Online Nonnegative CP-dictionary Learning for Markovian Data

Hanbaek Lyu

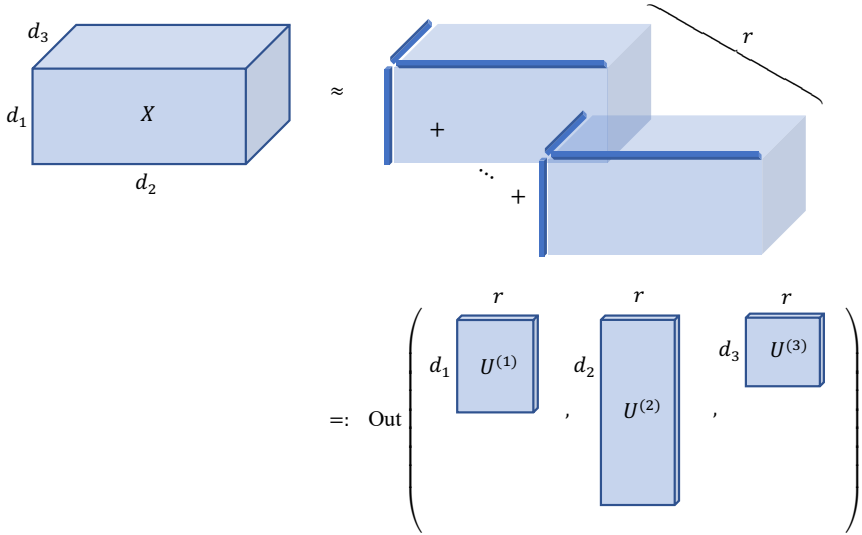
Department of Mathematics, IFDS
University of Wisconsin - Madison

Partially supported by NSF DMS #2206296 and #2010035

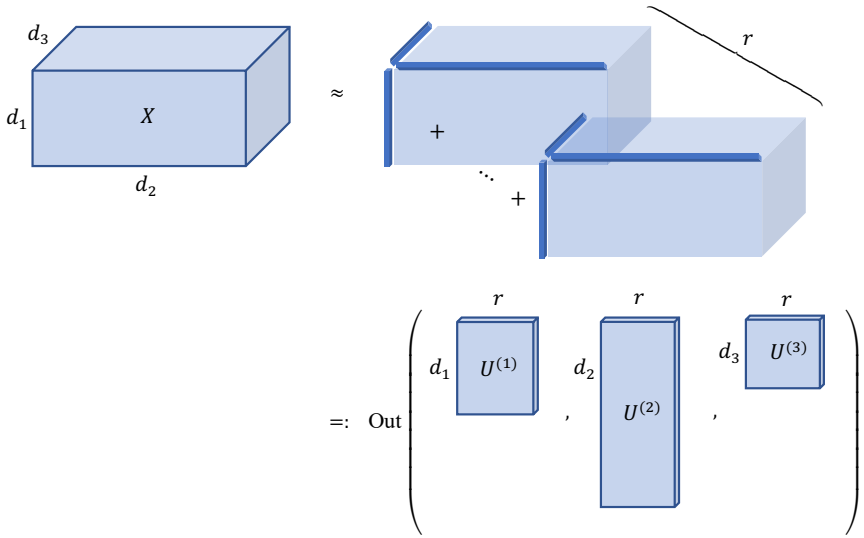
NeurIPS 2022

Joint work with Chris Strohmeier and Deanna Needell (JMLR '22)

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



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



- Nonnegative CP (CANDECOMP/PARAFAC) Decomposition

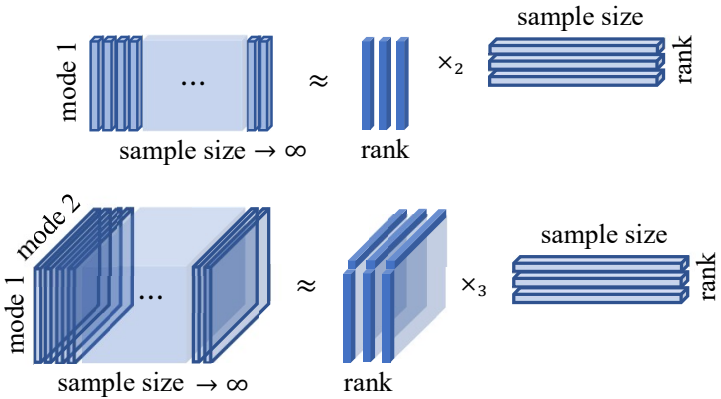
$$\underset{U^{(1)} \in \mathbb{R}_{\geq 0}^{d_1 \times r}, U^{(2)} \in \mathbb{R}_{\geq 0}^{d_2 \times r}, U^{(3)} \in \mathbb{R}_{\geq 0}^{d_3 \times r}}{\operatorname{argmin}} \| X - \mathrm{Out}(U^{(1)}, U^{(2)}, U^{(3)}) \|_F^2$$

- Online (Nonnegative) Matrix Factorization:

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

- Online (Nonnegative) CP-dictionary Learning:

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



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

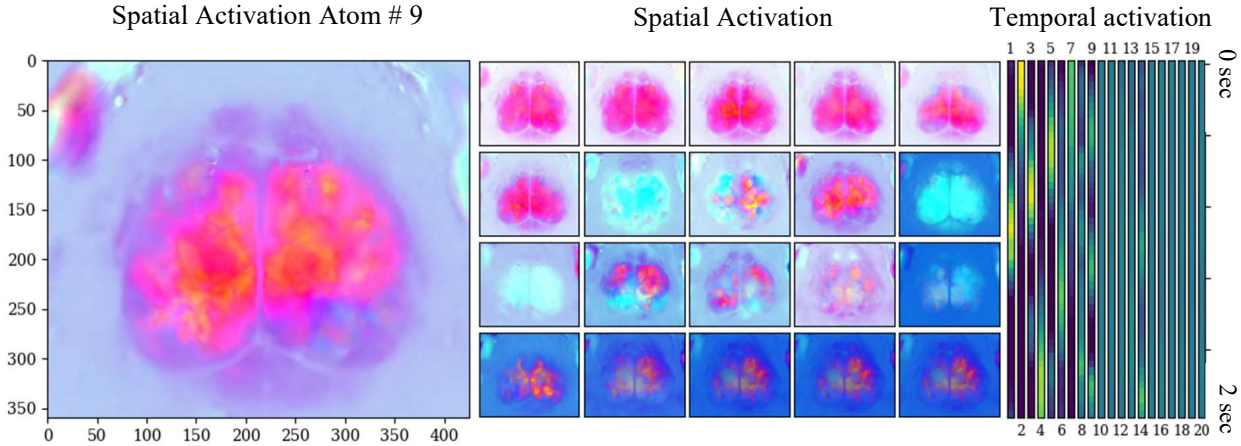
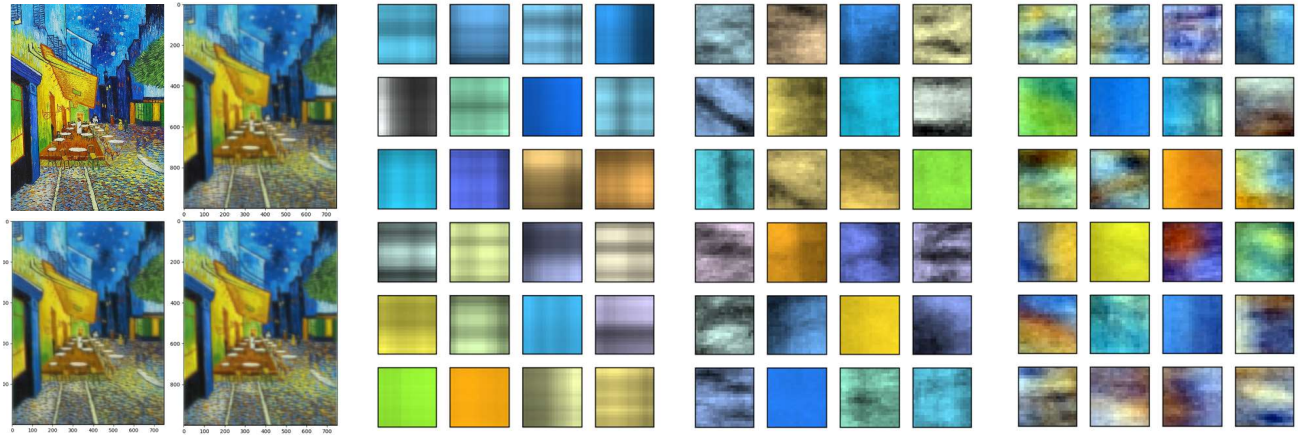


Figure: Temporal dictionary learned from mice brain activity video (Original data from Barson et al. *Nature methods* (2020))

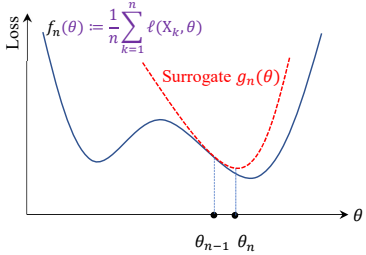
- Tensor reshaping + OnlinCPDL: Disentangle modes of choice



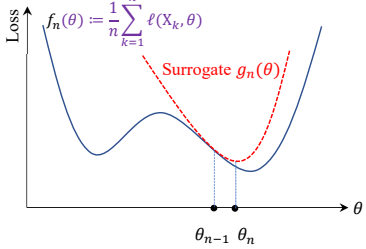
(a) (b) (c) (d)

Figure: Temporal dictionary learned from $N = 10^4$ patches of shape $20 \times 20 \times 3$ from a color image in top left of (a). Tensor reshaping: (b) $20 \times 20 \times 3 \times N$; (c) $20^2 \times 3 \times N$; (d) $20^2 \cdot 3 \times N$. Other subplots in (a) are reconstruction using CP-dictionaries in (b)-(c).

- Stochastic Majorization-Minimization (SMM) – Mairal [6]
  - Iteratively minimize majorizing surrogates g_n of the empirical loss f_n:



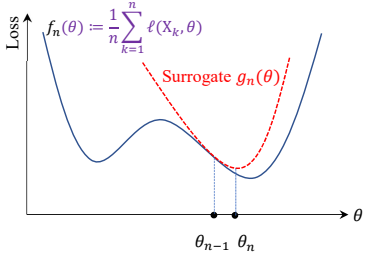
- Stochastic Majorization-Minimization (SMM) – Mairal [6]
  - Iteratively minimize majorizing surrogates g_n of the empirical loss f_n:



- Online NMF algorithm (Mairal, Bach, Saprio, Ponce [7]):

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

- Stochastic Majorization-Minimization (SMM) – Mairal [6]
  - Iteratively minimize majorizing surrogates g_n of the empirical loss f_n:

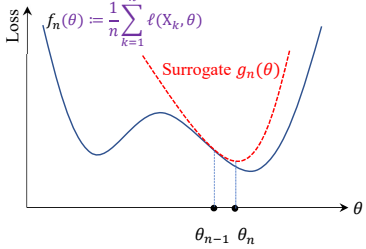


- Online NMF algorithm (Mairal, Bach, Saprio, Ponce [7]):

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

  - Algorithm converges to a stationary point almost surely

- Stochastic Majorization-Minimization (SMM) – Mairal [6]
  - Iteratively minimize majorizing surrogates g_n of the empirical loss f_n:



- Online NMF algorithm (Mairal, Bach, Saprio, Ponce [7]):

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

  - Algorithm converges to a stationary point almost surely
  - Convergence also holds when data sequence has Markovian dependence (L., Needell, Balzano [4])

- Online NTF algorithm (?):

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

- Online NTF algorithm (?):

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

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

- Online CP-dictionary learning alg. (this work) (L., Strohmeier, Needell [5]):

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

- Online CP-dictionary learning alg. (this work) (L., Strohmeier, Needell [5]):

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

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

Theorem (L. Strohmeier, Needell '22 [5])

Let ObsTensor$_t$ = function(X_t), X_t *a Markov chain* (*irreducible, aperiodic, countable states*) *with* $\|\pi - \pi(X_n|X_{n-r})\|_{TV} = O(r^{-\gamma})$ *for some* $\gamma > 0$. $\boldsymbol{\theta}_n$:=*output of OnlineCPDL*.
Θ = Set of constraints. *Under mild conditions,*

$$\boldsymbol{\theta}_n \to \{\text{stationary pts. of the expected loss over } \Theta\} \quad \text{a.s. as } n \to \infty$$

Theorem (L. Strohmeier, Needell '22 [5])

Let ObsTensor$_t$ = function(X_t), X_t *a Markov chain* (*irreducible, aperiodic, countable states*) *with* $\|\pi - \pi(X_n|X_{n-r})\|_{TV} = O(r^{-\gamma})$ *for some* $\gamma > 0$. $\boldsymbol{\theta}_n$:=*output of OnlineCPDL.* Θ = Set of constraints. *Under mild conditions,*

$$\boldsymbol{\theta}_n \to \{\text{stationary pts. of the expected loss over } \Theta\} \quad \text{a.s. as } n \to \infty$$

- Special case: ObsTensor$_t$ are i.i.d.. Convergence of online CP-decomposition even under this case was not known before

Theorem (L. Strohmeier, Needell '22 [5])

Let ObsTensor$_t$ = function(X_t), X_t *a Markov chain (irreducible, aperiodic, countable states) with* $\|\pi - \pi(X_n|X_{n-r})\|_{TV} = O(r^{-\gamma})$ *for some* $\gamma > 0$. $\boldsymbol{\theta}_n$:=*output of OnlineCPDL.*
Θ = Set of constraints. *Under mild conditions,*

$$\boldsymbol{\theta}_n \to \{\text{stationary pts. of the expected loss over } \Theta\} \quad \text{a.s. as } n \to \infty$$

- Special case: ObsTensor$_t$ are i.i.d.. Convergence of online CP-decomposition even under this case was not known before
- Why convergence guaratee under Markovian setting is useful?
  - One may not have direct access to the unknown data distribution π but Indirectly by MCMC algorithms
  - Recent interest in stochastic optimization for Markovian data from reinforcement learning context

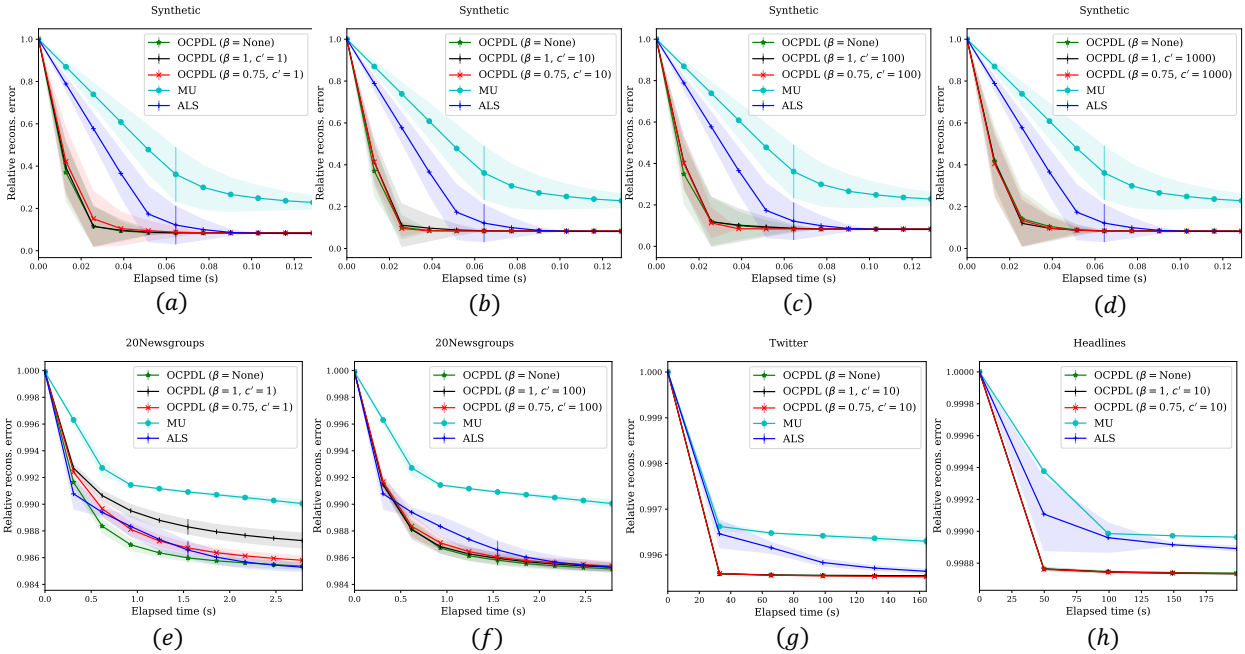
Theorem (L. Strohmeier, Needell '22 [5])

Let ObsTensor$_t$ = function(X_t), X_t *a Markov chain (irreducible, aperiodic, countable states) with* $\|\pi - \pi(X_n|X_{n-r})\|_{TV} = O(r^{-\gamma})$ *for some* $\gamma > 0$. $\boldsymbol{\theta}_n$:=*output of OnlineCPDL.* Θ = Set of constraints. *Under mild conditions,*

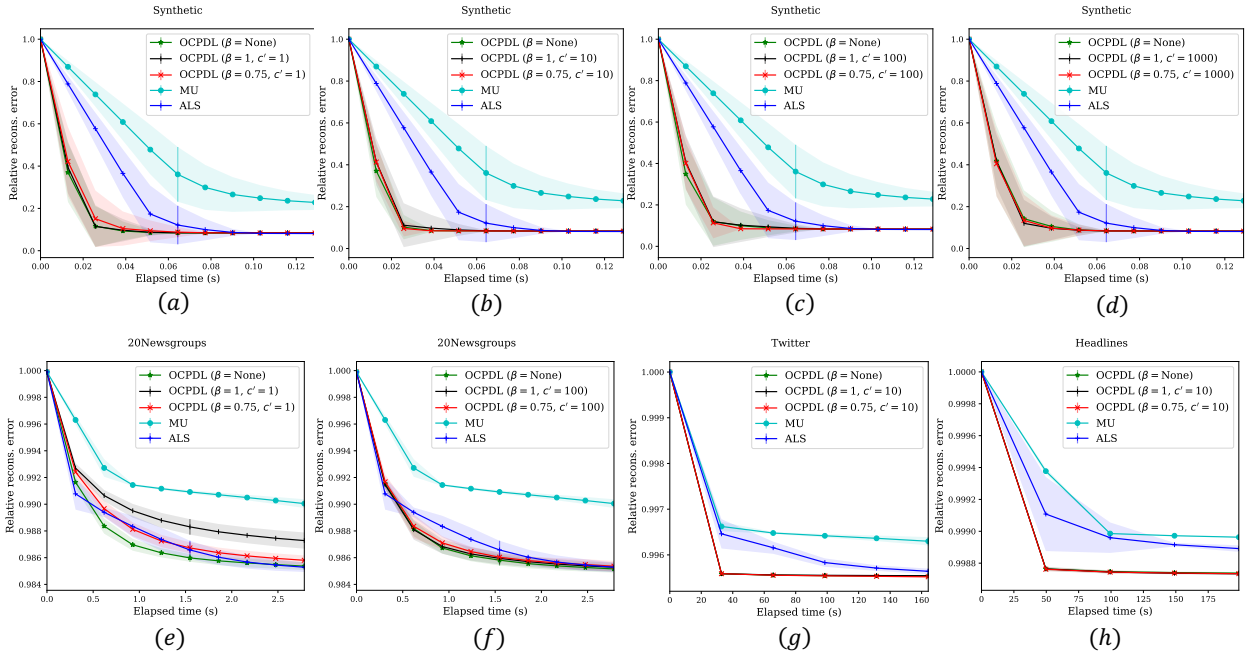
$$\boldsymbol{\theta}_n \to \{\text{stationary pts. of the expected loss over } \Theta\} \quad \text{a.s. as } n \to \infty$$

- Special case: ObsTensor$_t$ are i.i.d.. Convergence of online CP-decomposition even under this case was not known before
- Why convergence guaratee under Markovian setting is useful?
  - One may not have direct access to the unknown data distribution π but Indirectly by MCMC algorithms
  - Recent interest in stochastic optimization for Markovian data from reinforcement learning context
- More recent result: Rate of convergence $O((\log n)/n^{1/4})$ w.r.t. expected loss, $O((\log n)/n^{1/2})$ w.r.t. empiricla loss (L. [3])

- 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 CPDL converges faster than both ALS and MU on all 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. "Stochastic regularized majorization-minimization with weakly convex and multi-convex surrogates". In: *arXiv preprint arXiv:2201.01652* (2022).

[4] Hanbaek Lyu, Deanna Needell, and Laura Balzano. "Online matrix factorization for Markovian data and applications to network dictionary learning". In: *Journal of Machine Learning Research* 21.251 (2020), pp. 1–49.

[5] Hanbaek Lyu, Christopher Strohmeier, and Deanna Needell. "Online nonnegative CP-dictionary learning for Markovian data". In: *Journal of Machine Learning Research* 23.148 (2022), pp. 1–50.

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

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

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

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