Non-Negative Matrix Factorization Meets Time-Inhomogeneous Markov Chains

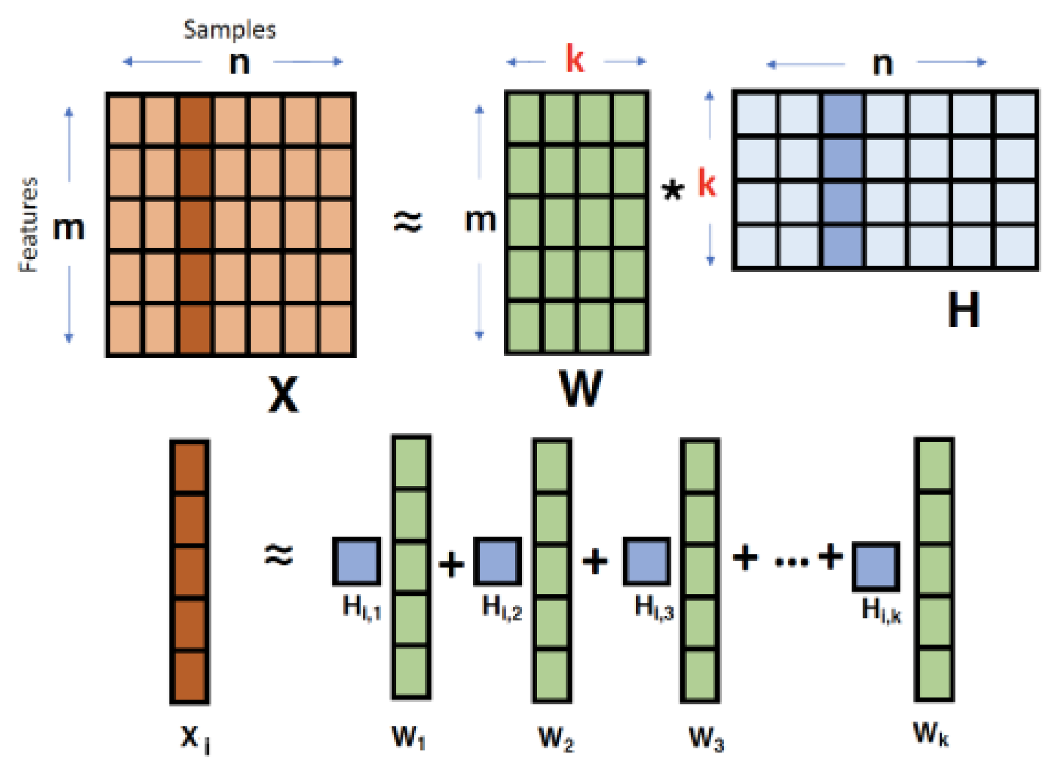
Ievgen Redko Marc Sebban Amaury Habrard

Université de Lyon, UJM-Saint-Etienne, CNRS, Institut d'Optique Graduate School
Laboratoire Hubert Curien UMR 5516, F-42023, Saint-Etienne, France

Non-negative matrix factorization 101

A **standard NMF** [3] is represented as the following optimization problem:

$$\min_{\mathbf{W},\mathbf{H}\geq 0} J(\mathbf{W},\mathbf{H}) = \min_{\mathbf{W},\mathbf{H}\geq 0} \|\mathbf{X}-\mathbf{W}\mathbf{H}\|_F^2. \quad (1)$$



Source: Nebgen et al. '20

Commonly optimized using **multiplicative update rules** (MURs):

$$\mathbf{W}^{(i+1)} = \mathbf{W}^{(i)} \circ \frac{\mathbf{X}\mathbf{H}^T}{\mathbf{W}^{(i)}\mathbf{H}\mathbf{H}^T}, \quad (2)$$

$$\mathbf{H}^{(i+1)} = \mathbf{H}^{(i)} \circ \frac{\mathbf{W}^T\mathbf{X}}{\mathbf{W}^T\mathbf{W}\mathbf{H}^{(i)}}, \quad (3)$$

✓ Guaranteed to **not increase** the objective function

✓ Due to their simplicity **widely used** by practitioners

but ...

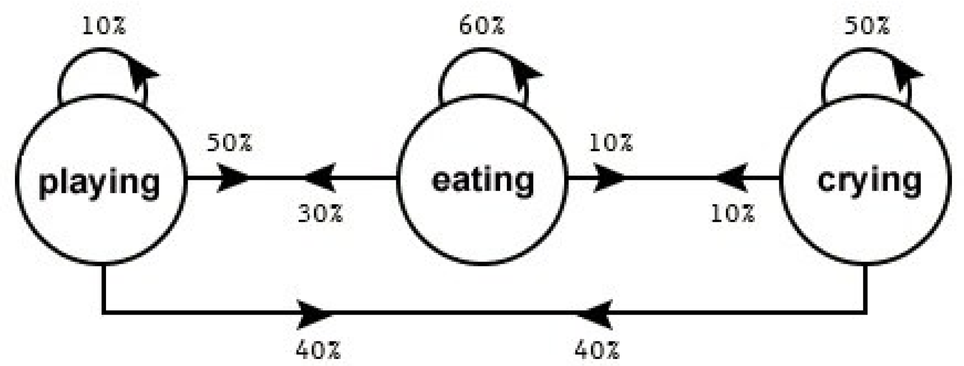
× MUR **may fail to converge** to a **local minimum** or a **stationary point** [4]

× Solution **is not unique** (without additional constraints)

Markov chains 101

- Set of **states** $S = \{s_1, s_2, \dots, s_c\}$ + **transition matrix** $\{\mathbf{P}(S_i,S_j)\}_{i,j=1}^c = P(S_j|S_i)$
- **Homogeneous** (static) Markov chain: $\mathbf{x}_t = \mathbf{x}_0\mathbf{P}^t$.
- **Heterogeneous** (dynamic) chain: $\mathbf{x}_t = \mathbf{x}_0\prod_{i=1}^t \mathbf{P}_i$, with $\mathbf{x}_{i+1} = \mathbf{x}_i\mathbf{P}_i$.

Markov state diagram of a child behaviour



Source: Luis Fok on Quora

Motivating example

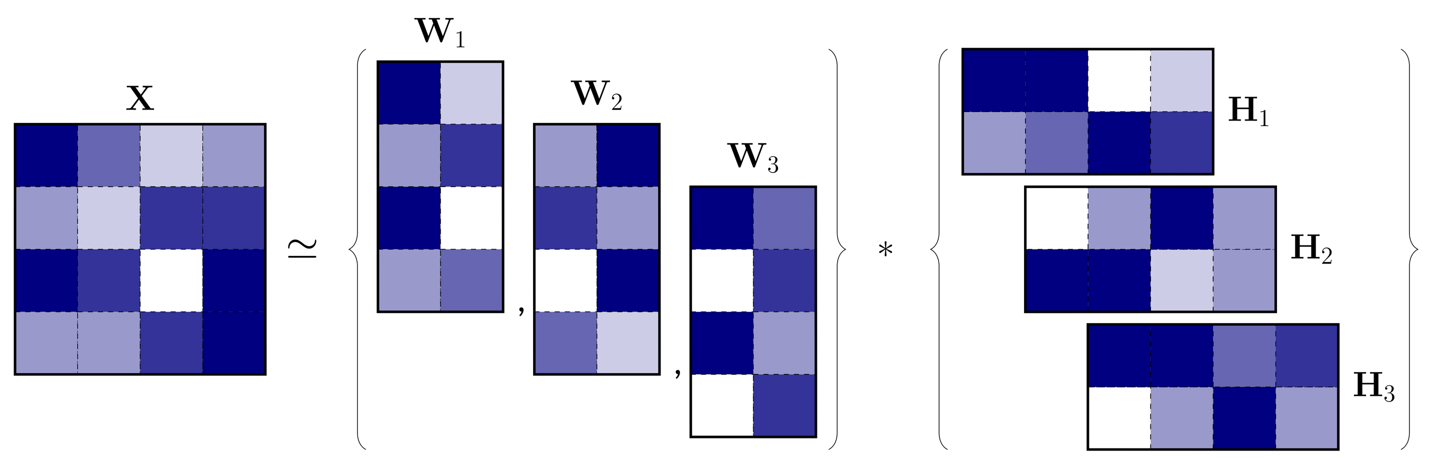


Figure 1. Illustration of 3 different solutions obtained using MUR with random initializations of $\mathbf{W}$ and $\mathbf{H}$.

Main results

Step 1. Show that $\{\mathbf{W}^{(i)}\}_{i=0}^\infty$ obtained with MUR can be **generated with at least one** time-heterogeneous Markov chain

Theorem

Consider the NMF problem given in Eq.1 and MUR updates given in Eq.2. Then, matrices $\{\mathbf{W}^{(i)}\}_{i=0}^\infty$ can be generated by a time inhomogeneous Markov chain with transition matrices $\{\mathbf{S}_i\}_{i=0}^\infty$ similar to Soules matrices with eigenvalues given by vec $\left(\mathbf{X}\mathbf{H}^T/\mathbf{W}^{(i)}\mathbf{H}\mathbf{H}^T\right)$.

× Can build **multiple Markov chains** generating the **same sequence** $\{\mathbf{W}^{(i)}\}_{i=0}^\infty$

× Markov chains can converge to **different stationary distributions**

Step 2. Use convergence theorems for time-heterogeneous Markov chain to understand when it **converges to the same solution regardless the initialization**

Theorem

MUR converge to the same solution regardless the initial initialization when one of the following conditions are verified:

1. *There is only one $\mathbf{S}_i$ at each iteration*
2. *Different $\mathbf{S}_i$ have the same stationary distribution.*

× Hard to ensure uniqueness (need to solve an open algebraic problem)

Step 3. Analyze what is the **speed of convergence** based on the properties of the transition matrix

Corollary

Let $\sigma_1(\mathbf{S}_i)$ to be the second largest singular value of $\mathbf{S}_i$. Then, we have that $\|\prod_{i=1}^n \mathbf{S}_i(j,\cdot) - \pi^*\|_2 \leq (\pi^*(j)-1)^{\frac{1}{2}}\prod_{i=1}^n \sigma_1(\mathbf{S}_i)$.

✓ Surprising dependence of the **speed of convergence** of MUR on the product of **the second largest singular values** of $\mathbf{S}_i$

Experimental results

- **Test 1**: synthetic example with **unique factorization**
- **Test 2**: mixture of Gaussian distributions
- **Baselines**: **random** initialization, **unique** factorization with **NNDSVD** [1] and **Gillis method** [2]

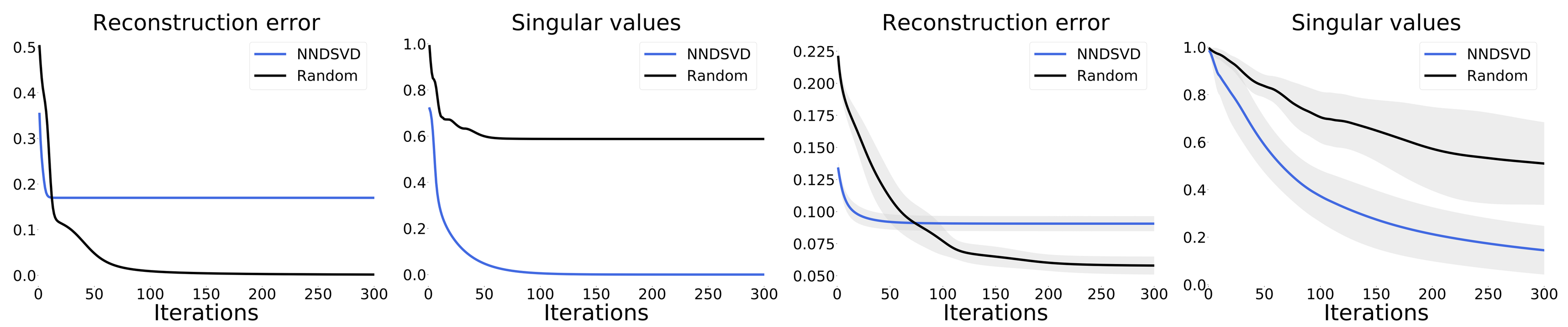


Figure 2. Results obtained with NNDSVD compared to random initialization: **(left)** reconstruction error and **(middle left)** product of second singular values of the transition matrices on the data admitting a unique factorization; **(middle right)** reconstruction error and **(right)** product of second singular values of the transition matrices on the mixture of 5 isotropic Gaussian distributions with $n = 20000$, $k = 5$ and $d \in \{10, \dots, 100\}$.

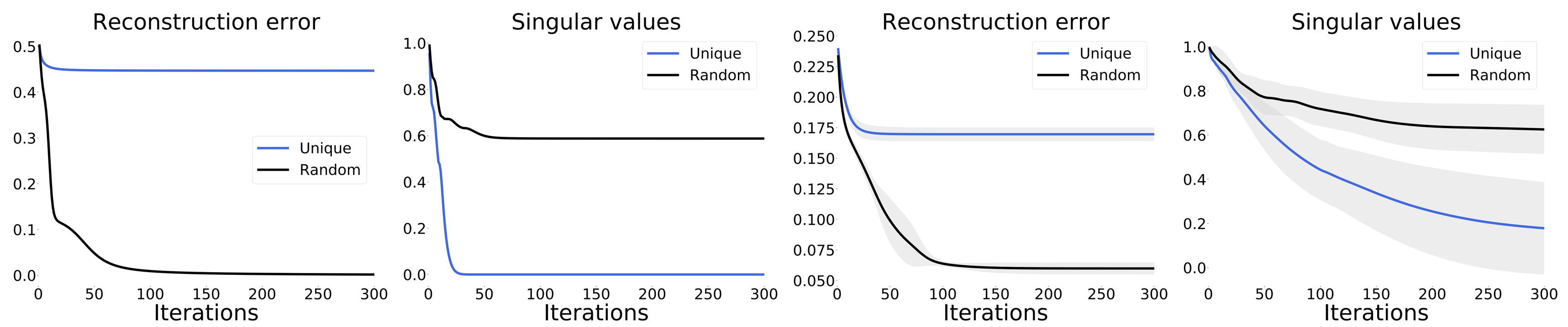


Figure 3. Results obtained with Gillis' pre-processing presented in the same order as above with $n = 200$ for the case of the mixture of Gaussian distributions. For both cases, the variance (shaded area) around the mean curve over varying d is represented only for the case of the mixture of Gaussians as for the unique factorization all the parameters remain fixed. The code to reproduce the two figures is given in the Supplementary material.

✓ **Theory is useful in practice** to asses **convergence speed** based on the second largest **singular values** of transition matrices

Conclusion

✓ Explaining the **lack of uniqueness** in NMF with MUR trough **equivalence to Markov chains**

✓ **Forcing uniqueness** with MUR requires solving an **open algebraic problem**

✓ **Convergence speed** depends on the **second largest eigenvalues** of the transition matrices: **confirmed in practice** for several methods and datasets

References

[1] C. Boutsidis and E. Gallopoulos. Svd based initialization: A head start for nonnegative matrix factorization. *Pattern Recogn.*, 41(4):1350–1362, 2008.

[2] Nicolas Gillis. Sparse and unique NMF through data preprocessing. *J. Mach. Learn. Res.*, 13(1):3349–3386, 2012.

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

[4] Chih-Jen Lin. Projected gradient methods for NMF. *Neural Comput.*, 19(10):2756–2779, 2007.