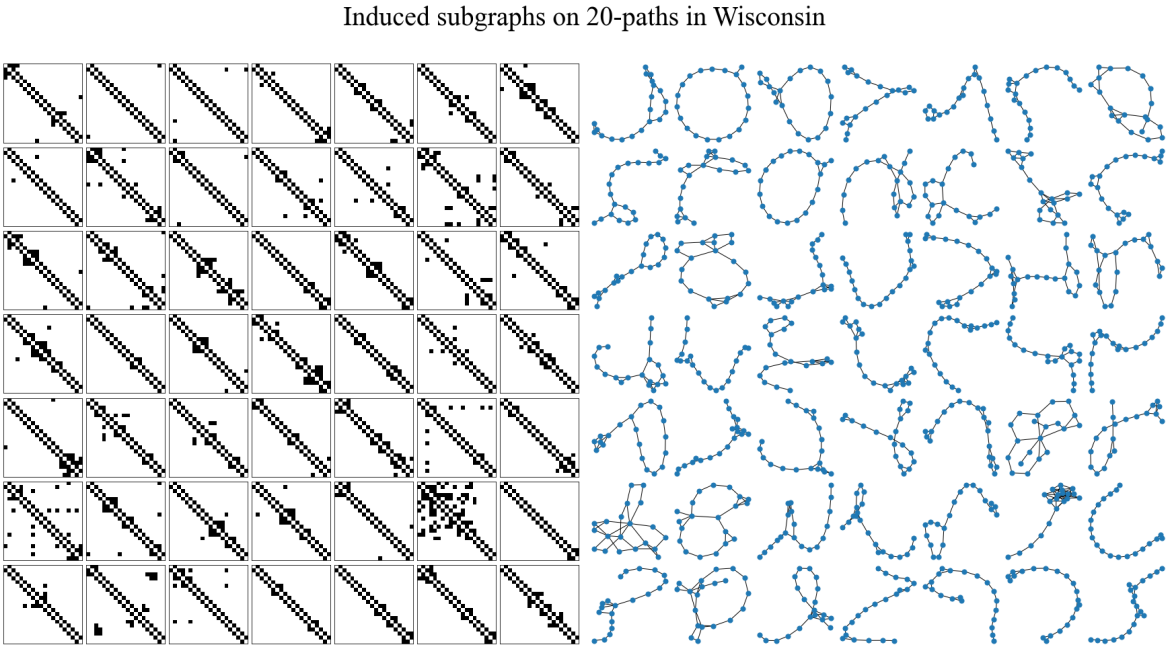


Figure: From L., Kureh, Vendrow, Porter '22+

- 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)

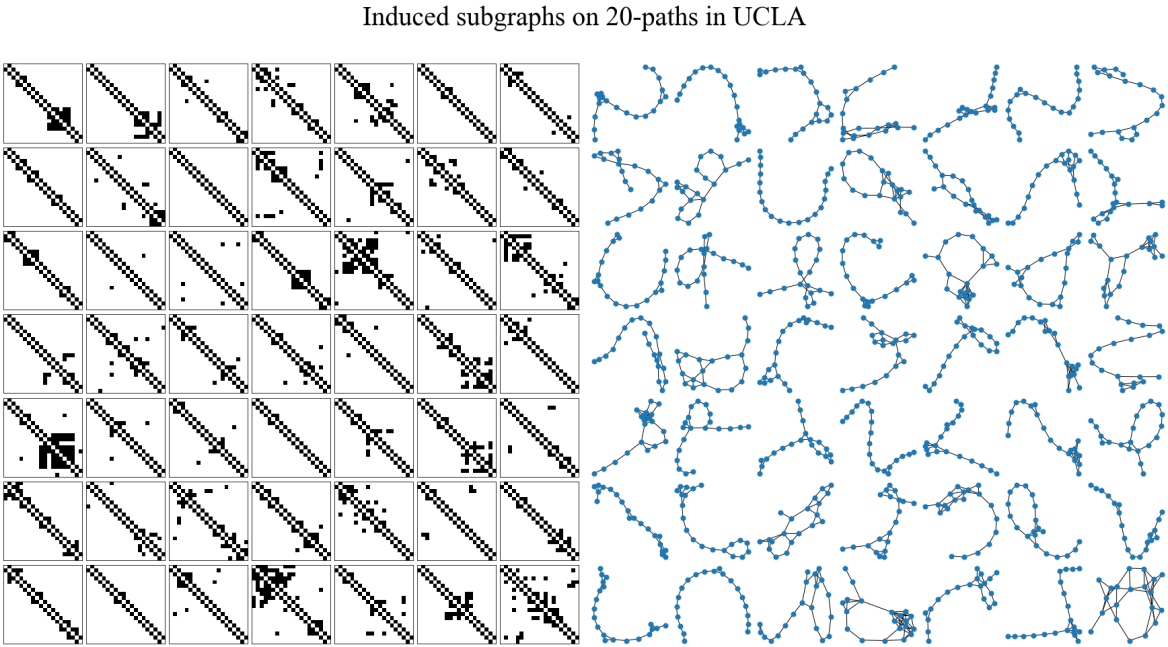
Dictionary Learning with Network Subgraphs

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



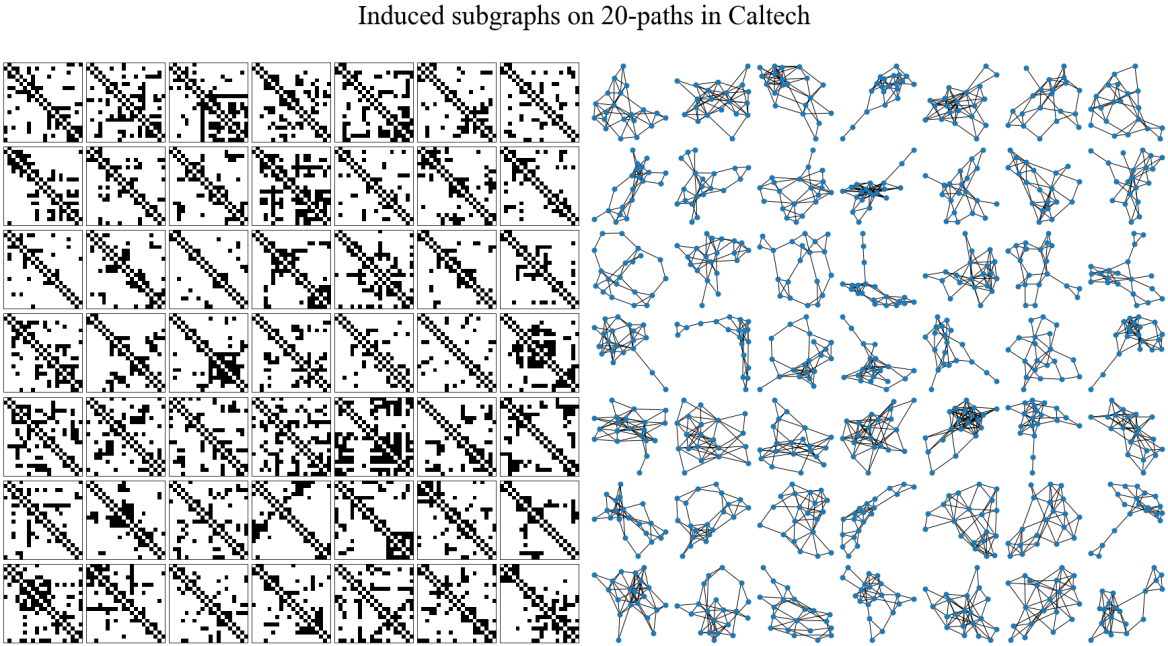
Dictionary Learning with Network Subgraphs

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



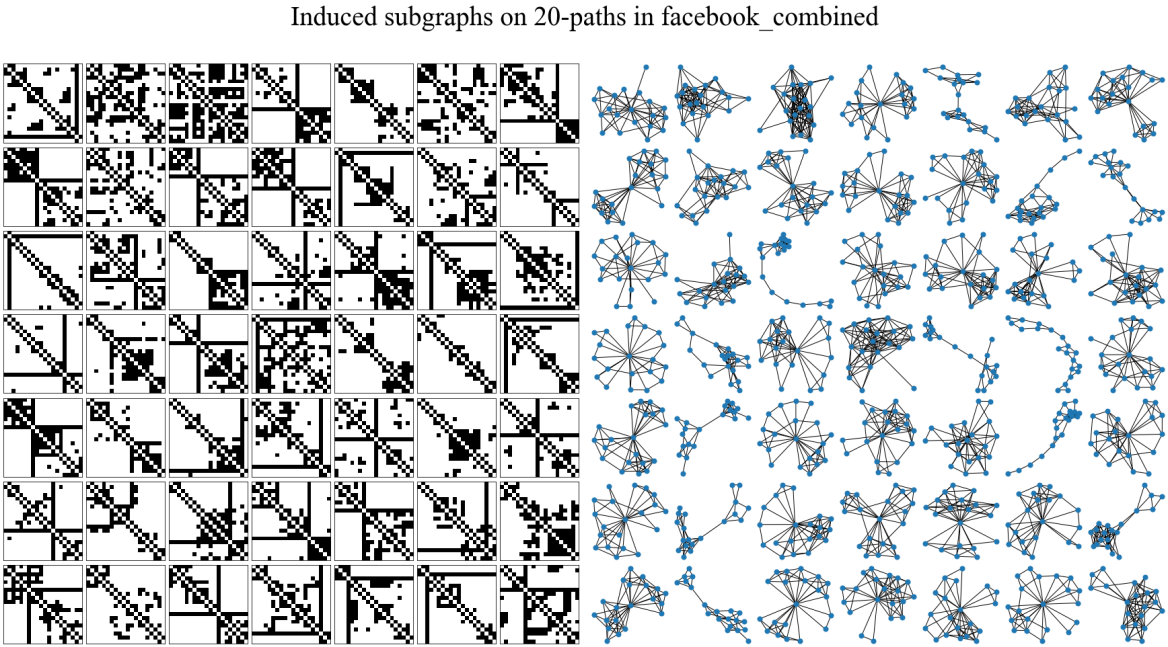
Dictionary Learning with Network Subgraphs

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



Dictionary Learning with Network Subgraphs

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



(a) arXiv

(b) Facebook

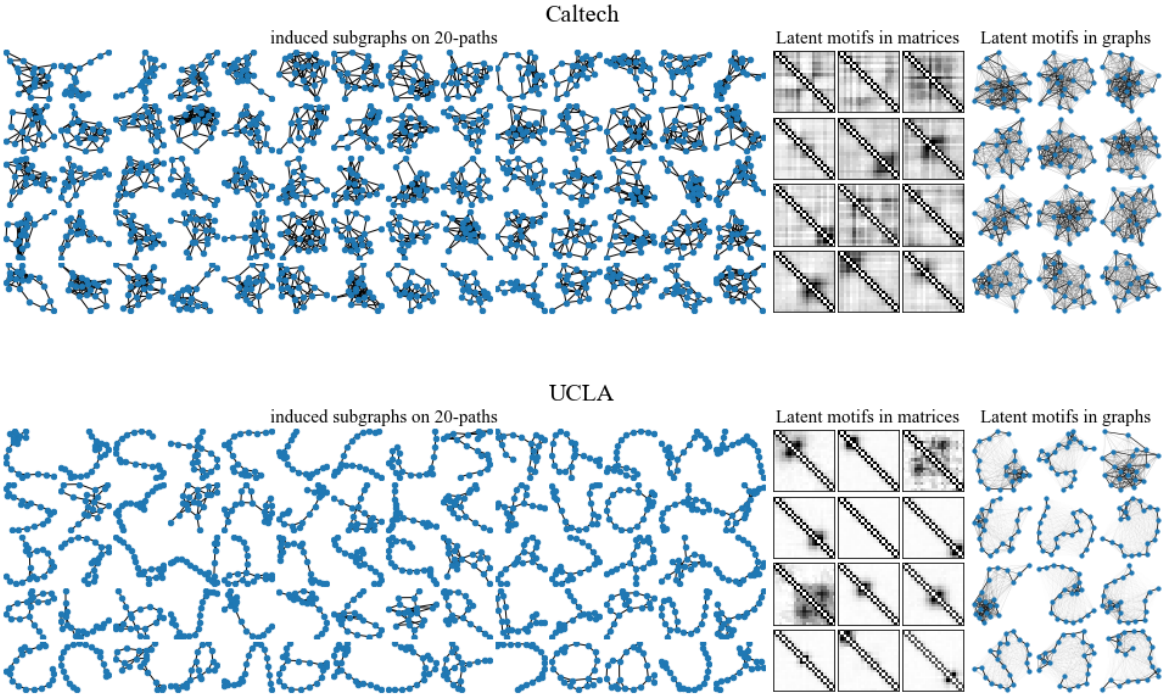
(c) Caltech

(d) UCLA

(e) UW-Madison

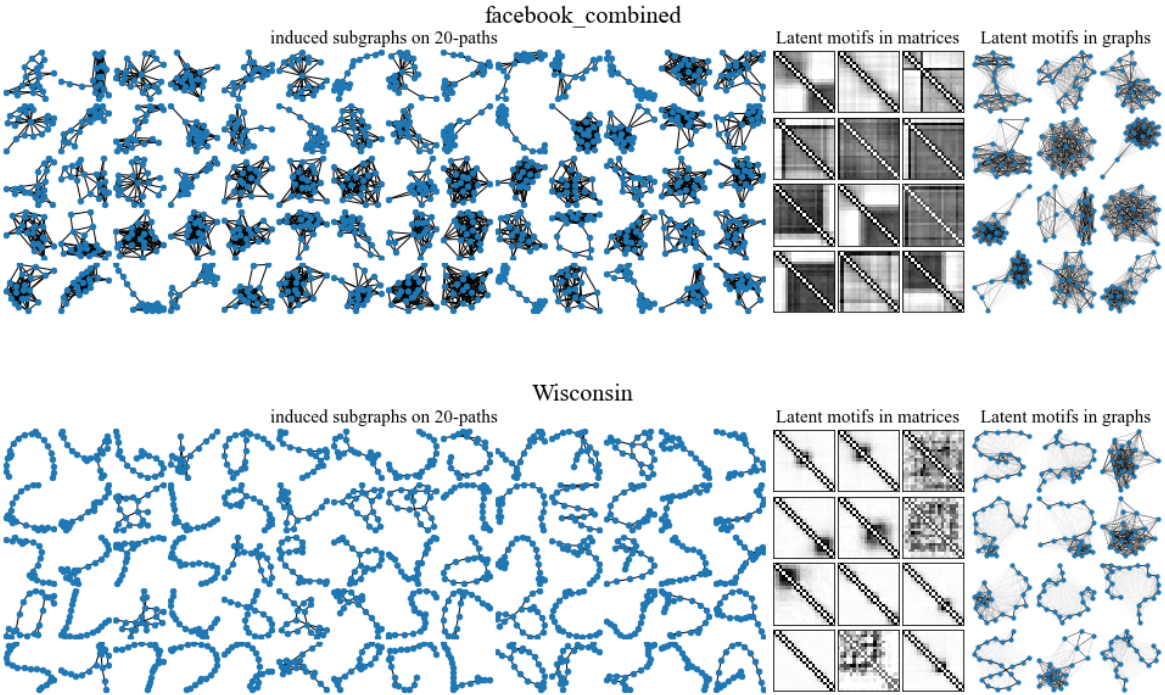
Dictionary Learning with Network Subgraphs

- Online NMF + MCMC subgraph sampling



Dictionary Learning with Network Subgraphs

- Online NMF + MCMC subgraph sampling



Network Dictionary Learning (NDL)

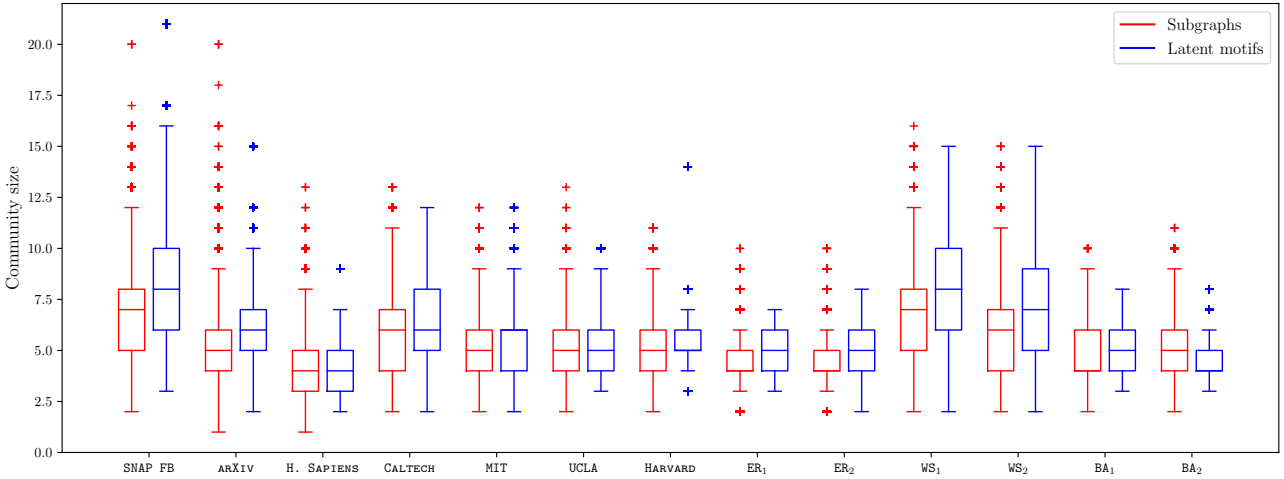


Figure: Comparing community sizes in 10K random subgraphs vs. 25 latent motifs

- NDL: Network data $\longrightarrow$ Latent motifs (nonnegative basis for subgraphs)
  - First introduced in L., Needell, Balzano [3]
  - Further developed in L., Kureh, Vendrow, Porter [5]

Outline

Empirical/Expected Loss Minimization

- Empirical Loss Minimization

Empirical/Expected Loss Minimization

- Empirical Loss Minimization
  - Given data sequence $(\mathbf{x}_n)_{n\geq 1}$, loss function $\ell(\mathbf{x},\boldsymbol{\theta})$, adaptivity weights $(w_n)_{n\geq 1}$:

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

Empirical/Expected Loss Minimization

- Empirical Loss Minimization
  - Given data sequence $(\mathbf{x}_n)_{n\geq 1}$, loss function $\ell(\mathbf{x},\boldsymbol{\theta})$, adaptivity weights $(w_n)_{n\geq 1}$:

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

  - Special cases:

Empirical/Expected Loss Minimization

- Empirical Loss Minimization
  - Given data sequence $(\mathbf{x}_n)_{n\geq 1}$, loss function $\ell(\mathbf{x},\boldsymbol{\theta})$, adaptivity weights $(w_n)_{n\geq 1}$:

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

  - Special cases:

$$w_n = \frac{1}{n} \quad\Longrightarrow\quad \bar{f}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{k=1}^{n} \ell(\mathbf{x}_k,\boldsymbol{\theta}) \qquad \text{Slow adaptation}$$

Empirical/Expected Loss Minimization

- Empirical Loss Minimization
  - Given data sequence $(\mathbf{x}_n)_{n\geq 1}$, loss function $\ell(\mathbf{x},\boldsymbol{\theta})$, adaptivity weights $(w_n)_{n\geq 1}$:

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

  - Special cases:

$$w_n = \frac{1}{n} \implies \bar{f}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{k=1}^n \ell(\mathbf{x}_k,\boldsymbol{\theta}) \qquad \text{Slow adaptation}$$

$$w_n \equiv \alpha \in (0,1) \implies \bar{f}_n(\boldsymbol{\theta}) = \sum_{k=1}^n \ell(\mathbf{x}_k,\boldsymbol{\theta})\alpha(1-\alpha)^{n-k} \qquad \text{Fast adaptation}$$

Empirical/Expected Loss Minimization

- Empirical Loss Minimization
  - Given data sequence $(\mathbf{x}_n)_{n\geq 1}$, loss function $\ell(\mathbf{x},\boldsymbol{\theta})$, adaptivity weights $(w_n)_{n\geq 1}$:

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

  - Special cases:

    $w_n = \frac{1}{n} \implies \bar{f}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{k=1}^n \ell(\mathbf{x}_k,\boldsymbol{\theta})$ Slow adaptation

    $w_n \equiv \alpha \in (0,1) \implies \bar{f}_n(\boldsymbol{\theta}) = \sum_{k=1}^n \ell(\mathbf{x}_k,\boldsymbol{\theta})\alpha(1-\alpha)^{n-k}$ Fast adaptation

- Expected Loss Minimization

Empirical/Expected Loss Minimization

- Empirical Loss Minimization
  - Given data sequence $(\mathbf{x}_n)_{n\geq 1}$, loss function $\ell(\mathbf{x},\boldsymbol{\theta})$, adaptivity weights $(w_n)_{n\geq 1}$:

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

  - Special cases:

$$w_n = \frac{1}{n} \implies \bar{f}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{k=1}^n \ell(\mathbf{x}_k,\boldsymbol{\theta}) \qquad \text{Slow adaptation}$$

$$w_n \equiv \alpha \in (0,1) \implies \bar{f}_n(\boldsymbol{\theta}) = \sum_{k=1}^n \ell(\mathbf{x}_k,\boldsymbol{\theta})\alpha(1-\alpha)^{n-k} \qquad \text{Fast adaptation}$$

- Expected Loss Minimization
  -

$$\boldsymbol{\theta}^* \in \operatorname*{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left(\underbrace{f}_{\text{expected loss}} (\boldsymbol{\theta}) := \mathbb{E}_{\mathbf{x}\sim\pi}[\ell(\mathbf{x},\boldsymbol{\theta})] \right).$$

Empirical/Expected Loss Minimization

- Empirical Loss Minimization
  - Given data sequence $(\mathbf{x}_n)_{n\geq 1}$, loss function $\ell(\mathbf{x},\boldsymbol{\theta})$, adaptivity weights $(w_n)_{n\geq 1}$:

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

  - Special cases:

$$w_n = \frac{1}{n} \implies \bar{f}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{k=1}^n \ell(\mathbf{x}_k,\boldsymbol{\theta}) \qquad \text{Slow adaptation}$$

$$w_n \equiv \alpha \in (0,1) \implies \bar{f}_n(\boldsymbol{\theta}) = \sum_{k=1}^n \ell(\mathbf{x}_k,\boldsymbol{\theta})\alpha(1-\alpha)^{n-k} \qquad \text{Fast adaptation}$$

- Expected Loss Minimization
  -

$$\boldsymbol{\theta}^* \in \operatorname*{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left(\underbrace{f}_{\text{expected loss}} (\boldsymbol{\theta}) := \mathbb{E}_{\mathbf{x}\sim\pi}[\ell(\mathbf{x},\boldsymbol{\theta})] \right).$$

  - Can only model the slowest time-scale — Stationary features

Empirical/Expected Loss Minimization

- Empirical Loss Minimization
  - Given data sequence $(\mathbf{x}_n)_{n\geq 1}$, loss function $\ell(\mathbf{x},\boldsymbol{\theta})$, adaptivity weights $(w_n)_{n\geq 1}$:

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

- Expected Loss Minimization
  -

$$\boldsymbol{\theta}^* \in \operatorname*{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left(\underbrace{f}_{\text{expected loss}} (\boldsymbol{\theta}) := \mathbb{E}_{\mathbf{x}\sim\pi}\left[\ell(\mathbf{x},\boldsymbol{\theta})\right] \right).$$

Empirical/Expected Loss Minimization

- Empirical Loss Minimization
  - Given data sequence $(\mathbf{x}_n)_{n\geq 1}$, loss function $\ell(\mathbf{x},\boldsymbol{\theta})$, adaptivity weights $(w_n)_{n\geq 1}$:

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

- Expected Loss Minimization
  -

$$\boldsymbol{\theta}^* \in \operatorname*{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left(\underbrace{f}_{\text{expected loss}} (\boldsymbol{\theta}) := \mathbb{E}_{\mathbf{x}\sim\pi}\left[\ell(\mathbf{x},\boldsymbol{\theta})\right] \right).$$

- These two problems are 'equivalent' in the slow adaptation regime:

Empirical/Expected Loss Minimization

- Empirical Loss Minimization
  - Given data sequence $(\mathbf{x}_n)_{n\geq 1}$, loss function $\ell(\mathbf{x},\boldsymbol{\theta})$, adaptivity weights $(w_n)_{n\geq 1}$:

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

- Expected Loss Minimization
  -

$$\boldsymbol{\theta}^* \in \operatorname*{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left(\underbrace{f}_{\text{expected loss}} (\boldsymbol{\theta}) := \mathbb{E}_{\mathbf{x}\sim\pi}[\ell(\mathbf{x},\boldsymbol{\theta})] \right).$$

- These two problems are 'equivalent' in the slow adaptation regime:
  -

$$\left(\begin{array}{c} (\mathbf{x}_n)_{n\geq 1} \text{ are i.i.d. (Markovian) } \sim \pi \\ \frac{1}{n} \leq w_n \ll \frac{1}{\sqrt{n}} \\ \ell \text{ smooth, } \boldsymbol{\Theta} \text{ compact} \end{array} \right) \quad \Longrightarrow \quad \|\bar{f}_n - f\|_\infty, \|\nabla \bar{f}_n - \nabla f\|_\infty \to 0 \text{ a.s.}$$

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:

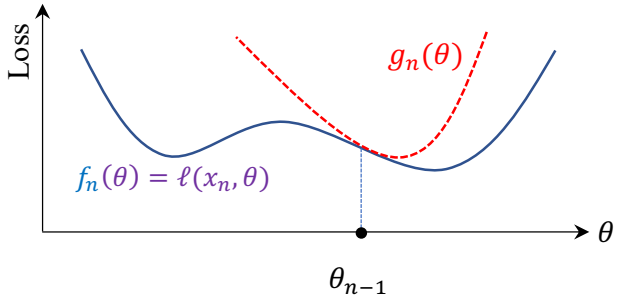
$$\text{(Dictionary Learning)} \qquad \ell(\mathbf{x}_n, \boldsymbol{\theta}) = \inf_{H} \|\mathbf{x}_n - \boldsymbol{\theta} H\|^2$$

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 $\boldsymbol{\theta} \mapsto \ell(\mathbf{x}_n, \boldsymbol{\theta})$:

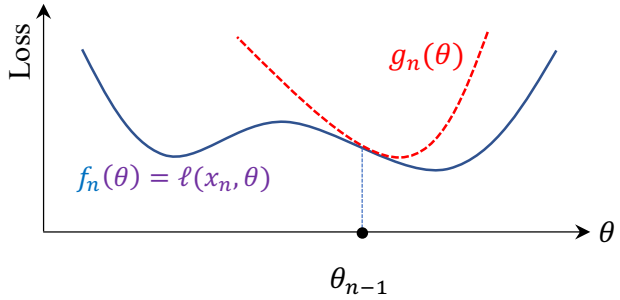


So how do we solve empirical loss minimization?

- 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 $\boldsymbol{\theta} \mapsto \ell(\mathbf{x}_n, \boldsymbol{\theta})$:



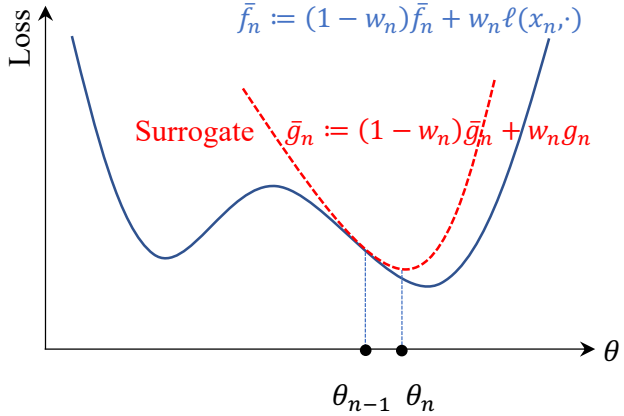
  - Ex: Gradient descent — Assuming ∇f_n is L-Lipschitz,

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

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)

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

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):

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

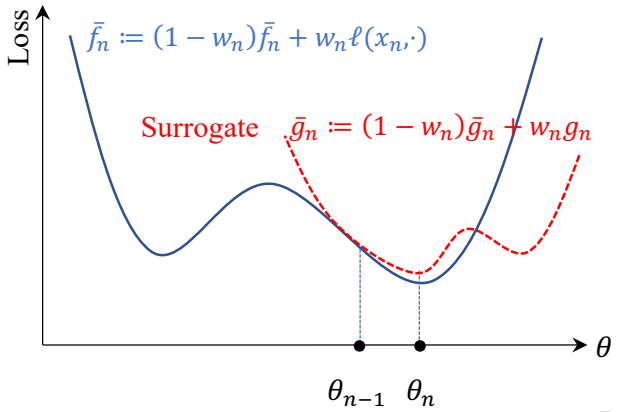
Stochastic (Block) Majorization-Minimization

- 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 [2], '20 [1])

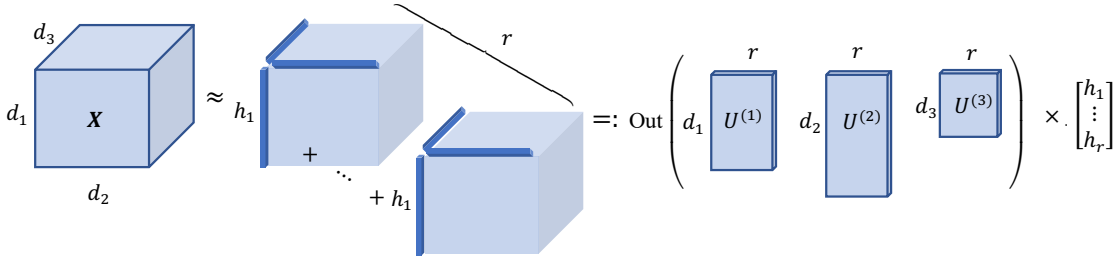
$$\textbf{(SBMM)} \quad \begin{cases} \text{Sample } \mathbf{x}_n \sim \pi(\cdot|\mathbf{x}_1,\dots,\mathbf{x}_{n-1}) \ ; \\ g_n \leftarrow \text{Block multi-convex majorizing surrogate of } f_n(\cdot) = \ell(\mathbf{x}_n,\cdot); \\ \boldsymbol{\theta}_n \approx \operatorname{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}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



Stochastic (Block) Majorization-Minimization

- Online CP-dictionary Learning (Lyu, Strohmeier, Needell '20 [4]):

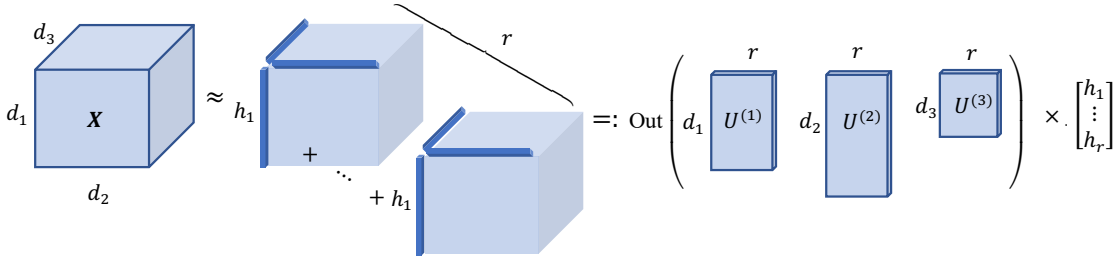
(CP-recons. error) $\ell(\underbrace{\mathbf{X}}_{m\text{-tensor}}, \mathbf{U} = \underbrace{[U^{(1)}, \dots, U^{(m)}]}_{\text{factor matrices}}, H) := \|\mathbf{X} - \underbrace{\mathrm{Out}(\mathbf{U})}_{\text{CP-dict.}} \times_{m+1} H\|_F^2$



Stochastic (Block) Majorization-Minimization

- Online CP-dictionary Learning (Lyu, Strohmeier, Needell '20 [4]):

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

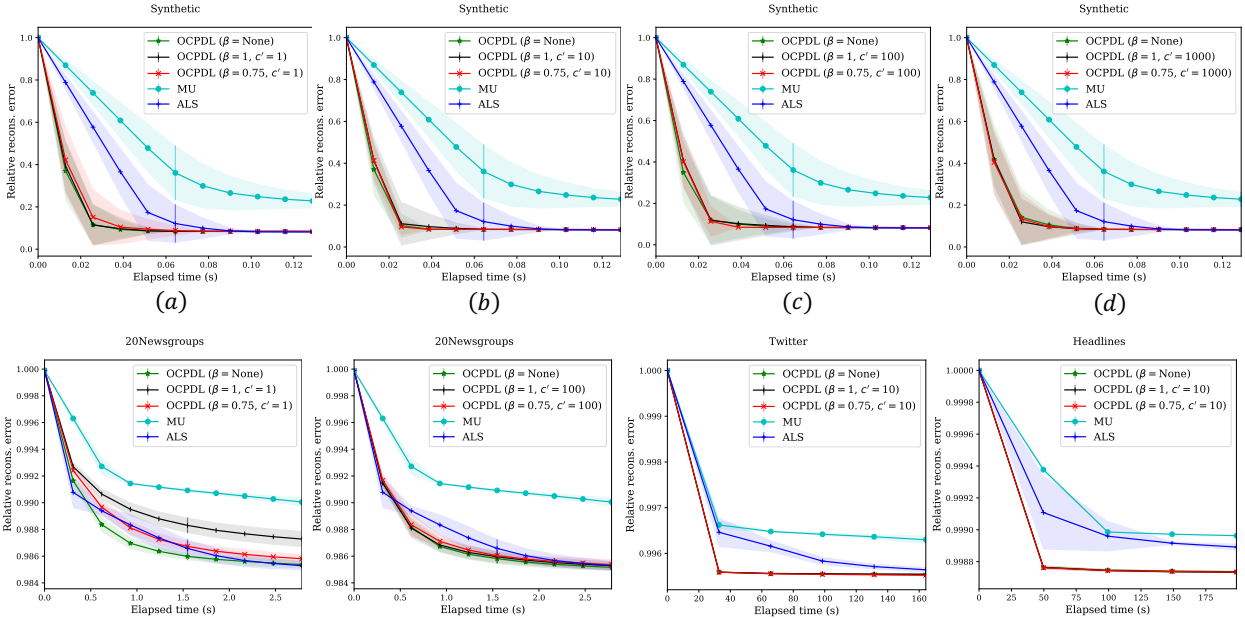


- Upon arrival of $\mathbf{X}_n \in \mathbb{R}^{d_1 \times \cdots \times d_m}$:

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

Stochastic (Block) Majorization-Minimization

- Online CP-dictionary Learning (Lyu, Strohmeier, Needell '20 [4]):
  - Only bounded memory to learn from infinitely many samples
  - Cheaper per-iteration cost than offline methods
  - Converges faster than offline methods (empirically)



Outline

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 $\boldsymbol{\theta}_n \in \text{interior}(\boldsymbol{\Theta})$ *for* $n \geq 1$ *and* $w_n = n^{-1/2}(\log n)^{1+\varepsilon}$,

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

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

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 $\boldsymbol{\theta}_n \in \text{interior}(\boldsymbol{\Theta})$ *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 $O(1/n^{1/4})$.

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,*

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

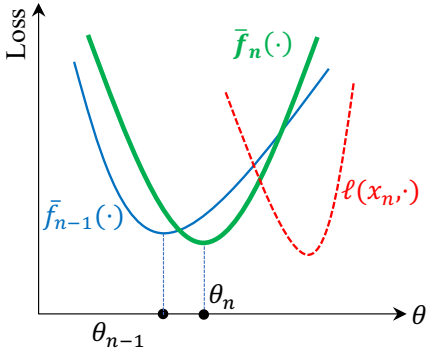
$$\min_{1\leq k\leq n}\left[-\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left\langle \nabla \bar{f}_k(\boldsymbol{\theta}_k), \frac{(\boldsymbol{\theta}-\boldsymbol{\theta}_k)}{\|\boldsymbol{\theta}-\boldsymbol{\theta}_k\|}\right\rangle\right]^2 = O\Big(\underbrace{\left(\sum_{k=1}^n w_k\right)^{-1}}_{\text{best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

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

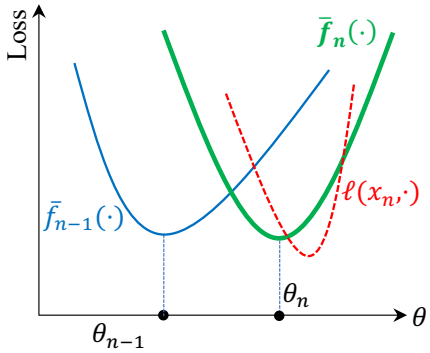
$$\min_{1\leq k\leq n}\Bigg[\underbrace{-\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left\langle \nabla f(\boldsymbol{\theta}_k), \frac{(\boldsymbol{\theta}-\boldsymbol{\theta}_k)}{\|\boldsymbol{\theta}-\boldsymbol{\theta}_k\|}\right\rangle}_{\text{optimality gap for constrained case}}\Bigg]^2 =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

Open questions

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



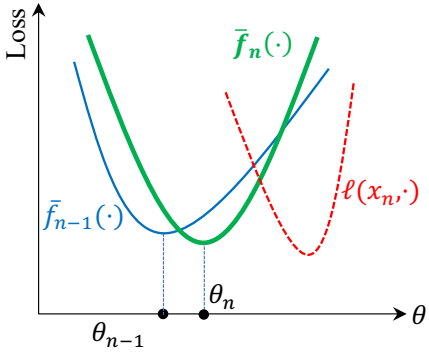
Slow adaptation



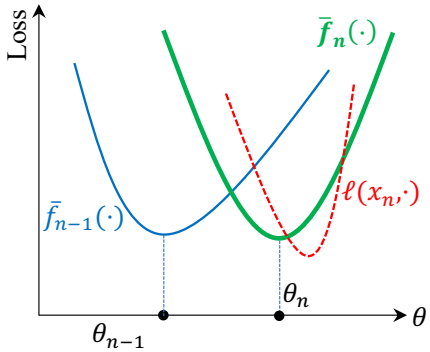
Fast adaptation

Open questions

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



Slow adaptation

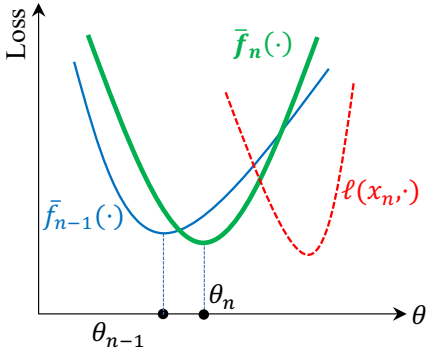


Fast adaptation

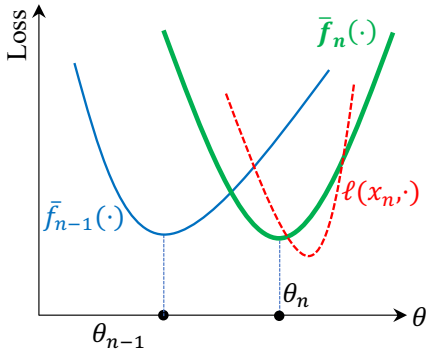
  - All theoretical analysis assumes slow adaptation regime $\frac{1}{n} \le w_n \ll \frac{1}{\sqrt{n}}$:

Open questions

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



Slow adaptation

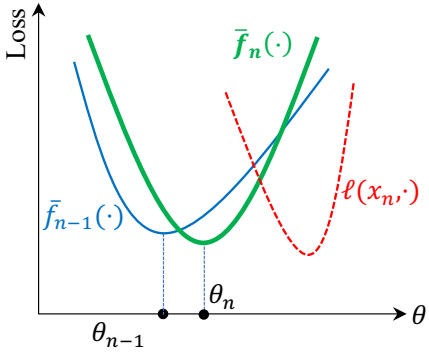


Fast adaptation

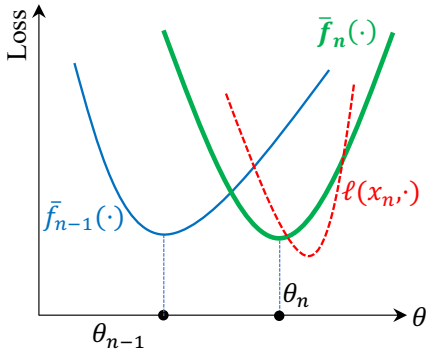
  - All theoretical analysis assumes slow adaptation regime $\frac{1}{n} \le w_n \ll \frac{1}{\sqrt{n}}$:
  - It allows to use weighted versions of SLLN and CLT: $\bar{f}_n \approx f$, $\nabla \bar{f}_n \approx \nabla f$

Open questions

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



Slow adaptation

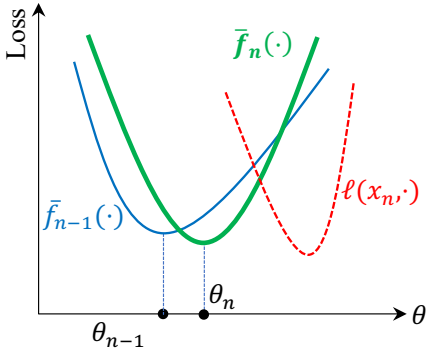


Fast adaptation

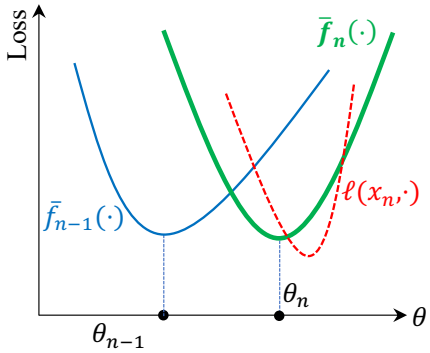
- All theoretical analysis assumes slow adaptation regime $\frac{1}{n} \le w_n \ll \frac{1}{\sqrt{n}}$:
- It allows to use weighted versions of SLLN and CLT: $\bar{f}_n \approx f$, $\nabla \bar{f}_n \approx \nabla f$
- Formulate the goal of learning non-stationary (short-time scale) features

Open questions

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



Slow adaptation



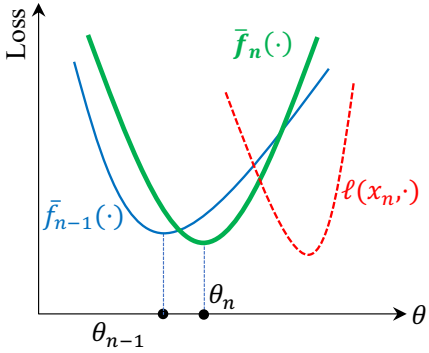
Fast adaptation

  - 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

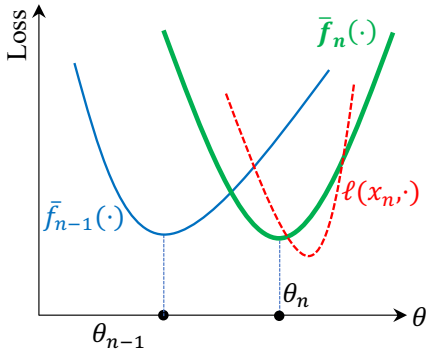
- Finding global minimizer for some online nonconvex problems?

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 $\sum_{n=1}^{\infty} r_n^2 < \infty$,

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

Proposition (Asymptotic first-order optimality)

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

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

Proposition (Finite first-order variation)

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

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

Proposition (Asymptotic first-order optimality)

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

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

- 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 $(\boldsymbol{\theta}_n)_{n\ge1}$ 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 $\boldsymbol{\Theta}$:

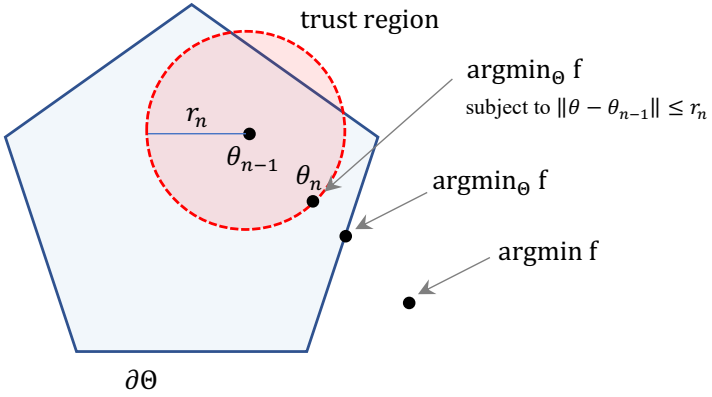
$$\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \langle \nabla f(\boldsymbol{\theta}_\infty), \boldsymbol{\theta} - \boldsymbol{\theta}_\infty \rangle \geq 0$$

- 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 $\boldsymbol{\Theta}$:

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

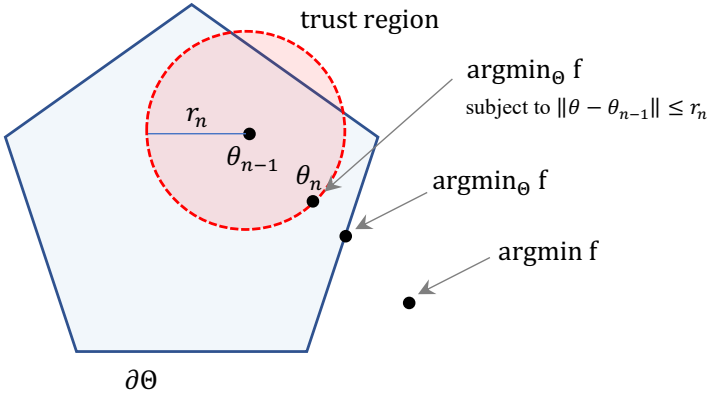
- Main difficulty: Show that the DR (also the PR) modification of BCD does not affect the asymptotic property of iterates
  - For BCD-DR: What if $\boldsymbol{\theta}_n$ touches the trust region boundary $\|\boldsymbol{\theta} - \boldsymbol{\theta}_n\| \leq r_n$ infintely often?



- 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?



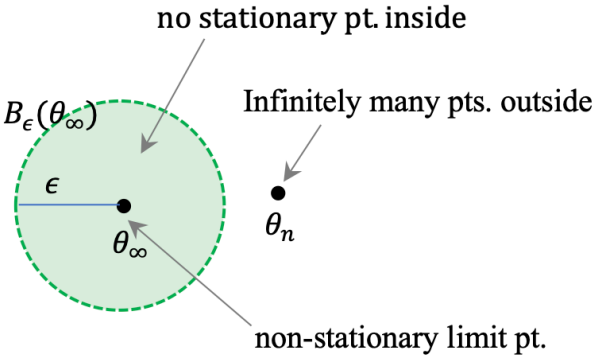
  - For BCD-PR: What if the PR term tilts the true gradient asymptotically?

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* $B_\varepsilon(\boldsymbol{\theta}_\infty)$.

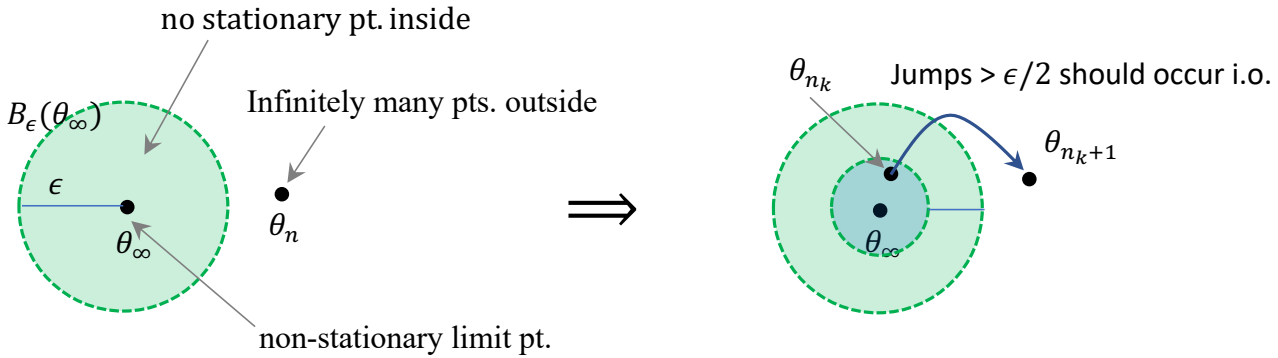


Proposition (Local structure of a non-stationary limit point)

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* $B_\varepsilon(\boldsymbol{\theta}_\infty)$.

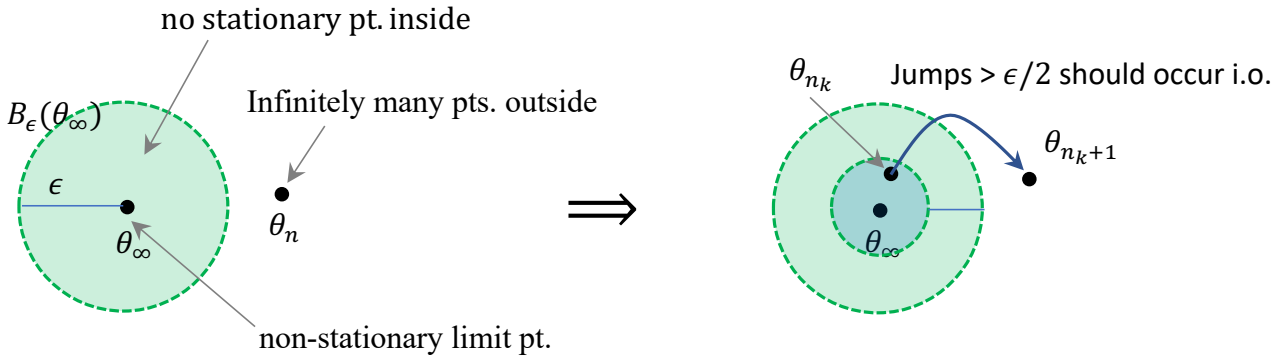


Proposition (Local structure of a non-stationary limit point)

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* $\boldsymbol{\Theta}$

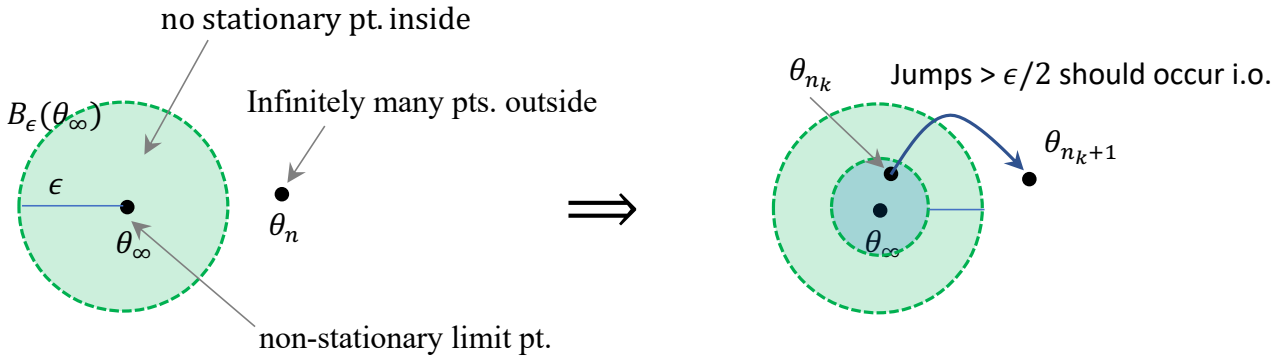
(b) *There exists infinitely many* $\boldsymbol{\theta}_n$*'s outside of* $B_\varepsilon(\boldsymbol{\theta}_\infty)$.



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

Proposition (Sufficient condition for stationarity II)

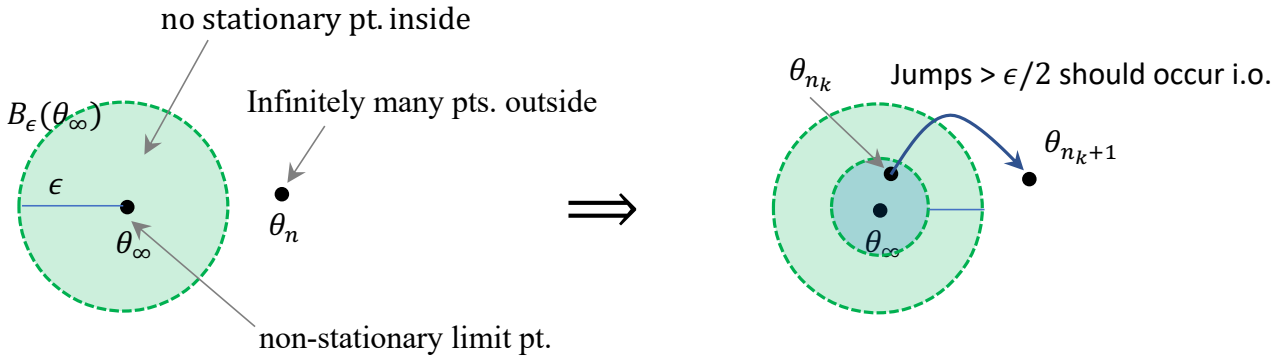
Suppose there exists a subsequence $(\boldsymbol{\theta}_{n_k})_{k\geq 1}$ such that $\sum_{k=1}^{\infty}\|\boldsymbol{\theta}_{n_k}-\boldsymbol{\theta}_{n_k+1}\|=\infty$. There exists a further subsequence $(s_k)_{k\geq 1}$ of $(n_k)_{k\geq 1}$ such that $\boldsymbol{\theta}_\infty := \lim_{k\to\infty}\boldsymbol{\theta}_{s_k}$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



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

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 $B_\varepsilon(\boldsymbol{\theta}_\infty)$, $\Rightarrow\Leftarrow$.

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

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

[3] 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.

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

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