Online Nonnegative Tensor Factorization and CP Decomposition

Hanbaek Lyu

University of California, Los Angeles

Partially supported by NSF DMS #2010035

May. 21 2021

Outline

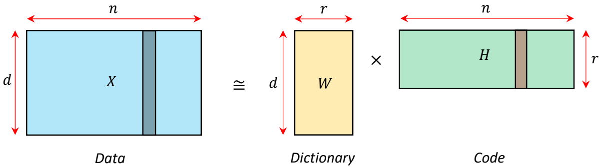
Offline Matrix and Tensor Factorization

Online Matrix and Tensor Factorization

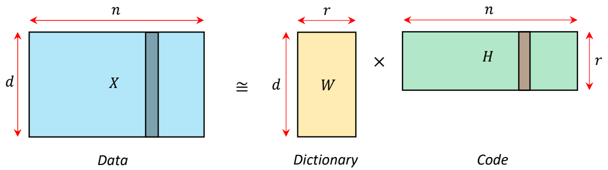
Algorithms for Online Matrix/Tensor Factorization

Matrix Factorization

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



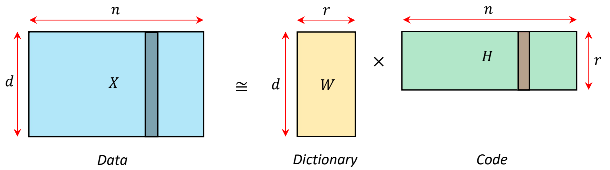
- Matrix factorization is a fundamental tool in dictionary learning problems.



- Formulated as a non-convex optimization problem:

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

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

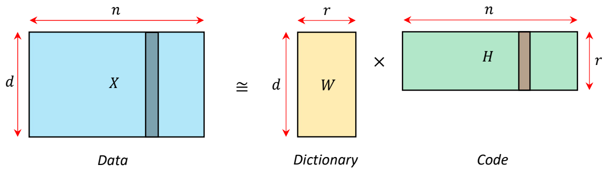


- Formulated as a non-convex optimization problem:

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

  - Nonnegative Matrix Factorization (NMF): $\mathcal{C} = \mathbb{R}_{\geq 0}^{d \times r}$, $\mathcal{C}' = \mathbb{R}_{\geq 0}^{r \times n}$, $\lambda = 0$,

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



- Formulated as a non-convex optimization problem:

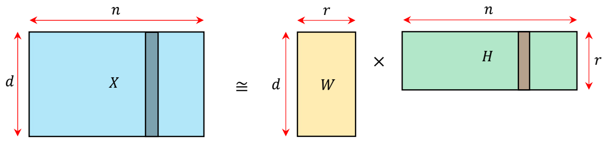
$$\begin{cases} \text{minimize} & \|X - WH\|_F^2 + \lambda \|H\|_1 \qquad \text{(Reconstruction error)} \\ \text{subject to} & W \in \mathcal{C},\ H \in \mathcal{C}' \qquad (\textit{Constraints}) \end{cases}$$

  - Nonnegative Matrix Factorization (NMF): $\mathcal{C} = \mathbb{R}_{\geq 0}^{d\times r}$, $\mathcal{C}' = \mathbb{R}_{\geq 0}^{r\times n}$, $\lambda = 0$,

Subspace clustering, Matrix Completion, Sparse PCA, Robust PCA, Poisson PCA, Heteroscedastic PCA,

Bilinear Inverse Problems, Max-Plus Factorization ...

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



- Formulated as a non-convex optimization problem:

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

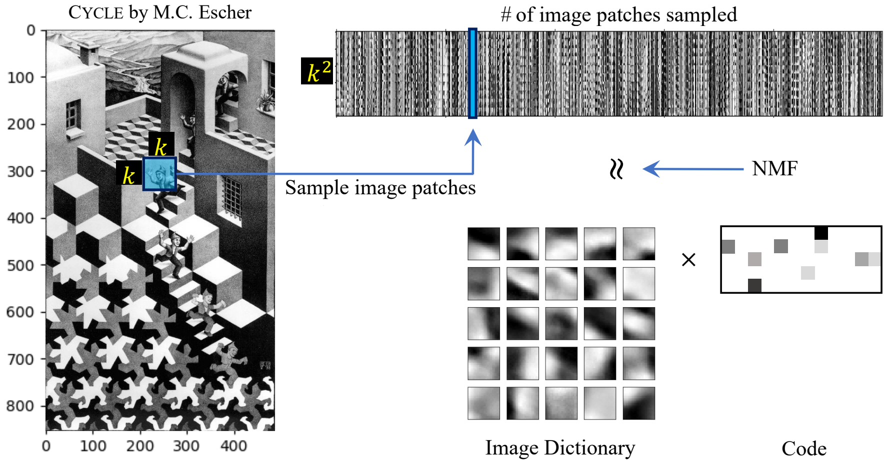
  - Nonnegative Matrix Factorization (NMF): $\mathcal{C} = \mathbb{R}_{\geq 0}^{d \times r}$, $\mathcal{C}' = \mathbb{R}_{\geq 0}^{r \times n}$, $\lambda = 0$,

    Subspace clustering, Matrix Completion, Sparse PCA, Robust PCA, Poisson PCA, Heteroscedastic PCA,

    Bilinear Inverse Problems, Max-Plus Factorization ...

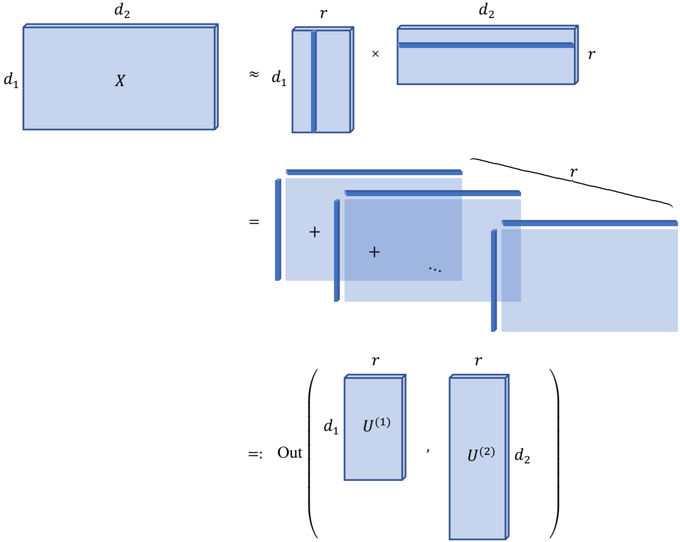
- Applications in text analysis, image reconstruction, medical imaging, bioinformatics, etc.

Example of NMF for Image dictionary learning

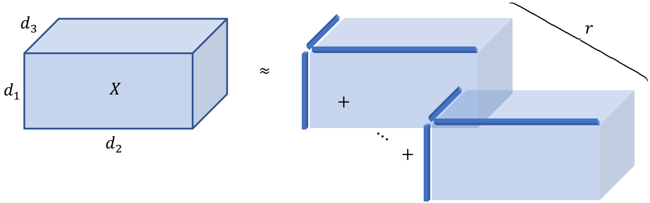


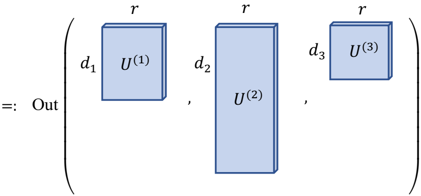
An alternative view of Matrix Factorization

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



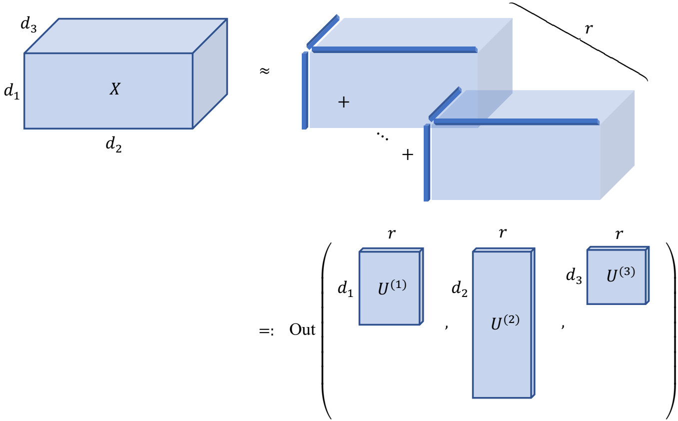
- $X \approx \mathrm{Out}(U^{(1)}, U^{(2)}, U^{(3)})$





Tensor Factorization (CP decomposition)

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



- Nonnegative Tensor Factorization

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

- Nonnegative Tensor Factorization (NTF)

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

- Nonnegative Tensor Factorization (NTF)

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

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

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

- Nonnegative Tensor Factorization (NTF)

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

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

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

  - BCD for NMF is guaranteed to converge to a stationary point of the loss function (Grippo, Sciandrone '00 [1])

- Nonnegative Tensor Factorization (NTF)

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

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

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

  - BCD for NMF is guaranteed to converge to a stationary point of the loss function (Grippo, Sciandrone '00 [1])
  - BCD for NTF is not guaranteed to converge to a stationary point of the loss function (Powell '73 [6])

- Nonnegative Tensor Factorization (NTF)

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

- Block Coordinate Descent (BCD) with Diminishing Radius (Lyu '20 [2]) for NTF

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

- Nonnegative Tensor Factorization (NTF)

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

- Block Coordinate Descent (BCD) with Diminishing Radius (Lyu '20 [2]) for NTF

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

  - (Diminishing Radius) $\sum_{t=1}^{\infty} w_t = \infty$, $\sum_{t=1}^{\infty} w_t^2 < \infty$

Block Coordinate Descent for Tensor Factorization

- Nonnegative Tensor Factorization (NTF)

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

- Block Coordinate Descent (BCD) with Diminishing Radius (Lyu '20 [2]) for NTF

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

  - (Diminishing Radius) $\sum_{t=1}^{\infty} w_t = \infty$, $\sum_{t=1}^{\infty} w_t^2 < \infty$
  - (Convergence Guarantee) This algorithm is guaranteed to converge to a stationary point of the loss function

- Nonnegative Tensor Factorization (NTF)

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

- Block Coordinate Descent (BCD) with Diminishing Radius (Lyu '20 [2]) for NTF

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

  - (Diminishing Radius) $\sum_{t=1}^{\infty} w_t = \infty$, $\sum_{t=1}^{\infty} w_t^2 < \infty$
  - (Convergence Guarantee) This algorithm is guaranteed to converge to a stationary point of the loss function
  - Rate of convergence: $O((\log n)^2/n)$

Outline

Offline Matrix and Tensor Factorization

Online Matrix and Tensor Factorization

Algorithms for Online Matrix/Tensor Factorization

- Online (Nonnegative) Matrix Factorization:

$$\operatorname*{argmin}_{W\in\mathbb{R}^{d\times n}_{\geq 0}} \left(\ell(W) := \mathbb{E}_{X\sim\pi}\left[\inf_{H\in\mathbb{R}^{r\times d}_{\geq 0}} \|X - WH\|_F^2 \right] \right)$$

- Online (Nonnegative) Matrix Factorization:

$$\operatorname*{argmin}_{W\in\mathbb{R}^{d\times n}_{\geq 0}} \left(\ell(W) := \mathbb{E}_{X\sim\pi}\left[\inf_{H\in\mathbb{R}^{r\times d}_{\geq 0}} \|X - WH\|_F^2 \right] \right)$$

- Online (Nonnegative) Tensor Factorization:

$$\operatorname*{argmin}_{U^{(1)}\in\mathbb{R}^{d_1\times r}_{\geq 0},\, U^{(2)}\in\mathbb{R}^{d_2\times r}_{\geq 0}} \left(\ell(U^{(1)}, U^{(2)}) := \mathbb{E}_{X\sim\pi}\left[\inf_{U^{(3)}\in\mathbb{R}^{d_3\times r}_{\geq 0}} \|X - \mathrm{Out}(U^{(1)},\, U^{(2)},\, U^{(3)})\|_F^2 \right] \right)$$

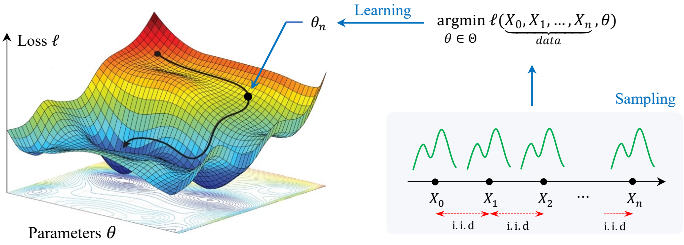
- Online (Nonnegative) Matrix Factorization:

$$\operatorname*{argmin}_{W\in\mathbb{R}^{d\times n}_{\geq 0}} \left(\ell(W) := \mathbb{E}_{X\sim\pi}\left[\inf_{H\in\mathbb{R}^{r\times d}_{\geq 0}} \|X - WH\|_F^2 \right] \right)$$

- Online (Nonnegative) Tensor Factorization:

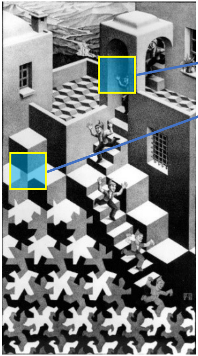
$$\operatorname*{argmin}_{U^{(1)}\in\mathbb{R}^{d_1\times r}_{\geq 0},\, U^{(2)}\in\mathbb{R}^{d_2\times r}_{\geq 0}} \left(\ell(U^{(1)}, U^{(2)}) := \mathbb{E}_{X\sim\pi}\left[\inf_{U^{(3)}\in\mathbb{R}^{d_3\times r}_{\geq 0}} \|X - \mathrm{Out}(U^{(1)},\, U^{(2)},\, U^{(3)})\|_F^2 \right] \right)$$

- Why "online"?



Example of Online NMF + i.i.d. sampling: Image dictionary learning

CYCLE by M.C. Escher



i.i.d. sampling

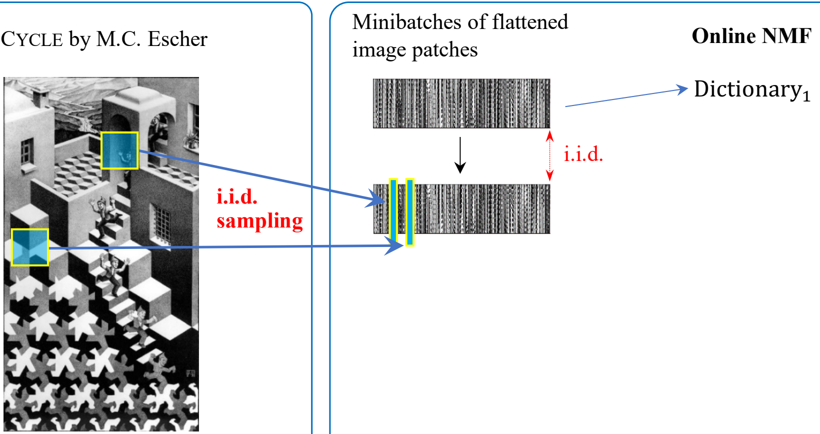
Minibatches of flattened image patches

Online NMF

Dictionary$_1$

Example of Online NMF + i.i.d. sampling: Image dictionary learning

CYCLE by M.C. Escher

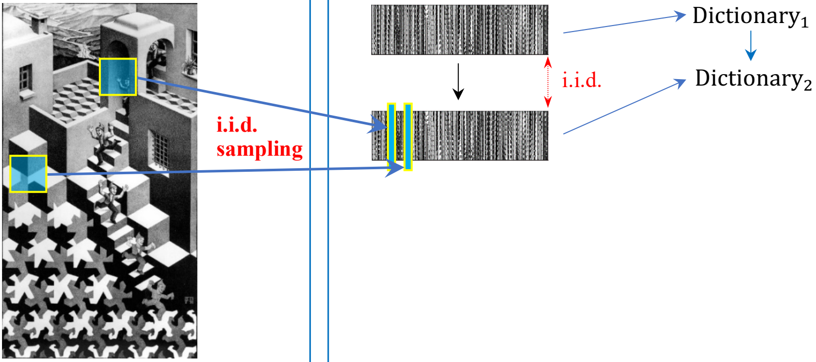


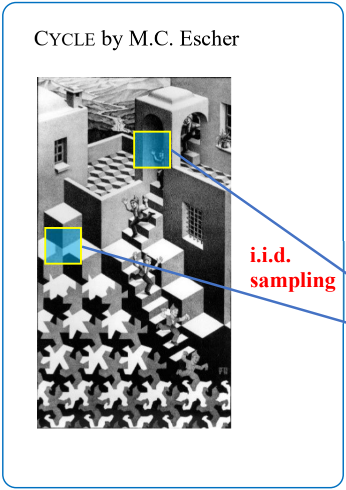
Example of Online NMF + i.i.d. sampling: Image dictionary learning

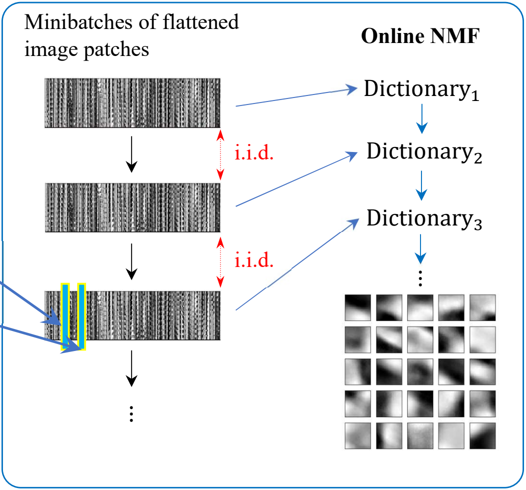
CYCLE by M.C. Escher

Minibatches of flattened image patches

Online NMF




CYCLE by M.C. Escher
i.i.d.
sampling


Minibatches of flattened image patches
Online NMF
i.i.d.
i.i.d.
Dictionary_1
Dictionary_2
Dictionary_3

- Going from matrix factorization to tensor factorization
  $\longrightarrow$ Learn also from the time dimension

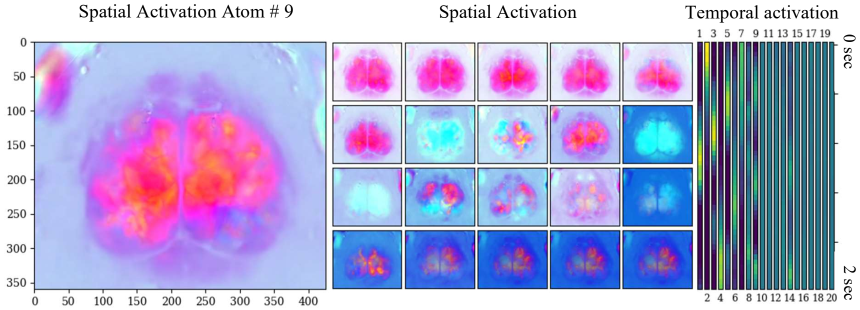


Figure: Temporal dictionary learned from mice brain activity video (Original data from Barson et al. *Nature methods* (2020))

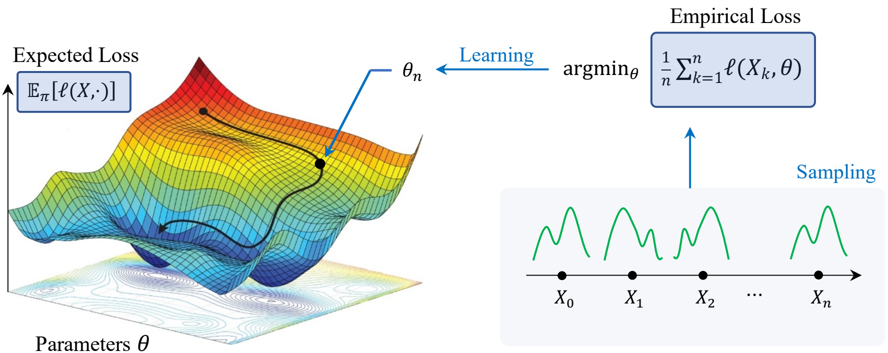
Outline

Offline Matrix and Tensor Factorization

Online Matrix and Tensor Factorization

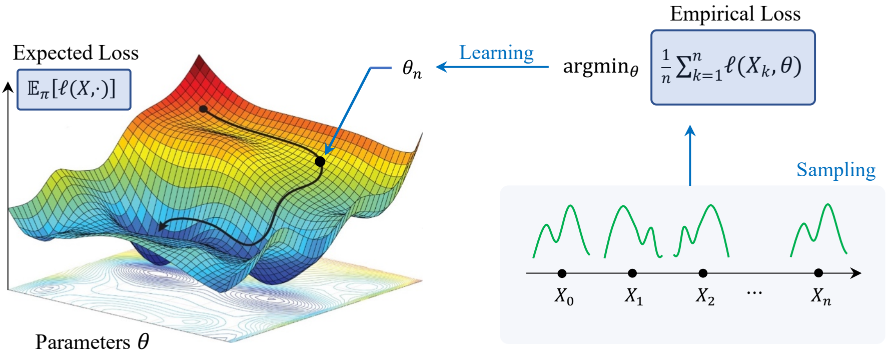
Algorithms for Online Matrix/Tensor Factorization

- Goal: Minimize the expected loss $\mathbb{E}_{X\sim\pi}[\ell(X,\theta)]$ given a loss function ℓ



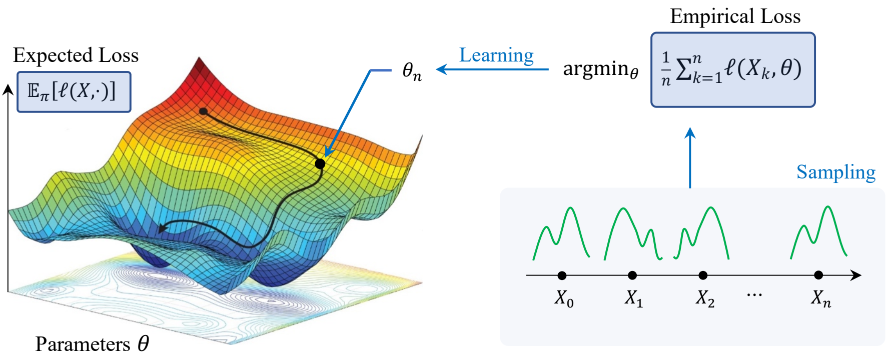
Empirical Loss Minimization

- Goal: Minimize the expected loss $\mathbb{E}_{X\sim\pi}[\ell(X,\theta)]$ given a loss function ℓ
- Polular approach: *Empirical Loss Minimization*

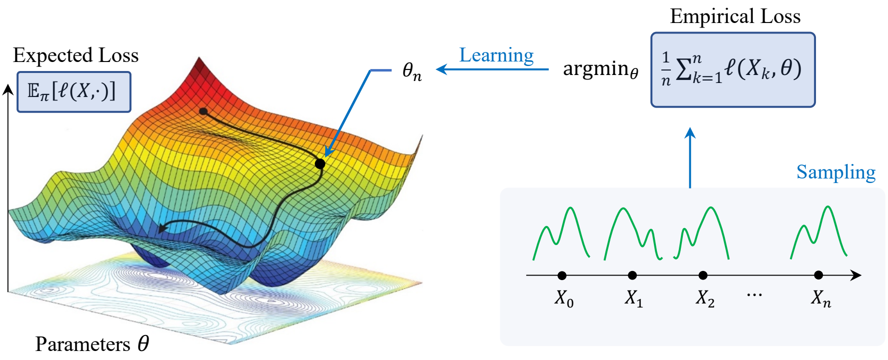


Empirical Loss Minimization

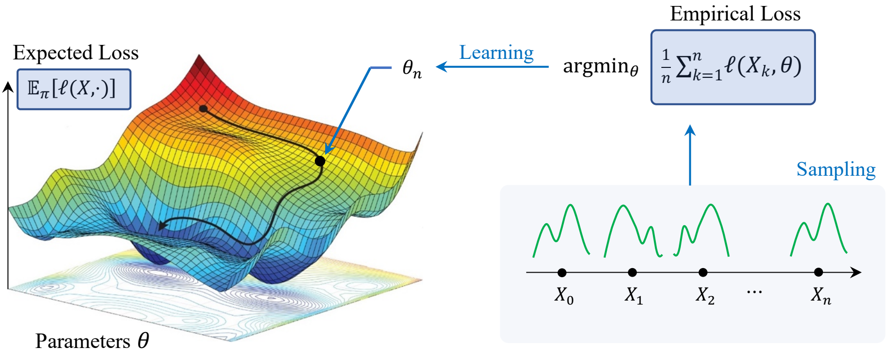
- Goal: Minimize the expected loss $\mathbb{E}_{X\sim\pi}[\ell(X,\theta)]$ given a loss function ℓ
- Polular approach: *Empirical Loss Minimization*
  - Background: $\lim_{n\to\infty}$ Empirical Loss = Expected loss



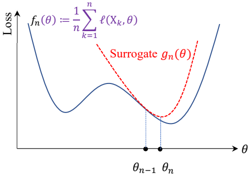
- Goal: Minimize the expected loss $\mathbb{E}_{X\sim\pi}[\ell(X,\theta)]$ given a loss function ℓ
- Polular approach: *Empirical Loss Minimization*
  - Background: $\lim_{n\to\infty}$ Empirical Loss = Expected loss
  - Not practical in many cases:



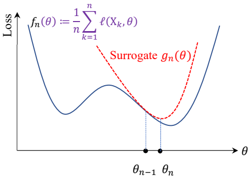
- Goal: Minimize the expected loss $\mathbb{E}_{X\sim\pi}[\ell(X,\theta)]$ given a loss function ℓ
- Polular approach: *Empirical Loss Minimization*
  - Background: $\lim_{n\to\infty}$ Empirical Loss = Expected loss
  - Not practical in many cases:
    - The *empirical loss is often hard to minimize* (e.g., Matrix/Tensor Factorization)



- Stochastic Majorization-Minimization (SMM) – Mairal [4]
  - Iteratively minimize majorizing surrogates g_n of the empirical loss f_n:



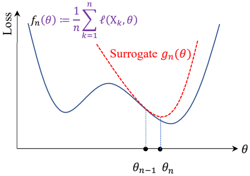
- Stochastic Majorization-Minimization (SMM) – Mairal [4]
  - Iteratively minimize majorizing surrogates g_n of the empirical loss f_n:



- Online NMF algorithm (Mairal, Bach, Saprio, Ponce [5]):

$$
\begin{cases}
H_t & \leftarrow \operatorname*{argmin}_{H \in \mathbb{R}^{r\times n}_{\geq 0}} \left[\|X_t - W_{t-1}H\|_F^2 \right] \qquad \textit{(Convex)} \\
\hat{f}_t(W) & \leftarrow (1-w_t)\hat{f}_{t-1}(W) + w_t \left(\|X_t - WH_t\|_F^2 \right) \qquad \text{(Surrogate empirical loss)} \\
W_t & \leftarrow \operatorname*{argmin}_{W \in \mathbb{R}^{d\times r}_{\geq 0}} \hat{f}_t(W) \qquad \textit{(Convex)}
\end{cases}
$$

- Stochastic Majorization-Minimization (SMM) – Mairal [4]
  - Iteratively minimize majorizing surrogates g_n of the empirical loss f_n:



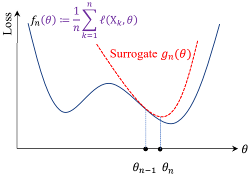
- Online NMF algorithm (Mairal, Bach, Saprio, Ponce [5]):

$$
\begin{cases}
H_t & \leftarrow \operatorname*{argmin}_{H \in \mathbb{R}^{r \times n}_{\geq 0}} \left[\|X_t - W_{t-1} H\|_F^2 \right] \qquad (\textit{Convex}) \\
\hat{f}_t(W) & \leftarrow (1 - w_t)\hat{f}_{t-1}(W) + w_t \left(\|X_t - W H_t\|_F^2 \right) \qquad (\text{Surrogate empirical loss}) \\
W_t & \leftarrow \operatorname*{argmin}_{W \in \mathbb{R}^{d \times r}_{\geq 0}} \hat{f}_t(W) \qquad (\textit{Convex})
\end{cases}
$$

- Algorithm converges to a stationary point almost surely

Stochastic Majorization-Minimization

- Stochastic Majorization-Minimization (SMM) – Mairal [4]
  - Iteratively minimize majorizing surrogates g_n of the empirical loss f_n:



- Online NMF algorithm (Mairal, Bach, Saprio, Ponce [5]):

$$\begin{cases} H_t & \leftarrow \operatorname*{argmin}_{H \in \mathbb{R}^{r\times n}_{\geq 0}} \left[\|X_t - W_{t-1}H\|_F^2 \right] \qquad \text{(Convex)} \\ \hat{f}_t(W) & \leftarrow (1-w_t)\hat{f}_{t-1}(W) + w_t \left(\|X_t - WH_t\|_F^2 \right) \qquad \text{(Surrogate empirical loss)} \\ W_t & \leftarrow \operatorname*{argmin}_{W \in \mathbb{R}^{d\times r}_{\geq 0}} \hat{f}_t(W) \qquad \text{(Convex)} \end{cases}$$

- Algorithm converges to a stationary point almost surely
- Convergence also holds when data sequence has Markovian dependence (Lyu, Needell, Balzano [3])

- Online NTF algorithm (?):

$$\begin{cases} U_t^{(3)} & \leftarrow \underset{U^{(3)} \in \mathbb{R}_{\geq 0}^{d_3 \times r}}{\operatorname{argmin}} \left[\left\| X_t - \mathtt{Out}(U_{t-1}^{(1)},\, U_{t-1}^{(2)},\, U^{(3)}) \right\|_F^2 \right] \qquad \color{red}{(\textit{Convex})} \\ \hat{f}_t(U^{(1)}, U^{(2)}) & \leftarrow (1 - w_t)\hat{f}_{t-1}(U^{(1)}, U^{(2)}) \\ & \quad + w_t \left(\left\| X_t - \mathtt{Out}(U^{(1)},\, U^{(2)},\, U_t^{(3)}) \right\|_F^2 \right) \qquad \color{red}{(\text{Surrogate empirical loss})} \\ (U_t^{(1)}, U_t^{(2)}) & \leftarrow \underset{U^{(1)} \in \mathbb{R}_{\geq 0}^{d_1 \times r},\, U^{(2)} \in \mathbb{R}_{\geq 0}^{d_2 \times r}}{\operatorname{argmin}} \hat{f}_t(U^{(1)}, U^{(2)}) \qquad \color{red}{\boldsymbol{(\textit{Non-convex})}} \end{cases}$$

- Online NTF algorithm (?):

$$\begin{cases} U_t^{(3)} & \leftarrow \underset{U^{(3)} \in \mathbb{R}_{\geq 0}^{d_3 \times r}}{\operatorname{argmin}} \left[\left\| X_t - \texttt{Out}(U_{t-1}^{(1)},\, U_{t-1}^{(2)},\, U^{(3)}) \right\|_F^2 \right] & \textit{(Convex)} \\ \hat{f}_t(U^{(1)}, U^{(2)}) & \leftarrow (1 - w_t)\hat{f}_{t-1}(U^{(1)}, U^{(2)}) & \\ & \quad + w_t \left(\left\| X_t - \texttt{Out}(U^{(1)},\, U^{(2)},\, U_t^{(3)}) \right\|_F^2 \right) & \text{(Surrogate empirical loss)} \\ (U_t^{(1)}, U_t^{(2)}) & \leftarrow \underset{U^{(1)} \in \mathbb{R}_{\geq 0}^{d_1 \times r},\, U^{(2)} \in \mathbb{R}_{\geq 0}^{d_2 \times r}}{\operatorname{argmin}} \hat{f}_t(U^{(1)}, U^{(2)}) & \textbf{\textit{(Non-convex)}} \end{cases}$$

- Surrogate empirical loss $\hat{f}_t$ is non-convex, so cannot directly find minimizing $(U_t^{(1)}, U_t^{(2)})$.

- Online NTF algorithm (??):

$$
\begin{cases}
U_t^{(3)} & \leftarrow \underset{U^{(3)} \in \mathbb{R}_{\geq 0}^{d_3 \times r}}{\operatorname{argmin}} \left[\left\| X_t - \mathtt{Out}(U_{t-1}^{(1)},\, U_{t-1}^{(2)},\, U^{(3)}) \right\|_F^2 \right] \qquad \text{(Convex)} \\
\hat{f}_t(U^{(1)}, U^{(2)}) & \leftarrow (1 - w_t)\hat{f}_{t-1}(U^{(1)}, U^{(2)}) \\
& \quad + w_t \left(\left\| X_t - \mathtt{Out}(U^{(1)},\, U^{(2)},\, U_t^{(3)}) \right\|_F^2 \right) \qquad \text{(Surrogate empirical loss)} \\
U_t^{(1)} & \leftarrow \underset{U^{(1)} \in \mathbb{R}_{\geq 0}^{d_1 \times r}}{\operatorname{argmin}} \hat{f}_t(U^{(1)}, U_{t-1}^{(2)}) \qquad \textbf{(Convex)} \\
U_t^{(2)} & \leftarrow \underset{U^{(2)} \in \mathbb{R}_{\geq 0}^{d_2 \times r}}{\operatorname{argmin}} \hat{f}_t(U_t^{(1)}, U^{(2)}) \qquad \textbf{(Convex)}
\end{cases}
$$

SMM+ BCD ?

- Online NTF algorithm (??):

$$
\begin{cases}
U_t^{(3)} & \leftarrow \underset{U^{(3)} \in \mathbb{R}_{\geq 0}^{d_3 \times r}}{\operatorname{argmin}} \left[\left\| X_t - \mathtt{Out}(U_{t-1}^{(1)},\, U_{t-1}^{(2)},\, U^{(3)}) \right\|_F^2 \right] \qquad \textit{(Convex)} \\
\hat{f}_t(U^{(1)}, U^{(2)}) & \leftarrow (1 - w_t)\hat{f}_{t-1}(U^{(1)}, U^{(2)}) \\
& \quad + w_t \left(\left\| X_t - \mathtt{Out}(U^{(1)},\, U^{(2)},\, U_t^{(3)}) \right\|_F^2 \right) \qquad \text{(Surrogate empirical loss)} \\
U_t^{(1)} & \leftarrow \underset{U^{(1)} \in \mathbb{R}_{\geq 0}^{d_1 \times r}}{\operatorname{argmin}} \hat{f}_t(U^{(1)}, U_{t-1}^{(2)}) \qquad \boldsymbol{\textit{(Convex)}} \\
U_t^{(2)} & \leftarrow \underset{U^{(2)} \in \mathbb{R}_{\geq 0}^{d_2 \times r}}{\operatorname{argmin}} \hat{f}_t(U_t^{(1)}, U^{(2)}) \qquad \boldsymbol{\textit{(Convex)}}
\end{cases}
$$

  - $(U_t^{(1)}, U_t^{(2)})$ does not minimize $\hat{f}_t$

- Online NTF algorithm (??):

$$\begin{cases} U_t^{(3)} & \leftarrow \underset{U^{(3)} \in \mathbb{R}_{\geq 0}^{d_3 \times r}}{\operatorname{argmin}} \left[\left\| X_t - \texttt{Out}(U_{t-1}^{(1)},\, U_{t-1}^{(2)},\, U^{(3)}) \right\|_F^2 \right] \qquad \textcolor{red}{(\textit{Convex})} \\ \hat{f}_t(U^{(1)}, U^{(2)}) & \leftarrow (1-w_t)\hat{f}_{t-1}(U^{(1)}, U^{(2)}) \\ & \quad + w_t \left(\left\| X_t - \texttt{Out}(U^{(1)},\, U^{(2)},\, U_t^{(3)}) \right\|_F^2 \right) \qquad \textcolor{red}{(\text{Surrogate empirical loss})} \\ U_t^{(1)} & \leftarrow \underset{U^{(1)} \in \mathbb{R}_{\geq 0}^{d_1 \times r}}{\operatorname{argmin}} \hat{f}_t(U^{(1)}, U_{t-1}^{(2)}) \qquad \textcolor{red}{(\boldsymbol{Convex})} \\ U_t^{(2)} & \leftarrow \underset{U^{(2)} \in \mathbb{R}_{\geq 0}^{d_2 \times r}}{\operatorname{argmin}} \hat{f}_t(U_t^{(1)}, U^{(2)}) \qquad \textcolor{red}{(\boldsymbol{Convex})} \end{cases}$$

  - $(U_t^{(1)}, U_t^{(2)})$ does not minimize $\hat{f}_t$
  - No convergence guarantee to stationary points

- Online NTF algorithm (!) (Strohmeier, Lyu, Needell [7]):

$$
\begin{cases}
U_t^{(3)} & \leftarrow \underset{U^{(3)}\in\mathbb{R}_{\geq 0}^{d_3\times r}}{\operatorname{argmin}} \left[\left\| X_t - \mathtt{Out}(U_{t-1}^{(1)},\, U_{t-1}^{(2)},\, U^{(3)}) \right\|_F^2 \right] \qquad \textcolor{red}{(\textit{Convex})} \\
\hat{f}_t(U^{(1)}, U^{(2)}) & \leftarrow (1-w_t)\hat{f}_{t-1}(U^{(1)}, U^{(2)}) \\
& \quad + w_t \left(\left\| X_t - \mathtt{Out}(U^{(1)},\, U^{(2)},\, U_t^{(3)}) \right\|_F^2 \right) \qquad \textcolor{red}{(\text{Surrogate empirical loss})} \\
U_t^{(1)} & \leftarrow \underset{U^{(1)}\in\mathbb{R}_{\geq 0}^{d_1\times r},\ \textcolor{blue}{\|U^{(1)}-U_{t-1}^{(1)}\|_F\leq w_t}}{\operatorname{argmin}} \hat{f}_t(U^{(1)}, U_{t-1}^{(2)}) \qquad \textcolor{red}{\boldsymbol{(Convex)}} \\
U_t^{(2)} & \leftarrow \underset{U^{(2)}\in\mathbb{R}_{\geq 0}^{d_2\times r},\ \textcolor{blue}{\|U^{(2)}-U_{t-1}^{(2)}\|_F\leq w_t}}{\operatorname{argmin}} \hat{f}_t(U_t^{(1)}, U^{(2)}) \qquad \textcolor{red}{\boldsymbol{(Convex)}}
\end{cases}
$$

SMM+BCD with Diminishing Radius

- Online NTF algorithm (!) (Strohmeier, Lyu, Needell [7]):

$$
\begin{cases}
U_t^{(3)} & \leftarrow \operatorname*{argmin}_{U^{(3)} \in \mathbb{R}_{\geq 0}^{d_3 \times r}} \left[\left\| X_t - \texttt{Out}(U_{t-1}^{(1)},\, U_{t-1}^{(2)},\, U^{(3)}) \right\|_F^2 \right] \qquad \textcolor{red}{(\mathit{Convex})} \\
\hat{f}_t(U^{(1)}, U^{(2)}) & \leftarrow (1 - w_t)\hat{f}_{t-1}(U^{(1)}, U^{(2)}) \\
& \quad + w_t \left(\left\| X_t - \texttt{Out}(U^{(1)},\, U^{(2)},\, U_t^{(3)}) \right\|_F^2 \right) \qquad \textcolor{red}{(\text{Surrogate empirical loss})} \\
U_t^{(1)} & \leftarrow \operatorname*{argmin}_{U^{(1)} \in \mathbb{R}_{\geq 0}^{d_1 \times r},\, \textcolor{blue}{\|U^{(1)} - U_{t-1}^{(1)}\|_F \leq w_t}} \hat{f}_t(U^{(1)}, U_{t-1}^{(2)}) \qquad \textcolor{red}{\boldsymbol{(\mathit{Convex})}} \\
U_t^{(2)} & \leftarrow \operatorname*{argmin}_{U^{(2)} \in \mathbb{R}_{\geq 0}^{d_2 \times r},\, \textcolor{blue}{\|U^{(2)} - U_{t-1}^{(2)}\|_F \leq w_t}} \hat{f}_t(U_t^{(1)}, U^{(2)}) \qquad \textcolor{red}{\boldsymbol{(\mathit{Convex})}}
\end{cases}
$$

  - $(U_t^{(1)}, U_t^{(2)})$ still does not minimize $\hat{f}_t$

SMM+BCD with Diminishing Radius

- Online NTF algorithm (!) (Strohmeier, Lyu, Needell [7]):

$$
\begin{cases}
U_t^{(3)} & \leftarrow \underset{U^{(3)} \in \mathbb{R}_{\geq 0}^{d_3 \times r}}{\operatorname{argmin}} \left[\left\| X_t - \texttt{Out}(U_{t-1}^{(1)},\, U_{t-1}^{(2)},\, U^{(3)}) \right\|_F^2 \right] \qquad \color{red}{(\textit{Convex})} \\
\hat{f}_t(U^{(1)}, U^{(2)}) & \leftarrow (1 - w_t)\hat{f}_{t-1}(U^{(1)}, U^{(2)}) \\
 & \quad + w_t \left(\left\| X_t - \texttt{Out}(U^{(1)},\, U^{(2)},\, U_t^{(3)}) \right\|_F^2 \right) \qquad \color{red}{\text{(Surrogate empirical loss)}} \\
U_t^{(1)} & \leftarrow \underset{U^{(1)} \in \mathbb{R}_{\geq 0}^{d_1 \times r},\, \color{blue}{\| U^{(1)} - U_{t-1}^{(1)} \|_F \leq w_t}}{\operatorname{argmin}} \hat{f}_t(U^{(1)}, U_{t-1}^{(2)}) \qquad \color{red}{\boldsymbol{(\textit{Convex})}} \\
U_t^{(2)} & \leftarrow \underset{U^{(2)} \in \mathbb{R}_{\geq 0}^{d_2 \times r},\, \color{blue}{\| U^{(2)} - U_{t-1}^{(2)} \|_F \leq w_t}}{\operatorname{argmin}} \hat{f}_t(U_t^{(1)}, U^{(2)}) \qquad \color{red}{\boldsymbol{(\textit{Convex})}}
\end{cases}
$$

  - $(U_t^{(1)}, U_t^{(2)})$ still does not minimize $\hat{f}_t$
  - But we can get a.s. convergence to stationary points [7]

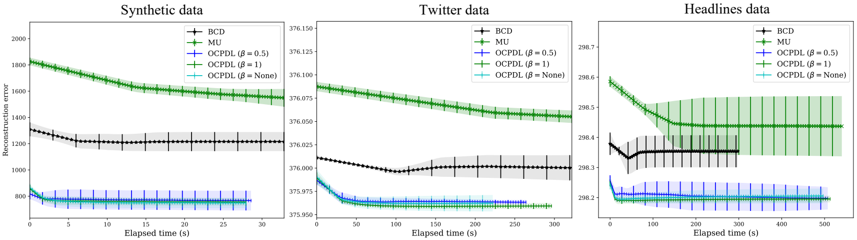
SMM+BCD with Diminishing Radius

- Online NTF algorithm (!) (Strohmeier, Lyu, Needell [7]):

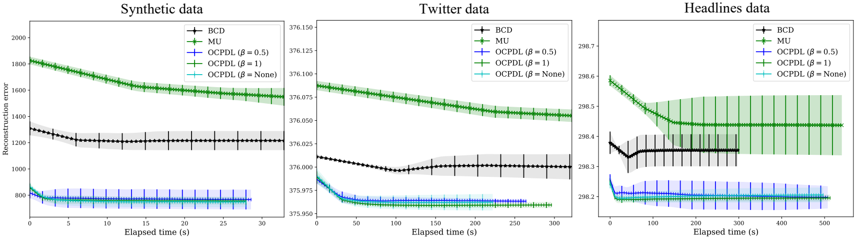
$$\begin{cases} U_t^{(3)} & \leftarrow \operatorname*{argmin}_{U^{(3)} \in \mathbb{R}_{\geq 0}^{d_3 \times r}} \left[\left\| X_t - \texttt{Out}(U_{t-1}^{(1)},\, U_{t-1}^{(2)},\, U^{(3)}) \right\|_F^2 \right] \qquad \textcolor{red}{(\textit{Convex})} \\ \hat{f}_t(U^{(1)}, U^{(2)}) & \leftarrow (1 - w_t)\hat{f}_{t-1}(U^{(1)}, U^{(2)}) \\ & \quad + w_t \left(\left\| X_t - \texttt{Out}(U^{(1)},\, U^{(2)},\, U_t^{(3)}) \right\|_F^2 \right) \qquad \textcolor{red}{(\text{Surrogate empirical loss})} \\ U_t^{(1)} & \leftarrow \operatorname*{argmin}_{U^{(1)} \in \mathbb{R}_{\geq 0}^{d_1 \times r},\ \textcolor{blue}{\| U^{(1)} - U_{t-1}^{(1)} \|_F \leq w_t}} \hat{f}_t(U^{(1)}, U_{t-1}^{(2)}) \qquad \textcolor{red}{(\textbf{\textit{Convex}})} \\ U_t^{(2)} & \leftarrow \operatorname*{argmin}_{U^{(2)} \in \mathbb{R}_{\geq 0}^{d_2 \times r},\ \textcolor{blue}{\| U^{(2)} - U_{t-1}^{(2)} \|_F \leq w_t}} \hat{f}_t(U_t^{(1)}, U^{(2)}) \qquad \textcolor{red}{(\textbf{\textit{Convex}})} \end{cases}$$

  - $(U_t^{(1)}, U_t^{(2)})$ still does not minimize $\hat{f}_t$
  - But we can get a.s. convergence to stationary points [7]
  - Rate of convergence: $O((\log n)/n^{1/4})$ (Upcoming paper)

- Comparison of Online NTF against BCD (ALS) and Multiplicative Update (MU) on offline NTF:



- Comparison of Online NTF against BCD (ALS) and Multiplicative Update (MU) on offline NTF:



- Online NTF outperforms both ALS and MU on all three datasets

Thanks!

[1] Luigi Grippo and Marco Sciandrone. "On the convergence of the block nonlinear Gauss–Seidel method under convex constraints". In: *Operations research letters* 26.3 (2000), pp. 127–136.

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

[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* 21.251 (2020), pp. 1–49.

[4] Julien Mairal. "Stochastic majorization-minimization algorithms for large-scale optimization". In: *Advances in Neural Information Processing Systems*. 2013, pp. 2283–2291.

[5] Julien Mairal et al. "Online learning for matrix factorization and sparse coding". In: *Journal of Machine Learning Research* 11.Jan (2010), pp. 19–60.

[6] Michael JD Powell. "On search directions for minimization algorithms". In: *Mathematical programming* 4.1 (1973), pp. 193–201.

[7] Christopher Strohmeier, Hanbaek Lyu, and Deanna Needell. "Online tensor factorization and CP-dictionary learning for Markovian data". In: *arXiv preprint arXiv:2009.07612* (2020).